Approximate single step genomic predictions for Norwegian Red cattle

Janez Jenko & Ismo Stranden

Interbull meeting

Lyon, 26. and 27. 8. 2023

Content

- Background
- Current single-step genomic prediction (ssGBLUP) for Norwegian Red
- Challenges & possible solutions
- Alternative methods to decrease computational costs
- Data & results
- Conclusions

Background

- We have introduced ssGBLUP for all traits in February 2016
  - 18,000 animals with genotype information
- We are genotyping approximately 35,000 animals each year
  - Today we have 217,000 animals genotyped
- We will soon reach RAM limits of our current computer

Current ssGBLUP for Norwegian Red

- Single trait single step mixed model equation (Christensen and Lund, 2010)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \lambda\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

where: $\lambda = \sigma_e^2/\sigma_a^2$

- The inverse relationship matrix is

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{G}_w)^{-1} - (\mathbf{A}_{22})^{-1} \end{bmatrix}$$

where: $\mathbf{A}^{-1}$ is based on pedigree relationships (sparse & easy to compute)

$(\mathbf{A}_{22})^{-1}$ is based on pedigree relationships for the genotyped animals

$(\mathbf{G}_w)^{-1} = ((1-w)\mathbf{G} + w\mathbf{A}_{22})^{-1}$ combines genomic information (dense & demanding to compute) and pedigree information with $w = 0.1$

How to solve this problem?

- Increase computer RAM
  - Expensive
  - Non-sustainable solution

- Remove genotypes from animals without phenotype
  - Challenging if most genotyped animals have at least one phenotype

- Alternative single step methods to decrease computational costs
  - **APY** (Misztal et al., 2014)
  - **SVD** (Ødegård et al., 2018)
  - ssGTBLUP (Mäntysaari et al., 2017)
  - ssSNPBLUP (Legarra and Ducrocq, 2012; or Liu et al. 2014)

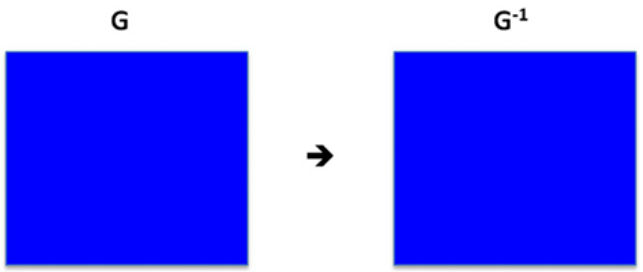
APY – Algorithm for Proven and Young

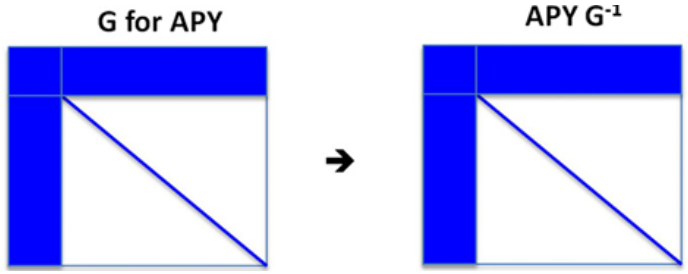
$$\mathbf{G}^{-1} \approx \begin{bmatrix} \mathbf{G}_{pp}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{pp}^{-1}\mathbf{G}_{py} \\ \mathbf{I} \end{bmatrix} \mathbf{M}^{-1} \begin{bmatrix} -\mathbf{G}_{yp}\mathbf{G}_{pp}^{-1} & \mathrm{I} \end{bmatrix}$$

where $\mathbf{G}$ is partitioned into proven/core (p) and young/non-core (y) animals

and $\mathbf{M}$ is a diagonal matrix with elements $m_{ii} = \left(\mathbf{G}_{yy} - \mathbf{G}_{yp}\mathbf{G}_{pp}^{-1}\mathbf{G}_{py}\right)$

