Updates on large-scale genomic analyses

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Daniela Lourenco

M. Bermann, S. Tsuruta, I. Misztal

September 17, 2024

4 million genotyped animals

4 M

30 M

VanRaden & Miller

45 M

https://www.usda.gov/media/blog/2020/06/18/data-saydairy-has-changed

- 5 breeds
- 3 traits
- > 200 M equations

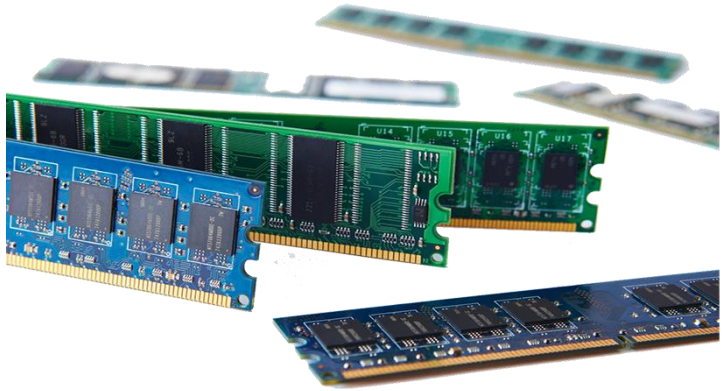
- Did it work?
- ssGBLUP with APY
- 72 hours
- 1 TB RAM

“Memory” exercise

- 4 M genotyped individuals

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$$

VanRaden (2008)



https://www.hp.com/us-en/shop/tech-takes/what-are-gigabytes-of-ram

RAM

$$\mathbf{RAM_{TB}} = \mathrm{N} * \mathrm{M} * 8/1024^4$$

$$\mathbf{RAM_{TB}} = 4\mathrm{M} * 4\mathrm{M} * 8/1024^4 = 116\ \mathrm{TB}$$

Algorithm for Proven and Young (APY)

- Realized relationship matrix in ssGBLUP

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

$\mathbf{G}^{-1}$

Dense ➔ $u_i | u_1 + u_2 + u_3, \ldots, u_{i-1} = \sum_{j=1}^{n-1} p_{ij} u_j + \varepsilon_i$

$\mathbf{G}_{\text{APY}}^{-1}$

Sparse ➔ $u_i | u_{c1} + u_{c2} + u_{c3}, \ldots, u_{ci} = \sum_{j=1}^{c} p_{ij} u_j + \varepsilon_i$

Misztal et al. (2014)
Fragomeni et al. (2015)
Lourenco et al. (2015)

Condition on a set of features or animals = CORE animals

➔

$\mathbf{G}_{\text{APY}}^{-1}$

- $\mathbf{G}_{\text{APY}}^{-1}$ sparse
- Efficient computations

Masuda et al. (2016)

ssGBLUP with 4 M genotyped animals

- 4M genotyped – 45k core
- 30M pedigree and 45M records
- 5 breeds

Dimensionality within each breed

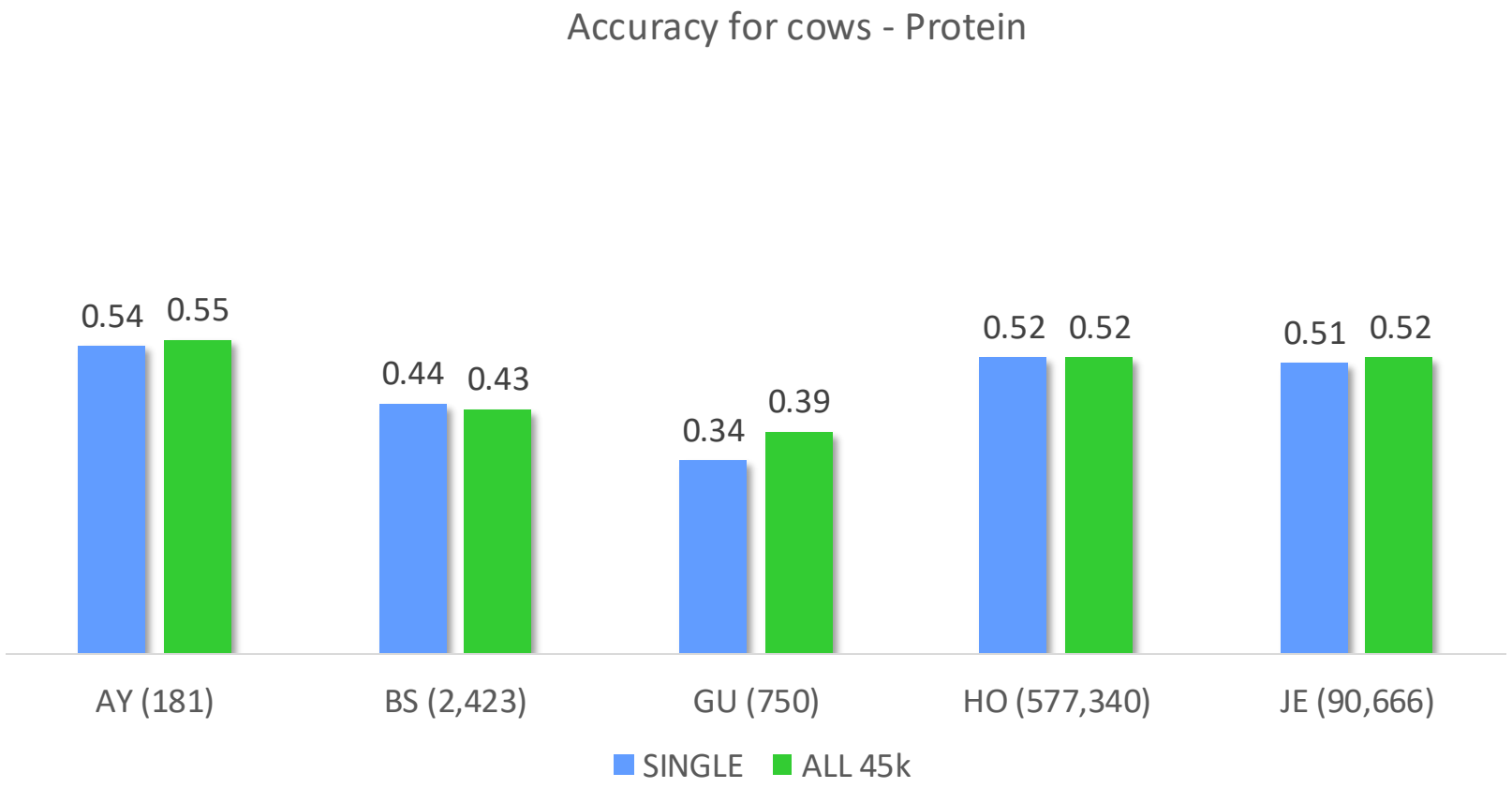
AY, BS, GU = 5k each

HO, JE = 15k each

UPG: altered-QP

UPG: breed, YOB, sex

Breed-specific fixed effects



- 2.7 days for solutions
- 5 days for computing $\mathbf{G}^{-1}_{\text{APY}}$ and $\mathbf{A}^{-1}_{22}$

