Single-step genomic evaluations

Esa A. Mäntysaari,
Minna Koivula & Ismo Strandén
Animal Genetics Research

Ten Years of Genomic Selection
Joint ADSA/Interbull Session 24. 6. 2019

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Challenges of genomic selection

Genomic selection is the main source of genetic progress in dairy cattle breeding

In theory evaluations ignoring genomic selection
(= Animal Model BLUP)
are biased

Still, AMBLUP results are used as input:

- Multi-step genomic evaluations
- International Evaluations (i.e. MACE)

The genomic selection is accounted in

Single-step GBLUP

Frequently ssGBLUP shows higher genetic trend in selected animals than the AMBLUP

Reasons not well understood:

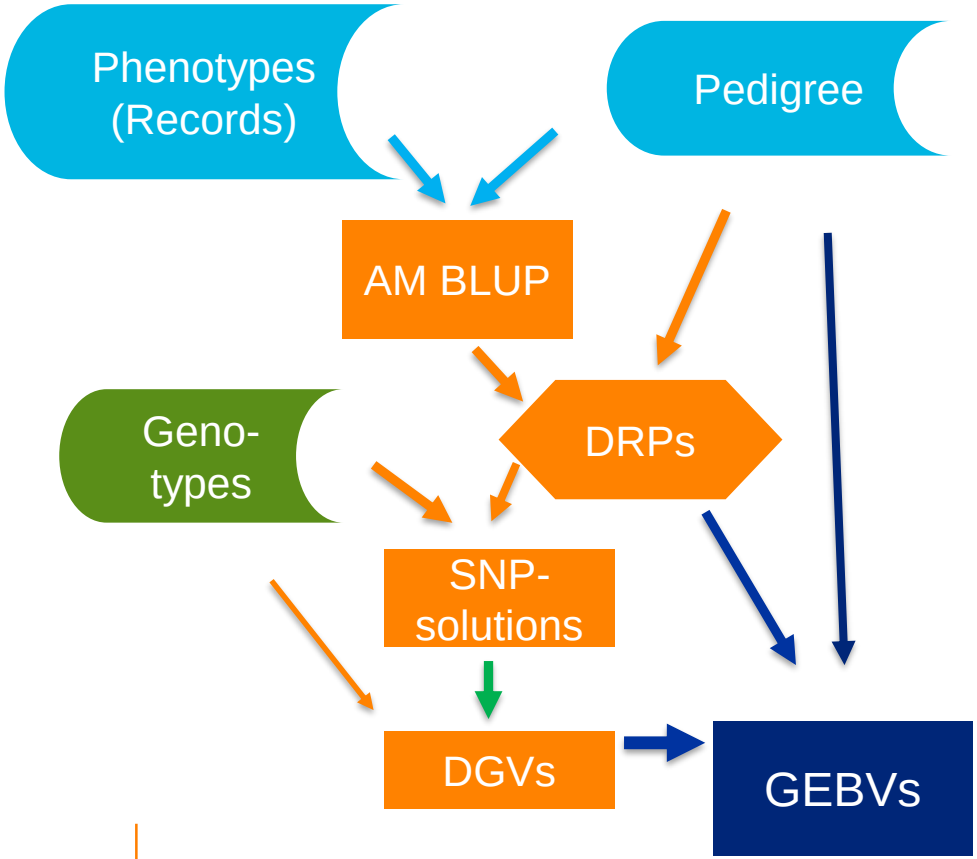
- AMBLUP are often assumed to find genetic progress from well connected overlapping data

