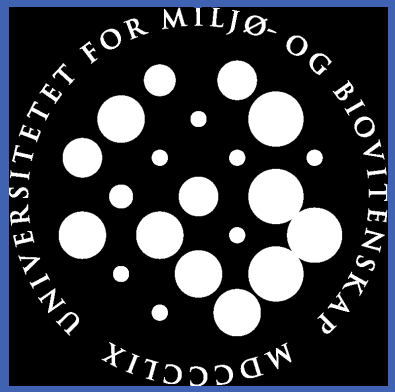
Using partial least square regression (PLSR) and principal component regression (PCR) in prediction of genome wide breeding values

T. R. Solberg[1], A.K. Sonesson[2], J. A. Woolliams[1,3], T. H. E. Meuwissen[1]

[1]University of Life Sciences, Dept. of Animal and Aquacultural Sciences, P.O. Box 5003, N-1430 Aas, Norway.

[2]AKVAFORSK (Institute of Aquaculture Research Ltd.), P. O. Box 5010, N-1432 Aas, Norway.

[3]Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK.



Introduction

- Dense molecular data available
  - Genomic selection (e.g. Meuwissen *et al.*, 2001; Gianola *et al.*, 2006; Schaeffer, 2006; Solberg *et al.*, 2006)
- SNP data is a typical case where the number of predictors can be larger than the number of records
- Partial Least Square Regression (PLSR) and Principal Component Regression (PCR) are methods designed for such situations
- Test the hypothesis that genome wide evaluation can be obtained using models of reduced dimensionality.
- Compared to the 'bayesB' method (Solberg *et al.*, 2006).

Introduction (cont.)

- PLSR and PCR mostly used within econometrics, chemometrics and social sciences (e.g. Wold, 1981; Wold, 1985; Martens & Næs,1991)

- Basic idea to reduce the number of predictors (principal components, PC)

Materials and methods

Population structure/genome

- Stochastic simulation study, Ne=100. Random mating for 1000 generations

Increased to 1000 animals in t = 1001 and t = 1002

Phenotypic values simulated by $P_i = TBV_i + \varepsilon_i$ ($\varepsilon \sim N(0,\sigma^2)$)

Animals genotyped in generation t = 1001 and t = 1002

BVs predicted on animals in generation t = 1002, using marker information in generation t = 1001.

M & M (cont.)

- Genome simulated with 10 chromosomes, each with a length of 100 cM
- 4 different marker densities evaluated, 1 cM, 0.5 cM, 0.25 cM and 0.125 cM spacing between the markers

(1010 mrk ⟹ 8080 mrk)

M & M (cont.)

- Marker genotype data organised as *m* x *p* matrix (**X**)

  (1000 x 1010 ⟹ 1000 x 8080)

- **y** vector and **X** matrix centered by subtracting their mean and scaled by deviding by their SD
- PCR create the PC by linear combinations among predictors
- PLSR create the PC by maximising the covariance between the predictors and the response

