J. Dairy Sci. 95:3390–3400
http://dx.doi.org/10.3168/jds.2011-4274

Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach

M. A. Pintus,* G. Gaspa,* E. L. Nicolazzi,† D. Vicario,‡ A. Rossoni,§ P. Ajmone-Marsan,† A. Nardone,# C. Dimauro,* and N. P. P. Macciotta*[1]
*Dipartimento di Scienze Zootecniche, Università di Sassari, Sassari 07100, Italy
†Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza 29100, Italy
‡Associazione Nazionale Allevatori Razza Pezzata Rossa Italiana (ANAPRI), Udine 33100, Italy
§Associazione Nazionale degli Allevatori di Razza Bruna (ANARB), Verona 37012, Italy
#Dipartimento di Produzioni Animali, Università della Tuscia, Viterbo 01100, Italy

ABSTRACT

The large number of markers available compared with phenotypes represents one of the main issues in genomic selection. In this work, principal component analysis was used to reduce the number of predictors for calculating genomic breeding values (GEBV). Bulls of 2 cattle breeds farmed in Italy (634 Brown and 469 Simmental) were genotyped with the 54K Illumina beadchip (Illumina Inc., San Diego, CA). After data editing, 37,254 and 40,179 single nucleotide polymorphisms (SNP) were retained for Brown and Simmental, respectively. Principal component analysis carried out on the SNP genotype matrix extracted 2,257 and 3,596 new variables in the 2 breeds, respectively. Bulls were sorted by birth year to create reference and prediction populations. The effect of principal components on deregressed proofs in reference animals was estimated with a BLUP model. Results were compared with those obtained by using SNP genotypes as predictors with either the BLUP or Bayes_A method. Traits considered were milk, fat, and protein yields, fat and protein percentages, and somatic cell score. The GEBV were obtained for prediction population by blending direct genomic prediction and pedigree indexes. No substantial differences were observed in squared correlations between GEBV and EBV in prediction animals between the 3 methods in the 2 breeds. The principal component analysis method allowed for a reduction of about 90% in the number of independent variables when predicting direct genomic values, with a substantial decrease in calculation time and without loss of accuracy.

Key words: single nucleotide polymorphism, genomic selection, principal component analysis, accuracy

INTRODUCTION

Advancements in genome sequencing technology have been implemented into high-throughput platforms able to simultaneously genotype tens of thousands of SNP markers distributed across the whole genome of livestock species (Van Tassell et al., 2008). Dense marker maps are today used in cattle breeding for genome-wide association studies (Price et al., 2006; Cole et al., 2009) and for predicting genomic breeding values (**GEBV**) of candidates to become sires and dams in genomic selection (**GS**) programs (Meuwissen et al., 2001). The basic framework of genomic selection involves 2 steps. First, effects of chromosomal segments are estimated in a set of reference animals with known phenotypes and SNP genotypes. Then, estimates are used to predict direct genomic values (**DGV**) of animals for which only marker genotypes are known. The DGV are usually blended with other measures of genetic merit such as official pedigree index (**PI**) to obtain the final GEBV (Ducrocq and Liu, 2009; VanRaden et al., 2009). Different countries have implemented GS programs to evaluate young bulls entering progeny testing, achieving reliabilities greater than those of PI (Hayes et al., 2009a; VanRaden et al., 2009). Expected benefits of GS are a reduction in generation interval, increase in EBV accuracy for females, and a cost reduction for progeny testing (Schaeffer, 2006; König et al., 2009).

However, several issues still need to be addressed in GS. Examples are the assessment of the frequency of marker effect re-estimation along generations (Solberg et al., 2009), the impact of population structure on estimated effects (Habier et al., 2010), and the choice of the most suitable mathematical model and dependent variable for the estimation step (Guo et al., 2010). Apart from situations in which the number of genotyped animals is quickly approaching or exceeding the number of markers used, as in the North American genomic project (VanRaden and Sullivan, 2010), the

Received February 15, 2011.
Accepted February 13, 2012.
[1]Corresponding author: macciott@uniss.it

huge imbalance between predictors and observations still represents the main constraint to the implementation of GS programs, especially for breeds other than Holstein.

One way to reduce this data asymmetry could be to combine data from different populations of the same breed or from different breeds in a common reference set, both within and across countries (Boichard et al., 2010). Reports on simulated and real data show some increases in DGV accuracy, but results are strongly dependent on the genetic similarity between breeds and on the trait analyzed (de Roos et al., 2009; Hayes et al., 2009b).

A different strategy is based on the reduction of the number of predictors used in the estimation equations. A straightforward approach is to perform a preliminary selection of markers based on their relationship with the phenotype or of their chromosomal location (Hayes et al., 2009a; Moser et al., 2010; Vazquez et al., 2010). An alternative is represented by the Bayes_B method, which retains markers with nonzero effects on phenotypes directly during the estimation step (Meuwissen et al., 2001; VanRaden, 2008). Other approaches of SNP selection have been proposed mainly for genome-wide association analyses (Gianola et al., 2006; Aulchenko et al., 2007; Long et al., 2007; Gianola and van Kaam, 2008). In the above-mentioned methodologies, selection of SNP is based on their relevance to the considered phenotype. Thus, specific sets of markers may be required for different traits.

