Multitrait across country genomic evaluations for EuroGenomics countries

Hanni Kärkkäinen[1], Vincent Ducrocq[2], Sören Borchersen[3],
Gert Aamand[4], Reinhard Reents[5], Esa Mäntysaari[1]

Interbull Open Meeting, June 22, 2019

[1]Natural Resources Institute Finland,
[2]French National Institute for Agricultural Research (INRA), France
[3]EuroGenomics Cooperation, Denmark
[4]Nordic Cattle Genetic Evaluation (NAV), Denmark
[5]Vereinigte Informationssysteme Tierhaltung w.V. (vit), Germany

The project

- EuroGenomics[1] multitrait across country evaluation -project[2] started on May 2018
- Contrary to the Melbourne project, EuroGenomics countries share bull genotypes

  $\Longrightarrow$ Possible to build a true multitrait across country SNP BLUP evaluation using pseudo phenotypes from all countries directly

[1] Germany (DEU), Nordic countries Denmark, Finland and Sweden (DFS), France (FRA), The Netherlands (NLD), Spain (ESP) and Poland (POL). Order and abbreviations from Interbull practice.

[2] Financed jointly by Luke, INRA, EuroGenomics COOP and German Livestock Association.

Shared EuroGenomics data

Genotypes

- 46,342 SNP genotypes for total of $\sim$ 35,000 bulls with a record
- Imputed genotypes received from NAV ($\rightarrow$ common set of markers)

Phenotypes

- Protein yield, somatic cell score and female fertility
- Around 11,000 – 3,700 records within countries
- EBVs the countries send to Interbull evaluation + EDC → DRP
- Heritability estimates from countries

Pedigree

Genetic correlation estimates from Interbull

First phase

SNP MACE

- Our first goal was to demonstrate and validate the performance of EuroGenomics SNP MACE
- We have accomplished that based on the shared bull genotypes, and shown that it is feasible and benefits the participants

SNP MACE Model

- Basic SNP MACE model $\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zg} + \mathbf{e}$

$$\iff \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_c \end{bmatrix} = \begin{bmatrix} \mu_1 \mathbf{1}^{n_1} \\ \vdots \\ \mu_c \mathbf{1}^{n_c} \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 \mathbf{g}_1 \\ \vdots \\ \mathbf{Z}_c \mathbf{g}_c \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_c \end{bmatrix} \tag{1}$$

- $\mathbf{y}_i \in \mathbb{R}^{n_i}$ is the pseudo phenotype (deregressed national breeding value, later DYD) for country $i \in [1, \ldots, c]$ with n_i observations
- μ_i the general mean for country i
- $\mathbf{1}$ vector of n_i ones
- $\mathbf{Z}_i \in \mathbb{R}^{n_i \times m}$ design matrix for genotypes (m is the number of markers, all countries have the same set of markers with same 0,1,2 coding)

- $\mathbf{g}_i \in \mathbb{R}^m$ estimated SNP effects for country i
- $\mathbf{e}_i \in \mathbb{R}^{n_i}$ residual effects for country i individuals
- $\mathrm{Var}(\mathbf{e}_i) = \sigma^2_{e_i} \mathrm{diag}(1/\mathrm{EDC}_{ik}) = \mathbf{R}_i \ \forall i$, for animals $k \in [1, \ldots, n_i]$
- $\mathrm{Cov}(\mathbf{e}_i, \mathbf{e}_{i^+}) = 0 \ \ \forall i \neq i^+$
- $\mathrm{Var}(\mathbf{g}_i) = \mathbf{I}^m \sigma^2_{s_i} \theta_i$, where $\theta_i = 1/\sum_{j=1}^{m} 2p_i(1 - p_{ij})$ with p_{ij} = allele frequency of locus j in country i, $\sigma^2_{s_i}$ = sire variance of country i and $\mathbf{I}^m \in \mathbb{R}^{m \times m}$ identity matrix
- $\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_{i^+}) = \mathbf{I}^m \sigma_{ii^+} \sqrt{\theta_i \theta_{i^+}}$, where $\sigma_{ii^+} = \rho_{ii^+} \times \sigma_{s_i} \sigma_{s_{i^+}}$, with ρ_{ii^+} = genetic correlation between countries i and i^+

SNP MACE Model – (Co)variance matrices

$$\text{Var}\begin{bmatrix}\mathbf{g}_1\\ \vdots\\ \mathbf{g}_c\end{bmatrix} = \begin{bmatrix}\mathbf{I}^m\sigma^2_{s_1}\theta_1 & \dots & \mathbf{I}^m\sigma_{1c}\sqrt{\theta_1\theta_c}\\ & \ddots & \vdots\\ symm. & & \mathbf{I}^m\sigma^2_{s_c}\theta_c\end{bmatrix} = \mathbf{D} \qquad (2)$$