ssGBLUP results cannot be used as input for

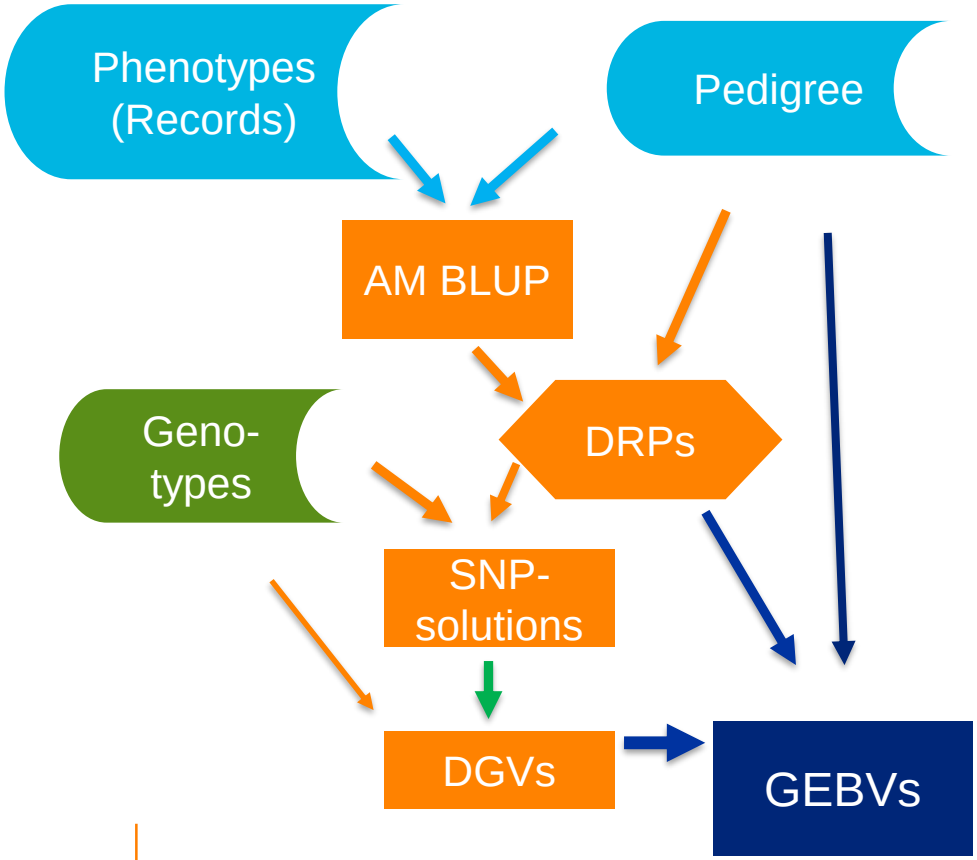
- Multi-step genomic evaluations
- MACE

Multi-step genomic evaluations

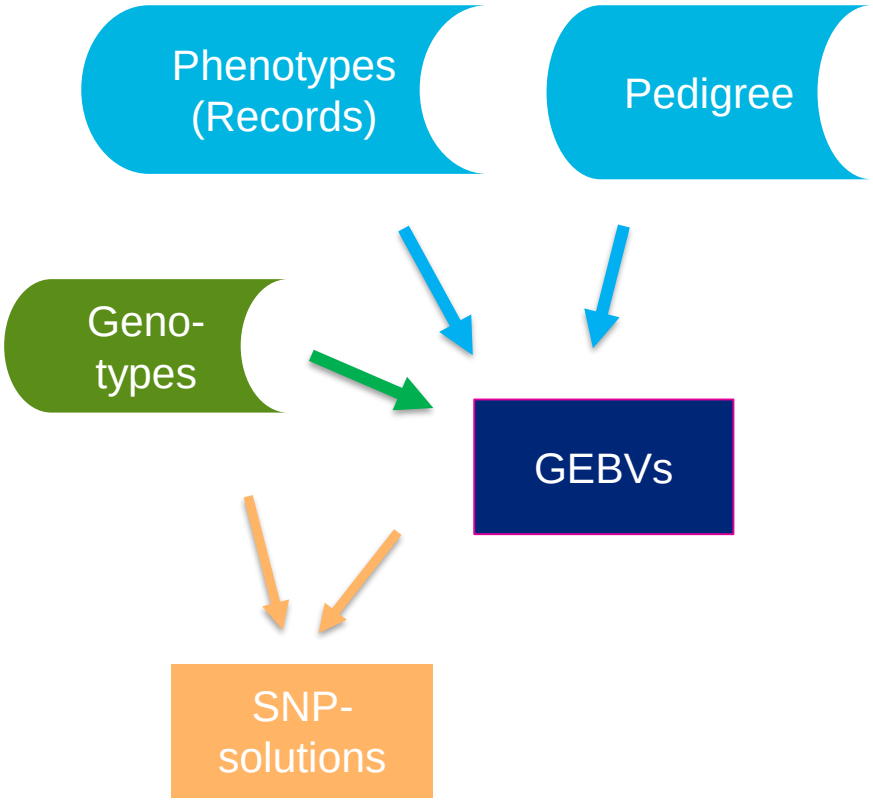
Single-step genomic evaluations



Multi-step genomic evaluations



Single-step genomic evaluations



GEBV – EBV comparison (Example I)

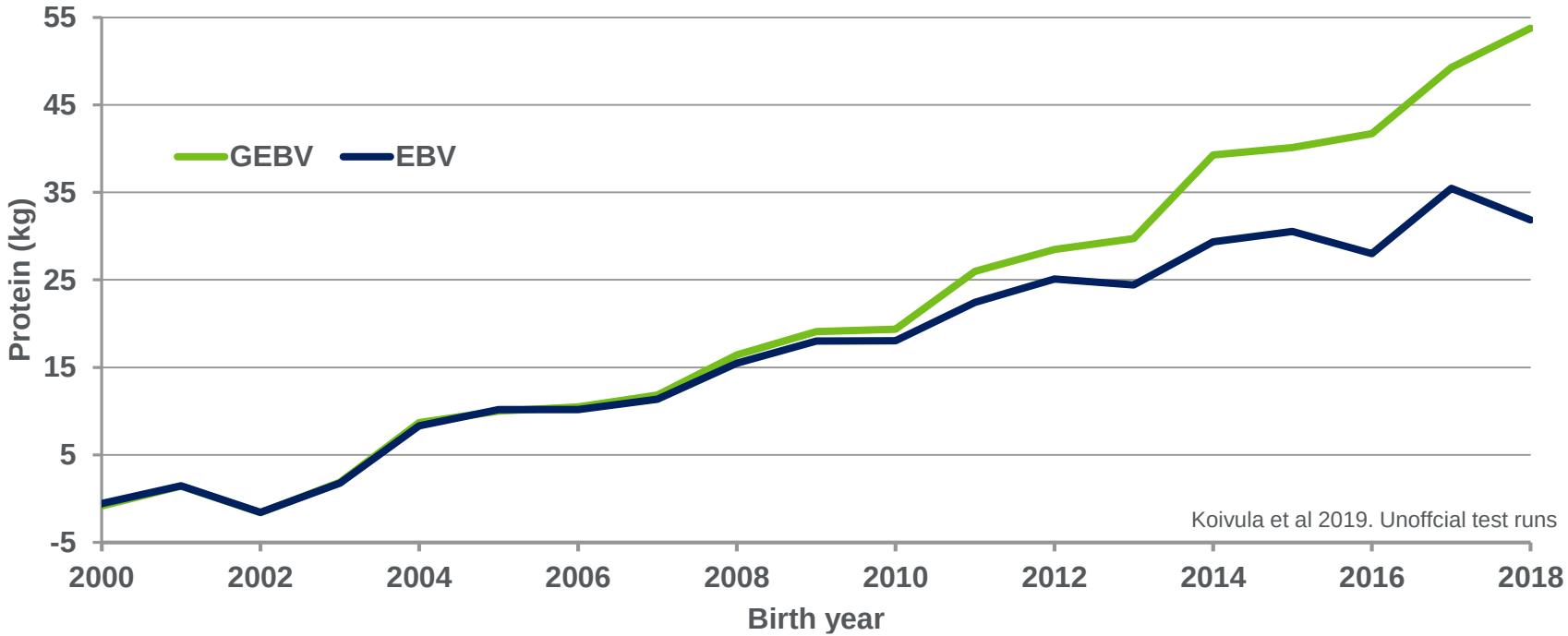
Tested:

Nordic Holstein milk production 305 data (milk, protein, fat), including

about 7.3 million cows in the data and
10 million animals in the pedigree

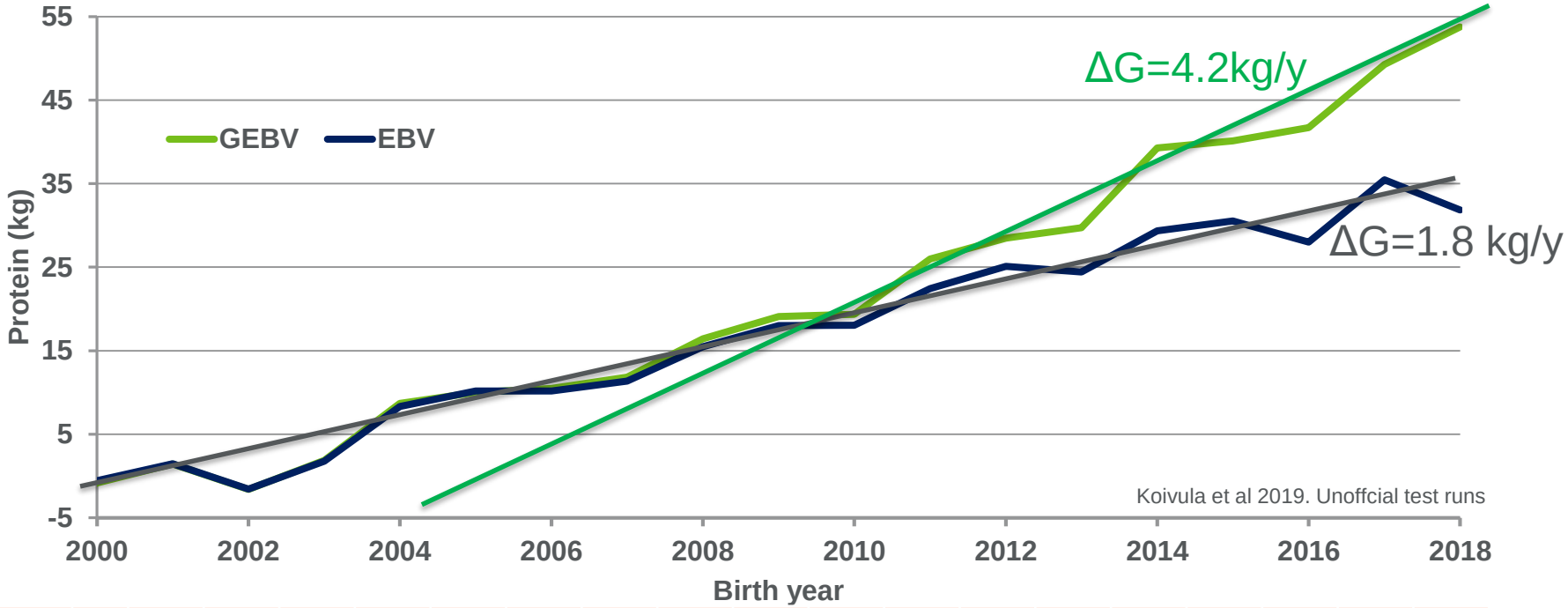
~ 178 000 genotyped animals

Protein trend - genotyped DFS HOLSTEIN bulls



Birthyear	2000		2002		2004		2006		2008		2010		2012		2014		2016		2018
N	276	264	298	258	317	341	386	304	250	213	195	192	211	173	131	97	78	70	42

Protein trend - genotyped DFS HOLSTEIN bulls



Birthyear	2000		2002		2004		2006		2008		2010		2012		2014		2016		2018
N	276	264	298	258	317	341	386	304	250	213	195	192	211	173	131	97	78	70	42

Single-step in (national) dairy cattle evaluations

ONLY FEW OFFICIAL SINGLE STEP EVALUATIONS !

- https://interbull.org/ib/nationalgenoforms (accessed 17.6.2019)

Single-step evaluations on phenotypes

- Czech Republic Test Day model 2016
- Norway 2019

Pseudo single-step

- Belgium Walloon Region
- New Zealand

(Zoetis, USA. Wellness evaluations)

Under development,
or to be released next
(not in particular order)

- DFS (Nordic evaluations)
- New Zealand,
- NDL, FRA, IRL, USA, ….

Single-step in (national) dairy cattle evaluations

ONLY FEW OFFICIAL SINGLE STEP EVALUATIONS !