An alternative to marker selection for reducing predictor dimensionality is represented by their synthesis via multivariate reduction techniques. In particular, principal component analysis (**PCA**) and partial least squares regression have been proposed for estimating DGV (Solberg et al., 2009). In fact, in the partial least squares regression approach, the extraction of latent variables from predictors is carried out by maximizing their correlation with the dependent variable(s). Thus, reduction of the system dimension is still based on the magnitude of the predictor effects on the considered trait. In contrast, PCA is based entirely on the factorization of the SNP (co)variance (or correlation) matrix. This technique allows for a huge reduction of the number of independent variables (>90%) in the estimation of DGV while achieving accuracies comparable to those obtained using all SNP genotypes (Solberg et al., 2009; Macciotta et al., 2010). A recent comparison highlighted the high accuracy of both dimension reduction techniques in predicting DGV for milk yield in US Holsteins (Long et al., 2011). Compared with other approaches of predictor reduction, PCA limits the loss of information because each SNP is involved in the composition of each principal component (**PC**). Moreover, extracted PC are orthogonal, thus avoiding multicollinearity problems. The PCA approach also allows the variance structure of predictors in the BLUP normal equations to be modeled by using eigenvalues as variance priors (Macciotta et al., 2010). Furthermore, PCA has been used in genome-wide association studies to reduce the number of dependent variables (Bolormaa et al., 2010).

The reduction of predictor dimensionality is a straightforward strategy when implementing GS with reference populations of limited size. This situation may occur in minor cattle breeds or in larger populations in the early stages of GS programs. This was the case for the SELMOL project recently started in Italy that involves different cattle breeds (both dairy and beef). The aim of this study was to calculate GEBV for dairy traits in populations of limited sizes of Italian Brown and Simmental bulls by using the PCA approach for reducing the number of predictors. The PCA-based method was compared with other approaches that directly fit all SNP genotypes available as predictors.

MATERIALS AND METHODS

Data

A total of 775 Italian Brown and 493 Italian Simmental bulls were genotyped at 54,001 SNP loci with the Illumina Bovine SNP50 54K bead chip (Illumina Inc., San Diego, CA). Considering the limited size of the sample, the priority in the edit was to keep the number of bulls as large as possible. A stringent selection was performed on markers. Edits were based on the percentage of missing data (<0.025), Mendelian inheritance conflicts, absence of heterozygous loci, minor allele frequency (<0.05), and deviance from Hardy-Weinberg equilibrium (<0.01; Wiggans et al., 2009). Edits on animals were based on the number of missing genotypes (<1,000) and on inconsistencies in the Mendelian inheritance (96 and 70 father-son pairs were included in the archives for Italian Brown and Simmental, respectively). An overall accuracy >99% was obtained by double-genotyping some animals. A summary of the initial and final number of bulls and SNP, together with the effect of the different elimination steps, is reported in Table 1. In the final data, missing genotypes (in general <0.5%) were replaced by the means of the observed genotypes at that specific locus.

Phenotypes used were multiple across-country evaluation (MACE) deregressed proofs (**DRPF**) provided by the 2 breed associations. Traits considered were

Table 1. Number of animals and markers discarded in the different edit steps

Data set/breed	Repeated[1]	Mendelian inheritance[2]	Missing[3]	MAF[4]	HW[5]	Final data set
Animals						
Brown	17	3	6			634
Simmental	6	2	6			469
SNP markers						
Brown		23	1,118	15,046	560	37,254
Simmental		21	999	12,215	587	40,179

[1]Number of animals genotyped twice to check genotyping quality.
[2]Animals that showed >2,000 Mendelian conflicts or SNP that showed Mendelian conflicts in >2.5% father-son pairs.
[3]Animals with >1,000 missing genotypes or SNP with >2.5% missing genotypes.
[4]SNP with a minor allele frequency (MAF) <0.05.
[5]SNP that deviate significantly ($P < 0.01$) from Hardy-Weinberg (HW) equilibrium.

milk, fat, and protein yields (kg), fat and protein percentages, and SCS. Average reliabilities of DRPF were 0.87 (±0.08) and 0.92 (±0.04) for Italian Brown and Simmental bulls, respectively.

Animals were sorted by year of birth and the data set split into reference (**REF**) and prediction (**PRED**) subsets, comprising older and younger animals, respectively. Three ratios of REF:PRED animals were considered (70:30, 80:20, 90:10). The distribution of years of birth in the different breeds is depicted in Figure 1.

Statistical Models

Principal component analysis was used to extract latent variables from the SNP data matrix **M** with m rows (m = number of individuals in the entire data set; i.e., REF plus PRED) and n columns (n = number of SNP retained after edits). Each element (i,j) corresponded to the genotype at the jth marker for the ith individual. Genotypes were coded as −1 and 1 for the 2 homozygotes, and 0 for the heterozygote, respectively. The

