International Genomic Evaluations for Young Bulls

P.G. Sullivan[1], B. Zumbach[2], J.W. Durr[2], J.H. Jakobsen[2]
[1]Canadian Diary Network, Guelph, ON, Canada
[2]Interbull Center, Dept of Animal Breeding and Genetics, SLU, S-750 07 Uppsala, Sweden

Abstract

Methods were developed to combine MACE solutions of proven bulls with individual GMACE equations for young bulls, to eliminate a double-counting of genomic information previously observed when applying population-based GMACE equations. With the new methods, two options were considered for variance parameters; the same MACE variances for both proven and young bulls, or separate variances for young bulls, estimated from GEBV data. Cross-validation was used to compare how well excluded local GEBV could be predicted from foreign data by each method. All validation results for comparisons among young genomically-tested bulls favoured the new methods and separate variances. Across all traits and scales of evaluation, the correlation with local GEBV increased from .86 for population-based GMACE to .90 for the new methods with separate variances, average absolute differences decreased from 0.13 to 0.06 and root mean squared differences decreased from 0.39 to 0.29. Errors of regression also decreased from 0.17 to 0.09, indicating less bias in comparisons of top versus average young bulls and likely more consistent ranking of top young versus top proven bulls. Approximate reliabilities, and variances of international evaluations were lower with the new methods, as expected with less double-counting of information.

Key words: genomics, international evaluation, MACE, GMACE

Introduction

Residual correlations among national genomic evaluations (GEBV) can be approximated as a function of genetic correlations, data shared by countries, and relative amounts of traditional data (EDC) and genomic data (GEDC) in the GEBV of individual bulls (Van Raden and Sullivan, 2010). The purpose of fitting residual correlations in GMACE is to avoid double-counting shared data at both the national and international levels. The GMACE approach performed well with ideal simulated data, but is expected to have problems in practice, as outlined by Sullivan and VanRaden (2010) and demonstrated by Sullivan (2011). Currently, there are at least two key problems.

The first problem is that upward bias of international reliability (double-counting) is not avoided when data are unbalanced among countries for a given bull. This is mainly a problem for proven bulls, with large EDC in only 1 or few countries, and is less of a concern for young genomically-tested bulls, with 0 EDC in all countries. Since international MACE evaluations are already available for proven bulls, and GMACE results have not been shown superior to MACE for proven bulls, the focus of the present study was on international genomic evaluations for young bulls only. From a commercial point of view the young genotyped bulls with no progeny are of primary interest for international genomic selection.

The second problem is that residual correlations exist at the family level, for example when sons are genotyped in multiple countries that share data for genomics, but residual correlations are only fit when the bull is itself genotyped in multiple countries. In initial efforts to implement an international genomic evaluation service (Zumbach *et al.*, 2011) input data were limited to a single GEBV per bull, in order to avoid the first problem above. However, this approach did not address the family-based residual correlations, which were confirmed to cause problems in the results from that study. VanRaden (2011) suggested a partitioning method to remove the upward bias noted in the GMACE approximate reliabilities. However, the bias in reliabilities is caused by a double-counting problem for the international GEBVs. Therefore, an objective of the present study was to extend the partitioning idea to the international GEBVs as well as the reliabilities,

by developing procedures for conversion of proofs for young bulls. Other objectives were to apply alternative procedures to traits with high and low correlations and heritabilities, and to perform model validations in order to recommend the best approach for a GMACE service.

Data

A data call for participation in the present study was sent to national evaluation centers passing the genomic validation test for protein (Nilforooshan *et al.*, 2011). Countries were invited to send data for five traits: protein (pro), stature (sta), somatic cells (scs), direct longevity (dlo) and female fertility (cow conception 1; cc1). Seven countries (CAN, DEU, DFS, FRA, NLD, POL, USA) provided GEBV data.

The data were edited to include only GEBVs of bulls born since 2006 with no progeny test, and for all other bulls the conventional EBVs that would normally be included in MACE, per Interbull Code of Practice (www.interbull.org). The EBVs included were the same as used for the Interbull routine genetic evaluation in April 2011 for all countries and traits except France, which sent new conventional data for cc1 and dlo. Numbers of GEBV and EBV records used for the present study are listed in Table 1, and ranges of heritabilities and correlations among countries are in Table 2.

