Use of partial least squares regression to predict single nucleotide polymorphism marker genotypes when some animals are genotyped with a low-density panel

Published version:

DOI:10.1017/S1751731110002600

Use of partial least-squares regression to predict single-nucleotide polymorphism marker genotypes when some animals are genotyped with a low density panel

C. Dimauro[a], R. Steri, M. A. Pintus, G. Gaspa and N. P.P. Macciotta

Dipartimento di Scienze Zootecniche, Università di Sassari, via De Nicola 9, 07100 Sassari, Italy

[a]Corresponding author: Corrado Dimauro, Dipartimento di Scienze Zootecniche, Università di Sassari, via De Nicola 9, 07100 Sassari, Italy. Tel. +39 079229298; fax +39 079 229302. E-mail: dimauro@uniss.it

Running head: Single nucleotide polymorphism prediction

Abstract

High density SNP platforms are currently used in Genomic Selection (GS) programs to enhance the selection response. However, the genotyping of a large number of animals with high throughput platforms is rather expensive and may represent a constraint for a large-scale implementation of GS. The use of low density marker platforms could overcome this problem, but different SNP chips may be required for each trait and/or breed. In this paper a strategy of imputation independent from trait and breed, is proposed. A simulated population of 5,865 individuals with a genome of 6,000 SNP equally distributed on six chromosomes was considered. First, reference and prediction populations were generated by mimicking high and low density SNP platforms, respectively. Then, the partial least squares regression (PLSR) technique was applied to reconstruct the missing SNP in the low density chip. The proportion of SNP correctly reconstructed by the PLSR method ranged from 0.78 to 0.97 when 90% and 50% of genotypes were predicted, respectively. Moreover, data sets consisting of a mixture of actual and PLSR-predicted SNP or only actual SNP were used to predict genomic breeding values (GEBV). Correlations between GEBV and true breeding values varied from 0.74 to 0.76 respectively. Results of the study indicate that the PLSR technique can be considered a reliable computational strategy for predicting SNP genotypes in a low density marker platform with reasonable accuracies.

Keywords: genomic selection; SNP prediction

Implications

In genomic selection programs, animals are genotyped with high-density SNP marker platforms with around 50-60K markers. However, being the number of phenotypes available markedly lower than the number of markers, several statistical shortcomings arise when data are analyzed. In this paper we propose the use of both high and low-density SNP marker platforms in combination with partial least squares regression (PLSR) technique to reconstruct the missing SNP in the low density chips. Savings obtained by using low density platforms could be used to enlarge the number of animals involved in the selection program.

Introduction

Traditional genetic evaluations for livestock combine phenotypic data with pedigree relationships to estimate the probability that genes are transferred to the next generations. Genomic selection (GS), on the contrary, exploits dense marker information represented by single nucleotide polymorphism (SNP) to evaluate genomic breeding values (GEBV) by estimating the effect of chromosome segments on phenotypes (Hayes and Goddard, 2008). Advances in high throughput technologies have led to the construction of dense SNP platforms that could trace the inheritance of individual genes. High density marker (HDM) platforms with 50 – 60 K SNP are currently used in GS programmes. However, the number of genotyped animals is considerably smaller than the number of markers. In dairy cattle, the ratio number of animals vs. number of markers is, on average, between 0.08-0.15, apart from USA and Canada where it is around 0.45 (VanRaden *et al.*, 2009). Such a data asymmetry results in several statistical shortcomings, as collinearity among predictors and issues in multiple testing procedures. Furthermore, the well known curse of multi-dimensionality should become now more relevant, due to the recent commercial availability of the 777 K SNP Illumina Bead-chip.

The use of low density marker platforms (LDM) may represent a interesting technical option to reduce the genotyping costs and enlarge the number of animals involved in GS programmes. However, the reduction of SNP density is expected to decrease GEBV accuracy. Weigel *et al*. (2009) reported a loss of about one-third in the gain of reliability of GEBV for lifetime profit in cattle when a low-density assays with 750-1,000 SNP was used. In this study, SNP were chosen either on the basis of their chromosomal location (evenly spaced) or for their relevance on the considered trait. Habier *et al*. (2007) combined the use of evenly spaced SNP and co-segregation information from LDM to track HDM inheritance within families. On simulated data, they found a reduction in GEBV accuracy ranging from 1 to about 25%, depending on the considered scenario.