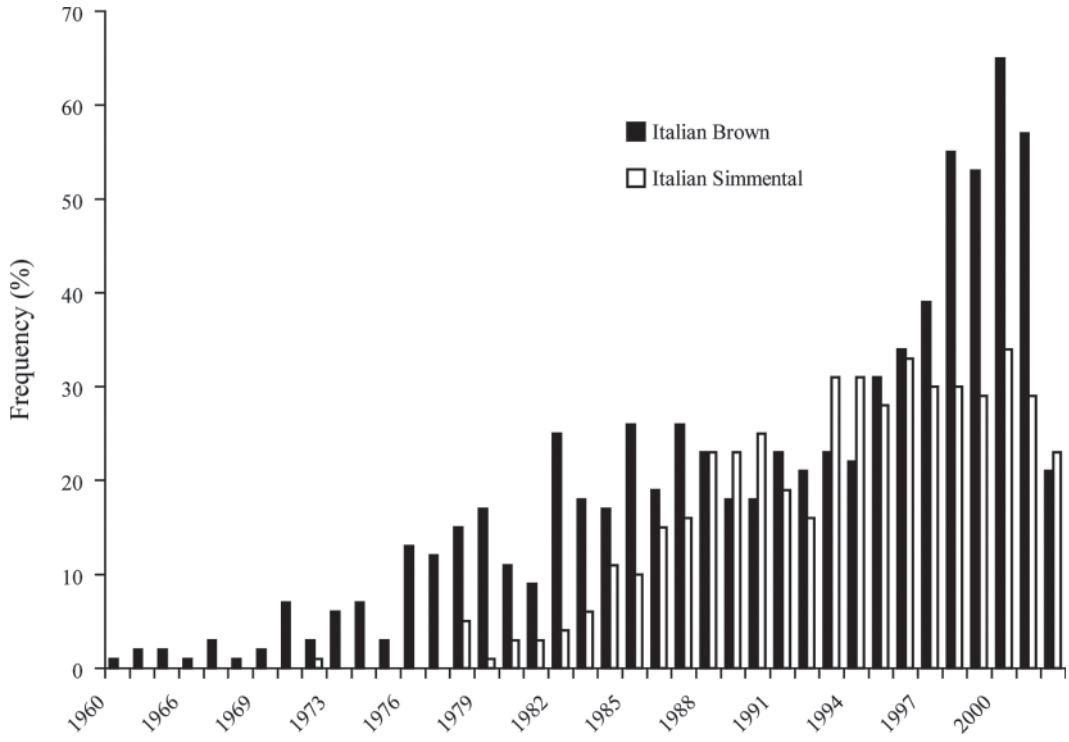

Figure 1. Distribution of number of bulls across year of birth.

PCA was performed separately for each chromosome. On simulated data, whether analyses were carried out on the whole genome simultaneously or separately by chromosome did not affect DGV accuracy (Macciotta et al., 2010). The PCA was carried out on the whole data set (REF + PRED) separately for each breed. The number of principal components retained (k) was based on the sum of their eigenvalues. An empirical threshold of 80% of explained variance was fixed according to indications of other authors (Bolormaa et al., 2010). Scores of the selected components were calculated for all individuals.

For each breed, the estimation of predictor effects on the REF data set was carried out using the following BLUP model (**PCA_BLUP**):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of DRPF, $\mathbf{1}$ is a vector of ones, μ is the general 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. Covariance matrices of random PC effects ($\mathbf{G}$) and residuals ($\mathbf{R}$) were modeled as diagonal $\mathbf{I}\,\sigma^2_{aj}\lambda$ and $\mathbf{I}\,\sigma^2_e$, respectively. In particular, the contribution of each jth principal component to the genetic variance was assumed to be proportional to its corresponding eigenvalue (λ); that is, $\sigma^2_{ji} = \left(\sigma^2_a\right) \times \lambda_j$ (Macciotta et al., 2010). Variance components were those currently supplied by breed associations for Interbull evaluations (http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm).
The BLUP solutions were estimated using Henderson's normal equations (Henderson, 1985) solved by using an LU (lower-upper) factorization, where the left-hand-side part of mixed model equations was decomposed into the product of a lower and a upper triangular matrix, respectively (Burden and Faires, 2005).

To evaluate the effect of the PCA reduction of predictors on DGV accuracy, the estimation step was also carried out with 2 methods that fit all available SNP genotypes, but with different assumptions on the distribution of their effects.

The first was the BLUP method (**SNP_BLUP**) that assumed an equal contribution of each marker locus to the variance of the trait, sampled from the same normal distribution (Meuwissen et al., 2001). In this case, $\mathbf{Z}$ was the matrix of SNP genotypes coded as 0, 1, and 2. Mixed model equations were solved using a Gauss-Seidel iterative algorithm.

The second was the Bayes_A method, which allowed variance to differ across chromosome segments on the assumption that a large number of SNP have small effects and few have a large effect (Meuwissen et al., 2001). The fitted model (BAYES_A) was

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{Wu} + \mathbf{e},$$

where $\mathbf{u}$ is a vector of polygenic breeding values assumed to be normally distributed, with $u_i \sim N\left(0, \mathbf{A}\sigma^2_a\right)$, where $\mathbf{A}$ is the average relationship matrix and σ^2_a is the additive genetic variance. Prior structure and hyperparameters were chosen according to Meuwissen et al. (2001). A scaled inverted chi-squared prior distribution was assumed for SNP specific variances, under the hypothesis that most of markers have almost zero effects (i.e., markers not linked to any QTL) and only a few have large effects. In total, 20,000 iterations were performed, discarding the first 10,000 as burn-in and considering no thinning interval. A residual updating algorithm was implemented to reduce computational time (Legarra and Misztal, 2008).

