AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Use of different statistical models to predict direct genomic values for productive and functional traits in Italian Holsteins

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1687051> since 2019-02-08T11:08:18Z

Published version:

DOI:10.1111/j.1439-0388.2012.01019.x

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

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

Pintus, M. , Nicolazzi, E. , Kaam, J. , Biffani, S. , Stella, A. , Gaspa, G. ,
Dimauro, C. and Macciotta, N. (2013).

Use of different statistical models to predict direct genomic values for
productive and functional traits in Italian Holsteins.

J. Anim. Breed. Genet., 130: 32-40.
doi:10.1111/j.1439-0388.2012.01019.x

The publisher's version is available at:

<https://onlinelibrary.wiley.com/doi/full/10.1111/j.1439-0388.2012.01019.x>

When citing, please refer to the published version.

Link to this full text:

<http://hdl.handle.net/2318/1687051>

39 **Use of different statistical models to predict direct genomic values for productive and**
 40 **functional traits in Italian Holsteins**

41 Maria Annunziata Pintus¹, Ezequiel Luis Nicolazzi², Johannes Baptist Cornelis Henricus Maria Van
 42 Kaam³, Stefano Biffani³, Alessandra Stella⁴, Giustino Gaspa^{*1}, Nicolò Pietro Paolo Macciotta¹

43

44 ¹ Dipartimento di Agraria–Sezione Scienze Zootecniche, Università di Sassari, Sassari, Italy, 07100.

45 ² Istituto di Zootechnica, Università Cattolica del Sacro Cuore, Piacenza, Italy, 29100.

46 ³ Associazione Nazionale Allevatori Frisone Italiana (ANAFI), 26100, Cremona, Italy.

47 ⁴ Istituto di biologia e Biotecnologia Agraria CNR, Milano, Milan, Italy 20133.

48

49

50 ^{*}Corresponding author: Giustino Gaspa, Dipartimento di Agraria – Sezione Scienze Zootecniche,
 51 Università di Sassari, via De Nicola 9, 07100 Sassari, Italy. Phone number: +39 079229308. Fax
 52 number: +39 079229302. e-mail: gigaspa@uniss.it

53

54 **Keywords:** genomic selection, SNP, principal component analysis, cattle breeding

55

Summary

One of the main issues in genomic selection is the huge unbalance between number of markers and phenotypes available. In this work, principal component analysis is used to reduce the number of predictors for calculating direct genomic breeding values (DGV) for production and functional traits. 2,093 Italian Holstein bulls were genotyped with the 54K Illumina beadchip and 39,555 SNP markers were retained after data editing. Principal Components (PC) were extracted from SNP matrix and 15,207 PC explaining 99% of the original variance were retained and used as predictors. Bulls born before 2001 were included in the reference population, younger animals in the test population. A BLUP model was used to estimate the effect of principal component on Deregressed Proof (DRPF) for 35 traits and results were compared to those obtained by using SNP genotypes as predictors either with BLUP or Bayes_A models. Correlations between DGV and DRPF did not substantially differ among the three methods except for milk fat content. The lowest prediction bias was obtained for the method based on the use of principal component. Regression coefficients of DRPF on DGV highlighted a relevant difference between methods being lower than one for the approach based on the use of PC and higher than one for the other two methods. The use of PC as predictors resulted in a high reduction of number of predictors (about 38%) and of computational time that was about the 9% of the time needed to estimate SNP effects with the other two methods. Accuracies of genomic predictions were in most of cases slightly higher than those of the traditional pedigree index.

Introduction

Genomic Selection (GS) allows for an early prediction of the genetic merit of selection candidates by combining genotypes of biallelic SNP markers and phenotypes (Meuwissen *et al.* 2001). In GS programs, the effects of a large number of SNP on the considered trait is estimated from a reference (REF) population and then used to predict Direct Genomic Values (DGV) in a test (TEST) population where only marker information is available (Meuwissen *et al.* 2001).

