# The use of Principal Component Analysis for Predicting Genomic Breeding Values

Christos Dadousis

Norwegian University of Life Sciences
Department of animal and aquaculture sciences
Master Thesis 30 credits 2012

# The use of Principal Component Analysis for Predicting Genomic Breeding Values

CHRISTOS DADOUSIS
Registration number: 830524-167-070 (WUR)/ 976553 (UMB)

THESIS ANIMAL BREEDING AND GENETICS
COURSE CODE: ABG-80430 (WUR)/M30-IHA (UMB)
June 2012


WAGENINGEN UNIVERSITY
WAGENINGEN UR

Department of Animal Breeding and Genetics

SUPERVISORS
Dr. ir. Mario Calus (WUR)
Dr. Bjorg Heringstad (UMB)

Education and Culture
Erasmus Mundus

# Contents

List of Tables ..... 5
List of Figures ..... 7
Preface ..... 8
Acknowledgments ..... 9
English summary ..... 10
Hellenic summary (Ελληνική Περίληψη) ..... 12
Abbreviations ..... 14
1. Introduction ..... 16
1.1 Principal Component Analysis ..... 18
2. Objective ..... 20
3. Material and methods ..... 21
3.1 Data ..... 21
3.2 Principal Component Regression model ..... 22
3.3 GBLUP model ..... 23
4. Results ..... 24
4.1 PCR model ..... 24
4.2 GBLUP model ..... 31
5. Discussion ..... 38
6. Conclusions and implications ..... 42
References ..... 44

# List of Tables

Table 1 Descriptive statistics of pre-adjusted testday data of the traits ..... 21

Table 2 Number of cows with phenotype and genotype from the four countries ..... 21

Table 3 Highest accuracies obtained for the three models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. ..... 25

Table 4 Cumulative proportion of the original variability captured by principal components 28

Table 5 Highest accuracies obtained for GBR total and the two GBR lines separately. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. ..... 28

Table 6 Highest accuracies obtained for principal component regression models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. Difference denotes the change on the accuracies between the two different PCA methods. ..... 30

Table 7 Number of principal components needed to achieve highest accuracies per trait and per population when principal component analysis applied on the reference part ..... 31

Table 8 Number of PCs needed to achieve highest accuracies per trait and per population when the entire dataset was used to perform principal component analysis ..... 31

Table 9 Highest accuracies obtained from GBLUP models with (GBLUP_PCs) and without principal components (GBLUP_basic) included. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. Difference indicates the change on the accuracies between the two GBLUP models with or without PCs. ..... 34

Table 10 Summarized table of highest accuracies for GBLUP with (GBLUP_PCs) and without principal components (GBLUP_basic) included and principal component regression (PCR (EIGEN)) models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. ..... 35

Table 11 Highest accuracies for GBLUP with (GBLUP_PCs) and without principal components (GBLUP_basic) included and principal component regression models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. ..... 36

## List of Figures

Figure 1 Graphical representation of principal components of a dataset, shown as arrows in the diagram ..... 19

Figure 2 Pie chart of individuals included in the dataset per country of origin ..... 22

Figure 3 Scoreplot of the first two principal components (PC1 vs. PC2). Principal component analysis was performed on the whole dataset ..... 24

Figure 4 Pattern of the accuracies for principal component regression models where the selection of PCs was based either on eigenvalues (left panel) or on sum of square contribution (right panel). An increasing number of PCs, one by one, up to 1,000 was fitted. ..... 26

Figure 5 Scoreplot of the first two principal components (PC1 vs. PC2). Principal component analysis was performed on the whole dataset, whereas GBR population was split into the two different genetic groups. ..... 27

Figure 6 Scoreplot of the first two principal components (PC1 vs. PC2). Principal component analysis was performed on the G matrix ..... 33

Figure 7 Pattern of the accuracies for GBLUP models for tesday milk, fat and protein yield for four countries. An increasing number of PCs (extracted from G matrix) was fitted, one by one, up to 1,000. ..... 35

Figure 8 Pattern of the accuracies for PCR models where the selection of PCs was based either on eigenvalues (left panel) or on sum of square contribution (right panel). An increasing number of PCs, one by one, up to 1,000 was fitted. PCA applied on the G-matrix. ..... 37

## Preface

This thesis marks the end of a two year MSc studies followed in the European Master in Animal Breeding and Genomics. From the beginning of this MSc the idea was to perform a research on the revolutionary topic of genomic predictions (genomic selection).

Some decades ago the idea of predicting the yearly milk amount that a cow would produce just by using DNA information would appear as a science fiction story. However, this idea became feasible since 2001 when firstly proposed by Meuwissen et al and is well known as "Genomic Selection, GS". GS has already been established by several breeding companies worldwide in different breeding schemes and species.

Nevertheless, there are still important questions to be addressed. The last few years numerous statistical models have been proposed, most of them "lost" in complexity in a way to achieve high accuracies (comparable to progeny test) as well as to "capture" genetic architecture of quantitative traits.

The idea of this research was to check the predictive ability of an easy to handle multiple linear model, principal component regression (PCR), where strong assumptions for the data are not required. In PCR the original regressors (SNPs in genomic data) have been transformed into a small number of orthogonal axes which can capture the original variability of the SNP data while at the same time are uncorrelated to each other and each one includes all the original variables. These axes are the so called principal components (PCs) obtained by principal component analysis (PCA) of the original data.

PCR was used to predict genomic values using real data provided by RobustMilk project. Predictive ability of PCR was compared to an ordinary GBLUP model.

## Acknowledgments

Firstly, I would like to thank the European Master in Animal Breeding and Genetics (EMABG) and especially Professor Johan van Arendonk as well as all the EMABG group members. EMABG program gave me the opportunity and funding (scholarship) to study animal breeding in two of the most well-known universities in the section. High quality of education offered in collaboration to the opportunity of exploring not only two different countries but also cultures, broading in this way my life experiences, resulted in two of the most beautiful years of my life.

My supervisor, Mario Calus from Animal Breeding and Genomic Center (Wageningen University), not only for his advanced scientific guidance but for his kind character as well. Mario allowed me to 'express' myself on my thesis while at the same time pointing out clear solutions and alternative views when difficulties appeared. Moreover, his fast replies resulted in a faster progress of the work as well as in a more in depth view. My co-supervisor Bjørg Heringstad from Norwegian University of Life Sciences (UMB) who through her experience pointed out basic questions and comments on the research. I could briefly describe Mario and Bjørg as following: "just two **brilliant persons**!"

The RobustMilk project for providing the data.

I am thankful to all the staff of the Animal Breeding and Genomics Center at WUR and especially to Roel Veerkamp, Ina Hulsegge and Yvonne Wientjes. Great thanks to Marcin Pszczola for his help with R software.

I could not forget my friends Agnese, Magnai and Merina for spending days and nights studying together during our first year of the MSc at UMB University, in Ås.

Finally, I would like express my acknowledgments to Zafeiris Abas (School of Agricultural Development, Democritus University of Thrace) and Peristera Paschou (Department of Molecular Biology, Democritus University of Thrace). Without their help I would have not been here.

***Thank you all***

Δαδούσης Χρήστος (Dadousis Christos)

## English summary

During the last few years the idea of predicting quantitative traits and diseases based on genotypic information has raised a major interest in animal and plant breeding as well as in human genetics. However, there are still important questions and problems that need to be addressed. Some of these problems are statistical. Statistical problems mainly concern multicollinearity basic derived from the huge amount of available data. In addition, the number of variables that needs to be estimated *(p)* is much larger than the number of observations *(n)* disabling least squares methodology. Principal component analysis (PCA) is a multivariate statistical method often used to deal with these problems.

The objective of this study was to investigate the use of PCA for predicting genomic breeding values. Data of 1,609 first lactation Holstein heifers were analysed including test-day milk, fat and protein yields. Animals originated from 4 countries, Ireland, United Kingdom, the Netherlands and Sweden and were genotyped within the RobustMilk project with the Illumina BovineSNP50 Beadchip. After editing, 37,069 SNPs remained.

Two different models were compared for genomic predictions i) Principal component regression (PCR) was used to directly estimate genomic breeding values. Selection of principal components (PCs) was based either on their eigenvalues or the regression sum of square (SS) contribution, ii) a best linear unbiased prediction model with genomic relationship matrix (GBLUP) was developed to compare accuracies to those obtained by PCR models. In a third case, PCs extracted from the G-matrix were added in the GBLUP model as fixed effects to investigate the impact of population structure when predicting genomic breeding values. The dataset was split in four training (reference populations) and testing parts for validation. Each testing subset included all animals from only one country. Predictive ability was calculated as Pearson correlation between the predicted genomic values and the phenotypes.

PCR where PCs selection was based on their eigenvalues resulted in considerably high accuracies and outperformed both PCR (SS) and GBLUP models. Accuracies varied between populations and traits. Interestingly, highest accuracies were obtained for the only genetically distinguished population (GBR), according to PCA, in the dataset with only the first or the first two PCs for protein and milk yield, respectively. In GBLUP models an increase of the accuracies (~40% on average) was observed in all cases when PCs were added in the model.

Simplicity of PCR method, fast computation, reduction of data dimension (>96%) as well as the ability of both predicting breeding values and identifying groups in the data are the main benefits of PCR. The above elements together with at least as accurate predictions as GBLUP, obtained with real data, marks PCR as an attractive tool for animal breeding. However, the variation on the number of PCs needed to achieve highest accuracies could be a drawback of the method. According to our results, where the highest accuracies obtained for the only group of animals genetically separated from the rest, we hypothesize that PCR could be tested for across breed genomic predictions.

## Hellenic summary (Ελληνική Περίληψη)

Κατά τη διάρκεια των τελευταίων χρόνων η ιδέα της πρόβλεψης ποσοτικών ιδιοτήτων και ασθενειών με βάση γενετική πληροφορία και μόνο έχει εξάρει το ενδιαφέρον στους τομείς της γενετικής βελτίωσης φυτών και ζώων, όπως επίσης και στον άνθρωπο. Παρ' όλα αυτά, υπάρχουν ακόμα σημαντικά ερωτήματα και προβλήματα προς απάντηση. Μέρος αυτών των προβλημάτων αφορούν τη στατιστική ανάλυση. Τα στατιστικά προβλήματα προέρχονται κυρίως από το μεγάλο όγκο των δεδομένων και αφορούν, κυρίως, πολυσυγγραμμικότητα (multicollinearity). Επιπλέον, αριθμός των μεταβλητών *(p)* μεγαλύτερος των παρατηρήσεων *(n)* καταργεί τη χρήση της μεθόδου των ελαχίστων τετραγώνων. Η principal component analysis (PCA) αποτελεί κλάδο της πολυπαραγοντικής ανάλυσης (multivariate analysis) και συχνά χρησιμοποιείτε για αντιμετώπιση τέτοιων προβλημάτων στη στατιστική ανάλυση.

Σκοπός της εργασίας ήταν η διερεύνηση της χρησιμότητας της PCA για πρόβλεψη κληροδοτκών αξιών των ζώων βάση μοριακών δεικτών. Γαλακτοπαραγωγή, λιποπαραγωγή και πρωτεΐνοπαραγωγή από αγελάδες 1,609 Holstein πρώτης γαλακτοπαραγωγικής περιόδου χρησιμοποιήθηκαν. Τα δεδομένα προήλθαν από το RobustMilk project στο οποίο συμμετέχουν πειραματικοί σταθμοί από Μεγάλη Βρετανία, Ολλανδία, Ιρλανδία και Σουηδία. Οι γονοτυπήσεις πραγματοποιήθηκαν με το Illumina BovineSNP50 Beadchip και 37,069 SNPs χρησιμοποιήθηκαν στην τελική ανάλυση.

