UNIONE EUROPEA
Fondo sociale europeo

Repubblica Italiana

REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA

FSE 2007-2013
POR SARDEGNA
obiettivo competitività regionale e occupazione

Fondazione di Sardegna

uniss
UNIVERSITÀ DEGLI STUDI DI SASSARI

UNIVERSITÀ DEGLI STUDI DI SASSARI

CORSO DI DOTTORATO DI RICERCA
Scienze Agrarie

Curriculum
Scienze e Tecnologie Zootecniche

Ciclo XXX

Use of multivariate discriminant methodologies in the analysis of phenotypic and genomic data of cattle

dr.ssa Elisabetta Manca

Coordinatore del Corso	Prof. Antonello Cannas
Referente di Curriculum	Dr. Gianni Battacone
Docente Guida	Dr. Corrado Dimauro

Anno Accademico 2016- 2017

UNIONE EUROPEA
Fondo sociale europeo

Repubblica Italiana

REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA

FSE 2007-2013
POR SARDEGNA
obiettivo competitività regionale e occupazione

Fondazione di Sardegna

uniss
UNIVERSITÀ DEGLI STUDI DI SASSARI

UNIVERSITÀ DEGLI STUDI DI SASSARI

CORSO DI DOTTORATO DI RICERCA
Scienze Agrarie

Curriculum
Scienze e Tecnologie Zootecniche

Ciclo XXX

Use of multivariate discriminant methodologies in the analysis of phenotypic and genomic data of cattle

dr.ssa Elisabetta Manca

Coordinatore del Corso	Prof. Antonello Cannas
Referente di Curriculum	Dr. Gianni Battacone
Docente Guida	Dr. Corrado Dimauro

Anno Accademico 2016-2017

Università degli Studi di Sassari

Dipartimento di Agraria

Dottorato di ricerca in Scienze Agrarie

Curriculum

Scienze e Tecnologie Zootecniche

XXX CICLO

La presente tesi è stata prodotta durante la frequenza del corso di dottorato in Scienze Agrarie dell'Università degli Studi di Sassari, anno accademico. 2016/2017 – XXX ciclo, con il sostegno di una borsa di studio cofinanziata con le risorse del **P.O.R. SARDEGNA F.S.E. 2007-2013** - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività l.3.1 "*Finanziamento di corsi di dottorato finalizzati alla formazione di capitale umano altamente specializzato, in particolare per i settori dell'ICT, delle nanotecnologie e delle biotecnologie, dell'energia e dello sviluppo sostenibile, dell'agroalimentare e dei materiali tradizionali*".

La tesi è stata prodotta, altresì, grazie al contributo della Fondazione di Sardegna.

Anno Accademico 2016-2017

A mio padre e mia madre

con amore e gratitudine

ACKNOWLEDGEMENTS

I would like to thank all members of Animal Science Section (Department of Agricultural Science, University of Sassari, Italy), professors, researchers and technicians, for their scientific and human support.

I would like to express my special thanks to my supervisor Dr. Corrado Dimauro for huge and constant human support, teachings and help through scientific advice and full availability.

I would to thank to Prof. Nicolò Pietro Paolo Macciotta, Dr. Giustino Gaspa, Dr. Silvia Sorbolini, Dr. Alberto Stanislao Atzori and Dr. Massimo Cellesi for their precious help during my experimental work.

I want to sincerely thank my colleagues for their essential help, collaboration, human sustain and for their sincere friendship.

Thanks to Dr. Attilio Rossoni and Dr. Enrico Santus (ANARB, Bussolengo, Italy), Dr. Enrico Costa (Società Bonifiche Sarde, Arborea, Italy) and Dr. Sergio Beltrame (CHJAVIZA Kft, Tiszaalpar, Hungary) for their essential collaboration and for data provided.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

GENERAL ABSTRACT

The present thesis deals with different application of multivariate discriminant procedures both in the analysis of phenotypic and genomic data. This dissertation is organized in 4 main chapters.

The Chapter 1 is the general introduction and essentially regards the use of the multivariate statistical techniques in animal science, with a particular emphasis on the discriminant analysis. This technique, specifically conceived to classify different observations in already existent groups, become very useful when classification is developed by using characters that singularly are not able to classify observations.

In Chapter 2, a new statistical method called Discriminant Association Method (DAM) was proposed. Data used in the present research were previously analyzed by Sorbolini et al. (2016) who carried out an ordinary GWAS on seven growth, carcass and meat quality phenotypes. Involved animals were 409 young Marchigiana bulls genotyped with the Illumina's 50K BeadChip. The DAM approach, developed by using multivariate statistical techniques, overcomes most of problems that affect the single SNP regression technique used in the ordinary GWAS. The DAM was able to highlight the associations reported by Sorbolini et al. (2016) and to propose new associated markers often related to interesting genes.

In Chapter 3, a new index to evaluate feed efficiency was defined: the residual concentrate intake (RCI). The RCI identifies efficient and inefficient bovines in converting the concentrate. Unlike the residual feed intake (RFI), the RCI is quite simple to measure and therefore it could be easily included in genomic breeding programs. A useful contribute

to breeding programs that include RCI could be offered by the detection of genomic regions and of candidate genes which regulate RCI. In the present research, in addition to the ordinary single SNP regression approach, the DAM method (previously explained in chapter II of this dissertation) was applied to develop a GWAS for selecting markers associated to RCI.

The research reported in Chapter 4 was aimed to develop an algorithm able to early identify dairy cows that, having a persistent lactation, might be destined to have a long lactation. Four different lactation curve models (Wood, Ali & Schaeffer, Legendre Polynomials and 4th Degree Polynomials) were fitted to individual lactations by using the first 90, 120 and 150 DIM (days in milking). Estimated regression parameters were used to develop two multivariate techniques: the canonical discriminant analysis (CDA) and the discriminant analysis (DA). The proposed algorithm combines the talent of curve models in depict features of the lactation and the ability of multivariate statistical techniques in distinguishing differences between groups. In this case, groups consisted of lactations with low (LC) and high (HC) persistency. Only milk production data recorded in early lactation (not more than 150 DIM) was used in all analyses. The algorithm developed could help farmers to early select a quota of their herd to be destined to a long lactation.

CHAPTER 1

INTRODUCTION

1.1 Multivariate Statistical Analysis

The main objective of univariate statistical analysis is to decompose the variance of a dependent character (y) in its components. The total variance of y is given by:

$$\partial^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{\mathrm{n} - 1}$$

The analysis of variance (ANOVA) is a statistical technique used to test differences between two or more means by decomposing the total variance into several components depending on one or more factors of variation. These factors can be both categorical or continuous variables. Despite several characters can be involved in the ANOVA, this technique is however numbered among the univariate statistical analyses because the object of the study is the variability of the y variable.

The bivariate statistical analysis is instead focused on two variables, x and y, that are analyzed simultaneously by decomposing their covariance defined as:

$$\partial_{xy} = \frac{\sum_i^n (x_i - \bar{x})\,(y_i - \bar{y})}{\mathrm{n} - 2}$$

The covariance is analogue to the variance in the univariate approach. The variance describes the variability of a certain character y, whereas the covariance explains how much the variability of y is influenced by the variability of x.

The variance is always positive *(∂^2>0),* while the covariance can be positive or negative. If, on average, when x increases also y increases then $\partial_{xy} > 0$. On the contrary, if, on average, when x increases y decreases then $\partial_{xy} < 0$.

Another index used to evaluate the relationship between two characters is the Pearson linear correlation (Person 1896):

$$\rho_{xy} = \frac{\partial_{xy}}{\partial_{xx}\partial_{yy}}$$

where ∂_{xy} is the covariance, ∂_{xx} and ∂_{yy} are the standard deviations (i.e. the square root of the variance) of the two variables. The correlation is able to evaluate and understand the linear links between two continuous variables (Mukaka, 2012). It assumes values ranging from -1 (perfect negative correlation), 0 (when there is not a correlation) to 1 (perfect positive correlation) (Mukaka, 2012). Table 1 shows the different degrees of correlation between two variables.

Table 1. General interpretation of correlation values (Mukaka 2012)

Range (negative or positive)		
From	**To**	**Interpretation (positive or negative)**
± 0.9	± 1.0	Very High Correlation
± 0.7	± 0.9	High Correlation
± 0.5	± 0.7	Moderate correlation
± 0.3	± 0.5	Low Correlation
0	± 0.3	Negligible Correlation

However, it is important to note that the interpretation of the correlation coefficient may change depending on the involved characters; hence, the interpretations in the table may vary. A statistical *t*-test can be developed to test if the two variables are significantly related or not (i.e. if $\rho\neq 0$ or not).

H_0: $\rho=0$

H_a: $\rho \neq 0$

The test is based on the *t-statistics* with *n*-2 degree of freedom (d. f.):

$$t = \frac{r\sqrt{n}-2}{\sqrt{1}-r^2}$$

where *r* is the value of the correlation and *n* is the number of involved data.

Multivariate analysis consists of a collection of methods that can be used when several characters are observed on the same individual or object.

This approach allows to gain information besides the study of the single variable. In the multivariate analysis, variables are simultaneously analyzed to highlight dependences among them (Ricci 2003). The main objectives of all multivariate techniques are both the synthesis of the data and the study of mutual relation among the different variables. As in univariate and bivariate analysis, the starting point of multivariate techniques is the variability, i.e. the matrix of variance and covariance of data:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \ldots\ldots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \ldots\ldots & \sigma_{2p} \\ \vdots & \vdots & \ldots\ldots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \ldots\ldots & \sigma_{pp} \end{bmatrix}$$

In this matrix σ_{ii} is the variance of the variable *i* and σ_{ij} is the covariance between the *i*th and *j*th variables.

The main multivariate techniques are:

1. Principal Component Analysis;

2. Multivariate Factor Analysis;
3. Partial Lest Squares Regression;
4. Cluster Analysis;
5. Discriminant Analysis.

In the present thesis, we focused our attention on the principal component analysis and the discriminant procedure. These two techniques were described in details.

1.2 Principal Components Analysis

Principal component analysis (PCA) is the most popular multivariate statistical technique and it has been widely exploited by almost all scientific disciplines. It is also the oldest among the multivariate techniques that are currently used. Kearl Pearson, in 1901, was the first to develop this technique. It was, however, formalized in its modern instantiation by Hotelling (1933) who also coined the term *Principal Components*. PCA is a statistical technique whose main objective is the reduction of the space variables. The basic idea of PCA is very simple and its development involves the matrix algebra. Suppose we have a set of *n*-variables, $x_1, x_2, ..., x_n$, measured on m objects. If we have only two characters, the profile of involved objects can be visualized by the scatter plot of x_1 v.s. x_2, as displayed in Figure 1a where each point represents one object under study.

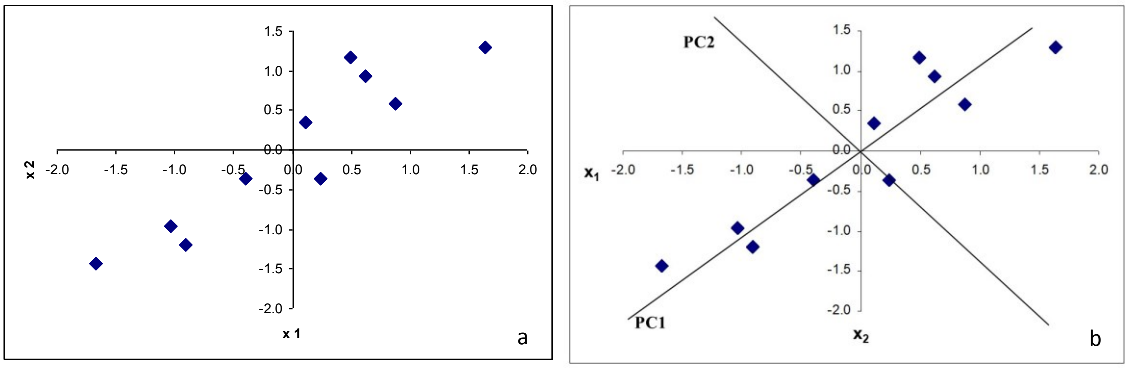

Figure 1. Scatterplot of (a) the original x_1 and x_2 variables and of (b) the PC1 and PC2

As generally happens, data-points in the graph are arranged on the direction of the maximum variation. In a multivariate space variable, with a number of variables *n* >3, points group as an ellipsoid whose axes represent the directions of maximum variation. PCA consists in a rotation of axes on the directions of the maximum variability. In the bivariate space (Figure 1b), PC1 and PC2 are the new rotated axes. The rotated variables, i.e. the principal components, can be obtained by solving an eigenvalues problem applied to the variance and covariance matrix of data. This algebraic procedure extracts new orthogonal axes, also called eigenvectors, whose direction clashes with the maximum variation of data. The total variation of data is therefore reallocated along the new directions and is given by the eigenvalues that are extracted as eigenvectors are obtained. The PC1, usually, summarizes most of the variability, the PC2 a lower value and so on with the others PCs. The number of PCs that are retained depends on the cumulative explained variation. Usually the procedure stops when the extracted components show a cumulative explained variation around 80-85%. The consequence is a drastic reduction of the number of variables. Scores of objects in the rotated axes are obtained as linear combinations of the original variables:

$$\mathrm{PC1} = \alpha_1 x_1 + \ldots + \alpha_n x_n$$

where α_i are the loadings of each extracted eigenvector (Macciotta et al., 2010).

Actually, PCA is often used to solve algebraic problems in developing more sophisticated multivariate techniques as principal component regression, factorial analysis or discriminant procedures.

1.3 Canonical Discriminant Analysis

The multivariate discriminant techniques were first formulated by Ronald A. Fisher in 1936. He applied those procedures starting from a dataset of fifty Iris flowers that belong to three different species: two (*I. setosa* and *I. versicolor*) coming from the same colony and one (*I. virginica*) coming from another colony. The considered characters were the sepal length, petal length and petal width. Singularly, the three variables were not able to separate the three groups. When they were analyzed simultaneously, in a discriminant analysis, the three groups were well highlighted, as showed in Figure 2.

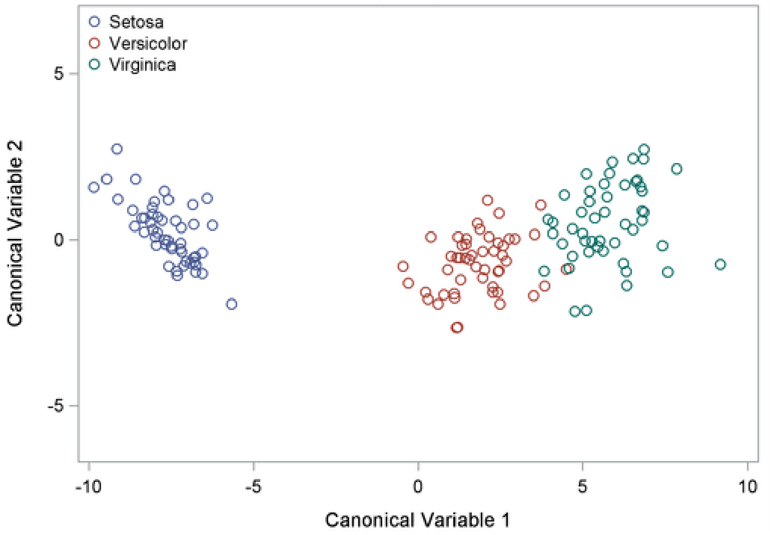

Figure 2. Fisher 1936, Iris data: Plot of Canonical Variables (www.supportsas.com)

The canonical discriminant analysis (CDA) is a dimension-reduction technique that is related to principal component analysis and canonical correlation. Given a classification variable and several interval variables, CDA derives a set of new variables, called canonical functions (CAN) that are linear combinations of the original interval variables. As in PCA, CANs are obtained by solving an eigenvalues problem. The substantial difference between the two multivariate techniques is that PCs summarize the total variation in the data, whereas CANs summarize the between-groups variation. PCA analyzes the variance-covariance matrix of data to rotate axes in the direction of the maximum variation. CDA analyzes a different variance-covariance matrix obtained by the combination (the ratio) of the between-groups and within-groups variance-covariance matrices. The new axes extracted by CDA highlight differences between groups better then PCs.

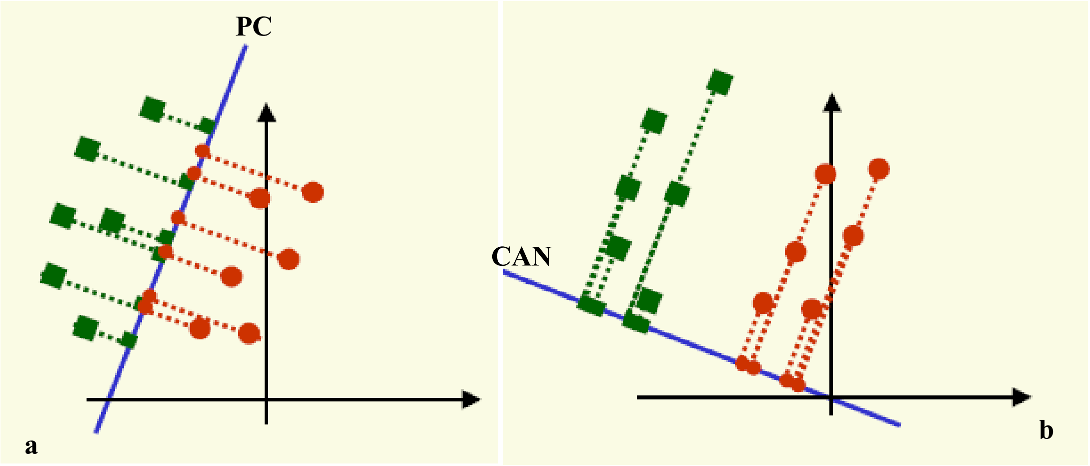

Figure 3. Differences between new axes extracted by PCA (a) and CAN (b)

Figure 3 displays these differences. Suppose we have two groups of objects in a bidimensional space variable. When objects are projected along the PC, the two groups appears are not separated (Figure 3a). On the contrary, when the CAN is extracted, the two groups are perfectly separated (Figure 3b). If *k*-groups are involved in the study, CDA derives *k* – 1 CANs, each one accounting for a decreasing quota of the between- groups variation. As in PCA, the procedure stops when the variation explained by CANs is around 80-85%. The distance among groups can be measured through the Mahalanobis distance (De Maesschalck et al., 2000). Finally, the effective separation between groups can be tested by using the Hotelling's t-square test (1933). This test, however, can be developed only if the (co)variance matrix is not singular. In a multivariate dataset, with objects in the rows and variables in the columns, the number of columns would be lower than the number of rows to obtain a full rank variance and covariance matrix. If this does not happen, the number of involved variables should be reduced (or the number of objects should be enlarged). Stepwise algorithms can help to solve the problem. Finally, CANs are also used to develop a discriminant criterion to classify observations into one of the involved groups. In practice, CANs are applied to each object and a discriminant score is produced. An individual is assigned to a particular group if its discriminant score is lower than the cutoff value obtained by calculating the weighted mean distance among group centroids (Mardia et al., 2000).

1.4 Stepwise Discriminant Analysis

The stepwise discriminant analysis (SDA) is a multivariate technique specifically conceived to reduce the number of variables involved in the CDA. The rationale behind SDA is similar to those in the stepwise regression. The objective of a regression is to predict, for each involved individual, values of an unknown continuous variable using one or more known continuous characters. The objective of CDA is to predict the group membership of involved individuals using one or more continuous variables. When a dataset presents a high number of variables, it is possible that some of them are not essential to predict the unknown variable or to assign objects. Moreover, some of those useless variables could ruin the analysis. To avoid this problem the stepwise technique is often applied. It reduces the number of variables erasing those that are not-informational, i.e. that do not add useful information to predict the unknown variable or to assign objects to the true group. SDA can be developed through three different algorithms:

-*Forward stepwise selection*: where variables are included into the model one at time and those that do not improve the model are not considered;

-*Backward stepwise selection*: which starts from the complete dataset and remove from the model useless variables;

-*Bidirectional stepwise selection*: that is a combination of the first two procedures.

1.5 Use of discriminant techniques in animal science

The use of discriminant techniques is becoming very popular in animal science, from animal genetics to food quality.

Herrera at al. (1996) applied the stepwise discriminant analysis (SDA), the canonical discriminant analysis (CDA) and the discriminant analysis (DA) to some zoometrical variables (withers height, chest depth, body length, shoulder point width, rump length and width, head length and width, chest girth, and shank circumference) of five Andalusian goat breeds (Malaga, Granada, Florida, Andalusian white and black breeds), to test their discriminating power. The DA was applied to estimate the probability to assign each animal to its breed of origin by using the considered variables. This study showed that zoometric measures as head length, shin circumference and rump length could be used as discriminant characters in differentiating these goat breeds, instead chest girth, chest depth and rump width are traits with low discriminate power. Furthermore, values of Mahalanobis distance showed that most different breeds were Florida and Malaga, whereas, Granada and Malaga were the most similar breeds.

Dossa et al., (2007) used the DA to discriminate four groups of goats raised in four different vegetation zones from South to North Benin (Africa) by using morphological characters. The best discriminant model used in this study has identified only five best discriminant variables on 12 considered (height at withers, neck length, rump height, tail length and auricular index). Mahalanobis distances among vegetation groups were significant and the two discriminant functions obtained by CDA correctly classified the 76.6% of animals to correct zone, showing that vegetation zone influence the goat ecotype.

Yakubu et al. (2010) used discriminant procedures on fifteen morphometric traits of two different Nigerian breeds of goat (West African Dwarf and Red Sokoto, both males and females). CDA selected seven most discriminant traits that were able to allocate, in average, the 99.7% of animals to respective breed (99.4% of West African Dwarf and 100% of Red Sokoto).

Several authors used DA to identify adulteration of products in dairy sector. For example, Gutiérrez et al. (2009) applied the DA on different triacylglycerol profile to distinguish between milk fat and other fats (not-milk fat in proportions of 5, 10, 15 and 20%). The discriminant procedures were able to discriminate 94.4 % of samples with level of adulteration <10%.

Dias et al. (2009) proposed a simple and economical procedure to ascertain whether a sample of goat milk is adulterated with adjunct of bovine milk. Authors developed an electronic system with 36 cross-sensibility sensors able to recognize the five different basic tastes. This system was applied on different raw skimmed milk samples of goat, cow and goat/cow. After the space variables was reduced by using the PCA, a linear discriminant model has been developed to obtain a differentiation among the different proportions of cow and goat milk. Errors in assigning some samples to goat and cow milks were probably due to the small number of the samples analyzed (19 and 16 respectively), compared to the large number of samples of goat/cow milks (142). Results of this work showed that this new electronic procedure, together with linear discriminant analysis, could be used to find adulterations in dairy industry.

