Inclusion of functional annotations into single-step genomic prediction

or marker-specific weights in single-step SNPBLUP

Ismo Strandén
Ismo.stranden@luke.fi

BovReg Training Course, September 28, 2023

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Introduction

Annotations → use different weights for SNP genotypes

In single-step models: the genomic relationship matrix can have different weights for the genotypes

Challenge: multi-trait models where every trait has different weights

- inverting multiple genomic relationship matrices takes time and memory
- current software often needs to be changed
- solving times will become large due to having to do the computations multiple times

A solution: use a single-step SNPBLUP model

Structure:

Single-trait single-step & marker weights
Multi-trait model with common weights
Multi-trait model with heterogeneous weights
A computational example

A single-step model: one trait

A single-trait single-step GBLUP (ssGBLUP) model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + \mathbf{e}$$

where

$\mathbf{b}$ is a vector of fixed effects,

$\mathbf{X}$ is a design matrix for the fixed effects,

$\mathbf{a}$ is a vector of random additive genetic effects,

$\mathbf{W}$ is an incidence matrix for the genetic effects,

and $\mathbf{e}$ is the random residual vector.

Standard assumptions: $\mathbf{a} \sim (\mathbf{0}, \mathbf{H}\sigma_a^2)$ and $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$.

Mixed model equations (MME): $\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + \mathbf{e}$

The animal covariance matrix is $\mathbf{H} = \begin{bmatrix} \mathbf{A}_{nn} + \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}(\mathbf{G}_C - \mathbf{A}_{gg})\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{G}_C \\ \mathbf{G}_C\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{G}_C \end{bmatrix}$

where

$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{nn} & \mathbf{A}_{ng} \\ \mathbf{A}_{gn} & \mathbf{A}_{gg} \end{bmatrix}$ and $\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} \end{bmatrix}$.

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93:743–52.

Christensen O, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42:2.

MME are

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \sigma_a^{-2}\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

where $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_C^{-1} - \mathbf{A}_{gg}^{-1} \end{bmatrix}$,

A= pedigree-based relationship matrix,

$\mathbf{A}_{gg}$= pedigree-based relationship matrix of the genotyped,

$\mathbf{G}_C$= genomic relationship matrix.

Single-trait weighting: genomic relationship matrix

Assume: $\mathbf{G}_C = \mathbf{G}_m + \mathbf{C}$

where $\mathbf{G}_m$ is the genomic and $\mathbf{C}$ is the regularization matrix.

Two common regularization matrices:

$\mathbf{C}_e = e\mathbf{I}$ e is a small number like 0.01,

$\mathbf{C}_w = w\mathbf{A}_{gg}$ w is the residual polygenic proportion.

Let $\mathbf{G}_m = \mathbf{Z}_c\mathbf{B}\mathbf{Z}_c'$

where $\mathbf{Z}_c$ is an n by m matrix of centered marker genotypes (n=the number of genotyped, m= the number of SNP markers) and

$\mathbf{B}$ is an m by m diagonal scaling and weighting matrix.

VanRaden method 1 assumption: $\mathbf{B}_e = \mathbf{I}\frac{1}{s}$ for $\mathbf{C}_e$ and $\mathbf{B}_w = \mathbf{I}\frac{1-w}{s}$ for $\mathbf{C}_w$

the scaling constant $s = 2\sum_{k=1}^{m} p_k(1-p_k)$.

Consider a diagonal weighting matrix $\mathbf{D}$: $\mathbf{B}_{D,w} = \mathbf{D}\frac{1-w}{s}$, where it is assumed that tr($\mathbf{D}$)=m.

Where to get weights?

Marker weights can be on 2 scales:

- variance scale, usable as such
- marker scale, usable after squaring

The weights need to be scaled: tr(**D**)=*m* ➔ average weight is 1.
Needed so that the genetic variance is accounted for correctly.

