This is the author's final version of the contribution published as:

GASPA G; M. A. PINTUS; E. L. NICOLAZZI; D. VICARIO; A. VALENTINI; C. DIMAURO; N. P. P. MACCIOTTA,
Use of Principal Component approach to predict Direct Genomic Breeding Values for Beef Traits in Italian Simmental Cattle,

The publisher's version is available at:

https://academic.oup.com/jas/article-abstract/91/1/29/4703008?redirectedFrom=fulltext

When citing, please refer to the published version.

Link to this full text:
http://hdl.handle.net/2318/1686992

1 **Use of Principal Component approach to predict Direct Genomic Breeding Values for Beef**

2 **Traits in Italian Simmental Cattle**

3

4 G. Gaspa†[1], M. A. Pintus†, E. L. Nicolazzi§, D. Vicario‡, A. Valentini¶, C. Dimauro†, N. P. P.

5 Macciotta†

6

7 †Dipartimento di Scienze Zootecniche, Università di Sassari,Sassari, Italy, 07100.

8 §Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza, Italy, 29100.

9 ‡Associazione Nazionale Allevatori Razza Pezzata Rossa Italiana (ANAPRI), Udine, Italy, 33100.

10 ¶Dipartimento di Produzioni Animali, Università della Tuscia, Viterbo, 01100.

11

12 Running Head: Genomic prediction by principal component analysis

13

14 [1]Corresponding author: Giustino Gaspa, Dipartimento di Scienze Zootecniche, Università di

15 Sassari, via De Nicola 9, 07100 Sassari, Italy. Phone number: 0039 079229308. Fax number: 0039

16 079229302. e-mail: gigaspa@uniss.it

17

ABSTRACT

In the current study, principal component (PC) analysis was used to reduce the number of predictors in the estimation of direct genomic breeding values (DGV) for meat traits in a sample of 479 Italian Simmental bulls. SNP marker genotypes were determined with the 54K Illumina beadchip. After edits, 457 bulls and 40,179 SNPs were retained. PC extraction was carried out separately for each chromosome and 2,466 new variables able to explain 70% of total variance were obtained. Bulls were divided into reference and validation population. Three scenarios of the ratio reference:validation were tested: 70:30, 80:20, 90:10. Effect of PC scores on polygenic EBVs was estimated in the reference population using different models and methods. Traits analyzed were daily live weight gain, size score, muscularity score, feet and legs score, beef index (economic index), calving ease direct effect, and cow muscularity. Accuracy was calculated as correlation between DGV and polygenic EBV in the validation bulls. Muscularity, feet and legs, and the beef index showed the highest accuracies calving ease the lowest. In general, accuracies were slightly higher when reference animals were selected at random and the best scenario was 90:10 and no substantial differences in accuracy were found among different methods. Accuracies of direct genomic values were higher than those of traditional PA. Results of the present study suggest possible advantages of the use of genomic index in the pre-selection of performance test candidates for beef traits.

Key Words: cattle, genomic selection, beef traits, principal component analysis

INTRODUCTION

In the last years, the development of high density SNP platforms has had a relevant impact on genetics and breeding research programs for many livestock species. Genotypes of thousands of marker loci are currently used in dairy cattle to search for genomic regions associated with yield and functional traits (Raadsma, et al., 2009; Bolormaa et al., 2010a; Cole et al., 2009) and for predicting genomic enhanced breeding values (GEBV) in genomic selection (GS) schemes. For beef cattle, most of studies have dealt with genome-wide scans for associations between SNP and beef traits such as residual feed intake, average daily gain, hip height, and carcass traits (Bolormaa et al., 2011b, Bolormaa et al., 2011c) or to detect signature of selection able to discriminate between beef and dairy cattle (Hayes et al., 2009a). Until now, less pressure has been put on the implementation of GS programs, even though this technology may represent a valuable option also for beef cattle, allowing to increase breeding value accuracy and to enlarge breeding goals by including traits that are difficult or expensive to measure routinely.

Possible constraints to the application of GS in beef cattle are the limited number of genotyped animals (Garrick, 2011) due to the limited size of male population, and the genotyping costs. The latter issue can be partially addressed by developing a low density SNP chip specific for beef breeds (Rolf et al., 2010), and imputing the 54k chip (Weigel et al., 2010, Berry and Kearney, 2011, VanRaden, 2011). An approach to deal with the disproportion between the limited sample size and SNP number, relevant also for GS programmes in dairy cattle, may be represented by the use of strategies able to reduce predictor dimensionality. Principal component analysis (PCA) and partial least squares regression have been suggested for reducing the number of predictors in DGV calculations both for simulated and actual data (Long et al., 2011; Moser et al., 2009; Solberg et al., 2009). In particular, PCA allows for a considerable reduction (>90%) of the number of independent

62   variables in DGV estimation with accuracies similar to those obtained using directly all SNP

63   genotypes available in simulated and real data (Macciotta et al., 2010a; Solberg et al., 2009; Long et

64   al., 2011).

65   Aim of this work was to calculate DGV for beef traits in the dual purpose Italian Simmental

66   cattle breed. A reduced set of predictors based on linear combinations of SNP genotyped on

67   Illumina platform was obtained by PCA. Moreover, this method was compared with two other

68   approaches commonly used to predict DGV in genomic selection programmes that use directly SNP

69   genotypes as predictors.

70

71   **MATERIALS AND METHODS**

72   *Data*

73   A total of 465 Italian Simmental bulls were genotyped at 54,001 SNP loci using the Illumina

74   Bovine SNP50TM bead-chip (Illumina, San Diego, CA). Animals with more than 1,000 missing

75   genotypes and with inconsistencies in the mendelian inheritance were excluded from the analysis.

76   SNP selection was more conservative and edits were based on the number of missing records ($<$

77   0.025), mendelian inheritance conflicts, absence of heterozygous individuals, minor allele

78   frequency ($>$ 0.05), deviance from Hardy-Weimberg equilibrium ($P < 0.01$) (Wiggans et al., 2009).