Δύο διαφορετικά μοντέλα δημιουργήθηκαν i) Principal component regression (PCR). Η επιλογή των PCs βασίστηκε είτε στις eigenvalues είτε στη συνεισφορά του αθροίσματος των τετραγώνων (regression sum of squares contribution), ii) άριστη γραμμική αμερόληπτη πρόβλεψη (GBLUP) για σύγκριση των αποτελεσμάτων με το PCR μοντέλο. Επιπλέον, αναπτύχθηκε ένα μοντέλο GBLUP όπου PCs εξαγόμενα από τον G-matrix προστέθηκαν ως σταθερές μεταβλητές με σκοπό τη διερεύνηση της σηματικότητας της δομής του πληθυσμού σε μοντέλα πρόβλεψης με γενετικούς δείκτες. Τα δεδομένα χωρίστηκαν σε τέσσερα διαφορετικά "εκπαίδευση-αξιολόγηση" μέρη. Το κάθε τμήμα αξιολόγησης περιείχε ζώα από μία χώρα. Οι ακρίβειες εκτίμησης των κληροδοτικών αξιών υπολογίστηκαν ως Pearson συσχετίσεις μεταξύ των εκτιμώμενων τιμών και των φαινοτύπων.

Η PCR όπου τα PCs επιλέχθηκαν βάση των eigenvalues έδωσε τα καλύτερα αποτελέσματα και υπερείχε των υπολοίπων μοντέλων. Οι ακρίβειες εκτιμήσεως κυμαίνονται και εξαρτώνται από τον πληθυσμό και το ποσοτικό γνώρισμα. Εντυπωσιακά, μεγαλύτερες ακρίβειες επιτεύχθηκαν για τον μοναδικό πληθυσμό που εν μέρει ξεχώριζε γενετικά από τους

υπολοίπους (βάση της PCA), μόνο με τη χρήση του πρώτου PC για τη γαλ/γή και πρωτεΐν/γή. Στα μοντέλα GBLUP αύξηση της ακρίβειας εκτιμήσεων των κληροδοτικών τιμών παρατηρήθηκε σε όλες τις περιπτώσεις με την προσθήκη PCs.

Η ευκολία εφαρμογής της PCR, ταχείς υπολογισμοί, ελάττωση του όγκου δεδομένων (>96%) όπως επίσης και οι δυνατότητες της σύγχρονης πρόβλεψης τιμών και εύρεσης πιθανών διαφορετικών ομάδων στα δεδομένα αποτελούν τα βασικά προτερήματα της PCR. Αυτά, σε συνδυασμό με επίτευξη ακριβειών εκτίμησης κληροδοτικών αξιών ζώων, με βάση γενετικούς δείκτες, τουλάχιστον ίσων με την GBLUP σε πραγματικά δεδομένα χαρακτηρίζουν τη μεθοδολογία της PCR ως ένα ελκυστικό εργαλείο στη γενετική βελτίωση των ζώων. Ωστόσο, η ποικιλία του αριθμού των PCs που πρέπει να προστεθούν στο μοντέλο για να επιτευχθούν μέγιστες ακρίβειες αποτελεί μειονέκτημα. Σύμφωνα με τα αποτελέσματά μας, όπου μέγιστες ακρίβειες επετεύχθησαν στο μοναδικό πληθυσμό που γενετικά διαχωριζόταν (εν μέρει) από τους υπολοίπους υποθέτουμε πως η PCR θα μπορούσε να δοκιμαστεί για προβλέψεις μεταξύ φυλών.

## Abbreviations

GBLUP: Best Linear Unbiased Prediction based on marker genomic relationship matrix (G)

G matrix: genomic relationship matrix

GS: Genomic selection

LD: Linkage Disequilibrium

PCA: Principal Component Analysis

PCR: Principal Component Regression

PCs: Principal Components (scores of the Principal Component Analysis)

PCR (EIGEN): Principal Component Regression where the selection of the PCs was based on their eigenvalues

PCR (SS): Principal Component Regression where the selection of the PCs was based on the regression sum of squares contribution

# 1. Introduction

Breeding programs have successfully been established in a variety of animal species and countries worldwide. Selection of the "best" animals (animals used to breed next generation) was based on phenotype and pedigree. The aim of breeding programs is population improvement on the desirable traits included in the breeding goal. One of the most remarkable changes in (dairy cattle) breeding programs was the implementation of progeny test. Progeny test was firstly implemented in Denmark and very soon spread out all over the world (Johanson, 1959). However, the first revolution in breeding programs was the introduction of artificial insemination (AI) which resulted in a fast transfer of the genetic gain to the whole (commercial) population (Vishwanath, 2003). To illustrate the value of animal breeding over the last decades, Amer et al (2011) estimates the aggregate benefits due to genetic improvement in UK dairy industry of around 2.2 -2.4 £ billion since 1980.

Some decades ago the idea of predicting the yearly milk amount that a cow would produce just by using DNA information would appear as a science fiction story. However, in animal breeding the idea of using molecular markers in connection with phenotypes to predict animal genetic merit is quite old (Neimann-Sorenson and Robertson, 1961), but not only until the recent advances in molecular techniques (whole genome scan, DNA microarrays) became this idea feasible. The release of whole genome sequence (e.g. bovine hapmap project), SNP polymorphisms and the technology of microarrays have been successfully collaborated allowing us to genotype individuals across the whole genome for tens or even hundreds of thousands markers in a considerably low cost (~100$).

In 2001, Meuwissen et al showed through simulations that genome-wide dense markers can adequately be used to estimate breeding values (EBVs; an estimate of the additive genetic merit for a particular trait that an individual will pass on to its descendants) for animals with a considerably high accuracy. This idea is what is called Genomic Selection (GS). In GS, DNA information is used to predict the genetic merit of young animals.The key point in GS is that with a genome-wide panel of dense markers all quantitative trait loci (location of a gene on the chromosomes that affects a quantitative trait; QTLs) are in linkage disequilibrium (non-random association of alleles at two or more loci; LD) with at least one marker.

Over the last few years GS has been implemented in dairy cattle breeding programs (Harris et al., 2008; Van der Linde and Wilmink, 2008; Wiggans et al., 2008; Berry et al., 2009; De Roos et al., 2009; Ducrocq, 2009; Schenkel et al., 2009a,b; Van Doormaal et al., 2009) and has fairly been described as **the most promising molecular application in livestock** (Sellner et al., 2007).

In practise, GS involves two steps. First, the effect of each marker (Single Nucleotide Polymorphism; SNP) is estimated in a reference population where animals with known phenotypes and genotypes are included. In the second step, genomic breeding values (GEBVs) of young animals are calculated by using only the information of the markers. The prediction of the breeding values through the markers can be derived from the following model:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij} b_j + e_i$$

Where, $y$ is a vector of phenotypic records, $\mu$ is the overall mean, $x$ is the code of genotype for SNP $j$ and $b$ is the additive effect of SNP $j$.

Despite the fact that several algorithms have been presented to solve the above model, there are still important questions and problems to be addressed. Some of these problems are statistical. Statistical problems mainly concern multicollinearity in the SNP dataset, that in genetic terms is interpreted as LD among markers, which leads to unstable estimates in least-squares regression. Moreover, a major problem in SNP datasets is that the number of variables that needs to be estimated *(p)* is much larger than the number of observations *(n)* disabling least squares methodology. Principal Component Analysis (PCA) is a powerful dimensionality reduction technique and together with its regression (PCR; Principal Component Regression) are methods often used to overcome these problems.

### 1.1 Principal Component Analysis

Principal component analysis belongs to the general framework of multivariate analysis and is one of the most famous and oldest multivariate techniques. It was introduced by Pearson (1901) but also independently developed by Hotelling (1933). However, only in recent years through the advanced computer technology PCA (like many multivariate techniques) became a useful tool and applicable in practise.

As stated above, the problem in SNP matrices mainly arrives from the huge size of the data where relatively few individuals (e.g. some thousands) are genotyped for many markers (e.g. hundreds of thousands). As a result unavoidably some markers will be in LD which in statistics creates multicollinearity. Assume a matrix X of order *(n x p)* where *n* individuals have been genotyped for *p* SNPs. The elements of this matrix may be 0, 1 or 2 representing the genotype of each individual for each SNP (0 and 2 for homozygotes and 1 for heterozygotes). The main idea of PCA is to reduce the number of variables in a dataset as well as solving the multicollinearity problem (high correlation among X-variables); so, find a small set *k (k<p)* of principal components (PCs) explaining as much as possible of the variability in the original X-variables. This is achieved through an orthogonal transformation (axes of variation) of the original dataset while at the same time including as much of the original variability as possible in the first few PCs (Figure 1). So, PCs are linear combinations of a set of random variables $X^t$ =[$X_1$, $X_2$,..., $X_p$]., such as $T = a^t X$ . The first principal component is defined as the variable $T_1 = a_1^t X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$ which has the maximum variance with the constraint that $a^t a = 1$. For all the combinations it stands that: $\text{cov}(PC_i, PC_j) = 0$ for all $i \neq j$ *(i,j= 1,2,…,p).*

The basis of PCA is either the spectral decomposition of the covariance (correlation) matrix or singular value decomposition (SVD) of a data matrix. Concerning the equality of these two techniques there is a disagreement in the literature. Some authors believe that they are identical, some that they differ in normalization strategies, while others state that PCA and SVD are completely different approaches (Skillicorn, 2007). However, it is important to note that SVD is less computationally demanding than PCA, especially with large datasets where *n<<p*. The reason is that SVD works directly on the matrix X *(n x p),* whereas PCA on the covariance (correlation) matrix *(p x p).*

Another interesting point about PCA is the impact of the rank of the variance-covariance matrix and its effect on PC extraction. Assuming a matrix X with dimensions *(n x p)*, where $n<<p$. While we would expect PCA to return as many PCs as the original variables are *(p)*, due to spectral decomposition of X´X, instead PCA will return *n-1* PCs. Thus, the total variability which has originally distributed along *p* axes now is compressed in *n-1* dimensions. This may result in "spurious" results (Bumb, 1982a,b; Bumb 1986) due to singular and positive semi-definite correlation matrix and has been observed in factor analysis when the number of variables exceeds the number of observations. However, in the literature there are other authors stating that still there is no problem on the interpretation of the results (Adelman and Morris, 1982a,b; Murrell, 1986). Dimauro et al (2011) addressed the above problem in the era of genomic predictions.