Updates in $\mathbf{A}_{22}$ for blending

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

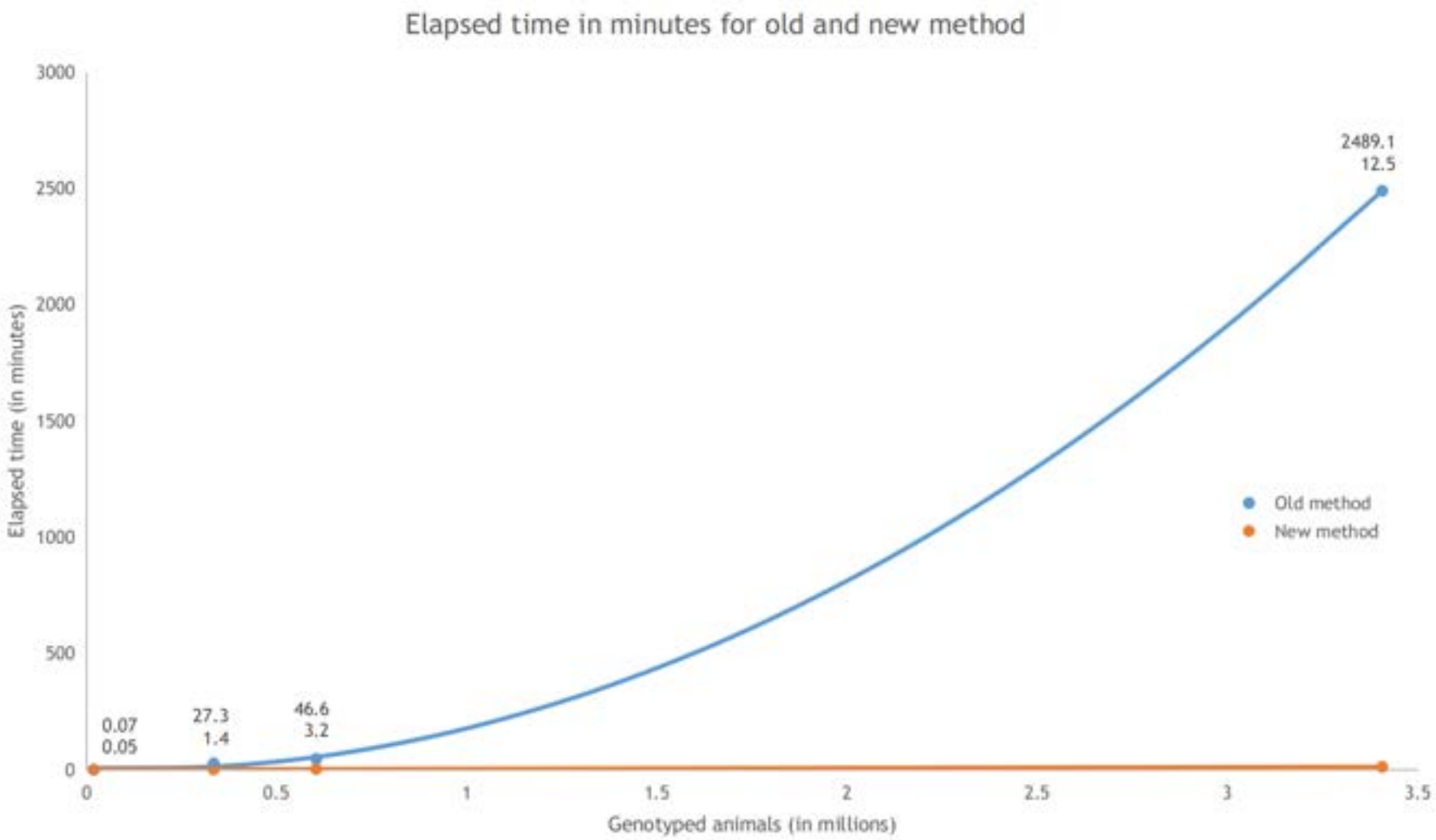
Matias Bermann

$\mathbf{A}_{22}^{-1}$ components: $\mathbf{A}_{22}^{-1} = \mathbf{A}^{22} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}$

APY $\mathbf{G}^{-1}$: $\mathbf{G}_{\mathrm{APY}}^{-1}=\begin{bmatrix}\mathbf{G}_{\mathrm{cc}}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}+\begin{bmatrix}-\mathbf{G}_{\mathrm{cc}}^{-1}\mathbf{G}_{\mathrm{cn}}\\ \mathbf{I}\end{bmatrix}\mathbf{M}_{\mathrm{nn}}^{-1}\begin{bmatrix}-\mathbf{G}_{\mathrm{nc}}\mathbf{G}_{\mathrm{cc}}^{-1} & \mathbf{I}\end{bmatrix}$

- Blending: $\mathbf{G}$ = 0.95 $\mathbf{G}^{*}$ + 0.05 $\mathbf{A}_{22}$
  - Colleau (2002)
  - Rearranging Colleau for core and noncore: from ~4 days to 12.5 minutes

Updates in $\mathbf{A}_{22}$ for blending



Alberto Cesarani

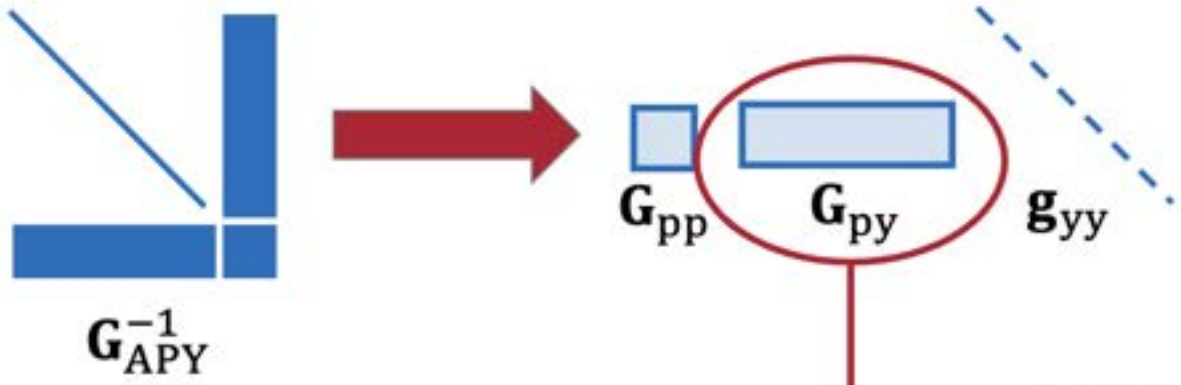
1TB of RAM

- From 5 days to 8 hours to compute $\mathbf{G}_{\mathrm{APY}}^{-1}$ and $\mathbf{A}_{22}^{-1}$ in ALL_45k

Better memory management

- Iteration on data
  - Data in disk and parallelization by splitting data and pedigree files
  - Genomic matrices in memory

