Preliminary Report from Interbull Task Force on the Role of Genomic Information in Genetic Evaluations

G. Banos[a], M. Calus[b], V. Ducrocq[c], J. Dürr[d], H. Jorjani[d], Z. Liu[e], E. Mäntysaari[f], P. Sullivan[g] and P. VanRaden[h]

[a] Faculty of Veterinary Medicine, Aristole University of Thessaloniki,
Box 393 GR-54124 Thessaloniki, Greece
[b] Animal Breeding and Genomics Centre, Animal Sciences Group, Wageningen University and Research Centre, 8200 AB Lelystad, The Netherlands
[c] Station de Génétique Quantitative et Appliquée, INRA, Domaine de Vilvert, 78352 Jouy-en-Josas, France
[d] Interbull Centre, Department of Animal Breeding and Genetics, SLU, Box 7023 – 750 07, Uppsala Sweden
[e] VIT, Heideweg 1, D-27283 Verden, Germany
[f] Agrifood Research Finland, Animal Production, 31600 Jokioinen, Finland
[g] Canadian Dairy Network,150 Research Lane, Suite 307, Guelph, N1G4T2, Ontario, Canada
[h] USDA – ARS, Animal Improvement Programs Laboratory, Building 5,
BARC-West, Beltsville, Maryland 20705, USA

Introduction

Following the joint session between Interbull and ICAR on the use of genomic data in Niagara Falls (June 2008), it became obvious that Interbull needed to address this issue in order to be in line with the fast developments happening at the member country level. Therefore, Interbull Steering Committee decided to establish a task force with two main objectives: i. to set the scientific framework for the use of genomic data in national and international genetic evaluations, and ii. to promote the idea and benefits of international collaboration, under the auspices of Interbull, with regards to genomic evaluation and selection.

A group of nine scientists accepted to participate in the task force. Terms of reference were defined, and the mandate included the following points:

i. Review and evaluate methods of using and/or incorporating genomic data into national and international genetic evaluations.
ii. Define the process for including the following data in Interbull evaluations:
    a. national genetic evaluations based on progeny performance of bulls intensively pre-selected using genomic data
    b. national genetic evaluations of bulls based on both progeny performance and genomic data
    c. national genetic evaluations of bulls based on pedigree and genomic data only
    d. bull genomic data.
iii. Propose methods that safeguard the quality and transparency of procedures incorporating genomic data to genetic evaluations.
iv. Outline the expected benefits from collaboration and sharing of genomic information between different countries and/or breeding organizations.
v. Prioritize future research needs.

The Task Force carried out an intense online discussion and finally met in Paris, in December 11 and 12, 2008, resulting in a preliminary collection of concepts on the use of genomic information in genetic evaluations that constitute the present report.

Genomic Data

1. *Terminology*

Standardization of a new terminology related to the use of genomic information is recommended. As a suggestion, the following terms will be adopted in this document:

- EBV – “conventional” estimated breeding value, free of genomic information
- SNP – single nucleotide polymorphism
- DGV – direct estimated genomic value

(based on genomic data only)

- GEBV – genomically enhanced estimated breeding value (combining EBV and DGV)

2. *Possible scenarios*

Methodology to be adopted in international genetic evaluations incorporating genomic information depends on what type of genomic information will be available to Interbull. It is possible to envision the following scenarios:

I. Interbull has access to national EBVs and bull genotypes

In this case, Interbull would maintain an international genotype database, which could optimize investments in genotyping if countries agreed to trade genotypic information through Interbull. Access to data would be restricted to Interbull Centre, which would be able to estimate SNP prediction equations in each country scale using all available national EBVs. This scenario is particularly attractive for small populations, which do not have enough data at the national database to estimate accurate SNP prediction equations.

II. Interbull has access to national EBVs, bull genotypes and SNP national prediction equations

Similar to the previous scenario, with the difference that Interbull would not have to estimate SNP effects, which simplifies significantly operational flow. Data access will still be absolutely restricted to Interbull Centre, which would apply the provided equations to estimate direct genomic breeding values (DGVs) of genotyped foreign animals in each country scale. In this case, countries/breeds with small populations would no longer benefit from Interbull genomic database to increase number of genotyped animals for SNP effect estimation, unless Interbull calculates SNP effects only for these breeds.

III. Interbull has access only to national EBVs and DGVs

In this scenario, the focus of Interbull activity would be on how to combine conventional and genomic breeding values into international genetic evaluations. Access to DGVs would also be restricted to Interbull Centre.

IV. Interbull has access only to national EBVs and GEBVs

In this scenario, the focus of Interbull would also be on how to combine conventional and genomic breeding values into international genetic evaluations.

V. Interbull has access only to national GEBVs

This scenario would jeopardize Interbull capability of estimating conventional international genetic evaluations and member countries would not have access to MACE EBVs as input for SNP effect estimation, only to MACE GEBVs.

Table 1. Requirements for the development of each possible scenario for Interbull including genomic information in international genetic evaluations.