The switch from traditional to GS breeding programmes should be justified by a higher reliability of DGV predictions compared to parent average (PA). Actually, DGV accuracy is primarily influenced by the REF population size and, to a lesser extent, by the estimation method. Early simulation studies highlighted that a few thousands of animals are needed in order to obtain DGV accuracies of 0.7 (Hayes *et al.* 2009b) and that about 30,000 unrelated individuals should be considered as REF to estimate DGV with the 800K chip (Meuwissen 2009). Such figures are rather difficult to achieve in practice, even in the case of major cosmopolite breeds and large international GS projects. Even in the USA, where the Holstein population is larger than in other countries, the REF population size in December 2010 was 16,293 (Wiggans 2011). Actually most studies on Holstein cattle have dealt with REF populations of about one (Berry 2009) or few thousands of animals (VanRaden *et al.* 2009; Habier *et al.* 2010; Liu 2011; Schenkel 2009; Su *et al.* 2010).

The increase of REF population size just by new genotyping is still rather expensive. This situation will be further exacerbated by the use of denser SNP platforms (i.e. 800K) or the whole genome sequence. Cooperation across countries represents a effective way to enlarge the size of reference population. Some experience has already been done. For example, United States, Canada, Italy and Great Britain shared their data (Olson 2011; VanRaden *et al.* 2011) and in Europe the EuroGenomics project allowed Germany, France, The Netherlands and Denmark, Finland and Sweden to join their datasets and obtain a REF population of about 18,000 bulls {Lund, 2011

100 #7516} . Similar experiences have occurred also in other breeds, as the Brown Swiss with the
 101 Intergenomics project (B. Zumbach *et al.* 2010).

102 Apart from the mathematical algorithms, the difference between methods used to predict
 103 DGV is mainly in the assumption on marker effect distribution. The BLUP approach fits an equal
 104 contribution of each SNP to the genetic variance of the trait (Meuwissen *et al.* 2001). It is
 105 equivalent to the use of an animal model with the additive genetic effect structured by the genomic
 106 relationship matrix {Hayes, 2009 #389}. On the other hand, Bayesian methods allow genetic
 107 variance to differ across chromosome segments, assuming that few SNPs have a large effect and
 108 many SNPs have a small effect on the trait, respectively (Hayes *et al.* 2009a; Meuwissen *et al.*
 109 2001; Su *et al.* 2010). Both approaches may implement a mixed inheritance by including a
 110 polygenic effect structured by pedigree relationship matrix to explain a part of the genetic variance
 111 (Habier *et al.* 2010; Berry 2009). In early studies based on simulated data, Bayesian methods
 112 usually outperformed BLUP (Meuwissen *et al.* 2001; Clark *et al.* 2011). On real data, such
 113 differences are no longer detectable except for traits for few genes with a larger effect has been
 114 detected (Hayes *et al.* 2009a; VanRaden *et al.* 2009).

115 A further issue on GS is represented by the adoption of techniques for reducing the huge
 116 unbalance between the number of phenotypes and genotypes available. It represents a basic
 117 requirement in the implementation of GS program in populations of limited size. However,
 118 reduction of predictor dimensionality may also be useful for large populations, as the Holstein
 119 breed, with the perspective of using a 800K SNP chip or the complete sequence in the near future.
 120 SNP pre-selection based on the relevance to the trait or the use of dimension reduction multivariate
 121 methods as principal component analysis (PCA) (Solberg *et al.* 2009; Macciotta *et al.* 2010;
 122 Vazquez *et al.* 2011, Pintus *et al.*, 2012) and partial least squares regression (Moser *et al.* 2009;
 123 Vazquez *et al.* 2011) represent the two main strategies adopted to address this issue). Compared to

124 SNP pre-selection, PCA reduction does not discard any SNP and the reduced panel of predictors is
 125 independent from the trait considered.

126 In this work, DGV of different production and functional traits for a sample of Italian
 127 Holstein bulls obtained by joining data generated in two GS research projects were predicted by
 128 using different types of predictors, i.e. the SNP genotypes or the scores of a reduced number of
 129 principal components. Moreover, also the assumptions on predictor effect are compared by using a
 130 Bayesian or a BLUP method.

