

Implementation of BOA model in Dairy Cross

Huiming Liu

SEGES Innovation, Avlsværdivurdering, Kvæg

Background

In the official genetic evaluation, the BOM model is used to calculate genomic estimated breeding values (GEBV) for crossbred animals. This calculation relies solely on the information from purebred animals, specifically the SNP solutions and the breed of origin of alleles. In contrast, the model currently under examination is the BOA model. Like the BOM model, the BOA model also requires the tracing of the breed of origin of alleles. However, it additionally allows for the incorporation of the performance data of crossbred animals. A key advantage of the BOA model is its capability to predict the breeding values of crossbred animals, even in cases where there is a lack of phenotypic data from a specific breed, such as in the case of MON crosses.

Data analyses

GENOTYPE (AUG 2023) AND PHENOTYPE DATA (305 DAYS DATA; AUG 2023)

Table 1: The number of HOL, JER, RDC and XXX animals with phenotype data and with genotype data are:

	# animals with pheno	# animals with geno	# animals with both
HOL	5 905 415	502 912	108 177
JER	1 088 720	136 544	54 781
RDC	865 475	249 223	22 514
MON		175	
XXX	748 132	13 705	8 481

*We need to remove the individual with pheno of -99999

PEDIGREE

- Trace pedigree for 5 generations
- Replace the nav_id with id_nor

MAP

- Merge the SNP maps of HOL (46343 SNP markers), JER (41898 SNP markers), and RDC (46915 SNP markers) based on the names and positions of the SNP markers. This merging process should create a comprehensive map that encompasses all 47586 SNP positions across the three pure breeds.

IMPUTATION AND PHASING

- Extract the genotypes of the purebred animals with both phenotypes and genotypes.
- Extract the genotypes of XXX animals
 - The sire and maternal grandsire need to be HOL, JER, RDC and MON, and the mother needs to be HOL, RDC, MON or XXX.
- Impute the purebred animals breed wise first and then impute crossbred animals.
 - The imputation is conducted using FImpute 3, following a two-step process. Initially, imputation is separately executed for the purebred animals, with an additional 20k animals per breed within each purebred group to improve the imputation accuracy of the purebred ancestors. Subsequently, the crossbred animals are imputed and phased by using the imputed purebred genotypes as reference.

PRE-CORRECTION OF PHENOTYPES

Model description (detailed)							
Trait		Model					
Trait / effect	: S MLK	=	MAN_GR	+	KLV_A_M	+	K_ALDER + NAV_ID + tothet + e / W_M
Type	: NOR		F		F		R FR R W
Input no.	: R1		I1		I4		I10 R8
Random mat. no.:							1

Table 2: the fixed effects included in the model

Name in phenotype	Explanation
MAN_GR (management group)	Herd*time*lactation
KLV_A_M	Calving*year*month
K_ALDER	Calving*age
Tothet	Heterozygosity

- The phenotypes of each trait were then corrected by subtracting fixed effects and the fixed regression (which equals EBV+error) for all individuals and were referred to corrected phenotypes.
- Run DMU again to get estimated breeding values (corrected phenotype~NAV_ID+e) and compare it with EBV from the previous model. The correlation of EBV from the two runs is very high (>0.99).

STEPS:

Step 1: Run the model for purebred animals and obtain the summary statistics.

Statistical model for purebred animals (HOL for example)

$$y = 1\mu + Z_{HOL}\beta_{HOL} + e$$

where y is the vector of corrected phenotypes (milk, yield or protein) of the n_{HOL} purebred animals ($n_{HOL} \times 1$), μ is the general mean, Z_{HOL} is the imputed genotype of HOL. The β_{HOL} is the vector of SNP effects for HOL, and e is the vector of residuals (Karaman et al., 2021).

The allele frequencies p_j and q_j were calculated for HOL, JER and RDC, where j represents the breed id.

A normal distribution prior was assigned for each SNP effects with mean of 0 and variance that equals $\text{var}(ebv_j)/\text{sum}(2p_jq_j)$ for each breed. The error variance was calculated as $\text{var}(\text{error}_j)$. ebv and error were obtained from DMU model for getting corrected phenotypes.

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

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

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

Where j represents the marker id.

The output from the Bayesian model includes the snp solutions $\hat{\beta}_j$ (column means of betaM1Out), prediction error variance $PEV(\hat{\beta}_{HOL})$ of all the markers (column variance of betaM1Out), estimated error variance (varEOut) and the SNP variance (varM1Out). We run the same model also for JER and RDC.