Pillonel et al. (2005) used the DA to assign 183 samples of Europen Emmental Cheeses to the respective regions of origin (Western Austria, Switzerland, South Germany, Finland, France Savoie, France Brittany and France East-Central). At the first, backward SDA procedure was used to select the best discriminant factors on the base of 25 factors previously analyzed. By using the selected variables, the DA correctly assigned the 95% of samples to the true geographic group in the validation set.

Vasta et al. (2011) conducted a research to evaluate the effect of different diets on the presence of volatile organic compounds in meat beef. The researchers used four different diets to fed different groups of heifers. Ninety-four volatile compounds were used as discriminant variables. The SDA selected 16 compounds able to separate the four diets and CDA was applied on these to obtain the respective CANs. The DA correctly assigned all samples to the true respective dietary group.

Sometime DA procedures have been exploited to validate others estimation methods (Basarab et al., 1993) or to help veterinary in predicting diseases (Hailemariam et al., 2014).

With the availability of high-throughput SNP platforms for several livestock species, the discriminant techniques have been also used to analyze genomic data. The enormous number of involved variables (the SNPs) limits however the use of DA in this field.

Jombart et al. (2010), to overcome this problem, introduced a new methodology of DA called Discriminant Analysis of Principal Components (DAPC) where, in developing DA, the original variables were replaced by PCs. Solberg et al. (2009) used the PCA to reduce the dimension of the space variable in applying Bayesian methods to evaluate the genomic breeding value.

Dimauro et al. (2013) applied discriminant procedures to SNP-genotypes data of three breeds of bulls: Holstein, Brown Swiss, and Simmental. The SDA selected 48 high discriminant SNPs that in a genome wide CDA yielded a significant separation among groups. Figure 4 displays the CAN1 vs. CAN2 scatter plot in which the three breeds are clearly differentiated. All animals were correctly assigned to the group of origin.

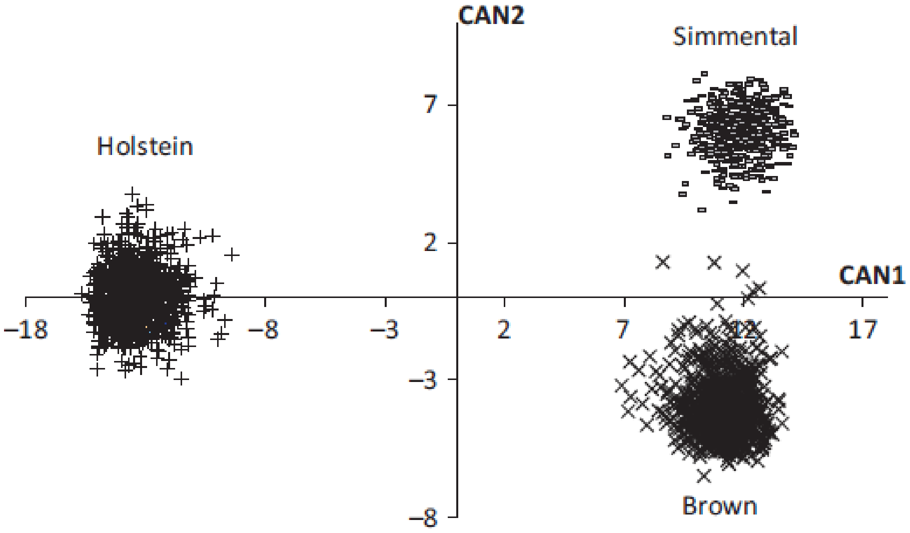

Figure 4. Plot of the two canonical functions (CAN1 and CAN2) obtained by using 48 high discriminant markers Brown and Simmental (Dimauro et al., 2013)

Nishimura et al. (2013) studied two cattle populations, the Japanese Black and the Holstein, to detect breed label falsification in retail beef. Eighteen highly discriminant

SNPs have been used to separate the Japanese Black from Holstein and Japanese Black x Holstein (F_1). The selected SNPs were able to separate the groups.

Biffani et al. (2015) used two multivariate statistical techniques to identify haplotype carriers in a cattle population. In this study, 3645 Italian Brown Swiss cows and bulls, genotyped with the Illumina's BovineSNP50 v2 (54k) BeadChip, were divided in two groups: carriers or non-carriers of the BH2 haplotype on BTA19. Authors used the backward SDA to select SNPs that better fit the model. The error rate of classification with linear DA it was around 1% (or lower) using both two panels of SNP-chips (7K and 54K).

1.6 References

Basarab, J. A., L. M. Rutter and, P. A. Day. 1993. The efficacy of predicting dystocia in yearling beef heifers: II. Using discriminant analysis. Journal of Animal Science. 71(6):1372-1380.

Biffani, S., C. Dimauro, N. P. P. Macciotta, A. Rossoni, A. Stella and, F. Biscarini. 2015. Predicting haplotype carriers from SNP genotypes in Bos taurus through linear discriminant analysis. Genetics Selection Evolution. 47(1):4.

De Maesschalck, R. D. Jouan-Rimbaud and, D. L. Massart. 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems. 50(1):1-18.

Dias, L. A., A. M. Peres, A. C. Veloso, F. S. Reis, M. Vilas-Boas and, A. A. Machado. 2009. An electronic tongue taste evaluation: Identification of goat milk adulteration with bovine milk. Sensors and Actuators B: Chemical. 136(1):209-217.

Dimauro, C., M. Cellesi, R. Steri, G. Gaspa, S. Sorbolini, A. Stella and, N. P. P. Macciotta. 2013. Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. Animal Genetics. 44(4):377-382.

Dossa, L. H., C. Wollny and, M. Gauly. 2007. Spatial variation in goat populations from Benin as revealed by multivariate analysis of morphological traits. Small Ruminant Research. 73(1):150-159.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Annals of Human Genetics.7:179-188.

Gutiérrez, R., S. Vega, G. Díaz, J. Sánchez, M. Coronado, A. Ramírez and, B. Schettino. 2009. Detection of non-milk fat in milk fat by gas chromatography and linear discriminant analysis. Journal of Dairy Science. 92(5):1846-1855.

Hailemariam, D., R. Mandal, F. Saleem, S. M. Dunn, D. S. Wishart and, B. N. Ametaj. 2014. Identification of predictive biomarkers of disease state in transition dairy cows. Journal of Dairy Science. 97(5):2680-2693.

Herrera, M., E. Rodero, M. J. Gutierrez, F. Pena and, J. M. Rodero. 1996. Application of multifactorial discriminant analysis in the morphostructural differentiation of Andalusian caprine breeds. Small Ruminant Research. 22(1):39-47.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. Journal Education Psychology. 25:417–441.

Jombart, T., D. Pontier and, A. B. Dufour. 2009. Genetic markers in the playground of multivariate analysis. Heredity. 102(4):330.

Macciotta, N. P. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro, C. Pieramati and, A. Cappio-Borlino. 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. Journal of Dairy Science. 93(6):2765-2774.

Mardia, K.V., J.T. Kent and, J. M. Bibby. 2000. Multivariate Analysis. Academic Press, London Morrison, F. 1976. Multivariate statistical methods. McGraw-Hill, New York, NY.

Mukaka, M. M. 2012. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. Malawi Medical Journal. 24(3):69-71.

Nishimura, S., T. Watanabe, A. Ogino, K. Shimizu, M. Morita, Y. Sugimoto and, A. Takasuga. 2013. Application of highly differentiated SNPs between Japanese Black and Holstein to a breed assignment test between Japanese Black and F1 (Japanese Black x Holstein) and Holstein. Animal Science Journal. 84(1):1-7.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2(11):559-572.

Pillonel, L., U. Bütikofer, H. Schlichtherle-Cerny, R. Tabacchi and, J. O. Bosset. 2005. Geographic origin of European Emmental. Use of discriminant analysis and artificial neural network for classification purposes. International Dairy Journal. 15(6):557-562.

Ricci, R. 2003.Appunti di Statistica. Università di Firenze, Facoltà di Psicologia, Corso di Laurea in Scienze e Tecniche di Psicologia del Lavoro e delle Organizzazioni.

Solberg, T. R., A. K. Sonesson, J. Woolliams and, T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genetic Selelection Evolution. 41(1):29.

Vasta, V., G. Luciano, C. Dimauro, F. Röhrle, A. Priolo, F. J. Monahan and, A. P. Moloney. 2011. The volatile profile of longissimus dorsi muscle of heifers fed pasture, pasture silage or cereal concentrate: Implication for dietary discrimination. Meat Science. 87(3):282-289.

Yakubu, A., A. E. Salako, I. G. Imumorin, A. O. Ige and, M. O. Akinyemi. 2010. Discriminant analysis of morphometric differentiation in the West African Dwarf and Red Sokoto goats. South African Journal of Animal Science. 40(4):381-387.

CHAPTER 2

A NEW MULTIVARIATE APPROACH FOR GENOME-WIDE ASSOCIATION STUDIES

2.1 Abstract

Traditionally, GWAS are carried out by using a single marker regression model. However, due to the multiple testing error rate, as the number of SNPs increases, the probability to obtain false positive associations enlarges. In this research, an alternative multivariate statistical method, called Discriminant Association Method (DAM), able to overcome those limitations was proposed. Genomic and phenotypic data of 409 young Marchigiana bulls, previously analyzed in a traditional GWAS were used. Seven growth, carcass and meat quality traits were measured: body weight, average daily gain, carcass weight, dressing percentage, shank circumference, head weight and pH at slaughter. Animals belonging to the tails of the phenotypic distribution (25th and 75th percentile) of each trait were selected and flagged as low (LP) or high phenotype (HP). The canonical discriminant analysis (CDA) was developed by using markers as predictors and the LP and HP groups as categorical variable. Around 190 markers for each trait were enough to significantly differentiate LP from HP. Considering SNPs selected in the ordinary GWAS, around 63% of them were confirmed by the DAM approach. The minimum number of DAM selected SNPs able to significantly discriminate groups ranged from 139 for average daily gain to 94 for body weight. The most significant markers, i.e. those with canonical coefficient greater than 0.2 were submitted to gene discovery. Thirty-three interesting loci were highlighted for the seven traits under study. This new information may be useful to better understand the genetic architecture of growth and body composition in cattle.

2.2 Introduction

Genome wide association studies (GWAS) are mainly aimed at understanding the genetic background of complex traits by relating large number of marker genotypes to observed phenotypes. Traditionally, GWAS are carried out by using a single marker regression model that includes both fixed and random effects. However, the larger the number of markers involved in the study, the greater the number of significant SNPs that could be false positives. One of the most popular methods to correct p-values for multiple testing is the Bonferroni's approach. In practice, if *k* is the number of statistical tests developed in the study, the Bonferroni's correction adjusts the α =0.05 significance threshold to α =0.05/*k*. So, with a SNP platform of 700 K, a marker can be declared significant if its p-value is lower than $7*10^{-9}$. The consequence is that few SNPs result significant. The Bonferroni's correction, however, requires that all tests are independent of each other (Bush and Moore, 2012). In the GWAS contest, this hypothesis generally does not fit because as the marker density increases, tests become more correlated, due to the linkage disequilibrium among adjacent SNPs. This leads to an overcorrection applying the Bonferroni's approach. A less-stringent chromosome-wide significance threshold is often considered: the classic p-value = 0.05 is divided by the number of SNPs in each chromosome (Li et al., 2015). So, for example, for BTA1, the significance threshold is reduced to $1.7*10^{-6}$.

A common alternative to the Bonferroni correction is the use of false discovery rate (FDR) (Benjamini and Hochberg, 1995; Osborne, 2006; Bolormaa et al., 2010). This procedure, essentially, corrects for the number of expected false discoveries. However, also FDR can be too much conservative, depending on the fraction of discoveries that are tolerated to be false.

The test statistic distribution could be calculated using permutations (Churchill and Doerge, 1994). Under the null hypothesis that a marker has no effect on the phenotype, data are permuted by randomly assigning phenotypes to each individual thus breaking the genotype-phenotype relationship in the dataset. The procedure is repeated a prefixed number of times (generally 5,000 or 10,000). However, when the number of markers is large as in medium or high density chips, the computational time required becomes a limiting factor.

Apart from the Bonferroni's correction, methods used to control the multiple testing error rate are, however, useful compromises that allows detecting some candidate regions that possibly affect the trait under study.

Each significant SNP obtained with the single marker regression approach explains only a small fraction of the genetic variance of quantitative traits (Maher B., 2008; Visscher et al., 2010). In fact, genetic differences usually are not located in a single locus but often involve also the surrounding part of the genome. Signatures of selection, for example, originates both from the selection pressure on a specific locus but also from the linkage disequilibrium with adjacent loci (Sorbolini et al., 2016). Thus, the analysis of the correlation structure between SNPs located in a particular genomic region or in a chromosome may offer useful insights for finding chromosomal segments associated to phenotypic expression of traits of interest. An alternative could be to develop a statistical method able to simultaneously analyze multiple markers thus accounting for most of the genetic variance (Hayes et al., 2010; Fan et al., 2011). One example, that have obtained encouraging results, is the Bayesian regression methods that, although originally proposed for whole genomic prediction (Meuwissen et al., 2001), can be used for GWAS as well (Fan et al., 2011; Sun et al., 2011; Erbe et al., 2012).

In this paper, a statistical approach able to analyze simultaneously hundreds of SNP-genotypes based on multivariate statistical analysis is proposed. The idea is that, individuals belonging to the tails of the phenotypic distribution of a trait of interest share different allelic combinations for genes involved in its determinism. In consequence, some genes, and related markers, would act differently in the two groups. Genetic differences could be highlighted by using the canonical discriminant analysis (CDA), a multivariate technique that is able to enhance the differences between predefined groups. CDA has already been used to detect pool of markers to be used for traceability purposes in cattle and sheep (Dimauro et al., 2013, 2015), to study signatures of selection (Sorbolini et al, 2016), and to detect carries of recessive haplotypes (Biffani et al., 2015).

In the present research, a method called Discriminant Association Method (DAM), which exploits multivariate statistical techniques, was proposed to highlight possible associations between seven meat phenotypes and SNP-markers. The proposed algorithm was developed and validated by using data previously analyzed in an ordinary GWAS (Sorbolini et al., 2016). The DAM and the GWAS results were then compared.

2.3 Material and methods

The data

Data used in the present research were analyzed by Sorbolini et al. (2016) who carried out a GWAS on 409 young Marchigiana bulls belonging to 117 commercial herds. Aim of that study was the detection of markers significantly associated with carcass and meat traits. Animals were slaughtered at an age ranging from 16 to 24 months. In the GWAS, the following seven out of ten traits investigated showed markers significantly associated: body weight (BW), average daily gain (ADG), carcass weight (CW), dressing percentage (DP), shank circumference (SC), head weight (HW) and pH at slaughter (pH). Only data belonging to these seven traits were analyzed by using the DAM algorithm.

Animals were genotyped by using the Illumina's 50 K BeadChip assay. After data editing, 43,313 markers were retained (for more details see Sorbolini et al., 2016). Phenotypes were adjusted as in the GWAS, by using the following mixed linear model:

$$Y = D + bAGE + a + h + e$$

where Y=the considered phenotype (7 traits); D=fixed effect of slaughter date (46 levels); $bAGE$ =fixed covariable of age at slaughter in month; a =random effect of animal; h =random effect of herd (117); e =random residuals. The animal effect was assumed to be distributed as ~N(0,σ_a^2) where G is the genomic relationship matrix and σ_a^2 is the additive genetic variance. G was calculated according to VanRaden (2008).

For each phenotype, animals belonging to the first and the last quartile were selected and flagged as the high (HP) and low phenotype (LP) group, respectively.

The DAM algorithm for SNP association

Two multivariate discriminant techniques were in sequence applied to the data: the canonical discriminant analysis (CDA) and the discriminant analysis (DA). The CDA is a multivariate dimension-reduction technique whose main objective is the determination of relationships among a categorical variable and a list of independent variables. In particular, CDA tests if the independent variables are able to identify groups listed in the categorical variable. In our research, categories were the HP and LP groups, whereas the independent variables were the SNP-genotypes. The CDA derives a set of new variables, called canonical functions (CAN) that are linear combinations of the original characters. In general, if *k*-groups are involved in the CDA, *k*-1 CANs are extracted. In this research, having two groups for each phenotype, only one CAN was obtained. Canonical coefficients (CC) are the correlations between CAN and original variables. The greater is a CC, the larger the SNP contribution to the CAN. The separation of the two groups was assessed by means of the Mahalanobis distance and the corresponding Hotelling's T-square test (De Maesschalcket al. 2000). This test, however, can be developed only if the pooled (co)variance matrix of data is not singular. In our research, the number of involved animals (rows of data matrix) was lower than the number of SNPs (columns), even for each single chromosome. In this condition, any multivariate technique becomes meaningless because the (co)variance matrix does not have a full rank (Dimauro et al. 2011). A reduction of the space-variables is, therefore, required. Following the suggestions of Dimauro et al. (2013), CCs of CAN extracted for each chromosome were ranked according to their absolute value. Then, SNPs whose CCs exceeded an arbitrary fixed threshold were retained. Markers selected in the 29 autosomes were joined and reranked according to their CCs. Given, in general, the matrix of data, if *n* denotes the number of row (the animals involved in the study), at best, only *n*-1 variables (the SNPs) are linearly independent (Dimauro et al., 2011). However, due to the very low variation of each marker (a SNP has only 0, 1 or 2 as values) the number of linearly independent variables could be lower than *n*-1. The optimum space of the variables was therefore obtained by deleting SNPs with the lower CCs in an iterative process. The process stopped when the maximum number of linearly independent markers, for each phenotype, was obtained. In this condition, when the GW-CDA is developed, both the Mahalanobis distance and the Hotelling's test can be evaluated.

The DA was used to classify animals in the two groups. In DA, the CAN is applied to each individual thus producing a discriminant score. An animal is assigned to a particular group if its discriminant score is lower than the cutoff value obtained by calculating the weighted mean distance among group centroids (Mardia et al., 2000).

Both CDA and DA were used to select the most discriminant markers. They were obtained by reducing, in a new recursive procedure, the number of SNP-variables till obtain the minimum subset of markers able to significantly discriminate the two groups (Hotelling's test p-value <0.0001) and to 100% correctly assign animals to the true group of origin.

Statistical analyses were developed by using the PROC MIXED, CANDISC and DISCRIM of SAS (SAS Institute, Inc.).

Annotation and gene discovery analysis

For all the considered phenotypes, a gene discovery was performed in the genomic regions located around most discriminant SNPs. Annotated genes were identified from the UCSC Genome Browser Gateway (http://genome.ucsc.edu./) and National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov) databases. Intervals of 0.25 Mb upstream and downstream of each SNP were considered. Gene-specific functional analyses were performed by GeneCards (www.genecards.org) and NCBI databases consultation. The biological function of each annotated gene (and related proteins) contained in the significant genomic regions was studied by means of an accurate literature search. Gene names and symbols were derived from the HUGO Gene nomenclature database (www.genenames.org).

2.4 Results

DAM selected SNPs

The DAM procedure selected, for the all the seven studied traits, 1,031 markers spanning the entire genome. As showed in Figure 1, their distribution was not uniform. The largest number of markers (73) was found on BTA2 followed by BTA6 (59). The lowest number (15) was located on BTA29. No significant markers were observed on BTA 5.

On average, around 190 linearly independent SNPs (Table 1) for each phenotype were retained. The subsequent GW-CDA developed for each trait significantly separated the

HP from the LP (p-value <0.0001) and the DA correctly assigned all animals to the true group.

Considering the GWAS selected markers (Table 1), around 63% of them were confirmed by DAM approach. For example, most of GWAS SNPs were identified by DAM for BW, DP and ADG. Only 1 over 5 markers for pH, and 6 over 13 SNPs for SC were obtained by DAM.

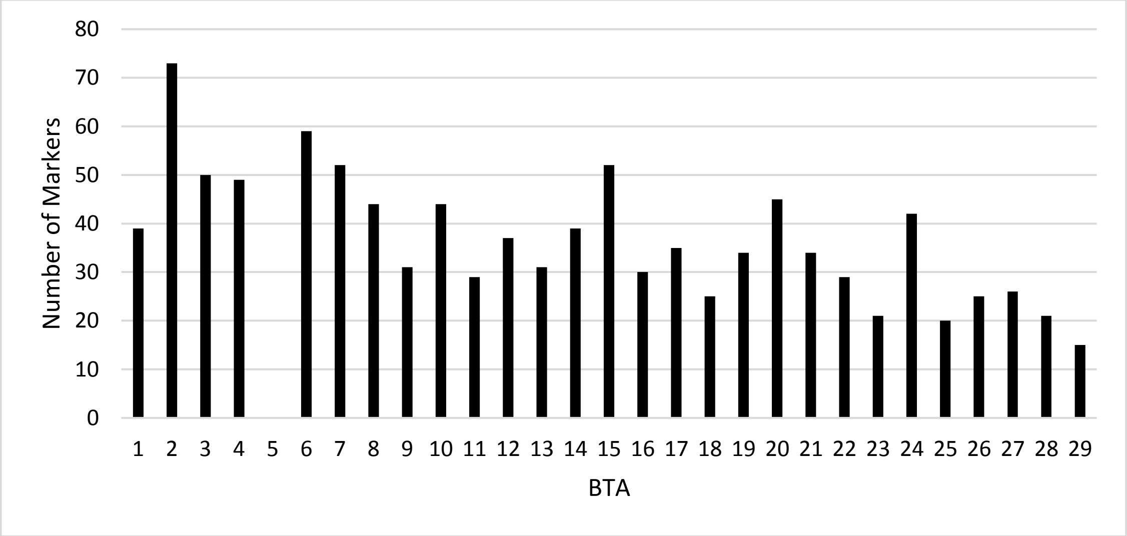

Figure 1. Distribution across the genome of 1,031 DAM selected markers for all seven studied traits

Some of DAM SNPs were shared by two phenotypes, as displayed in Table 2. For example, BW and ADG shared 113 SNPs, whereas for HW and DP, or PH and SC, no common marker was found. Observing the correlations between corrected phenotypes (Table 2), the higher the correlations the greater the number of shared markers.

The minimum number of DAM SNPs able to significantly discriminate groups is reported in Table 1. These markers represent, for each phenotype, the most discriminant SNPs and their number ranges from 139 for ADG to 94 for BW.

