Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis

MACCIOTTA N.P.P; GASPA G; STERI R; NICOLAZZI E.L; DIMAURO C; PIERAMATI C; CAPPIO-BORLINO A,

Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis,

JOURNAL OF DAIRY SCIENCE, 93: 2765:2774, 2010,

doi: 10.3168/jds.2009-3029

Interpretive Summary

Title: Using eigenvalues as variance priors in the prediction of Genomic breeding values by principal component analysis *By Macciotta et al.*

Principal component analysis with the use of eigenvalues as variance priors was effective in reducing the number of predictors up to 96% and saving computational resources for the prediction of individual genetic merit for a genome of 6 chromosomes and 6K SNP markers available. The same accuracy (0.76) was obtained when 279 principal components were used as predictors instead of 5,925 SNP markers. Moreover, one of the top principal components was able to depict the variation between individuals of different generations

PRINCIPAL COMPONENT ANALYSIS IN GENOMIC SELECTION

Using eigenvalues as variance priors in the prediction of Genomic breeding values by principal component analysis

N. P. P. Macciotta,[*1] G. Gaspa,[*] R. Steri,[*] E. L. Nicolazzi,[§] C. Dimauro,[*] C. Pieramati[†] and A. Cappio-Borlino[*]

[*]Dipartimento di Scienze Zootecniche, Università di Sassari, Sassari, Italy 07100

[§]Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza Italy 20100

[†]Centro di Studio del Cavallo Sportivo, Università di Perugia, Perugia, Italy 06100

[1]Corresponding author: Nicolò P.P. Macciotta, Dipartimento di Scienze Zootecniche, Università di Sassari, via De Nicola 9, 07100 Sassari, Italy. Phone number: 0039 079229298. Fax number: 0039 079229302. e-mail: macciott@uniss.it

ABSTRACT

Genome wide selection aims at predicting genetic merit of individuals by estimating the effect of chromosome segments on phenotypes using dense SNP marker maps. In the present paper, principal component analysis was used to reduce the number of predictors in the estimation of genomic breeding values for a simulated population. Principal component extraction was carried out either using all markers available or separately for each chromosome. Priors of predictor variance were based on their contribution to the total SNP correlation structure. The principal component approach yielded the same accuracy of predicted genomic breeding values obtained with the regression using SNP genotypes directly, with a reduction in the number of predictors of about 96% and computation time by 99%. Although these accuracies are lower than those currently achieved with Bayesian methods, at least for simulated data, the improved calculation speed together with the possibility of extracting principal components directly on individual chromosomes may represent an interesting option for predicting genomic breeding values in real data with a large number of SNPs. The use of phenotypes as dependent variable instead of conventional breeding values resulted in more reliable estimates, thus supporting the current strategies adopted in research programmes of genomic selection in livestock.

Key words: SNPs, genomic selection, principal component analysis, eigenvalues.

INTRODUCTION

Marker Assisted Selection (MAS) programs have had limited commercial applications till early 2000’s due to the fact that most of reported marker-QTL associations had been found within families but were in linkage equilibrium across the population (Dekkers, 2004; Hayes and Goddard, 2001; Khatkar et al., 2004). The availability of genome-wide dense marker maps for several animal species has recently allowed the prediction of genomic breeding values (GEBV) by estimating marker haplotype effects on phenotypes (Goddard and Hayes, 2007; Meuwissen et al., 2001). Genome wide selection relies on highly dense markers whose effects on phenotypes are estimated on a training population and then used to calculate GEBV both for training individuals and animals with only marker genotypes available (for example, young animals without phenotypes or estimated breeding values). A reduction in generation interval, an increase of accuracy in the cow side of the pedigree and a decrease of selection costs are the expected advantages of an efficient genome wide selection over traditional selection (Konig et al., 2009; Schaeffer, 2006).

High density SNP maps fulfill the basic requirement of genome wide selection, i.e. the analysis of genome bits having large and persisting population-wide linkage disequilibrium (Muir, 2007). However, the use of dense marker platforms results in a large number of effects to be estimated (many thousands) in comparison with the relatively small amount of phenotypes available (often just a few thousands). Such a data asymmetry raises several statistical issues, such as collinearity among predictors and multiple testing (Gianola and van Kaam, 2008). To cope with such a problem, several methods of reduction of the number of predictors without a large decrease in accuracy have been proposed.

Selection of relevant SNP by single marker regression on phenotypes may improve results in genome-wide association studies (Aulchenko et al., 2007; Long et al., 2007), but it leads to a decrease of GEBV accuracy (Meuwissen et al., 2001). Bayesian methods that select SNP by evaluating their individual contribution to the variance of the trait, such Bayes B method (Meuwissen et al., 2001; Fernando et al., 2007; VanRaden, 2008), usually give best GEBV accuracies when simulated data with few QTLs are modeled. However, results on actual data indicate that BLUP estimation, which assumes an equal contribution of all marker intervals to the genetic variance, performs only slightly worse than Bayesian methods in GEBV prediction (Hayes et al., 2009; VanRaden et al., 2009). Moreover in all the above mentioned techniques, markers are selected according to their relevance on the variability of the phenotype analyzed. Consequently, specific sets of markers may be required for different traits (Habier et al., 2009).

Multivariate dimension-reduction techniques may offer an alternative approach based on the evaluation of the contribution of each marker locus to the total SNP (co)variance structure. Principal component analysis (PCA) has been used for analyzing complex genetic patterns in human genetics (Cavalli Sforza and Feldman, 2003; Paschou et al., 2007) and for selecting markers in genome-wide association studies. Solberg et al. (2009) used principal component analysis and partial least squares regression (PLSR) to reduce the dimensionality of predictors in genomic selection. Both principal component (PC) and PLSR showed comparable accuracies with Bayes B when lower marker densities were fitted, whereas the gap between methods increased with the number of markers used. Solberg et al. (2009) concluded that reduction in computational complexity provided by multivariate methods did not counterbalance their lower accuracy compared to Bayes B. Such considerations are justified by the low cost of calculation time and by the computational speed that can be provided by optimized techniques such as parallel computing. On the other hand, it is reasonable to expect that denser SNP platforms will be very soon available for livestock species and dimensionality will again represent a relevant problem.