The general mean (μ) and the vector $\left(\hat{\mathbf{g}}\right)$ of the principal component or marker effects estimated either with BLUP (SNP_BLUP) or Bayes A (BAYES_A) in the REF population were used to calculate the DGV for the jth animal in the PRED subset for each breed:

$$DGV_j = \mu + \sum_{i=1}^{k} \mathbf{z}'_{ij}\,\hat{\mathbf{g}}_i,$$

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

The DGV obtained with PC_BLUP, SNP_BLUP, or BAYES_A were blended with PI to obtain GEBV, using the equivalent daughter contributions (EDC) as weighting factors:

$$GEBV_i = DGV_i \cdot edcG + PI_i \cdot edc_i,$$

where *edcG* and *edc* are the EDC for DGV or PI, respectively. Values of *edcG* were calculated from the approximate DGV reliabilities, obtained as $\mathrm{REL_{DGV}} = (\mathrm{r^2_{DGV,DRPF}})/\mathrm{REL_{DRPF}}$ (Hayes et al., 2009b):

$$edcG = k \times \mathrm{REL_{DGV}}/(1 - \mathrm{REL_{DGV}}),$$

where $k = (4 - \mathrm{h^2})/\mathrm{h^2}$. The same approach was used to calculate *edc* for PI. The procedure followed was that used to validate the international GEBV of Italian Simmental approved in November 2011 (http://www.interbull.org).

Finally, to evaluate the efficiency of genomic predictions versus the traditional polygenic evaluations in PRED individuals, squared correlations between GEBV and EBV ($\mathbf{R^2_{EBV\text{-}GEBV}}$) were computed and compared with those between PI and EBV ($\mathbf{R^2_{EBV,PI}}$). Bias was assessed by evaluating regression coefficients of EBV on predicted GEBV.

RESULTS

A common criterion for choosing the number of PC to retain is visual inspection of their eigenvalue pattern. As an example, Figure 2 reports the chromosome-wide variance explained by each successive component extracted from SNP located on BTA6 in the Brown breed. The eigenvalue was small for the top 2 components (about 7 and 5%, respectively), with a smooth decrease followed by a plateau reached at about 100 PC (86% of variance explained) for this chromosome. The number of retained PC genome-wide was 3,596 and 2,257 for the Simmental and Brown breeds, respectively. A similar number of components was retained by Long et al. (2011). In any case, a large reduction of predictor dimensionality (<10% of the number of original variables) was realized.

The extracted PC were able to distinguish Brown from Simmental bulls. Individual scores of the first PC of BTA6, for example, discriminated the 2 breeds, whereas the third PC highlighted a larger heterogeneity within the Brown sample (Figure 3). In PCA, the meaning of each extracted component is usually inferred by looking at eigenvector coefficients; that is, the weights of each original variable (in this case, the SNP genotype) in the component. However, it would be very hard to achieve an interpretation by examining thousands of correlations. The meaning of extracted variables could be assessed indirectly by looking at their relationships with other characteristics of the considered individuals. For example, the third PC for BTA6 in the Brown breed was negatively correlated with the observed average individual heterozygosity (−0.43), and its score average showed a progressive decrease across year of birth of bulls. Such an ability of PCA to cluster individuals based on causes of variation of SNP genotype frequency has also been reported for simulated data (Macciotta et al., 2010).

Squared correlations between GEBV or PI and EBV are reported in Tables 2 and 3 for the 2 breeds. The $R^2_{GEBV,EBV}$ values were substantially lower for Brown than for Simmental, except for fat and protein percentages, which showed the opposite behavior. Squared correlations for GEBV were generally higher than those for PI in the Brown breed. Similar behavior could also be observed for the Simmental, except for fat and protein percentages. The PC_BLUP and BAYES_A methods performed better than the SNP_BLUP method in

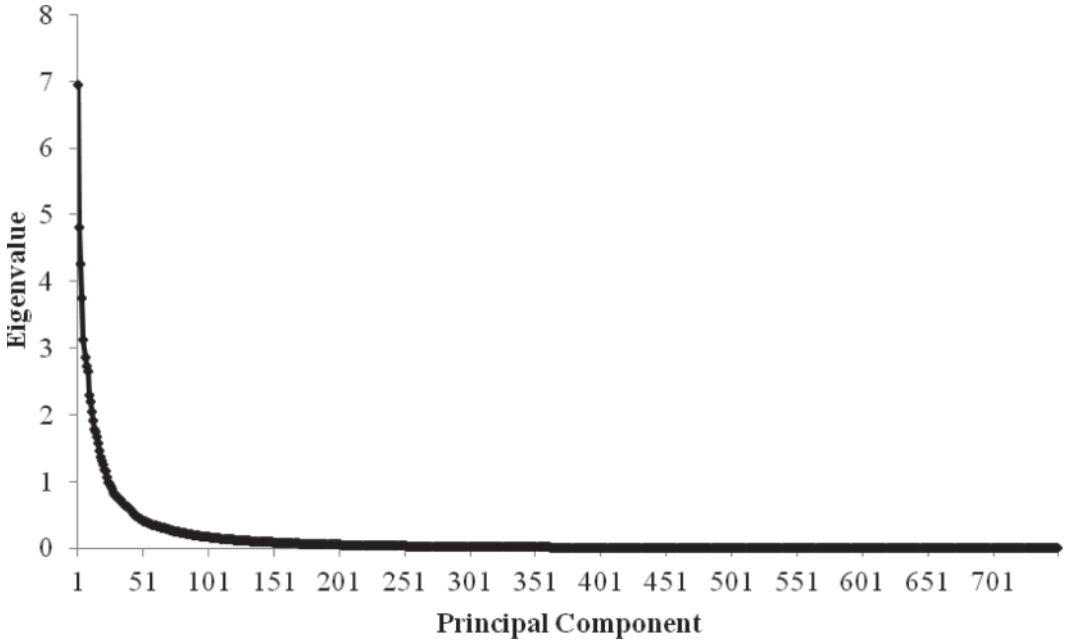

Figure 2. Pattern of the proportion of variance (%) accounted for by each successive principal component extracted from the correlation matrix of SNP markers for chromosome 6 in the Brown breed.