PLSR

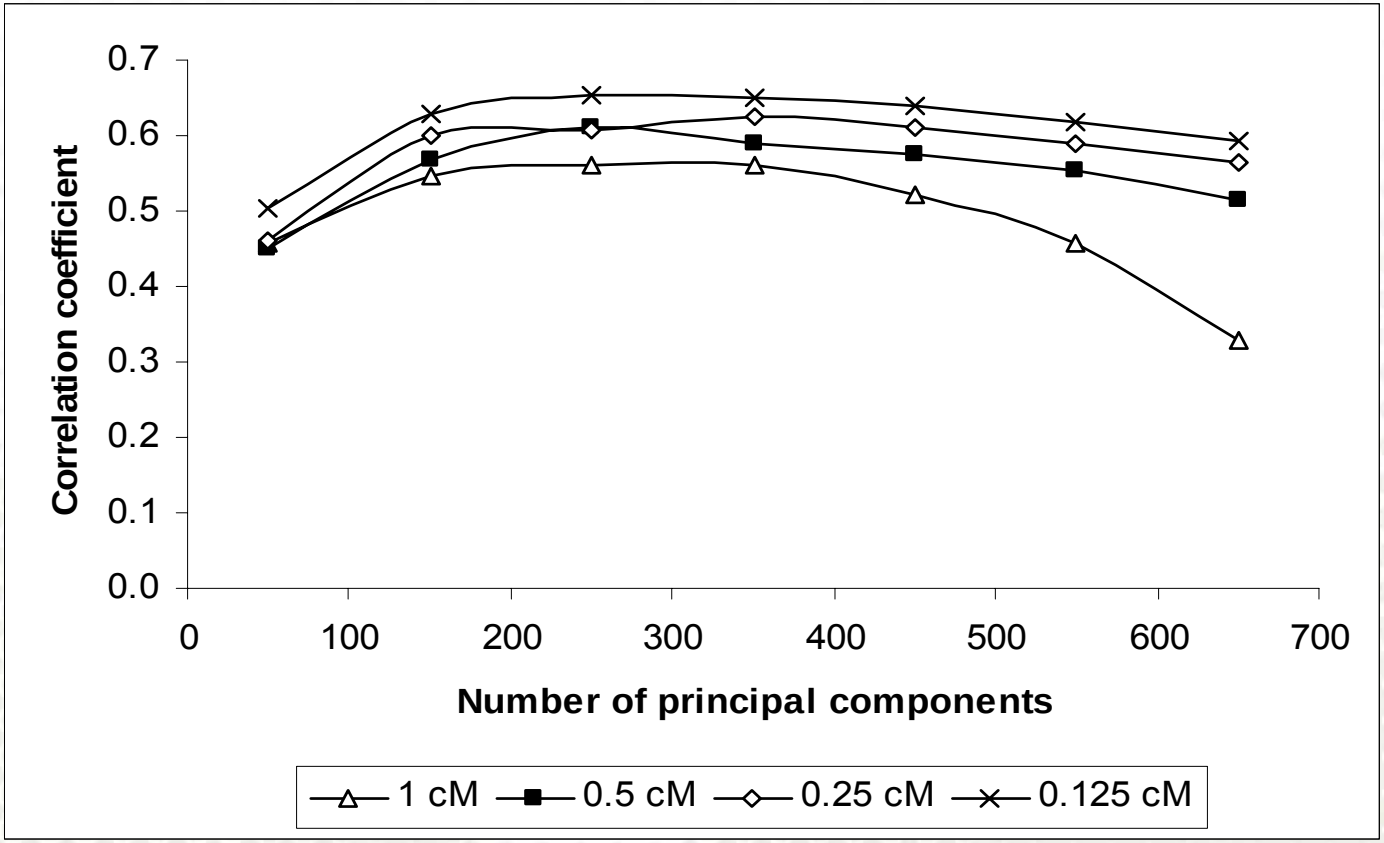
- SIMPLS algorithm performed (de Jong, 1993)
  - Compute dominant eigenvectors of the matrix **X** (**X'y**, since the data is univariate)
  - Deflating **X** matrix and **y** vector to compute the next PC

PCR

- Singular value decomposition of the **X** matrix to find the PC (Press *et al.*, 2003)