In their proposal, Solberg et al. (2009) regressed phenotypes on principal component scores extracted from the SNP matrix using the single value decomposition approach with an assumption of equal variance of each PC score. The choice of priors of marker effects represents a crucial point for genomic models (de Los Campos et al., 2009). On the other hand, the ordinary method for calculating PC relies on the eigenvalues of the correlation matrix of starting variables that measure the contribution of each PC to the original variance of predictors. Thus eigenvalues can be used as priors of predictor effect for the calculation of GEBV. It is worth remembering that eigenvalues have been already incorporated in mixed model algorithms to optimize calculations for variance component estimation (Dempster et al., 1984; Taylor et al., 1985).

In the present paper, principal component analysis is used to perform a BLUP prediction of GEBV in a simulated data set to test the ability of this technique to reduce the number of predictors without decreasing GEBV accuracy. Moreover, the feasibility of extracting PC from dense commercially available SNP platforms is tested.

MATERIALS AND METHODS

Data. The data set was generated for the XII QTLs – MAS workshop (http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html). The base population consisted of 100 individuals (50 males and 50 females). The genome had six chromosomes (total length 6 M), with 6,000 biallelic SNP, equally spaced at a distance of 0.1 cM. A total of 48 biallelic QTL were generated, with positions sampled from the genetic map of the mouse genome. QTL effects were sampled from a gamma distribution with parameters estimated by Hayes and Goddard (2002). Initial allelic frequencies of both SNP and QTL were set to 0.5. Then 50 generations of random mating followed. Generations 51 to 57 were used to create the experimental population of 5,865 individuals. Generations 51 to 54 (4,665 individuals, TRAIN data set) had pedigree, phenotype, and marker information available. For the last three generations (1,200 individuals, PRED data set) only pedigree and marker information were available. True breeding values (TBV) were considered as the sum of all QTL effects across the entire genome. Phenotypes were generated by adding environmental noise to the TBV. Further details on the simulation can be found in Lund et al. (2009).

Polygenic breeding values (EBV), being among the most frequently used dependent variable in GEBV prediction with real data, were also predicted. EBV, additive genetic **(**σ^2_a) and residual **(**σ^2_e) variance components were estimated with a single trait animal model that included the fixed effects of sex and generation, and the random additive genetic effect of the animal. The pedigree relationship matrix included 5,939 animals.

PCA analysis. Principal component analysis aims at synthesizing information contained in a set of n observed variables (M_1, ..., M_n) by seeking a new set of k (k<n) orthogonal variables (PC_1,…, PC_k) named principal components. PC are calculated from the eigen decomposition of the covariance (or correlation) matrix of M. The jth PC is a linear combination of the observed variables:

$$PCj = \alpha_{1j}M_1 + \ldots + \alpha_{nj}M_n$$

where coefficients α_{ij} are the elements of the eigenvector corresponding to jth eigenvalue. PC are usually extracted in a descending order of the corresponding eigenvalue that measures the quota of variance of original variables explained by each PC (Morrison, 1976; Krzanowsky, 2003).

A SNP data matrix **M** with m rows (m=5,865, the number of individuals in the entire data set) and n columns (n=5,925, the number of SNP markers that were found to be polymorphic) was created. Each element (i,j) corresponded to the genotype at the the jth marker for the ith individual. Genotypes were coded as -1, 0 or 1, according to the notation used by Solberg et al. (2009).

Data editing is usually recommended when handling dense marker maps (Wiggans et al., 2009), either to correct for data quality (i.e. genotyping not successfully performed) or to avoid possible estimation biases due to a severe unbalancement of genotypes. However, considering that in the present simulated data only 288 markers had minor allele frequency (MAF) <0.05, while 47 deviated significantly (P<0.01) from the Hardy-Weinberg equilibrium and this deviation may be attributable to drift, only the 75 monomorphic SNP were discarded from the analysis. Such a choice is, at least partially, supported by results of Chan et al (2008) that pointed out that SNP attributes commonly considered in SNP data editing, such as MAF or deviation from Hardy-Weinberg equilibrium, have actually a very small effect on overall false positive rate in genome-wide association studies.

PCA was carried out on **M** and the number of PC (k) retained for further analysis was based on both the sum of their eigenvalues and the obtained GEBV accuracy. PC extraction was performed either on all SNP simultaneously (PC_SNP_ALL) or separately for each chromosome (PC_SNP_CHROM). Scores of the *k* selected PC were calculated for all individuals. Marker haplotypes may be more efficient than genotypes in capturing marker-QTL association, especially in outbred populations where it may differ between families (Calus et al., 2008). Thus, PCA was performed also on haplotypes constructed from pairs of adjacent marker loci, either using all loci together (PC_HAP_ALL) or separately per chromosome (PC_HAP_CHROM).

Predictor effect estimation and GEBV calculations. Dependent variables used in the analysis were either phenotypes or polygenic EBV. For the estimation of the effects of predictors, records of the 4,665 individuals of the TRAIN data set were analysed with the following mixed linear model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$$

where **y** is the vector of either phenotypes or EBV, **X** is the design matrix of fixed effects (mean, sex=1,2; generation=1,2,3,4 for phenotypes; only mean for EBV), **b** is the vector of solutions for fixed effects, **Z** is the (m x k) design matrix of random effects, where each element corresponds to the score of the kth component for the mth animal of the training generations**,** **g** is the vector of solution for random regression coefficients of PC scores, **e** is the random residual. Covariance matrices of random PC effects (**G**) and residuals (**R**) were modeled as diagonal $\mathbf{I}(\sigma^2_{ai})$ and $\mathbf{I}(\sigma^2_e)$, respectively. BLUP methods used for estimating SNP effects usually assume an equal contribution of each SNP locus to the variance of the trait, sampled from the same normal distribution, i.e. $\sigma^2_{aj}=\sigma^2_a/n$ (Meuwissen et al.,2001; VanRaden et al., 2009). In the present work, two different options were compared. The first is the above mentioned equality of variances. The second starts from the consideration that PC scores were used as predictor variables and their contribution to the original SNP covariance structure is quantified by the corresponding eigenvalue (λ). Thus, variances of PC effects were calculated as $\sigma^2_{aj}=(\sigma^2_a/k) \times \lambda_j$.