The direct genomic values (DGV) for breed j were calculated by multiplying genotypes with the estimated SNP effect $\hat{\beta}_j$.

We extracted the DGV from the latest official test for purebred animals and calculated its correlation with DGV with new model.

Table 3. Preliminary results on correlations of DGV for purebred animals

Breed	Number of animals for comparison	Cor(DGV_new,DGV_official)		
		Milk	Protein	Fat
HOL	107 980	0.86	0.78	0.77
JER	54 662	0.90	0.90	0.97
RDC	22 472	0.78	0.69	0.69

The results look OK, because in official test all the three lactations were used, whereas only the first lactation was used when implementing BOA model.

STEP 2: Calculate GEBV in crossbreds.

Table 4. Scenarios in the analyses

Scenarios	Information used	Extra explanation	Model
1	SumStat H/J/R		BOM
2	SumStat H/J/R		BOA
3	SumStat (H/R) +XXX	JER -> MON	BOA_MON
4	SumStat(H/R) +XXX + remove most of JER genotypes	JER -> MON; only 140 JER genotypes were used for breed of origin because we only have genotypes of 140 MON sires; only run for milk	BOA_only140

Statistical model for crossbred animals

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

$$y = 1\mu + Xb + M_{HOL}\beta_{HOL} + M_{JER}\beta_{JER} + M_{RDC}\beta_{RDC} + e,$$

where y is the vector of phenotypes (milk, yield or protein) of the n crossbred animals ($n \times 1$) in the reference population, μ is the general mean, X is the matrix of breed proportions ($n \times 3$), b is the vector of fixed breed effects (3×1), M_{HOL} , M_{JER} and 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. Initially, the entry at a locus in, for instance M_{HOL} , for an animal is the number (missing, 0, 1 or 2) of counted alleles A originated from HOL. That is, when the animal has no allele originating from HOL, the corresponding entry is missing. When a HOL animal has an aa genotype, the corresponding entry is zero. The same applies to matrices M_{JER} and M_{RDC} . Subsequently, these matrices were further centered (see centering procedure section). The β_{HOL} , β_{JER} and β_{RDC} are SNP effects for HOL, JER and RDC respectively, and e is the vector of residuals (Karaman et al., 2021).

Breed of origin (BOA) assignment

Assignment of BOA is done by the Allor program (Eiríksson et al., 2021). The matrices M_{HOL} , M_{JER} and M_{RDC} matrices are formed using a self-written script in Fortran.

The centering procedure is as follows

The denominator is the sum of each column in breed-specific matrices. That is the total number of alleles coming from each breed for each locus (Figure 1).

Breed specific probability matrices

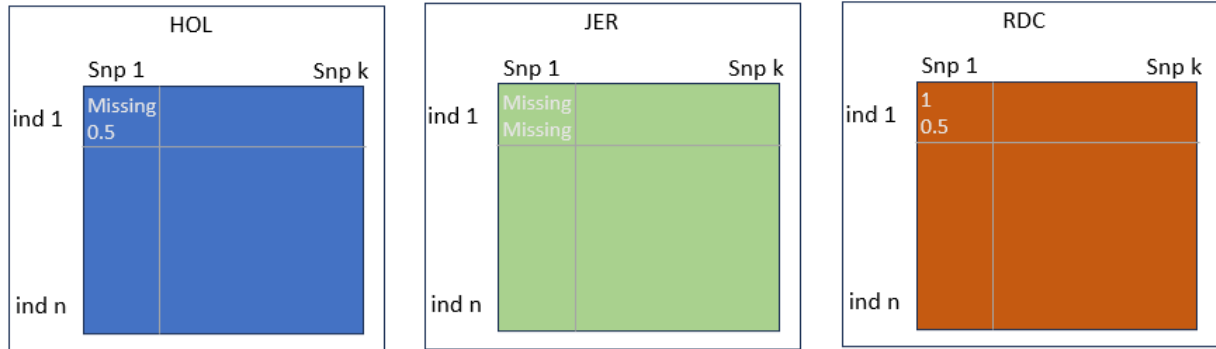


Figure 1. Breed specific probability matrices from AllOr program.

The numerator is the sum of each column of the Hadamard product of the breed-specific probability matrix and the haplotype (Figure 2). This is the total number of 'A' alleles for each breed (HOL, for example).