I assume the computation of weights was covered in
"**Bayesian methods in GS: BayesA, BayesB, BayesC, and RKHS Bayes**"

A simple quick and dirty approach to compute weights

1) Estimate marker effects by a regular single-step

2) Compute weights by (VanRaden, 2008):

$$\mathbf{w}_{i,k} = 1.25^{s_{i,k}-2}$$

where $s_{i,k} = |\hat{\mathbf{g}}_{i,k}|/sd(\hat{\mathbf{g}}_k)$,

$\hat{\mathbf{g}}_{i,k}$ is marker k (k=1,…,m) effect solution of trait i,

$sd(\hat{\mathbf{g}}_k)$ is standard deviation of $\hat{\mathbf{g}}_k$.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

3) Scale weights to be one on average within trait.

Fragomeni BO, Lourenco DA, Legarra A, VanRaden PM, Misztal I. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. J Dairy Sci. 2019;102:10012–9.

The approach can be iterated: marker solutions from single-step used to recalculate weights.

Simple weighting for a multi-trait model

Consider a multi-trait model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \begin{bmatrix}\mathbf{W}_{i,n} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{i,g}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i,n} \\ \mathbf{u}_{i,g}\end{bmatrix} + \mathbf{e}_i$, trait number i=1,...,T.

Subscript n is for the non-genotyped, g is for the genotyped.

Denote $\mathbf{u}_i = \begin{bmatrix}\mathbf{u}_{i,n} \\ \mathbf{u}_{i,g}\end{bmatrix}$ and $\mathbf{W}_i = \begin{bmatrix}\mathbf{W}_{i,n} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{i,g}\end{bmatrix}$ for the additive genetic effects of trait i.

Let $\mathbf{y} = \begin{bmatrix}\mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T\end{bmatrix}$, $\mathbf{b} = \begin{bmatrix}\mathbf{b}_1 \\ \vdots \\ \mathbf{b}_T\end{bmatrix}$, $\mathbf{u} = \begin{bmatrix}\mathbf{u}_1 \\ \vdots \\ \mathbf{u}_T\end{bmatrix}$ and $\mathbf{e} = \begin{bmatrix}\mathbf{e}_1 \\ \vdots \\ \mathbf{e}_T\end{bmatrix}$, and $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_T\end{bmatrix}$ and $\mathbf{W} = \begin{bmatrix}\mathbf{W}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W}_T\end{bmatrix}$.

Then, the model can be written as $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{u} + \mathbf{e}$

It is assumed $\mathbf{u} \sim MVN(\mathbf{0}, \mathbf{G}_0 \otimes \mathbf{H})$ and $\mathbf{e} \sim MVN(\mathbf{0}, \mathbf{R})$ with the additive genetic covariance matrix $\mathbf{G}_0$ and the residual covariance matrix $\mathbf{R}$.

Mixed model equations (MME): $\mathbf{y} = \mathbf{Xb} + \mathbf{Wu} + \mathbf{e}$

MME are

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \mathbf{G}_0^{-1} \otimes \mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

where $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_C^{-1} - \mathbf{A}_{gg}^{-1} \end{bmatrix}$,

A= pedigree based relationship matrix,

$\mathbf{A}_{gg}$= pedigree-based relationship matrix of the genotyped,

$\mathbf{G}_C$= genomic relationship matrix.

THUS: all can be done as for a single-trait model with a one genomic relationship matrix assumed to be valid for all traits.

Pros: existing software can be used, also for large evaluations with millions of genotyped.

Cons: Not ideal in case there are 1) different weights by trait, 2) weights for trait covariances.

Full multi-trait case

Consider a simple two-trait genetic variance with different weights:

$$\mathrm{Var}\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \begin{bmatrix}\mathbf{g}_{0,11}(\mathbf{Z}_c\mathbf{B}_{11}\mathbf{Z}_c' + \mathbf{C}) & \mathbf{g}_{0,12}(\mathbf{Z}_c\mathbf{B}_{12}\mathbf{Z}_c' + \mathbf{C})\\ \mathbf{g}_{0,21}(\mathbf{Z}_c\mathbf{B}_{21}\mathbf{Z}_c' + \mathbf{C}) & \mathbf{g}_{0,22}(\mathbf{Z}_c\mathbf{B}_{22}\mathbf{Z}_c' + \mathbf{C})\end{bmatrix}$$

$$= (\mathbf{I}_2 \otimes \mathbf{Z}_c)\begin{bmatrix}\mathbf{g}_{0,11}\mathbf{B}_{11} & \mathbf{g}_{0,12}\mathbf{B}_{12}\\ \mathbf{g}_{0,21}\mathbf{B}_{21} & \mathbf{g}_{0,22}\mathbf{B}_{22}\end{bmatrix}(\mathbf{I}_2 \otimes \mathbf{Z}_c') + \mathbf{G}_0 \otimes \mathbf{C}$$