131

132 **Materials and methods**

133 **Data**

134 Genotypes of 2,093 Italian Holstein bulls were generated in two Italian research projects: the
 135 SELMOL and the PROZOO. Birth years of the bulls ranged from 1979 to 2007, with an average
 136 number of 72 animals per year. Bulls born before or after 2001 were included in the REF and TEST
 137 populations, respectively. Distribution of REF and TEST bulls across birth years is illustrated in
 138 Figure 1

139 Animals were genotyped using the BovineSNP50 BeadChip (Illumina, San Diego, CA).
 140 Data editing procedure has been performed. SNP were discarded based on missing data (>0.025),
 141 minor allele frequency <0.05), existence of Mendelian inheritance conflicts, absence of
 142 heterozygous genotypic class, deviance from Hardy-Weimberg equilibrium (<0.01 bonferroni
 143 corrected). (Wiggans *et al.* 2009). Markers retained after edits were 39,555. Missing SNP alleles
 144 were replaced by the most frequent allele at that specific locus. A total of 86 bulls were discarded:
 145 48 samples were replicates or had inconsistent mendelian inheritance information, whereas 38
 146 samples had low overall call rate (>1000 missing SNPs).

147 Phenotypes were Deregressed EBV (DRPF) provided by the Italian Holstein Association
 148 ANAFI. Thirty-five productive and functional traits have been considered (Table 1). Not all
 149 phenotypes were available for all bulls, thus small differences in sizes of REF and TEST
 150 populations across traits occurred. On average, sizes of REF and TEST populations were of 1,314
 151 and 624 bulls, respectively, . For each traits, heritability, number of REF and TEST bulls and
 152 average reliability of DRPF are reported in table xx

153

154 **Methods**

155 Methodologies used to calculate DGV differed in the dimensionality of predictors (SNP
 156 genotypes vs. PC scores) and in the assumptions on marker effect distributions (BLUP vs
 157 Bayes_A).

158 **Reduction of predictor dimensionality by Principal Component Analysis**

159 PCA were used to extract latent variables from the SNP matrix (n x m) (where n=total
 160 number of animals, and m=number of SNPs retained after edits). Genotypes were coded as -
 161 $1/\sqrt{2p_i(1-p_i)}$ and $1/\sqrt{2p_i(1-p_i)}$ for two different homozygotes and 0 for heterozygotes,
 162 respectively, where p_i is the frequency of one of the two allele at locus i .{Luan, 2009 #230}.
 163 Principal components were extracted separately for each chromosome for computational reasons.
 164 Previous studies based on simulated data reported the same DGV accuracy for PCA carried out on
 165 the entire genome or separately per chromosome (Macciotta *et al.* 2010). The number of
 166 components to retain was based on the amount of original variance explained, calculated as sum of
 167 eigenvalues. In particular, five thresholds with regard to the amount of variance explained were
 168 considered with a corresponding number of extracted variables ranging from about 2,600 to 15,200

169 (Figure 2). Component scores for each animal were used as predictors in the further steps of DGV
 170 calculation and validation.

171

172 **BLUP**

173 The effect of predictors, either SNP (SNP_BLUP) or principal component scores
 174 (PC_BLUP), on phenotypes of the REF bulls was estimated with the following mixed linear model

$$175 \quad \mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{g} + \mathbf{e} \quad [1]$$

176 where $\mathbf{y}$ is the vector of Deregressed EBV, $\mathbf{1}$ is a vector of ones, μ is the general mean respectively,
 177 $\mathbf{Z}$ is the matrix of SNP genotypes or PC scores, $\mathbf{g}$ is the vector of their effects treated as random,
 178 and $\mathbf{e}$ is the vector of random residuals. Covariance matrices of random effects ($\mathbf{G}$) and residuals
 179 ($\mathbf{R}$) were modelled as diagonal $\mathbf{I}\sigma_{gi}^2$ and $\mathbf{I}\sigma_e^2$ respectively, where λ is σ_e^2/σ_{gi}^2 (where $\sigma_{gi}^2 = \sigma_a^2/n$ PC)
 180 assuming an equal contribution of each predictor to the additive genetic variance. Additive genetic
 181 σ_a^2 and residual σ_e^2 variances for all traits were provided by the Holstein association. BLUP
 182 solutions were estimated using Henderson's normal equations (Henderson 1985) and mixed model
 183 equations were solved using a Gauss-Seidel residual update (GSRU) iterative algorithm (Legarra
 184 and Mistral, 2008)

185

186 **BAYES_A**

187 A Bayes A method (BAYES_A) that assumes that most of markers have very small effects
 188 (e.g. markers not linked to any QTL) and only few have large effects was fitted to the REF data set
 189 with the same structure used in model [1]. Prior distributions and parameters were chosen
 190 according to Meuwissen *et al.* (2001). Twenty thousand iterations were performed, the first 10,000

191 were taken as burn in and thus discarded, and all the others were kept. A residual updating
 192 algorithm was used to solve the model (Legarra *et al.* 2008).