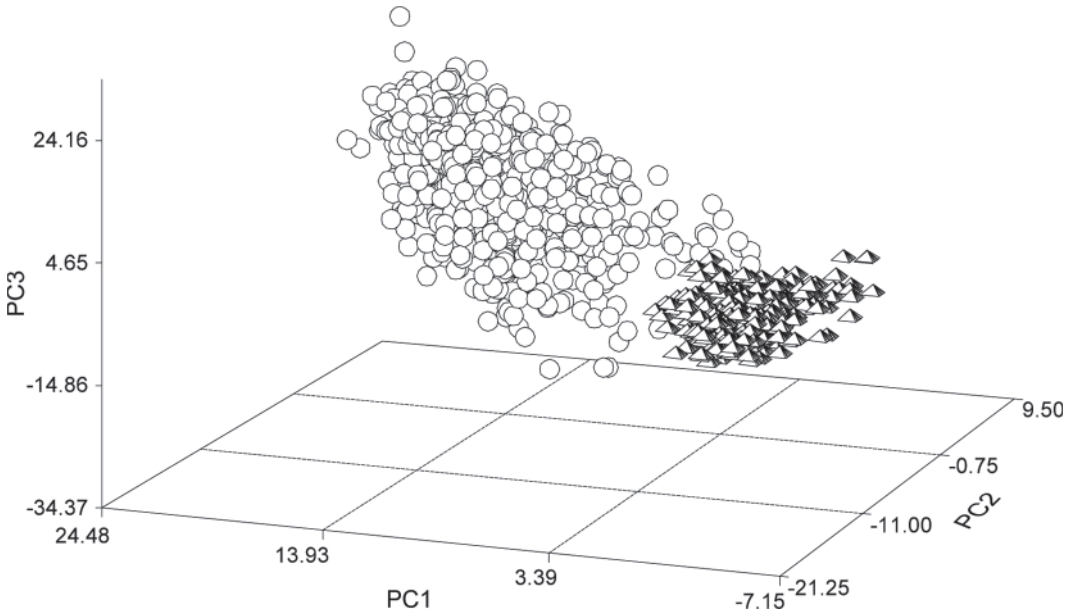

Figure 3. Plot of the individual scores of the first 3 principal components (PC1, PC2, and PC3) extracted from chromosome 6 in the 2 breeds (circles = Brown; pyramids = Simmental).

Brown bulls. Finally, increasing the ratio REF:PRED seemed to increase $R^2_{EBV,GEBV}$ in Brown, whereas no such effect was observed in Simmental.

In particular, squared correlations ranged from 0.01 to 0.39 for Italian Brown (Table 2). The lowest values were obtained for yield traits, in particular for milk and protein (on average <0.1). The highest $R^2_{EBV,GEBV}$ were observed for fat percentage, protein percentage, and SCS (on average 0.35, 0.32, and 0.15, respectively). Olson et al. (2011) reported the same value of genomic

Table 2. Squared correlations between genomic breeding values obtained using principal component scores (PC_BLUP) as predictors, or SNP genotypes with BLUP (SNP_BLUP) or Bayes A (BAYES_A) methods, or pedigree indexes (PI) and polygenic EBV for different scenarios in the Brown breed[1]

	Estimation method			
Scenario/trait	PC_BLUP	SNP_BLUP	BAYES_A	PI
Ref:Pred 70:30				
Milk yield	4.5	1.6	3.6	4.6
Fat yield	9.3	6.0	9.9	5.7
Protein yield	2.7	1.1	2.5	3.5
SCC	13.9	13.2	13.4	12.5
Fat percentage	35.1	30.4	35.2	25.6
Protein percentage	38.4	30.5	34.9	29.8
Ref:Pred 80:20				
Milk yield	9.0	4.6	7.8	8.6
Fat yield	9.7	8.1	10.4	6.3
Protein yield	2.4	1.0	2.3	2.2
SCC	11.7	11.2	10.9	9.7
Fat percentage	38.5	34.4	36.7	26.7
Protein percentage	34.2	28.8	30.6	24.5
Ref:Pred 90:10				
Milk yield	12.3	7.1	6.6	6.6
Fat yield	22.9	19.2	18.4	8.3
Protein yield	12.6	3.5	2.9	0.4
SCC	21.0	22.0	19.8	20.9
Fat percentage	36.7	34.1	34.5	28.4
Protein percentage	37.6	26.3	27.1	20.4

[1]Ref:Pred scenarios = ratio between number of animals included in the reference and prediction populations, respectively.

Table 3. Squared correlations between genomic breeding values obtained using principal component scores (PC_BLUP) as predictors, or SNP genotypes with a BLUP (SNP_BLUP) or Bayes A (BAYES_A) methods, or pedigree indexes (PI), and polygenic EBV for different scenarios in the Simmental breed[1]

	Estimation method			
Scenario/trait	PC_BLUP	SNP_BLUP	BAYES_A	PI
Ref:Pred 70:30				
Milk yield	36.6	35.4	35.8	34.5
Fat yield	34.3	33.8	33.9	33.3
Protein yield	35.3	34.1	34.4	34.1
SCC	20.1	20.4	20.3	20.5
Fat percentage	14.8	15.0	14.7	15.4
Protein percentage	20.2	19.0	19.4	21.0
Ref:Pred 80:20				
Milk yield	36.7	35.3	35.7	33.1
Fat yield	31.2	30.0	30.3	28.8
Protein yield	33.0	30.6	31.0	30.5
SCC	20.3	20.5	20.6	20.6
Fat percentage	12.7	14.9	14.1	15.9
Protein percentage	17.9	16.5	17.4	16.9
Ref:Pred 90:10				
Milk yield	36.6	30.4	31.8	24.8
Fat yield	29.4	27.3	27.8	23.4
Protein yield	32.7	24.0	25.1	20.5
SCC	18.2	18.3	17.8	18.2
Fat percentage	5.2	6.0	5.5	7.0
Protein percentage	11.9	15.2	13.3	15.0

