LfL

Bavarian State Research Center
for Agriculture

Approximation of Reliability in Single Step Models using the Interbull Standardized Genomic Reliability Method

M. Erbe, C. Edel, J. Dodenhoff, E. Pimentel, K.-U. Götz
Institute for Animal Breeding
Bavarian State Research Centre for Agriculture

Background & Motivation

- Background:
  - development of routine single step models
  - conformation traits in Fleckvieh breed with relatively simple model structure and manageable system size
  - unsolved: How to calculate/approximate reliabilities?

- Interbull Standardized Genomic Reliability Method
  - proposed by Liu et al. (2017)
  - designed not only for single step models, but looked applicable
- Aim of this study: Assess Liu et al.s' reliability approximation
  - in a small data set → whole system is invertible
  - in a routine-like data-set → aspects like computing time etc.

Description of data sets

	Small test data set	Routine-like data set
data basis	Subset of pig routine evaluation	Routine data set for conformation traits
# in pedigree	16‘500	3’300’000
# with phenotypes	4‘300	1’400’000
h^2 of the modeled trait	0.33	0.24
# with genotypes	5‘800	78’000
# with genotypes + phenotypes	180	5’500
# with genotypes + ≥ 1 non-genotyped, but phenotyped offspring	600	12’000

Steps of the approach of Liu et al.

Step 1:
Reliability of SNP genotypes

Step 2:
Reliability of DGV

Step 3:
Adjusting the theoretical reliabilities

Step 4:
Calculating the genomic EDC gain

Step 5:
Propagation of genomic information to non-genotyped relatives

Step 6:
Final reliabilities enhanced with genomic information

$$\boldsymbol{PEV}_g = \boldsymbol{Z}\begin{bmatrix} \boldsymbol{1'W^{-1}1} & \boldsymbol{1'W^{-1}Z} \\ \boldsymbol{Z'W^{-1}1} & \boldsymbol{Z'W^{-1}Z} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_{SNP}^2} \end{bmatrix}^{-1} \boldsymbol{Z'}$$

$$R^2_{DGV_i} = x_{imp}x_{poly}x_{val}R^2_{g_i}$$

$$\varphi_{gain_i} = \frac{R^2_{DGV_i}}{1 - R^2_{DGV_i}}\lambda - \frac{R^2_{A22_i}}{1 - R^2_{A22_i}}\lambda$$

$$\boldsymbol{PEV}_{prop} = approx\left(\begin{bmatrix} \boldsymbol{1'D^{-1}1} & \boldsymbol{1'D^{-1}K} \\ \boldsymbol{K'D^{-1}1} & \boldsymbol{K'D^{-1}K} + \boldsymbol{A^{-1}}\lambda \end{bmatrix}^{-1}\right)$$