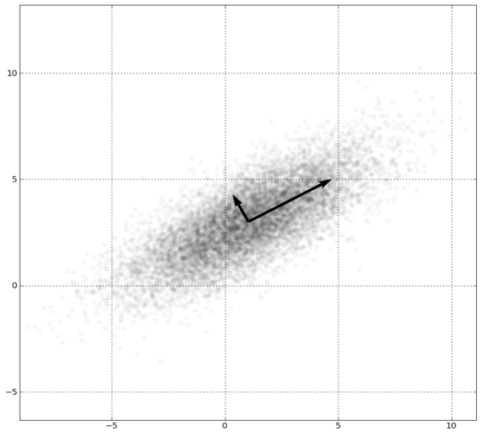



**Figure 1 Graphical representation of principal components of a dataset, shown as arrows in the diagram** (http://en.wikipedia.org/wiki/Principal_component_analysis)

In genetics PCA has mainly been used for population studies and has turned out to be a powerful tool for studying population structures, migration patterns and correcting for stratification in association studies by capturing genetic variation (Price et al, 2006; Patterson et al, 2006; Liu and Zhao, 2006; Paschou et. al., 2007; Novembre and Stephens, 2008; Novembre et al, 2008; Reich et al, 2008; Paschou et al., 2008; McVean, 2009; Drineas et. al., 2010). The first application of PCA in population genetics was in 1978 by Cavalli-Sforza in a human variation research. In this study, where PCA was used to produce maps of human genetic variation across mainland regions.

In animal breeding PCA has recently been used for inferring population clusters from different breeds (Lewis et. al., 2011) as well as to represent genotypes in genomic breeding value estimation (Pinto et al, 2006; Solberg et al, 2009; Macciotta et al, 2010a,b; Long et al, 2010; Dimauro et al, 2011). Daetwyler, et al (2011) used PCA in an attempt to show the impact of population structure on the accuracy of genomic breeding values (GEBVs) in a multi-breed sheep population. In all the above cases, the PCs added to prediction models of genomic breeding values only accounted for the variability captured in the original X-variables (SNPs) and not for the proportion of explained phenotypic variability (response variables – in this case phenotypes of animals in the reference population). However, it has been shown in statistical literature that in PCR the first principal components (accounting for most variation in the X-variables) can totally fail as predictors (accounting for the variation in the response variable) and that even components explaining little variance in the X-variables can be important in the regression (Jeffers, 1967; Jolliffe, 1982; Hawkins, 1973; Boneh and Mendietta, 1994; Hwang and Nettleton, 2002). For instance, Hadi and Ling (1998) have shown using Hald's data(1) that while the first three (out of four) PCs account for 99.96% of the variability in X, they contribute nothing (zero sum of squares) to the fit of the regression model; instead, the last PC alone contributes everything. Thus, they propose the selection of the PCs to be based *not only* on the variance decomposition of the co-variables *but* on the contribution of each PC to the regression sum of squares, as well.

## 2. Objective

The main objective of this research was to investigate the potential of PCA used for genomic prediction. Alternative techniques of PCs selection were considered and different models were developed (simple and mixed linear models) for prediction of genomic breeding values for yield traits in Holstein.

More precisely, the objectives of this study were i) to use PCR for genomic predictions and compare the predictive ability of PCR and GBLUP models, ii) to investigate the difference of PCs selection based on either their eigenvalues or the correlation with the response variable (predictive ability) when using PCR for genomic predictions and iii) to evaluate the impact of accounting for population structure using PCs on genomic predictions accuracies of different traits in a GBLUP model.

(1) Hald's data (published by Hald, 1952, pp. 635-639) is a very nice example for studying collinearity among variables, developing variable selection methods, model building as well as checking for outliers and influential observations. The data have been widely used in statistical literature.

# 3. Material and methods

## 3.1 Data

For this study 66,116 test-day records up to 45 weeks in lactation for milk, fat and protein yield from 1,609 first lactation Holstein heifers were used. Heifers originated from 4 countries, Ireland (Teagasc, Moorepark Dairy Production), United Kingdom (Scottish Agricultural College), the Netherlands (Wageningen UR Livestock Research) and Sweden (Swedish University of Agricultural Science). The phenotypes were pre-adjusted to account for mean lactation curve, herd, nutritional treatment, milking frequency, year-month of milk test by management group, and experimental treatments (for a full description, see Veerkamp et al, 2012). Descriptive statistics of the pre-adjusted phenotypes are shown in Table 1.

**Table 1 Descriptive statistics of pre-adjusted test-day data of the milk traits**

| Trait | Mean | Sd | sdErr | Min | Max | n |
|---|---|---|---|---|---|---|
| Milk_yield (kg) | 23.837 | 4.442 | 0.111 | 0.990 | 38.980 | 1609 |
| Fat_yield (kg) | 0.928 | 0.175 | 0.004 | 0.120 | 1.790 | 1609 |
| Protein_yield (kg) | 0.721 | 0.127 | 0.003 | 0.040 | 1.340 | 1609 |

All animals were genotyped within the RobustMilk project with the Illumina BovineSNP50 Beadchip (Illumina Inc., San Diego, CA) containing 54,001 SNPs. After quality control 37,069 SNPs remained (Table 2). The dataset was split in four training (reference populations) and testing subsets. Each testing subset included all animals from only one country. Thereby, each animal was allocated to one subset, such that each animal had its genomic breeding value predicted once. Predictive ability of each model was assessed through validation and calculated as Pearson correlation between the predicted genomic breeding values and the phenotypes.

**Table 2 Number of cows with phenotype and genotype from the four countries**

| Population | Animals with genotypes | Animals with phenotypes and genotypes | SNPs |
|---|---|---|---|
| GBR | 566 | 416 | **37,069** |
| IRL | 413 | 394 | |
| NLD | 638 | 618 | |
| SWE | 214 | 181 | |
| **TOTAL** | **1,831** | **1,609** | |

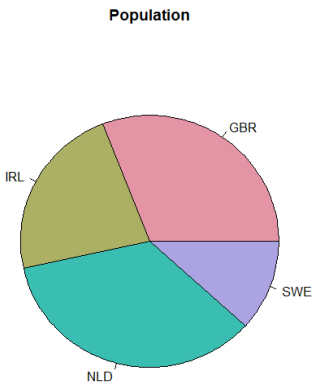



**Figure 2 Pie chart of individuals included in the dataset per country of origin**

### 3.2 Principal Component Regression model

The concept of principal component regression (PCR) i.e. the use of PCs in regression is not new. Kendal (1957) and Hotelling (1957) firstly proposed this idea, while in 1967 Jeffers published a well-known example. For the application in genomic prediction, the initial step for a PCR model is to perform PCA on the SNP matrix **X** *(n x p).* In our study, singular value decomposition via the function "prcomp" in R was performed. Let *r* be the rank of a matrix. The matrix **T** (*n x k*) of PCs is calculated as **T= XP**, where *k*<*r* and **P** *(p x k)* the loading matrix derived from SVD of **X** (which give the weights for the original variables). PCA was performed only in the reference population, the **P** matrix was extracted and the **T** matrix of components was calculated as $\mathbf{T_r = X_r P}$, where *r* in our case denotes the reference population. The **P** matrix extracted from the reference population was used to construct the **T** matrix of PCs for the test population such as $\mathbf{T_t = X_t P}$, where t denotes the test population.

Two different methods were tested for the selection of the PCs to be added in the PCR models. In a first approach, PCs were selected based on their eigenvalues (variation in the explanatory variables; genotypes) abbreviated as PCR (EIGEN). In a second model the selection of the PCs was based on their contribution to the sum of squares (SS) of the regression (variation in the response variable), PCR (SS). This contribution was developed in the training population through a PCR model where only the animals of the reference population were included (phenotypes and genotypes). In both cases, the effect of each PC was estimated in the reference population.

The PCR model is $y = \mu + Tg + e$, where **y** is the vector of phenotypic observations, **μ** is the overall mean, **T** is the matrix of principal components, **g** a vector of regression coefficients and ***e*** are the residuals.

In this case, genomic predictions are only based on the estimated effect of the PCs, so the genomic relationships among animals are not taken into account like in a GBLUP model. Moreover, by PCR the derived SNP effects (i.e. PCs) are treated as fixed effects and not as random effects as usual in genomic prediction models.

### *3.3 GBLUP model*

The following best linear unbiased prediction model with genomic relationship matrix (GBLUP) was fitted in ASReml-R (Butler et al, 2009): $y = Xb + Zu + e$, where, **y** is a vector of pre-adjusted records of test-day yield, **X** and **Z** are design matrices, **b, u** and **e** are vectors of fixed, additive genetic and residual effects, respectively. For additive and residual effects the following normal distributions are assumed $u \sim N(0, \sigma_u^2 G)$ and $e \sim N(0, I\sigma_e^2)$. The G matrix is a genomic relationship matrix calculated as described by VanRaden (2008). Fixed effects only include a mean effect, because phenotypic records were pre-adjusted. PCA was performed on the G matrix using the function "eigen" in R. The impact of accounting for population structure using PCs on genomic predictions accuracies was shown by adding an increasing number of PCs in the model as fixed effects. As a result, variance explained in G matrix is now entered to the model as PCs and thereby expected to be removed from the breeding values. The PCs were added one by one in the model and up to 1,000 PCs were included (accounting for 88.87 of the variability of the G matrix). So, the model becomes as follows:

$$y = GBLUP + \sum_{i=1}^{1,000} PCs + e$$

# 4. Results

## 4.1 PCR model

As a first step, PCA was performed on the whole dataset to check for any differences on the genotypic level between the four different populations. According to PCA graph (Figure 3) part of the GBR population can be distinguished from the rest with the first PC.

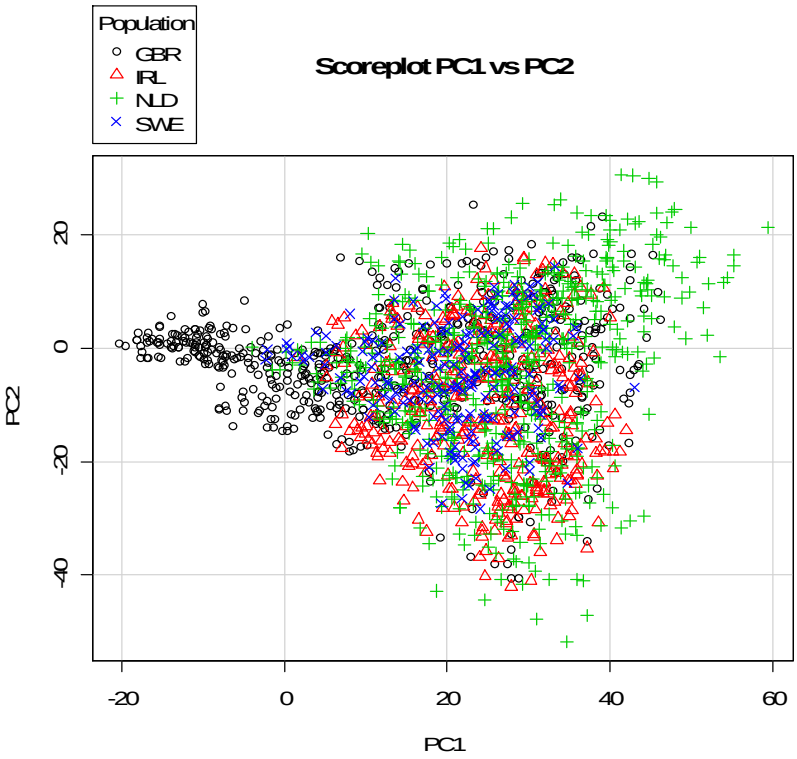



**Figure 3 Scoreplot of the first two principal components (PC1 vs. PC2). Principal component analysis was performed on the whole dataset**

Results from PCR and the basic GBLUP models for all three traits and four populations are shown in Table 3. The number of PCs included in the model with the highest accuracies obtained is presented as well. Two things are mainly interesting in these results. Firstly, the PCR (EIGEN) method outperforms (i.e. gave the highest accuracies) the PCR (SS) and the GBLUP_basic model in 10 out of 12 cases. Only in two cases slightly higher accuracies were obtained with PCR (SS) model. PCR (EIGEN) results had always higher accuracies compared to GBLUP. The difference was quite large and even doubled in some cases.