79   After editing, 8 animals (2 for mendelian inheritance conflicts, 6 for missing genotypes) and 13,822

80   SNP (21 SNP for mendelian inheritance conflict, 999 SNP with missing exceeding the threshold,

81   12,215 SNP with MAF$\leq$ 0.05 and 587 not in HW equilibrium) were discarded. Final number of

82   bulls and SNP used were 457 and 40,179 respectively. Missing genotypes were replaced with the

83   most frequent allele at that specific locus.

84   Phenotypes used were polygenic EBV provided by Italian Simmental association

85   (evaluation of December 2009). Seven traits were considered: average daily weight gain (ADWG,

86   kg/d), size score (SS), muscularity score (MS), feet and legs score (FLS), beef index (BI =

87   0.40*ADWG + 0.10*SS + 0.40*MS + 0.10*FLS), calving ease direct effect (CED), cow

88   muscularity score(CWM). Table 1 reports EBV average value and reliability. EBV for CED and

89   CWM were derived from progeny test whereas the other traits were measured on performance test.

90   The scale of EBV analyzed were equivalent for different traits (standardized with mean 100 and

91   genetic standard deviation 12).

92   Animals were sorted by year of birth (range 1972-2002) and the whole dataset was split into

93   two subsets, reference (REF) and validation (VAL), containing the oldest and youngest animals,

94   respectively. Different sizes of REF population were tested. Bulls born before 1999, 2000 or 2001

95   were included in the REF population (Figure 1), corresponding to the ratios REF/VAL of 70:30,

96   80:20 and 90:10 respectively.

97

98   ***Statistical model***

99   ***PC-BLUP (BLUP on Principal Components).*** Data matrix $\mathbf{M}_{nxm}$ of marker genotypes was set up (n

100 = total number of individuals, m = number of marker genotypes). Each element m_{ij} corresponded to

101 the genotype at the j-th marker for the i-th individual. Genotypes were coded as -1, 0 or 1, where -1

102 and 1 are the two homozygotes and 0 the heterozygote, respectively (Solberg et al., 2009). PC

103 extraction was carried out separately for each chromosome The number of PCs retained was based

104 on the percentage of variance explained (Macciotta et al., 2010a). Scores of the selected PC were

105 calculated for all individuals. The estimation of effects of the PC on the REF data set was carried

106 out using a BLUP model.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{g} + \mathbf{e} \qquad [1]$$

where $\mathbf{y}$ is the vector of polygenic EBVs, $\mathbf{1}$ is a vector of ones, μ is the overall mean, $\mathbf{Z}$ is the matrix of PC scores, $\mathbf{g}$ is the vector of PC regression coefficients treated as random, and $\mathbf{e}$ is the vector of random residuals. Random PC effects ($\mathbf{g}$) were assumed identically and normally distributed with $g_i \sim N(0, \mathbf{I}\sigma_{gi}^2)$ where $\sigma_{gi}^2 = \sigma_a^2/k$ (σ_a^2 = additive genetic variance, k=number of PC retained). Random residuals were assumed normally distributed with $e_i \sim N(0, \mathbf{I}\sigma_e^2)$. Variance components were supplied by breed associations. BLUP mixed model equations were solved by using Gauss-Seidel iterative method.

PC-BLUP_EIGEN. It is the same method as above, but the (Co)variance matrices of random PC effects (G) and residuals (R) were modeled as diagonal $\mathbf{I}\sigma_{gi}^2\lambda_j$ and $\mathbf{I}\sigma_e^2$ respectively. In particular, the contribution of each j-th principal component to the genetic variance was assumed to be proportional to its corresponding eigenvalue (λ_i) $\sigma_{gi}^2 = (\sigma_a^2/k)*\lambda_j$ (Macciotta et al., 2010a).

To evaluate the effect of the reduction of predictor dimensionality on genomic predictions DGV were calculated also with other two approaches that directly uses all markers available (R-BLUP and BAYES A), but with different theoretical assumptions on the distribution of marker effects. Hereafter, these are named "full models".

R-BLUP. In this model, marker effects were estimated using the same structure of model [1]. In this case, $\mathbf{Z}$ is the design matrix of SNP genotypes – coded as 0,1 and 2 according to the number of copies of the second allele. Marker effects were assumed to be sampled from the same normal distribution. (Co)variance matrix of SNP effects (G) was modelled as diagonal $\mathbf{I}\sigma_{gi}^2$, where $\sigma_{gi}^2 = \sigma_a^2/n$, with n equal to the number of SNP. Mixed model equations were solved using a Gauss-Seidel iterative algorithm until convergence.

129   ***BAYES A***. A Bayes A model (BAYES A) that allows for variance to differ across chromosome

130   segments (Meuwissen et al., 2001) was fitted:

131    $$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{Wu} + \mathbf{e} \qquad [2]$$

132   where **W** is the incidence matrix that allocate the animal with their phenotypic record and **u** is a

133   vector of polygenic breeding values assumed to be normally distributed, with $u_i \sim N(0, \mathbf{A}\sigma_a^2)$, where

134   **A** is the numerator relationship matrix and σ_a^2 is the additive genetic variance. The other symbols

135   were the same as in model [1]. Prior structure and hyper-parameters were chosen according to

136   Meuwissen et al., (2001). A scaled inverted chi-squared prior distribution was assumed for SNP

137   specific variances, under the hypothesis that most of markers have nearly zero effects and only few

138   have large effects. A total of 20,000 iterations were performed, discarding the first 10,000 as burn-

139   in and considering no thinning interval. A residual updating algorithm was implemented to reduce

140   computational time (Legarra and Misztal, 2008).