- https://interbull.org/ib/nationalgenoforms (accessed 17.6.2019)

WHY NOT YET:

1) Computational solution
   - still under development

2) Single-step Genomic models
   - still many open questions

- Computational challenge
- Convergence problems

- Prediction bias b_0,
- Over-dispersion b_1
- Model: GBLUP, Bayesian ”weights”, residual polygenic proportion, …

Background: ssGBLUP is a computational challenge

”Conventional” single-step GBLUP are iterative solutions from the MME
(Aguilar et al. 2010; Christensen and Lund 2010)

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

Here **H** represents the relationship matrix among animals

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where both the NMR relationship matrix $\mathbf{A}_{22}$ and **G** are
dense matrices of the size of *Number of genotyped animals*

Computational solutions / approaches when > 1,000,000 animals are genotyped

Never compute $\mathbf{G}^{-1}$, but instead:

1. Use sparse approximation $\mathbf{G}^{-1}_{APY}$

   or,

2. Never compute $\mathbf{G}^{-1}$, but instead, compute the two matrix products:

$$\mathbf{G}^{-1}\mathbf{d} \quad \text{as} \quad \mathbf{C}\,\mathbf{d} - \mathbf{T}_{\varepsilon}'\mathbf{T}_{\varepsilon}\mathbf{d}$$

Woodbury matrix identity

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Computational approaches - APY ss GBLUP

APY

Never compute $\mathbf{G}^{-1}$, but instead:

1. Use sparse approximation $\mathbf{G}^{-1}_{APY}$

   Divide genotyped animals to core (c) and non-core (y) animals.
   Imagine Cholesky decomposition for the **G** matrix

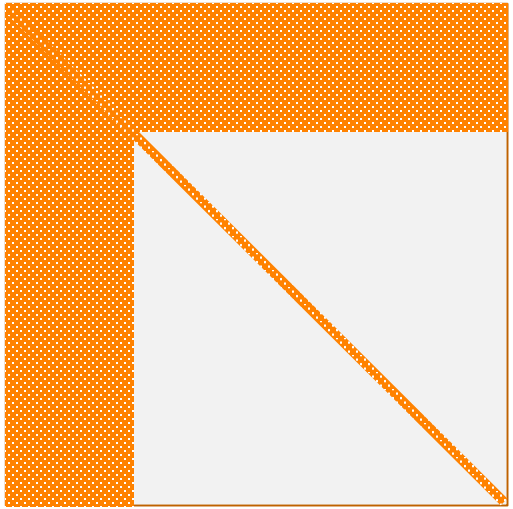
$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{L}_{cc} & 0 \\ \boldsymbol{L}_{yc} & \boldsymbol{L}_{yy} \end{bmatrix}, \quad \text{but use} \quad \boldsymbol{L}_{\boldsymbol{APY}} = \begin{bmatrix} \boldsymbol{L}_{cc} & 0 \\ \boldsymbol{L}_{yc} & diag(\boldsymbol{G}_{yy} - \boldsymbol{L}_{yc}\boldsymbol{L}'_{yc}) \end{bmatrix}$$

   Then

$$\boldsymbol{G}^{-1}_{APY} = \boldsymbol{L}^{-T}_{APY}\boldsymbol{L}^{-1}_{APY}$$

APY ss GBLUP

APY



1. Use sparse approximation $\mathbf{G}^{-1}_{APY}$

   – $\mathbf{G}^{-1}_{APY}$ is nice and sparse (has less non-zeros)
   i.e. ~ 2 n_c*(n_g – n_c/2), where n_g animals genotyped and
   n_c animal in core

   – Requires understanding of population structure to decide whom to choose to be core animals

Computational approaches - ss GTBLUP

ssGTBLUP

2. Never compute $\mathbf{G}^{-1}$, but instead compute the two matrix products:

$$\mathbf{G^{-1}d} \quad \text{as} \quad \mathbf{C^{-1}d - T'Td}$$

where **d** is the direction vector needed in PCG algorithm

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Computational approaches - T matrix in ssGTBLUP

ssGTBLUP

2. Never compute $\mathbf{G}^{-1}$, but instead compute the two matrix products:

$$\mathbf{G}^{-1}\mathbf{d} \quad \text{as} \quad \mathbf{C}^{-1}\mathbf{d} - \mathbf{T}'\mathbf{T}\mathbf{d}$$

ssGTBLUP is based on **Woodbury** matrix identity:

If $\mathbf{G}_C = \mathbf{G}_0 + \mathbf{C} = \mathbf{Z}\mathbf{Z}' + \mathbf{C}$ then $\mathbf{G}_C^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{C}^{-1}\mathbf{Z} + \mathbf{I})^{-1}\mathbf{Z}'\mathbf{C}^{-1}$

for example $\mathbf{G}_\varepsilon = \mathbf{Z}\mathbf{Z}' + \mathbf{I}\varepsilon$ then $\mathbf{G}_\varepsilon^{-1} = \mathbf{I}\varepsilon^{-1} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \varepsilon\mathbf{I})^{-1}\mathbf{Z}'\varepsilon^{-1}$

Computational approaches - T matrix in ssGTBLUP

ssGTBLUP

2. Never compute $\mathbf{G}^{-1}$, but instead compute the two matrix products:

$$\mathbf{G}^{-1}\mathbf{d} \quad \text{as} \quad \mathbf{C}^{-1}\mathbf{d} - \mathbf{T}'\mathbf{T}\mathbf{d}$$

ssGTBLUP is based on **Woodbury** matrix identity:

If $\mathbf{G}_C = \mathbf{G}_0 + \mathbf{C} = \mathbf{Z}\mathbf{Z}' + \mathbf{C}$ then $\mathbf{G}_C^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{C}^{-1}\mathbf{Z} + \mathbf{I})^{-1}\mathbf{Z}'\mathbf{C}^{-1}$

or $\mathbf{G}_w = (1-\mathrm{w})\mathbf{Z}\mathbf{Z}' + w\mathbf{A}_{22}$ then $\mathbf{G}_w^{-1} = \frac{\mathbf{1}}{w}\mathbf{A}_{\mathbf{22}}^{-1} - \frac{\mathbf{1}}{w}\mathbf{A}_{22}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{A}_{22}^{-1}\mathbf{Z} + \frac{1-w}{w}\mathbf{I}\right)^{-1}\mathbf{Z}'\mathbf{A}_{22}^{-1}$