Sire-dam pedigree was extracted from the Interbull database. Missing and conflicting birth years were resolved and pedigree was traced as far back as possible starting from animals with proofs for each of the five traits studied. Animals born before 1960 were set to missing, as were parents with unknown sire and dam and only one progeny, but keeping the information about breed, country and sex for the assigning of phantom parent groups. This resulted in 349716, 304598, 342708, 305404, and 254190 pedigree animals for pro, sta, scs, dlo and cc1, respectively.

Methods

In the previous GMACE pilot study, GEBV were included for all bulls, both young and proven, and rules were needed to select a single GEBV per bull (Zumbach *et al.*, 2011). In the present study, we allowed multiple GEBV per bull but restricted GEBV data to only the young genotyped bulls with no progeny test. For young bulls, the GMACE methods can work reasonably well to limit double-counting at the individual level.

In order to limit double-counting at the individual level it was assumed that data sharing was complete between countries in a group that share data for genomics (c_{ij}=100% within group). Two groups were identified; [CAN, USA] and [DEU, DFS, FRA, NLD, POL]. Between countries that do not share data a partial independence of residuals was assumed, to account for the fact that genomic evaluations based on 50K SNP estimates do not account for 100% of polygenic variance (~15% common residual error for the same proportion of polygenic variance not explained in all countries), and also because total information in GMACE is often inflated when input data (i.e. GEDC) are variable or inconsistent among countries (c_{ij} increased from 15% to 25% between groups).

The following new methods were implemented to avoid double-counting genomic information at the family level.

Genomic MACE Conversions (MCNV)

For each animal, a set of Mendelian Sampling (MS) mixed-model equations was set up (Lu=r), with u representing the vector of MS values for all countries. For the available GEBV, MS was calculated as GEBV – MACE parent average (PA), otherwise MS was MACE solution – MACE PA. Given the left-hand-sides ($L=[D+G^{-1}]$) and corresponding solutions (u), the right-hand-sides were derived (r=Lu), and noting that r=Dy, pseudo observations were

derived ($y=D^{-1}r$). Matrix D included traditional EDC and genomic GEDC data for the bull, and in this case EDC=0 for the young bulls. Finally, matrix D was replaced with matrix E* as described in Sullivan and VanRaden (2010), and the modified equations were solved to derive international MS estimates for all countries, to which we added back the MACE PA that were originally subtracted.

Reliability of MCNV included effective record contributions from both the MACE PA and the extra information added in from genomics (GEDC). Total effective records in each country were included in a set of mixed model LHS equations that were inverted to derive multivariate international genomic reliabilities for all countries.

Genomic Variances for Conversions (VCNV)

Countries attempt to generate GEBV that are directly comparable to EBV of non-genotyped bulls. However, different methods and assumptions for genomic prediction, and for deriving GEBV as a combination of direct genomic values (DGV) and EBV, can lead to different GEBV variances in each country. Additionally, the relative consistency between GEBV and MACE PA in each country can affect the variance of genomic MS deviations. To estimate and account for these differences, genomic variances were estimated as is routinely done for genetic or sire variances in regular MACE (Sullivan, 1999). Inputs required were MS estimates and prediction error variance (PEV) of MS. The latter term is a quadratic function of the PEV matrix for animal, sire and dam, which was taken as the corresponding matrix from MACE reliability approximation within the software (i.e. the inverted LHS after absorbing all other relatives), with rows and columns rescaled by the relative change in PEV due to additional GEDC. The relative change due to GEDC was the ratio of animal PEV that corresponds with total effective records (MACE+GEDC) divided by animal PEV that corresponds with effective records from only MACE.

Validation of Methods

Bulls with national GEBV from multiple countries were used to test each method of international genomic evaluation, using cross-validation. Assessing one country scale at a time, all local GEBV were deleted and then GMACE, MCNV and VCNV were applied to the local EBV plus the EBV and GEBV data from foreign scales to see how well the local (deleted) GEBV could be predicted. Correlations, regressions and the total of mean squared errors and bias were used as criteria to assess model performance. Approximate reliabilities were also compared for the 3 international approaches. All input data were standardized to a mean of 0 and variance of 1 prior to the international evaluations.