141   ***DGV estimation and accuracy assessment.*** The overall mean (μ) and the vector ($\hat{\mathbf{g}}$) of the PC

142   scores (or marker effects in full models) estimated in the REF animals with the above described

143   methods were used to calculate the DGV for VAL bulls as:

144    $$\hat{\mathbf{y}} = \mu + \mathbf{Z}\hat{\mathbf{g}}$$

145   where $\hat{\mathbf{y}}$ is the vector of DGV, **Z** is the matrix of PC scores (or marker genotypes in full models) for

146   validation bulls.

147   The accuracy of the genomic prediction in the validation set was evaluated through analysis of

148   Pearson correlation between EBV and DGV. To evalue the difference between DGV and traditional

149   polygenic evaluations, DGV accuracies were compared with correlations between EBV and Parent

150   Average (PA) calculated for beef traits included in the BI.

151 Bias was assessed by examining regression coefficient of EBV on predicted DGV, and 95%

152 confidence interval for b estimates was calculated. Mean squared error of prediction (MSEP) and its

153 partition in different sources of variation related to systematic and random errors (Tedeschi, 2006)

154 were used to evaluate the goodness of prediction.

155

156 **RESULTS**

157 *Accuracy of genomic prediction*

158 The number of principal components to retain was assessed based on the pattern of DGV

159 accuracies for increasing amounts of explained variance (Figure 2). A slight increase of DGV

160 accuracy can be observed for larger proportions of explained variance, with a peak at 0.70 for some

161 traits. This value, that corresponded to 2,466 extracted PC from the whole genome, was further used

162 in the study. Actually it minimized the computational demand of DGV estimation without losing in

163 accuracy. The distribution of extracted PC basically was proportional to the number of markers

164 present in the chromosome (Figure 3).

165 Table 2 reports the Pearson correlation coefficients between DGV and polygenic EBV

166 across four different estimation methods and for different REF:VAL ratios. Accuracies were

167 moderate to high except for CED, which showed lowest values (on average 0.24) across all

168 different validation sets and estimation methods. In particular, highest accuracies were obtained for

169 traits related to muscularity: average $r_{EBV, DGV}$ across estimation methods were 0.82, 0.73, 0.76 and

170 0.66 and for CWM, MS, FLS BI, respectively. ADWG and SS showed moderate values (0.45 and

171 0.51, respectively). Values for ADWG are higher than those reported by Rolf et al. (2010) for

172 Angus cattle. Accuracies found for SS were similar to those for stature reported by Olson et al.

173 (2011) in Brown Swiss using BAYES B. Liu et al. (2011) reported a values of 0.71 in German

174   Holstein. Values for CED were close to those reported for Piedmontese (Ajmone-Marsan et al.,

175   2010) and Brown Swiss (Olson et al., 2011). Higher values were reported for Angus bulls (Garrick,

176   2011; Saatchi et al., 2011) but with population sizes greater than 2,000 bulls.

177   In general, DGV accuracy tended to increase for larger REF:VAL ratios in almost all traits.

178   Best values were obtained with a ratio 90:10 (Table 2). A slight effect of the estimation method

179   could be observed, even though without a clear pattern. R-BLUP performed best for ADWG

180   (accuracy of 0.49 averaged across REF:VAL ratios) compared to the other methods. A similar

181   pattern can be observed for BI, due to the relevance of ADWG in its composition. The two methods

182   that used all the markers available showed better average accuracies than the PC based approaches

183   for size score (average values of 0.54 vs 0.48 respectively). No substantial differences can be

184   observed for the other traits. The use of eigenvalues of SNP covariance matrix as prior variance did

185   not result in higher DGV accuracy, except for CED. For this trait, accuracy ranged from 4% to 10%

186   passing from REF:VAL 70:30 to 90:10. In general, for the other traits the PC-BLUP_EIGEN

187   performed the same or slightly worse than PC-BLUP (the maximum difference between the two

188   methods was 7%).

189   Accuracies obtained with methods that used simultaneously all markers as predictors were

190   substantially equivalent. Basically, slightly higher accuracies were found using BAYES A with a

191   maximum difference of 6%. DGV accuracies were substantially higher than $r_{PA,EBV}$ for all traits

192   (Table 2). On average the mean correlation across traits was 0.60 (PC-BLUP), 0.58 (PC-

193   BLUP_EIGEN), 0.60 (R-BLUP) and 0.61 (BAYES A), and these figures were higher than the

194   average accuracy of PA (0.49).

195

196   ***Bias and goodness of prediction assessment.***

197   Regression coefficients between EBV and DGV were quite variable across methods (Figure

198   4). In particular, PC-BLUP and PC-BLUP_EIGEN estimates showed the smallest regression

199   coefficients, in most of cases lower than 1 (on average 0.82±0.27 and 0.89±0.28 respectively)

200   (Figure 4). On the contrary, the methods that use SNP genotypes showed $b_{EBV,DGV}$ higher than 1 (on

201   average 1.78±0.54 R-BLUP and 1.42±0.36 BAYES A) indicating that positive values of DGV

202   underpredict EBV and vice versa for negative DGV values. The effect on prediction bias of CED

203   was less defined compared to all other traits: regression slopes tended to be closer to one only for

204   the full models, whereas they became worse for the PC based approaches. Furthermore, Figure 4

205   shows the lowest variability of the regression coefficients of PC based approaches across different

206   traits in all REF:VAL ratios. Moreover, the PC-based estimates were less inflated than SNP based

207   estimates, in particular PC-BLUP-EIGEN performed slightly better than PC-BLUP, especially

208   when the reference population was larger (REF:VAL 90:10).

209   Table 3 reports the mean squared error of prediction of DGV and its decomposition for all

210   traits and estimation methods. MSEP did not show large variation among traits excepted for MS

211   (average of 60.8) that experienced the lower figure and BI with the highest MSEP (average of 32.7).