And this can be expressed as : $\mathbf{G}_w^{-1} = \frac{\mathbf{1}}{w}\mathbf{A}_{22}^{-1} - \mathbf{T}'\mathbf{T}$

Properties of single-step GTBLUP

ssGTBLUP

- ssGTBLUP is no approximation, but instead exact ssGBLUP
- It gives significant computational savings when $n_g > > > n_{snp}$

  i.e. the size of matrix **T** is $n_{snp} * n_g$, where n_{snp} number of SNPs
- The **T** matrix can be rank reduced
  - Koivula et al. 2018 used 14,038 eigenvalues for 101k genotyped

    (similar to 18,359 APY core animals in Masuda et al. 2018)

Computational approaches based on sparse $\mathbf{G}^{-1}$ –matrix

APY & ssGTBLUP

If you use sparse $\mathbf{G}^{-1}$ you do not want dense $\mathbf{A}_{22}^{-1}$

→ Computational methods without LARGE inverses

Never compute $\mathbf{A}_{22}^{-1}$, but instead, use two matrix times vector products:

$\mathbf{A}_{22}^{-1}\mathbf{d}$ in 2 pieces as $\mathbf{A}^{22}\mathbf{d} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}\mathbf{d}$

- Multiplications involving $\mathbf{A}^{22}$and $\mathbf{A}^{12}$ can be done using pedigree file
- Solving $(\mathbf{A}^{11})^{-1}$ $[\mathbf{A}^{12}\mathbf{d}]$ can be done using sparse matrix factor of $\mathbf{A}^{11}$

Example II
--ssGTBLUP

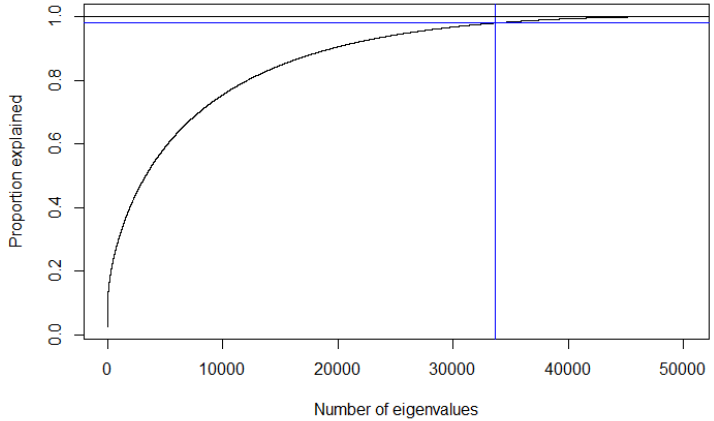
Computations when > 1,5 M animals are genotyped

Irish Cattle Breeding Federation (ICBF) evaluation for calving traits

Number of records 3.5 million rows

6 traits all with direct and maternal genetic effects
Number of pedigree animals: 10.26 million

Number of genotyped (used in the analysis): 1,498,984
Number of markers: 50,240



Three evaluations
(genetic groups as regression):

1) animal model (AM):
2) ssGTBLUP: 98% by eigen analysis (= 33,636 SNP equations)
   - **T** matrix in memory
3) APY33K with random core (33,636 core animals)
   - Inverse **G** matrix in memory

Note: - $ssGT_eBLUP$ (w_{RPG}=0.0)

- computations had 10 processors available
  but only ssGBLUP can fully take advantage of them.

Making **T** for ssGTBLUP and $\mathbf{G}^{-1}$ for APY

	Peak memory	Time	Most time consuming
ssGT$_e$BLUP(98%)	371GB	12.4h	**Z'Z**: 5.2h, eigen: 3.7h, $\mathbf{L}^{-1}\mathbf{Z}$: 3.2h
- Te, full	325GB	10.9h	
ss APY(core 33K)	592GB	14.2h	**G** make: 3.4h, inverse: 9h

Note: APY had to be implemented as memory efficient version, where **G** matrix is done in parts.

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Solving the MME

Case	Peak Memory	Time/iter	N itererations[1]	Total Time
AM BLUP	4.3GB	0.18m	239	43 min
ssGT$_e$BLUP(98%)	386.8GB	1.46m	334	8 h 8 min
ss APY(core 33K)	386.8GB	1.34m	440	9 h 50 min

[1]Convergence assumed $\text{when CR} = \sqrt{\frac{(\mathbf{Cx}-\mathbf{b})'(\mathbf{Cx}-\mathbf{b})}{\mathbf{b}'\mathbf{b}}} < 10^{-6}$

Note: - 6 traits all with direct and maternal genetic effect
- 236 milj equations
- genetic groups by regression coefficients 20/trait
Total computing time depends on chosen convergence

Computational approaches based on single-step marker models - single-step SNPBLUP

Marker Effect Model (ssMEM) by Fernando, Dekkers and Garrick, (2014)

→ ”Impute” expected SNPs to all non-genotyped animals

- Attractive simplicity
- Impractical data storage requirements….
  can be solved by *Imputation “on-the-fly” (*Taskinen et al. 2017)

Legarra and Ducrocq (2012) “Appendix A model”

- Re-derived by Fernando, Cheng, Golden and Garrick (2016)
- Named as single-step Hybrid Model (ssHM)
- A version with residual polygenic effect by Mäntysaari and Strandén (2016)