Requirement	**Scenario**				
	I	**II**	**III**	**IV**	**V**
Interbull getting access to genotypes	X	X			
Interbull getting access to national prediction equations		X			
Interbull getting access to national DGVs			X		
Interbull getting access to national GEBVs				X	X
Interbull getting access to national EBVs	X	X	X	X	
Developing, validating, testing and implementing the proper methodology	X	X	X	X	X
Changing current sire models used by Interbull into animal models in order to account for dam genomic information (very important for young sires)	X	X			
Implementing a completely new framework of analysis at Interbull Centre on top of current activities in a very short period	X	X			
Meeting countries' plans of estimating SNP effects several times per year	X	X			
Having to assume that all countries use the same SNP array to genotype the animals, which currently is not the case	X	X			

Table 1 shows the requirements for developing each of the proposed scenarios. Although scenarios I and II may be closer to an academically ideal situation, the scenarios III and IV seem more realistic in the short term. Scenario V is undesirable because of the limitations it imposes for national and international genetic evaluations.

Therefore, it is proposed to focus on the best possible methodology to combine national EBV, DGV and/or GEBV while, at the same time, identifying an international database of bull genotypes as a longer term goal for Interbull.

The discussion that follows on methodology will focus on incorporating DGVs and/or GEBVs into international genetic evaluations and the questions to be addressed to make it possible.

Methodology

1. Combining conventional EBV with DGV/GEBV

SNP data do not explain 100% of the variation in a trait. It gets closer to 100% as the number of independent and evenly spaced SNP increases, but has not reached 100% (yet). Because of this, it is not simply possible to combine genomic evaluations from different countries additively in the way conventional MACE applies to national daughter data. That is, the maximum reliability from a correlated trait is the squared genetic correlation (which is less than 100%). In terms of conventional MACE, genetic correlations between countries are used, however these maybe not the right ones to use for the genomic evaluations.

Interbull may have two different goals regarding GEBVs: a. convert GEBVs from each country to other country-scales and b. continue to provide non-genomic EBV as "training data" inputs for each country's genomic system. If MACE GEBVs are used as inputs to domestic genomic equations, then we will double count the genomic information and therefore biased estimates will be generated. Foreign bulls can increase the accuracy of domestic GEBVs a lot, but the equations expect EBVs rather than GEBVs as input data. Eventually, each country may need two files (both MACE EBV and GEBV) to meet the two goals. That may require running two MACE evaluations per trait group, one with and the other without genomic data.
Two separate MACE runs (one for conventional, and the other for DGV or GEBV) may provide a safeguard, in comparison to one single MACE run using combined information of both predicted genetic merit, in case there would be unexpected problems in national GEBVs from member countries. The current conventional MACE service is, therefore, not influenced by such problems from national genomic evaluation.

The GEBVs submitted to the second MACE run should be estimated using only national data, the same requirement we use for conventional MACE. This means that countries may need to run two national genomic evaluations, one for Interbull and a different one for national publication (same as happens now and will continue for conventional MACE, without and with foreign blending). If transparency is a concern here, it must also be a concern currently since this is the same approach used already by many countries, before genomics. For countries without genomic data, their input data is the same for the two MACE runs.

We need to modify the second MACE run to account for the fact that the genomic information from multiple countries cannot sum above the maximum proportion of variation that could be explained by genomics.

An alternative to avoid running two rounds of MACE per trait group would be to run a multi-trait MACE fitting EBVs and DGVs for the same character as different traits. In this case, it might be necessary to treat genomics as repeated records and simulation studies would need to be run to study what kind of correlation structure can/should be used.

Ultimately, using MT-MACE may seem preferable from the theoretical point of view, but it will not prevent the need of two MACE runs, since conventional MACE (free of genomic information) will still be necessary to provide input data for estimation of SNP effects at the national level.

2. *Avoiding double counting*

Currently MACE works with national data that are independent of each other. This allows assuming that across country residual are equal to zero.

Using MACE breeding values as input to genomic equations at national level blurs the distinction between national and international data. For example, a bull may have only daughters born in one country, but SNP solutions will include information from some foreign relatives. Thus, genomic evaluations from two different countries will be correlated in ways that the traditional EBVs were not. These extra correlations might safely be ignored for bulls that already have >100 daughters or for bulls that have a genomic EBV in only one country, but will become an issue as more countries report genomic evaluations.

If the Y vectors used by each country to estimate X (the SNP effects used to derive DGV) are residually independent (e.g. only within-country data are included in Y) there is no problem with double-counting. However, there will be a problem if the same data (Y) are used in multiple countries, for example if countries use MACE proofs to predict DGV. It would not matter if each country genotyped the same bull separately or if Interbull acted as the mediator in trading the bull's genotype between countries. The double-counting comes from vector Y, not X (the genotype).