- Large genomic datasets -> APY

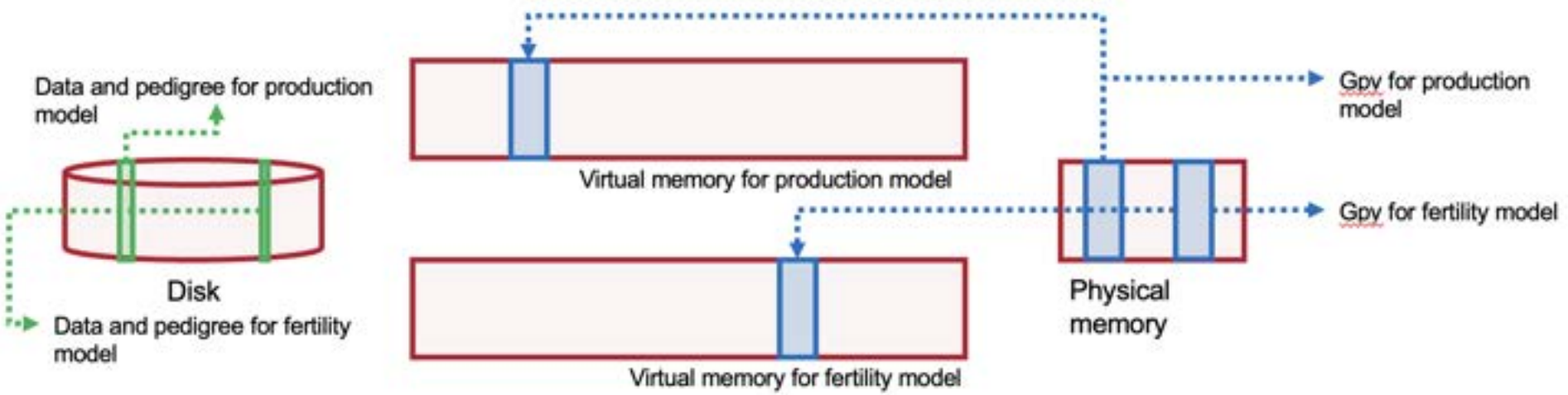


- Largest memory usage
- $n_p \times n_y \times 8\ byte \times 9.31{\times}10^{-10}\ GB/byte$
- For $n_p = 30{,}000$ and $n_y = 3{,}500{,}000$ -> 782.04 GB

Better memory management

Bermann et al. (Unpublished)

OLD



ssGBLUP with Memory Mapping

Memory mapping

- use "memory mapping" `mmap()` to handle $\boldsymbol{G}_{APY}^{-1}$
- A **memory-mapped file** is a segment of virtual memory[1] that has been assigned a direct byte-for-byte correlation with some portion of a file [...] this correlation between the file and the memory space permits applications to treat the mapped portion as if it were primary memory.
- 720 Gb RAM become 720 Gb disk
- modern alternative to "read from file and compute" iteration-on-data

12

Bermann et al. (Unpublished)

Credits: Andres Legarra

ssGBLUP with Memory Mapping

Tabet et al.
(Under review)

- 4 fertility traits: CCR, HCR, DPR, and EFC
- 50M records, 60M in pedigree, ~2M animals genotyped, ~500M equations
- ssGBLUP with APY: 45k core
  - Multi-breed

ssGBLUP with Memory Mapping

Running of APY

- PreGSf90: Set up $\boldsymbol{G}_{APY}^{-1}$ (with blending of [5% or 10%] $\boldsymbol{A}_{\Gamma 22}$).
  - RAM ≈ 720 Gigabytes [not using mmap()]
- Blup90iod3 (PCG iteration on data)
  - uses "memory mapping" `mmap()` to handle $\boldsymbol{G}_{APY}^{-1}$
  - As a result, only 120 Gb (non-genomic parts, including the 4 x 60M animals GEBVs...) are needed for the iteration
- accf90GS2 for reliabilities (Bermann et al 2022a) also uses `mmap()`

It also has memory mapping now!

- memory mapping assigns memory to disk space
- 720 GB of RAM -> 720 GB of disk
- 120 GB of RAM

13

Credits: Andres Legarra

Approximating reliabilities

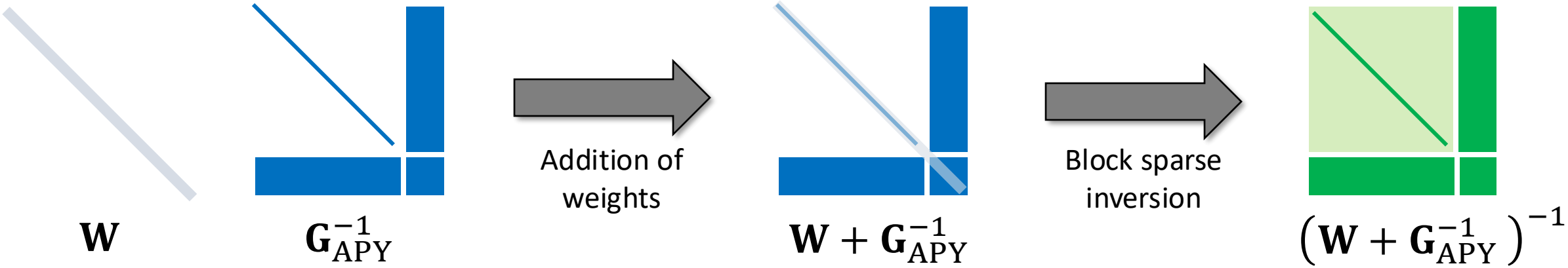
- Accuracy based on PEV
  - Approximated for large populations
  - Weights based on approximations
  - Block sparse inversion with APY

JOURNAL ARTICLE

Efficient approximation of reliabilities for single-step genomic best linear unbiased predictor models with the Algorithm for Proven and Young

Matias Bermann, Daniela Lourenco, Ignacy Misztal

Journal of Animal Science, Volume 100, Issue 1, January 2022, skab353, https://doi.org/10.1093/jas/skab353



$$diag\left(\mathbf{W}+\mathbf{G}^{-1}_{\mathrm{APY}}\right)^{-1} = diag\left((\mathbf{W}_{\mathrm{nn}}+\mathbf{M}^{-1}_{\mathrm{nn}})^{-1}+(\mathbf{W}_{\mathrm{nn}}+\mathbf{M}^{-1}_{\mathrm{nn}})^{-1}\mathbf{G}^{\mathrm{nc}}(\mathbf{W}_{\mathrm{cc}}+\mathbf{G}^{\mathrm{cc}}-\mathbf{G}^{\mathrm{cn}}(\mathbf{W}_{\mathrm{nn}}+\mathbf{M}^{-1}_{\mathrm{nn}})^{-1}\mathbf{G}^{\mathrm{nc}})^{-1}\mathbf{G}^{\mathrm{cn}}(\mathbf{W}_{\mathrm{nn}}+\mathbf{M}^{-1}_{\mathrm{nn}})^{-1}\right)$$

$$diag\left((\mathbf{W}_{\mathrm{cc}}+\mathbf{G}^{\mathrm{cc}}-\mathbf{G}^{\mathrm{cn}}(\mathbf{W}_{\mathrm{nn}}+\mathbf{M}^{-1}_{\mathrm{nn}})^{-1}\mathbf{G}^{\mathrm{nc}})^{-1}\right)$$

Equivalence APY ssGBLUP – ssSNPBLUP

- Equivalent models under same assumptions and data
  - Equal estimable functions
  - $\hat{\mathbf{u}} = \mathbf{Z}\hat{\boldsymbol{a}}$
  - $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$
  - $Var(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}'\mathbf{G}^{-1}(\mathbf{G} - \mathbf{C}^{\mathbf{u_2}\mathbf{u_2}})\mathbf{G}^{-1}\mathbf{Z}k$
  - $p\text{-}value_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$ → ssGWAS

Stranden and Garrick (2009)

Guladron-Duarte et al. (2014)

Equivalence APY ssGBLUP – ssSNPBLUP

- If using APY in ssGBLUP
- Equivalent APY ssSNPBLUP model
- $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$ ➔ $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}\hat{\mathbf{u}}_c$
- $Var(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}'\mathbf{G}^{-1}(\mathbf{G} - \mathbf{C}^{\mathbf{u_2u_2}})\mathbf{G}^{-1}\mathbf{Z}k$ ➔ $Var(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}\,(\mathbf{G}_{cc} - \mathbf{C}^{\mathbf{u}_{2c}\mathbf{u}_{2c}})\mathbf{G}_{cc}^{-1}\mathbf{Z}_c k$

Function of CORE animals
How to get it?