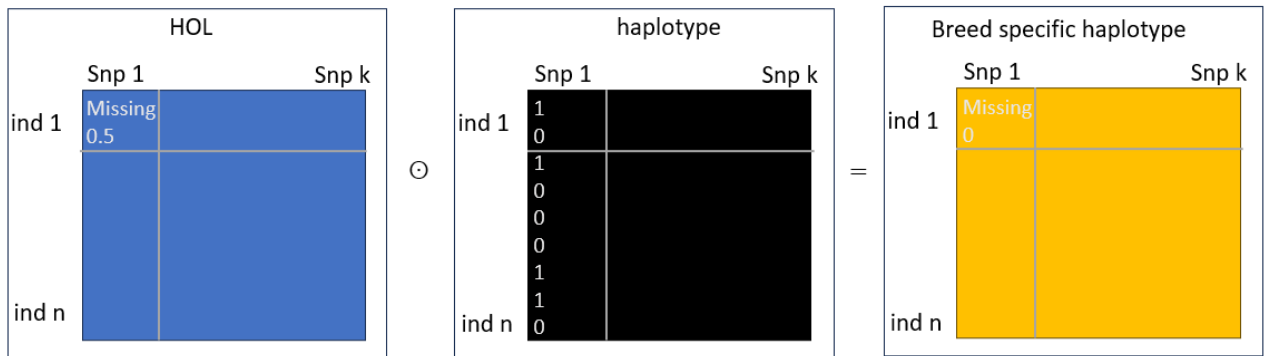


Figure 2. Construction of breed specific haplotype.

The allele frequencies for centering: $p_{HOL} = \frac{Numerator}{Denominator}$. Then in the breed specific haplotype matrix, the missing values were replaced with " p_{HOL} ". Then the two rows with haplotypes for each animal were summed to get genotype matrix M_{HOL} (Figure 3).

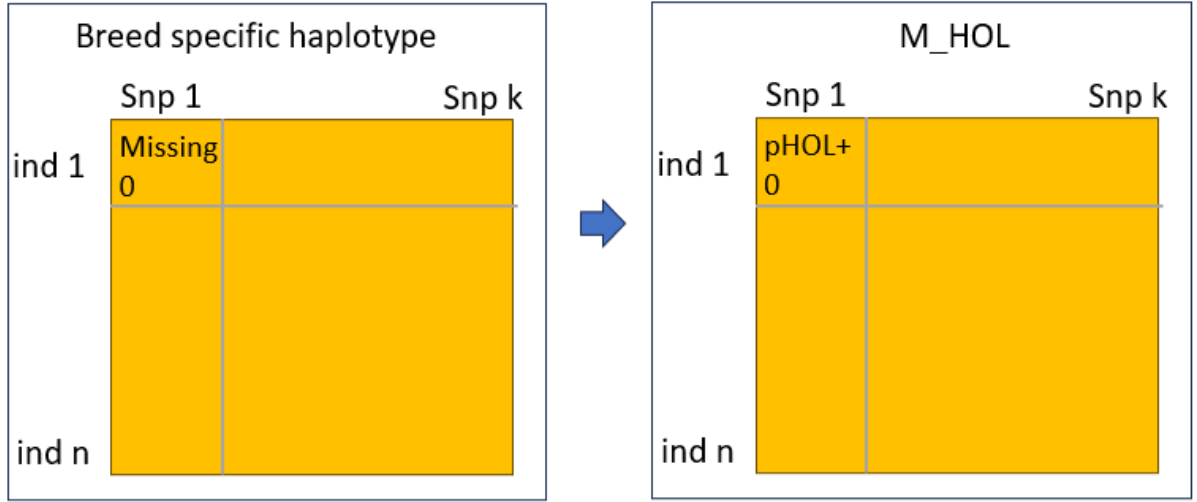


Figure 3. Construction of breed specific genotype matrix.

Finally, we subtracted 2pHOL for matrix M_{HOL} .

The use of summary statistics

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 pure breeds 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.

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

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

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

$$\beta_b \sim N(0, \sigma_{\beta_i}^2)$$

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

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

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

We need 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 ($PEV(\tilde{\beta}_b)$)
4. random residual variance (σ_e^2)
5. variance of SNP effects for each breed (B_b) which equals $\frac{var(DGV)}{\sum 2p(1-p)}$ where 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 .

We used assumption that SNP effects are breed-specific and uncorrelated across the breeds. The population was split into reference and testing populations based on the calving time. The animals with a calving month before 2021-04 were used as reference animals and the remaining animals were used as testing animals. We ran the Bayesian model using phenotype and genotype of the reference animals.