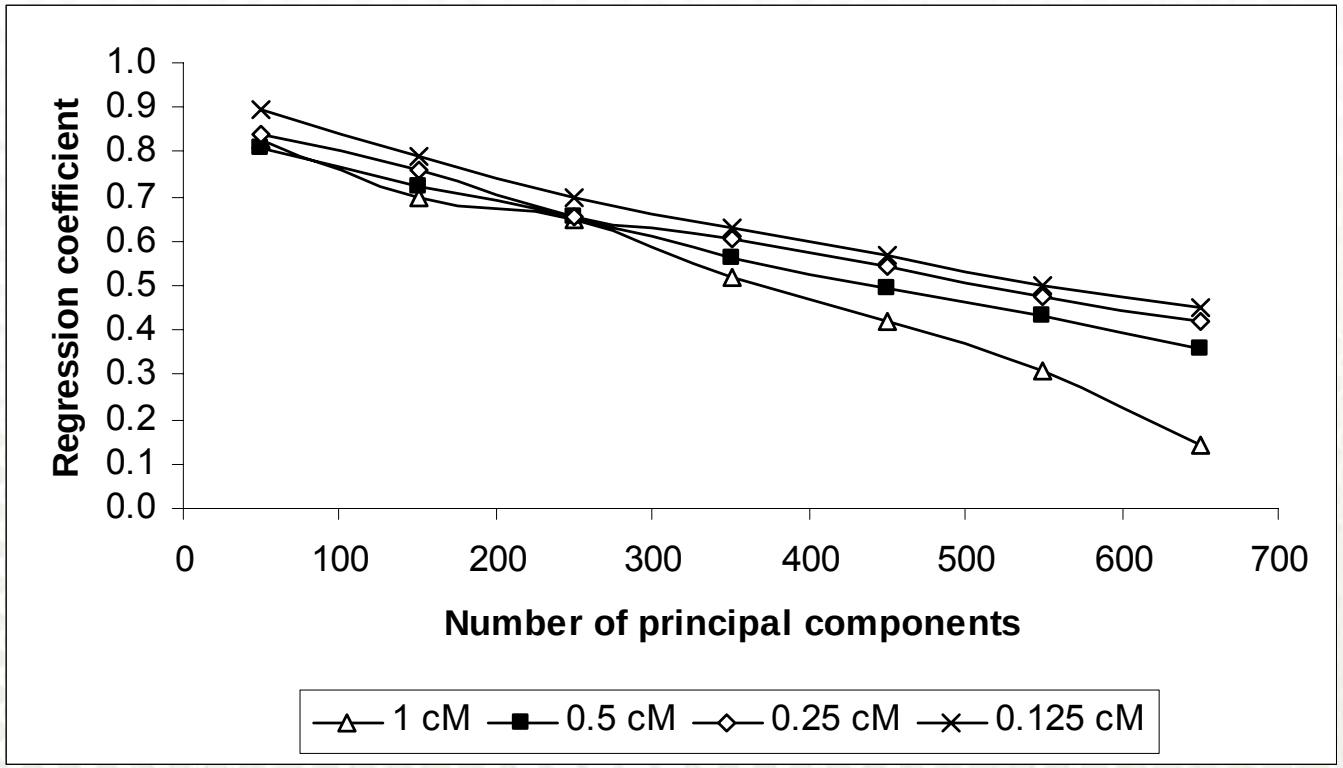
Results

- Optimum number of principal components using PCR (correlation)