- Computational costs
  - Core individuals: cubic computing and quadratic memory (as in ssGBLUP)
  - Noncore animals: linear computing and memory





geno | NORWEGIAN RED

SVD – Singular Value Decomposition

- Core animals are used to approximate correlations between markers
- When regularization is only a constant $\epsilon = 0.01$, $\mathbf{G}$ can be written as $\mathbf{G} = \mathbf{Z_c D Z_c'} + \mathbf{E}$ where $\mathbf{Z_c}$ is a centred marker matrix and scaling matrix $\mathbf{D} = \mathbf{I}\frac{1}{\mathrm{k}}$ with a scaling index $\mathrm{k} = 2\sum_{\mathrm{i}=1}^{\mathrm{m}} 0.5^2$ for m markers and $\mathbf{E}$ is $\epsilon\mathbf{I}$
- Using Woodbury matrix identity $\mathbf{G}^{-1}$ can be rewritten as $\mathbf{G}^{-1} = \frac{1}{\epsilon}\mathbf{I} - \mathbf{T}_\mathrm{C}'\mathbf{T}_\mathrm{C}$ where $\mathbf{T}_\mathrm{C} = \frac{1}{\epsilon}\mathbf{L}_\mathrm{C}^{-1}\mathbf{C}$ and the lower triangular matrix $\mathbf{L}_\mathrm{C}$ is Cholesky decomposition of $\mathbf{C}'\mathbf{C}\frac{1}{\epsilon} + \mathbf{I}\mathrm{k}$ that is $\mathbf{L}_\mathrm{C}\mathbf{L}_\mathrm{C}' = \mathbf{C}'\mathbf{C}\frac{1}{\epsilon} + \mathbf{I}\mathrm{k}$ where $\mathbf{C}$ is an approximation of $\mathbf{Z_c}$

Data & statistics

- Phenotypes, genotypes, and pedigree information from April 13, 2023 routine evaluation
  - 90 traits (29 single- or multi-trait mixed model equations)
  - 206,496 genotypes with 121,740 SNP markers

- Comparison statistics
  - Correlation between ssGBLUP predictions and approximate predictions
  - Linear regression coefficient and intercept when regressing predictions from ssGBLUP to the predictions from approximate approaches

Defining core individuals

- APY
  - Genotypes from 16,480 animals (AI sires, foreign animals, and animals with foreign sire)

- SVD
  - Genotypes from 5,186 AI sires
  - Singular values explaining 90% of genetic variation of the animals in the core
  - 43,917 components across genome (between 1029 and 2099 per chromosome)

Computational requirements

Preprocessing of:

$\mathbf{G}^{-1}$ for ssGBLUP and APY

$\mathbf{T_c}$ for SVD

Method	Memory (GB)	Time
ssGBLUP	670	23h 14min
APY	111	4h 21min
SVD	82	2h 03min

Solving mixed model equations

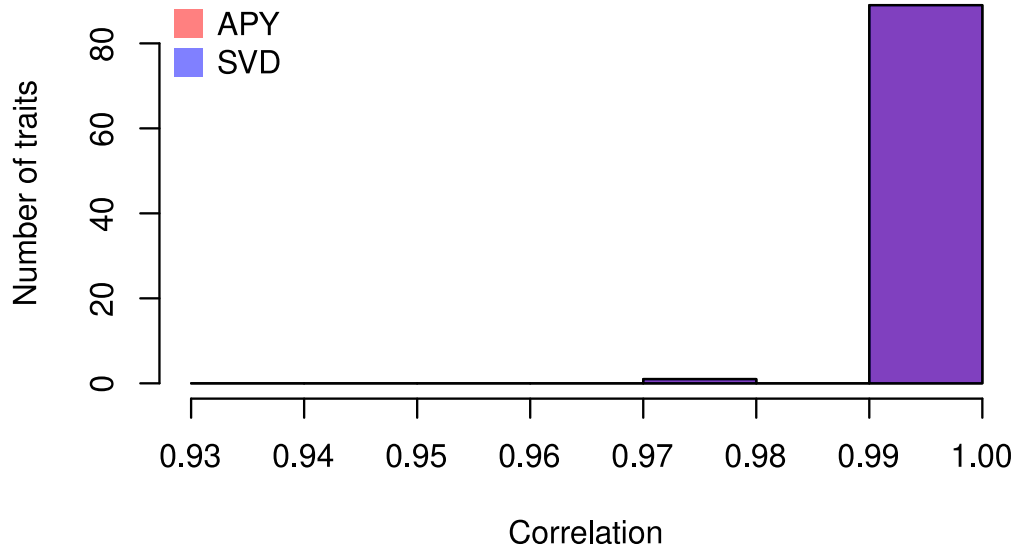
(RAM is not an issue here as relationship matrix is not stored in memory)