genotyped: $\varphi_{final_i} = \varphi_{konv_i} + \varphi_{gain_i}$

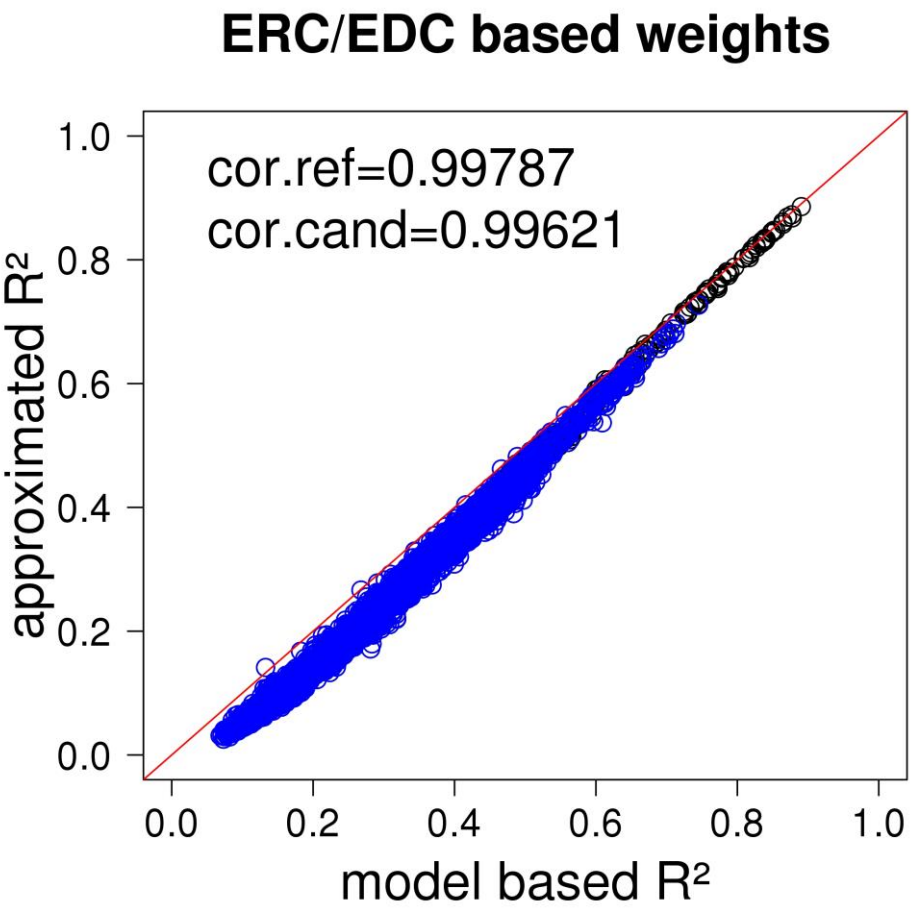
non-genotyped: $\varphi_{final_i} = \varphi_{konv_i} + \varphi_{prop_i}$

LfL Animal Breeding

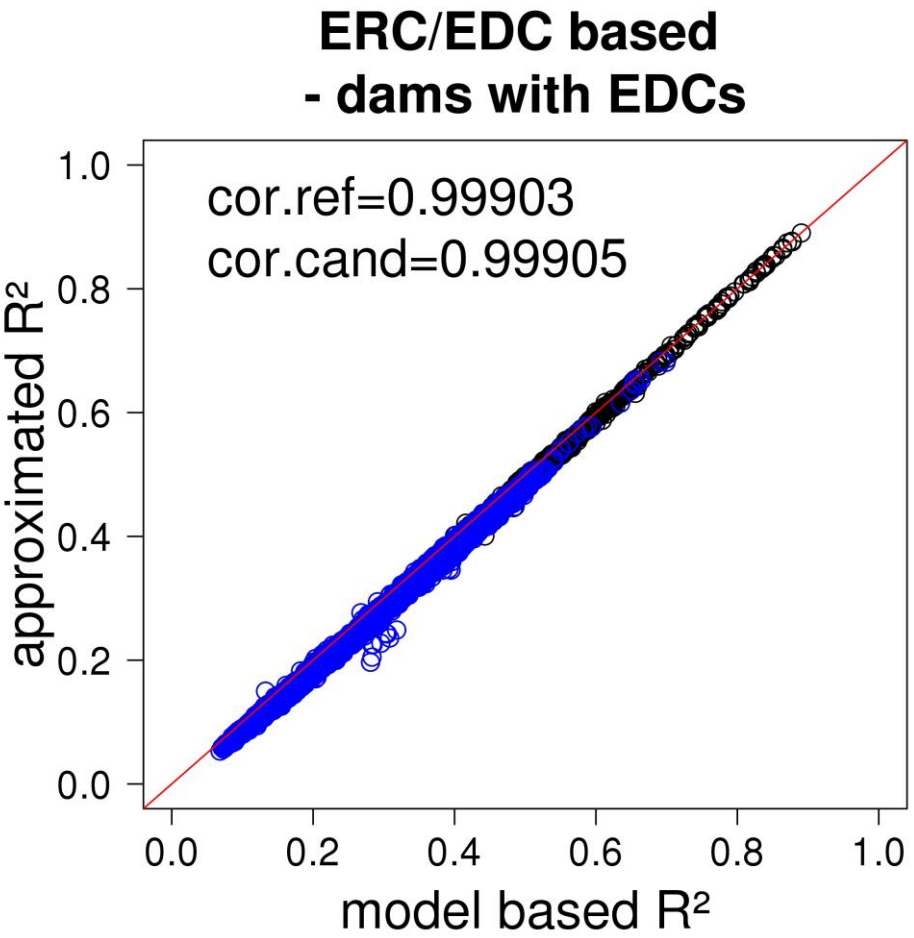
Weights / Reference set and their influence