G matrix diagonality, commonly implemented in BLUP methodologies for estimating SNP marker effects (Meuwissen et al., 2001; VanRaden, 2008), relies on the assumption that marker effects in a large population are uncorrelated (VanRaden et al., 2009). With the use of PC scores, such an assumption is consistent with the orthogonality between PC (Morrison, 1976). BLUP solutions were estimated using Henderson's normal equations (Henderson, 1985).

In order to have a comparison with the most straightforward estimation method, SNP effects were estimated directly by using the same mixed linear model but with **Z** indicating the design matrix of the 5,925 polymorphic SNP genotypes (coded as 0, 1 and 2, i.e. on the basis of the number of alleles). Covariance matrix **G** was assumed to be diagonal as **I(**σ^2_a/n). A Cholesky decomposition was used to solve mixed model equations (Harville, 1997).

Overall mean and effects of PC scores or SNP genotypes ($\hat{\mathbf{g}}$) estimated on the TRAIN data set were then used to predict GEBV both in TRAIN and PRED individuals. as

$$\mathbf{GEBV} = \mu + \mathbf{Z}\hat{\mathbf{g}}$$

where **GEBV** is the vector of predicted genomic breeding values and **Z** is the matrix of the PC scores or SNP genotypes of all individuals**.**

Accuracies of prediction where evaluated by calculating Pearson correlations between GEBV and TBV for the PRED generations. Bias of prediction was assessed by examining the regression coefficient of TBV on GEBV (Meuwissen et al., 2001). Goodness of prediction was evaluated also by the mean squared error of prediction (MSEP) calculated as

$$MSEP = \sum_{i=1}^{n} \frac{\left[TBV_i - GEBV_i\right]^2}{n}$$

where n is the number of individuals in the PRED generations, and by its partition in different sources of variation related to systematic and random errors of prediction (Tedeschi, 2006).

RESULTS

The pattern of eigenvalues of the correlation matrix of SNP genotypes obtained with PCA of all markers simultaneously is reported in Figure 1 (only the first 1,000 eigenvalues are plotted for brevity). A smooth decrease in the amount of variance explained by each successive PC can be observed, with a plateau between 250 and 300 PCs (about 84% of variance explained). A number of principal components between 200 and 300 could therefore be considered adequate for describing the original variance of the system.

GEBV accuracies for different numbers of retained PC (from 50 to 600) using all SNP simultaneously and eigenvalues as variance priors are reported in Figure 2. Accuracy for both training and prediction generations increases till a plateau, reached at about 250-300 PC. Increasing further the number of retained PC does not result in an increase of accuracy, probably due to the small amount of variance explained by each additional variable. Similar results were obtained by Solberg et al. (2009) that report best accuracies when 350 PC were extracted from 8,080 biallelic markers distributed on 10 chromosomes. However, Solberg et al. (2009) found a rather decreasing trend of the correlation between GEBV and TBV for larger numbers of PC. Based on the accuracy of GEBV prediction, 279 PCs (83% of the original variance) were retained in the present work for PC_SNP_ALL and PC_HAP_ALL approaches. In the analysis carried out on individual chromosomes, to keep the same number of predictors of the previous approach, 46 and 47 PC for chromosomes 1-3 and 4-6 were retained, respectively.

Average GEBV accuracies obtained using phenotypes are, for the three prediction generations, around 0.70 (Table 1) when an equal contribution of PC score on the variance of the trait is assumed, similar to those reported by Solberg et al. (2009). Accuracies increase by about 10% (to an average of 0.75) when eigenvalues are used in the diagonal of the $\mathbf{G}^{-1}$ matrix of mixed model equations. In general, results are of the same order as in previous literature reports for BLUP estimation on simulated (Fernando et al., 2007; Meuwissen et al., 2001; Meuwissen, 2009) and real data (Hayes et al., 2009; VanRaden et al., 2009). Correlations obtained when all SNP were used as predictors are equal to those obtained with PC with eigenvalues as priors. On the other hand, a remarkable difference in calculation speed between the two methods has been observed: about six hours for the SNP_ALL approach and 3 minutes for the principal components, using a computer with a dual core processor 2.33 GHz and 3.26 MB RAM. Slight differences can be observed between estimates of PC carried on all chromosomes or separately for each of them. Moreover, same results have been basically obtained when genotypes at single markers or haplotypes were used, in agreement with previous reports for high density markers (Calus et al., 2008; Hayes et al., 2007).

GEBV accuracies are larger when phenotypes instead of EBV are used as dependent variables (Table 1). This is particularly evident when all SNP are used as predictors (on average 0.75 vs 0.39). Also the drop of accuracy between TRAINING and PRED generations is more evident for EBV-based predictions (Figures 3 and 4). These findings are confirmed by values of regression coefficients of TBV on GEBV (Table 2). Moreover, *b* values for methods based on PC are similar to those reported by Solberg et al. (2009) when equal variances were assumed whereas they are closer to one (about 0.85) when eigenvalues are used as variance priors.

The decomposition of the mean squared error of prediction for some of the considered scenarios is reported in Table 3. MSEP is always smaller (about a half) when GEBV are calculated using phenotypes. Its partition highlights a great relevance of components related to the bias of prediction (i.e. mean bias, inequality of variances) in the approach that fits directly SNP genotypes (about 79%). Methods based on PC extraction are characterized by a prevalence (about 80%) of random terms, measured by the random error and by the incomplete covariation. The use of eigenvalues as variance priors results in the lowest MSEP and, compared to the other PC-based method, in a reduction of the slope bias and the highest relevance of random variation. These differences can be clearly seen from the plots of TBV versus GEBV for the PC_SNP_ALL approach using equal (Figure 5a) or eigenvalue-based (figure 5b) variance. The latter shows a regression slope closer to the equivalence line (y=x) and a smaller value for the intercept, that indicates a smaller systematic underestimation of TBV. The composition of MSEP becomes very similar across the different methods when EBV are used as dependent variables, with a reduced incidence of random components and a larger relevance of unequal variances compared to the phenotype-based estimates (Table 3). Actually, the comparison of plots of TBV versus GEBV estimated with the PC_SNP_ALL approach using phenotypes (Figure 5a) or EBV (Figure 5c), clearly shows a reduced range of variability and a higher underestimation (as evidenced by the larger value of the regression intercept) for EBV-based GEBV.