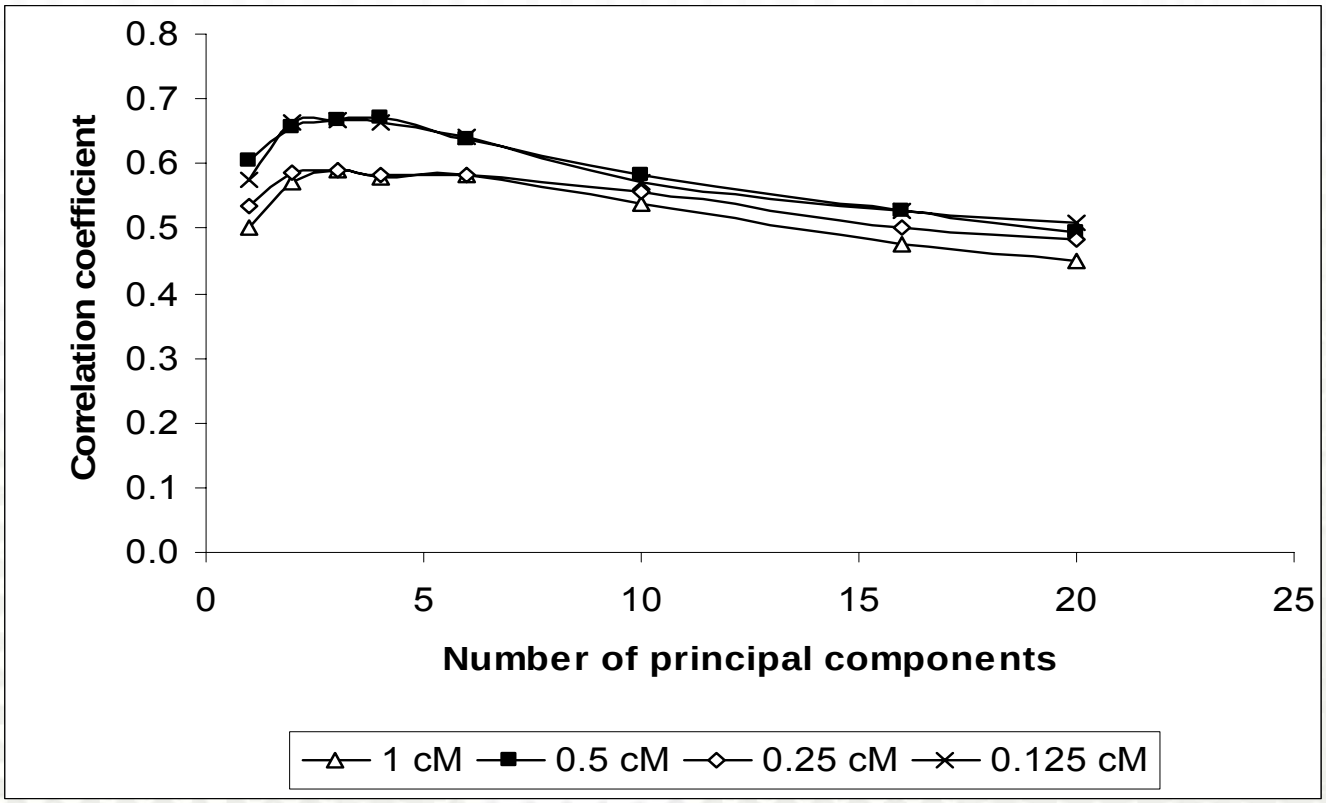
Results (cont.)

- Optimum number of principal components using PCR (regression)