Table 1. GWAS, DAM SNPs, number common markers and minimum number of discriminant DAM SNPs

Trait	GWAS SNPs	DAM SNPs	DAM v.s [a]GWAS SNPs	Minimum number [b]DAM SNPs
BW[1]	5	191	4	94
ADG[2]	45	193	30	139
CW[3]	9	191	4	98
DP[4]	12	191	10	98
SC[5]	13	193	6	108
HW[6]	7	192	5	99
pH[7]	5	190	1	120

BW[1] =body weight, ADG[2] =average daily gain, CW[3] =carcass weight, DP[4] =dressing percentage, SC[5] =shank circumference, HW[6] =head weight, pH[7] =pH at slaughter

[a]GWAS SNP =significant markers reported by Sorbolini et al., 2016

[b]DAM SNP =significant markers selected by the discriminant analysis method

Finally, a restricted group of top discriminant SNPs (105 for all traits) was selected by using a CC threshold equal to 0.25 (Table 3). Only these top discriminant markers were submitted to gene discovery.

Table 2. Number of selected markers shared by two phenotypes and, in bold, the Pearson correlations between corrected phenotypes.

	BW[1]	ADG[2]	CW[3]	DP[4]	SC[5]	HW[6]	PH[7]
BW[1]		**0.988**	**0.960**	**0.144**	**0.444**	**0.734**	**0.001**
ADG[2]	113		**0.946**	**0.131**	**0.441**	**0.722**	**-0.006**
CW[3]	88	94		**0.414**	**0.412**	**0.675**	**-0.049**
DP[4]	3	3	17		**0.016**	**0.002**	**-0.183**
SC[5]	8	12	13	4		**0.429**	**0.029**
HW[6]	33	33	26	0	8		**-0.015**
pH[7]	0	1	2	1	0	3	

BW[1] =body weight, ADG[2] =average daily gain, CW[3] =carcass weight, DP[4] =dressing percentage, SC[5] =shank circumference, HW[6] =head weight, pH[7] =pH at slaughter

2.5 Association analysis

Shank Circumference (SC)

Twenty-six significant markers were found to be associated with SC (Table 3). BTA 2 showed a large number of candidate genes: close to ARS-BFGL-NGS-71755 was found the gene *Titin* (TTN) that encodes a large protein of striated muscle; close to marker BTB-02054371 there is *Protein Activator Of Interferon Induced Protein Kinase* (EIF2AK2) and close to marker Hapmap25114-BTA-49906 there is *Oxysterol Binding Protein Like 6* (OSBPL6). Also still on the BTA2, close to marker BTB-000831208 (at 20 813 843 bp), several members of the *Homeobox* family (HOXD1, HOXD3, HOXD4, HOXD9, HOXD10, HOXD11, HOXD12, HOXD13) were located. Finally, close to marker ARS-BFGL-NGS-98126, there is the *Tripartite Motif Containing* (TRIM63), which plays a key role in the atrophy skeletal muscle.

Carcass Weight (CW)

Eighteen SNPs were found significantly associated with this trait (Table 3) but only one marker (UA-IFASA-6018 on BTA22) was associated to two interesting genes: the *Transketolase* (TKT) and *Protein Kinase C Delta* (PRKCD). This SNP was also found associated with the BW trait.

Average Daily Gain (ADG)

Seventeen top discriminant SNPs associated with ADG distributed across ten chromosomes were selected (Table 3). On BTA 3, four candidate genes close to marker ARS-BFGL-NGS-119955 were considered as interesting: *PAS Domain Containing Serine/Threonine Kinase* (PASK), the *High Density Lipoprotein Binding Protein* (HDLBP), the *Inhibitor Of Growth Family Member* (ING5) and the *Deoxythymidylate Kinase* (DTYMK). On BTA10 instead, two putative genes were found close to marker ARS-BFGL-NGS-116295: *Lactase Like* (LCT) and a member of *Small Nuclear RNA Activating Complex Polypeptide 5* (SNAPC5). Finally, on BTA15, close to marker BTB-00619772 the gene of *Apelin Receptor* (APLNR) is annotated.

Dressing Percentage (DP)

Seventeen SNPs (Table 3) were found significantly associated with DP. On BTA1 at 54.2 Mb the *Developmental pluripotency- associated protein 2* (DPPA2) gene is annotated. On BTAs 18 and 22 two loci were identified as candidate genes, the *Syntrophin beta 2* (SNTB2) and the *Solute carrier family 6 member 6* (SLC6A6), respectively.

Head Weight (HW)

In this study, twelve significant markers (Table 3) were found associated with HW. On BTA 7 at 44.8 Mb three putative candidate genes involved in the brain biology were annotated, the *Basigin* (*OK blood group*) (BSG), the *hyperpolarization activated cyclin* *nucleotide gated potassium channel 2* (HCN2) and the *Follistatin Like 3* (FSTL3). On BTA 19, two other sequences worthy of note were the *Sphingolipid transporter 3* (*putative*) (SPNS3) and the *Solute carrier family 26 member 11* (SLC26A11), respectively. Finally, on BTA 20 the *Solute carrier family 6 member 3* (SLC6A3) a dopamine transporter was annotated.

Body Weight (BW)

A total of nine significant SNPs distributed across six autosomes (Table 3) were found associated with BW. On BTA 4, the annotated sequence nearest the marker BTB-00182742 was the *Phosphoinositide 3- kinase gamma* (PIK3GC). On BTA 22 the marker Hapmap41774-BTA-121358 was already reported as significant for CW.

PH at slaughter (pH)

Only five top discriminant markers (Table 3) were found to be associated with this trait. On BTA18 the *Phosphorylase Kinase Regulatory Subunit Beta* (PHKB) was associated with the marker ARS-BFGL-NGS-24006 whereas on BTA 23 at 9.1 Mb, associated with the marker Hapmap38418-BTA-146026, the gene *Peroxisome Proliferator Activated Receptor Delta* (PPARD) was detected.

2.6 Discussion

GWASs have identified hundreds of common variants associated with production traits or disease risk. However, in those studies, the probability to obtain false negative associations is very high. This can be partially ascribed to the fact that most of production traits are controlled by a high number of genes, often associated to markers spanned across the entire genome, as in the present study (Figure 1). Singularly each gene has low effect on the trait and, sometimes, the single marker regression approach is not able to significantly differentiate the really associated SNPs from those that are not. Moreover, the correction of p-values to control the multiple testing error rate, both using severe methodologies (Macciotta et al., 2015) or low stringent techniques (Rolf et al., 2012; Do et al., 2017), increases the risk to obtain false negative associations. The result is that only few markers are often declared associated to production traits. Afterwards, to characterize genomic regions and identify candidate genes influencing the trait under study, a pathway analysis is often developed (Hamzic et al., 2015; Dadousis et al., 2017; Do et al., 2017). To enlarge the number of significant markers to be used in the pathway analysis also the also the so called "suggestive SNPs" (i.e. those SNPs that were near to the Bonferroni's significance threshold) are considered. In the DAM approach, the multivariate CDA was used to select a pool of markers able to discriminate animals belonging to two divergent groups, HP and LP. In the first step of DAM, SNP-variables belonging to each single chromosome were simultaneously analyzed and markers that, acting together, better discriminated groups, were selected. At the end of the procedure, 1,031 SNPs were obtained for the seven traits under study. Their distribution across the genome is displayed in Figure 1. Apart from BTA5 where no marker was detected, most of DAM markers were located in BTAs 2, 6, 7, and 15 with more than 50 SNPs everyone. The remaining markers were more or less uniformly distributed in the other chromosomes. For each trait, a maximum of around 190 linearly independent SNPs were selected (Table 1). The GW-CDA developed by using those markers perfectly separated the LP from the HP group and animals were 100% correctly assigned to the true group of origin by the DA. Among the DAM markers, 60 over 96 GWAS SNPs were found in common. This result indicates that the DAM method was able to capture most of the true associations highlighted by GWAS.

Considering one of the seven traits, BW for example, the 191 selected SNP-variables perfectly captured the differences between animals belonging to the LP and HP groups. However, not all SNPs equally weigh in separating groups. Markers with greater CC absolute values have more important role in discriminating groups than those with lower CCs. According to this suggestion, the minimum number of markers, for each phenotype, able to discriminate groups was also obtained (Table 1) by deleting SNPs with low CC values. For BW, 94 over 191 SNP variables were enough to significantly discriminate LP from HP and the DA correctly assigned all animals to the two groups. In our opinion, these SNPs would be considered "significantly" associated to the trait under study.

Among the "significant" markers, a total 105 most discriminant SNPs (markers whose CCs were greater than 0.25), around 15 for each trait, were selected. Only ten most discriminant markers were in common with the GWAS SNPs. This result indicates that the DAM approach captures the differences between HP and LP using those markers that, acting together, are able to better separate the two groups. Therefore, a singular SNP can be a very little impact on the trait but, acting together with other markers, it can be very important in discriminating groups.

Table 2 lists both the number of common markers and the correlations between corrected phenotypes for the traits under study. BW, ADG and CW showed correlations over than 95% and, consequently, a high number of common markers. In particular, 60 markers were found in common among those three traits. This phenomenon was expected because of pleiotropic effects of marker polymorphism correlated traits. Similar effects were already reported by several authors in beef cattle (Bolormaa et al., 2014; Saatchi et al.; 2014). On the contrary, phenotypes as pH, that is scarcely correlated with the other traits, shared with them few markers, from 0 to 3.

Gene discovery analysis conducted on the 105 most discriminant SNPs highlighted several interesting candidate genes for the considered beef and carcass traits. Table 3 displays, for each chromosome, the trait, the associated marker and the relative candidate genes. For example, on BTA2 at 20 813 843 bp the BTB-00083120 several members of the *Homeobox* gene family was significant associated with SC. These genes were involved in the differentiation and development of limb (Izpisúa-Belmonte and Duboule, 1992) and mutations in this gene have been associated with severe developmental defects on the anterior-posterior limb axis (Hawang et., 1998). Moreover, for ADG, CW and DP traits, some genes controlling nutrient metabolism were detected: APLNR and TKT loci are involved in the glucose metabolism, whereas SLC6A6 gene is related to the taurine transmembrane transport activity.

Table 3. Name, range of analysis and relative gene associated for 105 most discriminant markers

BTA	Trait	Marker	Range	Gene
1	BW	BTA-39405-no-rs	18929350 19429350	BTG3
	CW	ARS-BFGL-NGS-22768	142392593 142892593	BACE2
	DP	ARS-BFGL-NGS-24057	54024662 54524662	DPPA2, MORC1, TRAT1
2	DP	BTB-00077456	950474 1450474	AMER3, CYFIP1, IMP4, NIPA1, NIPA2, PTPN18 TUBGCP5
	SC	ARS-BFGL-NGS-71755	17807166 18307166	CCDC141, TTN
		BTB-02054371	17962187 18462187	DFNB59, FKBP7, PLEKHA3, PRKRA, TTN
		Hapmap25114-BTA-49906	18092153 18592153	DFNB59, FKBP7, OSBPL6, PLEKHA3, PRKRA
		BTB-00083120	20563843 21063843	HOXD1, HOXD3, HOXD4, HOXD9, HOXD10, HOXD11, HOXD12, HOXD13, LNPK, MTX2
		Hapmap47640-BTA-49632	18443590 18943590	OSBPL6, PDE11A, RBM45
	DP	ARS-BFGL-NGS-67309	118248848 118748848	TRIP12
		BTA-110873-no-rs	118248848 118748848	FBXO36, TRIP12
	SC	ARS-BFGL-NGS-98126	127509769 128009769	AUNIP, CATSPER4, EXTL1, FAM110D, MAN1C1 MTFR1L, PAFAH2, PAQR7, PDIK1L, SELENON, SLC30A2, STMN1, TRIM63, ZNF593
3	SC	ARS-BFGL-NGS-119921	14864778 15364778	CERKL, ITGA4, NEUROD1

Table 3. (Continued)

BTA	Trait	Marker	Range	Gene
3	SC	BTA-67383-no-rs	33384546 33884546	AHCYL1, ALX3, CSF1, GSTM3, KCNC4, RBM15. SLC16A4, SLC6A17, STRIP1, UBL4B
	PH	Hapmap60708-rs29011181	53780317 54280317	GBP5, LRRC8B
	HW	ARS-BFGL-NGS-65126	121025721 123226844	ATG4B, BOK, D2HGDH, DTYMK, GAL3ST2, ING5, GAL3ST2, GPR35, HDLBP, ING5, KIF1A, MTERF4, NEU4, PASK, PDCD1, PPP1R7, SEPT2, SNED1, STK25, THAP4
4	ADG, BW	BTB-00182742	47879941 48379941	PIK3CG
	SC	ARS-BFGL-NGS-21411	77028307 77528307	CCM2, H2AFZ, MYO1G, NACAD, OGDH, PURB, RAMP3, TBRG4, ZMIZ2
6	ADG, CW	ARS-BFGL-NGS-642	35593840 36093840	MMRN1, SNCA
	SC	BTA-77725-no-rs	107673653 108173653	ADD1, FAM193A, GRK4, HTT, MFSD10, MSANTD1, NOP14, SH3BP2, TNIP2
7	HW	BTB-00309643	44554507 45054507	AZU1, BSG, C2CD4C, CDC34, ELANE, FGF22, FSTL3, GZMM, HCN2, MADCAM1, MIER2, MISP, ODF3L2, PLPP2, POLRMT, PRSS57, PRTN3, PTBP1, RNF126, SHC2, THEG, TPGS1
	SC	Hapmap41358-BTA-79117	52212171 52712171	CXXC5, DNAJC18, ECSCR, MZB1, NRG2, PAIP2, PROB1, PSD2, SLC23A1, SPATA24, TMEM173, UBE2D3
		Hapmap27181-BTA-148757	75105176 75605176	GABRA6, GABRB2

Table 3. (Continued)

BTA	Trait	Marker	Range	Gene
7	ADG	ARS-BFGL-NGS-6029	63459982 63959982	ARSI, CAMK2A, CD74, NDST1, PDGFR8, SLC6A7,TCOF1,
8	PH	BTA-81053-no-rs	41618874 42118874	KCNV2, PUM3, VLDLR
	ADG	ARS-BFGL-NGS-1517	96300119 96800119	ABCA1, N1PSNAP3A, SLC44A1
9	HW	BTB-00380633	15391092 15891092	MYO6, SENP6
	CW	Hapmap40657-BTA-115707	21667247 22167247	FAM46A
10	ADG	BTB-01855834	2568174 3068174	KCNN2, YTHDC2
	ADG,BW	ARS-BFGL-NGS-57821	5998185 6498185	GCNT4
	ADG,BW	ARS-BFGL-NGS-116295	13065665 13565665	DIS3L, LCTL, MAP2K1, SMAD6, SNAPC5, TIPIN,
12	DP	BTA-115056-no-rs	23051763 23551763	FREM2, NHLRC3, PROSER1, STOML3
13	HW	Hapmap44369-BTA-32763	44469822 44969822	KLF6
	CW	ARS-BFGL-NGS-112445	53239743 53739743	PDYN, STK35, TGM3
	HW	Hapmap47850-BTA-118310	58883064 59383064	2BP1, CTCFL, PCK1, PMEPA1, RAE1, RBM38, SPC11
14	HW	Hapmap33635-BTC-049051	5068260 5568260	COL22A1
15	DP	BTB-01279624	67384199 67884199	C11orf74, COMMD9, PRR5L, RAG1, RAG2, TRAF6

Table 3. (Continued)

BTA	Trait	Marker	Range	Gene
15	PH	ARS-BFGL-NGS-119303	73903038 74403038	API5, TTC17
15	ADG	BTB-00619772	81297653 81797653	APLNR, LRRC55, OR5AK2
15	SC	ARS-BFGL-NGS-115316	76676260 77176260	C11orf94, CHST1, CREB3L1, CRY2, LARGE2, MAPK8IP1, PEX16, PHF21A, SLC35C1
		ARS-BFGL-NGS-28904	81598689 82098689	APLNR, LRRC55, P2RX3, PRG3, RTN4RL2, SSRP1, TNKS1BP1
16	SC	Hapmap48734-BTA-38315	23595036 24095036	BPNT1, EPRS, IARS2, LYPLAL1, SLC30A10
	HW	BTB-01495723	37866695 38366695	BLZF1, C1orf112, CCDC181, F5, KIFAP3, METTL18, SCYL3, SELE, SELL, SELP, SLC19A2
	SC	ARS-BFGL-NGS-57549	43049381 43549381	ANGPTL7, DISP3, EXOSC10, MASP2, MTOR, SRM, TARDBP, UBIAD1
17	DP	ARS-BFGL-NGS-38778	64202955 64702955	CSTF3, DEPDC7, PRRG4, QSER1, TCP11L1
18	DP	ARS-BFGL-NGS-24323	36024916 36524916	CDH1, CDH3, CHTF8, HAS3, SNTB2, TANGO6, UTP4, ZFP90
	PH	ARS-BFGL-NGS-24006	15721297 16221297	PHKB
19	HW	ARS-BFGL-NGS-93006	25227190 25727190	ANKFY1, CYB5D2, GGT6, GGT8,MYBBP1A, SMTNL2, SPNS3, TEKT1, UBE2G1, ZZEF1
19	HW	ARS-BFGL-NGS-117951	52912864 53412864	CARD14, CBX2, CBX4, CBX8, CCDC40, EIF4A3, GAA, SGSH, SLC26A11, TBC1D16
	SC	ARS-BFGL-NGS-112332	56247918 56747918	ACOX1, CASKIN2, CDK3, EVPL, FBF1, GALK1, H3F3B, ITGB4, LLGL2, MRPL38, RECQL5, SAP30BP, SMIM5, SMIM6, TMEM94, TRIM47, TRIM65, TSEN54, UNC13D, UNK, WBP2

Table 3. (Continued)

BTA	Trait	Marker	Range	Gene
20	SC	BTB-00+132:149778405	36853943 37353943	GDNF, NIPBL, NUP155, WDR70
20	HW	ARS-BFGL-NGS-117598	73246969 73746969	AHRR, BRD9, CCDC127, CEP72, CLPTM1L, EXOC3, LPCAT1, LRRC14, NKD2, PDCD6, SDHA, SLC12A7,SLC6A18, SLC6A19, SLC6A3, SLC9A3, TERT, TPP, TRIP13
21	DP	ARS-BFGL-NGS-115704	14260023 14760023	CHD2, FAM174B, RGMA
22	CW	UA-IFASA-6018	47956884 48456884	CHCHD5, DCP1A, PRKCD, TKT
	CW,BW	Hapmap41774-BTA-121358	48032179 48532179	CHCHD5, DCP1A, PRKCD, SFMBT1,TKT
	ADG	ARS-BFGL-NGS-76123	49243685 49743685	ABHD14B, ALAS1, DUSP7, GPR62, GRM2, IQCF1,IQCF2, IQCF3, IQCF5, IQCF6, PARP3, PCBP4, POC1A,
	CW	BTA-54868-no-rs	56618904 57118904	CAND2, EFCAB12, H1FOO, IFT122, MBD4, PLXND1, RHO, RPL32, TMCC1, TMEM40
	DP	Hapmap26413-BTA-146026	57577442 58077442	FGD5, MRPS25, NR2C2, RBSN, SYN2, TIMP4
23	PH	Hapmap38418-BTA-57213	8893924 9393924	DEF6, PPARD, SCUBE3, TCP11
24	CW	Hapmap38513-BTA-58574	50951345 51451345	AP3S1, ELAC1, MEX3C, SMAD4
24	DP	BTB-00856713	58184298 58684298	C3orf20, GRIP2, SLC6A6
29	SC	Hapmap54633-rs29021971	25012747 25512747	DBX1, HTATIP2, NAV2
		ARS-BFGL-NGS-4431	38121072 38621072	PAG5

2.7 Conclusions

In the present research, a new multivariate procedure to develop a GWAS on meet and carcass traits is proposed. The DAM algorithm selected around 60% of markers obtained with the traditional single SNP regression analysis. The proposed procedure assigns, to each marker, a CC that can be considered as an indicator of association: the greater the CC the more associated the related marker. For each trait, the minimum number of SNPs able to significantly discriminate groups was obtained. Their number swung from 94 for BW to 139 for ADG. Only these markers were declared associated with their own trait. However, among associated markers, those with higher CCs have a greater degree of association with the trait. Therefore, to explore the ability of DAM in selecting genomic regions harboring candidate genes, around 15 most discriminant SNPs (CC >0.25), for each trait, were selected and submitted to gene discovery analysis. Thirty-three putative genes were found thus confirming the goodness of proposed methodology.

2.8 References

Bolormaa, S., J. Pryce, B. J. Hayes and, M.E. Goddard. 2010. Multivariate analysis of a genome-wide association study in dairy cattle. Journal of Dairy Science. 93:3818–3833.

Bolormaa, S., J. Pryce, A. Reverter, Y. Zhang, W. Barendse, K. Kemper, B. Tier, K. Savin, B. J. Hayes and, M. E. Goddard.2014.A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS Genetics. 10(3):e1004198.

Benjamini, Y. and J. Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological). 57:289-300.

Biffani, S., C. Dimauro, N. P. P. Macciotta , A. Rossoni, A. Stella and, F. Biscarini. 2015. Predicting haplotype carriers from SNP genotypes in Bos taurus through linear discriminant analysis. Genetics Selection Evolution. 47:4.

Bush, W. S. and, J. H. Moore. 2012. Genome-wide association studies. PLoS Computational Biology. 8(12):e1002822.

Churchill, G. A. and, R. W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics. 138(3):963–971.

Dadousis, C., S. Pegolo , G. J. M. Rosa , D. Gianola , G. Bittante and, A. Cecchinato. 2017. Pathway-based genome-wide association analysis of milk coagulation properties, curd firmness, cheese yield, and curd nutrient recovery in dairy cattle. Journal of Dairy Science. 100:1223–1231.

De Maesschalck, R., D. Jouan-Rimbaud and, D. L. Massart. 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems. 50(1):1-18.

Dimauro, C., M. Cellesi, M. A. Pintus and, N. P. P. Macciotta. 2011. The impact of the rank of marker variance–covariance matrix in principal component evaluation for genomic selection applications. Journal of Animal Breeding and Genetics. 128(6):440-445.

Dimauro, C., M. Cellesi, R. Steri, G. Gaspa, S. Sorbolini, A. Stella and, N. P. P. Macciotta. 2013. Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. Animal Genetics. 44:377-382.

Dimauro, C., L. Nicoloso, M. Cellesi, N. P. P. Macciotta, E. Ciani, B. Moioli, F. Pilla and, P. Crepaldi. 2015. Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. Small Ruminant Research. 128:27-33.

Do, D. N., N. Bissonnette, P. Lacasse, F. Miglior, M. Sargolzaei, X. Zhao and, E. M. Ibeagha-Awemu. 2017. Genome-wide association analysis and pathways enrichment for lactation persistency in Canadian Holstein cattle. Journal of Dairy Science. 100:1955–1970.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason and, M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science. 95(7):4114-4129.