212   Within traits, MSEP of DGV obtained using PC as predictors were on average higher than those

213   calculated with SNP. Exceptions were observed for SS, FLS and CWM. PC-BLUP_EIGEN showed

214   MSEP always lower than PC_BLUP except for CWM. In any case, MSEP differences among

215   methods were rather small. On the other hand, larger differences in the MSEP decomposition can be

216   highlighted. In general, mean bias was not very high (highest average value, 0.33, was found for

217   ADWG) and for some traits it was close to zero. The systematic bias was very low for all traits

218   being the maximum obtained for CWM (27% and 23% of the MSEP for BLUP and BAYES A

219   respectively). A large incidence of random errors can be observed among traits with values ranging

220 from 60% (ADGW) to 98% (CED). Methods that use PC as predictors showed the lowest incidence

221 of components related to prediction bias, as inequality of variance, and the highest for sources of

222 random variation as incomplete co-variation.

223

224   **DISCUSSION**

225   In this paper, principal component analysis was used for reducing predictor dimensionality

226 and computational demand in calculating DGV for beef traits. The number of PC retained was

227 about 6% of the number of original variables. The magnitude of such a reduction was similar to the

228 one reported for US Holsteins by Long et al. (2011). The dimension of about 2,500 predictor is

229 quite recurrent in studies aimed at simplifying the predictor space in genomic selection application.

230 For example, Rolf et al. (2010) indicated a minimum threshold of 2,500 SNP markers for estimating

231 a reliable genomic relationship matrix in cattle population.

232   In general, DGV accuracies here obtained were moderate to high. Results on DGV accuracy

233 in literature are scarce and mainly related to feed efficiency and body weight. However, the

234 magnitude of correlations are in agreement with previous reports obtained on Angus (Garrick et al.,

235 2010; Rolf et al., 2010; Saatchi et al., 2011). An exception is represented by direct calving ease

236 which was much smaller in the present study if compared to aforementioned researches. It is rather

237 hard to relate DGV accuracy to some genetic features of the traits, i.e. h^2. However, best values

238 have been obtained for variables related to muscular development and to the robustness of legs.

239 Intermediate are those related to the size and weight of the animals. In any case, DGV accuracies

240 were higher than those of traditional parent averages, thus evidencing the superiority of the GS over

241 traditional evaluations.

242   Other possible interpretation of the presented DGV accuracy may be the effects of the

243   relatedness between reference and validation bulls which affects the accuracy as shown by Habier et

244   al. (2010) that split the observed accuracy into two component, one related to LD and the other due

245   to the relatedness of bulls in training and prediction population. Being 69 the number of sire-son

246   pairs a possible effect of the relatedness might be envisaged. A high number of phenotypic records

247   are needed to achieve reasonable accuracy as to overcome the curse of dimensionality and GS

248   implementation.

249   Among the factors that affected DGV accuracies, size of REF population and heritability of

250   the traits were the most important. The increase of the size of the reference population has been

251   widely reported to improve the accuracy of genomic prediction (Meuwissen et al., 2001; Liu et al.

252   2011). Also in the present study, for larger sizes of REF population a moderate increase of $r_{EBV,DGV}$

253   was observed. In general, the lower the heritability the larger the references population needs to be

254   (Hayes et al., 2009b). Simulation studies showed how the heritability of the trait affects positively

255   the estimation accuracy (Calus and Veerkamp, 2007; Kolbehdari et al., 2007) as confirmed also by

256   theoretical expectations (Daetwyler et al., 2008). The combination of low heritability and reduced

257   population size may be able to explain the results presented here on CED accuracy.

258   In general, no large differences in DGV accuracies were found between estimation methods

259   (on average 0.03, range 0.02-0.10). Methods used in this research basically differed in two aspects.

260   The first is the kind of predictors, i.e. SNP or PC scores. Results here obtained confirm the

261   substantial equivalence between the two approaches, already observed on simulated (Macciotta et

262   al., 2010a; Solberg et al., 2009) and real data for milk traits (Long et al., 2011; Macciotta et al.,

263   2010b). The second point deals with the distribution of predictor effects. Two methods, PC-BLUP

264   and R-BLUP, assume an equal contribution of each predictor (SNP or PC score) on the variance of

the trait whereas the BAYES A and PC-BLUP_EIGEN relies on a heterogeneity of variance across predictor effects. Early results on simulated data have highlighted the net superiority of the BAYES method over the BLUP approach, confirming the suitability of the finite locus model. However, also in the present work the two approaches yielded the same results, in agreement with reports on real data for dairy cattle (VanRaden et al., 20009).

On the other hand, difference between the kind of predictors was evident in the evaluation of prediction bias. PC based approaches were characterized by the lowest variability of $b_{EBV,DGV}$ within traits and by the predominance of the random components in the composition of the MSEP. These results are probably due to the orthogonality of PC scores that prevent problems of mullticollinearity between predictors. Apart from the relevant impact on calculation time (about 2 minute for PC-BLUP with 2.33 GHz Quad core processor and 4 Gb RAM; 3-8 hours for the R-BLUP 4x4 with Quad core processors and 128 Gb RAM; 3 hours for BAYES A using 3.2 GHz processor 8GB RAM), the PCA approach carried out by chromosome was effective also in reducing the gap between predictors and observations, which is a cause of bias for the application of multivariate techniques on non positive definite correlation matrices (Dimauro et al., 2011). Furthermore, PC-BLUP approach is a trait independent methods as the reduced set of variable may be used for different set of phenotypic measures.

CONCLUSIONS

Direct genomic values accuracies for some beef traits in the dual purpose Italian Simmental cattle breed exhibited high to moderate values. DGV accuracies were higher than those of PA. These figures may open interesting perspectives for the implementation of GS in this breed not only

287 for dairy but also for beef traits. The early availability of DGV with high or moderate accuracies

288 may allow for a better selection of young bulls entering performance test.