- Who is reference individual?
  - not so easy to define in single step environments
  - in this approach: only genotyped individuals can be reference
- Diagonal elements for $\mathbf{W}^{-1}$
  - avoid double counting
  - information of non-genotyped, but phenotyped offspring into genotyped reference individuals
- EDC/ERC-based approach
  - supplementary document by Interbull reliability working group
  - ERC for genotyped females with phenotypes and EDC for genotyped bulls from non-genotyped daughters with phenotypes
- Check final reliabilities for genotyped individuals

Weights / Reference set and their influence

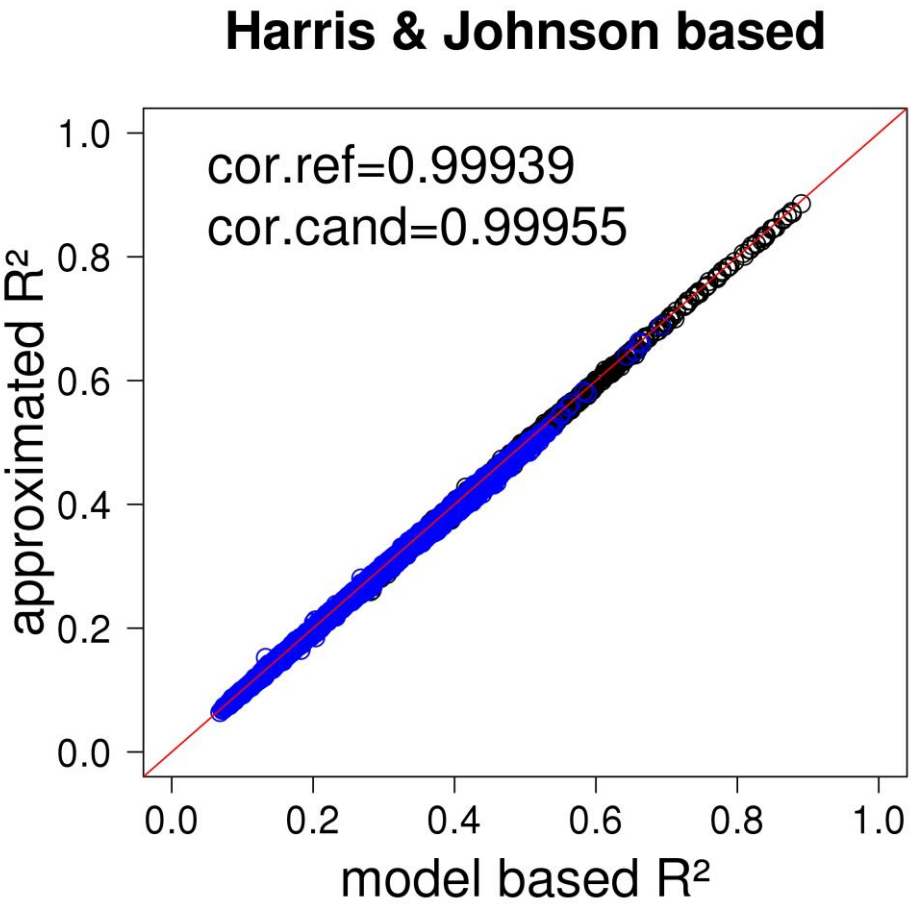


Weights / Reference set and their influence



- What could be missing?
  - information from non-genotyped, but phenotyped daughters not transmitted to genotyped dam

Weights / Reference set and their influence



- What could be missing?
  - information from non-genotyped, but phenotyped daughters not transmitted to genotyped dam
  - information from further generations not included in EDCs
  - → extend EDC calculation
  - → alternative: Harris & Johnson based approach e.g. as implemented in ApaX99

Steps of the approach of Liu et al.