The second interesting point is that for the GBR population higher accuracies were achieved compared to the rest of the groups. Moreover, the GBR accuracies were twice as large compared to the others or more in many cases. Finally, the aim of data reduction in the models was achieved (even up to 99%) because in all the cases very few regressors (PCs) were needed compared to the number of initial variables (SNPs).

The pattern of the accuracies, when an increasing number of PCs is added in the models, depends both on the trait and the population. Furthermore, it should be noted that in many cases accuracies very close to the highest ones may be obtained with very few PCs (usually less than 50) (Figure 4).

**Table 3 Highest accuracies obtained for the three models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes.***

| Test Population | Trait | GBLUP basic | PCR (EIGEN) | Number of PCs | PCR (SS) | Number of PCs | Size of reference population |
|---|---|---|---|---|---|---|---|
| GBR | Milk | 0.250 | 0.311 | 25 | 0.306 | 118 | 1,193 |
| | Fat | 0.259 | 0.294 | 812 | 0.272 | 811 | |
| | Protein | 0.266 | 0.294 | 244 | 0.181 | 396 | |
| SWE | Milk | 0.162 | 0.178 | 1112 | 0.161 | 1060 | 1,428 |
| | Fat | 0.089 | 0.220 | 46 | 0.101 | 991 | |
| | Protein | 0.062 | 0.114 | 265 | 0.076 | 790 | |
| IRL | Milk | 0.060 | 0.147 | 967 | 0.118 | 758 | 1,215 |
| | Fat | 0.081 | 0.123 | 954 | 0.142 | 572 | |
| | Protein | 0.043 | 0.120 | 749 | 0.09 | 245 | |
| NLD | Milk | 0.156 | 0.210 | 20 | 0.172 | 4 | 991 |
| | Fat | 0.152 | 0.172 | 794 | 0.186 | 400 | |
| | Protein | 0.133 | 0.173 | 7 | 0.161 | 8 | |

* GBLUP_basic denotes the GBLUP model with only the mean as fixed effect and without any PCs included. PCR (EIGEN) and PCR (SS) are PCR models where the selection of PCs was based either on the eigenvalues or the regression sum of square, respectively.

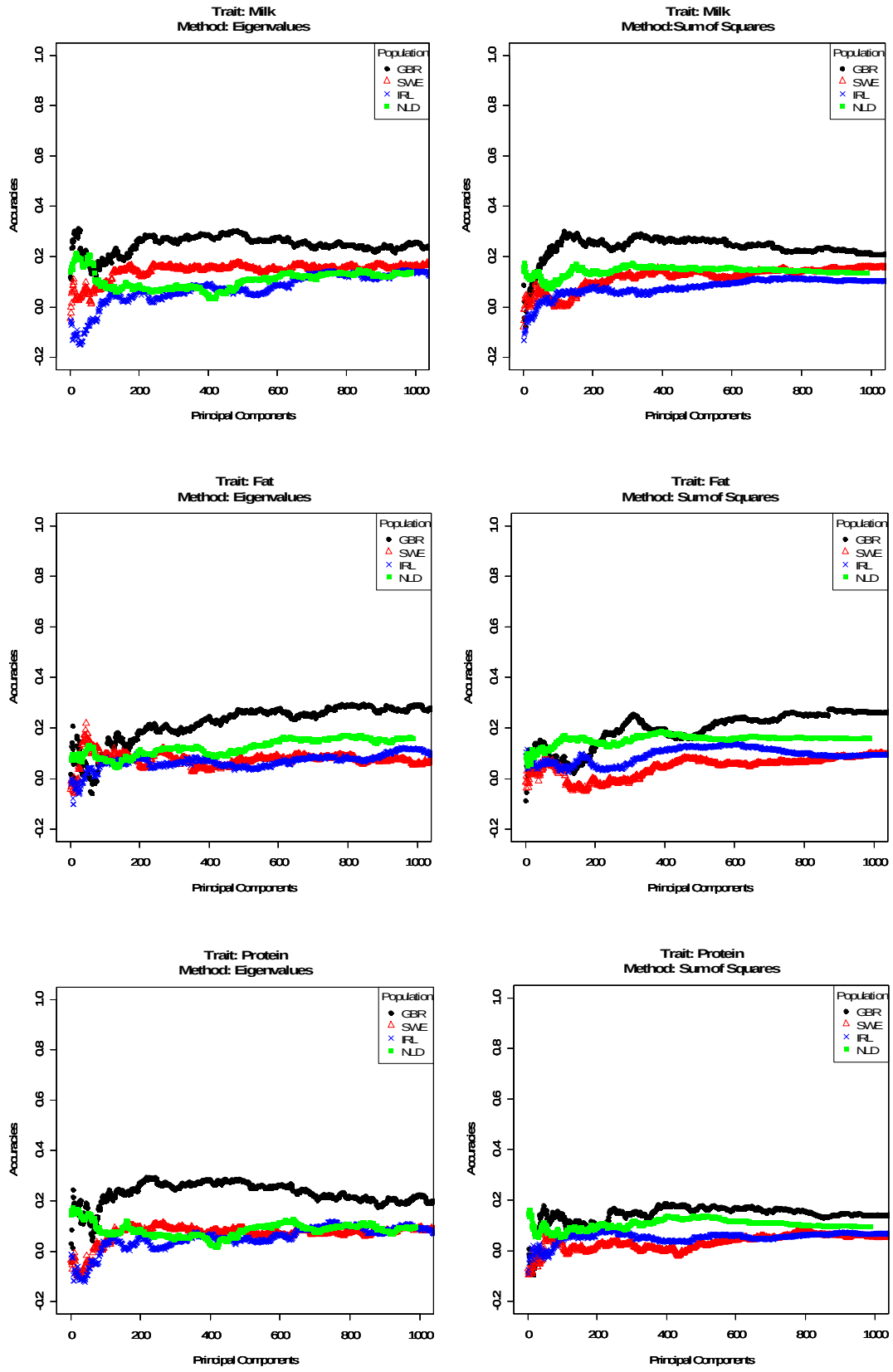



**Figure 4 Pattern of the accuracies for principal component regression models where the selection of PCs was based either on eigenvalues (left panel) or on sum of square contribution (right panel). An increasing number of PCs, one by one, up to 1,000 was fitted.**

Due to a partly differentiation of GBR population (Figure 3) it was interested to search inside this group and check exactly which animals were distinguished from the rest in the dataset. The reason is that GBR population actually consists of two different genetic lines (control vs. selection) as part of a still ongoing selection experiment established on 1992. This has been captured from PCA analysis, where the one genetic group is separated in the subspace of PCA with the first PC (Figure 5). However, it should be noted that the 1st PC captures only 1.5% while the first two PCs 4% of the total original variability of the SNP data (Table 4). Accuracies of predicted genomic breeding values for these lines are shown in Table 4. Again, PCR where the selection of PCs was based on eigenvalues gave the best results. The GBR_1 line had higher accuracies than GBR_2. However, the increase on the accuracies of GBR_1 compared to the whole GBR population was only present in milk yield. For fat and protein yield there was a decrease on the accuracies for GBR_1 compared to GBR.

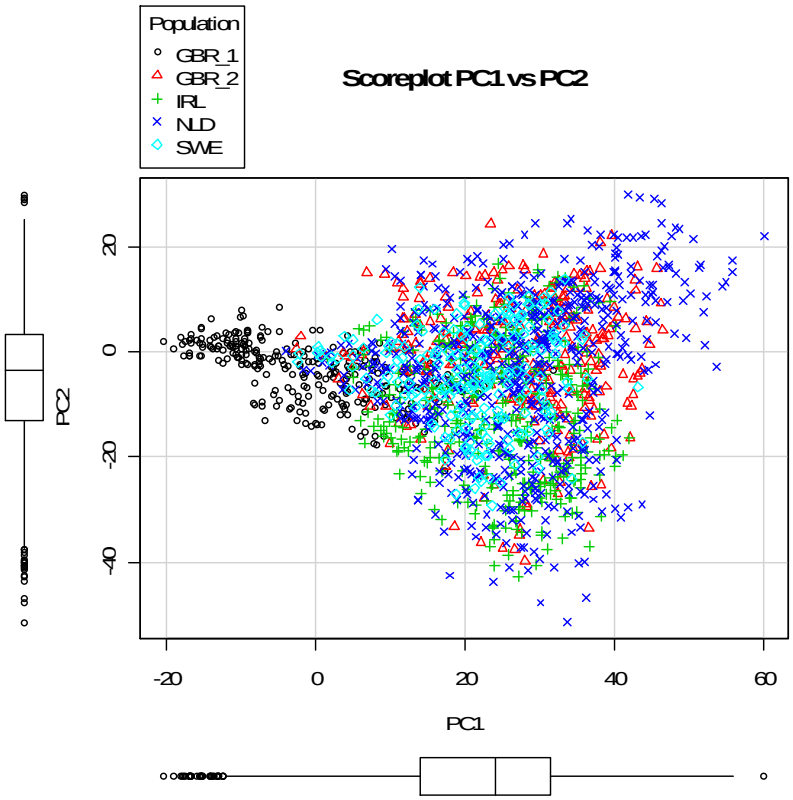



**Figure 5 Scoreplot of the first two principal components (PC1 vs. PC2). Principal component analysis was performed on the whole dataset, whereas GBR population was split into the two different genetic groups.**

**Table 4 Cumulative proportion of the original variability captured by principal components**

| Number of PCs | Cumulative Proportion (%) |
|---|---|
| 1 | 1.4 |
| 2 | 3.9 |
| 37 | 20 |
| 138 | 40 |
| 326 | 60 |
| 668 | 80 |
| 967 | 90 |

**Table 5 Highest accuracies obtained for GBR total and the two GBR lines separately. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes.***

| Test Population | Trait | GBLUP basic | PCR (EIGEN) | Number of PCs | PCR (SS) | Number of PCs | Size of reference population |
|---|---|---|---|---|---|---|---|
| GBR | Milk | 0.250 | 0.311 | 25 | 0.306 | 118 | 1,193 |
| | Fat | 0.259 | 0.294 | 812 | 0.272 | 811 | |
| | Protein | 0.266 | 0.294 | 244 | 0.181 | 396 | |
| GBR_1 | Milk | 0.261 | 0.436 | 14 | 0.355 | 2 | 1,403 |
| | Fat | 0.068 | 0.164 | 776 | 0.130 | 3 | |
| | Protein | -0.005 | 0.245 | 14 | 0.163 | 1 | |
| GBR_2 | Milk | 0.103 | 0.161 | 3 | 0.151 | 11 | 1,399 |
| | Fat | -0.021 | 0.082 | 1061 | 0.07 | 751 | |
| | Protein | 0.006 | 0.0662 | 1 | 0.078 | 2 | |

* GBLUP_basic denotes the GBLUP model with only the mean as fixed effect and without any PCs included. PCR (EIGEN) and PCR (SS) are PCR models where the selection of PCs was based either on the eigenvalues or the regression sum of square, respectively. GBR_1 and GBR_2 are the two genetic groups of GBR population.