Results and Discussion

As should be expected, estimates of genomic variance were generally close to the traditional genetic variances used in MACE, although in many cases the genomic estimates were a bit higher (genetic SD ratios greater than 1 in Table 3). Deviations from an SD ratio of 1 were smallest for pro and scs, the traits with most data. Higher estimates for genomic relative to traditional genetic variance is expected if GEBV for top young bulls are inflated relative to EBV of top proven bulls. Research that was conducted due to concerns about this in Canada led to a unique and relatively conservative national genomic evaluation system (Sullivan, 2009), which coincides with estimated genetic SD ratios being lowest for Canada in the present study. In France, genomic evaluations may be more conservative than in other countries due to a potentially stronger reliance on the national EBV system relative to estimates of SNP effects (Ducrocq *et al.*, 2009; Ducrocq and Patry, 2010), and hence the genetic SD ratios were generally closer to 1 for France than for any other country in the present study.

Two extreme estimates were observed, for the fertility trait cc1 in DFS and for stature in

POL, where genomic variance was 10 times the traditional genetic variance used for MACE (SD ratios of 3.28 and 3.17). In both of these cases the sample size was small and there were fewer GEBV available for these traits relative to others in these 2 countries. Although we have not investigated these 2 situations in detail, selection bias may have been an issue, and large sampling errors were expected due to the few GEBV records available. Extreme genomic variance estimates such as these could be a problem for routine applications by Interbull. Possible solutions are to impose constraints that limit how much genomic variance estimates can deviate from MACE variance, to develop validation tests related to genomic variance estimation and/or to impose minimum data requirements that countries would need to pass in order to participate in the genomic evaluation service.

Validation results for the alternative approaches are compared in Tables 4 and 5. For all comparison statistics, the genomic conversion method using estimated genomic variances (VCNV) out-performed both GMACE and the genomic conversion method using MACE variances (MCNV). International predictions were more highly correlated with the (deleted) local GEBV, regressions of local on predicted were generally closer to 1, biases of prediction were smaller (average absolute differences were closer to zero) and the square root of mean squared differences were closer to zero. The advantages of VCNV over GMACE were largest for fertility (cc1) and survival (dlo), the traits with lowest heritability and for which fewer data were available. Both MCNV and VCNV were much better than GMACE for these two traits.

The VCNV approach was generally better than MCNV for all traits, and for most but not all combinations of trait by country. The advantages of VCNV are dependent on good estimates of genomic variance. In the present study almost all of these estimates were in a reasonable range relative to MACE variances.

Reliabilities are presented in Table 6 for GEBV predictions based on foreign data, relative to the local GEBV. It is unlikely that predictions from foreign data would achieve reliabilities as high as for local estimates, but with GMACE this was quite common if GEBV for a bull were available from more than one foreign country. The GMACE reliabilities are biased upwards due to ignored residual correlations at the family level. Reliabilities were lower (and the same) for the two conversion methods MCNV and VCNV because these methods do not accumulate or double-count genomic information through relatives. The difference between conversion-based and GMACE reliabilities reflects the amount of bias in GMACE from ignoring the family-based residual correlations.

For comparison purposes, Table 6 also includes expected reliabilities from applying simple conversion equations (SCNV) to GEBV instead of converting MS deviations and adding back the parent averages (VCNV). Reliabilities were consistently lower for SCNV. For predictions from multiple foreign GEBVs, multivariate regression equations were used for SCNV, similar to the VCNV method but without the inclusion of PA contributions to total effective records if the bull did not have a national GEBV. For all methods, the reliability of predicted GEBVs increased in a similar way as the number of foreign GEBVs increased. This was expected, because if a bull had multiple foreign GEBVs, they were usually from both groups of data-sharing countries in the present study (North America and Europe), and all methods assumed the genomic data was partially (75%) independent between these groups of countries.