$\in \mathbb{R}^{(c\times m)\times(c\times m)}$ and it's inverse

$$\mathbf{D}^{-1} = \begin{bmatrix}\mathbf{D}^{11} & \dots & \mathbf{D}^{1c}\\ & \ddots & \vdots\\ symm. & & \mathbf{D}^{cc}\end{bmatrix} \qquad (3)$$

$\in \mathbb{R}^{(c\times m)\times(c\times m)}$.

$$\text{Var}\begin{bmatrix}\mathbf{e}_1\\ \vdots\\ \mathbf{e}_c\end{bmatrix} = \begin{bmatrix}\mathbf{R}_1 & \dots & \mathbf{0}\\ & \ddots & \vdots\\ symm. & & \mathbf{R}_c\end{bmatrix} = \mathbf{R} \qquad (4)$$

$\in \mathbb{R}^{n\times n}$, where $n = \sum_{i=1}^{c} n_i$ and

$\mathbf{R}_i = \sigma^2_{e_i}\text{diag}(1/\text{EDC}_{ik})$.

SNP MACE Model – Mixed Model Equations

$$\begin{bmatrix} \ddots & \vdots & & \vdots & \iddots \\ & \begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{Z}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{ii} \end{bmatrix} & \cdots & \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{ii^+} \end{bmatrix} & \cdots \\ & & \ddots & \vdots & \\ & & & \begin{bmatrix} \mathbf{1}'\mathbf{R}_{i^+}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_{i^+}^{-1}\mathbf{Z}_{i^+} \\ \mathbf{Z}_{i^+}'\mathbf{R}_{i^+}^{-1}\mathbf{1} & \mathbf{Z}_{i^+}'\mathbf{R}_{i^+}^{-1}\mathbf{Z}_{i^+} + \mathbf{D}^{i^+i^+} \end{bmatrix} & \cdots \\ & & & & \ddots \end{bmatrix} \times \begin{bmatrix} \vdots \\ \begin{bmatrix} \hat{\mu}_i \\ \hat{\mathbf{g}}_i \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \hat{\mu}_{i^+} \\ \hat{\mathbf{g}}_{i^+} \end{bmatrix} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{y}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{y}_i \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{y}_i \\ \mathbf{Z}_{i^+}'\mathbf{R}_i^{-1}\mathbf{y}_{i^+} \end{bmatrix} \\ \vdots \end{bmatrix} \quad (5)$$

Validation Method

- Data was split into learning and validation sets by bulls' birth date
  - The youngest 10% from each country → validation set
- Under SNP MACE the animal solutions (DGV) were computed as $\hat{a}_{ik} = \mathbf{z}_{ik}\hat{\mathbf{g}}_i$ for animal k in country i
- The bias b_1 was tested with a weighted linear regression of DRP_v on predicted DGV_v, using EDC_v as weights
- Validation reliability was defined as $R_v^2 = (\text{cor}(\text{DRP}_v, \text{DGV}_v))^2 / R^2_{\text{DRP}_v}$,
  - Records with $R^2_{\text{DRP}_v} \geq 0.5$ were used in validation
    (except for Poland fertility trait $R^2_{\text{DRP}_v} \geq 0.3$, due to limited no. of records)

Reference Estimation Methods for Validation

Two reference methods:

1. Country-wise single trait model
    - Compares to situation where country uses only their own geno- and phenotypes
2. Current EuroGenomics practice *i.e.* using MACE proofs for all exhange bulls
    1. Run MACE BLUP $\longrightarrow$ solutions for all animals
    2. Estimate reliabilities / EDC for all records
    3. $\longrightarrow$ Deregressed proofs for all animals
    4. $\longrightarrow$ National DGV:s by single trait GBLUP

Additional developments

Residual polygenic component

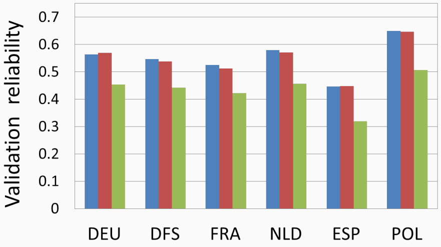
- Benefits all models (genomic MACE and both reference methods)
- We tested 10, 20 and 30% of polygenic effect $\Rightarrow$ On average 20% best choice
- Genomic MACE R_v^2 rises on average 6%, also bias diminishes noticeably

