SEGES INNOVATION

STØTTET AF Mælkeafgiftsfonden

Ministry of Food, Agriculture and Fisheries of Denmark gudp

Multi-breed genomic prediction using summary statistics in Dairy Cross

So far, we have only used estimated SNP effects and polygenic effects from pure breed genetic evaluations and available genotypes to compute genomic breeding values of crossbreds in our Dairy Cross routine evaluation. To maximize the use of all available information, we start to test a Bayesian approach by integrating all summary statistics from pure breed routine evaluations as well as phenotypes of crossbred animals to the data analysis for Dairy Cross. This is expected to result in higher reliability of genomic prediction for crossbred performance compared to the existing applied method.

Statistic model

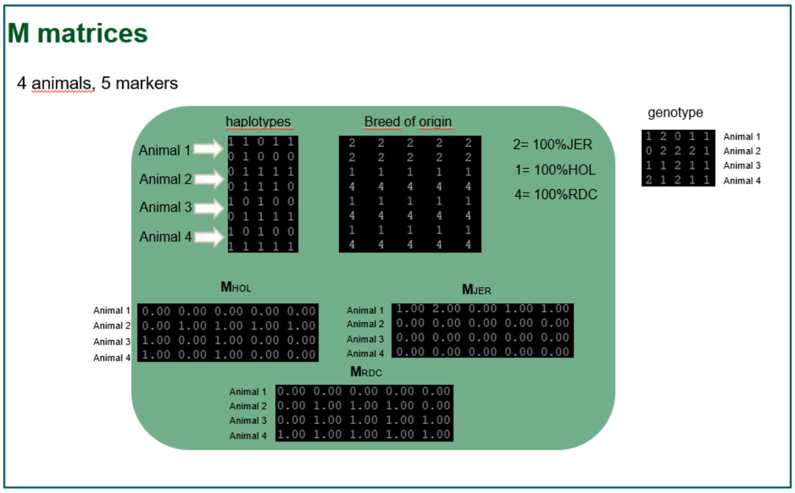
The model to estimate breed-specific SNP effects using breed-of-origin (BOA) is as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{b} + \boldsymbol{M_{HOL}}\boldsymbol{\beta_{HOL}} + \boldsymbol{M_{JER}}\boldsymbol{\beta_{JER}} + \boldsymbol{M_{RDC}}\boldsymbol{\beta_{RDC}} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of phenotypes (milk, yield or protein) of the n crossbred animals $(n \times 1)$, μ is the general mean, $\boldsymbol{X}$ is the matrix of breed proportions $(n \times 3)$, $\boldsymbol{b}$ is the vector of fixed breed effects (3×1), $\boldsymbol{M_{HOL}}$, $\boldsymbol{M_{JER}}$ and $\boldsymbol{M_{RDC}}$ are the matrices of breed specific allele content of SNPs ($n \times l$ where l is the number of SNPs) for HOL, JER and RDC, respectively. The entry at a locus in, for instance $\boldsymbol{M_{HOL}}$, for an animal is the number (0, 1 or 2) of counted alleles A originated from HOL. That is, when the animal has no allele originating from HOL, or when a HOL animal has an aa genotype, the corresponding entry is zero. The same applies to matrices $\boldsymbol{M_{JER}}$ and $\boldsymbol{M_{RDC}}$. The $\boldsymbol{\beta_{HOL}}$, $\boldsymbol{\beta_{JER}}$ and $\boldsymbol{\beta_{RDC}}$ are SNP effects for HOL, JER and RDC respectively, and $\boldsymbol{e}$ is the vector of residuals (Karaman et al., 2021).

BOA

As we do for monthly routine evaluation for Dairy Cross, the pedigree is traced back for 5 generations for the genotyped crossbred cows. The crossbred animals are included if sire and maternal grandsire is either HOL, JER or RDC and dam is either HOL, JER, RDC or crossbred. Genotypes are extracted on these crossbred cows and their genotyped purebred ancestors. Genotypes are imputed and phased using FImpute v2.2. Assignment to BOA is done by the Allor program (Eiríksson et al., 2021). The matrices $\boldsymbol{M_{HOL}}$, $\boldsymbol{M_{JER}}$ and $\boldsymbol{M_{RDC}}$ matrices are formed using a self-written script in Fortran (Figure 1).



M matrices

4 animals, 5 markers

haplotypes

```
1 1 0 1 1
0 1 0 0 0
0 1 1 1 1
0 1 1 1 0
1 0 1 0 0
0 1 1 1 1
1 0 1 0 0
1 1 1 1 1
```

Breed of origin

```
2   2   2   2   2
2   2   2   2   2
1   1   1   1   1
4   4   4   4   4
1   1   1   1   1
4   4   4   4   4
1   1   1   1   1
4   4   4   4   4
```

genotype

```
1 2 0 1 1   Animal 1
0 2 2 2 1   Animal 2
1 1 2 1 1   Animal 3
2 1 2 1 1   Animal 4
```

MHOL

```
Animal 1  0.00 0.00 0.00 0.00 0.00
Animal 2  0.00 1.00 1.00 1.00 1.00
Animal 3  1.00 0.00 1.00 0.00 0.00
Animal 4  1.00 0.00 1.00 0.00 0.00
```

MJER

```
Animal 1  1.00 2.00 0.00 1.00 1.00
Animal 2  0.00 0.00 0.00 0.00 0.00
Animal 3  0.00 0.00 0.00 0.00 0.00
Animal 4  0.00 0.00 0.00 0.00 0.00
```

MRDC

```
Animal 1  0.00 0.00 0.00 0.00 0.00
Animal 2  0.00 1.00 1.00 1.00 0.00
Animal 3  0.00 1.00 1.00 1.00 1.00
Animal 4  1.00 1.00 1.00 1.00 1.00
```

Figure1. The construction of $\boldsymbol{M_{HOL}}$, $\boldsymbol{M_{JER}}$ and $\boldsymbol{M_{RDC}}$ based on the haplotype phasing and BOA.