The use of the above mentioned methodologies can be useful to reduce the number of SNP but, separate chips for each trait and/or breed may be required. In this paper an alternative strategy, independent from trait or breed, is proposed. The method starts by creating a reference (REF) and a prediction (PRED) population of animals genotyped with HDM (containing *N* SNP) and LDM (*n* SNP) platforms, respectively (*N* > *n*). Missing *k*-markers (*k* = *N*-*n*) in PRED population are reconstructed by using a suitable mathematical tool and, as a final result, a PRED population with *N* SNP as in HDM is obtained. These markers are a mixture of actual and predicted SNP.

The most straightforward computational method for predicting unknown SNP markers in the LDM platform is the multivariate multiple regression. However, considering that adjacent SNP are highly correlated, the predictive capability of the model could be compromised by the multicollinearity among predictors (Draper and Smith, 1981). Partial least squares regression (PLSR), originally developed in the computational chemistry context (Hoeskuldsson, 1988), has become an established tool for modeling linear relations between multivariate measurements. It is characterized by an higher prediction efficiency compared to ordinary multivariate regression or principal component regression (Macciotta *et al*., 2006). PLSR has been already used in GS studies by Solberg *et al*. (2009) for reducing the dimensionality of predictors in the calculation of GEBV. In the present study, the PLSR technique is applied to predict missing SNP when animals are genotyped with a LDM platform. Actually, this statistical technique is particularly useful when a set of correlated dependent variables (**Y**) have to be predicted from a set of correlated independent variables (**X**). PLSR maximizes the correlation structures between **Y** and **X** and overcomes the multicollinearity problems by combining features of principal components analysis and multiple regression (Abdi, 2003).

The aim of this work is to test the ability of PLSR for predicting missing SNP genotypes when a PRED population is created by using a LDM platform of SNP markers.

Materials and methods

The data

Data were extracted from an archive generated for the XII QTLs – MAS workshop, freely available at: http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html. The base population consisted of 100 individuals (50 males and 50 females). A genome of six chromosomes (total length 6 M) with 6,000 biallelic SNP, equally spaced in the genome at a distance of 0.1 cM, was generated. A total of 48 biallelic QTLs were included, with positions sampled from the genetic map of the mouse genome and effects derived from a gamma distribution (Hayes and Goddard, 2001). Initial allelic frequencies of both SNP and QTL were set to 0.5. Then 50 generations of random mating followed. Generations from 51 to 57 were used to create the definitive archive of 5,865 individuals. For each generation 15 males and 150 females were randomly selected to be parents of the next generation. Each male had 100 sons and was mated to 10 females (10 sons for female). Animals belonging to the generations from 51 to 54 had pedigree, phenotype, and marker information available. For the last 3 generations only pedigree and marker information were available. These animals constituted the PRED population and were obtained by randomly selecting 400 animals for each generation (a total of 1200 individuals). True breeding values (TBV) were created as the sum of all QTL effects across the entire genome. Phenotypes were generated by adding to the TBV an environmental noise drawn from a normal distribution with mean zero and variance equal to the residual variance defined to obtain a heritability of 0.30. For further details on the data generation see Lund *et al*. (2009).

The PLSR technique

PLSR is a multivariate extension of the multiple regression analysis. It is particularly useful when (i) the number of predictor variables is similar to or higher than the number of observations and/or (ii) predictors are highly correlated (i.e. there is strong collinearity). The basic model is:

$$\mathbf{Y}=\mathbf{XB}+\mathbf{E}$$

where **Y** is a $n \times m$ response matrix, **X** is a $n \times p$ design matrix, **B** is a $n \times m$ regression coefficient matrix, and **E** is a $n \times m$ error term. In PLSR, matrices **X** and **Y** are simultaneously decomposed into a set of new variables (called latent factors). Factors are extracted in order to explain as much as possible of the covariance between **X** and **Y** and to minimize the covariance between variables inside each matrix. Extracted latent factors account for successively lower proportions of original variance and are defined as linear combinations of predictor and response variables (Hubert and Branden, 2003). Key elements in the different calculation steps of the PLSR are: the scores, i.e. values of the extracted latent factors both for the dependent (**U**) and independent variables (**T**), and factor loadings (**Q**) expressing correlations between extracted factors and original dependent variables. Considering a REF and a PRED population, latent factor scores ($\mathbf{T}_{\mathbf{ref}}$**)** extracted from $\mathbf{X}_{\mathbf{ref}}$, are used to predict scores of latent factors extracted from $\mathbf{Y}_{\mathbf{ref}}$ ($\mathbf{U}_{\mathbf{ref}}$**)**

$$\mathbf{U}_{\mathbf{ref}} = \mathbf{BT}_{\mathbf{ref}} \qquad (1)$$

Then, the estimated regression coefficients **B** are used to predict values of $\mathbf{Y}_{\mathbf{pred}}$ in the PRED population as:

$$\hat{\mathbf{Y}}_{\mathbf{pred}} = \mathbf{BT}_{\mathbf{pred}}\mathbf{Q'}_{\mathbf{ref}} \qquad (2)$$

where $\mathbf{Q'_{ref}}$ is the transposed matrix of factor loadings extracted from $\mathbf{Y_{ref}}$.