Single-step GWAS – now unlimited

- Genomic evaluation process
  - GEBV using APY ssGBLUP + accuracy using block sparse inversion
    - $\mathbf{C}^{\mathbf{u}_{2c}\mathbf{u}_{2c}} = \left(\mathbf{W} + \mathbf{G}_{\mathbf{APY}}^{-\mathbf{1}}\right)^{-\mathbf{1}}$
    - $Var(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}\left(\mathbf{G}_{cc} - \mathbf{C}^{\mathbf{u}_{2c}\mathbf{u}_{2c}}\right)\mathbf{G}_{cc}^{-1}\mathbf{Z}_c k$

Leite et al. Genetics Selection Evolution (2024) 56:59
https://doi.org/10.1186/s12711-024-00925-3

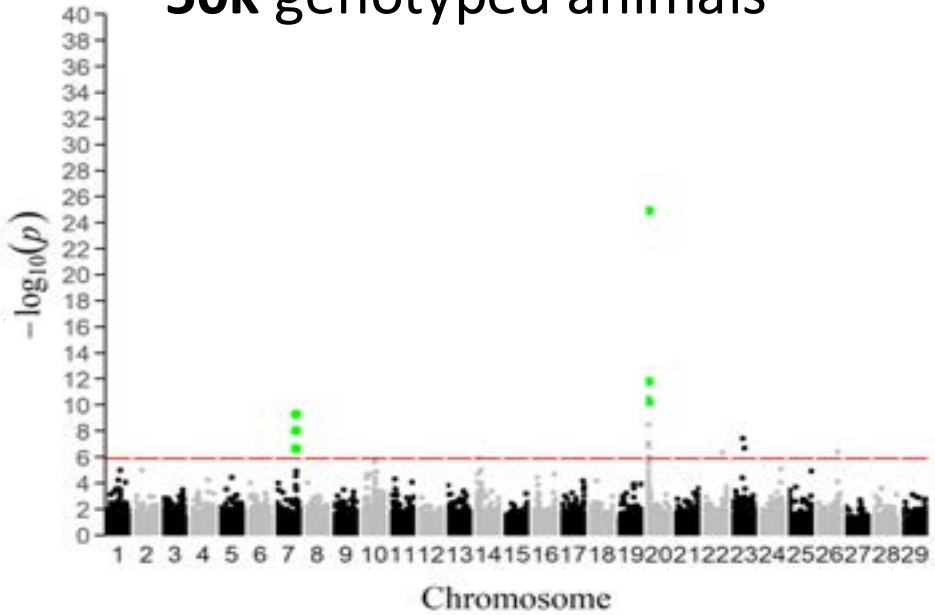
Genetics Selection Evolution

RESEARCH ARTICLE Open Access

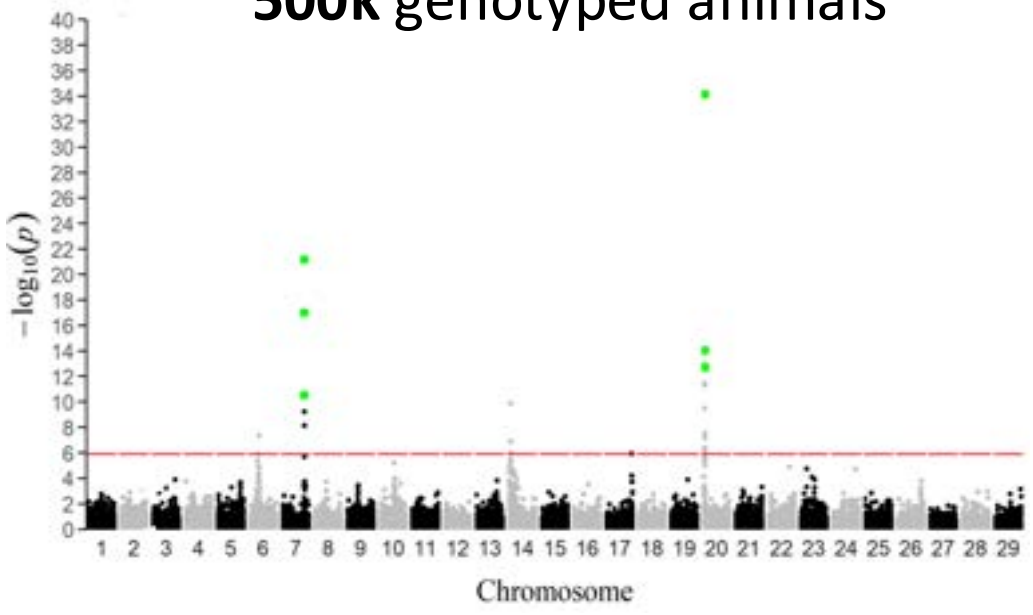
Marker effect p-values for single-step GWAS with the algorithm for proven and young in large genotyped populations

Natália Galoro Leite[1*], Matias Bermann[1], Shogo Tsuruta[1], Ignacy Misztal[1] and Daniela Lourenco[1]