Method	Number of iterations	Time
ssGBLUP	320	21h 45min
APY	792	2h 31min
SVD	830	35h 6min

Results – correlations

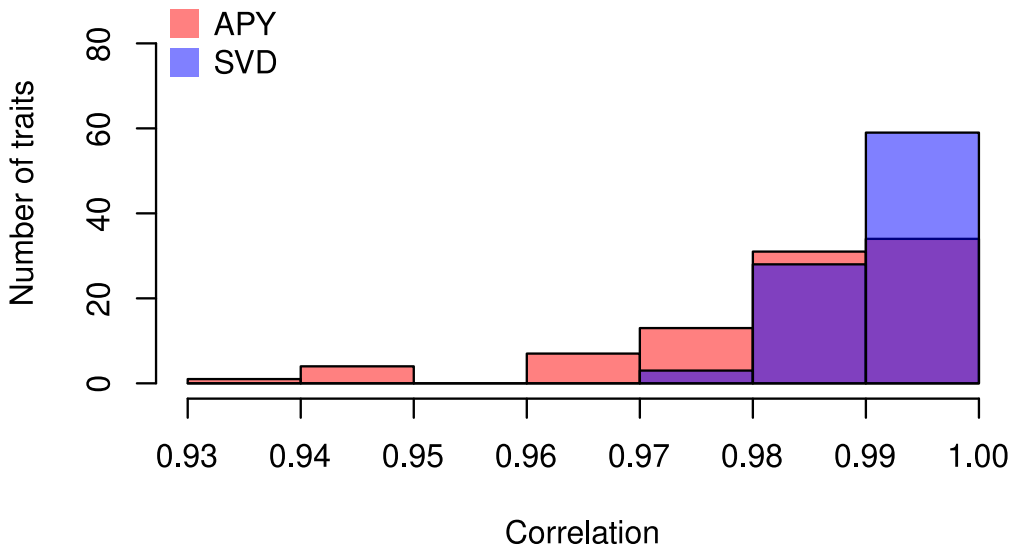
Individuals in pedigree

	Mean	SD	Min	Max
APY	0.998	0.003	0.976	1.000
SVD	0.997	0.003	0.971	1.000



Genotyped individuals born on October 1, 2021 or later

	Mean	SD	Min	Max
APY	0.983	0.013	0.940	0.995
SVD	0.990	0.004	0.977	0.995



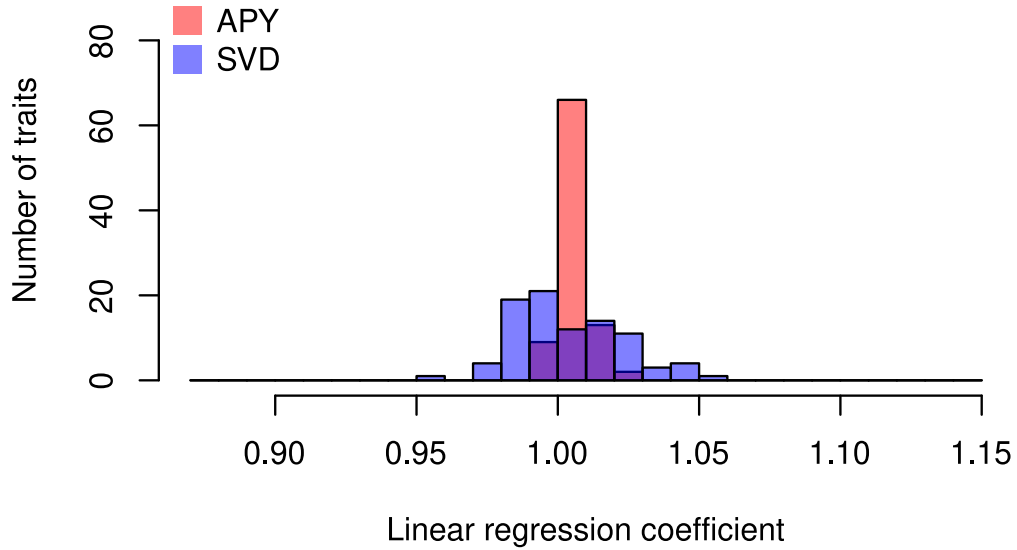
geno | NORWEGIAN RED

Results – linear regression coefficient

$$\text{ssGBLUP} \approx \mu + \beta * (\text{SVD or APY}) + e$$

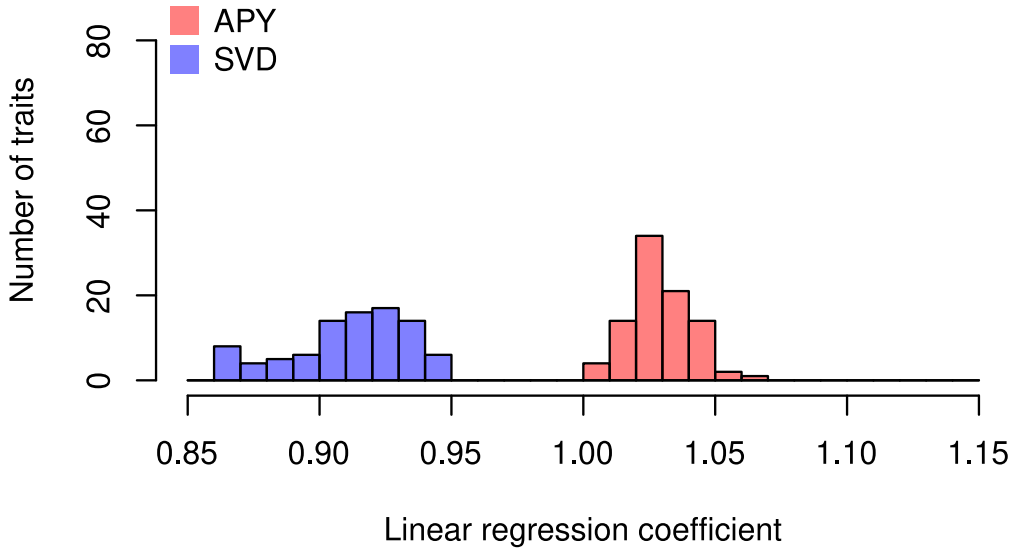
Individuals in pedigree

	Mean	SD	Min	Max
