Meta-model for genomic relationships of metafounders applied on large scale single-step random regression test-day model

M. Koivula[1], I. Strandén[1], G. P. Aamand[2], E. A. Mäntysaari[1]

1 Natural Resources Institute Finland (Luke), FI-31600 Jokioinen, Finland

2 NAV Nordic Cattle Genetic Evaluation, 8200 Aarhus N, Denmark

Allele frequencies (AF)

- Original single step approach used average population allele frequency (AF)
  - In this approach the smallest genomic inbreeding was in a point with a largest mass of genotypes
- Advised approach was to estimate "a base population AF"
  - This was done by essentially estimating the AF from genotyped animals that had missing parents or parents that have genotyped ancestors (details Gengler et al. 2007).
  - If many genotyped females at the end of the time span:
    --> The AF became the AF of the youngest females
- AF can be also estimated from different base populations (groups of animals with unknown parents)
  - Define groups into pedigree (for example breed or breed-origin and the birth decade)
  - Estimate AF in groups using e.g., Bpop (*Bpop, Strandén and Mäntysaari, 2020, AFSci Finland*)
    - We have considered group of animals born 1980s as a base population

- One option is to assume all AF= 0.5.
  This approach assumes the base population is many, many generations earlier

Genetic groups

- Genetic groups have significant effect on genetic trends, and, in the single-step genomic BLUP model, on convergence of iterative solving
- Genetic groups can be included into the evaluation model as birth year effects, or unknown parent contributions as regression coefficients
- Computationally more efficient approach is to re-express the parental genetic groups as unknown parent groups (UPG) resulting from QP transformation
- Originally in single step models, $\mathbf{A}_*^{-1}$ included the UPG, i.e., animals descended from different base populations.
  - This was not done for $\mathbf{A}_{22}^{-1}$ which, thus, assumed only one base population
  - Lead to convergence problems

Genetic groups

- The solution was the full QP transformation model where QP transformation is done to full $\mathbf{H}^{-1}$ matrix by inclusion of products

  $\mathbf{Q}'(\boldsymbol{G}_w^{-1} - \boldsymbol{A}_{22}^{-1})\mathbf{Q}$ and $-(\boldsymbol{G}_w^{-1} - \boldsymbol{A}_{22}^{-1})\mathbf{Q}$ into group equations in $\mathbf{A}_*^{-1}$

- Alternative option for accounting different base populations is by combining pedigree and genomic information using metafounders (MF)

Aim: to compare single step models using either QP transformation with different allele frequencies or MF approach

Data

- Official Holstein Nordic TD evaluation data for milk, protein and fat
- Genomic data:
    - 274 145 genotyped animals

FULL TD data

- 8.5 million animals with records, 10.9 million animals in the pedigree

REDUCED TD data for validation (four years of data reduction)

- 7.5 million animals with records

Metafounder approach

Single-step GBLUP assumes that the genomic and pedigree relationships are relative to a same base population

Alternatively, we could define the base population of $\mathbf{A}_{22}$ to a same base as in **G** and natural base population could be where the animals are unrelated and not inbred (AF=0.5)

MF steps:

1. Assume a base for **G** matrix to be in where the AF = 0.5
2. Define the base populations for $\mathbf{A}_*^{-1}$ (and $\mathbf{A}_{22}$) to be relative to the current genotyped animals (i.e., where the AF=0.5)
    - Estimate the allele frequencies in unknown parent genetic groups
    - Estimate **Γ** i.e., "genomic compliant relationships" among base population animals
    - Estimate inbreeding for all the animals using the **Γ**
3. Form $(\mathbf{A\Gamma})^{-1}$ (and $\mathbf{A\Gamma}_{22}$)

Metafounder tested approach

- Normally in dairy cattle there are > 100 genetic groups
  - in original NAV Holstein evaluation 446
- Define less genetic groups (from 446 to 176)

  Base breeds were assumed to be:
  - HOL divided into DNK, SWE, FIN, Other and RED
  - RDC, JER and "other" + a common trend by time

➡ the rank of the covariance function 9

- Assume metafounder **Γ**-matrix has a structure
  - Structure can be defined with covariance function kernel **K** (Kirkpatrik et al., 1994)
    - $\boldsymbol{\Gamma}_9 = \Phi_9 \boldsymbol{K} \Phi_9'$
    - $\boldsymbol{K} = (\Phi_9'\Phi_9)^{-1} * \Phi_9'\boldsymbol{\Gamma}_9\Phi_9 * (\Phi_9'\Phi_9)^{-1}$ (Tijani et al. 1999)
  - Covariance function covariables extend this structure to all groups $\boldsymbol{\Gamma}_{176} = \Phi_{176} \boldsymbol{K} \Phi_{176}'$