50k genotyped animals



500k genotyped animals

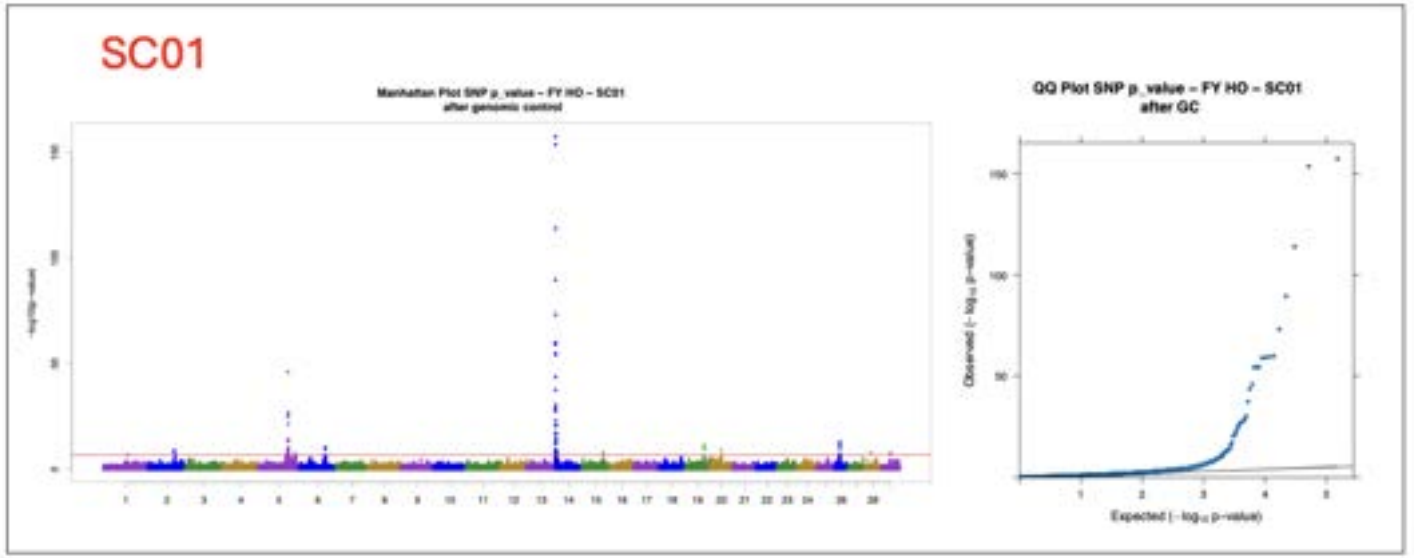


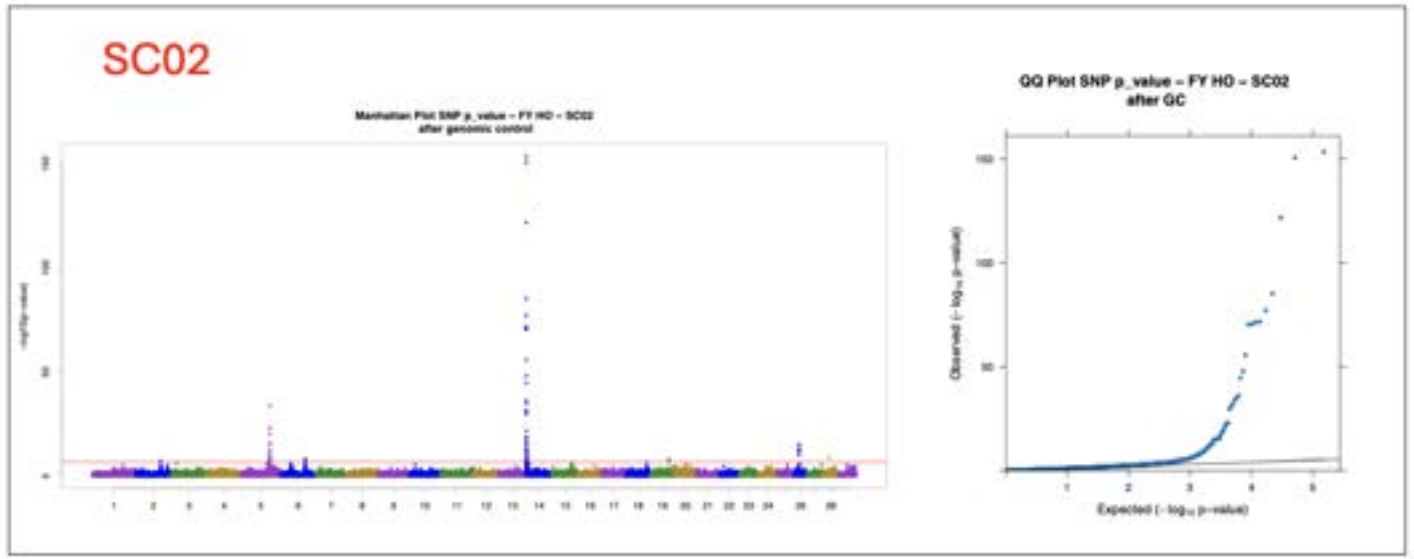
Single-step GWAS – now unlimited

- **Genotypes**
JE: 528,638
HO: 1,794,100
GU: 3,774
BS: 10,417
AY: 1,940
MULTI: 2,338,869

- **Pedigree (all breed)**
52,667,746

- **Phenotypes (all breed)**
106,982,064



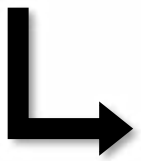


Carrara et al.
(In progress)

Reliability of IP

- Approximate reliability of Indirect Predictions

$$Var(\hat{\boldsymbol{a}}) = k\mathbf{Z}_c'\mathbf{G}_{cc}^{-1}\left(\mathbf{G}_{cc} - \mathbf{C}^{\mathbf{u}_{2c}\mathbf{u}_{2c}}\right)\mathbf{G}_{cc}^{-1}\mathbf{Z}_c k$$



Diagonal for GWAS

Full matrix for REL_{IP}

$$\mathrm{REL}_{\mathrm{IP}_j} = \frac{\mathbf{z}_j\, Var(\hat{\boldsymbol{a}})\mathbf{z}_j'}{\sigma_u^2}$$

Liu et al. (2017)

Holstein Dataset

- **Pedigree**: 2,240,568 animals
- **Milk Yield**: 1,422,330 Records
- **Genotypes**: Total: 33,338