193

194 **DGV estimation**

195 DGVs in the TEST population were calculated using the general mean ($\hat{\mu}$) and the vector
 196 ($\hat{\mathbf{g}}$) of the solution of predictors effects estimated with BLUP or BAYES_A in the previous step as:

$$197 \quad \text{DGV}_k = \hat{\mu} + \sum_{i=1}^m \mathbf{z}'_{ik} \hat{\mathbf{g}}_i$$

198 where $\mathbf{z}$ is the vector of PC scores or marker genotypes and m is the number of PC or
 199 markers used in the analysis.

200 The accuracy of direct genomic values DGV was assessed in TEST individuals by calculating
 201 Pearson correlations between DRPF and DGV. Bias were assessed by examining regression of
 202 DRPF on predicted DGV. Goodness of prediction was evaluated also by calculating the mean
 203 squared error of prediction (MSEP) and by its partition in different sources of variation related to
 204 systematic and random errors (Tedeschi 2006). Moreover, the accuracy of genomic predictions was
 205 compared to the realized accuracies of 2005 pedigree indexes (PI) of TEST individuals for some
 206 traits. PI from 2005 were chosen because nearly all animals in the TEST population did not have
 207 daughter records at that time.

208

209 **Results**

210 The effect of different thresholds of explained variance used in PC extraction on the DGV
 211 accuracy for seven traits in TEST bulls is reported in Figure 2. Basically, correlations between

212 DGV and DRPF exhibit a slight linear increase with increasing amounts of extracted components.
 213 This behavior can be observed for almost all traits except fat percentage. Thus the value of
 214 explained variance further considered in the study was 99%, with a corresponding number of
 215 15,199 extracted components.

216 Pearson correlations between predicted DGV and DRPF in TEST bulls for the different
 217 estimation methods are reported in Table 1. Values were low to moderate and different among traits
 218 and, to a lesser extent, among methods. Smallest accuracies were obtained for reproduction traits,
 219 especially calving ease, for which the correlation was 0.05. Milk composition traits, as protein and
 220 also somatic cell count showed highest values, ranging from 0.40 up to 0.64. Also some
 221 conformation traits as type, udder score and rump angle showed accuracies around 0.50. Yield traits
 222 had intermediate values of correlations (about 0.40-0.45).

223 Slight differences in $r_{\text{DGV,DRPF}}$ between methods were observed (Table 1). In general,
 224 accuracies of PC_BLUP and BAYES_A (for 21 and 12 traits out of 35, respectively) were slightly
 225 higher than those of BLUP method that uses SNP genotypes as predictors. On average, the
 226 maximum and the minimum value of accuracy for each trait differed about 0.04. A relevant
 227 exception is represented by fat percentage where BAYES_A markedly outperformed the other
 228 methods, yielding an accuracy greater than about 0.25 and 0.15 compared to the other approaches.
 229 Such a better performance was also observed for fat yield even though of a reduced magnitude. .

230 Comparison between accuracies of genomic predictions and of pedigree indexes shows a slight
 231 superiority for most of traits for genomic predictions

232 Table 2 shows the coefficient of determination (R^2), mean squared error of prediction and its
 233 decomposition of DGV calculated with the three methods for some selected traits: protein yield, fat
 234 percentage, somatic cell count, longevity, fertility, stature and udder support. The PC_BLUP

method showed the lowest values of MSEP across all the considered traits. Moreover, as far as the decomposition of the MSEP was concerned, for almost all traits this approach was characterized by the lowest incidence of components related to prediction bias, i.e. mean bias (on average 13% of the MSEP) and inequality of variances (22%), and highest for incomplete covariation (66%) and random error (85%), i.e. the sources of random variation. SNP_BLUP and BAYES_A had basically the same composition of the MSEP. Less defined is the pattern across traits. Protein yield, for example, had the highest value for mean bias but the lowest for inequality of variance. In any case, fat percentage and somatic cell count showed the largest incidence of random variation.