A fundamental concept with single trait MACE is that residually independent data sets are first evaluated with complicated national models, and the results can then be combined via MACE (a simpler model targeted to sires only). If the MACE input data are not derived from residually independent data sets (Y), we have violated a fundamental assumption of MACE, which will cause incorrect covariance estimates between countries, i.e. double-counting of information, etc.

When countries use only national data to estimate DGVs, they can be included in MACE without introducing double-counting problems. For other countries that use more than domestic information (potential contributor to the double-counting problem), Interbull should use only DGVs of their domestic bulls in MACE. With this strategy, the most important area of concern about double-counting will be for young sires that were jointly sampled in multiple countries.

Another source of double counting arises if major genes exist for a given trait, then the marker effects may act as repeated measures of the same portion of genetic merit rather than independent measures of total genetic merit.

Paul VanRaden and Peter Sullivan have proposed the use of a non-diagonal residual matrix to account for the first type of double counting (unpublished data). No methodology has been proposed for the second type yet.

3. *Computing reliabilities*

Initially theoretical reliabilities from matrix inversion were reported in the US, but 15,000 Holstein bulls are genotyped already, and direct inversion will not be possible in the near future. More recently, the theoretical reliabilities were adjusted downward to be more consistent with observed reliabilities, but further research is needed.

4. *Validation and Bias of national evaluations*

In order to ensure data quality, at least two issues need to be resolved, a) the effect of genomic selection on the distribution of Mendelian sampling terms (including the bias in its estimation) and b) completeness of information about selection decisions.

Different methods of validating national genetic evaluations (including Interbull validation tests) rely on unbiased estimate of genetic trend (which is assumed to be of reasonable magnitude) and/or unbiased estimates of variance of Mendelian sampling term (which is assumed to be of constant magnitude).

Further, currently available estimation theory (in its entirety) assumes the availability of all data that were used for making selection decisions. Incomplete data (known as “missing data” in statistical jargon) violates the assumptions of any mixed model methodology.

Can MACE be expected to work well on national EBVs although young sires are selected by DGV? This is not a problem that can be solved by MACE; it is a problem which has to be resolved at the national level. Thus, national evaluations will give wrong estimates of genetic trend versus environmental trend. This means that countries will not pass Interbull trend validation tests, and all parties need to deal with the inconveniences that causes. The pre-selection issue is a problem in MACE if only the above-average sires are given the opportunity for progeny evaluation. However, if the genomic evaluations used for pre-selection are available to national genetic evaluation centers (and eventually to Interbull), there are probably methods that can be used to account for the pre-selection, and eliminate the effects of selection bias on estimated trends sire variance.

One way to account for selection bias due to genomic selection is that countries should send DGVs for all young bulls included in the selection process (both selected and culled candidates). This is not only important in order to avoid selection bias on genomic evaluations, but for conventional evaluations as well. Methods and recommendations may be needed to account for pre-selection in national evaluation systems.

Data Exchange Benefits

One of the objectives of the Task Force is “to promote the idea and benefits of international collaboration, under the auspices of Interbull, with regards to genomic evaluation and selection”. The following arguments can be used in favor of genomic data exchange through Interbull:

1. Smaller breeds are expected to benefit more from data exchange since their global population will be substantially bigger than any single national one; this will allow more accurate estimation of SNP effects, DGV and GEBV.
2. Genomic data are not the same in different countries; SNP effects differ from country to country and DGV and GEBV must be calculated on different country-scales. Genotype-by-environment interaction still exists.
3. Interbull is the EU reference laboratory for bovine testing and genetic evaluation. The EU does not allow importations of bull semen if it is not evaluated by an internationally accepted body (Interbull). This supports the Interbull leading role in genomic evaluations. Assuring fair trade between countries is one of Interbull mandates.
4. More data-sharing benefits are expected for traits that are difficult to record (such as health-related, which become increasingly important in genetic evaluations).
5. Since many animals have progeny in more than one country, exchanging data would avoid that the same individuals are genotyped several times, optimizing investments.
6. Young bulls can receive a genomic evaluation for several countries as soon as genotypes are sent to Interbull Centre, which may allow different selection decisions in different environments.
7. If national and international genetic evaluations are biased due to genomic pre-selection, corrections or more robust models are needed and a minimum knowledge of pre-selection intensity is likely to be required. Interbull is the ideal place for sharing this knowledge.
8. Many alternative methodological approaches have been developed worldwide already, and Interbull is the logical forum for exchange of experience for validation of methods and results, and also for standardization of procedures. Test benchmarking and reference dataset simulations monitored by Interbull will help evaluate the methods.
9. Assuming that different SNP arrays and DNA screening techniques are likely to become available in the future, an international database of genotypes can be the basis for integrating genomic information of different sources.
10. Intensive use of genomic selection carries a danger of decreasing genetic variability within breeds and increasing overall inbreeding. Interbull is the only organization capable of monitoring genetic variability and inbreeding worldwide, given that a database of genotypes is in place.
11. As a neutral body, Interbull is prepared to maintain an international database of genotypes with restricted access, assuring confidentiality for all participants.