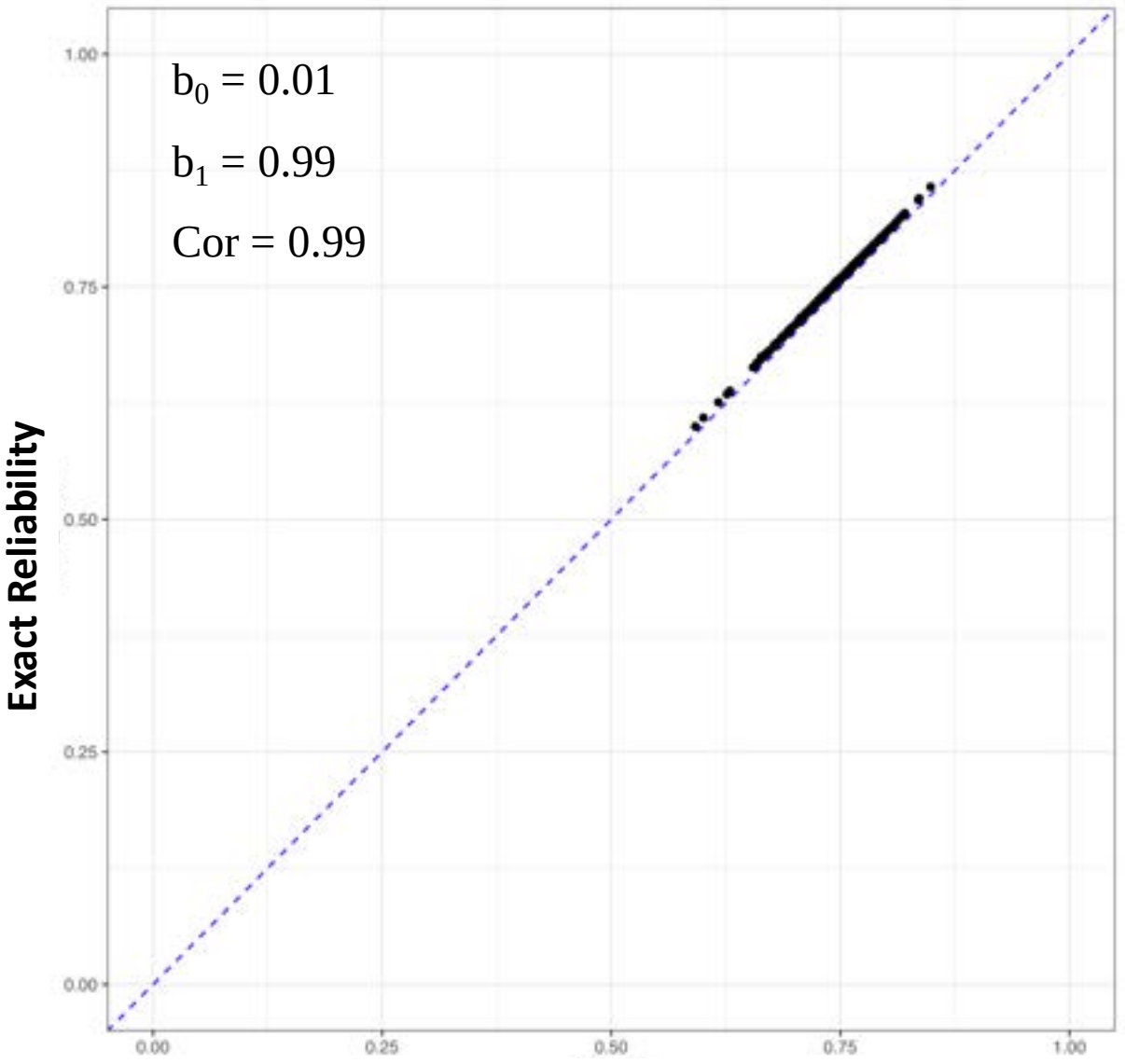
  Training: 32,570 bulls

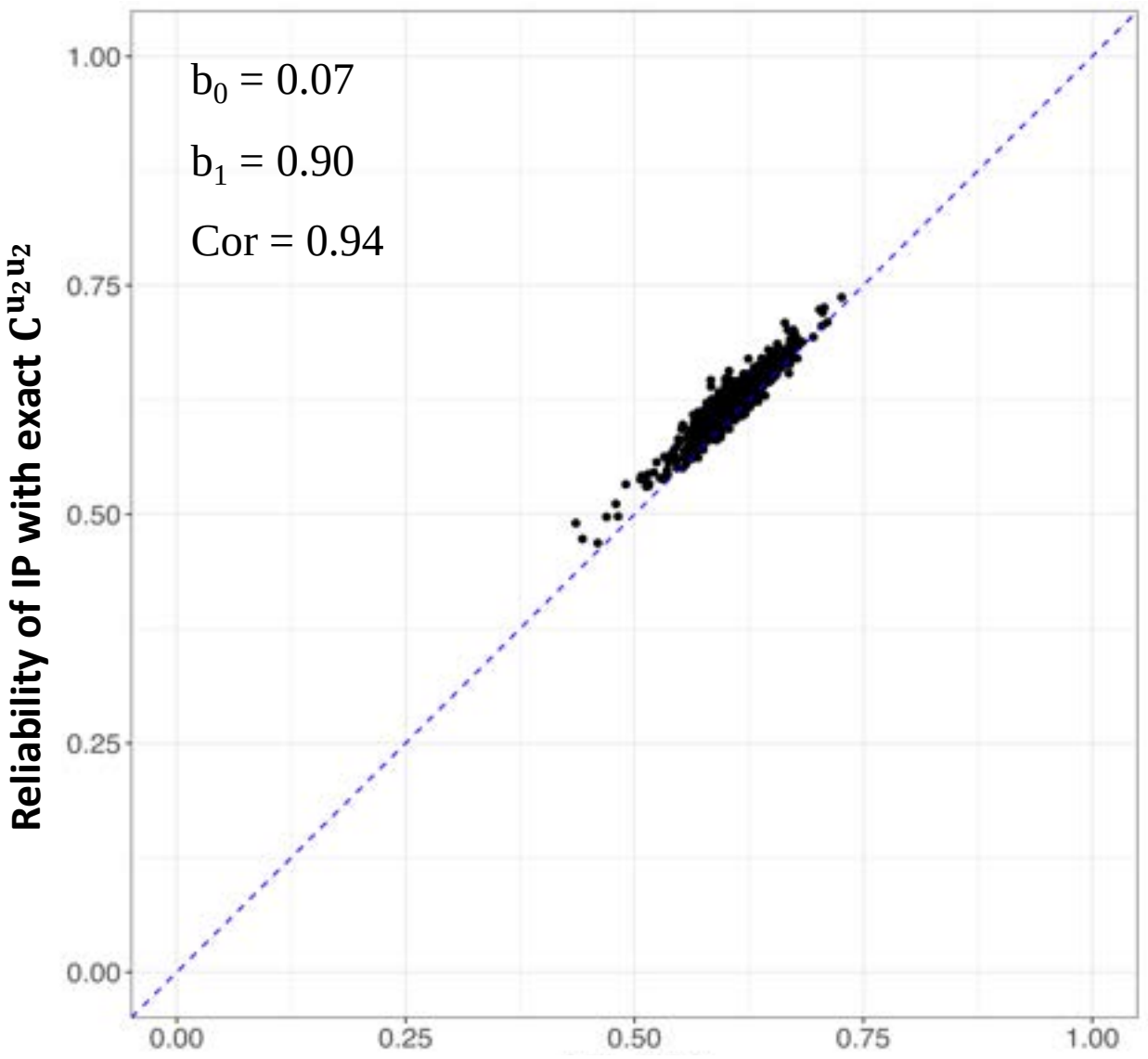
  Validation (2017): 768 bulls

Tabet et al.
(In progress)

- **Exact reliabilities based on the inverse** (training + validation)
- **Reliabilities of IP** (for validation) **with exact** $\mathbf{C^{u_2 u_2}}$ (from training)
- **Reliabilities of IP** (for validation) **with approximated** $\mathbf{C^{u_{2c} u_{2c}}}$ (from training)

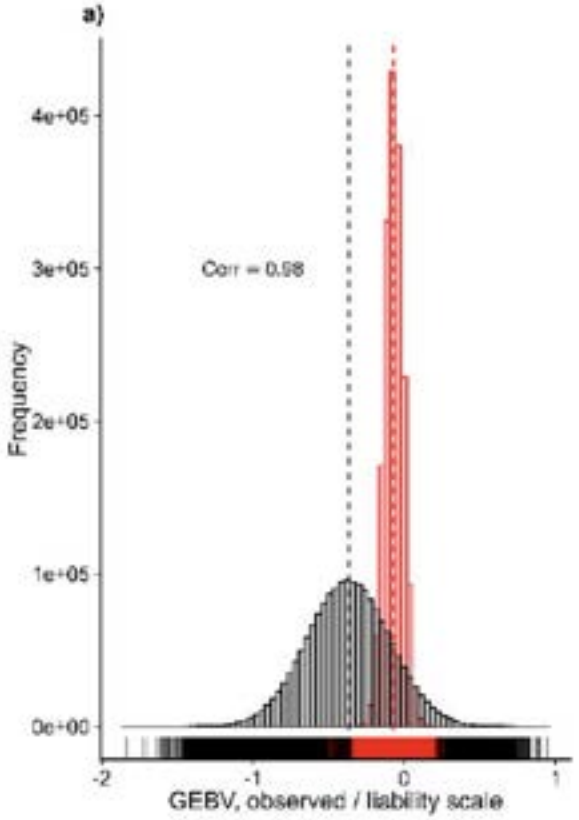
Reliabilities





Genomic predictions – binary/categorical

- Threshold models - Binary or categorical traits
  - > 10x more time to reach convergence
  - Liability solutions into probabilities
  - Linear solutions into probabilities (???)