[1]Ref:Pred scenarios = ratio between number of animals included in the reference and prediction populations, respectively.

prediction accuracy for SCS in a study on 1,188 Brown Swiss bulls. These authors observed higher values for yield traits. Accuracies for protein percentages reported in Table 2 agree with results obtained in Australian Holsteins and Jerseys using different approaches and a comparable size of reference population (Hayes et al., 2009b; Moser et al., 2009). The best results in genomic predictions for protein percentage were observed in US Holsteins (VanRaden et al., 2009).

The $R^2_{EBV,GEBV}$ obtained for the Simmental bulls ranged from 0.05 to 0.37 (Table 3). Values for milk yield were, on average, about 5 times that of the Brown breed (0.35 across all scenarios and methods). Yield traits had higher values compared with composition traits. For some scenarios, squared correlations for protein yield were similar to those recently reported for Fleckvieh cattle (Gredler et al., 2010). Intermediate values were obtained for SCS (0.20 on average). The PC_BLUP and BAYES_A methods slightly outperformed the SNP_BLUP approach. As in the case of Italian Brown cattle, PC_BLUP gave slightly larger values than BAYES_A for yield traits and smaller for composition traits, respectively. Other than values for fat and protein percentages, $R^2_{EBV,GEBV}$ were higher than $R^2_{EBV,PI}$ for all estimation methods.

Regression coefficients of EBV on GEBV (Table 4) showed variability across breeds, methods, and traits. Differences between breeds were evident for yield traits, with lower values for Brown bulls. For these traits, regression slopes were close to unity for all 3 methods and for all scenarios in the Simmental breed. For composition traits and SCS, regression coefficients were <1, indicating underprediction of EBV for high values and overprediction for low values. The opposite behavior was observed for Brown. The PC_BLUP method showed the lowest variability across traits.

DISCUSSION

In this paper, GEBV for some dairy traits were estimated by reducing the dimensionality of predictors with PCA. Such a reduction aimed at simplifying data handling and reducing computational burdens while retaining most of the information. The PCA approach was compared with two of the most popular methods used to predict GEBV—BLUP regression and Bayes A—that directly fit all marker genotypes available but with different theoretical assumptions on the distribution of their effects.

The BLUP methodology overcomes the problem of degrees of freedom in the estimation step by fitting SNP effects as random rather than fixed (Meuwissen et al., 2001; Muir, 2007). However, the curse of dimensionality still represents the most important theoretical constraint for GS implementation. This problem is enhanced when a small number of genotyped animals

Table 4. Regression coefficients of polygenic breeding values on genomic breeding values ($b_{EBV,GEBV}$) or pedigree index ($b_{EBV,PI}$) for dairy traits in Brown and Simmental prediction animals using principal component scores (PC_BLUP), SNP genotypes (SNP_BLUP), or Bayes A (BAYES_A) estimation methods for different proportions of reference and prediction population size (70:30, 80:30, or 90:10)[1]

		Brown			Simmental		
Trait	Method	70:30	80:20	90:10	70:30	80:20	90:10
Milk yield	PC_BLUP	0.49	0.66	0.86	1.09	1.00	0.96
	SNP_BLUP	0.26	0.45	0.59	1.12	1.10	1.01
	BAYES_ A	0.47	0.71	0.70	1.12	1.06	1.04
	PI	0.31	0.44	0.41	0.91	0.88	0.73
Fat yield	PC_BLUP	0.80	0.83	1.26	1.05	1.06	1.20
	SNP_BLUP	0.56	0.66	1.00	1.09	1.11	1.38
	BAYES_ A	0.93	0.99	1.34	1.09	1.11	1.38
	PI	0.39	0.43	0.48	0.93	0.94	1.05
Protein yield	PC_BLUP	0.42	0.41	1.01	1.00	0.99	1.10
	SNP_BLUP	0.22	0.23	0.47	1.02	0.99	1.04
	BAYES_ A	0.43	0.44	0.62	1.04	0.99	1.07
	PI	0.29	0.25	0.13	0.87	0.85	0.79
SCS	PC_BLUP	2.27	2.17	2.53	0.73	0.73	0.83
	SNP_BLUP	1.95	1.86	2.28	0.78	0.77	0.88
	BAYES_ A	2.28	2.15	2.57	0.78	0.77	0.87
	PI	0.80	0.73	0.94	0.73	0.72	0.81
Fat percentage	PC_BLUP	1.33	1.35	1.48	0.59	0.64	0.47
	SNP_BLUP	1.20	1.31	1.29	0.65	0.65	0.59
	BAYES_ A	1.46	1.54	1.46	0.64	0.64	0.56
	PI	0.78	0.80	0.80	0.53	0.54	0.46
Protein percentage	PC_BLUP	1.29	1.18	1.45	0.88	0.93	0.72
	SNP_BLUP	1.13	1.18	1.21	0.96	0.88	0.89
	BAYES_A	1.33	1.32	1.32	0.96	0.91	0.85
	PI	0.81	0.76	0.77	0.83	0.73	0.68

[1]Regression coefficients of polygenic breeding values and pedigree index (PI) are also reported.

is available, as in this study. In fact, PCA does not completely address such an issue because of the data structure. The SNP correlation matrix is singular and therefore the number of eigenvalues different from zero is equal to the number of animals (i.e., the rows) minus 1 (Bumb, 1982; Patterson et al., 2006; Dimauro et al., 2011). In this study, PCA was carried out separately for each chromosome. At this level, the gap between predictors and observations was reduced and the number of components retained per chromosome (on average 75 and 120 in Brown and Simmental, respectively) was markedly smaller than the number of markers and number of animals.