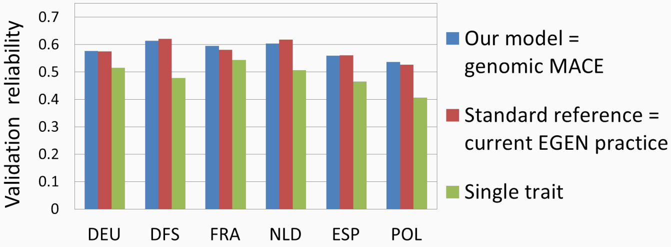
Estimation of genetic correlations

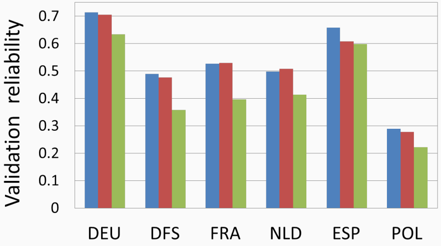
- Estimated with MTG2 program (Lee & *al.*)
- Done with the ”official Interbull style”
  $=$ variance ratio kept constant, genetic covariances and sire variances estimated
- Not much different values than Interbull estimates $\iff$ Not much different reliabilities

Results

Validation reliability R_v^2 of DGV predicted by the genomic MACE, current EuroGenomics MACE and a single trait model, all models with 20% polygenic effect and Interbull variance components.


Somatic cell score


Protein yield


Female fertility

First phase conclusions

After the first phase we have learned that

- Fitting genomic MACE with individual animal genotypes is feasible, and countries gain from cooperation
- The genomic MACE produces on average slightly higher validation reliability and is slightly less biased (higher b_1) than the current EuroGenomics MACE
- Under all of the tested models the equivalent GBLUP has better convergence properties than the SNP BLUP
- Residual polygenic component seems useful
- Genetic correlations estimated by Interbull can be utilized in genomic MACE

Second phase

Including cow information

- EuroGenomics countries want to include cow reference information *without* sharing the cow genotypes
- We are developing a method to use all the information, including cows
    - requires only SNP-solutions (computed with full national reference population) and PEVs of the shared bulls
- Procedure includes
    1. PEV $\xrightarrow{\text{iterative approximation}}$ EDC for bulls
    2. SNP-solutions & EDC & genotypes $\longrightarrow$ RHS
    3. RHS & EDC & genotypes $\longrightarrow$ Multitrait SNP-effects
- We call this ”SNP information approximation approach”

SNP MACE Model – Mixed Model Equations

$$\begin{bmatrix} \ddots & \vdots & & \vdots & \iddots \\ & \begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{Z}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{ii} \end{bmatrix} & \cdots & \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{ii^+} \end{bmatrix} & \cdots \\ & & \ddots & \vdots & \\ & & & \begin{bmatrix} \mathbf{1}'\mathbf{R}_{i^+}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_{i^+}^{-1}\mathbf{Z}_{i^+} \\ \mathbf{Z}_{i^+}'\mathbf{R}_{i^+}^{-1}\mathbf{1} & \mathbf{Z}_{i^+}'\mathbf{R}_{i^+}^{-1}\mathbf{Z}_{i^+} + \mathbf{D}^{i^+i^+} \end{bmatrix} & \cdots \\ & & & & \ddots \end{bmatrix} \times \begin{bmatrix} \vdots \\ \begin{bmatrix} \hat{\mu}_i \\ \hat{\mathbf{g}}_i \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \hat{\mu}_{i^+} \\ \hat{\mathbf{g}}_{i^+} \end{bmatrix} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{y}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{y}_i \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{y}_i \\ \mathbf{Z}_{i^+}'\mathbf{R}_i^{-1}\mathbf{y}_{i^+} \end{bmatrix} \\ \vdots \end{bmatrix}$$

Estimation of country wise EDC — Background

Since EuroGenomics countries share the bull genotypes, we can construct the left hand side matrices

$$\begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{Z}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}_i^{-1} \end{bmatrix}$$

— if we know the $\mathbf{R}_i^{-1}$

- Matrix $\mathbf{R}_i^{-1}$ holds weights for each bull
- Until now we have used $\mathbf{R}_i^{-1} = diag\{EDC_{ij}\sigma_{e_i}^{-2}\}$, with EDC_{ij} being the number of daughters of bull j in country i
- On the other hand, the prediction error variance of bull j from country i

$$PEV_{ij} \simeq [(\mathbf{R}_i^{-1} + \mathbf{G}_i^{-1}\sigma_{s_i}^{-2})^{-1}]_{j,j}, \text{ where } \mathbf{G}_i = \sigma_{s_i}^{-2} \times \mathbf{Z}_i\mathbf{D}_i\mathbf{Z}_i'$$