$$GEBV_{lia} \approx \frac{GEBV_{lin}}{\sqrt{\sigma^2_{e_{lin}} * \left(1 - \frac{h^2_{lin}}{h^2_{lia}}\right)}}$$

$$P_i = 1 - \Phi\left(\frac{t - \mu_u - u_i}{\sigma_e}\right)$$

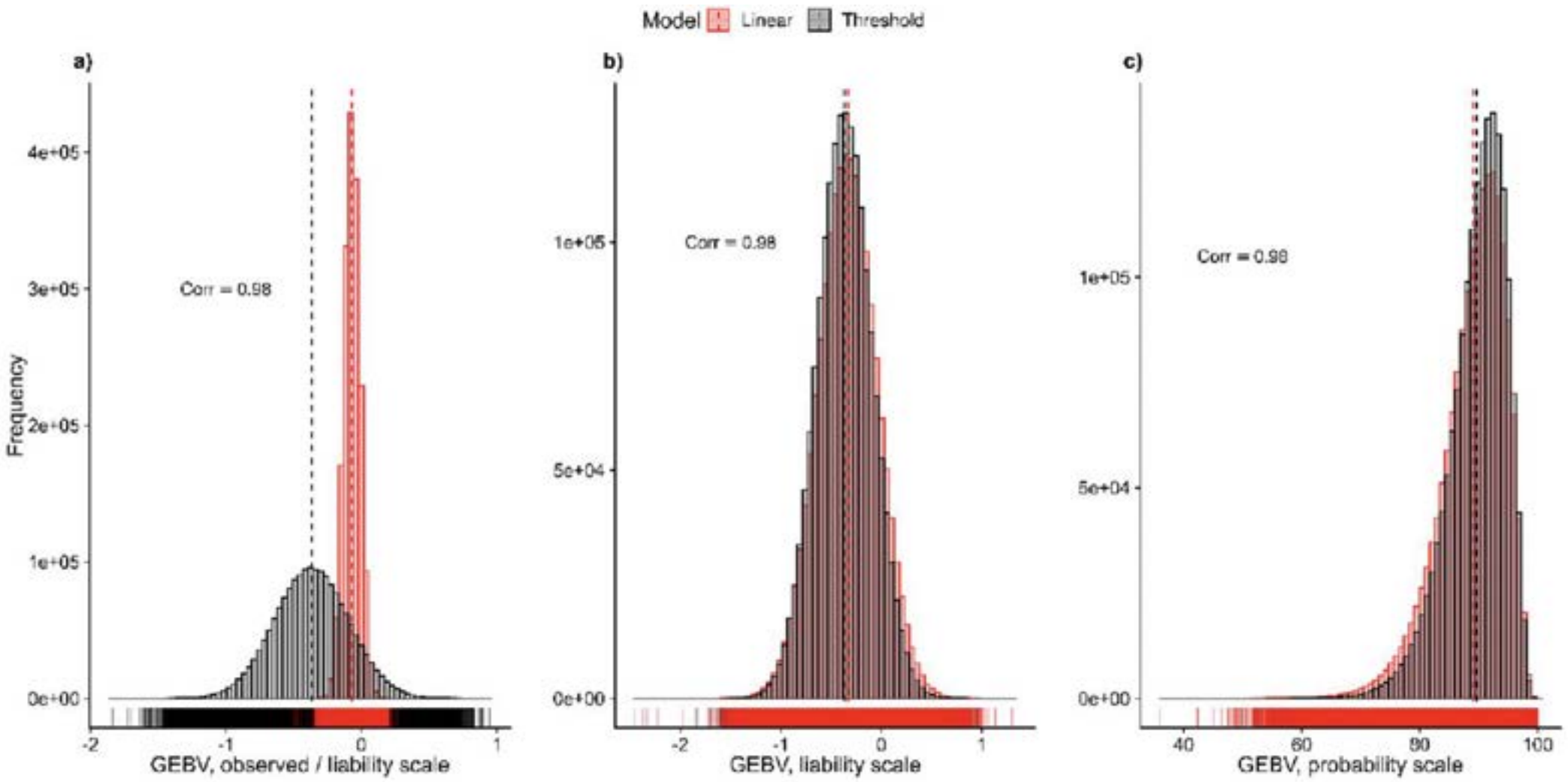
J. Dairy Sci. TBC
https://doi.org/10.3168/jds.2024-24767

Converting estimated breeding values from the observed to probability scale for health traits

Jorge Hidalgo,[1]* Shogo Tsuruta,[1] Dianelys Gonzalez,[2] Gerson de Oliveira,[2] Miguel Sanchez,[2] Asmita Kulkarni,[2] Cory Przybyla,[2] Giovana Vargas,[2] Natascha Vukasinovic,[2] Ignacy Misztal,[1] and Daniela Lourenco[1]
[1]Department of Animal and Dairy Science, University of Georgia, Athens, GA, 30602, USA
[2]Zoetis Genetics and Precision Animal Health, Kalamazoo, MI, 49007, USA

Mastitis
$h^2 = 0.09$
% = 27

Genomic predictions – binary/categorical



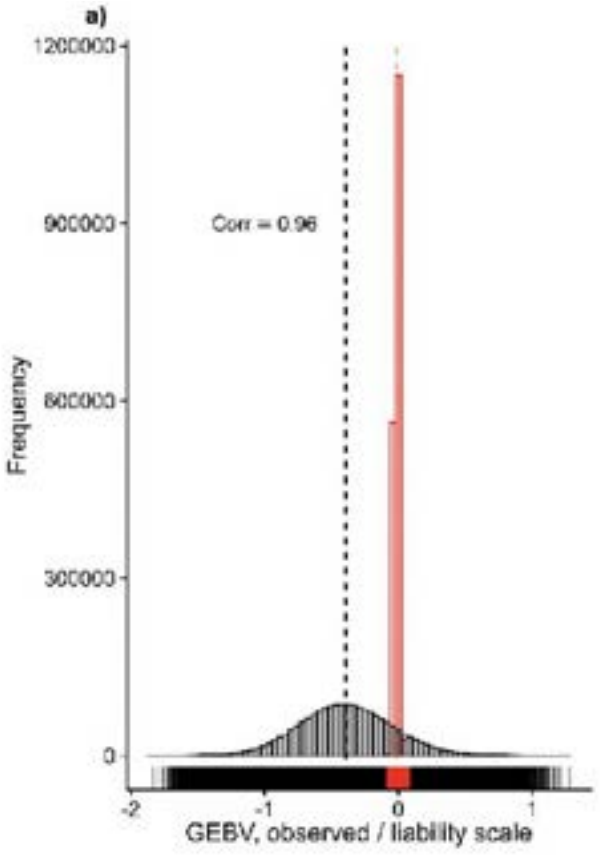
Mastitis
$h^2 = 0.09$
% = 27

Genomic predictions – binary/categorical

J. Dairy Sci. TBC
https://doi.org/10.3168/jds.2024-24767

Converting estimated breeding values from the observed to probability scale for health traits