In agreement with previous findings on both simulated and real data, PCA was able to efficiently describe the correlation matrix of SNP genotypes (80% of explained variance) with fewer than 10% of the original variables. Such a reduction had a straightforward effect on calculation time. The PC_BLUP approach required about 2 min using a personal computer with a 2.33-GHz quad core processor and 3.25 Gb of RAM. On the other hand, 6 to 9 h was needed, on average, for the SNP_BLUP and Bayes_A approaches, using a Linux server with 4 × 4 quad core processors and 128 Gb of RAM. The PCA required approximately 30 min, but it had to be done just once at the beginning of the work. Although calculation speed is not usually considered a technical priority for GS (compared with genotyping costs, for example), it is likely to become more relevant with the recent development of a larger (800K) SNP platform and the upcoming very low cost sequencing technologies (Van Raden et al., 2011).

Of great interest is that such a huge reduction of calculation time did not result in a lower value of squared correlations between GEBV and polygenic EBV. The similarity of results between the PC_BLUP approach and the other 2 methods considered in the present paper confirms previous findings obtained with another multivariate dimension reduction technique, partial least squares regression (Moser et al., 2009, 2010; Long et al., 2011). The reduction in predictor dimensionality obtained by selecting subsets of SNP based on their chromosomal location or on their relevance to the trait usually results in a decrease of GEBV accuracy (VanRaden et al., 2009; Vazquez et al., 2010). Compared with subset SNP selection, the multivariate reduction has the advantages of not discarding any marker and of using uncorrelated predictors. The latter feature is confirmed by the observed lower bias of the PCA method compared with the SNP_BLUP method.

The similar results obtained when using methods characterized by different theoretical foundations indi-

cates the need for further considerations. The BLUP assumption of an equal effect of all markers on the variance of the trait is commonly considered inadequate to fit the assessed distribution of QTL; that is, many loci with a small effect and a few with large effects (Hayes and Goddard, 2001). On the other hand, the superiority of the Bayesian approach that fits heterogeneous variances across chromosome segments is marked in simulations but not in real data (Hayes et al., 2009a; Moser et al., 2009; VanRaden et al., 2009). Genome-wide association studies on human height suggest that genetic variation is explained by many loci with small additive effects (Yang et al., 2010). Moreover, a superior predicting ability of GEBV for models that assume a heavy-tailed distribution of gene effects compared with finite locus models has recently been reported (Cole et al., 2009). Thus, BLUP methodology, even though not very accurate in terms of description of gene effect distribution, may offer robust DGV estimates (Goddard, 2009) with reasonable accuracy.

A possible criticism of the use of PCA is the lack of biological meaning in the extracted variables. Such a feature is in contrast to the general aims of the use of molecular markers in animal breeding; that is, overcoming the "black box" approach of traditional quantitative genetics. However, even though a clear interpretation based on eigenvectors is not feasible, some results obtained in this work are worth mentioning. The extracted PC scores were able to cluster animals of the 2 breeds, confirming the ability of this technique to capture genetic variation across and within populations, as has been highlighted in human genetic studies (Cavalli-Sforza and Feldman, 2003; Price et al., 2006; Paschou et al., 2007). Moreover, evidence was found of a relationship between one of the extracted PC and average individual heterozygosity. It is interesting to note that, in the case reported for BTA6, it was not the first extracted component that showed a relationship with heterozygosity but the third one. This is a distinguishing feature of PCA: the first extracted component seldom contains biologically relevant information, whereas it may be retrieved in components associated with smaller eigenvalues (Jombart et al., 2009).

In general, $R^2_{EBV,GEBV}$ were rather low, as expected because of the reduced size of the sample of bulls considered and their distribution across years of birth. In the Brown breed, composition traits showed larger values of squared correlations compared with yield traits. These results, in agreement with previous findings (Hayes et al., 2009a; VanRaden et al., 2009), may reflect some variation in the genetic determinism of the trait (Cole et al., 2009). In particular, genes with large effects for fat and protein percentages have been discovered (Grisart et al., 2002; Cohen-Zinder et al., 2005; Cole et al., 2009). Thus, considering that genomic predictions work by tracking the inheritance of causal mutations (VanRaden et al., 2009), the method may be more efficient for traits in which few loci affect a large proportion of the genetic variance.

Observed R^2 of genomic predictions were similar to or slightly higher than those of traditional pedigree indexes, except for fat and protein percentages for Simmental bulls. Even though genomic predictions have been reported to be more accurate than PI (VanRaden et al., 2009; de los Campos et al., 2010; Olson et al., 2011), these are expected results considering the limited size of the populations considered in this study.