No longer in the form of $\mathbf{G}_0 \otimes (\mathbf{Z}_c\mathbf{B}\mathbf{Z}_c' + \mathbf{C})$

→ Need to invert multiple genomic (relationship) matrices $(\mathbf{Z}_c\mathbf{B}_{ij}\mathbf{Z}_c' + \mathbf{C})$: takes time and memory

- current software may need large changes
- long solving times due to having to do the computations using different inverse matrices multiple times (matrix times vector products, instead of matrix times matrix products)

Vandenplas, J., ten Napel, J., Darbaghshahi, S. N., Evans, R., Calus, M. P., Veerkamp, R., et al. (2023). Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genet. Sel. Evol.* 55, 1–17. doi:10.1186/s12711-023-00808-z

An alternative:

- single-step GTBLUP: may be feasible but needs building and inverting many matrices of form $\mathbf{Z}_c'\mathbf{C}^{-1}\mathbf{Z}_c + \mathbf{B}_{ij}{}^{-1}$ → scalability may become an issue (each matrix takes ~20GB). Software changes required.

Full multi-trait case: basic ssSNPBLUP

Liu Z, Goddard ME, Reinhardt F, Reents R. A single-step genomic model with direct estimation of marker effects. J Dairy Sci. 2014;97:5833–50.

Vandenplas, J., ten Napel, J., Darbaghshahi, S. N., Evans, R., Calus, M. P., Veerkamp, R., et al. (2023). Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genet. Sel. Evol.* 55, 1–17. doi:10.1186/s12711-023-00808-z

Assuming no weights the MME are

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} & \mathbf{0} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \mathbf{G}_0^{-1} \otimes \mathbf{H}_C^{-1} & -\mathbf{G}_0^{-1} \otimes \mathbf{K}_C \\ \mathbf{0} & -\mathbf{G}_0^{-1} \otimes \mathbf{K}_C' & \mathbf{G}_0^{-1} \otimes \mathbf{K} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

where

$\mathbf{H}_C^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{-1} - \mathbf{A}_{gg}^{-1} \end{bmatrix}$, $\mathbf{K}_C = \begin{bmatrix} \mathbf{0} \\ \mathbf{C}^{-1}\mathbf{Z}_c \end{bmatrix}$ matrix is from the marker effects to genotypes, and $\mathbf{K} = \mathbf{Z}_c'\mathbf{C}^{-1}\mathbf{Z}_c + \mathbf{B}^{-1}$.

Note: 1) Genomic information is only in $\mathbf{K}_C$ and $\mathbf{K}$.

2) The weights are only in $\mathbf{K}$.

ssSNPBLUP and trait-wise weights

Liu Z, Goddard ME, Reinhardt F, Reents R. A single-step genomic model with direct estimation of marker effects. J Dairy Sci. 2014;97:5833–50.

The marker weights can be allowed to differ by trait.

Following Liu et al. (2014), the covariance matrix of the marker effects can be written as

$$\mathrm{Var}(\mathbf{g}) = \begin{bmatrix} \mathbf{g}_{0,11}\mathbf{B}_{11} & \mathbf{g}_{0,12}\mathbf{B}_{12} & \cdots & \mathbf{g}_{0,1T}\mathbf{B}_{1T} \\ \mathbf{g}_{0,21}\mathbf{B}_{21} & \mathbf{g}_{0,22}\mathbf{B}_{22} & \cdots & \mathbf{g}_{0,2T}\mathbf{B}_{1T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{0,T1}\mathbf{B}_{T1} & \mathbf{g}_{0,T2}\mathbf{B}_{T2} & \cdots & \mathbf{g}_{0,TT}\mathbf{B}_{TT} \end{bmatrix} = \mathbf{V}_g$$

where T is the number of traits, and $\mathbf{g}_{0,ij}$ is the genetic variance between traits i and j in $\mathbf{G}_0$.