Fan, B., S. K. Onteru, Z. Q. Du, D. J. Garrick, K. J. Stalder and, M. F. Rothschild. 2011. Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. PloS One. 6(2):e14726.

Hamzić, E., B. Buitenhuis, F. Hérault, R. Hawken, M. S. Abrahamsen, B. Servin and, B. Bed'Hom. 2015. Genome-wide association study and biological pathway analysis of the Eimeria maxima response in broilers. Genetics Selection Evolution. 47(1):91.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman and, M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PloS Genetics 6(9):e1001139.

Hwang, S. J., T. H. Beaty, I. McIntosh, T. Hefferon and, S. R. Panny. 1998. Association Between Homeobox-Containing Gene MSX1 and the Occurrence of Limb Deficiency. American Journal of Medical Genetics. 75:419–423.

Izpisúa-Belmonte, J. C. and, D. Duboule. (1992). Homeobox genes and pattern formation in the vertebrate limb. Developmental Biology. 152(1):26-36.

Li, X., A. J. Buitenhuis, M. S. Lund, C. Li, D.Sun, Q. Zhang, N. A. Poulsen and, G. Su. 2015. Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. Journal of Dairy Science. 98:8152–8163.

Macciotta, N. P. P., G. Gaspa, L. Bomba, D. Vicario, C. Dimauro, M. Cellesi and, P. Ajmone-Marsan. 2015. Genome-wide association analysis in Italian Simmental cows for lactation curve traits using a low-density (7K) SNP panel. Journal of Dairy Science. 98(11):8175-8185.

Maher, B. 2008. Personal genomes: The case of the missing heritability.Nature News. 456(7218):18-21.

Mardia, K.V., J.T. Kent and, J. M. Bibby. 2000. Multivariate Analysis. Academic Press, London Morrison, F. 1976. Multivariate statistical methods. McGraw-Hill, New York, NY.

Meuwissen, T. H. E., G. J. Hayes and, M. E.Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829.

Osborne, J. A. 2006. Estimating the false discovery rate using SAS. In SAS Users Group International Proceedings. 190:1-10.

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure, S. L. Northcutt, M. S. Kerley and, R. L. Weaber.2012. Genome-wide association analysis for feed efficiency in Angus cattle. Animal genetics. 43(4):367-374.

Saatchi, M., R. D Schnabel, J. F. Taylor and, D. J Garrick.2014. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics. 15:442.

Sorbolini, S., S. Bongiorni, M. Cellesi, G. Gaspa, C. Dimauro, A. Valentini and, N. P. P. Macciotta. 2016. Genome wide association study on beef production traits in Marchigiana cattle breed. Journal of Animal Breeding and Genetics. 134(1):43-48.

Sun, X., D. Habier, R. L. Fernando, D. Garrick, D. J. Garrick and, J. C. M. Dekkers. 2011. Genomic breeding value estimation and QTL mapping of QTLMAS-2010 data using Bayesian methods. In: BMC proceedings. BioMed Central. (pp.) S13.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science. 91(11):4414-23.

Visscher, P. M., J. Yang and, M. E. Goddard. 2010. A commentary on ‘common SINPs explain a large proportion of the heritability for human height’ by Young et al. (2010). Twin Research and Human Genetics. 13(6):517-524.

CHAPTER 3

GENOME-WIDE ASSOCIATION STUDY ON RESIDUAL CONCENTRATE INTAKE IN BROWN SWISS YOUNG BULLS BY USING THE MULTIVARIATE DAM APPROACH

3.1 Abstract

The genetic selection for feed efficiency has been mainly focused on the residual feed intake (RFI). This index, however, is difficult to measure because it requires, for each animal, the record of the total daily intake. Concentrates are the most expensive fraction of the bovine diet and, in intensive or semi-intensive farms that do not use the unifeed ration, the bovine diet consists of both concentrate and forages. The individual daily concentrate consumption can be easily recorded by using automated feeding systems. In the present research, a new index, the residual concentrate intake (RCI), was defined and a GWAS was developed on 736 genotyped Brown Swiss young bulls. Both the traditional single SNP regression and the DAM algorithm, developed in chapter II of this thesis, were used to select markers associated to the RCI. The regression approach highlighted only one associated marker, whereas the DAM selected 382 discriminant SNPs. These markers could be used as variables in a DA to assign Brown Swiss bovines to a low or high RCI group when their RCI phenotype is not known. Among the DAM selected markers, the most discriminant 88 SNPs were sufficient to significantly separate animals with low and high RCI values. Several putative genes, controlling directly or not the RCI were found in the genomic regions flagged by these markers.

3.2 Introduction

Feed costs weigh upon the farm budget for around 60-65% (Sainz et al., 2004). Generally, as the level of cow production increases costs of maintenance enlarge (Davis et al., 2014). However, animals that, at the same level of production or weight, efficiently convert nourishment into energy need less feed compared to inefficient animals (Green et al., 2013). Therefore, a genetic selection applied on traits associated to feed efficiency, both in dairy and beef cattle could reduce total farm costs (Sainz et al., 2004; Pryce et al., 2014).

Specialized literature reports several measures to evaluate the efficiency of feed utilization (FE). For example, the following indexes are often used: the average daily gain, the dry matter intake, the feed conversation ratio, the partial efficiency of growth, the residual body gain, the maintenance efficiency, the nutrient transformation, the residual intake gain and other (Carstens et al., 2006; Berry and Crowley, 2013; Crowley et al., 2010). Actually, the most used index to evaluate feed efficiency in cattle is the residual feed intake (RFI) (Berry and Crowley, 2013). It represents the amount of feed consumed, net of the animal requirements of body weight and production. In other words, for an individual, RFI is essentially the difference between the feed it eats and its predicted feed consumption. Efficient animals eat less than expected and have a negative RFI, while inefficient animals eat more than expected and have a positive RFI. Being, by definition, RFI independent from production and body weight, an individual with low RFI produces the same amount of products as its contemporaries eating less feed. RFI divergences in farm animals, both for beef (Herd and Arthur, 2009; de Oliveira et al., 2014; Rolf et al., 2012) and dairy cattle (Potts et al., 2015; Waghorn et al., 2012; Williams et al., 2011), have been well established. In addition, several researches have observed a very high repeatability of RFI across different diets (Potts et al., 2015) and across different periods of life of an animal (Macdonald et al., 2014).

Despite a RFI heritability around 0.20-0.40 (Bolormaa et al., 2013; Pryce et al., 2012; Connor et al., 2014), only few countries, at present date, include feed efficiency, in particular the RFI trait, in their breeding programs. One reason is that, in genomic selection, a congruent training population of bulls whose RFI is known would be created. The RFI evaluation, however, requires the measurement of the actual individual feed intake. Several researches have studied FE in cattle. However measure of feed intake is difficulty and expensive (Williams et al., 2011; Pryce et al., 2012), and for this reason, commonly, many studies have used only small groups of animals with, in consequence, a limited genetic variation and heritability.

The world human population is steadily growing and, in consequence, also the livestock sector is expected to increase (CAST 2013; van Zanten et al., 2016). The rising demand of cereals both for animals and humans feeding will determine an increase in their price. Several researches (Soder and Rotz, 2001; Steinfeld and Opio, 2010) have highlighted that, in all typologies of breeding systems, the most important part of the total farm costs is related to concentrates.

Generally, in intensive or semi-intensive farms that does not use the unifeed ration, the bovine diet consists of both concentrate and forages. As the individual consumption of total dry matter is difficult to evaluate, the amount of concentrates consumed by an animal can be easily obtained. Several farms are equipped with automated feed systems that allows recording the amount of concentrate consumed by the single animal with a good precision.

In the present research, a new index to evaluate feed efficiency, the residual concentrate intake (RCI) was defined. The RCI, as RFI, identifies efficient and inefficient individuals in converting, in this case, the concentrate. Since the RCI is quite simple to evaluate, it could take the place of RFI in genomic breeding programs. A useful contribute to breeding programs that include RCI could be offered by the detection of genomic regions and of candidate genes which regulate it. This objective can be achieved by means of genome wide association studies (GWAS). Traditionally these studies are developed by using a single SNP linear regression model which includes both fixed and random effects. However, especially for traits with low heritability, this approach leads to a great number of false negative SNPs. Moreover, as the number of markers involved in the study increases, the number of significant SNPs that could be false positives enlarges because of the multiple testing error rate.

In the present research, in addition to the ordinary single SNP regression approach, the DAM method explained in chapter II of this dissertation was applied to develop a GWAS for selecting markers associated to RCI.

3.3 Materials and methods

Animals

A total of 1092 Brown Swiss young bulls were involved in the study. Animals, originated from different farms, arrived at the age of 5 - 6 months in the ANARB genetic center, (Italian Association of Brown Swiss, Bussolengo, Italy). Bulls were housed in a quarantine pen for around one month and then were distributed among boxes with a maximum of six bulls/box. Each box was equipped with an automatic feeding system (Figure 1) able to recognize the individual animal and to record the daily concentrate it consumed. The diet offered consisted of 1.2 kg of concentrate for 100 kg of BW and of hay administered ad libitum. The concentrate composition is reported in Table 1. Animals remained in those boxes for around three months where the BW was monthly recorded. After this period, bulls were moved into single pens for mount training. From an initial 1092 young bulls, only 736 animals with at least three BW records were considered for statistical analysis.

Table 1. Chemical composition of concentrate diet

Compositions	% As Fed
Crude protein	18.00%
Crude oils and fats	3.20%
Crude cellulose	10.04%
Crude ash	7.63%
Sodium	0.37%

***Ingredients of the concentrate:** wheat bran, corn gluten feed, flaked corn, flaked barley, wheat meal, dehulled soybean meal, sunflower seed cake, flaked fava beans, corn germ meal, alfaalfa meal, dried beet pulp, dehulled sunflower cake, seed meal, dried carobs, corn, soy hulls, distillers, whey, calcium carbonate, beet molasses, palm oil, sodium chloride, sodium bicarbonate.

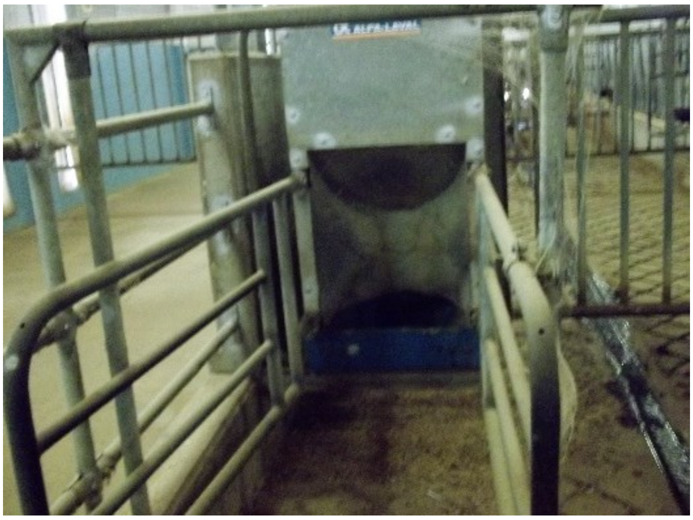

Figure 1. Automatic feeding system

Genomic data

Animals were genotyped by using the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA). SNP with more than 1% missing values or minor allele frequencies less than 5% were removed. The remaining missing genotypes were replaced with the most frequent allele at that specific locus. At the end of data editing, 41,183 SNP distributed in 29 autosomes were available for further analysis. Genotypes were coded as the number of copies of one SNP allele carries, that is, 0 (homozygous for allele A), 1 (heterozygous), or 2 (homozygous for allele B).

RCI evaluation

The RCI was evaluated by using the following simple equation:

$$RCI = CP - CC$$

where CC is the daily consumed concentrate, CP is the predicted daily concentrate intake calculated with the following equation:

$$CP = 1.2 * BW/100$$

where BW is the actual body weight, and 1.2 is the concentrate, in kg, for 100 kg of BW.

The obtained RCI values were adjusted by using the following linear mixed model:

$$RCI_{ijk} = \mu + M_i + Y_j + a_k + e$$

where μ =overall mean; M =fixed effect of the i^{th} birth month (12); Y =fixed effect of the j^{th} birth year (from 2002 to 2013); a =random additive effect of animals; and e =random residuals.

Animals were grouped into high (HRCI) and low (LRCI) RCI groups. One calf belonged to the HRCI if its RCI was higher than 0.5 SD of the mean RCI, whereas animals with RCI lower than 0.5 SD were classified in the LRCI group (Potts et al., 2015). For instance, calves belonging to LRCI and HRCI had divergent RCI, with the best and the worst efficient animals, respectively. Individuals that did not belong to the two groups were discarded.

Single SNP association analysis

The traditional genome wide association study was developed by regressing RCI phenotypes on SNP covariates with the following mixed linear model:

$$y_{ijkm} = \mu + M_i + Y_j + SNP_k + a_m + e$$

where, to respect the model used to correct RCI values, the fixed covariable of the k^{th} SNP marker genotype is included (for more details see Macciotta et al., 2015). The by chromosome

Bonferroni-corrected significance levels (Li et al., 2015) for SNP effects were calculated to account for multiple testing: uncorrected p-values were multiplied by the number of tests performed in each chromosome. One SNP was considered significantly associated when the corrected p-value was lower than 0.05.

The DAM algorithm for SNP association

Data were then arranged in a multivariate manner with one animal in one row and 41,184 columns: one for the classification variable indicating the RCI groups to which each animal belonged, and 41,183 for SNP-variables. The DAM algorithm was explained in detail in chapter II of this dissertation. Briefly, two multivariate techniques, the canonical discriminant analysis (CDA) and the discriminant analysis (DA) were applied to data. The CDA is a space variables reduction technique able to test if individuals belonging to *k* different groups can be correctly classified in those groups with a particular set of variables. With this aim, *k*-1 canonical functions (CAN), i.e. linear combinations of the original variables, are generated. In the present research, being *k* =2, only one CAN was extracted. Distances between groups (generally the Mahalanobis distance) are calculated and the effective separation of groups is assessed by means Hotelling's T-square test (De Maesschalcket al. 2000). In the DA, CANs are applied to each individual thus producing a discriminant score. An animal is assigned to a particular group if its discriminant score is lower than the cutoff value obtained by calculating the weighted mean distance among group centroids (Mardia et al., 2000).

The DAM algorithm was applied to data to select a pool of markers able to significantly separate the HRCI and the LRCI groups. To validate the derived discriminant functions, the complete dataset was randomly divided into training and validation dataset in the proportion of four to one. This partition of the dataset was iterated 5,000 times by using a bootstrap procedure (Efron, 1979). At each run, DA was applied to the training dataset to predict the RCI group of animals in the validation dataset. Finally, the minimum number of SNPs able to separate groups was identified by applying, repeatedly, the CDA and the DA. At each run, the number of involved markers was reduced till obtain an highly significant Hotelling's t-test in the CDA and, in the DA, a correct assignment of animals to the true group of origin.

Annotation and gene discovery analysis

The genomic regions located around most discriminant SNPs were analyzed to perform a gene discovery. Annotated genes were identified from the UCSC Genome Browser Gateway (http://genome.ucsc.edu./) and National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov) databases. Intervals of 0.25 Mb upstream and downstream of each SNP were considered. Gene-specific functional analyses were performed by GeneCards (www.genecards.org) and NCBI databases consultation. The biological function of each annotated gene (and related proteins) contained in the significant genomic regions was studied by means of an accurate literature search. Gene names and symbols were derived from the HUGO Gene nomenclature database (www.genenames.org).

3.4 Results

GWAS results

After the by chromosome Bonferroni correction, the traditional GWAS selected only one significant marker, ARS-BFGL-NGS-62299, located in BTA11 at 9 153 560 bp. Three genes associated with feed efficiency (Olivieri et al., 2016) were detected in the region surrounding the marker: the FHL2 (*Four And A Half LIM Domains 2*), the GPR45 (*G Protein-Coupled Receptor 45*) and the *TGFBRAP1* (*Transforming Growth Factor Beta Receptor Associated Protein 1*).

DAM results

The DAM selected a total of 382 markers spanned across the genome as displayed in Figure 2.

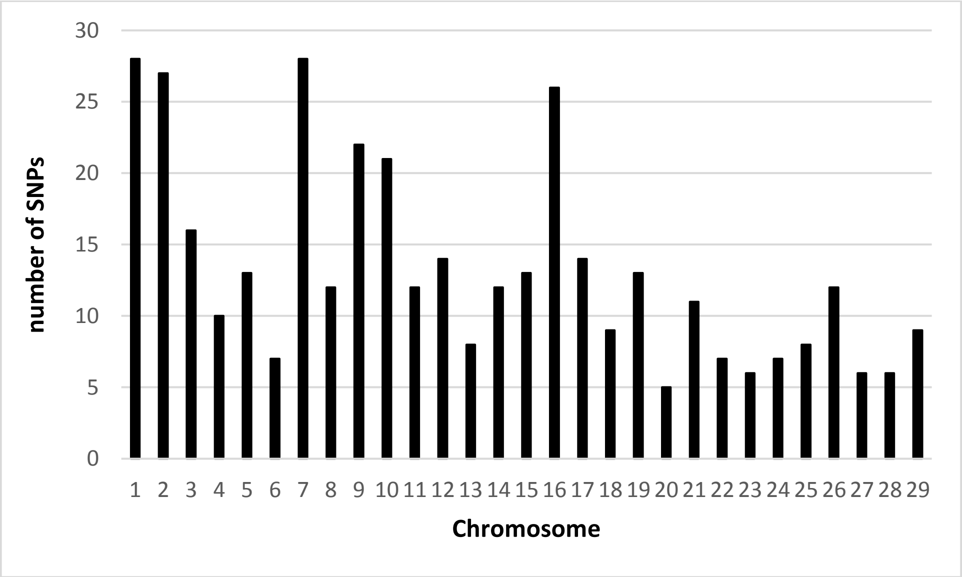

Figure 2. Distribution of DAM selected markers across the genome

Their distribution in each chromosome was not uniform. The largest number of markers was found on BTA1, BTA2, BTA7 and BTA16 with more than 25 SNPs each, followed by BTA3, BTA9 and BTA10 with 16, 22 and 21 markers, respectively. The residual markers were almost uniformly distributed in the remaining autosomes. The genome-wide CDA developed by using the 382 selected SNPs significantly separated LRCI from HRCI (Hotelling's p-value <0.0001) and the DA correctly assigned all animals to the true group of origin also in the bootstrap resampling procedure.

Figure 3 displays the plot of the CAN thus confirming the clear separation between the two RCI groups.

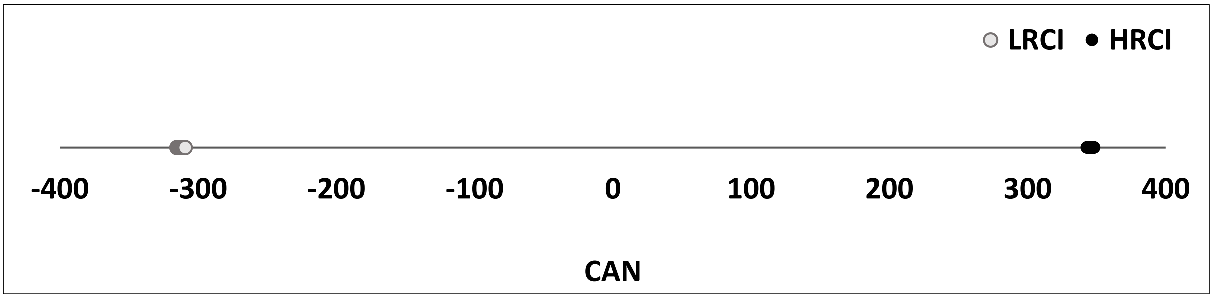

Figure 3. Graph of the canonical function (CAN) obtained in a genome-wide canonical discriminant analysis using a selected pool (382) of SNP variables

The minimum number of SNPs able to significantly discriminate LRCI from HRCI was fixed to 88 in the recursive procedure. Figure 4 shows the plot of the CAN where the separation between LRCI and HRCI is clearly depicted.

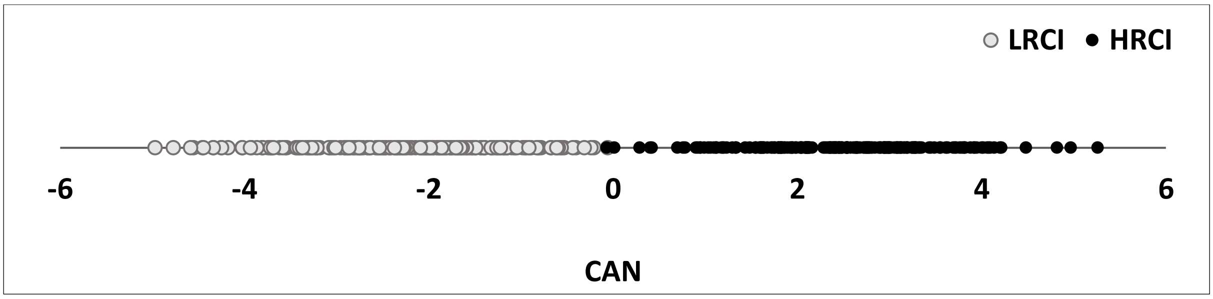

Figure 4. Graph of the canonical function (CAN) obtained in a genome-wide canonical discriminant analysis using a restricted pool (88) of SNP variables

The selected 88 SNPs were the most discriminant markers and, in consequence, they were considered associated to RCI. Given one of those markers, the greater the corresponding CC (canonical coefficient) the more associated with RCI. In Figure 5 the plot of CCs is displayed. Each point corresponds to a specific marker with its CC.

The selected 88 markers and their CCs are also reported in Table 2. Figure 5 and Table 2 should be considered together. Markers listed in Table 2 were divided in three classes. To the low class (L) were assigned those SNPs whose CCs were lower than -0.1 and were reported in black both Table 2 and in Figure 5: markers whose CCs ranged from -0.1 to 0.1 were assigned to the medium class (M) and reported in grey. The remaining SNPs were assigned, according to their CCs, to the L and H classes and reported in black.