Squared correlations were characterized by a relevant variation both within and between breeds. In particular, the Brown breed showed a higher variation in $R^2_{EBV,GEBV}$ across traits compared with the Simmental. Differences in genomic accuracies between traits have been reported in other papers (Hayes et al., 2009a; VanRaden et al., 2009; Su et al., 2010), although not of this magnitude. Moreover, most of the literature deals with Holstein cattle. Apart from the different genetic background of the considered traits, the sample size and the wide range of birth year of bulls can reasonably be considered the main causes of the present results. Reasons for the different behavior of the Simmental breed (less variation between traits, higher values for milk yield) remain unclear. A partial explanation can be found in the pattern of birth year of bulls, which was narrower for Simmental compared with Brown. Moreover, the lower accuracy for fat percentage compared with that in Brown could be ascribed to the known fixation of the favorable mutation at the acylCoA-diacylglycerol-acyltransferase 1 (*DGAT1*) locus in the Italian Simmental.

CONCLUSIONS

Principal component analysis was effective in reducing the number of predictors needed for calculating genomic breeding values for dairy traits in Brown and Simmental bulls. Such a reduction did not affect GEBV precision and allowed for a relevant decrease in calculation time. The obtained accuracies of squared correlations, although moderate to low mainly due to the number of animals considered, were of the same order or slightly higher than those of the traditional pedigree index. Moreover, some differences between traits and breeds were highlighted. Results of the present work suggest the PCA approach as a possible alternative to the use of SNP genotypes for predicting GEBV, especially for populations of limited size.

ACKNOWLEDGMENTS

This research was funded by the Italian Ministry of Agriculture (grant SELMOL).

REFERENCES

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

Boichard, D., V. Ducrocq, S. Fritz, and J. J. Colleau. 2010. Where is dairy cattle breeding going? A vision of the future. Interbull Workshop on the Use of Genomic Information in Genetic Evaluations. Paris, France. Interbull, Uppsala, Sweden.

Bolormaa, S., J. E. Pryce, B. J. Hayes, and M. E. Goddard. 2010. Multivariate analysis of a genome-wide association study in dairy cattle. J. Dairy Sci. 93:3818–3833.

Bumb, B. 1982. Factor analysis and development. J. Dev. Econ. 11:109–112.

Burden, R. L., and J. D. Faires. 2005. Numerical Analysis. Thomson Brooks/Cole, Belmont, CA.

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

Cohen-Zinder, M., E. Seroussi, D. M. Larkin, J. J. Loor, A. Everts-van der Wind, J. H. Lee, J. K. Drackley, M. R. Band, A. G. Hernandez, M. Shani, H. A. Lewin, J. I. Weller, and M. Ron. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 15:936–944.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92:2931–2946.

de los Campos, G., D. Gianola, and D. B. Allison. 2010. Predicting genetic predisposition in humans: The promise of whole-genome markers. Nat. Rev. Genet. 11:880–886.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553.

Dimauro, C., M. Cellesi, M. A. Pintus, and N. P. P. Macciotta. 2011. The impact of the rank of marker variance–covariance matrix in principal component evaluation for genomic selection applications. J. Anim. Breed. Genet. 128:440–445.

Ducrocq, V., and Z. Liu. 2009. Combining genomic and classical information in national BLUP evaluations. Interbull Bull. 40:172–177.

Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776.

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

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257.

Gredler, B., H. Schwarzenbacher, C. Egger-Danner, C. Fuerst, R. Emmerling, and J. Sölkner. 2010. Accuracy of genomic selection in dual purpose Fleckvieh cattle using three types of methods and phenotypes. Article n. 0907 in Proc. 9th World Congr. Genet. Appl. Livest. Prod. Eventlab, Leipzig, Germany.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222–231.

Guo, G., M. Lund, Y. Zhang, and G. Su. 2010. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. J. Anim. Breed. Genet. 127:423–432.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5.

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

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy of genomic breeding values in multibreed dairy cattle populations. Genet. Sel. Evol. 41:51.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443.

Henderson, C. R. 1985. Best linear unbiased prediction using relationship matrices derived from selected base populations. J. Dairy Sci. 68:443–448.

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

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

Legarra, A., and I. Misztal. 2008. Technical note: Computing strategies in genome-wide selection. J. Dairy Sci. 91:360–366.

Long, N., D. Gianola, G. J. M. Rosa, and K. A. Weigel. 2011. Dimension reduction and variable selection for genomic selection: Application to predicting milk yield in Holsteins. J. Anim. Breed. Genet. 128:247–257.

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

Macciotta, N. P. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro, C. Pieramati, and A. Cappio-Borlino. 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. J. Dairy Sci. 93:2765–2774.

Meuwissen, T. H. E., B. J. Hayes, and 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, and H. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

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

Olson, K. M., P. M. VanRaden, M. E. Tooker, and T. A. Cooper. 2011. Differences among methods to validate genomic evaluations for dairy cattle. J. Dairy Sci. 94:2613–2620.

Paschou, P., E. Ziv, E. G. Burchard, S. Chloudry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. 2007. PCA-correlated SNPs for structure identification in world-wide human populations. PLoS Genet. 3:e160.

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.

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

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

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

Su, G., B. Guldbrandtsen, V. R. Gregersen, and M. S. Lund. 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J. Dairy Sci. 93:1175–1183.

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W.

C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247–252.
VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.
VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. Weigel. 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43:10.
VanRaden, P. M., and P. G. Sullivan. 2010. International genomic evaluation methods for dairy cattle. Genet. Sel. Evol. 42:7.
VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.
Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–5949.
Wiggans, G. R., T. S. Sonstegard, P. M. Vanraden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431–3436.
Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569.