Single-step models compared

ssGTBLUP

- ssGTBLUP with AF from 1980 's considered as base population (Bpop, Strandén and Mäntysaari, 2020)
- 176 genetic groups and full QP transformation, RPG 30 %
- Pedigree inbreeding accounted in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}$
- Matrix **G** was scaled so that trace(**G**)==trace($\mathbf{A}_{22}$)

ssGTBLUP_AF05

- ssGTBLUP with AF 0.5
- 176 genetic groups and full QP transformation, RPG 30%
- Pedigree inbreeding accounted in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}$
- Matrix **G** was scaled so trace(**G**)==trace($\mathbf{A}_{22}$)

ssGTBLUP_MF

- Metafounder model, **G** with AF=0.5
- RPG 30 %
- MF based inbreeding accounted in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}$
- 176 meta-founders, $\boldsymbol{\Gamma}$ -matrix with CF

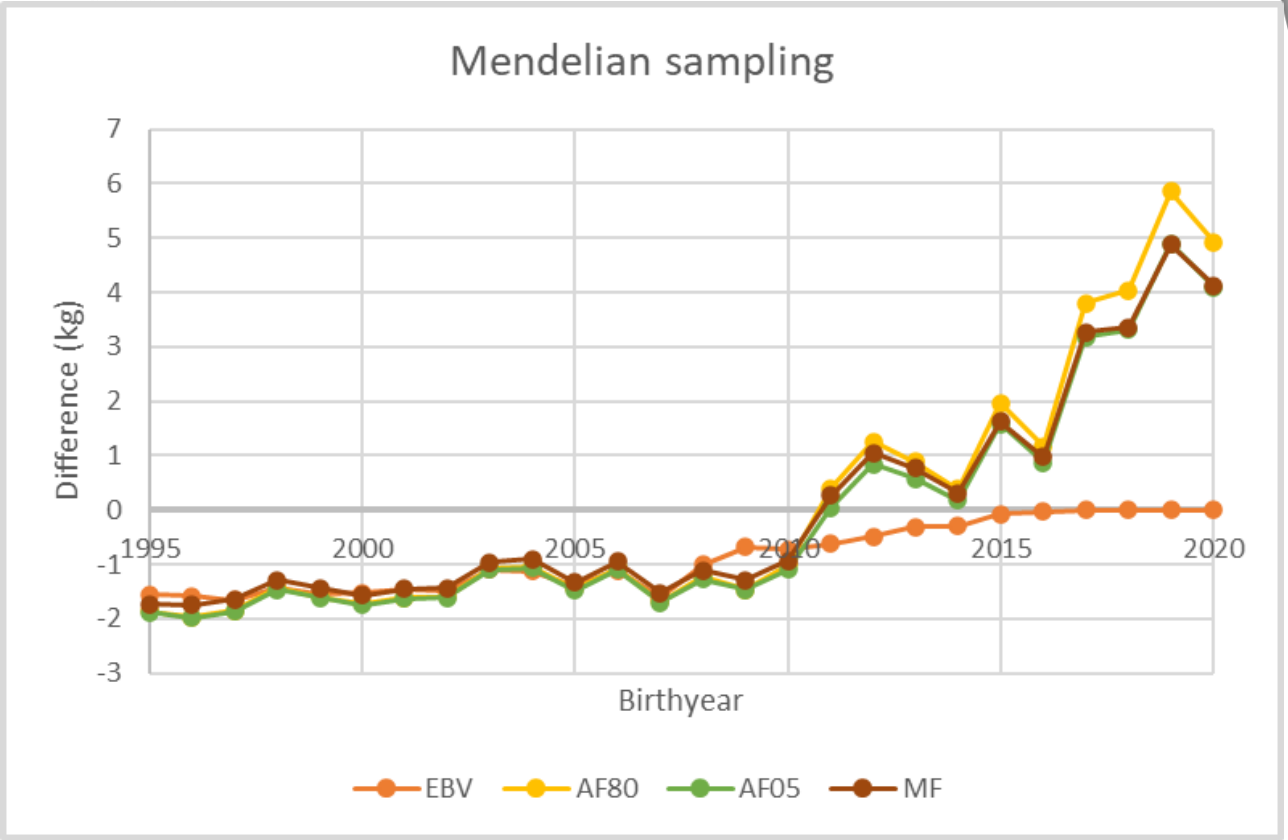
Legarra-Reverter regression

		b_0	b_1	R^2
Milk	PA	-101.7	0.84	0.32
	GEBV_AF80	-311.8	0.87	0.67
	GEBV_AF05	-319.8	0.87	0.67
	GEBV_MF	-272.3	0.89	0.68
Protein	PA	0.80	0.74	0.24
	GEBV_AF80	-10.81	0.82	0.63
	GEBV_AF05	-11.10	0.81	0.63
	GEBV_MF	-9.71	0.83	0.64
Fat	PA	-2.18	0.73	0.23
	GEBV_AF80	-15.81	0.82	0.64
	GEBV_AF05	-16.16	0.82	0.64
	GEBV_MF	-14.67	0.85	0.65

b_0=mean(Full_(G)EBV – reduced_(G)EBV)
GEBV_AF80 - ssGTBLUP with QP and RPG 0.30 and AF 1980
GEBV_AF05 - ssGTBLUP with QP and RPG 0.30 and AF 0.5
GEBV_MF - ssGTBLUP with RPG 0.30 and MetaFounders

