ANAFI

Associazione Nazionale Allevatori Frisona Italiana

Validation experiences in Italian Holstein Genomic Selection

Jan-Thijs van Kaam

May 2010

Holstein bull genotypes available May 2010

- Reference population: Italian proven bulls and their (grand)sires.
- Genotypes: 54.001 Illumina SNPs.
- 80% oldest bulls used for estimation, 20% youngest used for validation.

	Genotyped samples
Total genotypes	3032
Replicates	- 86
Unique bulls	2946
Removed in data editing	- 50
Left after data editing	2896
Young bulls	- 307
Proven bulls (kg milk)	2589

Preparation of genotype data

- Selection of samples:
  - Free of known identity errors
  - Merge (if matching) or reject (if not matching) replicate samples
- Selection of SNPs by removing SNPs with undesirable characteristics:
  - Unscorable (i.e. many missing genotypes)
  - Monomorphic
  - Not mapped
  - Low minor allele frequency (MAF)
  - Low minor genotype frequency (MGF) (Low MGF doesn't always imply low MAF)
  - Large deviation from Hardy-Weinberg equilibrium
  - Highly correlated with other SNPs
  - Non-autosomal

SNP selection

SNP selection criteria	Flag per criteria	Flag only for this criteria
Monomorphic	3464	0
Non-autosomal or unmapped	1491	376
% Missing	1344	588
Mendelian	1328	44
Minor Genotype Frequency	10834	793
Minor Allele Frequency	9280	43
Hardy-Weinberg	3477	566
Correlation	9331	1299
X-linked	1178	81
Any flags / No flags	**14757**	**39244**

- Very little difference between more lax and more stringent SNP selection.
- ‘Bad’ SNPs have more false positive AND false negative associations.

Estimation of SNP effects

- SNP effects estimated using a single trait genomic BLUP approach based on a preconditioned conjugate gradient algorithm with residual updating.
- Speed: 29 single traits in 10 minutes total.
- Direct Genomic Value as sum of SNP effects.
- Composite traits are composed based on single trait results.
- Might add Gibbs sampling to get individual reliabilities based on posterior distribution.

Validation system

- Use oldest bulls for training with EDPs from 3 years ago.
- Check if the SNP effects predicted with the training bulls are capable to predict the realized EDPs of the youngest bulls.

EDP = Effective Daughter Performance (Deregressed EBV)

Validation criteria

1. The regression coefficient b for
   - $EDP_{2010} = a + b * DGV_{2007}$
   - b should be close to 1 (Interbull)
   - b <1 with selective genotyping (VanRaden)
2. The increase in R^2, i.e. effective daughter contributions, from DNA info:
   - $EDP_{2010} = a + b * PI_{2007}$
   - $EDP_{2010} = a + b_1 * PI_{2007} + b_2 * DGV_{2007}$

EDP = Effective Daughter Performance (Deregressed EBV),

DGV = Direct Genomic Value, PI = Pedigree Index

Regression of EDP on DGV

- EDP, EBV and DGV are all estimates of TBV.
- EDP are EBV but deregressed.
- It is suggested that regression of EDP on DGV should have a regression coefficient close to 1.
- In reality when regressing EDP on DGV the regression coefficients were around 0.60. Probably this will increase when more bull genotypes will be available.
- SNP coefficient and variance both determine size of SNP effect. Increasing **Ve/Vm** increases the **b** coefficient, and hence one can get to the desired value.
- Vm = Vg/sum(2pq)

Effect of variance ratio Ve/Vm

	Bulls		REL	REL			a+b*DGV		a+b*PI		a+b1*PI+b2*DGV	
Trait	Pred	Val	PI	GEBV	EDCg	h2	b	R2	b	R2	R2	Gamma
kg fat	1945	431	33.4	47.6	5.2	0.29	0.63	0.24	0.67	0.13	0.25	0.5*Ve/Vm
kg fat	1945	431	33.4	47.6	5.2	0.29	0.72	0.24	0.67	0.13	0.25	2.0*Ve/Vm
kg fat	1945	431	33.4	46.7	4.8	0.29	0.83	0.23	0.67	0.13	0.24	5.0*Ve/Vm
kg fat	1945	431	33.4	45.4	4.2	0.29	0.95	0.22	0.67	0.13	0.23	10.*Ve/Vm
% fat	1942	426	33.4	65.9	10.0	0.50	0.87	0.43	0.73	0.15	0.43	0.5*Ve/Vm
% fat	1942	426	33.4	65.2	9.6	0.50	0.98	0.42	0.73	0.15	0.42	2.0*Ve/Vm
% fat	1942	426	33.4	62.3	8.1	0.50	1.10	0.40	0.73	0.15	0.40	5.0*Ve/Vm
% fat	1942	426	33.4	58.4	6.3	0.50	1.22	0.36	0.73	0.15	0.36	10.*Ve/Vm
% prot	1942	426	33.4	55.4	5.2	0.50	0.79	0.37	0.87	0.20	0.39	0.5*Ve/Vm
% prot	1942	426	33.4	55.0	5.0	0.50	0.90	0.37	0.87	0.20	0.38	2.0*Ve/Vm
% prot	1942	426	33.4	53.8	4.6	0.50	1.03	0.36	0.87	0.20	0.37	5.0*Ve/Vm
% prot	1942	426	33.4	52.2	4.1	0.50	1.18	0.35	0.87	0.20	0.36	10.*Ve/Vm
fert	1666	420	31.1	44.9	28.7	0.05	0.67	0.13	0.73	0.08	0.14	0.5*Ve/Vm
fert	1666	420	31.1	44.3	27.0	0.05	0.95	0.13	0.73	0.08	0.14	2.0*Ve/Vm
fert	1666	420	31.1	41.3	20.0	0.05	1.19	0.12	0.73	0.08	0.13	5.0*Ve/Vm
fert	1666	420	31.1	38.5	13.6	0.05	1.43	0.11	0.73	0.08	0.11	10.*Ve/Vm

What does R^2 mean?

- While moving from 38416 SNPs to 43385 SNPs, USDA gained 0.4% reliability on average across traits. (Wiggans, 2010)
- Did they actually gain when they add 5000 parameters and hardly increase the reliability?
- R^2 will go to 1 also if one adds a million random variables to a model!
- Fitted variance ≠ ‘Explained’ variance
- Some sort of information criterion needed which accounts for the number of parameters/SNPs.

North American blending

- $GEBV = w_1*PA + w_2*subset\text{-}PA + w_3*DGV$
- Weights based on reliabilities
- Subset-PA based on **A** matrix with only the genotyped ancestors. Added because genotypes are only available on a subset of sires and grandsires.

GEBV = Genome Enhanced Breeding Value,

PA = Parental Average

How to blend?

- GEBV =

  (EDCc*EBV + EDCg*DGV)/(EDCc+EDCg)
- Should variances of conventional index and Direct Genomic Value be the same?
- Or should they differ based on level of reliability?
- What is best to present?

EDCc = Conventional Effective Daughter Contributions

EDCg = Genomic Effective Daughter Contributions

Conclusions

- The R^2 depends mostly on the number of genotypes available.
- More stringent or lax selection of SNPs had a minimal effect on R^2.
- Increasing the variance ratio, i.e. reducing the marker variance, increases the b-value, while R^2 remains nearly equal.

End

- Thank you for your attention.
- Questions?
- Acknowledgement:
  - Thanks to all organizations, projects and persons involved.