For the Bayesian models, Markov-chain Monte Carlo (MCMC) algorithm were 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 burn-in cycles will be kept for posterior analysis, yielding 4000 posterior post 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 GEBV of crossbred animals were calculated by multiplying the original breed-specific matrices in BOA (M_{HOL} , M_{JER} and M_{RDC}) with estimated marker effects (breed-specific SNP effects in BOA), adding a fixed breed contribution $X\tilde{b}$.

$$GEBV = X\hat{b} + M_{HOL}\hat{\beta}_{HOL} + M_{JER}\hat{\beta}_{JER} + M_{RDC}\hat{\beta}_{RDC}$$

The same snp solutions $\hat{\beta}_j$ information from Bayesian analysis for purebred animals was implemented on BOM to make a fair comparison with BOA.

BOM model

$$• \quad GEBV_{BOM,i} = \underbrace{\sum_{b=1}^{N_b} \mu_b \frac{\sum s_{1,i,b} + \sum s_{2,i,b}}{2m}}_{\text{Intercept}} + \underbrace{\sum_{b=1}^{N_b} (v'_b(w_{i,1} \circ s_{1,i,b}) + v'_b(w_{i,2} \circ s_{2,i,b}))}_{\text{Snp solutions}} + \underbrace{a_i}_{\text{polygenic}}$$

Results

Table 5 The correlation between corrected phenotypes and GEBV in test population for milk.

Group	Number of animals	BOM	BOA	BOA_MON	BOA_140JER
		sumStat HOL JER RDC	sumStat HOL JER RDC +XXX	sumStat HOL RDC +XXX	sumStat HOL RDC Only 100 JER geno +XXX
All test animals	1533	0.60	0.60	0.60	0.59
> 50% JER	252	0.60	0.60	0.54	0.53
(JER x HOL)	(107)	0.59	0.58	0.48	0.49
(JER x RDC)	(6)	0.88	0.90	0.75	0.75
(JER x XXX)	(139)	0.51	0.50	0.47	0.45
> 50% HOL	1128	0.55	0.54	0.54	0.54
(HOL x RDC)	(364)	0.52	0.50	0.50	0.50
> 50% RDC	605	0.51	0.52	0.52	0.52

Table 6 The correlation between corrected phenotypes and GEBV in test population for protein.

Group	Number of animals	BOM	BOA	BOA_MON
		sumStat HOL JER RDC	sumStat HOL JER RDC +XXX	sumStat HOL RDC +XXX
All test animals	1533	0.46	0.47	0.46
> 50% JER	252	0.45	0.44	0.39
(JER x HOL)	(107)	0.47	0.46	0.36
(JER x RDC)	(6)	0.91	0.86	0.78
(JER x XXX)	(139)	0.34	0.31	0.30
> 50% HOL	1128	0.44	0.44	0.42
(HOL x RDC)	(364)	0.41	0.41	0.41
> 50% RDC	605	0.42	0.43	0.42

Table 7 The correlation between corrected phenotypes and GEBV in test population for fat.

Group	Number of animals	BOM	BOA	BOA_MON
		sumStat HOL JER RDC	sumStat HOL JER RDC +XXX	sumStat HOL RDC +XXX
All test animals	1533	0.44	0.42	0.42
> 50% JER	252	0.35	0.34	0.26
(JER x HOL)	(107)	0.45	0.46	0.39
(JER x RDC)	(6)	0.92	0.83	0.57
(JER x XXX)	(139)	0.25	0.24	0.20
> 50% HOL	1128	0.45	0.43	0.42
(HOL x RDC)	(364)	0.40	0.37	0.37
> 50% RDC	605	0.43	0.42	0.42