Step 1:
Reliability of SNP genotypes

$$\boldsymbol{PEV}_g = \boldsymbol{Z}\begin{bmatrix}\boldsymbol{1'W^{-1}1} & \boldsymbol{1'W^{-1}Z} \\ \boldsymbol{Z'W^{-1}1} & \boldsymbol{Z'W^{-1}Z} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_{SNP}^2}\end{bmatrix}^{-1}\boldsymbol{Z'}$$

Step 2:
Reliability of DGV

Step 3:
Adjusting the theoretical reliabilities

$$R^2_{DGV_i} = x_{imp}x_{poly}x_{ral}R^2_{g_i}$$

Step 4:
Calculating the genomic EDC gain

$$\varphi_{gain_i} = \frac{R^2_{DGV_i}}{1 - R^2_{DGV_i}}\lambda - \frac{R^2_{A22_i}}{1 - R^2_{A22_i}}\lambda$$

Step 5:
Propagation of genomic information to non-genotyped relatives

$$\boldsymbol{PEV}_{prop} = approx\left(\begin{bmatrix}\boldsymbol{1'D^{-1}1} & \boldsymbol{1'D^{-1}K} \\ \boldsymbol{K'D^{-1}1} & \boldsymbol{K'D^{-1}K} + \boldsymbol{A^{-1}}\lambda\end{bmatrix}^{-1}\right)$$

Step 6:
Final reliabilities enhanced with genomic information

genotyped: $\varphi_{final_i} = \varphi_{konv_i} + \varphi_{gain_i}$

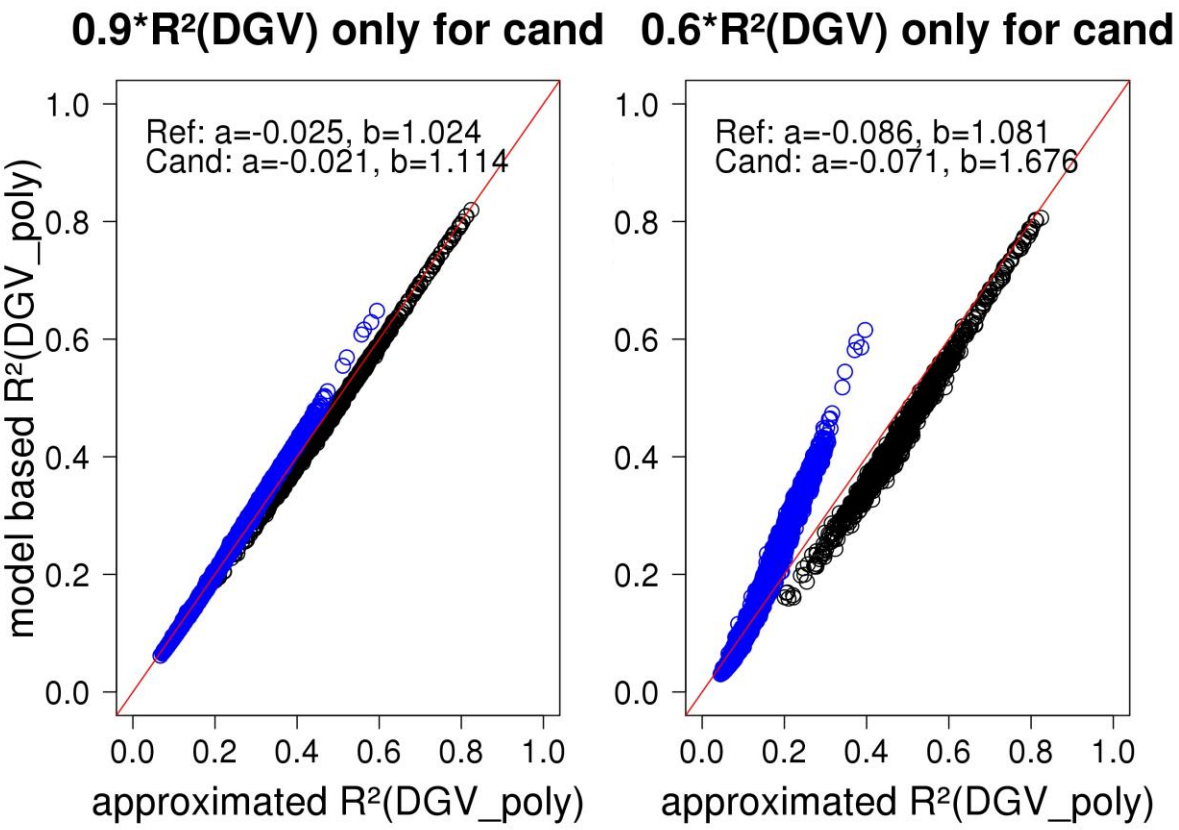
non-genotyped: $\varphi_{final_i} = \varphi_{konv_i} + \varphi_{prop_i}$