Recommendations

Method VCNV should be used as the first choice. However, the tests presented here should be repeated with an expanded call for data. In this study, not all available GEBVs were provided by all countries, which could affect the estimates of genomic variances. The number of bulls with multiple GEBV was also smaller than necessary, reducing the power of these tests. Additional research should focus on the potential for selection bias effects in the genomic variance estimates, perhaps by studying the distributions of genomic MS deviations from each country. Benefits of constraining genomic variance estimates to be relatively similar to MACE variances should be investigated, and also the possible need for new validation tests related to genomic variance

estimation and/or minimum data requirements to participate in VCNV.

Addressing the questions above should take higher priority than further development of a full-scale GMACE system, which could use as input national GEBV for all bulls (young and proven).

Acknowledgements

Helpful discussions with Gerrit Kistemaker were appreciated.

Ducrocq, V., Fritz, S., Guillaume, F. & Boichard, D. 2009. French report on the use of genomic evaluation. *Interbull Bulletin 39,* 17-22.

Ducrocq, V. & Patry, C. 2010. Combining genomic and classical information in national BLUP evaluation to reduce bias due to genomic pre-selection. *Interbull Bulletin 41,* 33-36.

Nilforooshan, M.A., Zumbach, B., Jakobsen, J., Loberg, A., Jorjani, H. & Dürr, J. 2011. Validation of national genomic evaluations. *Interbull Bulletin 42,* 56-61.

Sullivan, P.G. 1999. REML estimation of heterogeneous sire (co)variances for MACE. *Interbull Bulletin 22,* 146-148

Sullivan, P.G. 2009. Options for combining direct genomic and progeny-test results. *Dairy Cattle Breeding and Genetics Committee Meeting.* Guelph. Oct. http://cgil.uoguelph.ca/dcbgc/Agenda0910/agenda0910.htm

Sullivan, P.G. 2011. Accounting for residual correlations among regional genomic predictions via GMACE. Interbull Workshop, Feb. 27-28. Guelph, Canada. *Interbull Bulletin 43,* 16-21.

Sullivan, P.G. & VanRaden, P.M. 2010. GMACE implementation. *Interbull Bulletin 41,* 3-7.

VanRaden, P.M. 2011. Personal communication.

VanRaden, P.M. & Sullivan, P.G. 2010. International genomic evaluation methods for dairy cattle. *Gen. Sel. Evol. 42,* 7.

Zumbach, B. Jakobsen, J., Forabosco, F., Jorjani H. & Dürr, J. 2011. Data selection and pilot run on Simplified Genomic MACE (S-GMACE). Interbull Workshop, Feb. 27-28. Guelph, Canada. *Interbull Bulletin 43,* 11-18.

Table 1. Number of genomic (GEBV) and conventional EBV records for protein yield, stature, somatic cell count, direct longevity and female fertility (cow conception one; cc1).

Country	Protein		Stature		Somatic Cell		Direct Longevity		Fertility (cc1)	
	GEBV	EBV	GEBV	EBV	GEBV	EBV	GEBV	EBV	GEBV	EBV
AUS		5918		2924		5921		5889		
BEL		821		769		771		698		
CAN	11372	8730		8014	11399	8710		8537		5851
CHE		927		925		1067		1062		973
CHR		1554		1369		1597		1421		1333
CZE		2922		2625		2537		3174		2429
DEU	11481	20943	11486	19261	11336	20942	11481	18860	11481	20001
DFS	1168	10204	1149	9498	1166	10522	1343	9209	764	10274
ESP		2124		2032		2139		2071		
EST		695		324		688				
FRA	6051	13500	5936	13162	6051	13526	6299	12981	6238	12382
FRR		165		168		183				
GBR		5214		4729		4854		5345		4824
HUN		2348		1779		1879		2361		
IRL		1542				1518		1672		
ISR		975				967		960		940
ITA		8022		7306		8148		8019		7538
JPN		4055		3804		4108				
LTU		416				405				
LVA		528				393				
NLD	3795	11999	3712	11512	2883	12189	3716	11348	3794	11491
NZL		5561		4368		5540		5425		
POL	337	6334	244	5312	336	5295				4567
PRT		1682				1551				
SVK		807				780				
SVN		275								
USA	1102	28175		24785	653	27993	653	27029		12599
ZAF		1106		675		995				
Total	35306	147542	22527	125341	33824	145218	23492	126061	22277	95202

Table 2. Ranges of heritabilities and correlations for protein yield, stature, somatic cell count, direct longevity, and female fertility (cow conception one; cc1).