APY	1.006	0.006	0.993	1.027
SVD	1.004	0.019	0.953	1.055



Genotyped individuals born on October 1, 2021 or later

	Mean	SD	Min	Max
APY	1.029	0.011	1.005	1.061
SVD	0.912	0.022	0.866	0.949



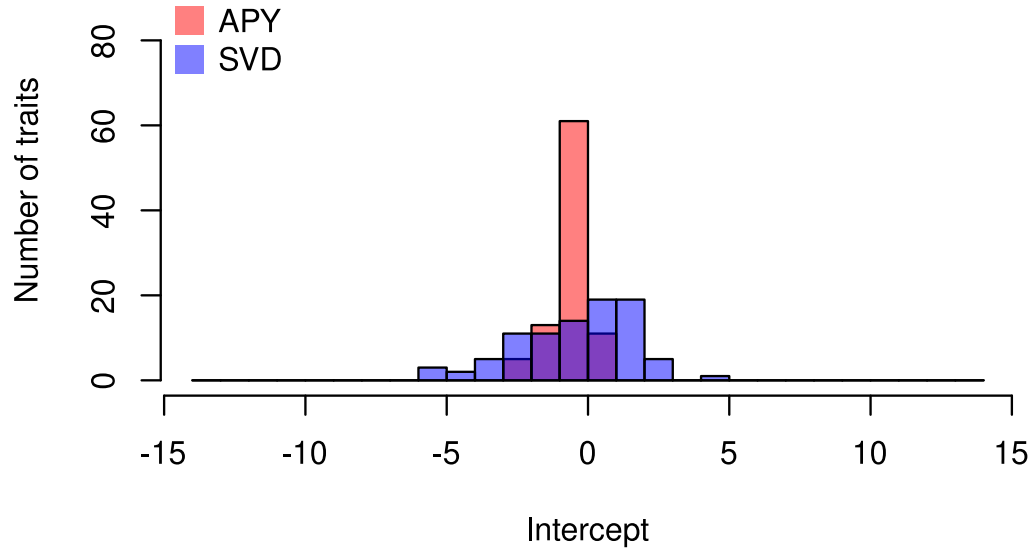
geno | NORWEGIAN RED

Results – intercept

$$\text{ssGBLUP} \approx \mu + \beta * (\text{SVD or APY}) + e$$

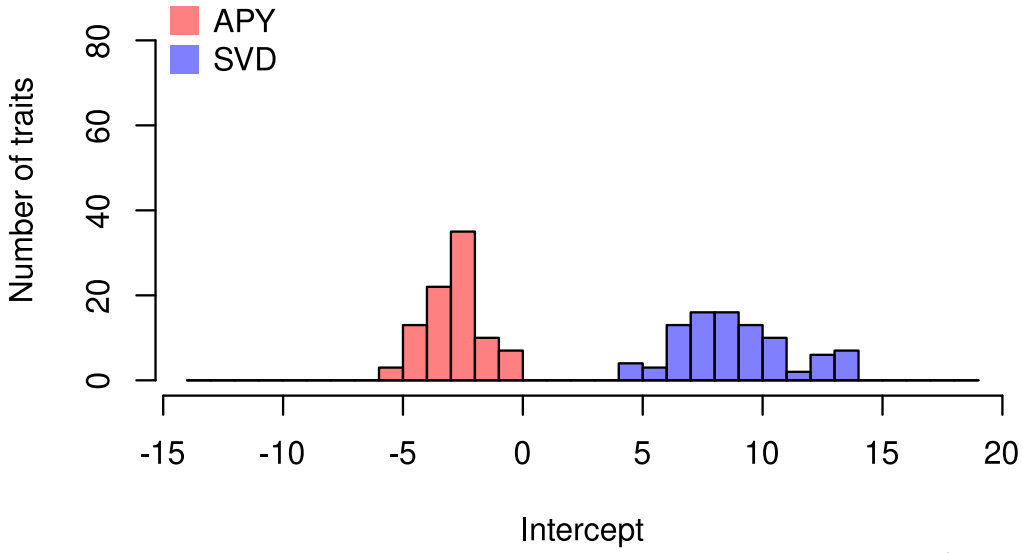
Individuals in pedigree

	Mean	SD	Min	Max
APY	-0.613	0.686	-2.929	0.928
SVD	-0.411	2.030	-5.803	4.943



Genotyped individuals born on October 1, 2021 or later

	Mean	SD	Min	Max
APY	-2.865	1.160	-5.848	-0.397
SVD	8.847	2.378	4.514	13.796



geno | NORWEGIAN RED

Conclusions

- Approximate single step methods solve computational issues of ssGBLUP
- Breeding values predictions from APY and SVD are highly correlated with the predictions from ssGBLUP
  - SVD has slightly higher correlations than APY for young genotyped individuals
- Young genotyped individuals will have overestimated predictions with SVD
- Continue working on the improvement approximate single step methods

Thank you ☺

We want to thank the Norwegian Research council for funding this research through the project 309611, «Large scale single step genomic selection in practice»

geno.no | norwegianred.com