Polygenic contribution

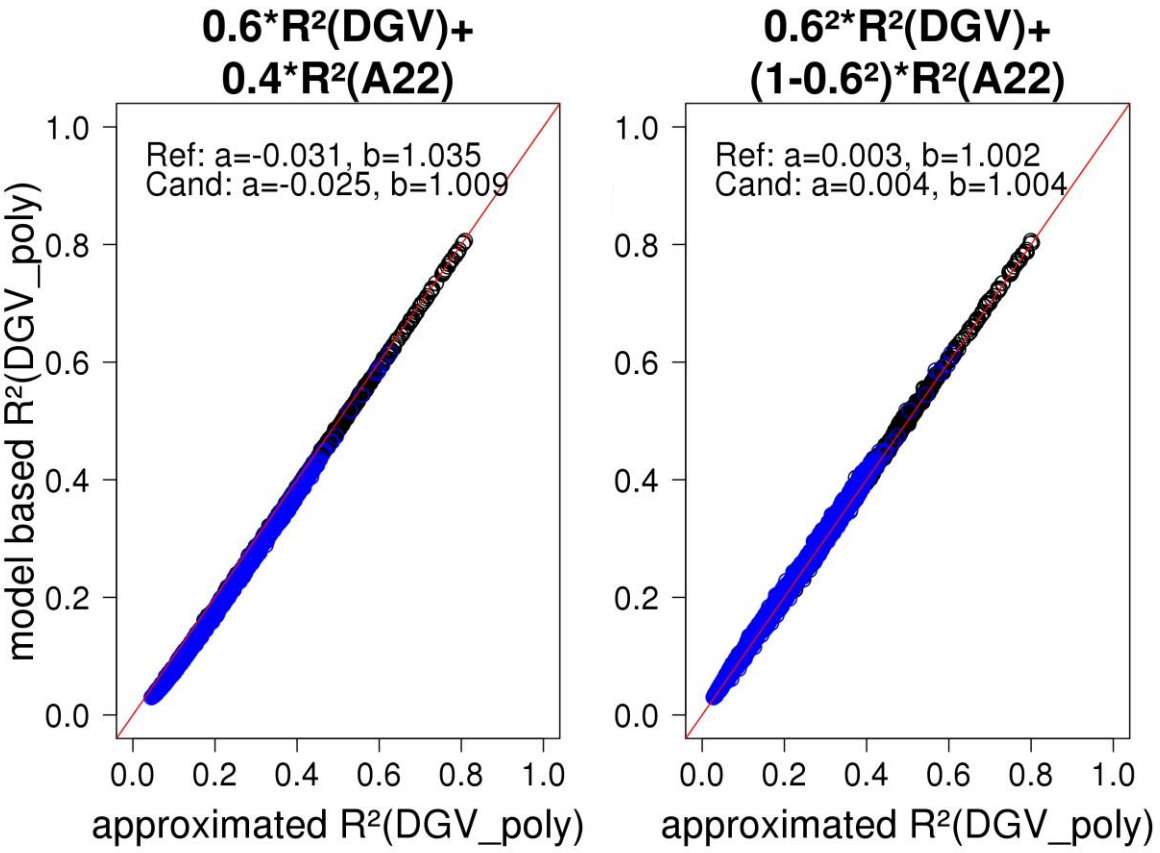
- Considering residual polygenic variation:
  - k is proportion of variance assumed to be not explained by markers
  - How to consider this when starting from a marker model?
  - proposal in Liu et al. (2017): $R^2_{DGVpoly\_approx} = \begin{cases} (1-k)R^2_{DGV} & for\ cand \\ R^2_{DGV} & for\ ref \end{cases}$

- Comparison of model-based $R^2_{DGVpoly}$ with $R^2_{DGVpoly\_approx}$

Polygenic contribution



Polygenic contribution



- Other approaches?
  - combination of raw DGV reliability and A22 reliability for all individuals
  - Proportion (1-k) and k
  - Proportion $(1-k)^2$ and $1-(1-k)^2$

Steps of the approach of Liu et al.