To summarize the above comparison of PCR and the basic GBLUP models i) higher accuracies were obtained for the most genetically diverged population (GBR), and ii) the PCR model where the PC selection was based on their eigenvalues resulted in higher accuracies (on average) compared to a PCR model where the PCs were selected based on their regression sum of square contribution. Moreover, PCR in all the cases outperforms GBLUP. If only the size of the reference population matters in order to have accurately predicted breeding values then we would expect the SWE and the IRL populations to have higher accuracies than the GBR and the NLD populations. On the opposite, in almost all cases the GBR followed by the NLD population resulted in higher accuracies. Moreover, given that part of GBR population

(more specifically the GBR_1 line) is genetically separated from the others we would expect to be more difficult to accurately estimate the GEBVs of this population. In contrast, this is the most accurately predicted population.

The explanation of the above results may simply be derived from the definition of principal components, that is "***axes of variation***" of the original data. As a result, the first PCs are going to the direction of the maximum variability in the data which means to the GBR population. However, in our approach the population for which the GEBVs were to be predicted was always excluded from the reference population which implies that the SNP variability in the evaluated population did not affect the definition of the PCs. To this purpose an alternative method to extract PCs was investigated.

In the new case the whole dataset (all animals included) was used to perform PCA and then the dataset was split into a training and test part by selecting the animals (rows) of interest. This resulted in an extra increase for the GEBVs for all traits, mainly to the GBR population and secondly to the GBR_1 incorporated with a further reduction on the number of the PCs needed to achieve the highest accuracies with the PCR (EIGEN) model (results shown in Table 6). Interestingly, accuracies of 0.502 and 0.465 for protein and milk yield were obtained with only the first or the first two PCs, respectively, for the GBR population. Even more interesting is that for the GBR_1 the first PC resulted in highest accuracies but for fat yield. From the results it is also clear that for the population that is more diverged from the rest (GBR) fewer PCs are needed in the model (Table 5). For the rest of the populations there was either an (substantial) increase or a decrease on the GEBVs accuracies. On average, this approach (extraction of PCs from the whole dataset) resulted in higher accuracies.

**Table 6 Highest accuracies obtained for principal component regression models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. Difference denotes the change on the accuracies between the two different PCA methods.***

| Test Population | Trait | PCR (EIGEN) | Difference | Number of PCs | PCR (SS) | Difference | Number of PCs |
|---|---|---|---|---|---|---|---|
| GBR | Milk | 0.311 | | 25 | 0.306 | | 118 |
| | Fat | 0.294 | | 812 | 0.272 | | 811 |
| | Protein | 0.294 | | 244 | 0.181 | | 396 |
| GBR_new | Milk | 0.465 | 49.20% | 2 | 0.247 | -19.28% | 81 |
| | Fat | 0.474 | 61.22% | 21 | 0.303 | 11.40% | 273 |
| | Protein | 0.502 | 70.75% | 1 | 0.244 | 34.81% | 1146 |
| GBR_1 | Milk | 0.436 | | 14 | 0.355 | | 2 |
| | Fat | 0.164 | | 776 | 0.130 | | 3 |
| | Protein | 0.245 | | 14 | 0.163 | | 1 |
| GBR_1_new | Milk | 0.450 | 3.21% | 6 | 0.23 | -35.21% | 458 |
| | Fat | 0.176 | 7.32% | 1 | 0.213 | 63.85% | 219 |
| | Protein | 0.280 | 14.29% | 7 | 0.092 | -43.56% | 1 |
| GBR_2 | Milk | 0.161 | | 3 | 0.151 | | 11 |
| | Fat | 0.082 | | 1061 | 0.07 | | 751 |
| | Protein | 0.0662 | | 1 | 0.078 | | 2 |
| GBR_2_new | Milk | 0.194 | 20.50% | 144 | 0.145 | -3.97% | 46 |
| | Fat | 0.120 | 46.34% | 593 | 0.09 | 28.57% | 1370 |
| | Protein | 0.100 | 51.06% | 151 | 0.082 | 5.13% | 233 |
| SWE | Milk | 0.178 | | 1112 | 0.161 | | 1060 |
| | Fat | 0.220 | | 46 | 0.101 | | 991 |
| | Protein | 0.114 | | 265 | 0.076 | | 790 |
| SWE_new | Milk | 0.210 | 17.98% | 365 | 0.165 | 2.48% | 344 |
| | Fat | 0.175 | -20.45% | 1425 | 0.177 | 75.25% | 871 |
| | Protein | 0.250 | 119.30% | 1424 | 0.196 | 157.89% | 1419 |
| IRL | Milk | 0.147 | | 967 | 0.118 | | 758 |
| | Fat | 0.123 | | 954 | 0.142 | | 572 |
| | Protein | 0.120 | | 749 | 0.09 | | 245 |
| IRL_new | Milk | 0.143 | -2.72% | 92 | 0.185 | 56.78% | 288 |
| | Fat | 0.155 | 26.02% | 790 | 0.122 | -14.08% | 965 |
| | Protein | 0.159 | 32.50% | 94 | 0.134 | 48.89% | 273 |
| NLD | Milk | 0.210 | | 20 | 0.172 | | 4 |
| | Fat | 0.172 | | 794 | 0.186 | | 400 |
| | Protein | 0.173 | | 7 | 0.161 | | 8 |
| NLD_new | Milk | 0.190 | -9.52% | 24 | 0.171 | -0.58% | 16 |
| | Fat | 0.171 | -0.58% | 585 | 0.176 | -5.38% | 78 |
| | Protein | 0.165 | -4.62% | 5 | 0.165 | 2.48% | 4 |

* Selection of the PCs was based either on the eigenvalues (PCR (EIGEN)) or the regression sum of square (PCR (SS)). Two different methods of applying principal component analysis (either separately for reference and test parts or on the whole dataset) were compared. The term "new" indicates the method where PCA performed on the whole dataset.

By considering the differences between the two different approaches of extracting PCA on the number of PCs needed per trait and per population a substantial decrease for GBR and GBR_1 for all traits was observed (Tables 7 and 8). However, it should be mentioned again that in many cases high accuracies very close to the highest ones were obtained with very few PCs (usually less than 50).

**Table 7 Number of principal components needed to achieve highest accuracies per trait and per population when principal component analysis applied on the reference part**

| Population | Milk | Fat | Protein | Average |
|---|---|---|---|---|
| GBR | 25 | 812 | 244 | 360 |
| GBR_1 | 14 | 776 | 14 | 268 |
| GBR_2 | 3 | 1061 | 1 | 355 |
| SWE | 1112 | 46 | 265 | 474 |
| IRL | 967 | 954 | 749 | 890 |
| NLD | 20 | 794 | 7 | 273 |
| **Average** | 356.83 | 740.50 | 213.33 | |

**Table 8 Number of PCs needed to achieve highest accuracies per trait and per population when the entire dataset was used to perform principal component analysis**

| Population | Milk | Fat | Protein | Average |
|---|---|---|---|---|
| GBR_new | 2 | 21 | 1 | 8 |
| GBR_1_new | 6 | 1 | 7 | 5 |
| GBR_2_new | 144 | 593 | 151 | 296 |
| SWE_new | 365 | 1425 | 1424 | 1071 |
| IRL_new | 92 | 790 | 94 | 325 |
| NLD_new | 24 | 585 | 5 | 204 |
| **Average** | 105.33 | 569.17 | 280.33 | |

### *4.2 GBLUP model*

The aim of the GBLUP model was to check the importance of population structure on the accuracy of genomic breeding values. PCA was performed on the G matrix and the PCs extracted were added to the basic GBLUP model as fixed effects. By adding PCs to the model an increase to the accuracies was observed for all populations as well as for all traits (from 6.2% up to 141.9%). So, information from the basic GBLUP model was extracted, transformed (in terms of PCs) and then added to the model as covariates. In this way no further improvement of the model would be expected, but perhaps rather a decrease in accuracy, because information may be removed from the breeding values. Nevertheless in all cases there was a significant improvement of the model with the contribution of PCs (Table 9, Figure 7). An adjective explanation would be that in GBLUP, somehow, we are losing

information concerning SNP variation. For instance, genomic relationship matrix (G) has been constructed based on genotypic information. Thus, G matrix includes information about SNP variation of the dataset. Nonetheless, if by PCA (technique based on data variation) we are able to increase the accuracies, although PCA has been applied to G matrix, this could reasonably imply that this loss of information is correlated to variation.

Based on changes in trend of the accuracies across the number of PCs added to the model (Figure 7) four main areas on the plots could be distinguished (0-200, 200-600, 600-800, 800-1,000 PCs). On average, accuracies were higher for GBR followed by NLD. For GBR there was a clear increase on the accuracies between 200-800 PCs, while a decrease started from 1 to 200 PCs. The pattern was the same for all traits. For NLD, SWE and IRL, accuracies were almost stabilized by fitting the first 200 PCs for all traits. For fat and protein we observe an interesting area between 200 and 600 PCs. In this range accuracies of SWE make a curve with a minimum point while at the same time accuracies of IRL make a curve with a maximum; implying that these PCs are more descriptive for IRL than SWE. Between 600-800 PCs accuracies remain stable for SWE and IRL basically for fat and protein. After 800 PCs for all traits we observe an unstable situation for GBR and NLD, stabilization of SWE while an increase of IRL. However, it is unknown what will happen if more PCs added till the last one. For this study it was not possible due to computation time.

According to the plot in Figure 6, no population can be distinguished by PCA when it is performed on the G matrix. This is different to what we have seen when performing the PCA on the genotypes. So, while the genotypes are informative and able to separate different populations the marker genetic relationships represent a more related view of the animals. This is logical because the spot of each individual in the (PCA) space is not only based on its own information (genotype) but on the others as well (genetic relationship) when G matrix is used to extract PCs.

Comparing GBLUP and PCR models the same pattern of predictions is observed concerning different populations such that in both cases GBR followed by NLD gave the highest accuracies. On the contrary, an opposite situation was observed for the number of PCs needed to achieve higher accuracies for GBR. In PCR the first few PCs were needed to achieve the highest accuracies, while in the GBLUP around 800 PCs were needed. However, PCR and GBLUP models have been developed for different purposes. The first is used for direct estimation of genomic breeding values while the second one to identify and remove

information that creates noise to the model. Table 10 summarizes the highest accuracies achieved in GBLUP models (with or without PCs) and the PCR (EIGEN) models in the two situations where either all animals were included to extract PCs or separately on the reference and test populations.