Single-step Hybrid Model

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'}_1\mathbf{W}_1 & \mathbf{X'}_2\mathbf{W}_2\mathbf{Z} \\ & \mathbf{W'}_1\mathbf{W}_1 + \lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{12}\mathbf{Z} \\ sym & & \mathbf{Z'W'}_2\mathbf{W}_2\mathbf{Z} + \lambda\mathbf{Z'}(\mathbf{A}_{22}^{-1} - \mathbf{A}^{22})\mathbf{Z} + \boldsymbol{\lambda}\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W}_{1'}\mathbf{y} \\ \mathbf{Z'W}_2'\mathbf{y} \end{bmatrix}$$

- Hybrid of snp-BLUP for genotyped animals and animal model of non-genotyped
- Number of random equations: $n_{SNP} + n_{ng}$
- After adding residual polygenic effect nbr of random equations:
  $n_{SNP} + n_{anim}$ (Mäntysaari and Strandén, 2016; EAAP)
  $n_{SNP} + n_{anim} + n_{ng}$ (Garrick et al. 2018, WCGALP)

Single-step animal model with marker effects

- Liu and Goddard augmented the SNP random effects to the vector of animal breeding values, and inverted corresponding H_a matrix (Gengler et al. EAAP 2012; Liu et al. J. Dairy Sci. (2014)

- This $\mathbf{H}_a^{-1}$ can be added to standard AM BLUP model with minimal changes
  - no need to change RHS etc.
  - Convergence has been found problematic

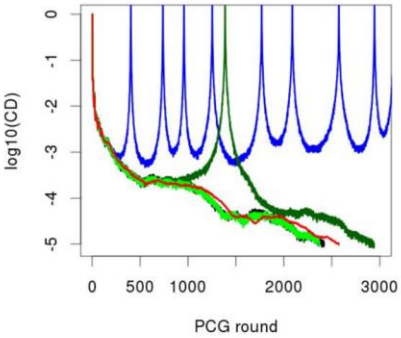
$$\mathbf{H}_a^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} & 0 \\ & \mathbf{A}^{22} + \left(\frac{1}{w} - 1\right)\mathbf{A}_{22}^{-1} & -\frac{1}{w}\mathbf{A}_{22}^{-1}\mathbf{Z} \\ sym & & \frac{1}{w}\mathbf{Z}'\mathbf{A}_{22}^{-1}\mathbf{Z} + \mathbf{B}^{-1} \end{bmatrix}$$

Convergence

Compared to AMBLUP all the single-step MMEs have large condition numbers

==> Poor convergence

Model	PCG rounds
BLUP	2,420
ssGBLUP	16,282
$ssGBLUP_{QP}$	2,941
$ssGBLUP_{QP\_Inb}$	2,373
$ssGBLUP_{QP\_Inb\_APY}$	2,573



Matilainen et al. 2016. Interbull, Puerto Varas, Chile

Some of the problems have been solved by

- Accounting inbreeding
- By the manner genetic groups are handled

Generally ssGBLUP always faster, ssMEM slower

- Large improvements using deflated PCG (Vandenplas et al. 2018) or ”second level preconditioner” (Vandenplas et al. 2019)

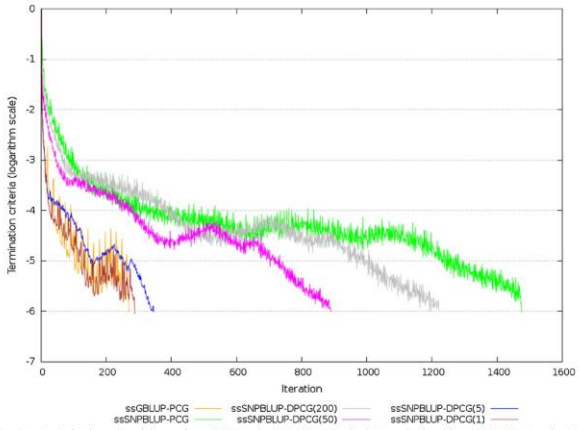
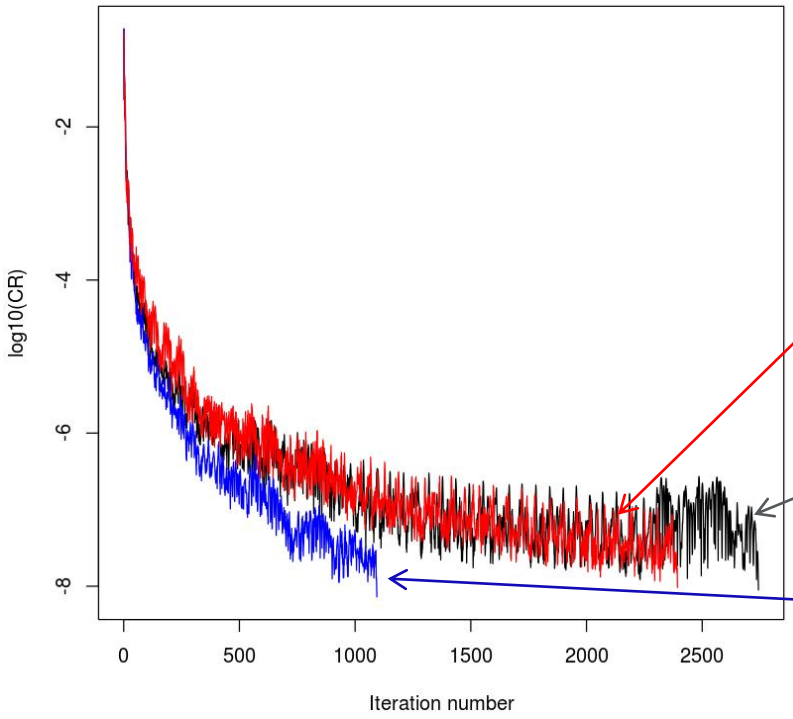


Fig. 2 Termination criteria for the reduced dataset for ssGBLUP and ssSNPBLUP using the PCG method and for ssSNPBLUP using the DPCG method. Number of SNP effects per subdomain is within brackets