289   The reduction of predictor dimensionality by using principal component had a relevant

290 impact in reducing computational time without reduction in accuracies. Difference in assumptions

291 of predictor effect distribution does not seem to affect DGV accuracies

292 <div align="center">**AKNOWNLEDGMENT**</div>

294

295 <div align="center">**REFERENCES**</div>

296 Ajmone-Marsan, P., Macciotta, N.P.P., Pintus, M.A., Gaspa, G., Pieramati, C., Nicolazzi, E.,
297   Albera, A., Nardone, A., Valentini, A. 2010. Accuracies of Direct Genomic Breeding Values
298   for calving ease estimated on Italian Piedmontese bulls with a principal component approach
299   page 62 in Proc. Internationlal Conference in Animal Genetics, ISAG, Edinburgh, UK.
300 Berry D.P. and J.F. Kearney. 2011. Imputation of gentoypes from low- to high-densisty genotyping
301   platforms and implications for genomic selection. Animal, *(in press)*
302 Bolormaa, S., J. E. Pryce, B. J. Hayes, and M. E. Goddard. 2010a. Multivariate analysis of a
303   genome-wide association study in dairy cattle. J. Dairy Sci 93: 3818-3833.
304 Bolormaa, S. B. J. Hayes, K. Savin, R. Hawken, W. Barendse, P. F. Arthur, R. M. Herd & M. E.
305   Goddard. 2011b. Genome-wide association studies for feedlot and growth traits in cattle. J.
306   Anim Sci.: jas.2010-3079.
307 Bolormaa, S. L. R. Porto Neto, Y. D. Zhang, R. J. Bunch, B. E. Harrison, M. E. Goddard & W.
308   Barendse. 2011b. A genome wide association study of meat and carcass traits in Australian
309   cattle. J. Anim Sci.: jas.2010-3138.
310 Calus, M. P. L., and R. F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring
311   the polygenic effect in genomic breeding value estimation with a marker density of one SNP
312   per cM. J. Anim. Breed. Genet. 124: 362-368.
313 Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel,
314   J. F. Taylor & G. R. Wiggans. 2009. Distribution and Location of Genetic effects for Dairy
315   traits. J. Dairy Sci., 92: 3542-3542.
316 Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic
317   risk of disease using a genome-wide approach. Plos One 3: e3395.
318 Dimauro, C., M. Cellesi, M. A. Pintus, and N. P. Macciotta. 2011. The impact of the rank of marker
319   variance-covariance matrix in principal component evaluation for genomic selection
320   applications. J Anim Breed Genet 128: 440-445.
321 Garrick, D. J. 2011. The nature, scope and impact of genomic prediction in beef cattle in the United
322   States. Genet. Sel. Evol. 43.

323   Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic
324   relationship information on genomic breeding values in German Holstein cattle. Genet. Sel.
325   Evol. 42.
326   Hayes, B. J., A. J. Chamberlain, S. Maceachern, K. Savin, H. McPartlan, I. MacLeod, L.
327   Sethuraman & M. E. Goddard. 2009a. A genome map of divergent artificial selection
328   between Bos taurus dairy cattle and Bos taurus beef cattle. Animal Genetics, 40: 176-184.
329   Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection
330   by using the realized relationship matrix. (vol 91, pg 47, 2009). Genet Res 91: 143-143.
331   Kolbehdari, D., L. R. Schaeffer, and J. A. B. Robinson. 2007. Estimation of genome-wide
332   haplotype effects in half-sib designs. J. Anim. Breed. Genet. 124: 356-361.
333   Legarra, A., and I. Misztal. 2008. Technical note: Computing strategies in genome-wide selection.
334   J. Dairy Sci. 91: 360-366.
335   Liu, Z. T., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller & R. Reents. 2011. Impacts of both
336   reference population size and inclusion of a residual polygenic effect on the accuracy of
337   genomic prediction. Genet. Sel. Evol., 43:19
338   Long, N., D. Gianola, G. J. M. Rosa, and K. A. Weigel. 2011. Dimension reduction and variable
339   selection for genomic selection: application to predicting milk yield in Holsteins. J. Anim.
340   Breed. Genet.:128:247-257.
341   Macciotta, N. P. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro, C. Pieramati & A. Cappio-
342   Borlino. 2010a. Using eigenvalues as variance priors in the prediction of genomic breeding
343   values by principal component analysis. J. Dairy Sci., 93:2765-2774.
344   Macciotta, N. P. P., M. A. Pintus, R. Steri, C. Pieramati, E. L. Nicolazzi, E. Santus, D. Vicario, J. T.
345   van Kaam, A. Nardone, A. Valentini & P. Ajmone-Marsan. 2010b. Accuracies of direct
346   genomic breeding values estimated in dairy cattle with a principal component approach. J.
347   Dairy Sci., 93 (suppl 1):532-533 (Abstract).
348   Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using
349   genome-wide dense marker maps. Genetics 157: 1819-1829.
350   Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five
351   methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers.
352   Genet. Sel. Evol. 41:56.
353   Olson, K. M., P. M. VanRaden, M. E. Tooker, and T. A. Cooper. 2011. Differences among methods
354   to validate genomic evaluations for dairy cattle. J. Dairy Sci. 94: 2613-2620.
355   Raadsma, H.W., Khatkar, M.S., Moser, G., Hobbs, M., Crump, R., Cavanagh, J.A.L. and B.Tier.
356   2009. Genome wide association studies in dairy cattle using high density snp scans. Proc.
357   Assoc. Advmt. Anim. Breed. Genet. 18:151-154
358   Rolf , M.M., .J.F Taylor, R.D. Schnabel, S.D. McKay, M.C. McClure, S. L. Northcutt, M. S. Kerley
359   and R.L. Weaber1. 2010. Impact of reduced marker set estimation of genomic relationship
360   matrices on genomic selection for feed efficiency in Angus cattle. BMC genetics 11:24.
361   Saatchi, M., M. McClure, S. McKay, M. Rolf, J. Kim, J. Decker, T. Taxis, R. Chapple, H. Ramey,
362   S. Northcutt, S. Bauck, B. Woodward, J. Dekkers, R. Fernando, R. Schnabel, D. Garrick &
363   J. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle
364   using K-means clustering for cross-validation. Genet. Sel. Evol., 43: 40.
365   Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing
366   dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41: 29.
367   Tedeschi, L. O. 2006. Assessment of the adequacy of mathematical models. Agr Syst 89: 225-247.