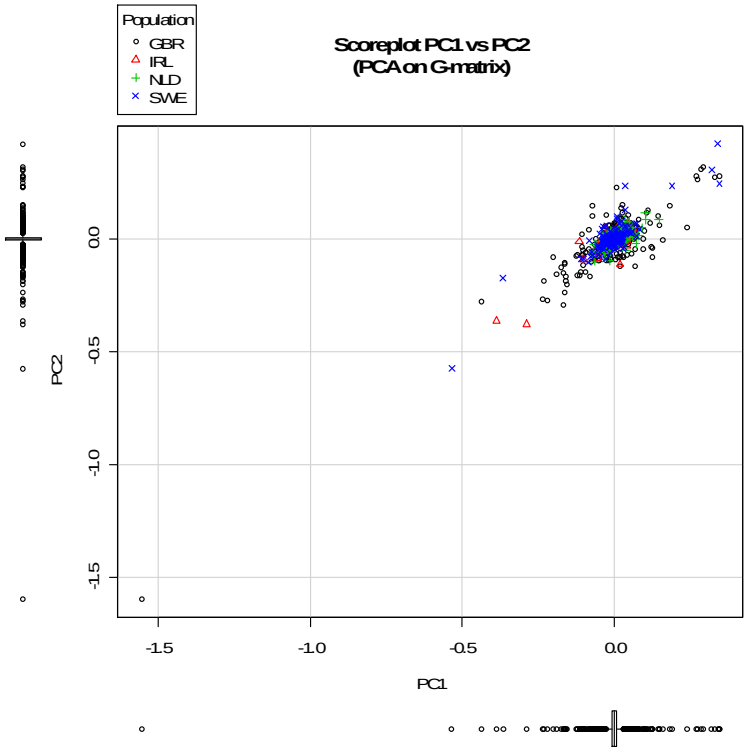
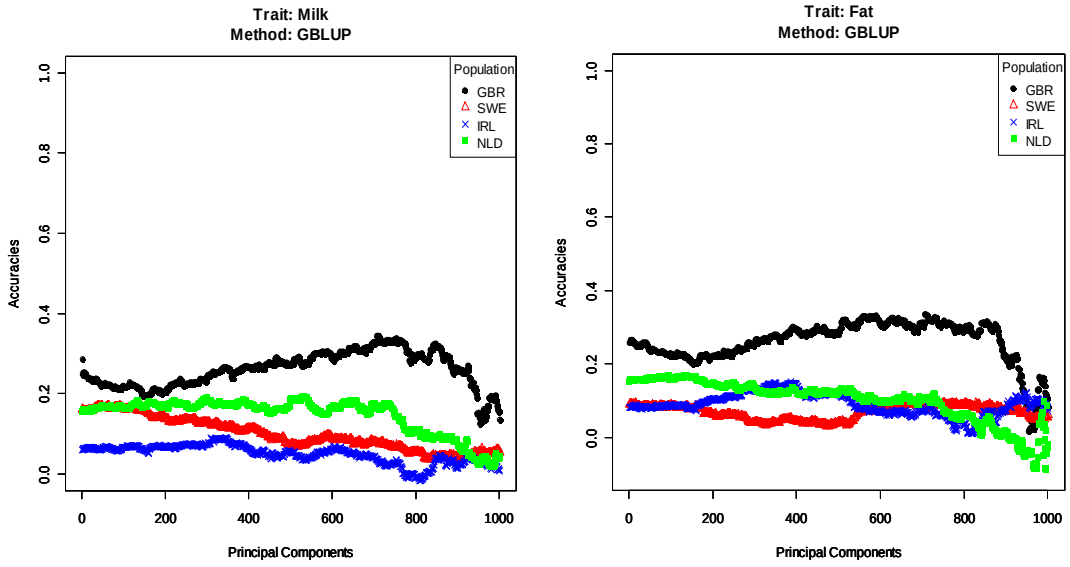



**Figure 6 Scoreplot of the first two principal components (PC1 vs. PC2). Principal component analysis was performed on the G matrix**

**Table 9 Highest accuracies obtained from GBLUP models with (GBLUP_PCs) and without principal components (GBLUP_basic) included. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes. Difference indicates the change on the accuracies between the two GBLUP models with or without PCs.**

| Population | Trait | GBLUP basic | GBLUP PCs | Number of PCs | Difference (%) |
|---|---|---|---|---|---|
| GBR | Milk | 0.250 | 0.344 | 708 | 38 |
| | Fat | 0.259 | 0.336 | 709 | 30 |
| | Protein | 0.266 | 0.325 | 708 | 22 |
| SWE | Milk | 0.162 | 0.172 | 64 | 6 |
| | Fat | 0.089 | 0.098 | 682 | 10 |
| | Protein | 0.062 | 0.084 | 93 | 35 |
| IRL | Milk | 0.060 | 0.090 | 346 | 50 |
| | Fat | 0.081 | 0.151 | 387 | 86 |
| | Protein | 0.043 | 0.104 | 986 | 142 |
| NLD | Milk | 0.156 | 0.192 | 531 | 23 |
| | Fat | 0.152 | 0.168 | 138 | 11 |
| | Protein | 0.133 | 0.160 | 132 | 20 |



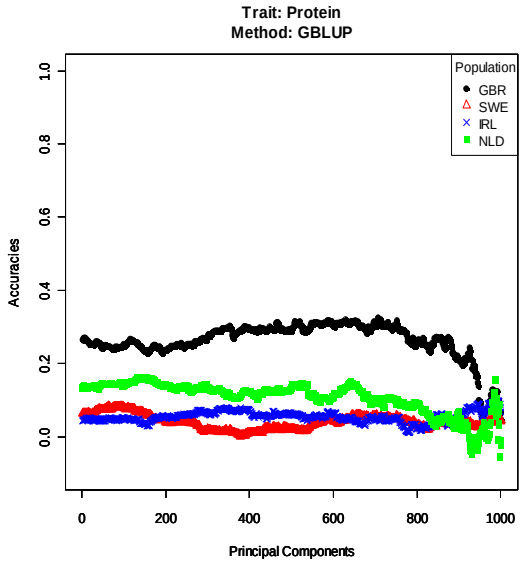



**Figure 7 Pattern of the accuracies for GBLUP models for test-day milk, fat and protein yield for four countries. An increasing number of PCs (extracted from G matrix) was fitted, one by one, up to 1,000.**

**Table 10 Summarized table of highest accuracies for GBLUP with (GBLUP_PCs) and without principal components (GBLUP_basic) included and principal component regression (PCR (EIGEN)) models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes.***

| Population | Trait | GBLUP basic | GBLUP PCs | Number of PCs | PCR (EIGEN) | Number of PCs | PCR (EIGEN_new) | Number of PCs |
|---|---|---|---|---|---|---|---|---|
| GBR | Milk | 0.25 | 0.344 | 708 | 0.311 | 25 | 0.464 | 1 |
| | Fat | 0.259 | 0.336 | 709 | 0.294 | 812 | 0.474 | 21 |
| | Protein | 0.266 | 0.325 | 708 | 0.294 | 244 | 0.502 | 1 |
| SWE | Milk | 0.162 | 0.172 | 64 | 0.178 | 1112 | 0.21 | 365 |
| | Fat | 0.089 | 0.098 | 682 | 0.220 | 46 | 0.175 | 1425 |
| | Protein | 0.062 | 0.084 | 93 | 0.114 | 265 | 0.25 | 1424 |
| IRL | Milk | 0.06 | 0.09 | 346 | 0.147 | 967 | 0.143 | 92 |
| | Fat | 0.081 | 0.151 | 387 | 0.123 | 954 | 0.155 | 790 |
| | Protein | 0.043 | 0.104 | 986 | 0.12 | 749 | 0.159 | 94 |
| NLD | Milk | 0.156 | 0.192 | 531 | 0.210 | 20 | 0.19 | 24 |
| | Fat | 0.152 | 0.168 | 138 | 0.172 | 794 | 0.171 | 585 |
| | Protein | 0.133 | 0.16 | 132 | 0.173 | 7 | 0.165 | 5 |

* PCR (EIGEN) denotes that selection of the PCs was based on their eigenvalues. The term “new” indicates that PCA was performed on the whole dataset.

To have a better view of the different GBLUP models (with or without PCs) a model with only PCs included as fixed effects but not the marker genetic relationships (G matrix) was developed. This means that the mixed model becomes a simple linear model with fixed effects the PCs extracted from the G matrix. For GBR and NLD highest accuracies where very low and around zero (Table 11). Surprisingly, for SWE and IRL similar or even higher accuracies were achieved compared to GBLUP models. Moreover, in 7 out of 12 cases PCR (SS) outperformed PCR (EIGEN). So, in the case of PCs extracted from G matrix results are opposite from previous (PCA applied on the genotypes) with SWE and IRL resulting in higher accuracies than GBR and NLD and sum of square contribution method of selecting PCs being, on average, as good as eigenvalues in PCR models.

Another interesting part deriving from Table 11 is that if add the values of GBLUP basic and PCR (EIGEN), similar values to the ones in the GBLUP model with PCs are obtained. This supports the idea of loss of information in the ordinary GBLUP model. However, this happens only for the GBR and NLD data and this sum of accuracies of two different models seems quite arbitrary.

The above results are consistent to the literature. Habier et al (2007) has shown through simulations that genetic relationships alone can affect GEBVs accuracies even if there is no LD between QTL and the marker. Their research was driven by Fernando 1998 who stated that additive relationships can also be captured by the markers. Thus, part of the genomic accuracies is only due to genetic relationships. In our case, for SWE and IRL, a fixed regression with only additive genetic relationships (in terms of PCs) results in at least same accuracies as the GBLUP model.

**Table 11 Highest accuracies for GBLUP with (GBLUP_PCs) and without principal components (GBLUP_basic) included and principal component regression models. Accuracies were calculated as Pearson correlation between the predicted genomic breeding values and the observed phenotypes.***

| Population | Trait | GBLUP basic | GBLUP PCs | Number of PCs | PCR (EIGEN) | Number of PCs | PCR (SS) | Number of PCs |
|---|---|---|---|---|---|---|---|---|
| GBR | Milk | 0.250 | 0.344 | 708 | 0.062 | 1119 | 0.104 | 854 |
| | Fat | 0.259 | 0.336 | 709 | 0.069 | 1190 | 0.074 | 10 |
| | Protein | 0.266 | 0.325 | 708 | 0.081 | 154 | 0.086 | 989 |
| SWE | Milk | 0.162 | 0.172 | 64 | 0.140 | 55 | 0.158 | 1399 |
| | Fat | 0.089 | 0.098 | 682 | 0.196 | 3 | 0.147 | 358 |
| | Protein | 0.062 | 0.084 | 93 | 0.126 | 1426 | 0.133 | 1323 |
| IRL | Milk | 0.060 | 0.09 | 346 | 0.081 | 218 | 0.101 | 1199 |
| | Fat | 0.081 | 0.151 | 387 | 0.120 | 1175 | 0.116 | 1209 |
| | Protein | 0.043 | 0.104 | 986 | 0.115 | 308 | 0.088 | 719 |
| NLD | Milk | 0.156 | 0.192 | 531 | 0.056 | 322 | 0.060 | 30 |
| | Fat | 0.152 | 0.168 | 138 | 0.033 | 3 | 0.024 | 949 |
| | Protein | 0.133 | 0.160 | 132 | 0.053 | 3 | 0.059 | 1 |

* Principal components were extracted from the G matrix. For the principal component regression the selection of the PCs was based either on the eigenvalues (PCR (EIGEN)) or the regression sum of square (PCR (SS)).