An interesting feature of principal component analysis is the possible technical interpretation of extracted variables. Figure 6 reports score averages for the first two PC that together explain about 5% of the original variance of the system, calculated for each generation. Averages of the second PC ranged gradually from negative values for the first three generations to positive for the last three generations. A possible explanation of the ability of the second PC to distinguish individuals of different generations can be found in its negative correlation with the average observed heterozygosity per animal (-0.26) that tends to decrease from older to younger generations (Figure 7).

DISCUSSION

Main objectives of the work are to assess the effect of reducing predictor dimensionality in genomic breeding value estimation using PCA and to test the effect of structuring the variance contribution of PC with their eigenvalues

PCA allows an efficient description of the correlation matrix of biallelic SNP with a markedly smaller number of new variables (4.7%) compared to the original dimension of the system. Such a huge decrease has a straightforward impact on the calculation speed of GEBV, with a reduction of more than 99% of computing time achieving the same accuracy of predicted GEBV using all SNP. Compared to other methods of reduction of predictors where SNP are selected based on their position along the chromosome (VanRaden et al., 2009) or their relevance with the trait considered (Hayes et al., 2009), the multivariate reduction approach limits the loss of information because each SNP is involved in the composition of each PC.

GEBV accuracies obtained in the present work agree with a previous report on the use of PCA to estimate genomic breeding values (Solberg et al., 2009) when an equal contribution of each principal component to the variance of phenotypes is assumed. This approach follows the common BLUP assumption of equality of variance of predictors, usually criticized for its inadequacy to fit the widely assessed distribution of QTL i.e,. many loci with a small effect and very few with large effect (Hayes and Goddard, 2001). However, when eigenvalues are used as prior of PC variance, accuracies increase by about 10%. These figures highlight the importance of an accurate modeling of the variance structure of random effects in GEBV estimation. Bayesian methods estimate variances of different chromosome segments combining information from prior distribution and data (Meuwissen et al., 2001). These methods usually give the best performance (accuracies >80%) when simulated data are fitted, whereas results obtained on real data seem to indicate a substantial equivalence with the BLUP approach (Hayes et al., 2009; VanRaden et al., 2009). A common explanation is that, in Bayes method, assumptions on prior distributions of parameters are more difficult to infer when real data are handled. The use of eigenvalues as variance priors rely only on data, i.e. the SNPs correlation structure, and does not require assumptions on prior distribution.

A potential drawback in the calculation of GEBV using PCA is represented by PC extraction. In the present work, about 40 minutes were needed to process a SNP data matrix of 5,865 rows and 5,925 columns. The commercially available SNP panel for cattle has 54K marker loci, although about 40K are retained on average after editing (Hayes et al., 2009). Such a marked increase of columns, usually not accompanied by a comparable increase of rows (i.e. phenotypic records), may lead to statistical and computational problems if PC are extracted treating all SNP simultaneously. However, results of the present study indicate that PC may be calculated separately for each chromosome, keeping the same GEBV accuracy. It should be remembered that the number of SNP per chromosome is not far from current dairy data (on average 1,200-1,300) (Hayes et al., 2009;

Van raden et al., 2009; Wiggans et al., 2009). Thus PCA carried out on individual chromosomes may be of great interest for real data, also considering the substantial biological orthogonality among chromosomes. The availability of denser marker maps (i.e. 500K SNP) will represent a challenge for the method, although the number of PC to be retained does not seem to increase linearly with the number of original variables. Missing genotypes is a potential problem for computation of PCA, which requires data in each cell. Although edits that are normally carried out on SNP data leave only a few missing cells per animal, they are spread across different markers and this may lead to a severe reduction in the number of records. Missing data can be reconstructed using appropriate algorithms as those described by Gengler et al. (2007) or others implemented in softwares of common use such as PHASE or PLINK.

Of particular interest is the difference in GEBV accuracy obtained when using phenotypes vs. polygenic EBV as dependent variable. Polygenic EBV are phenotypes corrected for additive relationships among animals based on pedigree information. On the other hand, in GEBV predictions the genetic similarity between animals is accounted for by the specific combination of marker genotypes possessed by each individual. Therefore, the use of EBV as dependent variable in GEBV prediction may be regarded as redundant in terms of exploitation of genetic relationships. This behavior is particularly evident for the regression using all SNP markers. In this form, the calculation of GEBVs is equivalent to the use of an animal model with the additive genetic effect structured by the genomic relationship matrix (Goddard, 2009). Such a double counting of genetic relationship resulted in a evident reduction of the variability of GEBV compared to true breeding values. From a statistical standpoint, EBV are model predicted values and may not be suitable as dependent variable in further analyses (Tedeschi, 2006). Results of the present study, although obtained on simulated data, may more accurately reflect the reality of genomic selection programmes in cattle. In previous studies, EBV were generally the dependent variable. This is because true breeding values are not available on real data and EBV estimated with a high accuracy (>0.90) may represent a sort of golden standard for cross validations. However, the tendency now seems to move toward the use of partially corrected phenotypes such as de-regressed proofs or Daughter Yield Deviations (VanRaden et al., 2009; Hayes et al., 2009).

Finally, an interesting side product of PCA used to reduce the dimensionality of predictors in genome wide selection is represented by the extraction of synthetic variables that can have a technical meaning. Researches in human and animal genetics have highlighted the role of PC as indicators of population genetic structure: for example, the top eigenvectors of the covariance matrix show often a geographic interpretation (Chessa et al., 2009; Price et al., 2006). Usually, the meaning of the ith PC in terms of relationship with the original variables is inferred from the structure of its eigenvector. In the present study, such an evaluation was not feasible, probably due to both the relatively small amount of variance explained by each PC and the large number of original variables considered (i.e. the 5,925 SNP). However, one of the top PC was able to reflect the genetic variation among generations, although the discrimination between individuals of different generations was rather fuzzy, as expected, given the small amount of variance explained. However, this last point deserves some additional consideration. An assessed criterion in choosing which PC to retain is to look at their eigenvalues. However, sometimes the PC associated with the largest eigenvalue does not have a defined meaning whereas successive PC characterized by smaller eigenvalues may contain more relevant or biological information (Jombart et al., 2009). In the case of the present work, a meaning of the second PC as indicator of genetic drift, which should be the only reason of variation of genotypic frequencies in the simulated generations (Lund et al., 2009) could be hypothesized.