The standard algorithms for computing latent factors are nonlinear and iterative (NIPALS and SIMPLS algorithms, for example) and require the use of dedicated software (for more details see Wold *et al*., 2001; de Jong, 1993). In this work, the PLS procedure of SAS-STAT software (SAS Institute INC, Cary, NC) was used.

The PLSR method for SNP genotypes prediction

To simulate a PRED population genotyped with a LDM platform, the first *k*-SNP were assumed to be not known. SNP from *k*+1 to 1,000 represented the predictors (i.e. $\mathbf{X_{ref}}$ and $\mathbf{X_{pred}}$) and were known both for REF and PRED population. SNP from 1 to *k* were known in REF ($\mathbf{Y_{ref}}$) and were used to calculate the matrix of regression coefficients **B** (equation 1). Then, using the equation (2), the $\mathbf{\hat{Y}_{pred}}$ matrix was predicted. Being that the genotype at each SNP is coded as the number of allele 1 copies, i.e. 0, 1 or 2, results (columns in $\mathbf{\hat{Y}_{pred}}$ each containing the predicted SNP genotype) were rounded to the nearest integer. The goodness of SNP prediction was evaluated by calculating correlations between real ($\mathbf{Y_{pred}}$) and PLSR predicted ($\mathbf{\hat{Y}_{pred}}$) SNP genotypes. Considering that for *k* predicted SNP *k* correlations were calculated, the average value of these correlations, for each prediction scenario, was considered. Moreover, percentage of correct predictions across SNP and mean percentage of corrected SNP predictions for each animal were calculated.

A crucial point in PLSR modeling is how many latent factors should be retained to correctly define the complexity of one experiment. When several and correlated predictors are used, the risk of obtaining a model able to fit data well but with a very poor predictive power is rather high. This problem is known as model “over-fitting”. It is usually handled by testing the predictive significance of the successive extracted factors. Cross-validation in combination with PRESS statistics is commonly used to this purpose (Wold *et al.*, 2001). However, in the present study several scenarios involving a great number of predictors are compared and, therefore, the use of the above cited tests become problematic in terms of computation time and resources. For these reasons, the best number of extracted latent factors in each scenario was fixed empirically by comparing the obtained results with real data (the procedure will be explained in the next section).

Setup of the PLSR method

Location of missing SNP along the chromosome, number of latent factors to be extracted for each scenario, number of SNP to be predicted and the minimum number of genotyped animals to use as REF population are relevant aspects for the method be efficiently performed in practice. They were tested in successive steps during the development of the PLSR method. All the computations were done separately per chromosome .

Step 1: four scenarios of chromosome location of SNP to be predicted (k =100) in PRED population were tested: at the beginning (SNP1 – SNP100), in the middle (SNP451-SNP550), at the end (SNP901 – SNP1,000), or evenly spaced in the chromosome.

Step 2: once the best SNP location was assessed, the optimum number of latent factors to be extracted was evaluated. In PLSR procedure, the number of factors can not exceed the number of the independent variables. Therefore, for each chromosome, several simulations were performed where 100 SNP were predicted with a number of factors ranging from 10 to 900.

Step 3: prediction accuracy for different number of SNP to be predicted was investigated using the following proportions for missing SNP in PRED population: 10%, 25%, 50%, 75% and 90%. At the end of the PLSR procedure, a series of new data sets for PRED population, each containing 10%, 25%, 50%, 75% and 90% of PLSR predicted SNP, were produced.

Step 4: the effect of the SNP reduction in the estimation of genomic breeding values was tested by evaluating GEBV's either in original and in five data sets, generated in step 3, which contain the mixture of actual and PLSR predicted SNP. Effects of SNP markers on phenotypes in the REF population were estimated with a mixed linear model that included the fixed effects of mean, sex (1,2) and generation (1,2,3,4), and the random effects of SNP genotypes (Meuwissen *et al*. 2001). Overall mean and effects of SNP genotypes were then used to predict GEBV in PRED population (Macciotta *et al*., 2010). Accuracies were evaluated by calculating Pearson correlations between GEBV and true breeding values.