Regression coefficients ($b_{\text{DGV,DRPF}}$) of DGV on DRPF are shown in Figure 3. A relevant difference between methods can be observed. Values are lower than one in almost all traits for the PC_BLUP method (on average 0.74 ± 0.21), indicating that positive values of DGV overpredict DRPF and vice versa for negative DGV values. On the contrary, all methods that use directly SNP genotypes showed ($b_{\text{DGV,DRPF}}$) almost always greater than one (except for calving ease): 1.23 ± 0.35 , 1.22 ± 0.37 , for SNP_BLUP and BAYES_A, respectively. Moreover, among all methods, the PC_BLUP showed the lowest degree of accuracy (Figure 3). A definite pattern across traits could not be identified, except for the very low values for calving ease and the rather high (>1.30) for some conformation traits.

Discussion

As expected, due to the limited size of the reference population, prediction accuracies for direct genomic values were low to moderate. For example, squared correlations reported for US Holstein (VanRaden *et al.* 2009) obtained by used a REF population of 3,576 bulls are on average 0.2 higher than those reported in the present work for a set of 23 common traits. Similar differences have been observed with reliabilities reported by Su *et al.* (2010) on a 3,330 Danish Holsteins. In VanRaden *et al.* (2009), the R^2 for Net merit has been calculated also with REF population sizes of

259 1,151 and 2,130. Values were similar to those here reported, i.e. 0.12 and 0.17 vs 0.16, respectively.
 260 Accuracies obtained in the present work were similar to those reported by Moser *et al.* (2010) with
 261 a REF population of 1,847 bulls. All the above mentioned figures confirm the importance of the
 262 reference animals for the realized accuracy of genomic predictions. In any case accuracies of DGV
 263 in this study were equal or in many cases higher than realized accuracies of traditional pedigree
 264 indexes.

265 The reduction of predictor dimensionality from 39555 to 15207 by principal component
 266 analysis did not reduce accuracy of DGV predictions compared to methods that use directly all SNP
 267 genotypes available. In most of cases the PC-BLUP approach gave the best accuracies even if
 268 differences from the other methods were rather small. Such results confirm previous reports on
 269 simulated (Solberg *et al.* 2009; Macciotta *et al.* 2010) and real data (Long *et al.*, 2011; Pintus *et al.*,
 270 2012). The reduction performed in this study was of a lower magnitude compared to some of the
 271 above mentioned research, and the number of PC to be retained was not fixed a priori but based on
 272 the test of different thresholds of explained variance (the number of PC variables were about 38%
 273 of the original variables). However, the effect on computation demand was evident. The average
 274 computation time using GSRU for the PC-BLUP method was about 1,21 min (from 1.14 to 2.81
 275 depending on the trait) 2 hours (from 50 min to 4 h depending on the trait), whereas 1 h 36 min
 276 (from 59 min to 2 h) whereas 18 hours (from 9 h to 29 h) were needed on average with the SNP-
 277 BLUP and BAYES_A approaches using a Linux server with 4 x 4 quad core processors and 128 Gb
 278 RAM.

279 DGV predictions obtained with the PC-BLUP methods were characterized by the lowest
 280 bias. This result has been also confirmed by the decomposition of the mean squared error of
 281 prediction, that highlighted a less bias for the PC-based method compared to the other approaches
 282 Moreover, the comparison between the two BLUP-based methods showed slightly better accuracies

for the PC_BLUP than for the SNP_BLUP (magnitude of difference was always lower than 8%). These results may be ascribed to better numerical properties of the extracted variables compared to the direct use of SNP genotypes. Actually principal components are uncorrelated and this feature prevents problems of multicollinearity that are likely to occur because of linkage disequilibrium between loci when dense marker genotypes are used as predictors (Long *et al.* 2011).