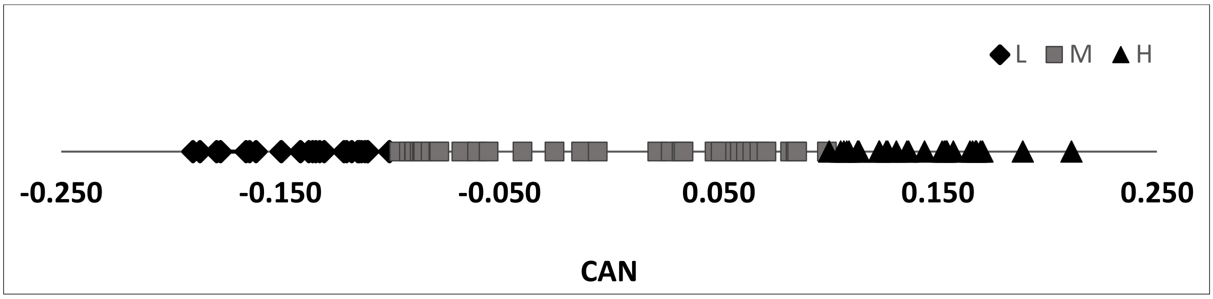

Figure 5. Graph of the canonical coefficients of the canonical function (CAN) obtained in a genome-wide canonical discriminant analysis using a selected number (88) of SNP variables of low (L), medium (M) and high (H) canonical coefficient classes

Table 2. List of 88 selected markers and their canonical coefficients (CC). Negative and positive black markers refer to the low (L) and high (H) class whereas grey markers indicate the medium (M) class

Name marker	BTA	Position	CCs
ARS-BFGL-NGS-7567	5	28331294	**-0.255**
Hapmap43797-BTA-17694	5	92527538	**-0.190**
BTA-101550-no-rs	18	38357061	**-0.187**
ARS-BFGL-NGS-111791	7	716475	**-0.179**
ARS-BFGL-NGS-79829	28	8501840	**-0.177**
ARS-BFGL-NGS-33663	18	63512359	**-0.166**
ARS-BFGL-NGS-104517	11	2716257	**-0.164**
ARS-BFGL-NGS-44030	12	11865727	**-0.161**
ARS-BFGL-NGS-114949	23	50170679	**-0.149**
BTB-01281817	3	72000950	**-0.141**
ARS-BFGL-NGS-10428	10	13822864	**-0.137**
BTA-83229-no-rs	9	10569900	**-0.135**
Hapmap41799-BTA-21550	6	34635944	**-0.134**
Hapmap24381-BTA-150366	22	9869416	**-0.132**
BTA-57141-no-rs	23	8513266	**-0.130**
BTA-68404-no-rs	6	61523195	**-0.121**
ARS-BFGL-NGS-84420	17	66725091	**-0.120**
BTA-28903-no-rs	12	7766553	**-0.117**
ARS-BFGL-NGS-41209	6	94589635	**-0.115**
ARS-BFGL-NGS-7003	5	54745039	**-0.114**
BTB-01746933	27	34704855	**-0.113**
ARS-BFGL-BAC-15732	13	50950127	**-0.111**
ARS-BFGL-NGS-110046	4	85188654	**-0.110**
BTB-01020342	29	31899837	**-0.110**
Hapmap42354-BTA-89491	17	52156863	**-0.100**

Table 2. (Continued)

Name marker	BTA	Position	CCs
BTB-01982674	3	62174708	**-0.096**
ARS-BFGL-BAC-27305	2	13202377	**-0.091**
ARS-BFGL-NGS-2658	2	5370008	**-0.089**
ARS-BFGL-NGS-1777	9	82563961	**-0.089**
BTA-54658-no-rs	22	47465329	**-0.086**
Hapmap32784-BTA-124466	1	99156076	**-0.085**
Hapmap51456-BTA-47469	2	42137116	**-0.085**
Hapmap42294-BTA-69421	3	6716282	**-0.085**
ARS-BFGL-NGS-3949	29	29795486	**-0.081**
BTB-00174824	4	34856797	**-0.078**
ARS-BFGL-NGS-5408	3	91757984	**-0.077**
ARS-BFGL-NGS-73909	25	38232904	**-0.067**
Hapmap33998-BES10_Contig543_926	8	109651024	**-0.060**
BTA-88724-no-rs	7	98194828	**-0.055**
UA-IFASA-4065	12	11824890	**-0.039**
ARS-BFGL-NGS-60387	2	1795004	**-0.025**
ARS-BFGL-NGS-75066	17	51097333	**-0.013**
ARS-BFGL-NGS-42296	28	9175452	**-0.005**
ARS-BFGL-NGS-2406	19	20395816	**0.022**
ARS-BFGL-NGS-39696	16	27629566	**0.028**
BTA-108009-no-rs	14	41636471	**0.033**
ARS-BFGL-NGS-111222	15	58796348	**0.034**
Hapmap35103-BES3_Contig455_1055	11	92297411	**0.048**
ARS-BFGL-NGS-117800	16	62871926	**0.051**
ARS-BFGL-NGS-43284	21	55096333	**0.057**
ARS-BFGL-NGS-97895	25	24877266	**0.060**
BTB-00623849	16	3429932	**0.063**
ARS-BFGL-NGS-102933	17	19315294	**0.065**
ARS-BFGL-NGS-25970	7	17945352	**0.069**
Hapmap45640-BTA-113575	11	95637394	**0.072**
Hapmap58782-rs29016179	27	10565495	**0.083**
ARS-BFGL-NGS-63916	26	34436316	**0.085**
Hapmap53367-rs29014082	2	134281005	**0.086**
BTA-43816-no-rs	18	55256431	**0.099**
UA-IFASA-6791	19	5311065	**0.101**
ARS-BFGL-NGS-34422	15	40835081	**0.106**
BTA-108798-no-rs	7	97393157	**0.107**
ARS-BFGL-NGS-43521	7	111684918	**0.108**
BTB-00807137	1	102442760	**0.110**
BTB-01697487	7	66645827	**0.113**
BTA-102427-no-rs	13	14665246	**0.114**
ARS-BFGL-NGS-114754	9	104926053	**0.123**
BTA-57038-no-rs	23	5931664	**0.123**
Hapmap47842-BTA-115522	15	37264052	**0.127**

Table 2. (Continued)

Name marker	**BTA**	**Position**	**CCs**
Hapmap50033-BTA-56544	23	38177588	**0.127**
BTA-61759-no-rs	26	45430774	**0.131**
BTB-01130157	7	58333017	**0.136**
ARS-BFGL-NGS-61091	5	114698428	**0.137**
ARS-BFGL-NGS-42945	21	45019363	**0.144**
Hapmap25578-BTA-148357	1	121818219	**0.152**
ARS-BFGL-NGS-34288	29	25286242	**0.153**
Hapmap38245-BTA-69529	3	111745557	**0.154**
Hapmap41666-BTA-86780	10	79455690	**0.154**
ARS-BFGL-BAC-16277	12	39972336	**0.157**
ARS-BFGL-BAC-7467	13	21590166	**0.165**
ARS-BFGL-NGS-5507	28	41306221	**0.166**
ARS-BFGL-NGS-94157	24	7380047	**0.167**
ARS-BFGL-NGS-116025	10	79496312	**0.168**
BTA-42767-no-rs	18	20947807	**0.168**
Hapmap43138-BTA-107007	2	28994690	**0.170**
Hapmap50324-BTA-34690	14	39640732	**0.170**
BTB-01145402	2	113946668	**0.189**
ARS-BFGL-NGS-4023	16	54134054	**0.211**

Gene discovery

Table 3 displays the complete list of genes found in genome zones surrounding the 88 most discriminant markers. Several genes controlling feed efficiency were found. The greatest number of genes surrounding a single SNP was found in BTA 7 with 20 putative genes associated with ARS-BFGL-NGS-25970 (on the region 17 695 352 -18 195 352 bp), followed by the marker ARS-BFGL-NGS-2406 in BTA 19 that presented 17 genes (in the region ranging from 20 145 816- 20 645 816 bp).

Table 3. List of the 88 top discriminant markers and relative genes surrounding them. Superscripts L, M and H refers to CC classes (L =low, M =medium, H =high)

BTA	Marker	Range	Gene
1	Hapmap32784-BTA-124466^{M}	98906076 99406076	MECOM
	BTB-00807137^{H}	102192760 102692760	BCHE
	Hapmap25578-BTA-148357^{H}	121568219 122068219	ZIC1, ZIC4
2	ARS-BFGL-NGS-60387^{M}	1545004 2045004	ARHGEF4, FAM168B, PLEKHB2
	ARS-BFGL-NGS-2658^{M}	5120008 5620008	BIN1, CYP27C1, ERCC3, MAP3K2, NAB1
	BTB-01145402^{H}	24737192 25237192	CYBRD1, DCAF17, DYNC1I2,METTL8
	Hapmap43138-BTA-107007^{H}	28744690 29244690	XIRP2
	Hapmap53367-rs29014082^{M}	134031005 134531005	ALDH4A1, MRTO4, PAX7, TAS1R2, UBR4
3	Hapmap42294-BTA-69421^{M}	6466282 6966282	CCDC190, DDR2, HSD17B7, UAP1
	ARS-BFGL-NGS-5408^{M}	91507984 92007984	BSND, DHCR24, TMEM61, USP24
	Hapmap38245-BTA-69529^{H}	111495557 111995557	GJA4, GJB3, GJB4, GJB5
4	ARS-BFGL-NGS-110046^{L}	84938654 85438654	KCND2
5	ARS-BFGL-NGS-7567^{H}	28081294 28581294	ACRV1B, ACVRL1, ANKRD33, FIGNL2, SCN8A, SLC4A8
	ARS-BFGL-NGS-7003^{L}	54495039 54995039	LRIG3
	Hapmap43797-BTA-17694^{L}	92277538 92777538	RERGL
	ARS-BFGL-NGS-61091^{H}	114448428 114948428	MCAT, PACSIN2, SAMM50, SCUBE1, TSPO, TTLL1, TTLL12
6	BTA-68404-no-rs^{L}	61273195 61773195	APBB2, NSUN7
	ARS-BFGL-NGS-41209^{L}	94339635 94839635	FRAS1, MRPL1
7	ARS-BFGL-NGS-111791^{L}	466475 966475	CNOT6, GFPT2, MAPK9

Table 3. (Continued)

BTA	Marker	Range	Gene
7	BTB-01697487^{H}	66395827 66895827	GRIA1
	BTA-108798-no-rs^{H}	97143157 97643157	ARSK, FAM81B, RFESD, RHOBTB3, SPATA9, TTC37
	BTA-88724-no-rs^{M}	97944828 98444828	PCSK
	ARS-BFGL-NGS-43521^{H}	111434918 111934918	MAN2A1
	ARS-BFGL-NGS-25970^{M}	17695352 18195352	CAMSAP3, CCL25, CERS4, CLEC4G, CTXN1, ELAVL1, MCEMP1, EVI5L, FBN3, FCER2, LRRC8E, MAP2K7, PCP2, RETN, SNAPC22, STXBP2, TIMM44
9	BTA-83229-no-rsL	10319900 10819900	OGFRL1
	ARS-BFGL-NGS-1777^{M}	82313961 82813961	PLAGL1, SF3B5, STX11, UTRN, ZC2HC1B
	ARS-BFGL-NGS-114754^{H}	104676053 105176053	C6H6orf120, ERMARD, PHF10, TCTE3, THBS2, WDR27
10	ARS-BFGL-NGS-10428^{L}	13572864 14072864	AAGAB SMAD3
	ARS-BFGL-NGS-116025^{H}	79246312 79746312	ATP6V1D, EIF2S1, FAM71D, MPP5, PLEK2, TMEM29B
	Hapmap41666-BTA-86780^{H}	79205690 79705690	ATP6V1D, EIF2S1, FAM71D, MPP5, PLEK2, TMEM229B
11	ARS-BFGL-NGS-104517^{L}	2466257 2966257	ACTR1B, ANKRD23, ANKRD39, ARID5A,CNNM3, CNNM4, FER1L5, LMAN2L,KANSL3,NEURL3,SEMA4, FAM178B
	Hapmap45640-BTA-113575^{M}	95387394 95887394	ADGRD2, ARPC5L, GOLGA1, NEK6, NR5A1, NR6A1, OLFML2A, PSMB7, WDR38
12	UA-IFASA-4065^{M}	11574890 12074890	RGCC, VWA8
	ARS-BFGL-NGS-44030^{L}	11615727 12115727	RGCC, VWA8
13	ARS-BFGL-BAC-7467^{H}	21340166 21840166	PLXDC2
	ARS-BFGL-BAC-15732^{L}	50700127 51200127	HAO1
14	Hapmap50324-BTA-34690^{H}	39390732 39890732	GDAP1, JPH1
15	ARS-BFGL-NGS-34422^{H}	40585081 41085081	MICALCL, MICAL2, PARVA
	ARS-BFGL-NGS-111222^{M}	58546348 59046348	BBOX1, CCDC34, LGR4, LIN7C
16	BTB-00623849^{M}	3179932 3679932	ELK4, MFSD4A, NUCKS1, PM20D1, RAB7B, RAB29, SLC26A9, SLC41A1, SLC45A3
	ARS-BFGL-NGS-39696^{M}	27379566 27879566	CAPN2, CAPN8, CCDC185, SUSD4, TP53BP2

Table 3. (Continued)

BTA	Marker	Range	Gene
16	ARS-BFGL-NGS-117800^{M}	62621926 63121926	ACBD6, CEP350, LHX4, QSOX1
17	ARS-BFGL-NGS-84420^{L}	66475091 66975091	CMKLR1, CORO1C, FICD, ISCU, SART3, SELPLG, TMEM119, WSCD2
18	BTA-42767-no-rs^{H}	20697807 21197807	TOX3
	BTA-101550-no-rs^{L}	38107061 38607061	ZFHX3
	BTA-43816-no-rs^{M}	55006431 55506431	BICRA, CABP5, CCDC114, CRX, EHD2, C19orf68, ELSPBP1, SELENOW, LIG1, TMEM143, ZNF114
	ARS-BFGL-NGS-33663^{L}	63262359 63762359	CNOT3, TSEN34, LENG1, MBOAT7, NLRP8, OSCAR, NLRP5, NLRP13, PRPF31, RPS9, TARM1, TFPT, TMC4, ZNF444, ZNF787
19	UA-IFASA-6791^{H}	5061065 5561065	TOM1L1, COX11
	ARS-BFGL-NGS-2406^{M}	20145816 20645816	ALDOC, FOXN1, IFT20, KIAA0100, NLK, PIGS, POLDIP2, SARM1, SEBOX, SLC13A2, SLC46A1, SPAG5, TMEM97, TMEM199, TNFAIP1, UNC119, VTN
21	ARS-BFGL-NGS-42945^{H}	44769363 45269363	EGLN3
	ARS-BFGL-NGS-43284^{M}	54846333 55346333	C14orf28, FKBP3, KLHL28, PRPF39, TOGARAM1
22	Hapmap24381-BTA-150366^{L}	9619416 10119416	ARPP21
	BTA-54658-no-rs^{M}	47215329 47715329	ACTR8, CHYDH, IL17RB, SELENOK
23	BTA-57141-no-rs^{L}	8263266 8763266	C6orf106,HMGA1,NUDT3,PACSIN1, RPS10, SNRPC, SPDEF
	BTA-57038-no-rs^{H}	20098228 20598228	MLIP, TINAG
	ARS-BFGL-NGS-114949^{L}	49920679 50420679	BPHL, ECI2, FAM217A, PXDC1, PRPF4B, PSMG4, TUBB2A, TUBB2B
24	ARS-BFGL-NGS-94157^{H}	7130047 7630047	CD226, RTTN, SOCS6
25	ARS-BFGL-NGS-73909^{M}	37982904 38482904	BAIAP2L1, BHLHA15, LMTK2, NPTX2
26	ARS-BFGL-NGS-63916^{M}	34186316 34686316	CASP7, DCLRE1A, HABP2, NHLRC2, NRAP, PLEKHS1
26	BTA-61759-no-rs^{H}	45180774 45680774	EDRF1, BCCIP, DHX32, TEX36, UROS
27	BTB-01746933^{L}	34454855 34954855	ADAM18, C8orf4, IDO1, IDO2
28	ARS-BFGL-NGS-79829^{L}	8251840 8751840	B3GALNT2, GNG4, LYST, NID1
	ARS-BFGL-NGS-42296^{M}	8925452 9425452	ACTN2, EDARADO, ERO1B, GPR137B, HEATR1, LGALS8

Table 3. (Continued)

BTA	Marker	Range	Gene
28	ARS-BFGL-NGS-5507^H	41056221 41556221	WAPL
29	ARS-BFGL-NGS-34288^H	25036242 25536242	NAV2
	ARS-BFGL-NGS-3949^M	29545486 30045486	CDON, DCPS,DDX25,FOXRED1, HYLS1, PATE2, PUS3, RPUSD4, SRPRA TIRAP

Genes reported in Table 3, already discussed in literature that could control feed efficiency are listed in Table 4. Four of these genes (the *CCR4-NOT Transcription Complex Subunit 6* (CNOT6), *Mitogen-Activated Protein Kinase 9* (MAPK9), the *Proprotein Convertase* Subtilisin*/Kexin Type 1* (PCSK1) and the *Resistin* (RETN)) are harbored in BTA 7. In detail the CNOT6 (in the region between 466 475-966 475 bp close to marker ARS-BFGL-NGS-111791), the RETN (located in the region ranged 17 695 352-18195352 close to ARS-BFGL-NGS-25970) are two important genes directly associated to feed efficiency.

BTA11 and BTA16 present other three interesting genes (*Fer-1 Like Family Member 5* (FERIL5), *Lectin*, *Mannose Binding 2 Like* (LMAN2L) and the *Proteasome Subunit Beta 7* (PSMB7) on BTA 11, *Calpain 8* (CAPN8), *LIM Homeobox 4* (LHX4) and *Solute Carrier Family 45 Member 3* (SLC45A3) on BTA 16). On BTA 1 were found two candidate genes associated with RFI (*Zic Family Member 1* (ZIC1) and *Zic Family Member 4* (ZIC4) close to marker Hapmap25578-BTA-148357 in the genome region ranging from 121 568 219 to 122 068 219 bp. On BTA3, close to the marker Hapmap42294-BTA, was annotated the gene UAP1 (*UDP-N-Acetylglucosamine Pyrophosphorylase 1*) that influences the sugar metabolism. On BTA15 (in the region between 58 546 348- 59 046 348 bp) the *Lin-7 Homolog C, Crumbs Cell Polarity Complex Component* (LIN7C) also was flagged. On BTA16, the *Calpain 8* (CAPN8) and the *Tumor Protein P53 Binding Protein 2* (TP53BP2) associated with the marker ARS-BFGL-NGS-39696 (on region between 27 379 566 and 27 879 566 bp) were found. Even on BTA 16, the annotated sequence close to marker BTB-00623849 showed the presence of SLC45A3 (*Solute Carrier Family 45 Member 3*) related with transport of glucose and other sugars. In the BTA 18, the gene TSEN34 (*TRNA Splicing Endonuclease Subunit 34*), near the marker ARS-BFGL-NGS-33663 in the region from 63 262 359 to 63 762 359 bp was found. It has potential roles in gene transcription. Finally, in the BTA 27, close to marker BTB-01746933(in the genome region from 34 454 855 to 34 954 855 bp) the ADAM18 (*DAM Metallopeptidase Domain 18*) that is associated with muscle development was found.

Table 4. Candidate protein-coding genes within 2.5 Mb of significant Markers for traits underlying nutrient repartitioning

	Gene	BTA	Category	Refence
Beef Cattle				
ADG[1]	TMEM229B	10	Nellore steers and young bulls	Santana et al., 2014
	CAPN8, TP53BP2	16	Nellore steers and young bulls	Olivieri et al., 2016
BW[2]	TSEN34	18	Multibreeds steers	Kern et al., 2016
FCR[4]	LIN7C	15	Nellore steers	de Oliveira et al., 2014
FE[5]	XIRP2	2	Multibreeds steers	Seabury et al., 2017
HRFI[6]	SLC45A3	16	Nellore steers	Tizioto et al., 2015
MBW[7]	CNOT6	7	Multibreeds steers	Seabury et al., 2017
RFI[8]	ZIC1 ZIC4	1	Nellore	Olivieri et al., 2016
	SPDEF	23	Angus sters	Weber et al., 2015
RIG[9]	FER1L5	11	Multibreeds steers	Serão et al., 2013
GT[11]	LHX4	16	Chinese multibreeds	Liu et al., 2010 Ren et al., 2014
	LRIG3	5	Beef and Dairy Multibreeds	Xu et al., 2014
	PAX7	2	Multibreeds	Coles et al., 2015
	PCSK1	7	Jiaxian calves	Sun et al., 2014
	RETN		Chinese multibreeds	Gao et al., 2011
Dairy Cattle				
DMI[3]	PLEKHS1	26	Holstein	Hardie et al., 2017
HRFI[6]	MAPK9	7	Holstein and Jersey	Salleh et al., 2017
MBW[7]	CABP5 CCDC114 ELSPBP1 TMEM143 ZNF114	18	Holstein	Hardie et al., 2017
NT[10]	UAP1	3	German Holstein and Charolaise bulls	Schwerin et al.2006
RFI[8]	LMAN2L	11	Holstein Mid-lactation	Yao et al., 2013
	ADAM18	27	Holstein	Hardie et al., 2017
GT[11]	PSMB7	11	Holstein bulls	Sadkowski et al., 2008
	MICAL2	15	Multibreeds	Taye et al., 2017

ADG[1] =average daily gain; BW[2] =body weight; DMI[3] =dry matter intake; FCR[4] =feed conversation ratio; FE[5] =feed efficiency; HRFI[6] =high residual feed intake; MBW[7] =metabolic body weight; RFI[8] =residual feed intake; RIG[9] =residual intake gain; NT[10] =nutrient transformation; GT[11]=other growth traits.

3.5 Discussion

Marker selection

In the present research, the DAM approach to develop a GWAS for RCI was proposed. The algorithm was able to overcome two of the most important drawbacks that affect the traditional single SNP regression approach. The first regards the correction of p-values to control the multiple testing error rate. A severe correction enlarges the probability to obtain false negative associations. On the contrary, a weak correction could produce a great number of false positive associations. In the DAM no test for the single SNP is developed and, in consequence, no p-value correction is due. The second problem regards the small fraction of the genetic variance explained by each single SNP (Visscher et al., 2010) when quantitative traits are analyzed. The DAM, being based on multivariate techniques, handle all markers simultaneously thus accounting for most of the genetic variance. Results of the present study confirm the goodness of the DAM algorithm. The GWAS developed with the traditional single marker regression selected only one SNP significantly associated with RCI. This marker flagged a region of the genome harboring three putative genes that regulate, even if indirectly, the feed efficiency (Olivieri et al., 2016). However, it does not appear realistic that only a single region of genome was related to RCI. Probably this result is linked with the above mentioned drawbacks that affect the univariate approach.