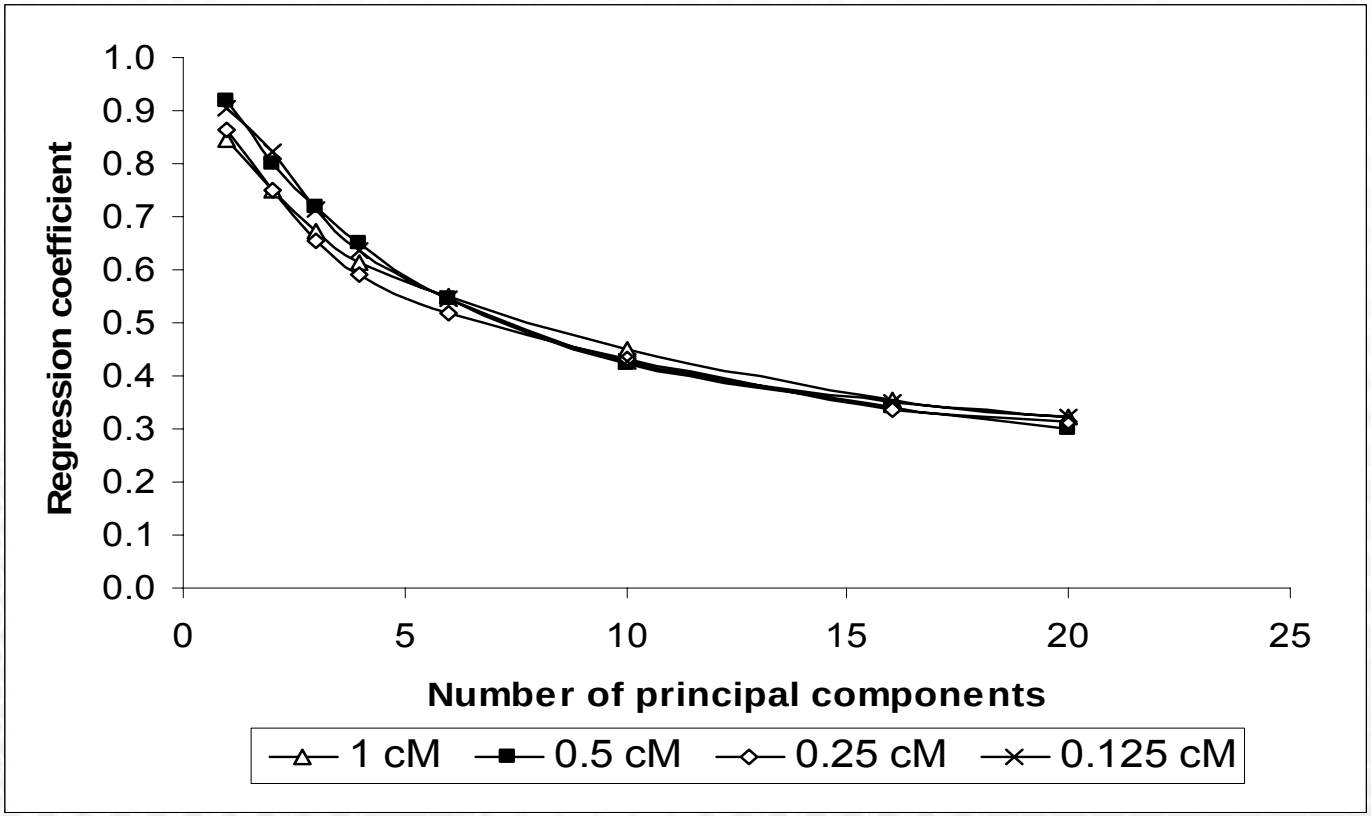
Results (cont)

- Optimum number of principal components using PLSR (correlation)



Results (cont.)

- Optimum number of principal components using PLSR (regression)