As far as the effect of the assumption on marker effect distribution is concerned, BAYES_A yielded substantially the same accuracies as BLUP methods for almost all traits. These figures do not agree with simulation studies where Bayesian methods performed better than BLUP methods (Meuwissen *et al.* 2001; Habier *et al.* 2007). On the other hand, they are similar to those obtained from real data (Moser *et al.* 2009; Su *et al.* 2010; VanRaden *et al.* 2009). A relevant exception is the genomic predictions of fat percentage. For this trait, the accuracy of the BAYES_A method was markedly higher (>30%) than in BLUP methods. A possible explanation can be found in the genetic structure of the trait. It is well known that fat content is largely influenced by single genes with major effect, DGAT1 (Grisart *et al.* 2004). Previous studies reported that methods that assume heterogeneity of variance across chromosome segments usually perform better than those that assume an equal contribution of all markers to the genetic variation in case of traits influenced by few genes. (VanRaden *et al.* 2009; Hayes *et al.* 2010).

Some differences across traits were evidenced, although no definite trend between categories (e.g. yield, conformation, udder, etc.) was observed. Highest values were observed for milk composition, for some conformation and yield traits. Lowest values were found for calving ease, fertility and most conformation traits. Such different behavior between traits is in agreement with reports on North American (Schenkel 2009; VanRaden *et al.* 2009; Olson 2011) and German (Liu 2011) Holsteins. These figures seem to be related, even if roughly, to the heritability of the trait even if some exception have been observed, as somatic cell count. Liu *et al.* (2011), partially

307 explained the lower genomic accuracies for traits with low heritability as a consequence of the
308 lower accuracies of their conventional EBV in the REF population.

309

310 **Conclusions**

311 In this work direct genomic breeding values of Italian Holstein bulls for productive and
312 functional traits have been calculated using different methods and types of predictors. Realized
313 accuracies of genomic predictions are low to moderate, conforming the importance of the size of
314 the REF populations. However, DGV accuracies were similar or, in many cases, slightly higher than
315 those of pedigree indexes. The use of dimension reduction techniques did not result in a decrease of
316 accuracy of genomic prediction compared to methods that uses all SNP available. Assumptions on
317 distribution of marker effect had a relevant influence in the efficiency of the genomic selection for
318 traits that are known to be affected by a limited number of genes with a large effect.

319

320 **Acknowledgments**

321 Research funded by the Italian Ministry of Agriculture (grant SELMOL) and by the Fondazione
322 CARIPLO (grant PROZOO)

323

324

325

326

References

- Berry, D. P., Kearney F., Hennis B.L. (2009). Genomic Selection in Ireland. *Interbull Bull.*
- Bolormaa, S., J. E. Pryce, B. J. Hayes, M. E. Goddard (2010). Multivariate analysis of a genome-wide association study in dairy cattle. *Journal of Dairy Science* **93**, 3818-3833.
- Clark, S. A., J. M. Hickey, J. H. J. van der Werf (2011). Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution* **43**.
- Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim, A. Kvasz, M. Mni, P. Simon, J. M. Frere, W. Coppieters, M. Georges (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**, 2398-2403.
- Habier, D., R. L. Fernando, J. C. M. Dekkers (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-2397.
- Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, G. Thaller (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution* **42**.
- Hayes, B., M. Goddard (2010). Genome-wide association and genomic selection in animal breeding. *Genome* **53**, 876-883.
- Hayes, B. J., P. J. Bowman, A. J. Chamberlain, M. E. Goddard (2009a). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**, 433-443.
- Hayes, B. J., P. M. Visscher, M. E. Goddard (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). *Genetics Research* **91**, 143-143.
- Henderson, C. R. (1985). Best Linear Unbiased Prediction Using Relationship Matrices Derived from Selected Base Populations. *Journal of Dairy Science* **68**, 443-448.
- Legarra, A., I. Misztal (2008). Technical note: Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360-366.
- Liu, Z., Seefried, F. R., Reinhardt, F., Rensing S., Thaller, G., Reents, R. (2011). Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genetic Selection Evolution* **43**.
- Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel (2011). Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics*, **128**, 247-257
- Macciotta, N. P. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro, C. Pieramati, A. Cappio-Borlino (2010). Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *Journal of Dairy Science* **93**, 2765-2774.
- Meuwissen, T. H. (2009). Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet Sel Evol* **41**, 35.
- Meuwissen, T. H. E., B. J. Hayes, M. E. Goddard (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-1829.
- Moser, G., M. Khatkar, B. Hayes, H. Raadsma (2010). Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genetics Selection Evolution* **42**, 37.
- Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, H. W. Raadsma (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution* **41**.
- Olson, K. M., VanRaden, P. M., Tooker, M. E. and Cooper, T. A. (2011). Differences among methods to validate genomic evaluations for dairy cattle. *Journal of dairy science* **94**, 2613-2620.
- Pintus, M.A , G. Gaspa, E.L. Nicolazzi, D. Vicario, A. Rossoni, P. Ajmone-Marsan, A. Nardone, C. Dimauro, N.P.P. Macciotta. (2012). Prediction of Genomic Breeding Values for dairy traits in Italian Brown and Simmental Bull using a Principal Component Approach. *Journal of Dairy Science (in press)*.