Weights included in the diagonal matrices $\mathbf{B}_{ij} = \mathbf{D}_{ij}\frac{1}{s}$ for trait i and j with a marker weight matrix $\mathbf{D}_{ij}$.

Note: all marker weights need to be larger than zero. The scaling of weights for a weight matrix is such that average weight is one within every weight matrix $\mathbf{D}_{ij}$.

Full multi-trait case: ssSNPBLUP

Liu Z, Goddard ME, Reinhardt F, Reents R. A single-step genomic model with direct estimation of marker effects. J Dairy Sci. 2014;97:5833–50.

Weighted ssSNPBLUP MME are

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} & \mathbf{0} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W}+\mathbf{G}_0^{-1}\otimes\mathbf{H}_C^{-1} & -\mathbf{G}_0^{-1}\otimes\mathbf{K}_C \\ \mathbf{0} & -\mathbf{G}_0^{-1}\otimes\mathbf{K}_C' & \mathbf{G}_0^{-1}\otimes\mathbf{Z}_c'\mathbf{C}^{-1}\mathbf{Z}_c+\mathbf{V}_g^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

where

$\mathbf{H}_C^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{-1} - \mathbf{A}_{gg}^{-1} \end{bmatrix}$, $\mathbf{K}_C = \begin{bmatrix} \mathbf{0} \\ \mathbf{C}^{-1}\mathbf{Z}_c \end{bmatrix}$ matrix is from the marker effects to genotypes.

Weights are included in $\mathbf{V}_g$ which includes also the genetic covariance $\mathbf{G}_0$.

The $\mathbf{V}_g$ matrix is easy to invert because it is block diagonal (Liu et al. 2014) having blocks of size *T* for each marker *k*=1,…,*m*:

$$\mathbf{V}_{g,k} = \begin{bmatrix} \mathbf{g}_{0,11}\mathbf{B}_{11,k} & \mathbf{g}_{0,12}\mathbf{B}_{12,k} & \cdots & \mathbf{g}_{0,1T}\mathbf{B}_{1T,k} \\ \mathbf{g}_{0,21}\mathbf{B}_{21,k} & \mathbf{g}_{0,22}\mathbf{B}_{22,k} & \cdots & \mathbf{g}_{0,2T}\mathbf{B}_{1T,k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{0,T1}\mathbf{B}_{T1,k} & \mathbf{g}_{0,T2}\mathbf{B}_{T2,k} & \cdots & \mathbf{g}_{0,TT}\mathbf{B}_{TT,k} \end{bmatrix}$$

Theoretical summary

- Single trait single-step models can accommodate marker weights by almost any approach

- Multi-trait single-step models can accommodate marker weights if these are the same over traits by almost any approach

- Multi-trait single-step models having different weights by trait (and trait-by-trait "covariance") can become computationally infeasible due to a) large RAM use, b) computing needs by many models:
    - The best choice seems to be single-step ssSNPBLUP where these increased demands seem tolerable.

A simple case study about scalability

Data:

Irish Cattle Breeding Federation (ICBF) data: 6 calving difficulty traits

6 direct genetic and 6 maternal genetic effects per individual

9.54 million in pedigree

5.76 million animals with data records

965,868 genotyped with 50,240 SNP markers

360. Indirect genomic prediction reduces computational costs in large-scale single-step evaluations

I. Strandén, J. ten Napel, R.F. Veerkamp, R. Evans, S. Naderi, E.A. Mäntysaari, J. Vandenplas

Pages: 1506 - 1509

OPEN ACCESS

https://doi.org/10.3920/978-90-8686-940-4_360

Data and models

MiX99
Solving Large Mixed Model Equations

DEPENDENT VARIABLES:

TR	TR-NAME	N-OBS	MEAN	SD	MINIMUM	MAXIMUM
1	DH	956190	1.3856	0.64910	1.0000	4.0000
2	DC	3071396	1.2539	0.54494	1.0000	4.0000
3	BH	270063	1.6218	0.83900	1.0000	4.0000
4	BC	1066897	1.4001	0.68742	1.0000	4.0000
5	bsize	716777	3.1630	0.72641	1.0000	5.0000
6	bwt	167228	41.306	7.5369	20.000	115.00

Direct h2 =	0.16	0.083	0.16	0.16	0.27	0.41
Maternal h2 =	0.045	0.021	0.086	0.089	0.063	0.082

Models: 1) Standard single-step, no weights to markers
2) Single-step with all weights equal to one (checking purposes, gives same as 1)
3) Use trait-wise weights computed by a simple approach

A simple quick and dirty approach to get weights

1) Estimate marker effects by a regular single-step

2) Compute weights by (VanRaden, 2008):

$$\mathbf{w}_{i,k} = 1.25^{s_{i,k}-2}$$

where $s_{i,k} = \left|\hat{\mathbf{g}}_{i,k}\right| / sd(\hat{\mathbf{g}}_k)$

and $\hat{\mathbf{g}}_{i,k}$ is marker k effect solution of trait i.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Assuming correlation of one between traits, the covariance $\mathbf{w}_{ij,k} = \sqrt{\mathbf{w}_{i,k}\mathbf{w}_{j,k}}$.

3) Scale weights to be one on average within trait.

MiX99 instruction code

```
DATAFILE        ../../data/data_mix99_calving.txt_good
INTEGER anim block sire dam damp byr sex hyDH hyDC hyBH hyBC hycalv dampar sirexh
REAL DH DC BH BC bsize cvig bwt mort1 mort2 gest agedam HO FR JE MO NO RR ID DY1 DY2 AA AU BA BB CH HE LM PI PT SA SH SI IB damHO damFR damJE
damNO damRR damID damDY1 damDY2 damAA damAU damBA damBB damCH damHE damLM damPI damPT damSA damSH damSI damIB FRxJEh FRxHOh HOxJEh DYxDYh TRxT
COh OBxOBh COxDYh COxTRh COxOBh DYxTRh DYxOBh TRxOBh

MISSING   -99
DATASORT  BLOCK=block PEDIGREECODE=anim

PEDFILE         ../../data/ICBF_6tr_pruned_mod.ped
PEDIGREE        G am
PARFILE         ../../data/covar_calv_6trt.txt

RANDOM          damp hyDH hyDC hyBH hyBC hycalv

SNPMATRIX  USE=PACK FIRST=2 LAST=50241 FORMAT='(i10,26x,50240i1)'  CENTER=p SCALE=p DWEIGHT=T
SNPFILE    ../../geno_data_nocand/ICBF_2018_10_genotypes_in_ped_ref.dat
CENTERFILE base_af_2col.dat
SSSNPBLUP  GTA 0.20
WEIGHTFILE SNPweights.dat
iA22File   PEDIGREE

INBRFILE   ../../data/ICBF_6tr_pruned.inbr
INBREEDING PEDIGREECODE=1 FINBR=3

TMPDIR          ./tmp

WITHINBLOCK G damp hycalv hyBC hyBH hyDC hyDH

PRECON b b b b b b b f

MODEL
DH    =  HO FR JE MO NO RR ID   -   -  AA AU BA BB CH HE LM PI PT SA SH SI IB damHO damFR damJE damMO damNO damRR damID   -      -      -
-     -     -     -     -     -     -     -     -     -     -   FRxJEh FRxHOh HOxJEh DYxDYh   -      -      -      -      -      -      -
  -     agedam            -        byr sex   -  hyDH  -     -     -       -   G(anim dam)
DC    =  HO FR JE MO NO RR ID   -   -  AA AU BA BB CH HE LM PI PT SA SH SI IB damHO damFR damJE damMO damNO damRR damID   -      -      -
-     -     -     -     -     -     -     -     -     -     -   FRxJEh FRxHOh HOxJEh DYxDYh   -      -      -      -      -      -      -
  -       -     gg(agedam|dampar) byr sex damp  -   hyDC  -     -       -   G(anim dam)
```