Step 5: finally, considering a possible application of the method on real data, accuracy of the PLSR predictions were tested for different sizes of the REF population, from 5,000 to 600 individuals. In all the simulations, the size of PRED population was kept constant (600).

Results and discussion

Step 1: the effects of SNP location on prediction accuracy can be observed in Table 1 where average correlations between actual and PLSR-predicted SNP genotypes for different scenarios are reported. Lowest correlations were obtained when markers to be predicted are located at the beginning or at the end of the chromosome. A slight increase of accuracy can be observed when SNP are located in the middle of the chromosome. The highest value was found for evenly spaced missing SNP. These results were expected, considering the decaying pattern of correlation between loci for increasing distances, and are in agreement with figures reported by Habier *et al*. (2009) who had already used evenly spaced SNP to simulate low density marker panels. In any case, the value of the mean correlation for the best scenario is notably high and may represent a useful indication for constructing a LDM platform without trait or breed constraints.

Step 2 : Figure 1 displays pattern of mean correlations between 100 actual and PLSR predicted SNP for increasing number of extracted latent factors for the first chromosome. There is a rapid increase of prediction accuracy from 10 up to 100 factors (from 47% to 93%). A plateau of 98% is then reached when about 150 - 200 factors are extracted. These results indicate that the number of latent factors to be extracted should be higher or, at least, equal to the number of predicted SNP.

Step 3: the variation of prediction accuracy for different number of SNP to be predicted is reported in Table 2. Moving from 10% to 75% missing SNP, there is small decrease (about 6%) in the average correlation between actual and predicted genotypes. In any case, prediction accuracy is higher than 90% even when two-third of the SNP are predicted. It slightly falls below 0.80 when 90% of SNP have to be predicted. However, even in this case, the accuracy can be considered satisfactory. If confirmed on real data, results of the present study may indicate that a chip with 5.4 K SNP evenly spaced across the genome could represent a suitable base for reconstructing, with a reasonable accuracy, the profile of an high density platform of 54 K SNP (i.e. the one currently used for cattle). In a recent study carried out with the bovine 54 K SNP, Weigel *et al*. (2010) using the algorithm implemented in fastPHASE 1.2 software (University of Washington TechTransfer Digital Ventures Program, Seattle, WA), reported a proportion of correctly reconstructed missing SNP of about 0.88 when 90% SNP were predicted. Druet and Georges (2010) combined fastPHASE and Beagle (Browning and Browning, 2007) algorithms to take into account both population (linkage disequilibrium) and familial (Mendelian segregation and linkage) information to predict missing genotypes. They found, with 50% missing genotypes, an imputation error of 3% and 1% for sparse and dense marker map, respectively. In the present work, the proportion of correctly reconstructed SNP for 90% and 50% missing genotypes was 0.86 and 0.98, respectively (Table 2).

The SNP genotype profile of each animal was also well reconstructed by the PLSR method. When 90% SNP were predicted, more than 84% of animals presented a percentage of corrected SNP reconstruction ranging from 80 to 100%. Moreover, when predicted SNP were lower then 75%, all animals had a proportion of corrected reconstructed SNP ranging from 95 to 100%.

Step 4: accuracies displayed in Table 3 indicate that the use of PLSR-predicted SNP does not affect the estimation of genomic breeding values. Correlations between true breeding values and GEBV remain basically the same moving from the scenario where all used SNP are actual to the one where 90% of marker genotypes are PLSR-predicted (Table 3). These results are similar to those obtained by Habier *et al*. (2009) who reported a reduction in GEBV accuracy of about 4% moving from a SNP panel density of 0.05 cM to 10cM.

Step 5: finally, Figure 2 displays accuracies of SNP prediction obtained with different sizes of REF population. As the number of fully genotyped animals becomes smaller, correlations between actual and predicted SNP slowly decrease reaching a value of 93% when the number of REF animals is twice (2,000) the total number of SNP per chromosome. Correlations dramatically drop (<70%) for a number of fully genotyped animals equal to 600. Considering that on real data each bovine chromosome has on average 1000-1200 SNP after data editing, a minimum number of 2,000-2,500 fully genotyped animals could be enough to obtain reliable predictions from the PLSR method.