ACKNOWLEDGMENTS

Research funded by the Italian Ministry of Agriculture, grant SELMOL. Authors thank organizers of the XII QTL-MAS workshop for providing simulated data. Discussion with P. Ajmone-Marsan is gratefully acknowledged. Helpful comments and suggestions from the two reviewers are also acknowledged.

REFERENCES

Aulchenko, Y. S. , D. J. de Koning and C. Haley. 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide predigree-based quantitative trait loci association analysis. Genetics 177:577-585.

Calus, M., T. H. E. Meuwissen, A. P. W. de Roos and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553-561.

Cavalli-Sforza, L. and M. W. Feldman. 2003. The application of molecular genetic approaches to the study of human evolution. Nat. Genet. 33: 266-275

Chan E. C. F., R. Hawken and A. Reverter. 2008. The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Anim. Genet. 40: 149-156

Chessa, B., F. Pereira, F. Arnaud, A. Amorim, F. Goyache et al. 2009. Revealing the history of sheep domestication using retrovirus. Science 324: 532

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci. 82(E. Suppl.):E313-E328.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375-385.

Dempster, A.P., C.M. Patel, M.R. Selwyn and A.J. Roth. 1984. Statistical and computation aspects of mixed model analysis. Appl. Stat. 33:203-214.

Fernando, R. L., D. Habier, C. Stricker, J. C. M. Dekkers and L. R. Totier. 2007. Genomic selection. Acta Agr. Scand. A-AN 57: 192-195

Gengler, N., P. Mayeres and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1: 21-28.

Gianola, D., and van Kaam J.B.C.H.M. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289-2303.

Goddard, M. E. and B. J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124: 323-330

Goddard, M. E. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245-257

Habier, D., R. L. Fernando and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182: 343-353

Harville, D. A. 1997. Matrix algebra from a statistician's perspective. Springer-Verlag, New York

Hayes, B. J. and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33: 209-229

Hayes, B. J., A. J. Chamberlain, H. M. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker assisted selection with single markers and markers haplotypes in cattle. Genet. Res. 89: 215-220

Hayes, B. J., P. J. Bowman, A. J. Chamberlain and M. E. Goddard. 2009. Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433

Henderson, C.R. 1985. Best Linear Unbiased Prediction using relationship matrices derived from selected base population. J. Dairy Sci. 68:443-448.

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

Khatkar, M. S., P. C. Thomson, I. Tammen and H. W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. Evol. 36: 163-190

Konig, S., H. Simianer and A. Willam. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci. 92: 382-391.

Krzanowsky, W. J. 2003. Principles of multivariate analysis. Oxford University Press Inc., New York.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel and S. Avendano. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. 124: 377-389.

Lund, M. S., G. Sahana, D. J. de Koning, G. Su and Ö. Carlborg. 2009. Comparison of analyses of QTLMAS XII common dataset. I: genomic selection. BMC proc. 3(suppl. 1): S1

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

Meuwissen, T. H. E. 2009. Accuracy of breeding values of ‘unrelated’ individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41: 35

Morrison, F. 1976. Multivariate statistical methods. McGraw-Hill, New York.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124: 342-355.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLos Genetics 3: 1672-1686.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weimblatt, N. A. Shadick and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. 38: 904-909.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218-223.

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

Taylor, J.F., B. Bean, C.E. Marshall and J.J. Sullivan. 1985. Genetic and environmental components of semen production traits of artificial insemination Holstein bulls. J. Dairy Sci.: 2703-2722.

Tedeschi, L. O. 2006. Assessment of adequacy of mathematical models. Agr. Syst. 89: 225–247.

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

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstengard, R. D. Schnabel, et al. 2009. Reliability of genomic predictions for north American Holstein bulls. J. Dairy Sci. 92: 4414-4423.

Wiggans, G. R., T. D. Sonstengard, P. M. VanRaden, L. K. Matukumalli, R.D. Schnabel, J.F. et al. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92: 3431-3436.

Table 1. Pearson correlations between predicted genomic breeding values and true breeding values, for different estimation methods, using either phenotypes or polygenic breeding values (EBV) for the PREDICTION generations and assuming either equal variance contribution for each PC or eigenvalues as variance priors.

Method	Phenotypes	EBV
SNP_ALL	0.76	0.41
	Equal variance	
PC_SNP_ALL	0.69	0.53
PC_SNP_CHROM	0.70	0.55
PC_HAP_ALL	0.68	0.54
PC_HAP_CHROM	0.71	0.56
	Eigenvalues	
PC_SNP_ALL	0.76	0.57
PC_SNP_CHROM	0.73	0.56
PC_HAP_ALL	0.75	0.56
PC_HAP_CHROM	0.73	0.55

(SNP_ALL = all 5,925 SNPs; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PC_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PC_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PC_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome)

Table 2. Regression coefficients ($b_{TBV,GEBV}$) of True breeding Value on Predicted Genomic Breeding Value (GEBV) for the different estimation methods using either phenotypes or polygenic breeding values (EBV) for the PREDICTION generations and assuming either equal variance contribution for each PC or eigenvalues as variance priors.

	Trait			
Method	Phenotypes		EBV	
	$b_{TBV,GEBV}$	s.e.	$b_{TBV,GEBV}$	s.e.
SNP_ALL	1.08	0.027	1.15	0.073
		Equal variance		
PC_SNP_ALL	0.63	0.019	1.08	0.049
PC_SNP_CHROM	0.67	0.019	1.13	0.048
PC_HAP_ALL	0.61	0.019	1.08	0.049
PC_HAP_CHROM	0.65	0.018	1.11	0.047
		Eigenvalues		
PC_SNP_ALL	0.88	0.021	1.33	0.055
PC_SNP_CHROM	0.84	0.022	1.28	0.055
PC_HAP_ALL	0.88	0.022	1.32	0.056
PC_HAP_CHROM	0.83	0.023	1.26	0.056

(SNP_ALL = all 5,925 SNPs; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PC_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PC_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PC_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome)

Table 3. Mean squared error of prediction (MSEP) decomposition (%) and coefficient of determination (r^2) for the PREDICTION generations in some scenarios using either phenotypes or polygenic breeding values (EBV) .