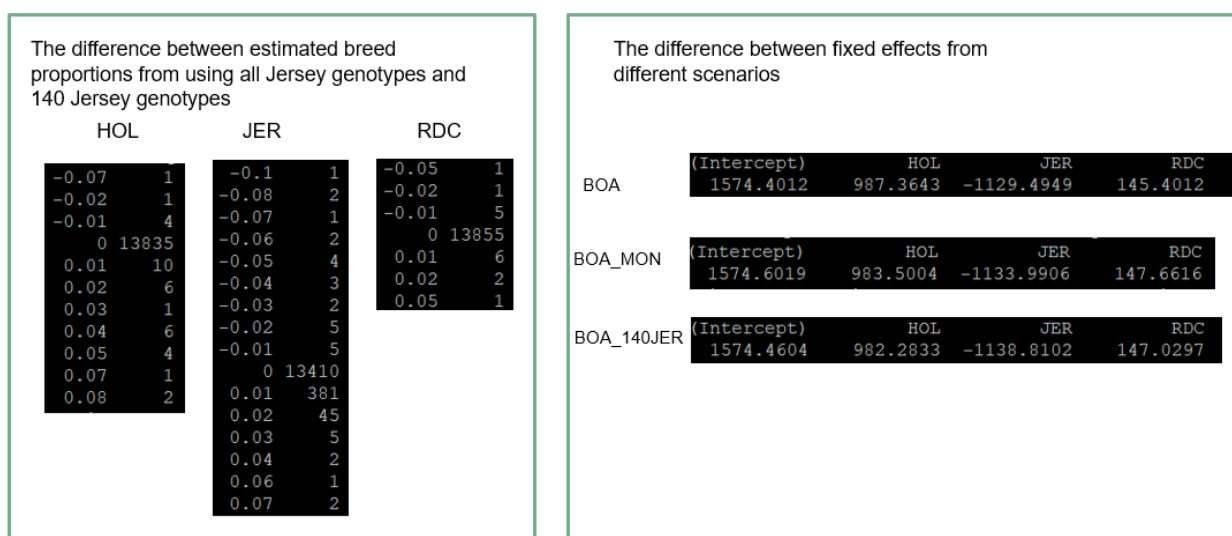
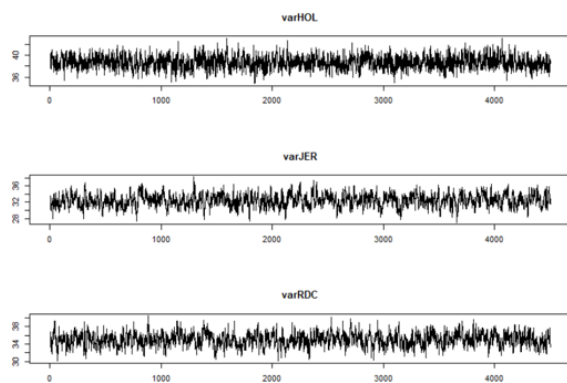


Figure 4. The difference between breed proportions from BOM and BOM_140JER scenarios, and the difference between fixed effects from different scenarios.

CONVERGENCE

BOA



BOA_MON

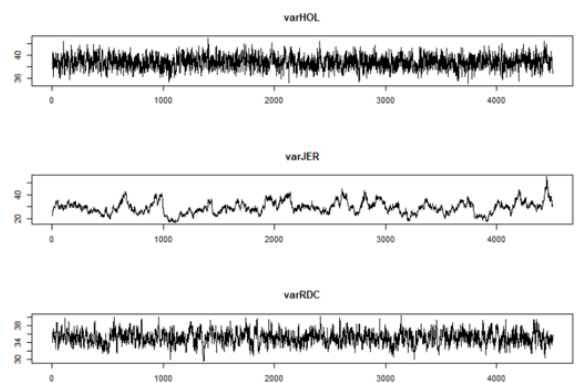


Figure 5. Trace plots SNP variances for BOA scenario.

Table 8. The correlation between GEBV from BOM and BOA models for all test animals

	BOA	BOA_MON	BOA_100JER
BOM	0.96	0.93	0.91
BOA		0.98	0.96
BOA_MON			0.99

Table 9. The correlation between GEBV from BOM and BOA models for test animals with >50% JER breed proportion

	BOA	BOA_MON	BOA_100JER
BOM	0.96	0.86	0.84
BOA		0.91	0.90
BOA_MON			0.99

Conclusions

- BOA performs as well as BOM when we have summary statistics from all purebreds and only limited number of crossbred animals.
- BOM cannot be used to evaluate cross when purebred info is missing, but BOA can be used.
- We only have 140 MON genotypes as purebred reference to trace the breed-of-origin. 140JER results show that BOA can definitely be used to evaluate MON crosses even though we don't have any MON phenotypes.