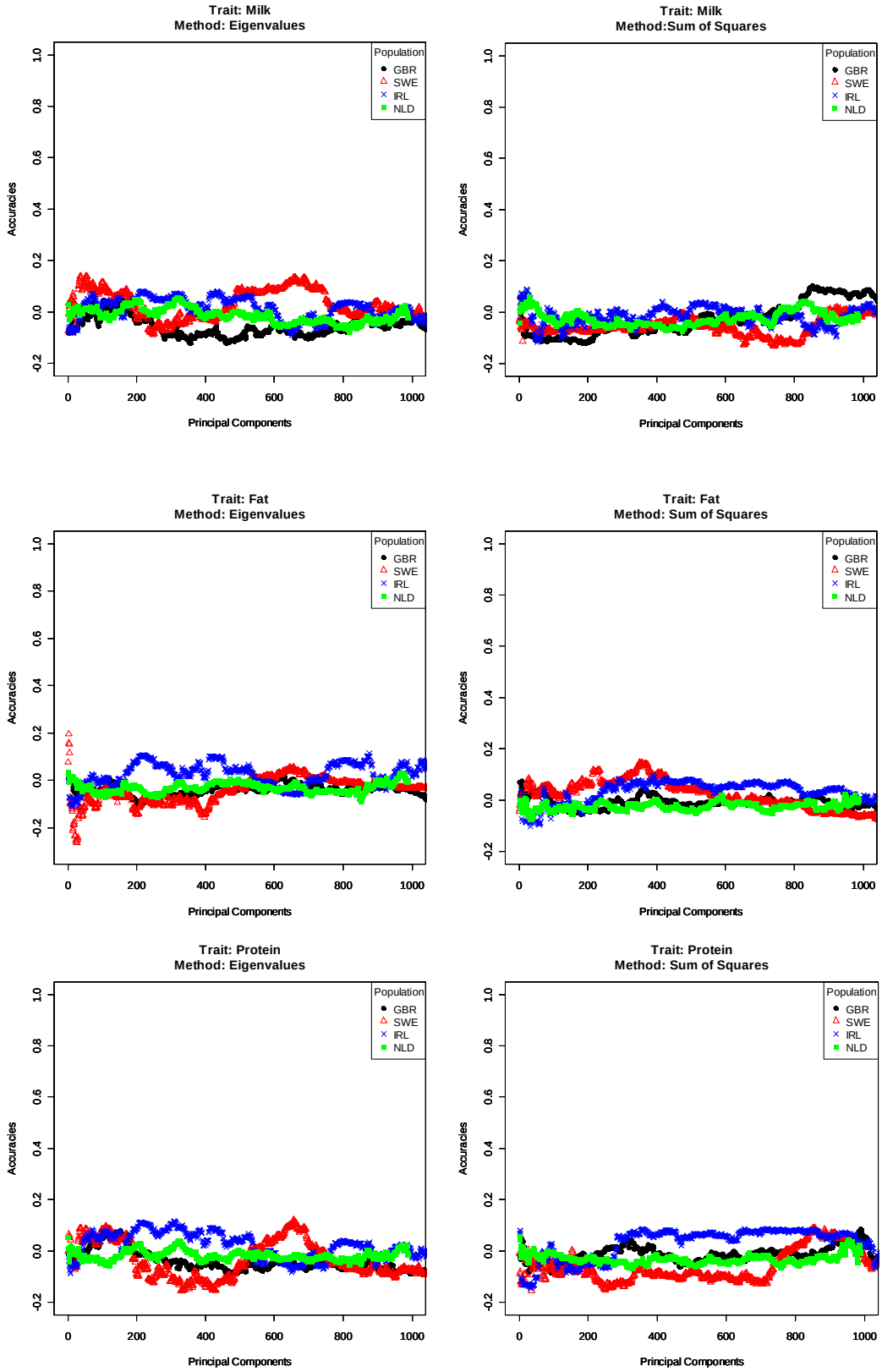



**Figure 8 Pattern of the accuracies for PCR models where the selection of PCs was based either on eigenvalues (left panel) or on sum of square contribution (right panel). An increasing number of PCs, one by one, up to 1,000 was fitted. PCA applied on the G-matrix.**

## 5. Discussion

Principal component analysis belongs in the general framework of multivariate statistical analysis techniques. Principal component regression is an alternative method to perform data reduction in the model as well as to solve problems of dependencies among variables (multicollinearity). Due to these characteristics PCA and its regression (PCR) were selected to be used as an alternative way for genomic predictions. The contribution of PCR to model improvements mainly derives from the ability of the PCA to capture the original variability of the dataset in a small set of PCs while these PCs are by default uncorrelated. In this study singular value decomposition was used to extract PCs from a given X matrix including the genotypes of the animals.

Results show that considerably high accuracies can be achieved with a multiple regression model (PCR) where SNPs are entered as fixed effects (in the sense of principal components). Breeding values of the animals were not available in our data, thus genomic prediction accuracies were back correlated to the phenotypes. Therefore, even higher accuracies would be expected if breeding values were used. Data reduction was at least 96% of the original data. Highest accuracies were achieved for a wide number of PCs that were fit in PCR model, from only one to more than one thousand. This is a wider range than found in the literature were it was shown through simulations that highest accuracies can be achieved in the range between 250 to 350 PCs (Solberg et al, 2009; Macciotta et al, 2010a). By fitting PCs one by one in the model it is shown that most of the PCs have a contribution to the model (either positive or negative) and thus the trend line of the accuracies is not a stable curve but fluctuates. As a result, empirical thresholds for selecting PCs (e.g. by keeping PCs that explain 80% of the original variability according to eigenvalues) may not provide the highest accuracies that can be achieved through the components. Thus, the way PCs are selected and kept in a regression model should be reconsidered.

Principal component regression outperforms a GBLUP model for predicting genomic breeding values in all cases in our analysis. It is impressive that a multiple linear model with very few regressors (compared to the original variables) resulted in higher accuracies than a mixed linear model where further information, in terms of genetic relationships, is included. This may be explained by the fact that in a fixed regression the assumption of equal contributions of each SNP does not hold as in a BLUP model. These findings are in

agreement with the literature (Pintus et al, 2012), although in that case PCs were treated as random variables in a BLUP model.

Moreover, the selection of PCs in a PCR model based on their eigenvalues instead of sum of squares contribution resulted in higher accuracies (on average) in our analysis. However, it is unknown what the case will be if the number of observations is increased (e.g. some thousands of animals in the reference population), so more information will be added in terms of sum of squares.

The dataset on which PCA is performed is also of importance. Therefore, when genotypes of test population were included in the extraction of the principal components the accuracies significantly increased, especially for the most genetically diverged population (GBR). This indicates that the weights of each SNP (in the P matrix of eigenvectors) were better estimated from the model. However, this was not always the case and decreases in accuracies were also observed in some cases. Furthermore, it should be noted that in this analysis phenotypic records from different countries and experimental stations were used. The records were already pre-adjusted (more details in Veerkamp et al, 2012). However, it cannot be 100% certain that all environmental effects are excluded.

The accuracies of a PCR model depend not only on the population structure but on the trait as well. Differences of genomic accuracies between traits as well as breeds have also been reported in other studies (Pintus et al, 2012). In addition, the pattern of the accuracies for each population and each trait differs.

Accuracies of 0.502 and 0.465 for protein and milk yield were obtained with only the first or the first two PCs, respectively, for the GBR population. Even more interesting is that for the one genetic line of GBR (GBR_1) that is separated from the rest animals in the data the first PC resulted in highest accuracies but for fat yield. This indicates that perhaps it is not clear what kind of information is captured in the PCs in terms of traits (QTLs). In other words, variation in the genome (at least as it can be captured by PCA) may not be informative for the expression of quantitative traits. The highest accuracy achieved in our analysis was 0.5 for GBR population in protein yield. In this case the genotypes of the test population were included in the extraction of the PCs. Interestingly, the GBR population was the one genetically separated from the rest by PCA. Thus, we wouldn't expect to accurately predict GBR genomic breeding values. The explanation may simply be derived from the definition of principal components, that is "***axes of variation***" of the original data. As a result, the first PCs

are going to the direction of the maximum variability in the data which means to the GBR population in our case. This drive us to hypothesize that perhaps PCA could be a helpful tool in across-breeds predictions.

It should be noted that for all models developed in the analysis only one replicate (test dataset) per country and trait was used. However, by using different randomly selected test parts and several cross-validations may yield in different results and ranking of the methods.

Concerning the GBLUP models, firstly a basic GBLUP model was fitted with only the mean as fixed effect because the phenotypic records were pre-adjusted. Then, the model was expanded by adding PCs extracted from genomic relationship matrix (G-matrix). In all cases an increase on the accuracies (~40% on average) observed compared to the basic GBLUP model. This is in agreement with findings in the literature were similar models were developed in across breeds genomic predictions. Daetwyler et al (2011) using a similar GBLUP model also found a substantial increase (300%, from 0.05 to 0.2) on the accuracy of one population in across breed predictions in sheep but only for one of the two traits investigated. However, this trend of increased accuracies when PCs are fit is unexpected. The reason is that variance explained in G matrix is now entered to the model as PCs and thereby expected to be removed from the breeding values. In other words, by correcting for population structure in the model a decrease of the predicted values would be expected. However, this increase in the accuracies indicates that there is a loss of information in the basic GBLUP model. According to this analysis it was not clear why this happened. Therefore more research is needed in this direction to have a better insight.

On the other hand, when a fixed regression applied with PCs derived from the G-matrix nonzero accuracies obtained. In addition, for SWE and IRL accuracies where at least as high as the GBLUP with PCs. The explanation has already been stated from Fernando (1998) and Habier et al (2007). According to those authors part of the GEBVs accuracies can be only due to additive genetic relationships even if there is no LD between markers and QTLs. The reason is that marker effects which are used for genomic predictions capture additive genetic relationships as well.

As a further research, it would be interesting to check the performance of a GBLUP model where the inverse of G matrix would be substituted by the inverse matrix of the PCs. Thus, variation on the genotypes will directly be used and take the place of additive genetic relationships in the model.

PCA is a useful and easy tool for preliminary analysis of the data. In an initial step it can be used to check for grouping in the data. Especially in a dataset where different breeds are included PCA can give an overview through easy to understand graphical representation of genetic similarities or differences among different breeds. In across breed predictions it would be interesting to check accuracies if only PCs that are descriptive for each breed and can separate it from the rest are used to predict genomic breeding values of the animals of the specific breed.

In a further step, it would be interesting to investigate which SNPs dominate each component, especially the ones that can differentiate the groups in the dataset and try to use this information for breed specification or correlation with a specific trait. PCA has already been used for genome-wide association studies (Bolormaa et al, 2010), thus broaden its application in a more thoroughgoing perspective in genomic data of breeding programs. However, we should always consider the amount of original variation captured by each PC before extracting conclusions of biological meaning.

Principal component analysis and multivariate analyses in general has been used in several studies to extract information from markers and have been proved to be a nice tool for capturing genetic variability. Moreover, as explanatory statistical methods they do not hold on strong assumptions of the data. However, we should still be very careful when applying multivariate analysis in genomic data and especially when interpreting the results. Jombart et (2009) gives a nice overview of wrongly used multivariate analysis in different datasets as well as fallacies during the interpretation of the results. On this direction, Edwards (2003) discusses “erroneous conclusions” derived from wrongly interpretation of genetic markers information in human genetic diversity studies. A lot of information can be derived by the plethora of genetic markers, still the way this information is used has to be optimized.