	Phenotype		
	SNP_ALL	PC_SNP_ALL1	PC_SNP_ALL 2
MSEP	1.55	1.48	1.02
Mean Bias (U_M)	72.2	53.5	56.9
Unequal variances (U_S)	6.9	0.6	1.9
Incomplete covariation (U_C)	21.9	45.9	41.2
Slope bias (U_R)	0.22	11.1	1.1
Random errors (U_D)	27.6	35.4	42.0
r^2	0.57	0.48	0.57
	EBV		
MSEP	2.96	2.88	2.72
Mean Bias (U_M)	72.0	75.1	74.6
Unequal variances (U_S)	13.9	8.9	11.9
Incomplete covariation (U_C)	14.1	16.0	13.5
Slope bias (U_R)	0.01	0.00	0.7
Random errors (U_D)	27.9	24.9	24.7
r^2	0.17	0.28	0.33

(SNP_ALL= all 5,925 SNPs; PC_SNP_ALL 1= principal components extracted from all SNP genotypes simultaneously and equal contribution of each SNP to the variance of the trait; PC_SNP_ALL 2 principal components extracted from all SNP genotypes simultaneously and contribution of each SNP to the variance of the trait proportional to the eigenvalue

Note that U_M+ U_S+ U_C= U_M+ U_R+ U_D=100%

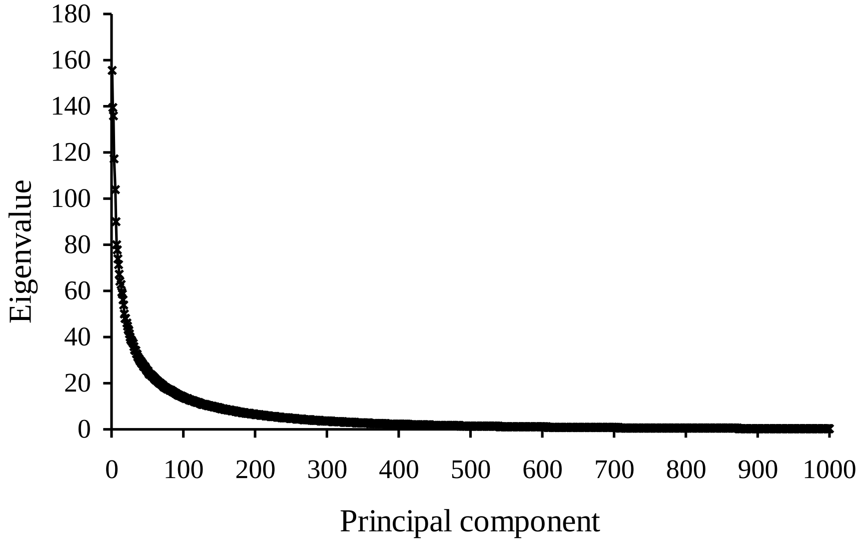

Figure 1. Pattern of the eigenvalues of the correlation matrix of SNP markers.

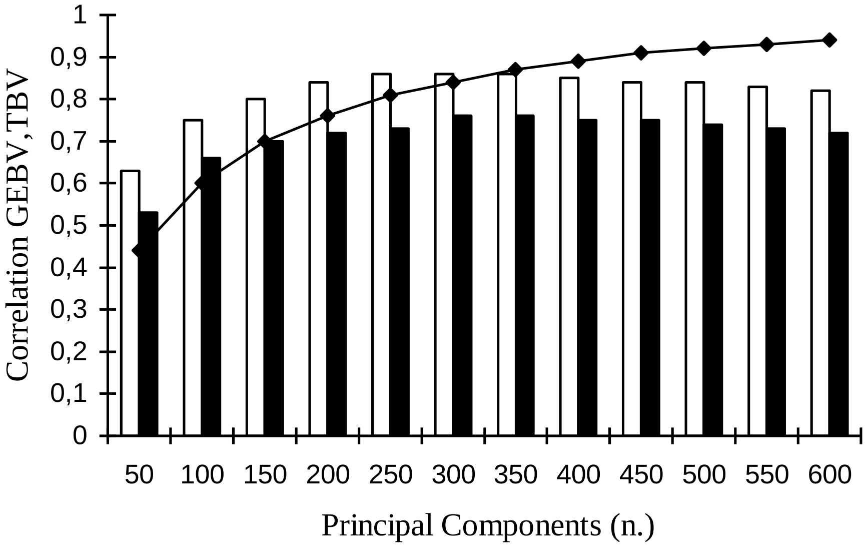

Figure 2. Pattern of correlations between genomic breeding values (GEBV) and true breeding values (TBV) when principal components are extracted from all SNP genotypes simultaneously and eigenvalues are used as priors, for different number of retained PC (white bars = training individuals, black bars = prediction individuals). The continuous line represents the amount of variance explained by the corresponding number of PC.