368 VanRaden, P. M. et al. 2009. Invited review: Reliability of genomic predictions for North American
369   Holstein bulls. J. Dairy Sci. 92: 16-24.
370 VanRaden, P. M. O'Connell, J.R., Wiggans, G.R. and Weigel, K.A. 2011. Genomic evaluations
371   with many more genotypes. Gen. Sel. Evol., 43:10
372 Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M. and Wiggans, G.R. 2010.
373   Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using
374   reference panels and population-based imputation algorithms. J. Dairy Sci., 93: 2229-2238.
375 Wiggans, G. R., T. S. Sonstegard, P. M. Vanraden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor,
376   F. S. Schenkel & C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and
377   quality of genotypes used in genomic evaluation of dairy cattle in the United States and
378   Canada. J. Dairy Sci., 92:3431-3436.
379
380
381

Table 1. Heritability of average daily weight gain (ADWG), feet and leg score (FLS), Calving Ease direct (CED), Beef Index (BI), Muscularity Score (MS), Size Score (SS) and Cow Muscularity (CWM). Mean and standard deviation of EBV used as phenotypes and their average reliability

Trait	h^2	Mean EBV[a] ± SD	Mean Reliability ± SD
ADWG[b]	0.35	104.08 ± 6.57	0.43 ± 0.12
SS[b]	0.32	103.07 ± 6.45	0.43 ± 0.12
MS[b]	0.61	106.45 ± 9.17	0.60 ± 0.16
FLS[b]	0.25	104.72 ± 7.31	0.42 ± 0.12
BI[c]	-	104.99 ± 6.29	0.43 ± 0.12
CED[d]	0.05	99.13 ± 6.98	0.59 ± 0.17
CWM[d]	0.36	100.76 ± 9.10	0.71 ± 0.21

a) all traits are reported as standardized breeding values with mean 100 and genetic standard deviation 12
b) EBV estimated in performance test
c) Aggregate index of ADWG, SS, MS and FLS
d) EBV estimated in progeny test

Table 2. Correlation coefficient between DGV on EBV of average daily weight gain (ADWG), feet and leg score (FLS), Calving Ease direct (CED), Beef Index (BI), Muscularity Score (MS), Size Score (SS) and Cow Muscularity (CWM) for three estimation methods tested and 3 composition ratios of reference/validation set.

| Trait[1]	PC-BLUP	PC-BLUP_EIGEN	R-BLUP	BAYES A	395 $r_{PA\text{-}EBV}$
REF:VAL 70:30					
ADWG	0.39	0.39	0.43	0.41	0.24
SS	0.43	0.44	0.49	0.50	0.19
MS	0.73	0.67	0.73	0.73	0.72
FLS	0.72	0.73	0.70	0.72	0.61
BI	0.63	0.59	0.67	0.67	0.64
CED	0.23	0.27	0.18	0.23	-
CWM	0.80	0.73	0.80	0.81	-
REF:VAL 80:20					
ADWG	0.36	0.35	0.45	0.39	0.23
SS	0.47	0.47	0.53	0.53	0.08
MS	0.67	0.64	0.70	0.72	0.71
FLS	0.74	0.70	0.74	0.76	0.63
BI	0.57	0.54	0.66	0.64	0.64
CED	0.23	0.27	0.20	0.20	-
CWM	0.85	0.84	0.83	0.85	-
REF:VAL 90:10					
ADWG	0.53	0.51	0.58	0.54	0.24
SS	0.53	0.53	0.61	0.60	0.21
MS	0.81	0.79	0.78	0.81	0.71
FLS	0.85	0.84	0.79	0.83	0.60
BI	0.74	0.71	0.75	0.76	0.64
CED	0.24	0.34	0.22	0.27	-
CWM	0.83	0.81	0.81	0.83	-

397 **Table 3.** Mean squared error of prediction (MSEP) of DGV and its decomposition for beef traits in

398 the validation bulls using different estimation method.

	MSEP[1]	RMSEP	MB	UV	IC	SB	RE
Methods			**ADWG**				
PC-BLUP	44.68	6.68	0.33	0.05	0.63	0.08	0.60
PC-BLUP_EIGEN	41.04	6.41	0.30	0.08	0.63	0.06	0.65
BLUP	38.79	6.23	0.33	0.39	0.28	0.01	0.66
BAYES A	41.14	6.41	0.37	0.26	0.38	0.00	0.64
			SS				
PC-BLUP	43.71	6.61	0.09	0.21	0.71	0.02	0.90
PC-BLUP_EIGEN	42.42	6.51	0.08	0.27	0.66	0.01	0.92
BLUP	44.92	6.70	0.08	0.72	0.20	0.10	0.82
BAYES A	42.93	6.55	0.11	0.57	0.33	0.05	0.85
			MS				
PC-BLUP	63.15	7.95	0.23	0.17	0.61	0.00	0.77
PC-BLUP_EIGEN	61.84	7.86	0.10	0.28	0.63	0.01	0.90
BLUP	59.66	7.72	0.06	0.57	0.38	0.17	0.79
BAYES A	58.70	7.66	0.10	0.47	0.44	0.11	0.79
			FLS				
PC-BLUP	40.01	6.33	0.33	0.11	0.56	0.00	0.67
PC-BLUP_EIGEN	34.50	5.87	0.22	0.25	0.54	0.03	0.76
BLUP	39.73	6.30	0.18	0.46	0.37	0.11	0.72
BAYES A	40.75	6.38	0.27	0.35	0.39	0.07	0.67
			BI				
PC-BLUP	36.25	6.02	0.36	0.08	0.56	0.01	0.64
PC-BLUP_EIGEN	32.76	5.72	0.25	0.15	0.61	0.00	0.75
BLUP	29.93	5.47	0.23	0.42	0.35	0.08	0.70
BAYES A	31.86	5.64	0.31	0.28	0.41	0.03	0.66
			CED				
PC-BLUP	49.13	7.01	0.02	0.14	0.85	0.13	0.86
PC-BLUP_EIGEN	46.54	6.82	0.02	0.17	0.82	0.09	0.89
BLUP	44.79	6.69	0.04	0.69	0.28	0.00	0.97
BAYES A	43.44	6.59	0.03	0.55	0.43	0.00	0.98
			CWM				
PC-BLUP	42.02	6.48	0.01	0.23	0.77	0.02	0.98
PC-BLUP_EIGEN	55.16	7.43	0.02	0.33	0.66	0.04	0.96
BLUP	58.39	7.64	0.03	0.64	0.33	0.27	0.70
BAYES A	51.04	7.14	0.01	0.59	0.41	0.23	0.77