Step 1:
Reliability of SNP genotypes

Step 2:
Reliability of DGV

Step 3:
Adjusting the theoretical reliabilities

Step 4:
Calculating the genomic EDC gain

Step 5:
Propagation of genomic information to non-genotyped relatives

Step 6:
Final reliabilities enhanced with genomic information

$$\boldsymbol{PEV}_g = \boldsymbol{Z}\begin{bmatrix}\boldsymbol{1'W^{-1}1} & \boldsymbol{1'W^{-1}Z} \\ \boldsymbol{Z'W^{-1}1} & \boldsymbol{Z'W^{-1}Z} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_{SNP}^2}\end{bmatrix}^{-1}\boldsymbol{Z'}$$

$$R^2_{DGV_i} = x_{imp}x_{poly}x_{val}R^2_{g_i}$$

$$\varphi_{gain_i} = \frac{R^2_{DGV_i}}{1-R^2_{DGV_i}}\lambda - \frac{R^2_{A22_i}}{1-R^2_{A22_i}}\lambda$$

$$\boldsymbol{PEV}_{prop} = approx\left(\begin{bmatrix}\boldsymbol{1'D^{-1}1} & \boldsymbol{1'D^{-1}K} \\ \boldsymbol{K'D^{-1}1} & \boldsymbol{K'D^{-1}K} + \boldsymbol{A^{-1}}\lambda\end{bmatrix}^{-1}\right)$$

genotyped: $\varphi_{final_i} = \varphi_{konv_i} + \varphi_{gain_i}$

non-genotyped: $\varphi_{final_i} = \varphi_{konv_i} + \varphi_{prop_i}$

LfL Animal Breeding

Propagation to non-genotyped individuals

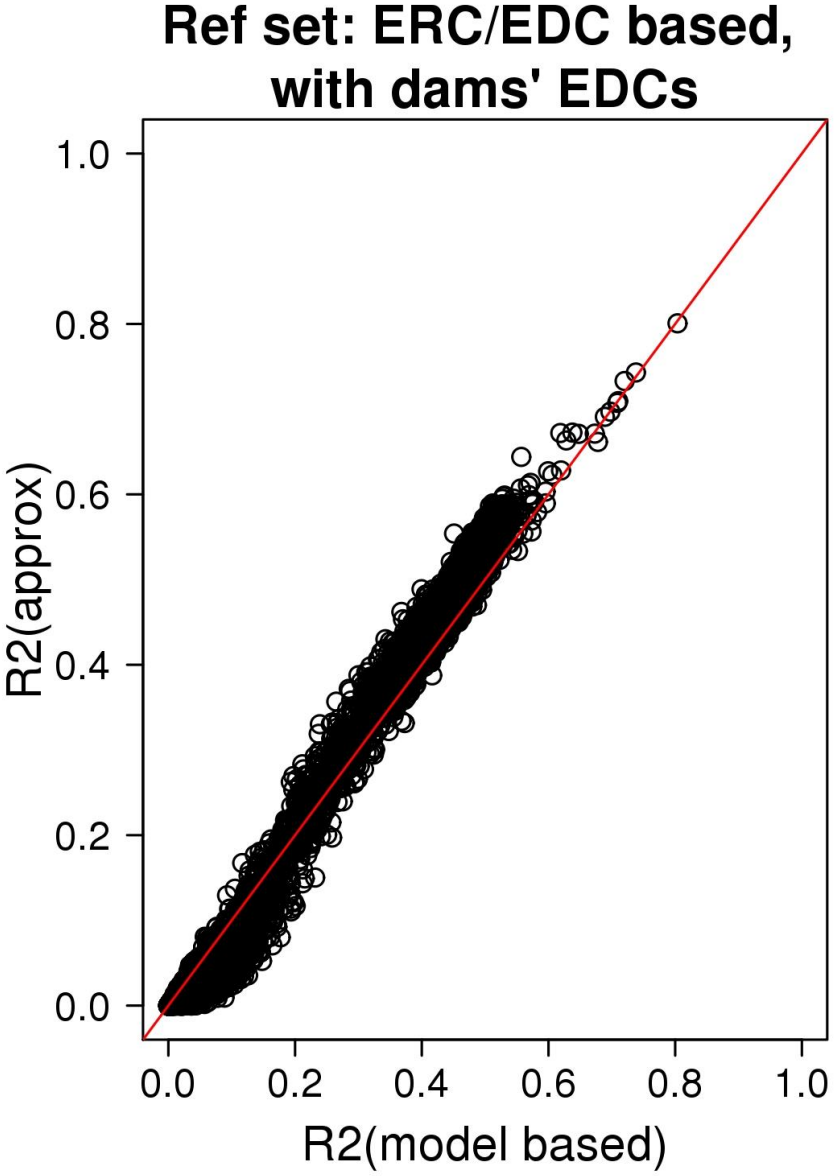
- optional step to propagate gain of genotyped individuals to non-genotyped ones

