GMACE Implementation

P.G. Sullivan[1] and P.M. VanRaden[2*]
[1]Canadian Dairy Network, Guelph, ON, Canada
[2]Animal Improvement Programs, USDA Agricultural Research Service, Beltsville, MD, USA

Abstract

Programs to compute multi-trait across country evaluations (MACE) were adapted to include national genomic evaluations. Strategies to account for regional genotype sharing were compared using a simulated data set in which two groups of countries shared data within but not between regions. Methods worked well unless countries within a region report different progeny equivalents from the same shared data, but a modified covariance matrix reduced this problem. Gains in reliability were large if young bull genotypes were evaluated in each country or region, but were much smaller if access to genotypes was limited. Exchange of national evaluations for young bulls becomes much more important as reliabilities of genomic evaluations increase.

Key words: genomics, international evaluation, MACE, GMACE

Introduction

A modified MACE model for genomic data (GMACE) was recently presented (Sullivan and VanRaden, 2009). The main difference from regular MACE was to fit residual correlations among the national genomic evaluations of a bull from multiple countries. The residual correlations account for common information that is shared among countries for national genomic predictions, e.g. the overlap of genomic training data if domestic and MACE proofs are used by each country or if the genomic evaluations are derived from a regional data set. Residual correlations could also prevent an over-accumulation of genomic data from multiple countries when genomic predictions (SNP effects) explain less than 100% of the total genetic variance. Major gene tests and low-density SNP panels can explain only a small percentage of the total genetic variance, and even a 50K SNP panel can probably explain no more than 90% of the variance. The accumulation of genomic information across multiple countries would be unlimited with regular MACE, and could theoretically accumulate up to 100% of the genetic variance, which would be incorrect.

The purposes of the present study were a) to test the GMACE model using simulated data and b) to develop software that Interbull could use for a routine international genomic evaluation service.

Methods

The GMACE methodology is described in detail by VanRaden and Sullivan (2010). However, some of the methods have been refined based on new knowledge from the present study, and these refinements are presented below.

Let **D** be a diagonal matrix of within-country residual variances for de-regressed animal EBV ($\mathbf{D} = \left[\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\right]^{-1}$). Matrix **E** is block-diagonal by animal, with all countries included in a block. The diagonals of **E** are the same as **D** while the off-diagonals reflect residual covariances caused by sharing of data for genomics. Let δ represent the progeny equivalents in each average without genomics and δ_g the additional progeny equivalents with genomics included. The additional progeny equivalents δ_g can be calculated a number of different ways, with 3 possibilities described by VanRaden and Sullivan (2010). The MME for GMACE are:

$$(\mathbf{E}^{-1} + \mathbf{A}^{-1} \otimes \mathbf{T}^{-1})\hat{\mathbf{g}} = (\mathbf{E}^{-1})\mathbf{y} \qquad [1]$$

Matrices **A** and **T** contain animal relationships and genetic covariances among traits, respectively. For comparison, the MME for regular MACE are:

$$(\mathbf{D}^{-1} + \mathbf{A}^{-1} \otimes \mathbf{T}^{-1})\hat{\mathbf{g}} = (\mathbf{D}^{-1})\mathbf{y} \qquad [2]$$

For a given animal, residual variance for country *i* (E_i) is equal to $R_i/(\delta + \delta_g)$, and residual covariances between countries (E_{ij}) are a function of the proportion of total progeny equivalents from genomics ($\gamma = \frac{\delta_g}{\delta+\delta_g}$), the amount of shared genomic information (*c*) between countries and the genetic correlation (r_g) between countries:

$$E_{ij} = r_g c\sqrt{\gamma_i \gamma_j E_i E_j}$$

The descriptions above match VanRaden and Sullivan (2010). Results from initial testing of these methods were very promising, but expanded tests revealed undesirable results (shown below) when values for δ_g were not the same in all countries. When there are different priors (input values) of δ_g for a bull, the GMACE equations generate results as though there is some independent genomic data in some or all of the countries, and all of the posterior estimates of δ_g increase accordingly.

Differences in δ_g among countries can be avoided by choosing only 1 value for each bull, to use in the off-diagonals of **E**, for example the maximum value ($\delta_{\max}$). This replaces γ with $\gamma^* = \frac{\delta_g}{\delta+\delta_{\max}}$, re-defining the off-diagonals in **E** as follows:

$$E^*_{ij} = r_g c\sqrt{\gamma^*_i \gamma^*_j E_i E_j}$$

The off-diagonals of **E** and $\mathbf{E}^*$ can be re-written as:

$$E_{ij} = r_g c\sqrt{\gamma_i \gamma_j R_i R_j} \,/\, \sqrt{(\delta_i + \delta_{g_i})(\delta_j + \delta_{g_j})}$$

$$E^*_{ij} = r_g c\sqrt{\gamma_i \gamma_j R_i R_j} \,/\, \sqrt{(\delta_i + \delta_{\max})(\delta_j + \delta_{\max})}$$

Note that if δ_g is the same in all countries then $\mathbf{E}^*$=**E**. Matrix $\mathbf{E}^*$ uses δ_g for variances on the diagonals, but $\delta_{\max}$ as a shared-data value for the covariances.