MiX99 instruction code

```
SNPMATRIX  USE=PACK FIRST=2 LAST=50241 FORMAT='(i10,26x,50240i1)'  CENTER=p SCALE=p DWEIGHT=T
SNPFILE    ../../geno_data_nocand/ICBF_2018_10_genotypes_in_ped_ref.dat
CENTERFILE base_af_2col.dat
SSSNPBLUP  GTA 0.20
WEIGHTFILE SNPweights.dat
iA22File   PEDIGREE
```

Genotypes in ICBF_2018_10_genotypes_in_ped_ref.dat (SNPFILE)

50240 marker columns (First=2, Last=50241),

Genotypes will be packed in RAM (USE=PACK),

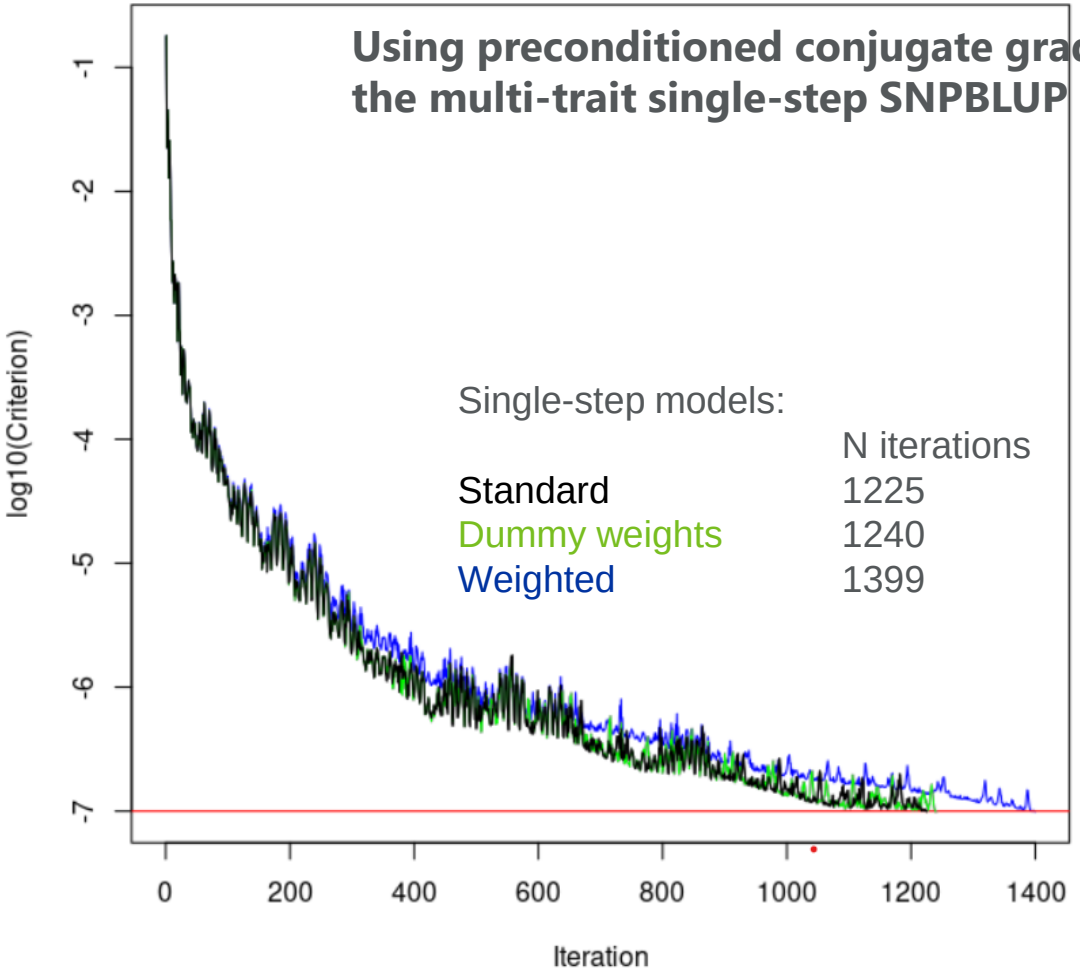
Genotype marker centering by allele frequencies in base_af_2col.dat (CENTER=p, CENTERFILE),

Scaling by $k = 2\sum_{i=1}^{m} p_i(1 - p_i)$ (SCALE=p, CENTERFILE),

Single-step SNPBLUP with the residual polygenic proportion of 20% (SSSNPBLUP GTA 0.20),

Marker weights from the SNPweights.dat file (WEIGHTFILE SNPweights.dat) assumed to have a weight column for every trait and a row for every marker (DWEIGHT=T).