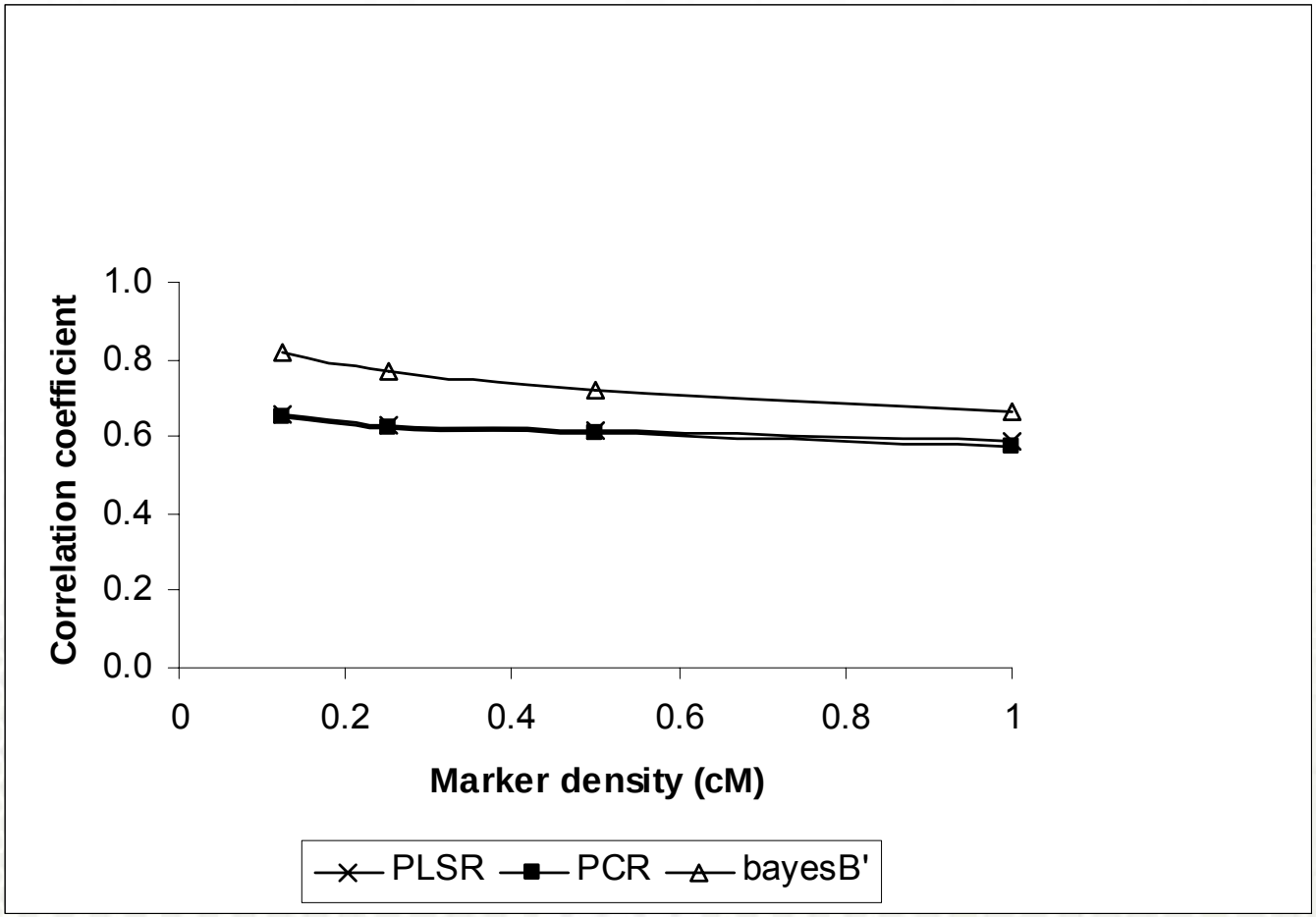
Results (cont.) Correlation

	PCR	PLSR	'bayesB'
Marker density	$r_{TBV;EBV} \pm SE$	$r_{TBV;EBV} \pm SE$	$r_{TBV;EBV} \pm SE$
1 cM	0.570 ± 0.014	0.588 ± 0.014	0.663 ± 0.019
0.5 cM	0.606 ± 0.009	0.612 ± 0.011	0.717 ± 0.014
0.25 cM	0.620 ± 0.010	0.630 ± 0.009	0.771 ± 0.010
0.125 cM	0.649 ± 0.013	0.658 ± 0.012	0.820 ± 0.017

Results (cont.) Regression

	PCR	PLSR	'bayesB'
Marker density	$b_{TBV;EBV} \pm SE$	$b_{TBV;EBV} \pm SE$	$b_{TBV;EBV} \pm SE$
1 cM	0.645 ± 0.011	0.749 ± 0.012	0.880 ± 0.016
0.5 cM	0.642 ± 0.011	0.754 ± 0.012	0.894 ± 0.018
0.25 cM	0.668 ± 0.013	0.782 ± 0.011	0.909 ± 0.012
0.125 cM	0.692 ± 0.010	0.801 ± 0.010	0.910 ± 0.012

Results (cont.)



Conclusions

- PLSR and PCR computationally very fast (hours)
- Simpler and less assumptions compared to ‘bayesB’
- Possible to predict breeding values for next generation (reasonable high accuracy)
- Accuracy increased as the marker density increased
- Why use it?
- Simple, computationally fast, non-parametric method with no assumptions