Selected combinations of δ and δ_g were assumed for a single sire and with complete sharing of data for national genomic evaluations, to compare and understand the implications of using **D** (MACE), **E** (GMACE) or $\mathbf{E}^*$ (GMACE*). The 3 models were also applied to a simulated international population of Brown Swiss data (details in VanRaden and Sullivan, 2010). There were 50K SNP and 10K polygenes simulated for a single trait, and phenotypic measurements in each of 9 countries. Genetic correlations among countries were the recent estimates by Interbull for Protein yield, and genetic variances near unity were assumed.

To test the ability of GMACE to account for sharing of data for genomics, the 9 countries were divided into two regional groups, composed of 4 and 5 countries respectively. Possible inputs to GMACE were thus national EBV, national DGV (sum of SNP effects) or regional DGV. For national DGV, values of δ_g were assumed the same for all genotyped bulls within a country, with country-specific values ranging from 0.25 to 33 (h^2=*1/3*). For regional DGV the country-specific values of δ_g were all increased by 16.5, assuming equal benefits from sharing data for all countries. These are very crude approximations of δ_g, but good for testing GMACE because better approximations may not be available in practice. Data sharing was assumed to be 100% within a region and 0% between regions for all GMACE models, regardless of the input data used. Thus for national DGV data, MACE is a better fit to the data and for regional DGV data, GMACE is a better fit.

Results

Traditional daughter equivalents (δ) in one country will increase δ in other countries through (G)MACE, but δ_g should not increase among the countries if there is complete genomic data sharing. Results of tests for these data patterns are in Table 1. The first scenario is the simplest, as it only involves δ_g (i.e. $\delta = 0$), which should not change in

and out of GMACE in any of the countries. MACE would clearly overestimate δ_g in all countries, by double-counting the correlated information already included in the input δ_g. GMACE only reduces the double-counting problem for countries with smaller input δ_g, and actually increases the problem for countries with larger input δ_g. With either MACE or GMACE, the double-counting problem scales up with the number of countries, and could become a serious concern for the Holstein breed for example, as more countries begin to offer a domestic genomic evaluation service. In contrast, δ_g from GMACE* are much closer to expectation, regardless of the number of countries involved.

The first scenario represents a young genotyped calf or perhaps an embryo. Scenarios 2 and 3 are slightly more complicated, representing a young genotyped sire with 1st crop daughters in a single country (e.g. $\delta = 100$ in either country 1 or country 3). The expectations for scenarios 2 and 3 are that the δ portion should increase for countries without daughters (as happened with regular MACE and conversions to foreign scales, prior to the genomics era). The genomic portion δ_g however, should remain the same in and out of GMACE, and ideally should not affect the amount of correlated information predicted on foreign scales for the 100 traditional daughters (δ). Relative to expectations, observations and conclusions were the same for all 3 scenarios, for MACE, GMACE and GMACE*.

Results in Table 2 show clear advantages over traditional national evaluation systems for genomic, regional and international evaluations. Reliabilities of GMACE* evaluations were nearly as high as from a full-scale global genomic evaluation system, even if assumptions were poor for δ_g or for the extent of data sharing among countries for national genomic evaluations. GMACE was only slightly better than GMACE* with the best input data and assumptions, and was much worse when assumptions were less ideal. It is expected that increasing the value of *c* (for data sharing) should decrease double-counting of information, but often the opposite was observed for GMACE under a number of different scenarios (results not shown). With GMACE*, however, the expected patterns relative to *c* were more consistently observed.

The simple approximations of δ and δ_g in this study may have been too high, which could explain upward bias in approximated reliabilities. Additionally, the approximations rely on an assumption that the EBV are BLUP and that regressions of BV on EBV should therefore equal 1.0, which was not the case for many of the models and data considered. When deviating more from BLUP, the resulting observed reliabilities decreased while approximated reliabilities increased.

These patterns are consistent with results in Table 1, where inflated variance of EBV for GMACE of NG for example, was likely due to double-counting of information. Double-counting will increase the variance of EBV without a corresponding increase in the covariance with BV. The approximate reliabilities can only follow variance of EBV because BV is unknown.

Relative to MACE, the results for GMACE* were more consistent than GMACE. When MACE assumptions were appropriate (i.e. national DGV as input), the MACE results were closer to BLUP, and when GMACE assumptions were appropriate (i.e. regional DGV as input) the GMACE* results were closer to BLUP. In general, the approaches that used suitable methodology and included all international data (e.g. MACE of national EBVs, GMACE or GMACE* of regional DGVs and Global DGV) generated results that were closest to BLUP (regression of BV on EBV closest to 1).