Development of pipeline to for GEBV calculations in MON crossbred animals

STEP 1 - Genotype Extraction

- Find all the genotyped crossbred animals.
- Tracing the pedigree back for 5 generations, and only keep the animals if the sire is HOL, JER, RDC or MON, the dam is HOL, JER, RDC, MON or XXX, and the maternal grand-sire is HOL, JER, RDC or MON.
- Find all the ancestors (obtained from the last step) with genotypes.
- Final all the genotyped MON males. We have limited number of MON animals (130 MON bulls) with genotypes, so we should use all of them for imputation and tracing the breed of origin.
- Merge the snp maps of HOL (46343 snp markers), JER (41898 snp markers) and RDC (46915 snp markers) according to the name and the positions of the snp markers, so that the merged map covers all the snp positions (47586 markers) of all the 3 pure breeds.

STEP 2 - Imputation using Flmpute 3

- The imputation was done in two steps:
 - Imputation for purebreds. We extracted 20 000 HOL, JER and RDC genotyped animals at similar age as the crossbred animal's parents. They were also typed with MD chips. These extra animals were used to improve the imputation of the purebreds. These extra animals were added to the genotyped ancestors, and then the imputation was done breed-wise. For MON purebred imputation, it is done based on all the available genotyped MON purebred males. It is possible that one or two markers are excluded from MON imputation, then the marker will be removed and a new map (Map_MON) is generated.
 - Imputation for crossbreds. The imputed genotypes from each purebred were used as reference for imputing the crossbreds. The map used for the imputation is Map_MON.

STEP 3 - Trace the breed of origin using AllOr program

- Utilize AllOr to split haplotypes by chromosome, incorporating purebred HOL, JER and RDC genotyped ancestors and all MON males are used as reference.
- Create breed code for the XXX animals by finding the information from breed codes in the IDs in the pedigree, considering the first 3 characters of the ID as breed code.
- Run AllOr chromosome by chromosome, generating matrices of M_{HOL} , M_{JER} , M_{RDC} and M_{MON} representing breed specific allele content of SNPs ($n \times l$ where l is the number of SNPs) for HOL, JER, RDC and MON, respectively.

STEP 4 - Pre-correction of phenotypes

The 305-day phenotypes (yield traits including milk, protein and fat) were corrected by subtracting fixed effects and the fixed regression (which equals EBV+error) for all individuals and were referred to corrected phenotypes.

STEP 5 - Generate summary statistics from HOL, JER and RDC

Identify all the purebred animals with both genotypes and phenotypes. Implement Bayesian model breed wise,

$$y = 1\mu + Z_{HOL}\beta_{HOL} + e$$

for example, to generate summary statistics for HOL, JER and RDC.

STEP 6 -Statistical model for crossbred animals

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

$$y = 1\mu + Xb + M_{HOL}\beta_{HOL} + M_{JER}\beta_{JER} + M_{RDC}\beta_{RDC} + M_{MON}\beta_{MON} + e,$$

where y is the vector of phenotypes (milk, yield or protein) of the n crossbred animals ($n \times 1$) in the reference population, μ is the general mean, X is the matrix of breed proportions ($n \times 4$), b is the vector of fixed breed effects (4×1), M_{HOL} , M_{JER} , M_{RDC} and M_{MON} 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 β_{HOL} , β_{JER} , β_{RDC} and β_{MON} are SNP effects for HOL, JER and RDC respectively, and e is the vector of residuals. The summary statistics for running this Bayesian model come only from HOL, JER and RDC.

STEP 7 - GEBV calculation

Calculate GEBV of crossbred animals by multiplying the original breed-specific matrices in BOA (M_{HOL} , M_{JER} , M_{RDC} and M_{MON}) with estimated marker effects (breed-specific SNP effects estimated from BOA), adding a fixed breed contribution $X\tilde{b}$.

$$GEBV = X\tilde{b} + M_{HOL}\hat{\beta}_{HOL} + M_{JER}\hat{\beta}_{JER} + M_{RDC}\hat{\beta}_{RDC} + M_{MON}\hat{\beta}_{MON}$$

For all the three traits, a rolling base population consisting of crossbred cows that are 1-7 years of age at the expected date of publication is applied. The average direct genomic breeding values (DGV; index unit) of these cows is calculated as Mean. The final GEBV for each trait is therefore calculated as $(GEBV - Mean) \times \text{Standardization factor of Holstein} + 100$.



SEGES Innovation

Agro Food Park 15, 8200 Aarhus N

T: +45 8740 5000 - F: +45 8740 5010 - E: info@seges.dk

Ansvar: Informationerne på denne side er af generel karakter og søger ikke at løse individuelle eller konkrete rådgivningsbehov. SEGES er således i intet tilfælde ansvarlig for tab, direkte såvel som indirekte, som brugere måtte lide ved at anvende notatets informationer.