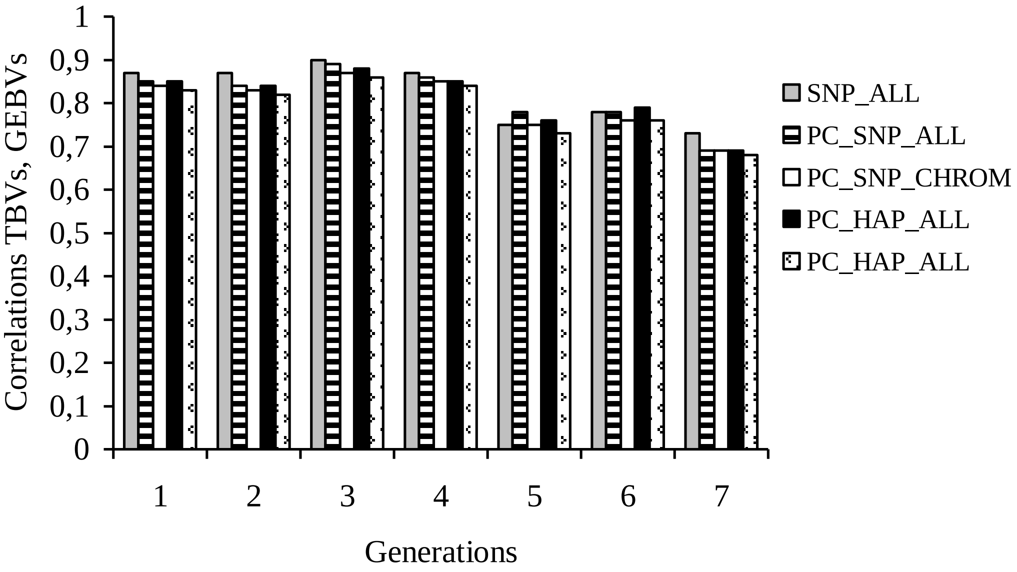

Figure 3. Correlations between genomic breeding values (GEBV) and true breeding values (TBV) in the different approaches when phenotypes were used as dependent variables (SNP_ALL = all 5,925 SNP; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PCA_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PCA_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PCA_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome).

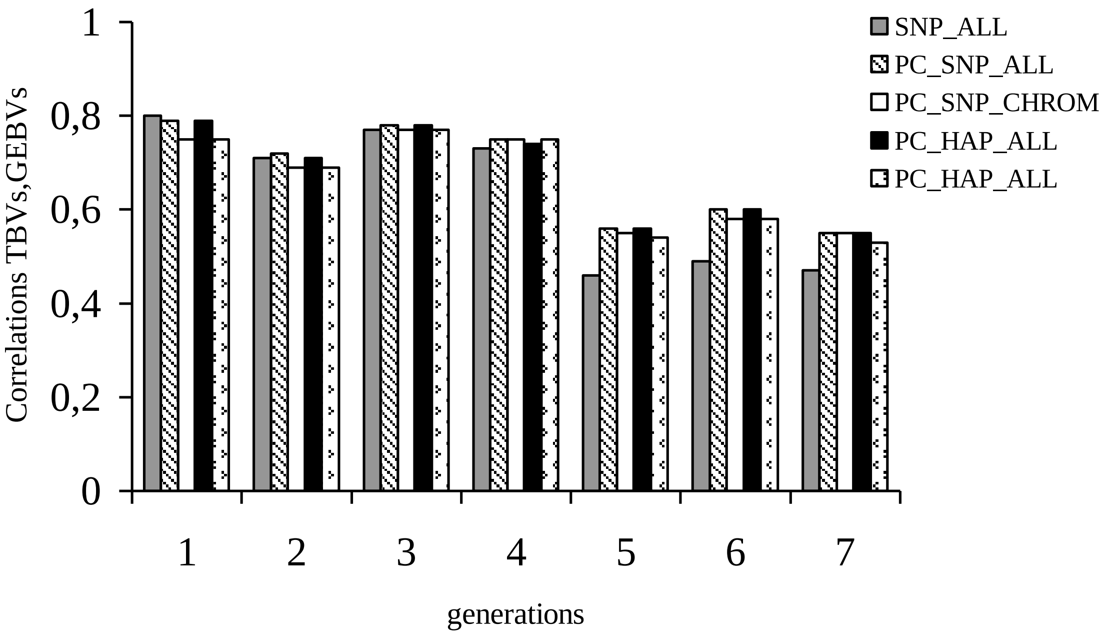

Figure 4. Correlations between genomic breeding values (GEBV) and true breeding values (TBV) in the different approaches when EBV were used as dependent variables (SNP_ALL = all 5,925 SNP; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PCA_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PCA_HAP_ALL = principal components extracted from all SNPS haplotypes simultaneously; PCA_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome).

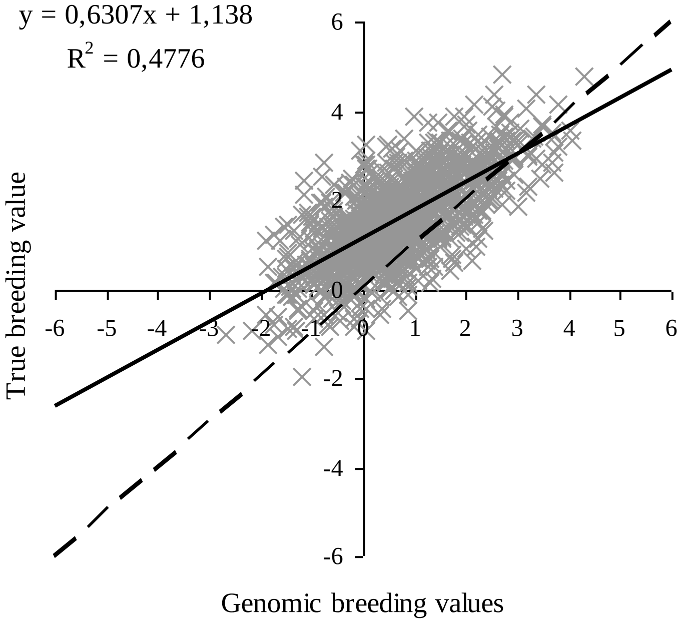

Figure 5a. Plot of true breding values versus genomic breeding values predicted using phenotypes when principal components are extracted from all SNP genotypes simultaneously and variance contribution of the PC scores in the estimation step is assumed equal (continuous line= regression line of TBV on GEBV; dotted line= equivalence line, y=x).