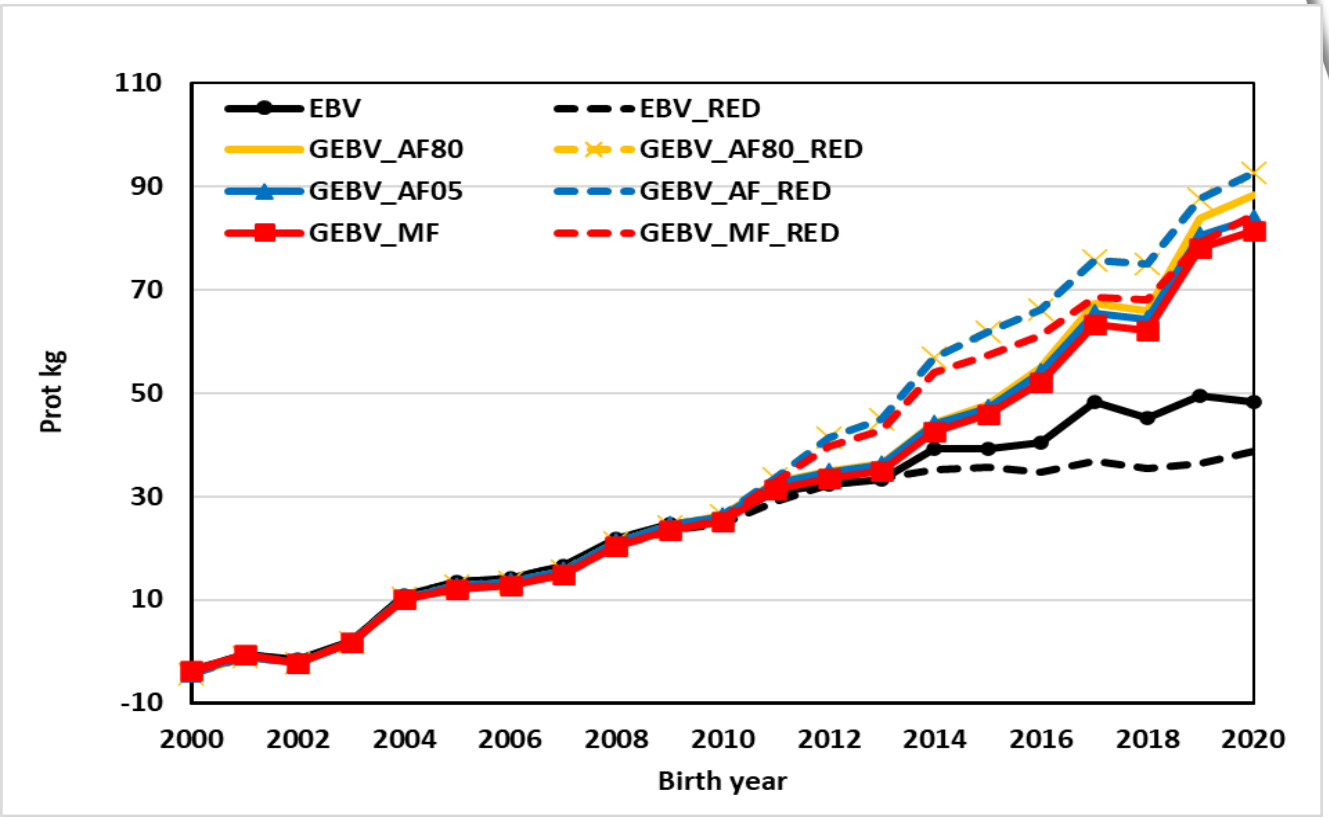
Regression of (G)EBV
on PA or GEBV_red

Mendelian sampling term bulls (protein)



AF80 - ssGTBLUP with QP and RPG 30 and AF 1980
AF05 – ssGTBLUP with QP and RPG 30 and AF 0.5
MF – ssGTBLUP with MetaFounders and RPG 30

Protein trend Nordic Holstein bulls
Full vs Reduced runs



GEBV_AF80 - ssGTBLUP with QP and RPG 30 and AF 1980
GEBV_AF05 – ssGTBLUP with QP and RPG 30 and AF 0.5
GEBV_MF – ssGTBLUP with MetaFounders and RPG 30

Conclusions

- ssGTBLUP with AF 0.5 is easy to use
  – no need to calculate AF with a different program

- ssGTBLUP with base population AF
  - need to calculate AF with e.g., Bpop program and decide what base AF to use (we used AF from 1980s)
    - Theoretically more correct as base population AF

These two above appear to have about the same inflation (b_1) and prediction reliability (R^2)

- ssGTBLUP with MF is theoretically more sophisticated way to combine pedigree and genomic information – also $\mathbf{A}^{-1}$ is modified according to genomic information
  - Does not increase the trend of young animals as much other single step methods tested
  – Marginally better validation results for inflation (b_1) and prediction reliability (R^2) than other approaches.

Thank you

Luke
NATURAL RESOURCES
INSTITUTE FINLAND


Luke
NATURAL RESOURCES
INSTITUTE FINLAND