Vandenplas et al. 2018. GSE(50):51

Convergence ICBF 6 trait model (1.5M genotyped)



Convergence statistic: $CR = \sqrt{\frac{(\mathbf{Cx}-\mathbf{b})'(\mathbf{Cx}-\mathbf{b})}{\mathbf{b}'\mathbf{b}}}$

Single step, groups as regression coefficients

Animal model, QP transformation

Animal model, groups as regression coefficients

All group regression coefficients in a preconditioner block

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Model developments
The bias in single-step evaluations

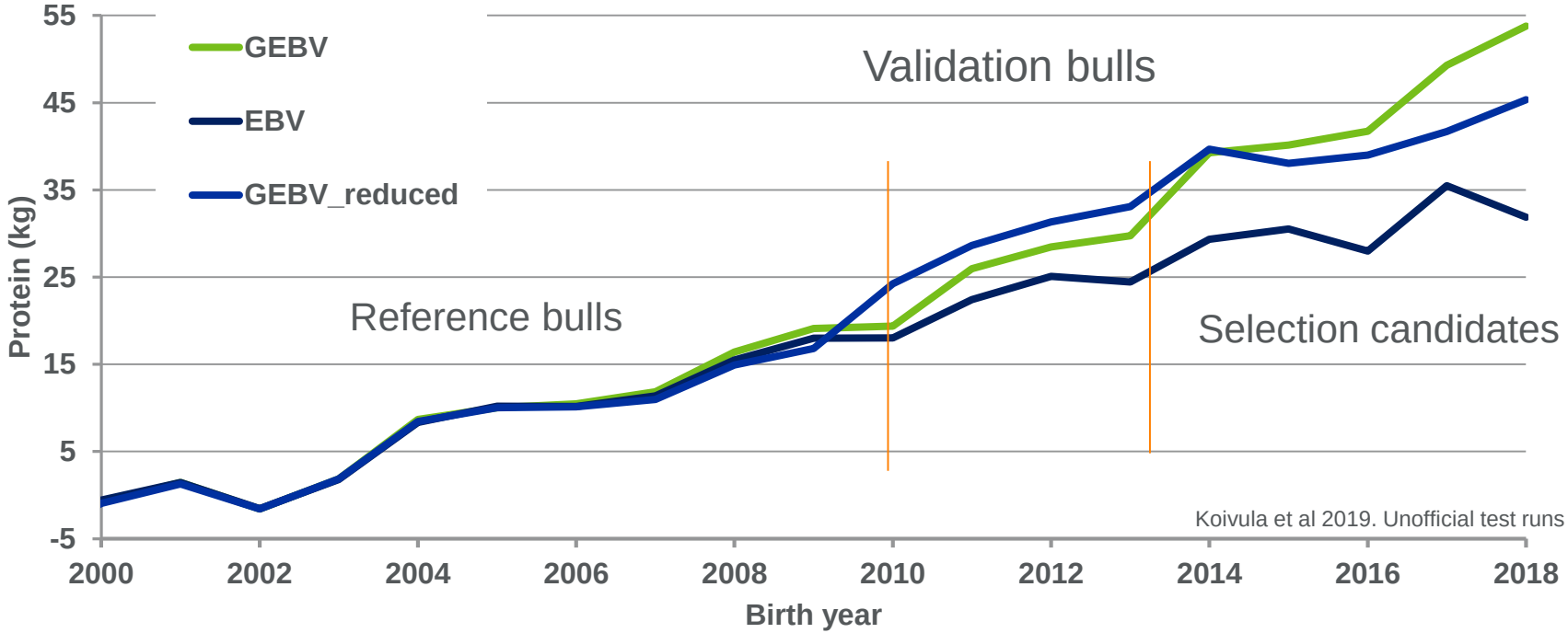
Genomic evaluations are known to over-value the genomic information

- Interbull GEBV validation test b_1:
  - Estimates the over-dispersion of GEBVs,
    i.e. how much of each unit of GEBV in bull calf will be seen in their progeny means
  - Generally Interbull requires $b_1 \geq 0.9$

  - b_1 value can be “fine-tuned” by changing RPG, scaling, blending PA, etc.
  - It is not critical in multi-step GEBV, because comparison is within same stage of animals

When GEBVs are over-valued, also selection is over-valued,
and young bulls are on average over evaluated

Protein trend - genotyped DFS HOLSTEIN bulls
Example I revisitted



Birthyear	2000		2002		2004		2006		2008		2010		2012		2014		2016		2018
N	276	264	298	258	317	341	386	304	250	213	195	192	211	173	131	97	78	70	42

GEBV = ssGTBLUP with QP transformation, w=30%, and including 178000 genotypes, FULL data
GEBV_reduced = ssGTBLUP with QP transformation, w=30%, and including 178000 genotypes, data REDUCED – 4 years

GEBV validation test results for protein (593 Holstein validation bulls)