374 Schenkel, F. S., Sargolzaei, M., Kistemaker, G., Jansen, G. B., Sullivan, P. Van Doormaal, B. J., Van Raden, P.
375 M. and Wiggans, G. R. (2009). Reliability of genomic evaluation of holstein cattle in canada.
376 *interbull Bull* **39**.

377 Solberg, T. R., A. K. Sonesson, J. A. Woolliams, T. H. E. Meuwissen (2009). Reducing dimensionality for
378 prediction of genome-wide breeding values. *Genetics Selection Evolution* **41**, -.

379 Su, G., B. Guldbbrandtsen, V. R. Gregersen, M. S. Lund (2010). Preliminary investigation on reliability of
380 genomic estimated breeding values in the Danish Holstein population. *Journal of Dairy Science* **93**,
381 1175-1183.

382 Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agricultural Systems* **89**, 225-
383 247.

384 VanRaden, P., J. O'Connell, G. Wiggans, K. Weigel (2011). Genomic evaluations with many more genotypes.
385 *Genetics Selection Evolution* **43**, 10.

386 VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, F. S. Schenkel
387 (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal*
388 *of Dairy Science* **92**, 16-24.

389 Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, D. B. Allison (2011). Predictive
390 ability of subsets of single nucleotide polymorphisms with and without parent average in US
391 Holsteins (vol 93, pg 5942, 2010). *Journal of Dairy Science* **94**, 537-537.

392 Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S.
393 Schenkel, C. P. Van Tassell (2009). Selection of single-nucleotide polymorphisms and quality of
394 genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *J Dairy Sci*
395 **92**, 3431-3436.

396 Wiggans, G. R., Van Raden, P. M., Cooper, T. A. (2011). The genomic evaluation system in the United States:
397 Past, present, future *Journal of Dairy Science* **94**, 3202-3211.

398

399 B. Zumbach, H. Jorjani and J. Dürr., Brown Swiss Genomic Evaluation. INTERBULL BULLETIN NO. 42. Riga,
400 Latvia, May 31 - June 4, 2010

401

402 Table 1. Pearson correlations between predicted DGV and DRPF, for different estimation methods, for the
 403 test animals.