- Equivalently the same can be attained using SNP model

$$PEV_{ij} \simeq \mathbf{z}_j[(\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}_i^{-1})^{-1}\mathbf{z}_j'$$

Estimation of country wise EDC – II

- However, we are restricted to genotypes and EDCs that countries exchange $\Longrightarrow$ can't compute $(\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}_i^{-1})^{-1}$ the countries use
- But, if we would get for each exchanged bull

  $$PEV_{ij} = \mathbf{z}_j[(\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}_i^{-1})^{-1}\mathbf{z}_j'$$

  from the countries
- We could equate it to

  $$PEV_{ij} = [(\Re_i^{-1} + \mathbf{G}_i^{-1}\sigma_{s_i}^{-2})^{-1}]_{j,j}$$

  where $\Re_i^{-1}$ would consists of weights of the (exchange) bulls that would lead into the same PEV that the country has computed with all animals (including) females in national genomic evaluation.

Estimation of $\Re_i^{-1}$

- Values of $\Re_i^{-1}$ can be estimated iteratively
- We have used a Newton method, where the new EDC are estimated as:

$$edc_{k+1} = edc_k - \mathbf{C}_k^{-1}(PEV_k - PEV), \text{ where}$$

  - edc_k and edc_{k+1} are the current and the subsequent EDC estimates, respectively,
  - PEV_k is the PEV computed as diag(LHS^{-1}) using the current (kth) estimate of EDC,
  - PEV consists of the PEVs the country has computed with the full national reference population and
  - $\mathbf{C}$ is the value of the partial derivative of $(PEV_k - PEV)$ with respect to edc at point edc_k, that can be simplified into $\mathbf{C} = LHS^{-1} \circ LHS^{-1}$,

corresponding to the general description of the Newton method

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

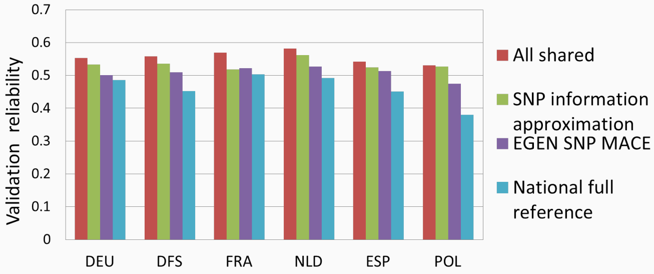
Pilot testing the model with current EuroGenomics data

1. Divide data for every country:
   - Validation (youngest 10%) & "full national reference population"
   - "Full reference" $\xrightarrow{\text{randomly 1:1}}$ "Shared bulls" + "Cows"
2. SNP-solutions and PEVs by country using the "full national reference"

   → "shared SNP-solutions" & "PEVs of the shared bulls"
3. Country i PEVs $\xrightarrow{\text{Newton iteration}}$ Country i EDCs
4. Country i SNP-solutions → Country i RHS
   - using "shared" genotypes (= half of the animals)
   - estimated EDC_i^{-1} as weights
5. All country wise RHS → multitrait SNP BLUP → SNP-solutions → DGV
6. Validate by using the youngest 10%

Pilot for protein yield

Validation reliability R_v^2 of DGV predicted by multitrait SNP BLUP with full reference population, SNP information approximation, MT SNP BLUP with shared reference and national full reference SNP BLUP.



MULTITRAIT ACROSS COUNTRY

- **All shared**: countries would share everything, including cow genotypes
- **SNP information approximation**: countries share bull genotypes (as currently) + SNP-solutions & PEVs of shared bulls
- **EGEN SNP MACE**: countries share bull genotypes — phase I model

SINGLE TRAIT

- **National full reference**: countries use all national information, but do not share anything

Discussion

EDC estimation

- The Newton iteration is computationally feasible
  - Requires 2 inverses of matrix size *number_of_animals* / iteration round
- In the pilot study the iteration was run until convergence

National SNP-solutions → RHS → MT SNP BLUP

- Is implemented for testing purposes as part of our MiX99 suite
- Works quite nicely
  - Converges and behaves similar to "normal" multitrait SNP BLUP runs
- Could be even more advantageous for low heritability traits
- **Pilot study!** ⇒ With actual cow data behaves probably differently

EURO Genomics
for reliable cattle breeding

Luke
NATURAL RESOURCES
INSTITUTE FINLAND

BUNDESVERBAND
RIND UND SCHWEIN E.V.

Thank You !

INRA
SCIENCE & IMPACT

vit
Service & Daten aus einer Quelle