Bayesian analysis

A Bayesian approach was considered in the parameter estimation, which requires assigning prior distributions to the unknowns of the model. When all the data from crossbreds and the summary statistics from the pure breeds are available, these two information sources can be integrated within a Bayesian framework, such that the summary statistics are used to form prior distributions for the model parameters. Normal distribution priors are assigned for the mean, breed effects and SNP effects for each breed b and residuals.

$$\mu \sim N\left\{\frac{1}{n_{HOL}+n_{JER}+n_{RDC}}\left(n_{HOL}\bar{y}_{HOL}+n_{JER}\bar{y}_{JER}+n_{RDC}\bar{y}_{RDC}\right),\frac{1}{n_{HOL}+n_{JER}+n_{RDC}}\sigma_e^2\right\}$$

$$b \sim N\left\{\begin{bmatrix} n_{HOL} & 0 & 0 \\ 0 & n_{JER} & 0 \\ 0 & 0 & n_{RDC} \end{bmatrix}^{-1}\begin{bmatrix} n_{HOL}\bar{y}_{HOL} \\ n_{JER}\bar{y}_{JER} \\ n_{RDC}\bar{y}_{RDC} \end{bmatrix},\begin{bmatrix} n_{HOL} & 0 & 0 \\ 0 & n_{JER} & 0 \\ 0 & 0 & n_{RDC} \end{bmatrix}^{-1}\sigma_e^2\right\}$$

$$\beta_b \sim N\left\{\left[PEC(\widetilde{\beta_b})^{-1}-B_b^{-1}\right]^{-1}\left[PEC(\widetilde{\beta_b})^{-1}\tilde{\beta}_b\right],\left[PEC(\overline{\beta_b})-B_b^{-1}\right]^{-1}\right\}$$

$e \sim N(0, D\sigma_e^2)$ [D_a: identity matrix]

The variance of SNP effects and variance of error effects were further assigned a scaled inverse chi-square prior with a number of degrees of freedom (v) and a scale (S) parameter:

$\sigma_{\beta_i}^2 \sim \chi^{-2}(v_{\beta_i}, S_{\beta_i})$

$\sigma_e^2 \sim \chi^{-2}(v_e, S_e)$

We can see that we need to prepare the following data from the pure breed evaluations to form the priors for the Bayesian analysis:

1. the number of animals from each breed (n_b where b represents breed b)
2. mean phenotype of each breed ($\bar{y}_b$)
3. prediction error covariances of estimated SNP effects for each pure breed ($PEC(\widetilde{\beta_\iota})$)
4. random residual variance (σ_e^2)
5. variance of SNP effects for each breed (B_b) which equals $\frac{var(\boldsymbol{DGV})}{\sum 2p(1-p)}$ where $\boldsymbol{DGV}$ is the direct genomic values for breed b and p is the allele frequency of allele one at a locus in the population of breed b.

The program to prepare data 1, 2 and 5 is ready. We are currently working on 3 and 4, which requires some changes in purebred routine evaluations to obtain the output of prediction error covariances and estimated residual variance.

We started with the assumption that SNP effects are breed-specific and uncorrelated across the breeds. The priors can also be assigned such that the marker effects are breed-specific but correlated between the breeds. That is, a multivariate normal distribution is assigned for each sub-vector of SNP effects. This will be tested after the test with the assumption of uncorrelated SNP effects across the breeds.

The Markov-chain Monte Carlo (McMC) algorithm will be used to obtain samples of each parameter from its full-conditional posterior distribution. The chain length for the analyses consisted of 50,000 cycles, of which the first 10,000 were discarded as burn-in. Every 10th sample of the post burn-in cycles will be kept for posterior analysis, yielding 4000 posterior samples. The mean value of the posterior samples will be used as the estimate of each parameter, and a program (written in R) to compute this mean value has been done.

After estimating marker effects, the DGV of crossbred animals will be calculated by multiplying the breed-specific matrices in BOA ($\boldsymbol{M_{HOL}}$, $\boldsymbol{M_{JER}}$ and $\boldsymbol{M_{RDC}}$) with estimated marker effects (breed-specific SNP effects in BOA), adding a fixed breed contribution $\boldsymbol{X\tilde{b}}$.

References

Eiríksson, J.H., Karaman, E., Su, G. et al. Breed of origin of alleles and genomic predictions for crossbred dairy cows. Genet Sel Evol 53, 84 (2021). https://doi.org/10.1186/s12711-021-00678-3.

Karaman, E., Su, G., Croue, I. and Lund, M. S. 2021. Genomic prediction using a reference population of multiple pure breeds and admixed individuals. Genet. Sel. Evol. 53:46. https: / / doi .org/ 10.1186/ s12711 - 021 -00637 -y.