Trait	Methods			
	SNP-BLUP	PC-BLUP	Bayes_A	PI
PFT	0.42	0.42	0.39	0.41
Milk Yield	0.43	0.43	0.46	0.45
Fat Yield	0.41	0.42	0.49	0.34
Protein Yield	0.39	0.39	0.38	0.40
Fat %	0.44	0.47	0.64	0.45
Protein %	0.51	0.53	0.55	0.50
SCC	0.54	0.54	0.52	
Longevity	0.34	0.35	0.31	
Fertility	0.27	0.28	0.28	
Type	0.51	0.51	0.51	0.43
Overall Conformation Score	0.43	0.42	0.40	
Overall Udder Score	0.48	0.49	0.46	0.41
Overall Feet & Leg Score	0.35	0.35	0.36	
Stature	0.47	0.48	0.46	0.50
Strength	0.36	0.37	0.35	0.13
Body Depth	0.39	0.41	0.37	0.46
Angularity	0.45	0.44	0.44	0.41
Rump Angle	0.52	0.53	0.49	0.43
Rump Width	0.44	0.42	0.43	0.54
Rear leg side view	0.35	0.35	0.34	0.39
Foot Angle	0.38	0.38	0.37	0.35
Rear leg rear view	0.33	0.32	0.34	
Locomotion	0.45	0.44	0.45	
Fore Udder Attachment	0.45	0.45	0.44	0.38
Rear Udder Attachment Height	0.46	0.46	0.44	0.39
Rear Udder Attachment Width	0.26	0.25	0.26	0.30
Udder Cleft	0.41	0.41	0.41	0.41
Udder Depth	0.43	0.45	0.42	0.37
Front Teat Placement	0.42	0.41	0.41	0.26
Teat Length	0.33	0.34	0.32	0.20
Rear Teat Placement	0.36	0.35	0.36	
Direct Calving Ease	0.05	0.05	0.05	
Maternal Calving Ease	0.04	0.04	0.05	
Production Persistency	0.29	0.30	0.30	
Maturity rate	0.34	0.34	0.34	
Average across traits (n=35)	0.39	0.39	0.39	
Average across traits (PA n=24)	0.42	0.43	0.43	0.39

404

405

Table 2. Mean squared error of prediction (MSEP) and its decomposition (%), and coefficient of determination (r^2) of Deregressed Proof on direct Genomic Breeding values for some traits in the PREDICTION animals using different estimation method.

Protein Yield	r^2	MSEP	mean bias	unequal variances	incomplete (co)variation	Systematic bias	Random errors
PC-BLUP	0.15	312.20	0.24	0.10	0.66	0.06	0.70
SNP-BLUP	0.15	327.31	0.31	0.15	0.54	0.02	0.67
Bayes_A	0.14	356.88	0.36	0.19	0.45	0.01	0.63
Fat %							
PC-BLUP	0.22	0.04	0.00	0.26	0.74	0.01	0.99
SNP-BLUP	0.19	0.04	0.00	0.38	0.62	0.00	1.00
Bayes_A	0.42	0.03	0.00	0.20	0.80	0.00	1.00
Somatic Cell Count							
PC-BLUP	0.29	25.34	0.01	0.29	0.70	0.00	1.00
SNP-BLUP	0.29	25.75	0.00	0.42	0.57	0.01	0.99
Bayes_A	0.29	26.49	0.00	0.54	0.46	0.04	0.96
Longevity							
PC-BLUP	0.12	63.37	0.22	0.18	0.60	0.03	0.75
SNP-BLUP	0.11	61.55	0.21	0.29	0.49	0.01	0.78
Bayes_A	0.09	61.46	0.19	0.53	0.28	0.01	0.80
Fertility							
PC-BLUP	0.08	81.05	0.09	0.24	0.67	0.04	0.87
SNP-BLUP	0.07	80.04	0.11	0.36	0.54	0.01	0.88
Bayes_A	0.07	82.37	0.14	0.49	0.37	0.00	0.86
Stature							
PC-BLUP	0.23	1.58	0.21	0.27	0.52	0.00	0.79
SNP-BLUP	0.22	1.74	0.27	0.36	0.38	0.01	0.73
Bayes_A	0.20	1.98	0.32	0.41	0.27	0.02	0.66
Udder support							
PC-BLUP	0.17	1.80	0.11	0.21	0.69	0.02	0.87
SNP-BLUP	0.17	1.83	0.14	0.32	0.54	0.00	0.86
Bayes_A	0.16	2.00	0.21	0.43	0.37	0.01	0.79

- 411 Figure 1. Distribution of number of bulls per birth year in the reference and test population.
- 412 Figure 2. Pearson correlations between predicted direct genomic breeding values and deregressed proof, for
413 the PC-BLUP method using a different number of Principal components (PC) explaining the given proportion
414 of the variance, for the PREDICTION animals.
- 415 Figure 3. Regression coefficients ($b_{DRPF,DGV}$) of Deregressed Proof on direct Genomic Breeding Values
416 estimated with PC-BLUP, SNP-BLUP and BAYES_A methods, and on Parent Average for all traits
417 considered in test animals
- 418

421

422 **FIGURE 1**

423

FIGURE 2

FIGURE 3