Using preconditioned conjugate gradient (PCG) iteration to solve the multi-trait single-step SNPBLUP model



Single-step models:

	N iterations	Time (h)	RAM (GB)	preprocessing (h)
Standard	1225	16	20.3	0.5
Dummy weights	1240	18	20.4	0.6
Weighted	1399	18	20.4	0.6

Convergence criterion: Cs=r

$$C_r = \sqrt{\frac{\left(\mathbf{Cs}_1^{[k]} - \mathbf{r}\right)'\left(\mathbf{Cs}_1^{[k]} - \mathbf{r}\right)}{\mathbf{r}'\mathbf{r}}}$$

Differences in solutions were small

Marker solution correlations between the original and weighted single-step were from 97.5% to 99.2%.

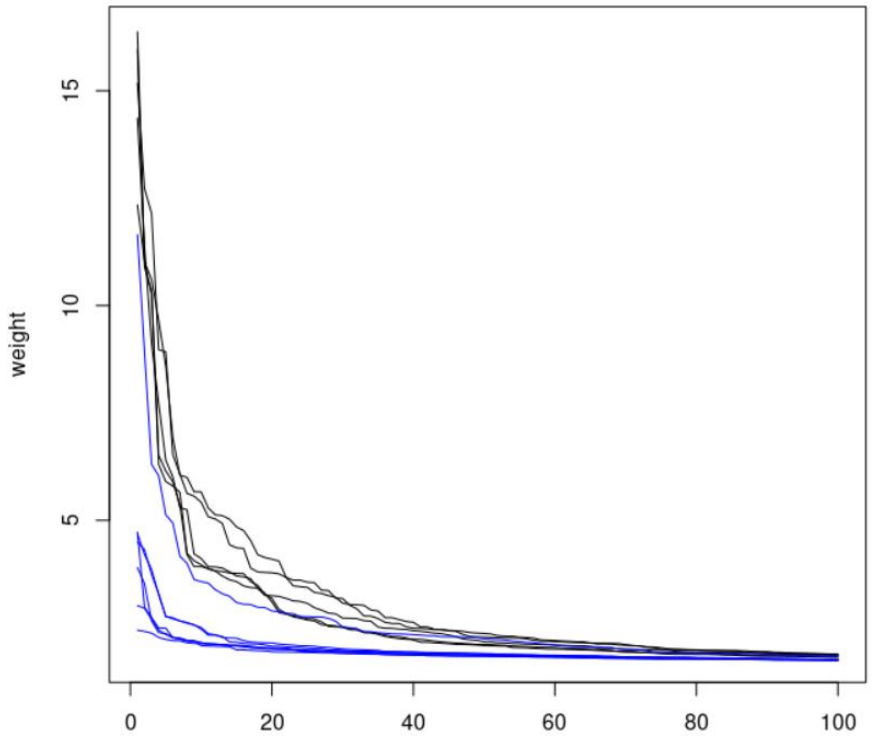
Median of weight ~0.96 in all traits

Minimum weight ~0.83 in all traits

Plot of 100 largest weights by trait:

Black= direct genetic

Blue= maternal genetic



Conclusions

- Single-step model with individual marker weighting is simple when weights are the same over the traits
- Weighting by trait in a multi-trait model can be done using a single-step SNPBLUP model
    - Alternatives exist but can be computationally more challenging
- Weighting increases slightly computational work and the number of iterations until convergence
    - Larger deviations in weights may give poorer convergence
- The usefulness of different marker weights in a large single-step evaluation needs to be quantified with better-derived weights
- Current approach limited to models for which weights can be computed by a similar model:
    - Ex: how to include weights to the test-day model when the weights were estimated by a 305-day data model?