The DAM selected 382 markers distributed across the genome (Figure1) that significantly separated the LRCI from the HRCI group in the CDA (p-value <0.0001). Figure 2 highlights a clear separation between groups. In particular, distances between the group centroids were much larger than those within groups. For this reason, groups depicted in Figure 2 appear nearly as a single point. The DA developed by using the 382 SNPs correctly assigned all bulls to the true group of origin, also in the bootstrap procedure. Actually, both Brown Swiss bulls and cows are currently genotyped by breeder associations or research centers. The 382 DAM selected SNPs could be used as variables in a DA to assign animals to LRCI or HRCI when their RCI phenotype is not known. This could be useful to select individuals that efficiently convert the concentrate fraction of the diet.

Among the 382 DAM selected markers, the most discriminant 88 SNPs were sufficient to obtain a significant (p-value <0.0001) separation between the two RCI groups. The graph in Figure 3 depicts the two RCI groups. Individuals belonging to LRCI are in the negative side of the graph, whereas animals belonging to HRCI are in the positive side. This structure indicates that the CAN, when is applied to the single individual in the DA, produces a negative score for animals belonging to LRCI and positive scores for the others. Table 3 lists the 88 SNPs with their CCs classified in three groups according to their values: low (L), medium (M) and high (H). In Figure 4, the same CC values are plotted. Figures 3 and 4 and Table 3 should be analyzed simultaneously. The M group, reported in gray both in Figure 4 and in Table 3, contains markers whose CCs have a low absolute value. These SNPs cannot be considered acting mainly on the LRCI or on the HRCI group. The L and H groups, reported in black, are composed by markers with high CC absolute values. In particular, L markers (the negatives) have a greater weight in composing the CAN for animals belonging to LRCI, whereas H markers act on the contrary. These results suggest that L markers favor the efficiency in converting the concentrate fraction of the diet (the more negative the RCI the more the efficiency) whereas the H markers oppose to it.

Gene discovery

The genetic variability existing among individuals of a population in ingesting, digesting and assimilating foods may be the cause of the different growth potentials. In livestock, nutrient repartitioning and utilization as well as growth and fat accumulation are considered physiological actions regulated by endogenous factors (for example hormones) and exogenous factors (for example diet). Since, nutritional turnover is crucial in productive performance in both dairy and meat bovines, identify QTL and genes responsible for the genetic variation in nutrient transformation represents a challenge for animal breeding. The present study aims at identifying candidate genes for concentrate transformation in Italian Brown growing calves. In this survey, several chromosome regions associated with nutrient transformation traits were identified using the DAM approach.

Among candidate genes identified as associated with feed efficiency (Table 4), some of them were already reported in literature. The *UDP-N-acetylglucosamine pyrophosph-orylase 1* (*UAP1*) was found differentially expressed in liver tissue of growing Charolais compared with German Holstein (Schwerin et al., 2006) and involved in the regulation of hormonal levels in Holstein (Xi et al., 2015). *Calpain 8* (*CAPN8)* and *Tumor Protein P53 Binding Protein 2* (TP53BP2) are two loci associated with the average daily gain in a study looking for genomic regions associated with feed efficiency traits in Nellore cattle (Olivieri et al., 2016). In the same paper also *Zic family member 1* (ZIC1) and *Zic family member 4* (ZIC4) were found associated with RFI. The gene *Solute Carrier Family 45 Member 3* (SLC45A3) involved in the transport of glucose and other sugars was also found by Tizioto et al., (2015). Authors divided Nellore cattle in two groups, one with high residual feed intake (HRFI) and one with low residual feed intake groups (LRFI). The gene was found down-regulated in the (HFRI) group. The gene *Lin-*

7 Homolog C, Crumbs Cell Polarity Complex Component (LIN7C), also discussed by de Oliveira et al. (2014) plays a role in the modulation of adiposity in mammals and was found in the feed conversion ratio trait in a study of genomic regions associated with feed efficiency.

In this study, several interesting gene were found associated with body gain traits (GT). Among this, the most fascinating was the *Resistin* (RETN). This gene was an adipose-specific secreted protein down regulated in adipose tissue in mouse. The protein encoded by this locus is involved in the mechanisms of obesity-related insulin resistance in mammals (Steppan and Lazzar 2002). Recently, this RETN was also found associated with meat quality traits in Chinese Bos taurus (Gao et al., 2011).

In the LCI group the marker ARS-BFGL-NGS-33663L identified the region where the TSEN34 locus was annotated. This gene was already reported as significantly associated with feed intake and body gain traits in a comparative study between low and high gain phenotypes in beef cattle (Kern et al., 2016). In addition, Kern et al. (2016) identified the same gene as down-regulated in low gain group.

A positional candidate gene associated with the LCI group was the CNOT6. In mammals, this gene is involved in the cellular growth and senescence (Morita at al., 2007; Mittal et al., 2011). In cattle, Seabury et al. (2017) found CNOT6 as involved in the metabolic body weight at midpoint of trial. This parameter is important in studies about the feed efficiency, because, commonly it is used to calculate the RFI.

Finally, one gene worthy of note associated with RFI, the ADAM18, was obtained in the present survey. This gene has been implicated in a variety of biologic processes such as fertilization, muscle development, and neurogenesis in mammals. Hardie et al. (2017) found the ADAM18 as candidate protein-coding genes in a study regarding the genetic and biological basis of feed efficiency in mid-lactation Holstein dairy cows.

3.6 Conclusions

The use of the DAM algorithm in developing a GWAS gave good results in selecting markers associated with RCI. The 382 DAM selected SNPs were able to correctly classify animals in the two RCI groups, also after the bootstrap resampling cross validation. These SNPs could be used to develop a DA on genotyped animals without a known RCI phenotype, to assign individuals to the LRCI or HRCI group. The minimum number of markers able to correctly discriminate the two groups was 88. Only these SNPs were considered associated to the RCI and, inconsequence, submitted to gene discovery. A great number of putative genes were found in the regions flagged by the 88 most discriminant markers, thus confirming the effectiveness of the DAM algorithm for GWAS.

3.7 References

Berry, D. P. and, J. J. Crowley. 2013. Cell biology symposium: genetics of feed efficiency in dairy and beef cattle. Journal of Animal Science. 91(4):1594-1613.

Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes, W. Barendse and, B. E. Harrison. 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in, and composite beef cattle. Journal of Animal Science. 91(7):3088-3104.

Carstens, G. E. and, L. O. Tedeschi. 2006. Defining feed efficiency in beef cattle. In Proceedings of Beef Improvement Federation 38th Annual Research Symposium and Annual Meeting, Choctaw, Mississippi (pp. 12-21).

Coles, C. A., J. Wadeson, C. P. Leyton, J. P. Siddell, P. L. Greenwood, J. D. White and, M. B. McDonagh. 2015. Proliferation rates of bovine primary muscle cells relate to liveweight and carcase weight in cattle. PloS One. 10(4):e0124468.

Council for Agricultural Science and Technology (CAST). 2013. Animal Feed vs. Human Food: Challenges and Opportunities in Sustaining Animal Agriculture Toward 2050. Issue Paper 53. CAST, Ames, Iowa.

Connor, E. E., J. L. Hutchison, H. D. Norman, K. M. Olson, C. P. Van Tassell, J. M. Leith and, R. L. Baldwin. 2013. Use of residual feed intake in Holsteins during early lactation shows potential to improve feed efficiency through genetic selection. Journal of Animal Science. 91(8):3978-3988.

Crowley, J. J., M. McGee, D. A. Kenny, D. H. Crews, R. D. Evans and, D. P. Berry. 2010. Phenotypic and genetic parameters for different measures of feed efficiency in different breeds of Irish performance-tested beef bulls. Journal of Animal Science. 88(3):885-894.

Davis, S. R., K. A. Macdonald, G.C. Waghorn and, R. J. Spelman. 2014. Residual feed intake of lactating Holstein-Friesian cows predicted from high-density genotypes and phenotyping of growing heifers. Journal of Dairy Science. 97(3):1436-1445.

de Oliveira, P. S., A. S. Cesar, M. L. do Nascimento, A.S. Chaves, P. C. Tizioto, R. R. Tullio and, J. M. Reecy. 2014. Identification of genomic regions associated with feed efficiency in Nelore cattle. BMC Genetics. 15(1):100.

Efron, B. 1979. Bootstrap Methods: Another look at the jackknife. The Annals of Statistics. 7.1:1−26.

Gao, L., J. A. Ujan, L. Zan, M. Xue and, A. Camus. 2011. A novel polymorphism of resistin gene and its association with meat quality traits in Chinese Bos taurus. African Journal of Biotechnology. 10(57):1252-12256.

Green, T. C., J. G. Jago, K. A. Macdonald and, G. C. Waghorn. 2013. Relationships between residual feed intake, average daily gain, and feeding behavior in growing dairy heifers. Journal of Dairy Science. 96(5):3098-3107.

Hardie, L. C., M. J. VandeHaar, R. J. Tempelman, K. A. Weigel, L. E. Armentano, G. R. Wiggans and, M. D. Hanigan. 2017. The genetic and biological basis of feed efficiency in mid-lactation Holstein dairy cows. Journal of Dairy Science. 100(11):9061-9075.

Herd, R. M. and, P. F. Arthur. 2009. Physiological basis for residual feed intake. Journal of Animal Science. 87(E. Suppl.):E64‑E71.

Kern, R. J., A. K. Lindholm-Perry, H. C. Freetly, W. M. Snelling, J. W. Kern, J. W. Keele and, P. A. Ludden. 2016. Transcriptome differences in the rumen of beef steers with variation in feed intake and gain. Gene. 586(1):12-26.

Liu, J. X., G. Ren, H. Chen, F. Li, X. Y. Lan, M. J. Li and, J. Q. Wang. 2011. Five novel SNPs of the bovine LHX 4 gene and their association with growth traits in native Chinese cattle breeds. Animal Science Papers and Reports. 29(1):19-28.

Macciotta, N. P. P., G. Gaspa, L. Bomba, D. Vicario, C. Dimauro, M. Cellesi and, P. Ajmone-Marsan. 2015. Genome-wide association analysis in Italian Simmental cows for lactation curve traits using a low-density (7K). Journal of Dairy Science. 98(11):8175-8185.

Macdonald, K. A., J. E. Pryce, R. J. Spelman, S. R. Davis, W.J. Wales, G. C. Waghorn and, B. J. Hayes. 2014. Holstein-Friesian calves selected for divergence in residual feed intake during growth exhibited significant but reduced residual feed intake divergence in their first lactation. Journal of Dairy Science. 97(3):1427-1435.

Mardia, K.V., J.T. Kent and, J. M. Bibby. 2000. Multivariate Analysis. Academic Press, London Morrison, F. 1976. Multivariate statistical methods. McGraw-Hill, New York, NY.

Mittal, S., A. Aslam, R. Doidge, R. Medica and, G. S. Winkler. 2011. The Ccr4a (CNOT6) and Ccr4b (CNOT6L) deadenylase subunits of the human Ccr4–Not complex contribute to the prevention of cell death and senescence. Molecular biology of the cell. 22(6):748-758.

Morita, M., T. Suzuki, T. Nakamura, K. Yokoyama, T. Miyasaka and, T. Yamamoto. 2007. Depletion of mammalian CCR4b deadenylase triggers elevation of the p27Kip1 mRNA level and impairs cell growth. Molecular and cellular biology. 27(13):4980-4990.

Olivieri, B. F., M. E. Z. Mercadante, J. N. D. S. G. Cyrillo, R. H. Branco, S. F. M. Bonilha, L. G. de Albuquerque and, F. Baldi. 2016. Genomic regions associated with feed efficiency indicator traits in an experimental Nellore cattle population. PloS One. 11(10):e0164390.

Potts, S. B., J. P. Boerman, A. L. Lock, M. S. Allen and, M. J. VandeHaar. 2015. Residual feed intake is repeatable for lactating Holstein dairy cows fed high and low starch diets. Journal of Dairy Science. 98(7):4735-4747.

Pryce, J. E., J. Arias, P. J. Bowman, S. R. Davis, K. A. Macdonald, G. C.Waghorn and, B. J. Hayes. 2012. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. Journal of Dairy Science. 95(4):2108-2119.

Pryce, J. E., O. Gonzalez-Recio, J. B. Thornhill, L. C. Marett, W. J. Wales, M. P. Coffey and, B. J. Hayes. 2014. Validation of genomic breeding value predictions for feed intake and feed efficiency traits. Journal of Dairy Science. 97(1):537-542.

Ren, G., Y. Z. Huang, T. B. Wei, J. X. Liu, X. Y. Lan, C. Z. Lei and, H. Chen. 2014. Linkage disequilibrium and haplotype distribution of the bovine LHX4 gene in relation to growth. Gene. 538(2):354-360.

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure, S. L. Northcutt and, R. L. Weaber. 2012. Genome-wide association analysis for feed efficiency in Angus cattle. Animal Genetics. 43(4):367-374.

Sadkowski, T., M. Jank, L. Zwierzchowski, E. Siadkowska, J. Oprządek and, T. Motyl. 2008. Gene expression profiling in skeletal muscle of Holstein-Friesian bulls with single-nucleotide polymorphism in the myostatin gene 5'-flanking region. Journal of Applied Genetics. 49(3):237-250.

Salleh, M. S., G. Mazzoni, J. K. Höglund, D. W. Olijhoek, P. Lund, P. Løvendahl and, H. N. Kadarmideen. 2017. RNA-Seq transcriptomics and pathway analyses reveal potential regulatory genes and molecular mechanisms in high-and low-residual feed intake in Nordic dairy cattle. BMC Genomics. 18(1):258.

Sainz, R.D. and, P. V. Paulino. 2004. Residual feed intake.UC Davis: Sierra foothill research and extension center. Retrieved http://escholarship.org/uc/item/9w93f7ks.

Santana, M. H. A., Y. T. Utsunomiya, H. H. R. Neves, R. C. Gomes, J. F. Garcia, H. Fukumasu and, J. B. S. Ferraz. 2014. Genome-wide association study for feedlot average daily gain in Nellore cattle (Bos indicus). Journal of Animal Breeding and Genetics. 131(3):210-216.

Schwerin, M., C. Kuehn, S. Wimmers, C. Walz and, T. Goldammer. 2006. Trait-associated expressed hepatic and intestine genes in cattle of different metabolic type–putative functional candidates for nutrient utilization. Journal of Animal Breeding and Genetics. 123(5):307-314.

Seabury, C. M., D. L. Oldeschulte, M. Saatchi, J. E. Beever, J. E. Decker, Y. A. Halley and, H. Yampara-Iquise. 2017. Genome-wide association study for feed efficiency and growth traits in US beef cattle. BMC Genomics.18(1):386.

Serão, N. V., D. González-Peña, J. E. Beever, D. B. Faulkner, B. R. Southey and, S. L. Rodriguez-Zas. 2013. Single nucleotide polymorphisms and haplotypes associated with feed efficiency in beef cattle. BMC Genetics. 14(1):94.

Soder, K. J. and, C. A. Rotz. 2001. Economic and environmental impact of four levels of concentrate supplementation in grazing dairy herds. Journal of Dairy Science. 84(11):2560-2572.

Steinfeld, H. and, C. Opio. 2010. The availability of feeds for livestock: Competition with human consumption in present world. Advances in Animal Biosciences. 1(2):421-475.

Steppan, C. M. and, M. A. Lazar. 2002. Resistin and obesity-associated insulin resistance. Trends in endocrinology & Metabolism. 13(1):18-23.

Sun, J., L. Shan, C. Zhang and, H. Chen. 2015. Haplotype combination of the bovine PCSK1 gene sequence variants and association with growth traits in Jiaxian cattle. Journal of Genetics. 94(1):123-129.

Taye, M., W. Lee, S. Jeon, J. Yoon, T. Dessie, O. Hanotte and, H. K. Lee. 2017. Exploring evidence of positive selection signatures in cattle breeds selected for different traits. Mammalian Genome. 1-14.

Tizioto, P. C., L. L. Coutinho. J. E. Decker, R. D. Schnabel, K. O. Rosa, P. S. Oliveira and, D. P. Lanna. 2015. Global liver gene expression differences in Nelore steers with divergent residual feed intake phenotypes. BMC Genomics. 16(1):242.

van Zanten, H. H., H. Mollenhorst, C. W. Klootwijk, C. E. van Middelaar and, I. J. de Boer. 2016. Global food supply: land use efficiency of livestock systems. The International Journal of Life Cycle Assessment. 21(5):747-758.

Visscher, P. M., J. Yang and, M. E. Goddard. 2010. A commentary on 'common SNPs explain a large proportion of the heritability for human height'by Yang et al.(2010). Twin Research and Human Genetics. 13(6):517-524.

Waghorn, G. C., K. A. Macdonald, Y. Williams, S. R. Davis and, R. J. Spelman. 2012. Measuring residual feed intake in dairy heifers fed an alfalfa (Medicago sativa) cube diet. Journal of Dairy Science. 95(3):1462-1471.

Weber, K. L., B. T. Welly, A. L. Van Eenennaam, A. E. Young, L. R. Porto-Neto, A. Reverter and, G. Rincon. 2016. Identification of gene networks for residual feed intake in Angus cattle using genomic prediction and RNA-seq. PloS One. 11(3):e0152274.

Williams, Y. J., J. E. Pryce, C. Grainger, W. J. Wales, N. Linden, M. Porker and, B. J. Hayes. 2011. Variation in residual feed intake in Holstein-Friesian dairy heifers in southern Australia. Journal of Dairy Science. 94(9):4715-4725.

Xi, Y. M., Z. Yang, F. Wu, Z. Y. Han and, G. L. Wang,. 2015. Gene expression profiling of hormonal regulation related to the residual feed intake of Holstein cattle. Biochemical and Biophysical Research Communications. 465(1):19-25.

Xu, L., D. M. Bickhart, J. B. Cole, S. G. Schroeder, J. Song, C. P. V. Tassell and, G. E. Liu. 2014. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. Molecular Biology and Evolution. 32(3):711-725.

Yao, C., D. M. Spurlock, L. E. Armentano, C. D. Page, M. J. VandeHaar, D. M. Bickhart and, K. A. Weigel. 2013. Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. Journal of Dairy Science. 96(10):6716-6729.

CHAPTER 4

USE OF DISCRIMINANT ANALYSIS TO EARLY DETECT LACTATION'S PERSISTENCY IN DAIRY COWS

4.1 Abstract

The aim of the present research was to develop an algorithm to early identify dairy cows that, having a persistent lactation, might be destined to have a long lactation. The insemination of these cows could be delayed in order to obtain a number of health benefits and an improvement of fertility. Data consisted of 2,294 lactations belonging to primiparous (1,015) and multiparous (1,279) cows. They were grouped into three production classes based on the milk yield at 305 DIM: low class (LC) with milk production <20 kg, middle class (MC) with milk production between 20 kg and 32 kg, and high class (HC) with milk production >32 kg. The lactations considered suitable to become a long lactation belonged either to MC or to HC.

Four different lactation curve models (Wood, Ali & Schaeffer, Legendre Polynomials and 4th Degree Polynomials) were fitted to individual lactations by using the first 90, 120 and 150 DIM. The regression coefficients obtained in each model were used as variables in two multivariate discriminant techniques. For each parity, the Canonical Discriminant Analysis (CDA) was used to test for possible differences between the two production classes. The Discriminant Analysis (DA) was exploited to assign the animals in the two extreme production classes (LC and HC). In order to validate the results, the dataset was randomly divided into training and validation datasets. This partition was iterated 5,000 times by using a bootstrap procedure. The CDA significantly separated the two production classes for each parity. Among the different lactation models, the 4th degree polynomials were those that better assigned animals in the bootstrap procedure. In particular, by using the first 150 days of lactation, the error in assigning animals to the two production classes was 10% for primiparous and 13% for multiparous. Error slightly increased when 120 days of lactation were used: 12% and 17% for primiparous and multiparous, respectively.

4.2 Introduction

In intensive dairy farms, maintain a seasonal calving pattern is becoming difficult. The one-year-one-lactation per cow is possible only if the lactation length is around 305 days which means that cows should be inseminated approximately two months after parturition (Steri et al., 2012). However, there are different opinions about the optimal time for the first insemination, because both premature and late inseminations can cause milk losses (De Vries et al., 2006). Table 1 summarizes some of these opinions.

Table 1. Optimum period for fist insemination

DIM	Authors	
30-60	Dijkhuizen et al., 1985; Holmann et al., 1984; Strandberg and Oltenacu, 1989	All parities
70 or 110-130	Bar-Anan and Soller, 1990; Weller et al., 1985	Primiparous
41-90	Bar-Anan and Soller, 1990; AA.VV., Cited by Arbel et al., 2001	Pluriparous
80-120	Stevenson et al.,2007	All parities
150	Arbel et al., 2001	Seasonal
It depends	Heimann, 1984; Van Amburgh et al., 1997; Allore and Erb, 2000; De Vries, 2006; Inchaisri et al., 2011	Individual

According to Table 1 the suggested periods of time for the first insemination are very different and ranges from 30 to 150 DIM (days in milking), depending also by parity. If a cow, for different reasons, become pregnant after 120-150 DIM or more, it will have a lactation over the traditional 305 days. In ordinary dairy cow herds this happens quite frequently. VanRaden et al. (2006) reported that more than 55% of US Holsteins have lactations longer than 305 days. However, if the cow does not maintain a high milk production after the traditional temporal limit, profit losses could occur. In high-yielding Holstein primiparous cows, Mellado et al. (2016) found that the average milk production per day was around 32 kg during the first 305 DIM and 30 kg in the subsequent 253 DIM. Similar proportion of milk yield was reported for multiparous cows (around 35 and 32 kg for the first and second lactation period, respectively). However, both for primiparous and multiparous cows, a standard deviation around 10 kg was reported. This result indicates that, in an ordinary herd after the standard 305 DIM, a number of animals could have a production too low to have an economically convenient long lactation.

Besides an extension of the lactation after the standard period, a long lactation can be achieved through a voluntary delay of insemination. Furthermore, this practice can bring to a number of benefits for the farm. First, a cow could have longer time at disposal to restart the normal ovarian cyclicity with a consequent reduction of hormonal treatments to control anestrus. Butler et al. (2010) studied the reproductive failure in Holstein cows and found that, with a calving interval of 12 and 24 months, the mean number of services per cow was 2.8 and 1.8, respectively. Inchaisri et al. (2010) reported that the probability of insemination successes tend to increase with the passing of lactation. They also found that positive inseminations before peak yield (PY) were 6% lower than those after the peak. No difference between primiparous and multiparous cows was found.