399 1) MB = Mean Bias; UV = Unequal variances; IC = Incomplete covariation; SB = Slope bias; RE = Random
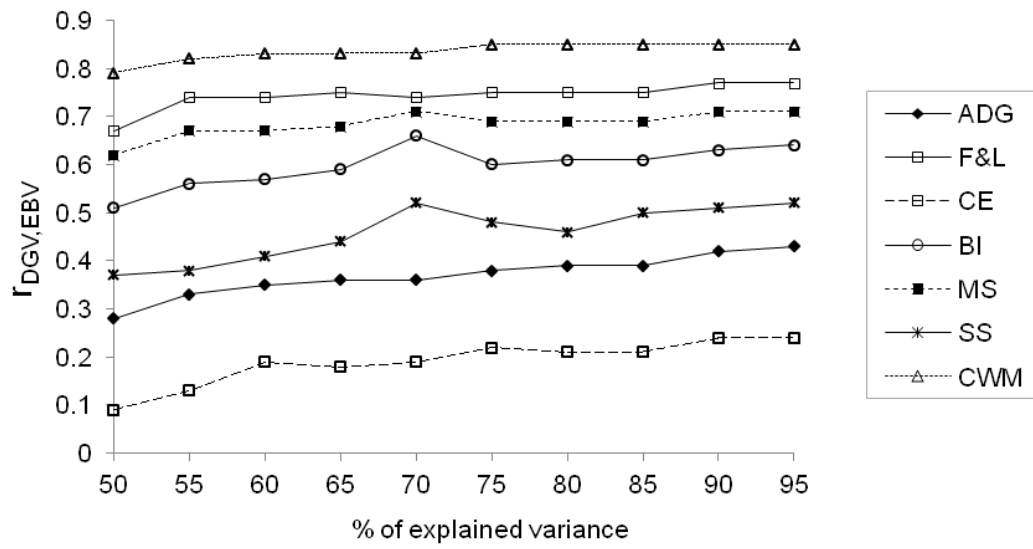
400 errors. Note that MB + UV+ IC= MB + SB + RE = 1
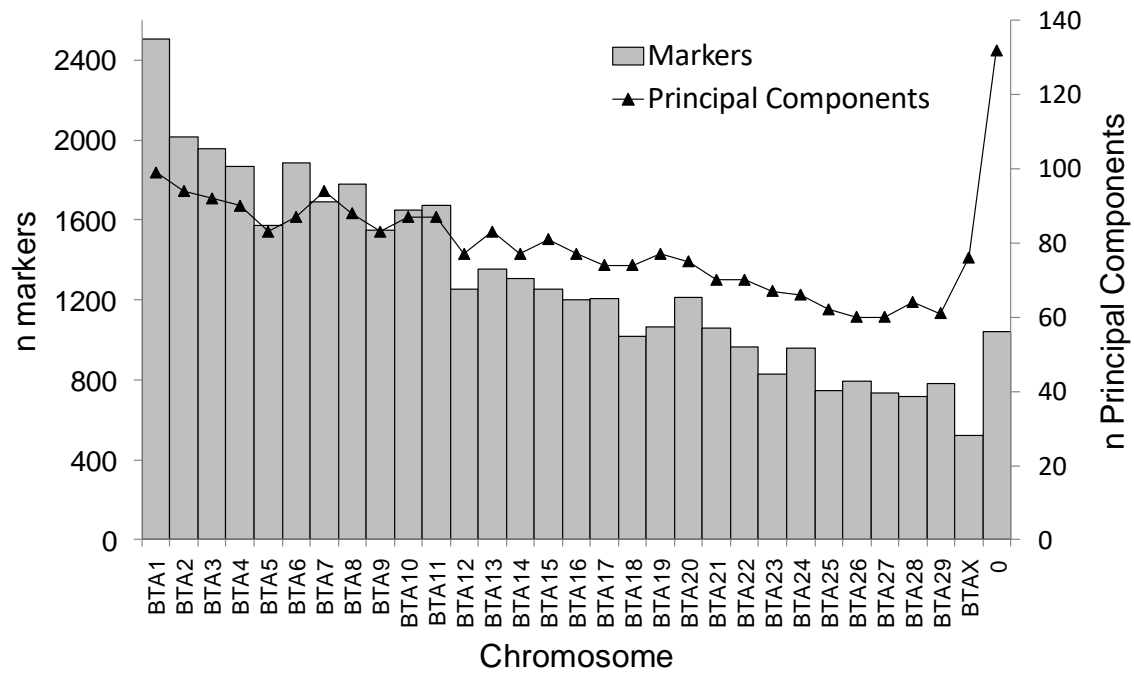
401
402
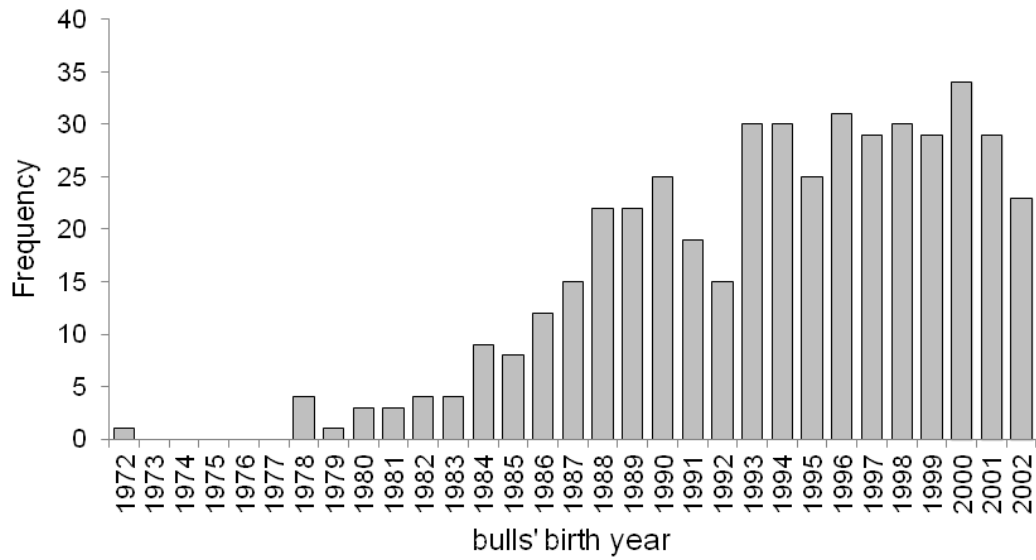
Figure 1