Regression of DRP on PA or GEBV

	b_0	b_1	R_v^2	MSE
PA	-1.4	0.73	0.19	309.7
GEBV	7.8	0.78	0.45	290.6

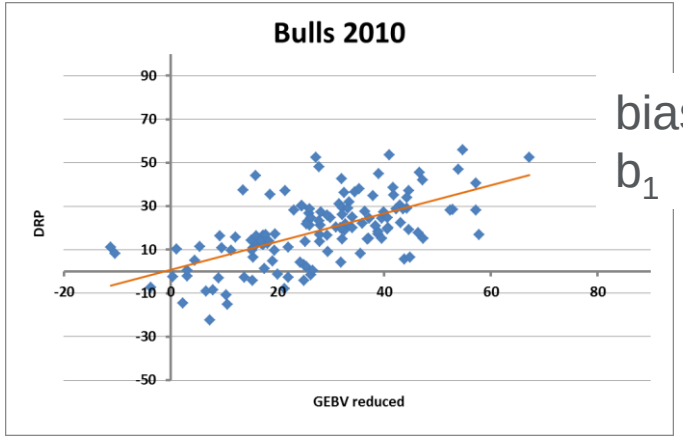
expressed as:
b_0 given b_1=1.0

$$R_v^2 = \frac{R^2_{model\,1}}{\bar{w}}$$

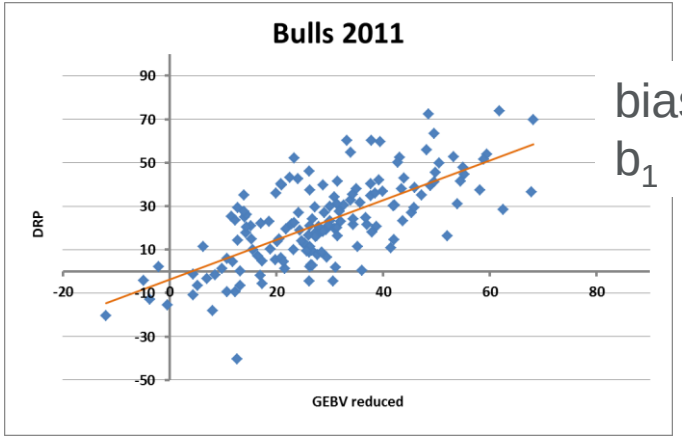
$$b_0 = bias = \frac{\sum_{i=1}^{n}(PA_i - DRP_i)}{n}$$

Luke
NATURAL RESOURCES INSTITUTE FINLAND

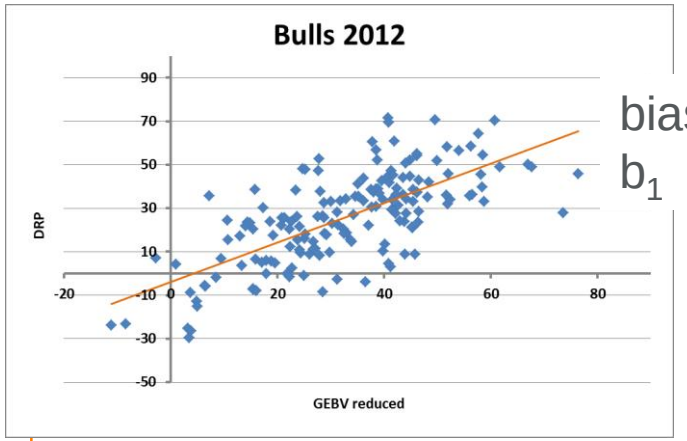
Validation bull DRPs vs $GEBV_r$ by birthyear



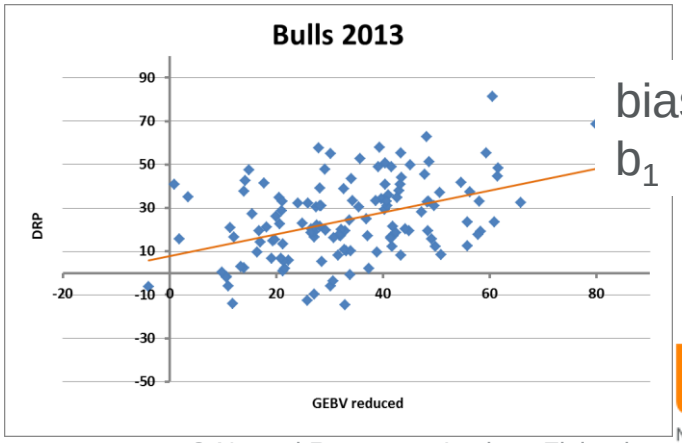
bias 9.2
b_1 0.66



bias 6.3
b_1 0.92



bias 7.0
b_1 0.91



bias 8.8
b_1 0.51

Model developments
Alternative genomic models

Basic single-step GBLUP assumption:

- all SNPs can potentially have effect

  == Same as computing genomic relationships using all SNP markers
- Useful assumption especially for multi-trait models

  -- If SNPs have a'priori different effects,
  the genomic relationships are different for different traits

  -- Difficult to implement in $ssG_{APY}BLUP$ or ssGTBLUP

Bayesian models or models with different weights for SNPs

- Much easier to utilize single-step marker effect models
- Especially if multi-trait models

Models under development
Single-step models with meta-founders

Meta-founders by Legarra, Christensen et al. Genetics (2015)

Matrices $\mathbf{A}_{22}$ and **G** should be compatible with the same base population definition

- Estimate ”genomic self- and across relationships” ($\boldsymbol{\Gamma}$) in base populations
- Build and use $\mathbf{A}_{22}^{\Gamma}$ and $(\mathbf{A}^{\Gamma})^{-1}$ according to $\boldsymbol{\Gamma}$
- Meta-founders will replace the unknown genetic groups

- Promising approach for cross-breed or across-breeds evaluations

Summary

Single-step genomic evaluations are needed to maintain the unbiasedness of genetic evaluations also in the future

The computational solving cost is not the biggest hinderance of implementation

- The easiest are GBLUP methods (ssGTBLUP and ssG$_{APY}$BLUP)
- ssMEM and ssHM are good options if causative variants are to be used

Overprediction, typical to genomic models, will show out more in single-step evaluations

- Finding the best model, testing, validating etc. can be time consuming

THANK YOU

Acknowledgements

Data used in the examples were from

Nordic Cattle Genetic Evaluation NAV,
Viking Genetics
and
Irish Cattle Breeding Federation

Their inputs and support are greatly appreciated