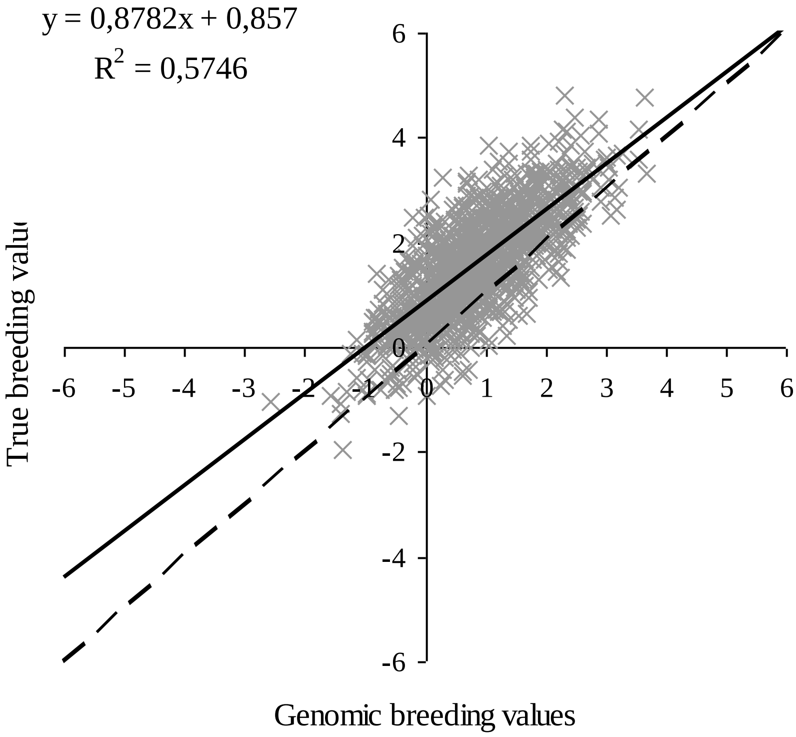

Figure 5b. Plot of true breeding values versus genomic breeding values predicted using phenotypes when principal components are extracted from all SNP genotypes simultaneously and variance contribution of the PC scores in the estimation step is based on their eigenvalues (continuous line= regression line of TBV on GEBV; dotted line= equivalence line, y=x).

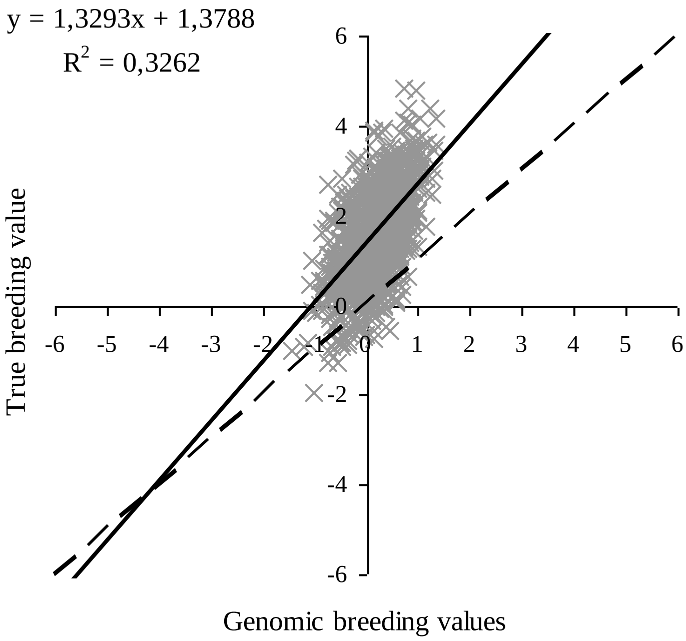

Figure 5c. Plot of true breeding values versus genomic breeding values predicted using phenotypes when all SNP genotypes are used as predictors (continuous line= regression line of TBVs on GEBVs; dotted line= equivalence line, y=x).

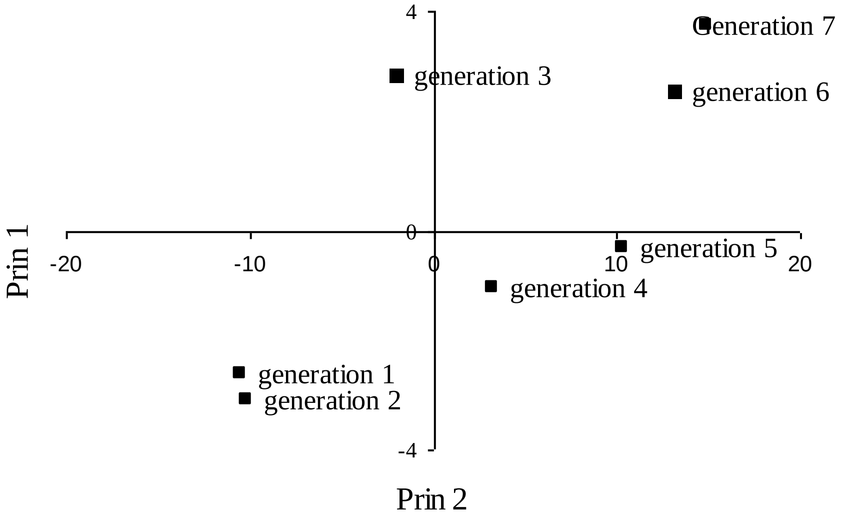

Figure 6. Plot of the average scores of the first two principal components for seven generations.

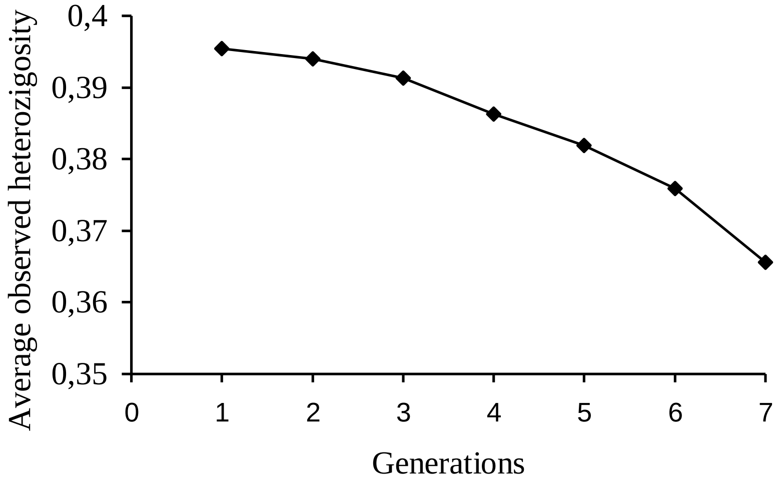

Figure 7. Pattern of the average observed heterozygosity in different generations.