$$\boldsymbol{PEV}_{prop} = approx\left(\begin{bmatrix} \mathbf{1}'\boldsymbol{D}^{-1}\mathbf{1} & \mathbf{1}'\boldsymbol{D}^{-1}\boldsymbol{K} \\ \boldsymbol{K}'\boldsymbol{D}^{-1}\mathbf{1} & \boldsymbol{K}'\boldsymbol{D}^{-1}\boldsymbol{K} + \boldsymbol{A}^{-1}\lambda \end{bmatrix}^{-1}\right)$$

  - Weighting factors in $\boldsymbol{D}^{-1}$ are gains $\boldsymbol{\varphi}_{gain}$ from previous step.

- check final reliabilities for non-genotyped individuals

Propagation to non-genotyped individuals



LfL Animal Breeding

Implementation steps in routine-like data set

Step 1:
Reliability of SNP genotypes

Step 2:
Reliability of DGV

Step 3:
Adjusting the theoretical reliabilities

Step 4:
Calculating the genomic EDC gain

Step 5:
Propagation of genomic information to non-genotyped relatives

Step 6:
Final reliabilities enhanced with genomic information

Implementation steps in routine-like data set

Step 1:
Reliability of SNP genotypes

Step 2:
Reliability of DGV

Step 3:
Adjusting the theoretical reliabilities

given that validation
has been run before

Step 4:
Calculating the genomic EDC gain

Step 5:
Propagation of genomic information
to non-genotyped relatives

Step 6:
Final reliabilities enhanced
with genomic information

ApaX99

basic arithmetic with scripts in R

snp_blup_rel

LfL
Animal Breeding

Implementation steps in routine-like data set

Step 1:
Reliability of SNP genotypes

Step 2:
Reliability of DGV

Step 3:
Adjusting the theoretical reliabilities

Step 4:
Calculating the genomic EDC gain

Step 5:
Propagation of genomic information
to non-genotyped relatives

Step 6:
Final reliabilities enhanced
with genomic information

~ 30 min
computationally demanding

solved in 1-2 minutes,
very little CPU & memory

LfL
Animal Breeding

Implementation with snp_blup_rel

- snp_blup_rel for the routine example:
  - applied on a Linux-Server with 96 threads and 512 GB RAM
  - program used different number of threads
  - no writing of MME output (takes long)
  - Total # of genotyped individuals: 78k

# reference animals	# SNPs	peak virtual memory	time in total	time for inversion	time for reliabilities
78k	41k	48 GB	35 min	9 min	12 min
16k	41k	38 GB	25-30 min	10 min	8 min
11k	41k	38 GB	25 min	10 min	8 min

Conclusions/Outlook

- Results from small test data set:
  - Reference set definition and way of weighting have an impact on results.
  - Correct way of considering polygenic contribution?
  - very promising results for genotyped individuals
  - Propagation to non-genotyped individuals not satisfying in this data set.
- Computing issues from routine-like data set:
  - step 1 demanding, but feasible – best only from time to time and/or for a small number of traits
  - All other steps are less critical in terms of memory, CPU and time.
- More general/'philosophical' implementation questions:
  - Any solutions for summarizing different traits in one run?

$$\begin{bmatrix} \mathbf{1}'\boldsymbol{W}^{-1}\mathbf{1} & \mathbf{1}'\boldsymbol{W}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{W}^{-1}\mathbf{1} & \boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{Z} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_{SNP}^2} \end{bmatrix}$$

  - Integration of non-genotyped, but implicitly imputed individuals to the reference set?
  - What about continuous evaluations for candidates e.g. in short time intervals on database level?

Thank you for your attention!