## 6. Conclusions and implications

Our results indicate that considerably high accuracies in genomic predictions can be obtained with a simple linear regression model (PCR) where very few components are fitted. Accuracies obtained by PCR were at least as high as a GBLUP model. However, the accuracies as well as the number of PCs to be fit into the PCR model depend both on the dataset and the trait. Due to the fact that the Holsteins populations used in this study originated from four different countries but highest accuracies obtained for the only one genetically diverged (GBR) from the rest, according to PCA graphs, it is hypothesized that PCR methodology could also be tested in across breed genomic predictions. However, the variance of the number of PCs added to the model in which highest accuracies occurred (from one to more than 1,000 PCs) is a drawback of the method.

It is proposed PCA to be applied on the whole SNP data and then split to training and testing parts for cross-validation to estimate accuracies of genomic breeding values. According to our analysis the selection of PCs based on their eigenvalues and not on the regression sum of square contribution (correlation to the trait) resulted in better results. However, this cannot be generalised for genomic data, especially when some thousands of phenotypes will be included, so more information will be added in terms of sum of squares. Furthermore, the first few PCs alone should be taken into consideration when fitting a regression model, even if they do not capture a significant amount of the total original data variability. It was shown in the analysis that highest accuracies could be achieved even with the first PC. Thus the methodology of keeping PCs that contribute over 80% of the variability of the SNP data should be reconsidered.

The simplicity of the method, the (considerably) fast computation, dimensionality reduction while at the same time keeping all the original variables in the dataset as well as the ability of both predicting and identifying groups in the data (pattern recognition) could be stated as the main advantages of PCR. The above elements together with nice performance in predictive ability of the model with real data characterizes PCR as an attractive tool for animal breeding.

## References

Adelman, I. and C.T. Morris, 1982a, Factor analysis and development: A reply, Journal of Development Economics 11, no. 1.

Adelman, I. and C.T. Morris, 1982b, Factor analysis and development: A rejoinder to a rejoinder, Journal of Development Economics 11, no. 1, 129.

Amer P.R., Wall E., Nuhs J., Winters M. and Coffey M.P., 2011. Sources of benefits from genetic improvement in the UK dairy industry and their impacts on producers and consumers. Interbull bulletin no. 40. Stavanger, Norway, August 26-29.

Berry PD, Kearney F, and Harris BL. 2009. Genomic selection in Ireland. Proceedings Interbull Genomic Selection Workshop. Uppsala, Sweden

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

Boneh, S., and Mendieta, G.R., 1994. "Variable selection in regression models using principal components," Communications in Statistics- Theory and Methods, 23, 197-213.

Bumb B., 1986. A note on variables and observations in factor analysis: A Reply. Journal of Development Economics, 24: 197-200

Bumb, B., 1982a, Factor analysis and development: a note, Journal of Development Economics 11, no. 1, 109-112.

Bumb, B., 1982b, Factor analysis and development: a rejoinder. Journal of Development Economics 11, no. 1, 125-128

Butler D.G., Cullis B.R., Gilmour A.R., and Gogel B.J., 2009. ASReml- R reference manual Version 3.0. Queensland Department of Primary Industries and Fisheries.

Daetwyler H.D., Kemper K.E., Van der Werf J.H.J. and Hayes B.J., 2011. The importance of population structure on the accuracy of genomic prediction in a multi-breed sheep population. Proc. Assoc. Advmt. Breed. Genet. 19:327-330.

De Roos APW, Schrooten C, Mullaart E, Van der Beek S, De Jong G. and Voskamp W. 2009. Genomic selection at CRV. Proceedings Interbull Genomic Selection Workshop. Uppsala, Sweden

Dimauro C., Cellesi M., Pintus M.A., and Macciotta N.P.P., 2011. The impact of the rank of marker variance-covariance matrix in principal component evaluation for genomic selection application. J. Anim Breed Genet. 440-445.

Drineas P, Lewis J, Paschou P., 2010 Inferring Geographic Coordinates of Origin for Europeans Using Small Panels of Ancestry Informative Markers. PLoS ONE 5(8): e11892. doi:10.1371/journal.pone.0011892

Ducrocq V, Fritz S, Guillaume F, and Boichard D. 2009. French report on the use of genomic evaluation. Proceedings Interbull Genomic Selection Workshop. Uppsala, Sweden

Edwards A.W.F., 2003. Human genetic diversity: Lewontin's fallacy. BioEssays 25: 798–801.

Fernando, R. L., 1998. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, NSW, Australia, Vol. 26, pp. 329–336.

Habier D., Fernando R. L., Dekkers J. C. M., 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Hadi AS and Ling RL., 1998. Some cautionary notes on the use of principal component regression. The American Statistician, 52: 15-19

Hald, A.,1952. Statistical Theory with Engineering Applications, New York: Wiley.

Harris BL, Johnson DL and Spelman RJ., 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. 36th ICAR Session, Niagara Falls 16 – 20 June, pp 325 - 331

Hawkins, D.M., 1973. "On the Investigations of Alternative Regressions by Principal Component Analysis," Applied Statistics, 22, 275-286.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol., 24, 417–441, 498–520.

Hotelling, H., 1957. The relations of the newer multivariate statistical methods to factor analysis.B rit. J. Stat. Psychol., 10, 69-79.

Hwang J.T.G,, Nettleton D., 2002. Principal components regression with data-chosen components and related methods. Technometrics. 45(1): 70-79.

Jeffers, J.N.R., 1967. Two Case Studies in the Application of Principal Component Analysis, Applied Statistics, 16, 225-236.

Johanson, I., 1959. Progeny testing methods in Europe. Page 706 – 713. Proc. In American Dairy Science Associations Meeting, University of Illinois, Urbana.

Jolliffe, I.T., 1982. "A note on the use of principal components in regression". Applied Statistics, 31, 300-303.

Jombart, T., Pontier D., and Dufour A.B., 2009. Genetic markers in the playground of multivariate analysis. Heredity 102:330–341.

Kendall, M. G., 1957. A Course in Multivariate Analysis. London: Griffin.

Lewis J, Abas Z, Dadousis C, Lykidis D, Paschou P, Drineas P., 2011. Tracing Cattle Breeds with Principal Components Analysis Ancestry Informative SNPs. PLoS ONE, 6(4): e18007. doi:10.1371/journal.pone.0018007

Liu N. and Zhao H., 2006. A non-parametric approach to population structure inference using multilocus genotypes. Hum Genomics 2: 353–364.

Macciotta N., Gaspa G., Steri R., Nicolazzi E., Dimauro C., Pieramati C., and Cappio-Borlino A., 2010a. Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. J. Dairy Sci., 93, 2765–2774.

Macciotta N., Pintus M., Gaspa G., Nicolazzi E., Rossoni A., Vicario D., van Kaam J., Nardone A., Valentini A., and Marsan P., 2010b. Use of a principal component approach for estimating direct genomic breeding values for somatic cell score in dairy cattle. 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany, August 1-6, 2010.

McVean G., 2009. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet, 5(10): e1000686. doi:10.1371/journal.pgen.1000686

Menozzi P, Piazza A, Cavalli-Sforza L., 1978. Synthetic maps of human gene frequencies in Europeans. Science 201: 786–792.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. E., 2001. 'Prediction of total genetic value using genome-wide dense marker maps.', Genetics 157(4), 1819-1829.

Murrell, P., 1986. A note on variables and observations in factor analysis, Journal of Development Economics 21, no. 2, 319-325.

Niemann-Sorenson, A., and A. Robertson., 1961. The association between blood groups and several production characteristics in three Danish cattle breeds. Acta Agric. Scand. 11:163–196.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A., Indap A., King K.S., Nelson S.B.M.R, Stephens M. and Bustamante C.D., 2008. Genes mirror geography within Europe. Nature 456: 98–101

Novembre, J. and Stephens, M., 2008. Interpreting principal component analyses of spatial population genetic variation. Nature Genet. 40, 646–649

Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, Joshua D., Smith J.D., Ridker P.M., Chasman D.I., Krauss R.M., and Ziv E., 2008. Tracing sub-structure in the European American population with PCA informative markers. PLoS Genet 4: e1000114.

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney M.W., and Drineas P., 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 3: 1672–86.

Patterson N, Price A. and Reich D., 2006. Population structure and eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190

Pearson, K., 1901. "On lines and planes of closest fit to systems of points in space". Philosophical Magazine 2 (6): 559–572.

Pinto LFB, Packer IU, De Melo CMR, Ledur MC, Coutinho L.L. 2006. Principal component analysis applied to performance and carcass traits in the chicken. Anim Res, 55:419-425.

Pintus, M.A., G. Gaspa, E.L. Nicolazzi, D. Vicario, A. Rossoni, P. Ajmone-Marsan, A. Nardone, C. Dimauro, and N.P.P. Macciotta. 2012. Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach. J. Dairy Sci. 95:3390–3400.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, and Reich D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.

Reich D, Price AL, Patterson N., 2008. Principal component analysis of genetic data. Nat Genet, 40: 491–492

Schenkel FS, Sargolzaei M, Kistemaker G, Jansen GB, Sullivan P, Van Doormaal BJ, VanRaden PM, Wiggans GR. 2009a. Genomic Evaluation of Holstein Cattle in Canada Utilizing MACE Proofs. Proceedings of the 2009 ADSA-CSAS-ASAS Joint Annual Meeting, Montreal, Quebec, Canada

Schenkel FS, Sargolzaei M, Kistemaker G, Jansen GB, Sullivan P, Van Doormaal BJ, VanRaden PM and Wiggans GR. 2009b. Reliability of genomic evaluation of Holstein cattle in Canada. Proceedings Interbull Genomic Selection Workshop. Uppsala, Sweden 2009

Sellner E.M., Kim J.W., McClure M.C., Taylor K.H., Schnabel R.D. and Taylor T.F., 2007. Board-invited review: Applications of genomic information in livestock. J. Anim. Sci., 85, pp. 3148–3158.

Skillicorn D., 2007. Understanding complex datasets: Data mining using matrix decompositions. Boca Raton (Florida): CRC Press.

Solberg T., Sonesson A., Woolliams J., and Meuwissen T., 2009. Reducing dimensionality for prediction of genome- wide breeding values, Genet. Sel. Evol., 41:29

Van der Linde R, and Wilmink H. 2008. Status of genomic selection in the Netherlands. 36th ICAR Session, Niagara Falls 16 – 20 June

Van Doormaal BJ, Sargolzaei M, Kistemaker G, Sullivan P, Schenkel FS. 2009. Canadian Implementation of Genomic Evaluations. Interbull Meeting, Barcelona, Spain, August 21-24

VanRaden, P., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91:4414–4423.

Veerkamp R.F., M.P. Coffey, D.P. Berry , Y. de Haas, E. Strandberg, H. Bovenhuis, M.P.L. Calus, and E. Wall. 2012. Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy cows from experimental research herds in four European countries. Animal, (in press-accepted).

Vishwanath, R., 2003. 'Artificial insemination: the state of the art.', Theriogenology 59(2), 571-584.

Wiggans GR, Sonstegard TS, VanRaden PM, Matukumalli LK, Schnabel RD, Taylor JF, Chesnais JP, Schenkel FS and Van Tassel., 2008. Genomic selection in the United States and Canada: A collaboration. 36th ICAR Session, Niagara Falls 16 – 20 June, pp 347 – 355