A longer calving interval leads to better insemination performances and, therefore, to a reduction of cows culled because they do not become pregnant. This result can be very important for the farm management, because the replacement heifers are not often available when the involuntary culling occurs.

However, a voluntary lengthening of the calving interval has to be economically sustainable for farmers. Apart from sanitary savings, animals should have a suitable milk production over the standard 305 DIM. Van Raden et al. (2006) estimated that a daily milk yield higher than 13.6 kg could be considered economically viable.

A farmer who accepts to have some animals with a long lactation in his herd should know in advance which cows will have a high persistent lactation in order to both assign them in the respective group and inseminate them later. Mathematical models that are currently used to describe the lactation curve could be useful to develop an algorithm able to early ascertain if a cow would have a highly persistent lactation or not. Persistency, however, depends on the slope of the curve after peak yield and, by using the ordinary statistical techniques (the regression, for example), it can be evaluated only in the late lactation.

In the present research, a new multivariate statistical approach to early estimate the persistency of dairy cows was proposed. The algorithm combines the talent of curve models in depict features of the lactation and the ability of multivariate statistical techniques to distinguishing differences between groups. In this case, groups were lactations with low and high persistency. Only milk production data recorded in early lactation (not more than 150 DIM) was used in all analyses.

4.3 Materials and methods

The data

Data consisted of individual milk test days (TD) supplied by two farms located in Italy (Arborea, Oristano) and in Hungary (Tiszaalpar, Bács-kiskun). The first farm contributed to data with 1,526,934 TD recorded from 2001 to 2008, the second with 271,359 TD recorded from 2008 to 2016. In each farm, the milk production was supervised by the software Afifarm (Afimilk, Kibbutz Afikim Israel) which allows to obtain production data directly from the milking machine. In addition to the daily milk production, the software provides information about parity, fertility, health status and any events that can occur in cow's life. A lactation was considered for further analyses if it had records ranging from the 10th DIM to the 305th DIM. The 2,294 lactations that matched these requirements were divided in two groups. The first group contained 1,015 lactations that belonged to primiparous cows (first parity group, FPG), the second group contained the remaining 1,279 lactations that belonged to multiparous cows (multiparous parity group, MPG).

Lactation curves model

Four mathematical models available in literature were selected to fit the average and individual lactation curves.

The incomplete gamma function of Wood (1967) $Y_t = at^b e^{ct}$ that is the most popular model among lactation curves (Silvestre et al., 2006; Steri et al., 2012). Parameters *a, b, c* define the shape and the height on the ordinate axis and can be combined to calculate some of the lactation curve characteristics:

- Peak yield (PY): $PY = -\frac{a}{\left(\frac{c}{b}\right)^{b} e^{b}}$
- Time to peak yield (T_{PY}): $T_{PY} = -\frac{b}{c}$
- Persistency: $PERS = -(b+1)\ln c$

The five - parameter polynomial regression of Ali and Schaeffer (1987) (A&S):

$$Y_t = a + b\ (t/305) + c\ (t/305)^2 + ln\ (305/t) + k[ln\ (305/t)]^2$$

where *a* is a parameter associated with PY, *b* and *c* are parameters associated with slope in the decreasing phase, instead *d* and *e* are parameters associated with increasing slope in the phase until PY (Silvestre et al., 2006);

The fourth-order Legendre orthogonal polynomial (Legendre):

$$Y_t = a_0P_0 + a_1P_1 + a_2P_2 + a_3P_3 + a_4P_4$$

where α_i are parameters to determine and P_j were calculated with values published by Schaeffer (2004).

The 4th degrees polynomials (4th Polynomials):

$$Y_t = a + b\ t + c\ t^2 + d\ t^3 + e\ t^4$$

where *a, b, c, d* and *e* are parameters to determine and *t* are the DIM.

Lactations were first explored by using Wood. The model was applied to each lactation and the milk yield at 305 DIM was calculated. Based on these values, lactations were grouped into three classes of production: low class (LC) with milk production lower than 20 kg, middle class (MC) with milk production between 20 kg and 32 kg, and high class (HC) with milk product greater than 32 kg. A lactation was considered suitable to become a long lactation if it belonged to MC or HC. The regression Wood's parameters and their combinations were submitted to ANOVA to test for possible differences between parities and, for each parity, among classes of production.

The four models were also used to fit individual lactations. The goodness of fit was evaluated using the adjusted coefficient of determination (AdjRSQ) calculated by the following equation:

$$AdjRSQ = \frac{(n-1)R^2 - (p-1)}{n-p}$$

where n is the number of observations and p is the number of parameters of the model.

Discriminant procedures

The objective of this research was to develop an algorithm to early predict if a cow, after 305 DIM, might have an high milk production by using milk data recorded until 90, 120 and 150 DIM. For this reason, two multivariate statistical techniques were exploited: the canonical discriminant analysis (CDA) and the discriminant analysis (DA). CDA is a multivariate statistical technique which allows researchers to ascertain, by using a particular set of variables, if two or more groups of objects belongs to different populations or not. Unlike the cluster analysis, in the CDA the group an individual belongs is known. If k is the number of involved groups, the CDA derives $k-1$ linear equations, called canonical functions (CAN) that are used to assign objects to groups. The statistical significance in group separation can be evaluated by means of the Mahalanobis distance and the corresponding Hotelling's T-square test (De Maesschalck et al., 2000).

DA is a multivariate technique capable to classify objects into one of the involved groups. In this case, an individual is assigned to a particular group if its discriminant score produced by the CANs is lower than the cutoff value obtained by calculating the weighted mean distance among group centroids (Mardia et al., 2000).

The four lactation models where repeatedly applied to data by using only the first 90, 120 and 150 DIM and the estimated regression parameters were used as variables in the discriminant procedures. To validate the obtained CANs, the complete dataset was randomly divided into training and validation, in the proportion of four to one. This partition of data was iterated 5000 times by using a bootstrap procedure (Efron, 1979). At each run, DA was applied to the training dataset to assign animals in the validation dataset. Errors in assigning individuals to the two groups were recorded.

According to the aims of this research, it is crucial that errors in assigning a low yielding cow to HC are avoided. In other words, we could tolerate incorrect assignations of high yielding cows because the practical consequence will be only a long lactation less in the herd. On the contrary, if a low yielding cow is assigned to HC, possible losses of profit can occur. Finally, a cow belonging to MC can be assigned, without distinction, to LC or HC. For this reason, only animals which belonged to LC and HC were involved in the discriminant procedures.

4.4 Results

The mean lactation curves for FPG and MPG are displayed in Figure1. FPG had a PY (around 32 kg) lower than MPG (around 40 kg) with a greater PERS (7.3 vs 6.7). FPG reached the PY at the 90th day after calving, whereas the MPG showed the PY after the 50th day.

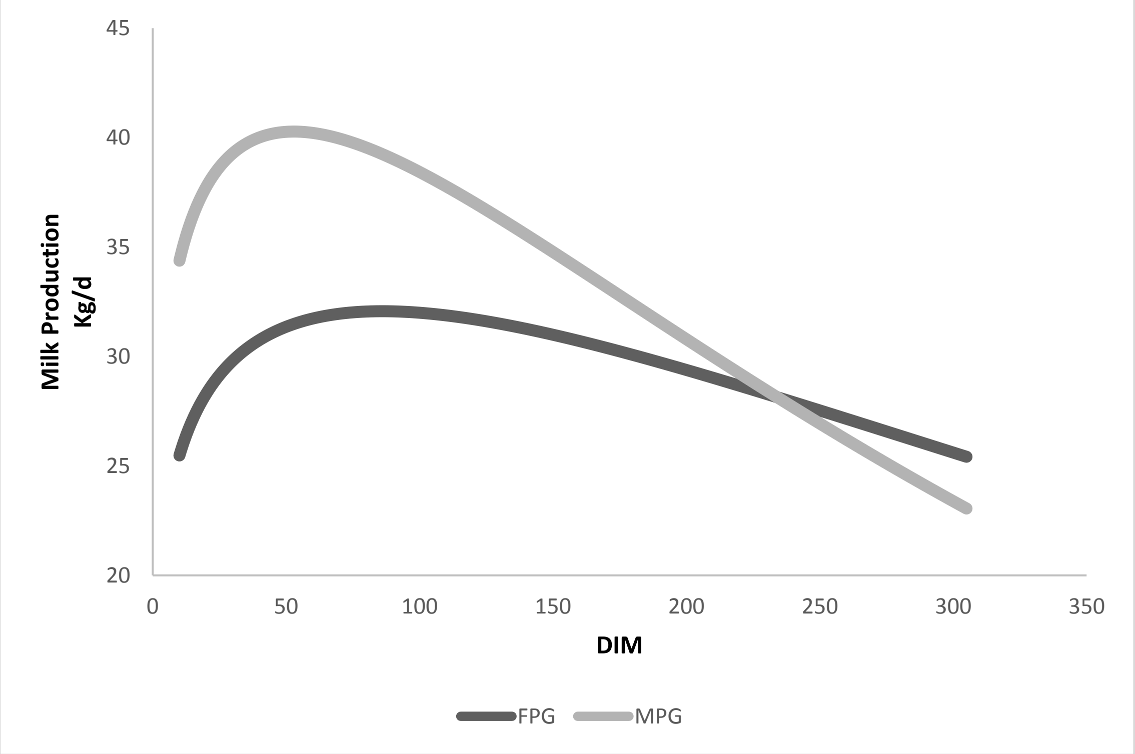

Figure 1. Average lactation curves obtained by fitting the Wood's model for FPG and SPG

Figure 2 shows the distribution of lactations in the three production classes. Most of the FPG lactations belonged to the MC class. Only a small percentage of them was in LC (around 17%) and in HC (around 7%). For the MPG group, half lactations were in MC and around 40% in LC. Few lactations, around 7%, were in the HC class.

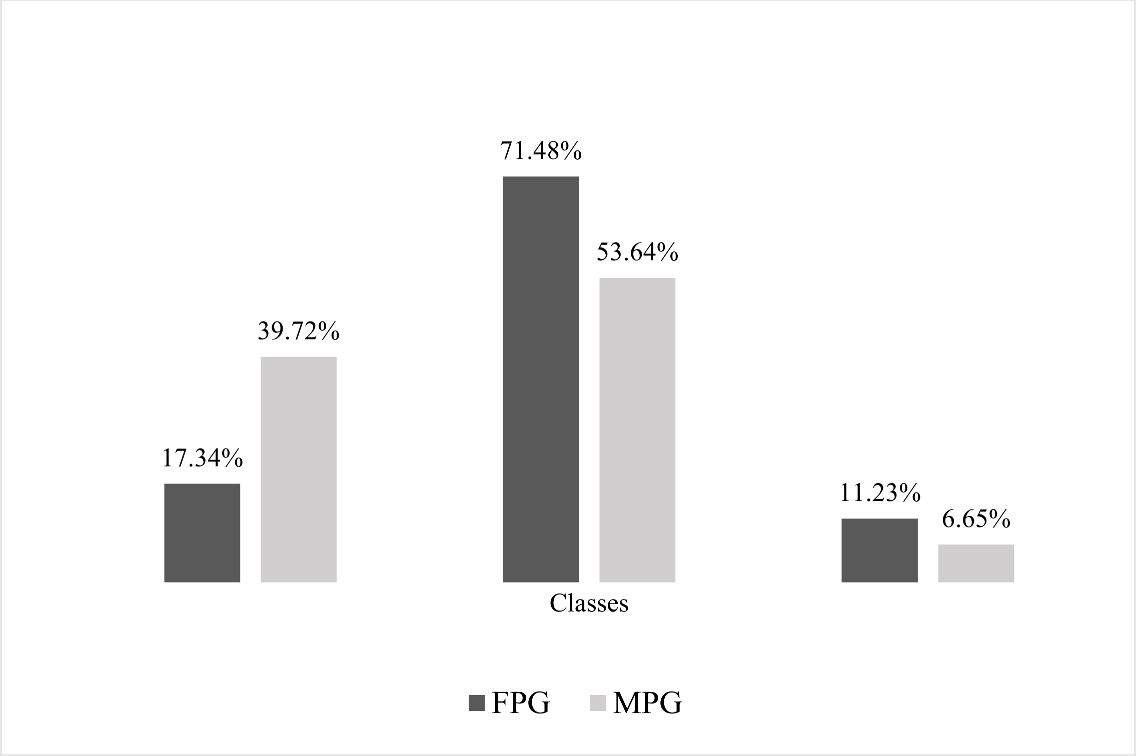

Figure 2. Lactation distribution among the three classes of production (LC <20 kg, MC >20 kg and >32 kg, HC >32 kg) at 305 DIM, for the two parity groups (SPG and MPG)

Figures 3 and 4 depict the pattern of lactation curves for the three production classes, separately for FPG (Figure 3) and MPG (Figure 4).

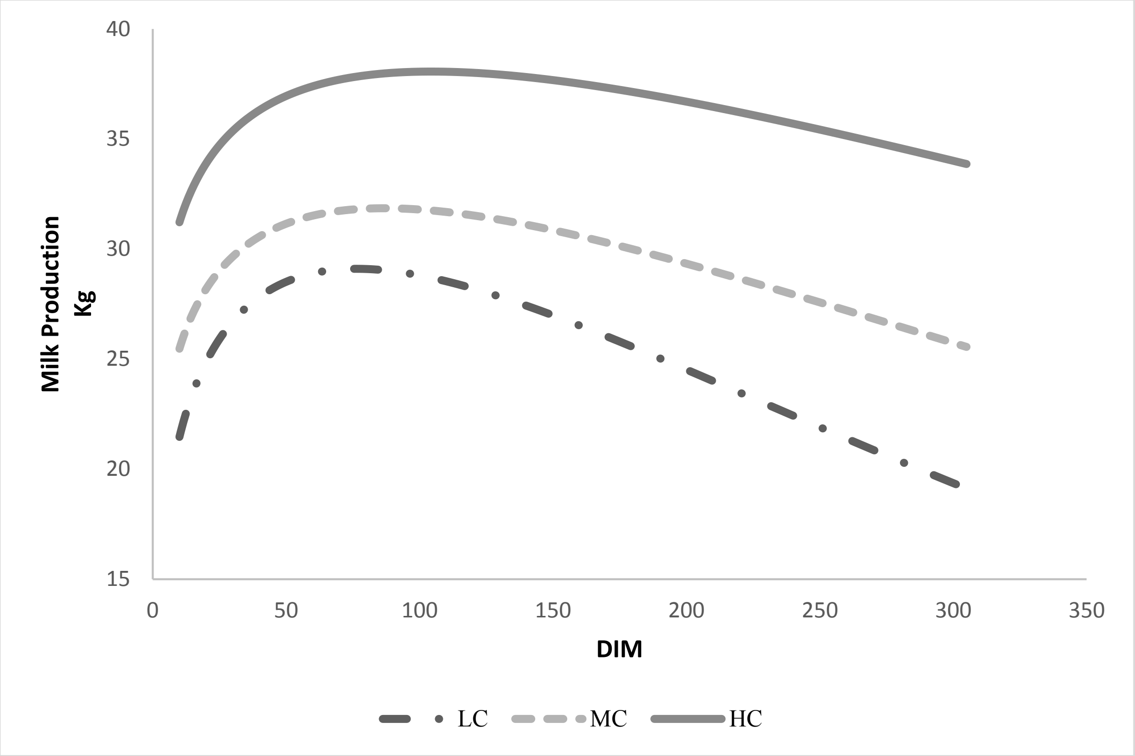

Figure 3. Average lactation curves obtained by fitting the Wood's model to the three classes of production in FPG

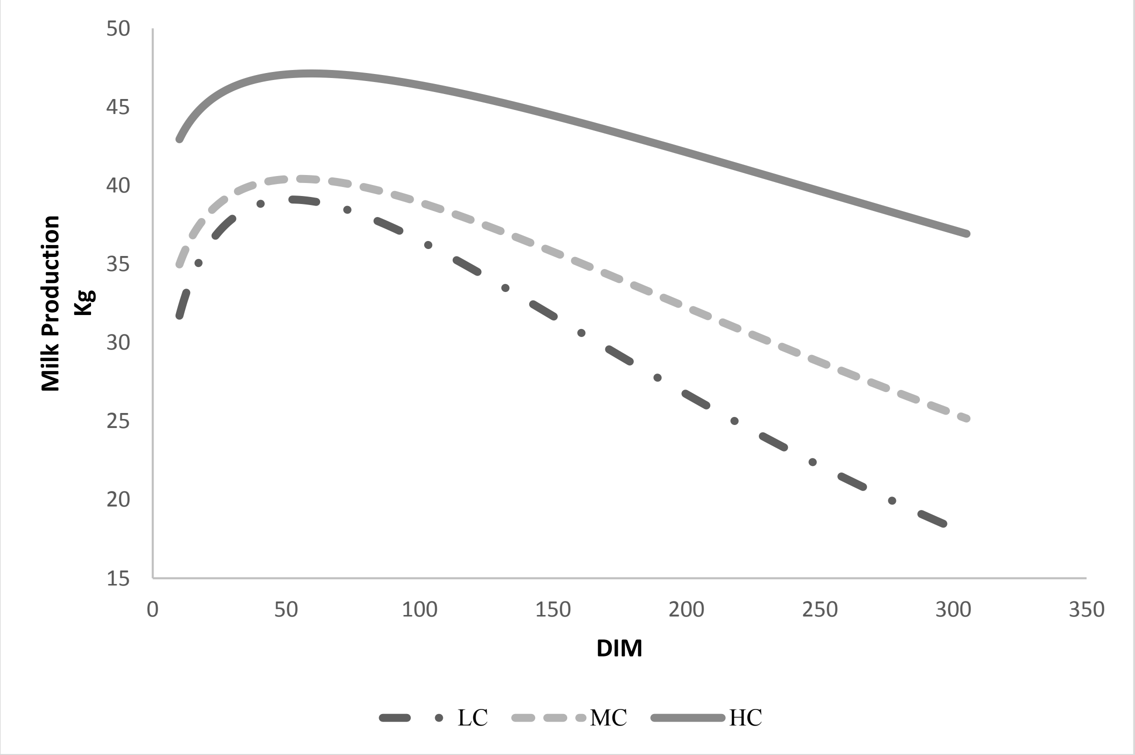

Figure 4. Average lactation curves obtained by fitting the Wood's model to the three classes of production in MPG

Correlations between the milk yield at 305 DIM and those at 400, 500 and 600 DIM were almost all above 90% (Table 2), except for the correlation between 305 and 600 DIM for primiparous cows which was 85%. This result indicates that the production at 305 DIM can be considered a good indicator for the production in the subsequent stages of lactation.

Table 2. Correlations between milk yield at 305 DIM and at 400, 500 and 600 DIM.

		FPG	
	400 DIM	500 DIM	600 DIM
305 DIM	0.97	0.91	0.85
		MPG	
305 DIM	0.98	0.94	0.90

The Wood's model was fitted to each single lactation and the regression parameters, including their combinations, were submitted to an ANOVA model to test possible differences between different groups. Table 3 shows the effect of parity on regression parameters and on their combinations. Except for *b*, all parameters were significantly affected by parity (p <0.0001).

Table 4 shows the effect of production classes on regression parameters and their combinations. All regression coefficients were significantly different among the three classes both within FPG and MPG (p <0.0001). All items reported in Table 3 increased with parity, except for *b* and *c* that were lower. In particular, (Table 3) PY is greater in the MPG than in FPG (41.29 kg vs 32.78 kg). On average, FPG reached the PY about 1 month after MPG (T_{PY} =92 and 58 days). In FPG, lactations of MC reached the PY about 17 days after LC (T_{PY} =93 and 76 days, respectively), whereas HC (T_{PY} =121 days) reached the PY about 27 days after MC. In the MPG, lactations of MC (T_{PY} =59 days) reached the PY about 6 days after LC (T_{PY}

=53 days) and lactations of HC (T_{PY} =77 days) reached the PY about18 days after MC (Table 4). Persistence and milk production after 305 DIM were greater in FPG than in MPG (7.49 v.s. 6.91, for PERS, and 24.89 kg v.s. 22.66 kg for milk production at 305, respectively). Total milk yield until 305 DIM (Y_m) was greater in MPG than in FPG (9,840.92 kg v.s. 8,728.58 kg) and increased with class of production in both groups.

Table 3. Differences of Wood's parameters and their combinations between lactations belonging to primiparous (FPG) and multiparous (MPG) cows

Item	FPG	MPG	p-value
a	16.20	22.05	<.0001
b	0.236	0.239	0.7269
c	-0.0026	-0.0040	<.0001
[1]PY(Kg/d)	32.78	41.29	<.0001
[2]T_{PY}(d)	92.34	58.46	<.0001
[3]PERS	7.49	6.91	<.0001
[4]Prod305	24.89	22.66	<.0001
[5]Y_m	8,728.38	9,840.92	<.0001

[1] PY =Wood's measure of milk yield at peak
[2] T_{PY} =Wood's measure of time at peak
[3] PERS =Wood's measure of persistency
[4] Prod305 =Wood's measure of predicted milk yield after 305 DIM
[5] Y_m =Total milk predicted with Wood's parameters

Table 4. Differences of Wood's parameters and their combinations between lactations belonging to low (LC), medium (MC) and high (HC) production classes for primiparous (FPG) and multiparous (MPG) cows

	FPG			
Item	LC	MC	HC	p-value
a	12.20	16.53	21.28	<.0001
b	0.31	0.22	0.19	<.0001
c	-0.0040	-0.0024	-0.0016	<.0001
PY(kg/d)	28.68	32.81	40.26	<.0001
T_{PY}(d)	75.95	92.64	120.96	<.0001
PERS	7.32	7.49	7.78	<.0001
Prod305	17.49	25.40	35.03	<.0001
Y_m	7,110.19	8,804.29	11,204	<.0001
	MPG			
	LC	MC	HC	
a	17.86	23.55	30.84	<.0001
b	0.30	0.21	0.15	<.0001
c	-0.005	-0.003	-0.002	<.0001
[1]PY	39.24	41.67	48.29	<.0001
[2]T_{PY}(d)	52.99	59.48	77.01	<.0001
[3]PERS	6.81	6.92	7.28	<.0001
[4]Prod305	16.49	24.70	36.79	<.0001
[5]Y_m	8,619.18	10,211	12,907	<.0001

[1] PY =Wood's measure of milk yield at peak

[2] T_{PY} =Wood's measure of time at peak

[3] PERS =Wood's measure of persistency

[4] Prod305 =Wood's measure of predicted milk yield after 305 DIM

[5] Y_m =Total milk predicted with Wood's parameters

Table 5 and Table 6 showed the average values of adjusted R^2 (AdjRSQ) obtained by fitting the four different models to individual lactations, for the FPG and the MPG, respectively. The four lactation models were fitted both for the first 90, 120 and 150 DIM and for the entire standard lactation, 305 DIM.