Conclusions

The use of LDM platforms in combination with a suitable computational algorithm able to predict the missing genotypes with respect to HDM chips is an option for reducing genotyping costs in GS programs. Savings could be used to enlarge the genotyped population thus enhancing the efficiency of the breeding scheme. In this paper, the ability of PLSR technique for predicting missing SNP genotypes in LDM platforms was tested. The method correctly assigned from 86 to 98% of missing genotypes, when 90 and 50% SNP were predicted, respectively. Moreover, only a slight difference (2%) in GEBV accuracies was observed using actual SNP or a mixture of actual and predicted SNP. Finally, a size of around 2,000-2,500 fully genotyped animals with a 54 K SNP chip was found to be a reliable REF population to reconstruct the SNP profile of a PRED population of animals genotyped with a LDM chip containing 5,4 K evenly spaced SNP.

Acknowledgments

Research was funded by the Italian Ministry of Agriculture (Rome, Italy), grant SELMOL.

References

Abdi H 2003. Partial least squares (PLS) regression. In Encyclopaedia of social sciences research methods (eds M Lewis–Beck, A Bryman and T Futing) pp. 1-7. Sage Publication, Thousand Oaks, CA.

Browning SR and Browning BL 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81, 1084-1097.

De Jong S 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18, 251–263.

Draper NR and Smith H 1981. Applied regression analysis. John Wiley and Sons, New York.

Druet T and Georges M 2010. Hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184, 789-798.

Habier D, Fernando RL and Dekkers JCM 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177, 2389-2397.

Habier D, Fernando RL and Dekkers JCM 2009. Genomic selection using low-density marker panels. Genetics 182, 343-353.

Hayes BJ and Goddard M E 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution 33, 209-229.

Hayes BJ and Goddard ME 2008. Technical note: prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science 86, 2089-2092.

Hoeskuldsson A 1988. Partial least squares PLS methods. Journal of Chemometrics 88, 211-228.

Hubert M and Branden KV 2003. Robust methods for partial least squares regression. Journal of Chemometrics 17, 537-549.

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

Macciotta NPP, Dimauro C, Bacciu N, Fresi P and Cappio-Borlino A 2006. Use of a partial least-squares regression model to predict test day of milk, fat and protein yields in dairy goats. Animal Science 82, 463-468.

Macciotta NPP, Gaspa G, Steri R, Nicolazzi E, Dimauro C, Pieramati C and Cappio-Borlino A 2010. Use of principal component analysis to reduce the number of predictor variables in the estimation of Genomic Breeding Values. Journal of Dairy Science 93, 2765-2774.

Meuwissen THE, Hayes BJ and Goddard ME 2001. Prediction of total genetic values using genome-wide dense marker maps. Genetics 157, 1819-1829.

Solberg TR, Sonesson AK, Woolliams J and Meuwissen THE 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genetics Selection Evolution 41, 29.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstengard TS, Schnabel RD *et al* 2009. Reliability of genomic predictions for north American Holstein bulls. Journal of Dairy Science 92, 4414-4423.

Weigel KA, De Los Campos G, González-Recio O, Naya H, Wu L, Long N, Rosa GJ and Gianola, D 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science 92, 5248-5257.

Weigel KA, Van Tassell CP, O'Connell JR, VanRaden PM and Wiggans GR 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science 93, 2229-2238.

Wold S, Michael Sjöström M, Eriksson L 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58,109–130.

Table 1 *Mean correlations (and related standard deviations) between 100 actual and predicted SNP in each chromosome*

Missing SNP position	Correlations	
	Mean	St. Dev.
First 100	0.57	0.17
Middle 100	0.75	0.11
Last 100	0.68	0.14
One every 10	0.93	0.09

Table 2 *Mean correlations (and related standard deviations) between actual and predicted SNP for increasing percentage of predicted SNP. Proportions of correct SNP prediction are also reported*

Percentage of predicted SNP	Correlations		Proportion of correct SNP prediction
	Mean	St. Dev.	
10%	0.98	0.07	0.99
25%	0.98	0.07	0.99
50%	0.97	0.08	0.98
75%	0.92	0.08	0.95
90%	0.78	0.13	0.86

Table 3 *GEBV accuracies for different ratio of available/predicted SNP.*

Real SNP	Predicted SNP	GEBV accuracy
100%	0%	0.76
75%	25%	0.76
50%	50%	0.76
25%	75%	0.75
10%	90%	0.74

Figure captions:

Figure 1 Pattern of the mean correlations between actual and predicted SNP for increasing number of extracted factors during the PLSR procedure

Figure 2 Mean correlations between actual and predicted SNP for different numbers of fully genotyped animals