Software

A genomic MACE system was added to software previously provided to Interbull, and used for multiple traits and countries evaluation (MT-MACE) of udder health and fertility traits. The updated software offers sire-mgs and animal model options for GMACE and MT-MACE. Model-specific deregression, evaluation and reliability

approximations are included. The software does not include a variance or covariance estimation feature. Inputs required include national proofs (e.g. EBV or GEBV), weighting factors (δ and δ_g), pedigree that includes user-defined genetic groups, covariance matrices and genomic data-sharing parameters among countries. The plan is to provide updated versions of this software, such that changes to Interbull systems and training needs can be minimize, as improved international evaluation methods become available.

Future Research and Development

This research is ongoing as we have not yet examined potentially complicating factors such as different values for heritability or genetic variance among countries. Genetic correlations were also quite high for this study and GMACE* should be tested for lower correlation structures, as observed for various conformation and fertility traits.

Modeling options not yet considered include:

- Restricting input to a single GEBV per bull in MACE rather than fitting covariances among multiple GEBV with GMACE,
- Regional instead of national de-regression of GEBVs and MT-MACE instead of national deregression and GMACE.
- Use of de-regressed δ_g for variances, but regressed $\delta_{\max}$ for covariances (regression from relatives via matrix **A**).
- GEBV instead of DGV as input to GMACE.
- Robust enhancements to the GMACE model to minimize propagation of detectable bias in national GEBVs to other countries.
- Model adjustments for V(SNP) < V(G).
- Implications when large $\delta_{foreign}$ are included in δ_g because DGV estimates include MACE proofs but EBV in regular MACE do not.
- Optimal choice of data-sharing parameters.

Acknowledgements

Discussion and suggestions from the Interbull genomics task force were much appreciated.

References

Sullivan, P.G. & VanRaden, P.M. 2009. Development of genomic GMACE. *Interbull Bulletin 40,* 157-161.

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

Table 1. Daughter equivalents (DE= $\delta + \delta_g$) in and out of MACE and GMACE for a trait with h^2=*1/3* and all r_g=*.90*. Traditional DE (δ = 0 or 100) were limited to the first 3 countries, while all countries had genomic DE (δ_g = 5, 10, or 20).

Model	3 countries			9 countries				27 countries			
Input DE	5	10	20	5	10	20	5…	5	10	20	5…
MACE	22	25	30	46	50	58	46	73	78	87	73
GMACE	13	19	35	25	35	59	25	52	70	110	52
GMACE*	13	14	20	15	16	20	15	15	16	20	15
Expected	5	10	20	5	10	20	5	5	10	20	5
Input DE	105	10	20	105	10	20	5…	105	10	20	5…
MACE	122	46	54	146	61	70	57	173	81	91	77
GMACE	109	37	58	125	55	87	41	150	83	130	62
GMACE*	109	38	46	109	38	50	36	109	38	52	36
Expected	105	40	50	105	40	50	35	105	40	50	35
Input DE	5	10	120	5	10	120	5…	5	10	120	5…
MACE	40	43	130	56	60	158	56	76	81	187	76
GMACE	33	45	125	42	57	142	42	62	84	172	62
GMACE*	35	37	120	35	37	120	35	35	37	120	35
Expected	35	40	120	35	40	120	35	35	40	120	35

Table 2. Observed reliabilities (empirical squared correlations between simulated true and estimated BV), theoretical reliabilities approximated by GMACE software (extension of Harris and Johnson approximation) and regressions of BV on EBV (Expectation is 100% if BLUP), averaged across 9 countries.

	Observed Reliability			Approximated Reliability			Regression of BV on EBV (*100%)		
Model	Calf[a]	Yng Sire[b]	All[c]	Calf	Yng Sire	All	Calf	Yng Sire	All
Nat EBV (NE)	6	20	19	19	28	24	41	94	90
Nat DGV (NG)	29	30	26	42	46	46	97	105	92
Reg DGV (RG)	52	52	48	65	67	67	97	102	94
MACE of NE	13	61	62	30	70	71	70	94	95
MACE of NG	60	61	60	77	80	80	99	104	107
GMACE of NG	48	50	51	80	82	82	71	83	87
GMACE* of NG	61	62	61	73	78	79	106	106	109
MACE of RG	61	63	62	85	86	86	86	90	92
GMACE of RG	61	65	65	81	84	84	100	101	101
GMACE* of RG	62	65	64	81	83	84	93	96	96
Global DGV	60	67	67	na	na	na	95	99	99

[a]Born since 2005 (n=120), [b]Born 2000-2004 (n=1518), [c]All genotyped males (n=8193)