Trait	Range of heritabilities	Range of correlations	No of countries
Protein Yield (pro)	0.136 – 0.508	0.751 – 0.949	28
Stature (sta)	0.370 - 0.630	0.697 – 0.991	21
Somatic Cell Count (scs)	0.062 – 0.433	0.753 - 0.972	27
Direct Longevity (dlo)	0.016 – 0.223	0.299 – 0.934	18
Female Fertility (cc1)	0.010 – 0.067	0.517 – 0.961	13

Table 3. Ratio of genetic standard deviation (SD) estimates (genomic / traditional), and in parentheses the number of GEBV used to estimate genomic SD, for all traits (defined in the text).

Country	cc1		dlo		pro		scs		sta	
CAN					0.91	(10594)	0.83	(10621)		
DEU	1.08	(11340)	0.89	(11341)	1.17	(11350)	1.21	(11207)	1.14	(11345)
DFS	3.28	(764)	1.89	(1343)	1.10	(1167)	1.26	(1166)	1.11	(1149)
FRA	0.92	(1374)	1.28	(1435)	0.98	(1189)	1.04	(1189)	1.09	(1073)
NLD	0.61	(3790)	0.75	(3712)	1.07	(3791)	1.10	(2880)	1.10	(3708)
USA			1.26	(542)	1.19	(1025)	1.04	(582)		
POL					1.74	(132)	1.16	(132)	3.17	(89)

Table 4. Average correlation ($\frac{1}{\sum n}\sum\{n * r_{\text{Obs,Pred}}\}$), and error from expected regression ($\sqrt{\frac{1}{\sum n}\sum\{n*(b_{\text{Obs,Pred}}-1)^2\}}$), of local GEBV (Obs) versus prediction (Pred) from a single foreign GEBV.

		Correlation			Regression Error		
Trait[z]	n	GMACE	MCNV	VCNV	GMACE	MCNV	VCNV
cc1	249	0.62	0.88	0.89	0.58	0.20	0.12
dlo	260	0.42	0.81	0.82	0.74	0.19	0.19
pro	2572	0.90	0.90	0.90	0.13	0.15	0.10
scs	1734	0.91	0.91	0.92	0.08	0.15	0.06
sta	274	0.83	0.82	0.86	0.24	0.26	0.13
all	5089	0.86	0.89	0.90	0.17	0.16	0.09

[z]as defined in the text.

Table 5. Average absolute difference ($\frac{1}{n}\sum\sqrt{(\text{Obs}-\text{Pred})^2}$) and root mean squared difference ($\sqrt{\frac{1}{n}\sum(\text{Obs}-\text{Pred})^2}$) between local GEBV (Obs) and predictions from a single foreign GEBV (Pred).

		Average Absolute Difference			Root Mean Squared Difference		
Trait[z]	n	GMACE	MCNV	VCNV	GMACE	MCNV	VCNV
cc1	249	0.76	0.08	0.07	1.19	0.44	0.40
dlo	260	0.65	0.21	0.22	1.21	0.45	0.44
pro	2572	0.07	0.09	0.06	0.27	0.27	0.25
scs	1734	0.04	0.02	0.01	0.34	0.36	0.31
sta	274	0.15	0.16	0.13	0.39	0.41	0.34
all	5089	0.13	0.07	0.06	0.39	0.32	0.29

[z]as defined in the text.

Table 6. Approximated local GEBV reliabilities, and corresponding reliabilities for international predictions from a single (or multiple) foreign GEBVs (traits defined in the text).

Trait	n	Local	GMACE	MCNV,VCNV	SCNV
cc1	252 (16)	47 (48)	44 (49)	33 (39)	25 (30)
dlo	266 (19)	52 (54)	39 (45)	28 (36)	21 (26)
pro	2572 (135)	73 (72)	66 (71)	63 (67)	58 (63)
scs	1734 (83)	71 (71)	67 (73)	64 (69)	59 (66)
sta	274 (16)	71 (71)	69 (76)	65 (73)	61 (70)
all	5098 (269)	70 (69)	64 (69)	60 (64)	55 (60)