Table 5. Average AdjRSQ for each model among classes at 150, 120 and 90 DIM for FPG

	AdjRSQ.						
FPG	**Production Class**						
	LC		MC		HC		
90 DIM	Mean	S.D	Mean	S.D.	Mean	S.D	Mean
Wood	0.43	0.24	0.45	0.25	0.47	0.23	0.45
A&S	0.46	0.24	0.47	0.24	0.51	0.21	0.48
Legendre	0.49	0.25	0.49	0.23	0.53	0.21	0.50
4thPolynomials	0.48	0.24	0.48	0.23	0.50	0.22	0.49
120 DIM							
Wood	0.41	0.23	0.41	0.24	0.40	0.22	0.41
A&S	0.44	0.22	0.44	0.23	0.47	0.21	0.45
Legendre	0.46	0.22	0.46	0.22	0.49	0.20	0.47
4thPolynomials	0.46	0.22	0.44	0.22	0.45	0.22	0.45
150 DIM							
Wood	0.38	0.22	0.38	0.23	0.38	0.21	0.38
A&S	0.42	0.22	0.43	0.22	0.46	0.21	0.41
Legendre	0.45	0.22	0.44	0.21	0.46	0.21	0.45
4thPolynomials	0.45	0.22	0.42	0.21	0.44	0.23	0.41
305 DIM							
Wood	0.55	0.17	0.39	0.20	0.30	0.18	0.41
A&S	0.64	0.16	0.46	0.20	0.38	0.19	0.49
Legendre	0.63	0.16	0.49	0.20	0.42	0.20	0.51
4thPolynomials	0.62	0.16	0.45	0.20	0.39	0.21	0.49

Table 6. Average AdjRSQ for each model among classes at 150, 120 and 90 DIM for MPG

	AdjRSQ						
MPG	**Production Class**						
	LC		MC		HC		Total
90 DIM	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean
Wood	0.35	0.25	0.35	0.25	0.34	0.23	0.35
A&S	0.39	0.23	0.41	0.23	0.42	0.22	0.41
Legendre	0.40	0.23	0.45	0.22	0.45	0.19	0.43
4thPolynomials	0.40	0.23	0.43	0.22	0.41	0.22	0.41
120 DIM							
Wood	0.40	0.22	0.41	0.24	0.40	0.22	0.40
A&S	0.41	0.22	0.41	0.22	0.40	0.20	0.41
Legendre	0.42	0.22	0.44	0.21	0.43	0.19	0.43
4thPolynomials	0.42	0.22	0.41	0.21	0.38	0.21	0.40
150 DIM							
Wood	0.38	0.23	0.38	0.23	0.38	0.21	0.38
A&S	0.45	0.21	0.44	0.21	0.42	0.20	0.44
Legendre	0.47	0.21	0.46	0.20	0.43	0.19	0.45
4thPolynomials	0.48	0.21	0.42	0.21	0.38	0.21	0.43
305 DIM							
Wood	0.76	0.14	0.62	0.18	0.42	0.24	0.60
A&S	0.80	0.12	0.68	0.16	0.52	0.23	0.66
Legendre	0.80	0.12	0.70	0.16	0.53	0.23	0.68
4thPolynomials	0.79	0.13	0.66	0.18	0.48	0.24	0.64

In general, AdjRSQ values (Tables 5 and 6) evaluated on the entire lactation (305 DIM) were higher in MPG than in FPG. They ranged from 0.30 to 0.80, depending on the curve model and the production class. The goodness of fit was fairly lower when lactation curves were applied to the first 90, 120 and 150 DIM. The mean values of AdjRSQ for each model ranged from 0.35 (for Wood in MPG, at 90 DIM) to 0.50 (for Legendre in the FPG, at 90 DIM). In particular, in the FPG among production classes, the greater value was obtained with Legendre at 90 DIM (AdjRSQ =0.53) in HC, followed by the A&S model at 90 DIM in the same class (AdjRSQ =0.51); the lowest values were obtained by the Wood model at 150 DIM, in all classes (AdjRSQ =0.38). In the MPG the greatest value was obtained at 150 DIM with 4thPolynomials (AdjRSQ =0.48) in the LC, whereas the lowest values were obtained with the Wood's model in all classes at 90 DIM (AdjRSQ =0.35).

Discriminant procedures

The CDA was applied to the three sub-datasets obtained by fitting the four lactation models to data until 90, 120 and 150 DIM. Continuous variables were the estimated regression parameters of each lactation model, whereas the two classes of production, LC and HC, were considered as class variables. Tables 7 and 8 showed the Mahalanobis distances among group centroids for FPG and MPG, respectively. Except for Wood at 90 DIM in FPG (Table 7), the CDA significantly separated the two production classes for both FPG and MPG. In FPG, the greatest distance was obtained with the 4th Polynomials at 150 DIM between LC and HC, followed by the Legendre Polynomials in the same situation. The lowest Mahalanobis distances were instead obtained with Wood. In MPG the highest distance value was obtained by Wood at 150 DIM between HC and LC, followed by the 4th Polynomials at 150 DIM between the same classes of production (HC and LC). The lowest distance was obtained with the Legendre Polynomials at 90 DIM between LC and HC.

Table 7. Mahalanobis distances between lactations belonging to the low (LC) and the high (HG) classes of production evaluated at 90, at 120 and 150 DIM for FPG

	Wood	Legendre	A&S	4th Polynomials
90 DIM	HC	HC	HC	HC
LC	1 *0.4593*	55 *<.0001*	64 *<.0001*	54 *<.0001*
120 DIM				
LC	4 *0.0076*	64 *<.0001*	64 *<.0001*	78 *<.0001*
150 DIM				
LC	12 *<.0001*	76 *<.0001*	65 *<.0001*	89 *<.0001*

Table 8. Mahalanobis distances between lactations belonging to the low (LC) and the high (HG) classes of production evaluated at 90, at 120 and 150 DIM for MPG

	Wood	Legendre	A&S	4th Polynomials
90 DIM	HC	HC	HC	HC
LC	41 *<.0001*	27 *<.0001*	41 *<.0001*	29 *<.0001*
120 DIM				
LC	54 *<.0001*	33 *<.0001*	47 *<.0001*	46 *<.0001*
150 DIM				
LC	82 *<.0001*	53 *<.0001*	53 *<.0001*	64 *<.0001*

In FPG, the best separation was obtained with the 4th Polynomials at 150 DIM (89) between LC and HC followed by the Legendre Polynomials in the same situation (80). The lowest Mahalanobis distances were obtained with Wood whereas in MPG the best value was obtained by Wood at 150 DIM between HC and LC (82.436), followed by the 4th Polynomials at 150

DIM among the same classes of production (HC and LC). The lowest value was obtained with the Legendre Polynomials at 90 DIM between LC and HC.

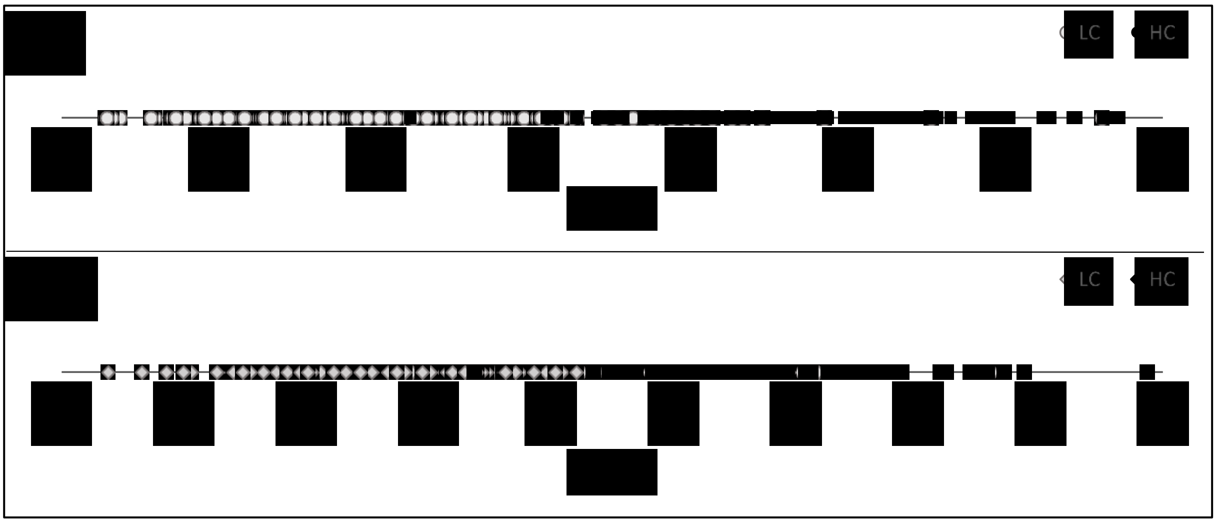

Figure 5. Plot of the CAN for the two parities (FPG and MPG) at 150 DIM between LC and HC

Figure 5 displays the plot of the CAN for FPG and MPG at 150 DIM. The ability of CAN in distinguishing LC from HC was greater in FPG than in MPG. This is also confirmed by the greatest distance between the two classes of lactation for FPG than MPG (Tables 7 and 8).

Due to the low performances of CDA in assigning lactations when the Wood's parameters were used, only the three polynomial models were exploited in the subsequent DA. Table 9 displays errors in assigning lactations to LC and HG after the bootstrap resampling procedure.

Table 9. Percentage of incorrect assignment of lactations belonging to LC and HC at 150,120 and 90 DIM for FPG and MPG

FPG	Legendre	A&S	4th Polynomials
150 DIM	12	16	10
120 DIM	14	16	12
90 DIM	17	16	14
MPG			
150 DIM	18	20	14
120 DIM	22	22	17
90 DIM	24	23	24

The DA with 4th Polynomials, correctly assigned 90% of lactations, by using the first 150 DIM. With only 120 and 90 DIM, the corrected assignations were of 88% and 86%, respectively. In the MPG, the DA, with the 4th Polynomials, correctly assigned 86% of lactations at 150 DIM, 83% at 120 DIM and 76% at 90 DIM. The effective error in assigning lactations, however, should be halved, because we consider true error only the incorrect assignment of lactations belonging to LC.

4.5 Discussion

In the present research, an algorithm to early detect if a lactation will have a persistency suitable to become a long lactation and, therefore, continue the production over the standard 305 DIM was developed. Milk production at 305 DIM was estimated and three different classes of production (LC, MC and HC) were considered. These three classes were good indicator of the future milk production as confirmed by the correlations on average higher than 90% between the milk yielded at 305 DIM and that produced at 400, 500 and 600 DIM (Table 2). A lactation was considered a candidate for a long lactation if it belonged to MC or to HC. The milk yielded at 305 DIM was, in these two classes, greater than 20 kg.

Data were explored by modeling lactations by using Wood. The results, substantially, confirmed what is reported in literature about PY, TPY, PERS and total milk yield at 305 DIM (Y_m) for Holstein cows. Regarding the regression coefficients obtained by fitting Wood to individual lactations, the increase of the parameter *a* by parity (Table 3) was in agreement with previous reports existent in the literature (Macciotta et al., 2005a). It influences the height of the curve that is lower in primiparous than multiparous. The coefficient *b* indicates the growing rate and the magnitude of the curvature of the lactations (Macciotta et al., 2005b). The lack of differences observed between parities indicated that both the growing rate of milk production and the magnitude of curvature do not depend on parities. The absolute value of the parameter *c*, which controls the declining rate of the curve (Macciotta et al., 2005b), decreases from the first to the multiparous parity, suggesting that lactations belonging to FPG have a higher persistence than MPG. The shape and duration of curve quite differ in the two parities (Hansen et al., 2006). The increase of PY with parity observed in this research is in accordance with previous studies on cows (Hansen et al., 2006; Van Raden et al., 2006; Dematawewa et al., 2007). Moreover several authors (Dematawewa et al., 2007; Steri et al., 2012) showed that first-parity cows reach the peak of lactation later than the multiparous. The results of the present study are in agreement with these researches. The lower PY and the higher lactation persistence observed for FPG compared to MPG can be ascribed to the ‘elastic’ properties of the lactation curve: if production at peak decreases, then the production in the last part of curve increases, and conversely. These findings can be biologically explained taking into account that animals at first parity are still subjected to growth processes, which determine a lower milk yield in the early lactation compared to older animals (Stanton et al., 1992; Pulina et al., 2005). According to several previous studies (Lee et al., 2006;Van Raden et al., 2006; Mellado et al., 2016), the total milk production after 305 DIM was greater in MPG (984,092 kg) than in FPG (872,858 kg). Lactations of HC tend to have higher PY that is reached later than lactations of LC, in agreement with literature (Dematatewa et al., 2007; Steri et al., 2012).

Considering a standard lactation of 305 DIM, the values of AdjRSQ observed in the present study (Tables 5 and 6) were not as high as values reported in literature due to the large variability of individual patterns (Steri et al., 2012b). In general, as expected, the goodness of fit decreased from 305 DIM to 150 DIM until 120 and 90, which showed the lowest values, except for Wood in FPG (where average values of AdjRSQ at 90 DIM were higher than others considered DIM). Legendre was the model that best fitted curves in the first part of lactation, both by parities and by classes of production. Stating that the models with a greater number of parameters have better fitting performances, these findings were expected, compared to the models with few parameters (Bouallegue et al., 2015; Steri et al., 2009; Macciotta et al., 2005b). However, apart from Wood, both A&S and 4th Polynomials little differentiate in AdjRSQ values with Legendre. For this reason, with the exclusion of Wood, all models were used in the subsequent analyses.

Discriminant procedures

It is very important to avoid incorrect assignation of lactations belonging to LC to the HC group. Actually, if a LC lactation is destined to be a long lactation, the animal would not be involved in the ordinary cycle of insemination. However, after 305 DIM, the cow will not continue the lactation having a too low milk production. This effect could compromise the good management of the farm. For this reason, in DA, only the incorrect assignations of LC lactations to HC were considered true errors.

The high statistical significance of Mahalanobis distances (Tables 7 and 8) highlighted a clear separation between the two classes of production. The lowest discriminant error (10%) was obtained in FPG (Table 9) by using the 4th Polynomials at 150 DIM. The inspection of lactation incorrectly assigned reveals that only half of them belonged to LC and were incorrectly assigned to HC. In consequence, the true error was around 5%. Errors slightly increased in scenarios involving 120 and 90 DIM, with a total error of 12% and 14%, respectively.

In MPG, errors in assignment increased to respect FPG. However, also in this case, the 4th Polynomials was the model that better contributed to obtain a good classification. The error in assignment could be acceptable in the 150 DIM and 120 DIM (14% and 17%, respectively). A total error of 24% was observed in the 90 DIM (Table 9) scenario.

4.6 Conclusions

The algorithm developed in the present study could help farmers to early select a quota of their herd to be destined to a long lactation. In practice, a database with former complete lactations should be firstly created. It represents the basic dataset where the CAN is obtained. Then, as a new lactation proceeds, the recorded milk production data can be fitted by using the 4th Polynomials model and the estimated parameters submitted to the DA. The entire procedure could be automated by implementing, for example, the Afifarm's report with a statistical computer software. The lactation is assigned to one of the two production classes with an error in assignment as reported in Table 7. Actually, those errors would be halved because only the incorrect assignment of a LC lactation to the HC class should be avoided.

4.7 References

Ali, T. E. and, L. R. Schaeffer. 1987. Accounting for covariances among test day milk yields in dairy cows. Canadian Journal of Animal Science. 67(3):637-644.

Allore, H. G. and, H. N. Erb. 2000. Simulated effects on dairy cattle health of extending the voluntary waiting period with recombinant bovine somatotropin.Preventive veterinary medicine. 46(1):29-50.

Arbel, R., Y. Bigun, E. Ezra, H. Sturman and, D. Hojman. 2001. The effect of extended calving intervals in high lactating cows on milk production and profitability. Journal of Dairy Science. 84(3):600-608.

Bar-Anan and, R. M. Soller.1979. The effects of days-open on milk yield and on breeding policy post partum. Animal Science. 29(1):109-119.

Boullalleguer, M., R. Steri and, M. B. Hamouda. 2015. Modelling of individual lactation curves of Tunisian Holstein-Friesian cows for milk yield, fat, and protein contents using parametric, orthogonal and spline models. Journal of Animal and Feed Sciences. 24:11−18.

Butler, T., L. Shalloo and, J. J. Murphy. 2010. Extended lactations in a seasonal-calving pastoral system of production to modulate the effects of reproductive failure. Journal of Dairy Science. 93:1283−1295.

De Maesschalck, R., D. Jouan-Rimbaud and, D.L. Massart. 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems. 50:1−18.

Demetawewa, C. M. B., R.E. Pearson and, P. M. VanRaden. 2007. Modelling Extended Lactations of Holsteins. Journal of Dairy Science. 90:3924−3936.

De Vries, A. 2006. Economic value of pregnancy in dairy cattle. Journal of Dairy Science. 89(10):3876-3885.

Dijkhuizen, A. A., J. Stelwagen and, J.A. Renkema. 1985. Economic aspects of reproductive failure in dairy cattle. I. Financial loss at farm level. Preventive Veterinary Medicine. 3(3):251-263.

Efron, B. 1979. Bootstrap Methods: Another look at the jackknife. 1979. The Annals of Statistics. 7(1):1−26.

Hansen, J. V., N. C. Friggens and, S. Højsgaard. 2006. The influence of breed and parity on milk yield, and milk yield acceleration curves. Livestock Science. 104:53−62.

Heiman, M. M. 1984. Results of an integrative computer program of fertility and production data from an AI cattle population. In 10° international congress on animal reproduction and artificial insemination, University of Illinois at Urbana-Champaign, Illinois (USA), 10-14 Jun 1984. University of Illinois at Urbana-Champaign.

Holmann, F. J., C. R. Shumway, R. W. Blake, R. B. Schwart and, E. M. Sudweeks. 1984. Economic Value of Days Open for Holstein Cows of Alternative Milk Yields with Varying Calving Intervals1. Journal of Dairy Science. 67(3):636-643.

Inchaisri, C., R. Jorritsma, P. L. A. M. Vos, G. C. Van der Weijden and, H. Hogeveen. 2010. Economic consequences of reproductive performance in dairy cattle. Theriogenology. 74:835−846.

Inchaisri, C., R. Jorritsma, P. L. A. M. Vos, G. C. Van Der Weijden and, H. Hogeveen. 2011. Analysis of the economically optimal voluntary waiting period for first insemination. Journal of Dairy Science. 94(8):3811-3823.

Lee, J. Y. and, I.H. Kim. 2006. Advancing parity is associated with high milk production at the cost of body condition and increased periparturient disorders in dairy herds. Journal of Veterinary Science. 7(2):161-166.

Macciotta, N. P. P., C. Dimauro, R. Steri and, A. Cappio-Borlino. 2008a. Mathematical modelling of goat lactation curves. Dairy goats feeding and nutrition. (pp.) 31-46.

Macciotta, N. P. P., D. Vicario and, A. Cappio Borlino. 2005b. Detection of Different Shapes of Lactation Curve for Milk Yield in Dairy Cattle by Empirical Mathematical Models. Journal of Dairy Science. 88:1178−1191.

Mardia, K.V., J.T. Kent and, J. M. Bibby. 2000. Multivariate Analysis. Academic Press, London Morrison, F. 1976. Multivariate statistical methods. McGraw-Hill, New York, NY.

Mellado, T. L., J. M. Flores, A. de Santiago, F. G. Veliz , U. Macías-Cruz, L. Avendaño-Reyes and, J. E.García. 2016. Extended lactationinhigh-yielding Holstein cows: Characterization of milk yield and risk factors for lactations >450 days. Livestock Science. 159:50−55.

Pulina, G., A. Nudda, N.P.P. Macciotta, G. Battacone, S. Fancellu and, C. Patta. 2005. Non-nutritional strategies to improve lactation persistency in dairy ewes. 11th Annual Great Lakes dairy sheep symposium: proceedings, November 3 – 5, 2005, Burlington (VT), USA. Madison, University of Wisconsin. (pp.) 38-68.

Schaeffer, L. R. 2004. Application of random regression models in animal breeding. Livestock Production Science. 86(1):35-45.

Stanton, T. L., L. R. Jones, R. W. Everett and, S. D. Kachman. 1992. Estimating milk, fat, and protein lactation curves with a test day model. Journal of Dairy Science. 75(6):1691-1700.

Strandberg, E. and, P. A. Oltenacu, 1989. Economic consequences of different calving intervals. Acta Agriculturae Scandinavica. 39(4):407-420.

Silvestre, A. M., F. Petim-Batista and, J. Colaço. 2006. The Accuracy of Seven Mathematical Functions in Modeling Dairy Cattle Lactation Curves Based on Test-Day Records From Varying Sample Schemes. Journal of Dairy Science. 89:1813−1821.

Steri, R., A. Cappio-Borlino and, N.P.P. Macciotta. 2009. Modelling extended lactation curves for milk production traits in Italian Holsteins. Italian Journal of Animal Science. 8(2):165-167.

Steri, R., C. Dimauro, E. Nicolazzi and, N. P. P. Macciotta. 2012a. Analysis of lactation shapes in extended lactations. Animal. 6(10):1572–1582.

Steri, R. 2012b. The mathematical description of the lactation curve of Ruminants: issues and perspectives (Doctoral dissertation, PhD Thesis. Università Degli Studi di Sassari, Sassari, Italy).

Stevenson, J. S., M. A. Portaluppi, D. E. Tenhouse, A. Lloyd, D. R. Eborn, S. Kacuba and, J. M. DeJarnette. 2007. Interventions after artificial insemination: conception rates, pregnancy survival, and ovarian responses to gonadotropin-releasing hormone, human chorionic gonadotropin, and progesterone. Journal of Dairy Science. 90(1):331-340.

Van Amburgh, M. E., D. M. Galton, D. E. Bauman and, R.W. Everett. 1997. Management and economics of extended calving intervals with use of bovine somatotropin. Livestock Production Science. 50 (1):15-28.

Van Raden, P.M., C. M. B. Dematawewa, R. E. Pearson and, M.E. Tooker. 2006. Productive Life Including All Lactations and Longer Lactations with Diminishing Credits. Journal of Dairy Science. 89:3213−3220.

Weller, J. I., R. Bar-Anan and, K. Osterkorn. 1985. Effects of days open on annualized milk yields in current and following lactations. Journal of Dairy Science. 68(5):1241-1249.

Wood, P. D. P. 1967. Algebraic model of the lactation curve in cattle. Nature. 216(5111):164-165.