404
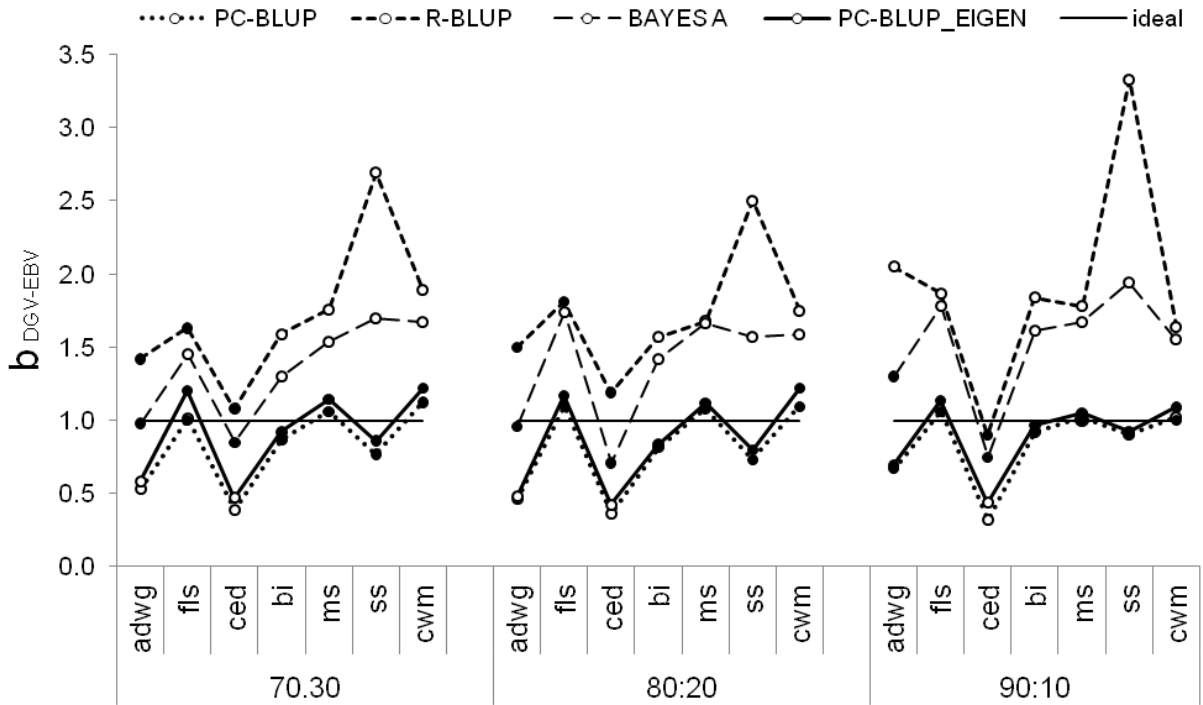**Figure 1**

406

407
408   **Figure 2.**
409

410

411   **Figure 3.**

412

413

414

415

416   Open circle = values of regression coefficient (b) out of the 95% CI including b=1 (p-value <0.001)

417   Solid circle = values of regression coefficient (b) inside the 95% CI including b=1 (p-value <0.001)

418   **Figure 4.**

419   **Figure 1.** Distribution of bulls by birth's year.

420   **Figure 2.** Number markers and number of PC components retained by chromosome.

421   **Figura 3.** Pattern of DGV correlation ($r_{DGV,EBV}$) function of % of variance explained by the PC of 7

422   meat traits (ADWG=average daily weight gain, FLS=Feet and leg score, CED=calving ease direct

423   effect, MS=muscularity score, SS=Size Score, CWM=cow muscularity).

424   **Figura 4.** Pattern of regression coefficient of EBV vs DGV ($b_{EBV,DGV}$) of 7 meat traits

425   (ADWG=average daily weight gain, FLS=Feet and leg score, CED=calving ease direct effect,

426   MS=muscularity score, SS=Size Score, CWM=cow muscularity) both for estimation methods and

427   different REF:VAL ratios.

428