Jorge Hidalgo,[1]* Shogo Tsuruta,[1] Dianelys Gonzalez,[2] Gerson de Oliveira,[2] Miguel Sanchez,[2] Asmita Kulkarni,[2] Cory Przybyla,[2] Giovana Vargas,[2] Natascha Vukasinovic,[2] Ignacy Misztal,[1] and Daniela Lourenco[1]
[1]Department of Animal and Dairy Science, University of Georgia, Athens, GA, 30602, USA
[2]Zoetis Genetics and Precision Animal Health, Kalamazoo, MI, 49007, USA

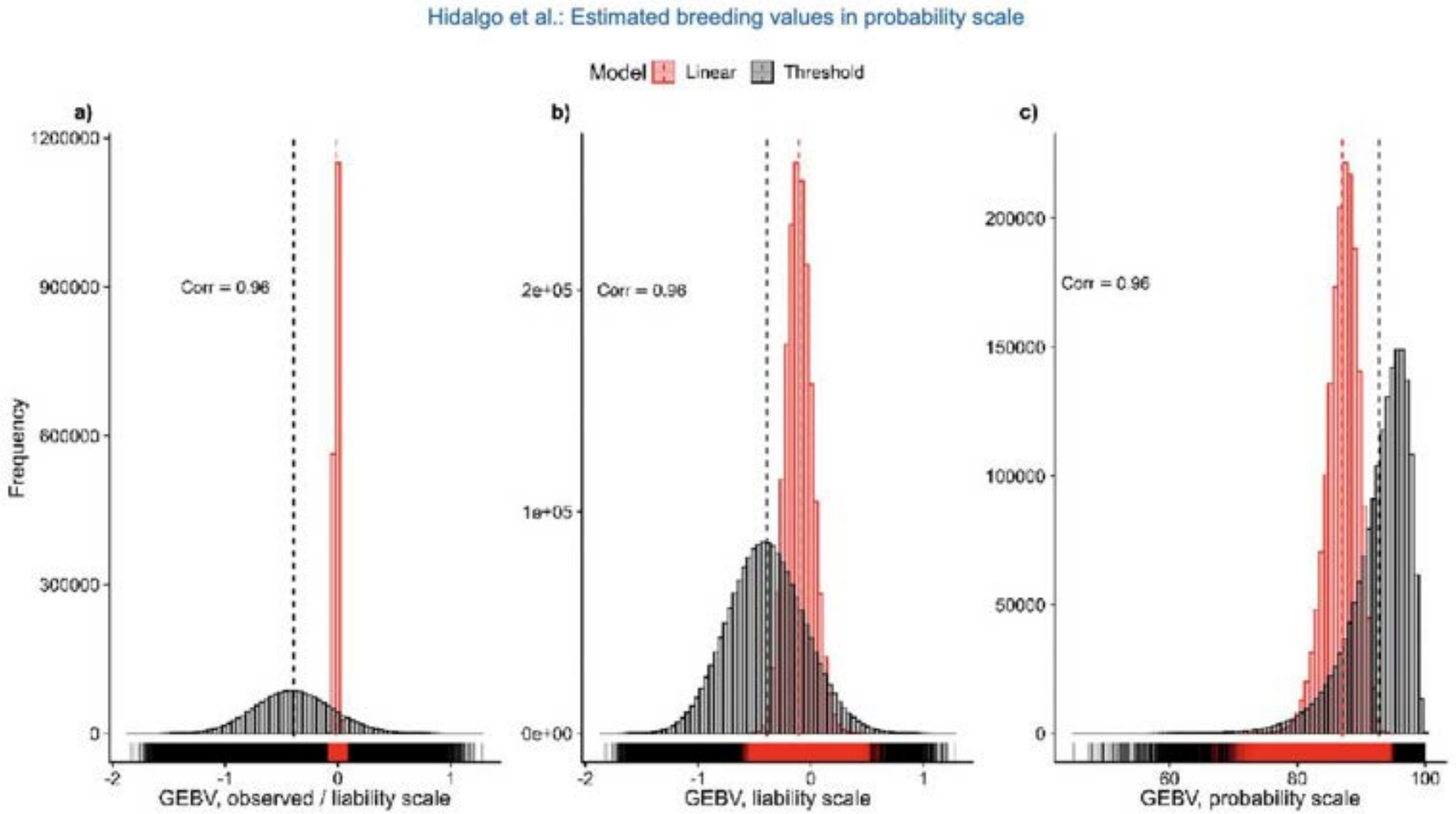


$$GEBV_{lia} \approx \frac{GEBV_{lin}}{\sqrt{\sigma^2_{e_{lin}} * \left(1 - \frac{h^2_{lin}}{h^2_{lia}}\right)}}$$

$$P_i = 1 - \Phi\left(\frac{t - \mu_u - u_i}{\sigma_e}\right)$$

Displaced abomasum
$h^2 = 0.09$
% = 2

Genomic predictions – binary/categorical



Displaced abomasum
$h^2 = 0.09$
% = 2

Genomic predictions – binary/categorical

- Second approach for categorical/binary traits:
  1) Linear model until convergence (BLUP)
  2) Compute pseudo-phenotypes based on residuals (EM approach; Quaas, 1994)
  3) Iterate back to 1 using pseudo-phenotypes
  4) Do it until pseudo-phenotypes “do not change anymore”

- CATEGF90 as a wrapper for BLUP90IOD3
  - Benefit: allows for multiple categorical/binary traits

Jennifer Richter

Andres Legarra

Fernando Bussiman

Genomic predictions – binary/categorical

Trait	# of Records	1	2	3 to 7
Ascites (AC)	163,971	161,950 (98.8%)	2,021 (1.2%)	-
Tibial Dyschondroplasia (TD)	59,124	57,995 (98.1%)	1,129 (1.9%)	-
Mortality (MT)	180,998	167,389 (92.4%)	13,609 (7.5%)	-
Femoral Head Necrosis (FN)	16,870	13,112 (77.7%)	2,295 (13.6%)	1,463 (8.7%)

	GIBBS	CBLUP	CATEGF90	GIBBS	CBLUP	CATEGF90
	AC			TD		
GIBBS	1	0.99	0.99	1	0.99	0.99
CBLUP		1	1		1	1
CATEGF90			1			1
	MT			FN		
GIBBS	1	1	1	1	1	1
CBLUP		1	1		1	1
CATEGF90			1			1

Jennifer Richter

Andres Legarra

Fernando Bussiman

Genomic predictions – binary/categorical

Trait	# of Records	1	2	3 to 7
Ascites (AC)	1,092,037	1,083,836 (99.2%)	8,201 (0.8%)	-
Tibial Dyschondroplasia (TD)	365,676	354,143 (96.8%)	11,533 (3.2%)	-
Mortality (MT)	1,191,175	1,093,992 (91.8%)	97,183 (8.2%)	-
Femoral Head Necrosis (FN)	88,012	58,817 (66.8%)	16,935 (19.2%)	12,260 (13.9%)

Jennifer Richter

Andres Legarra

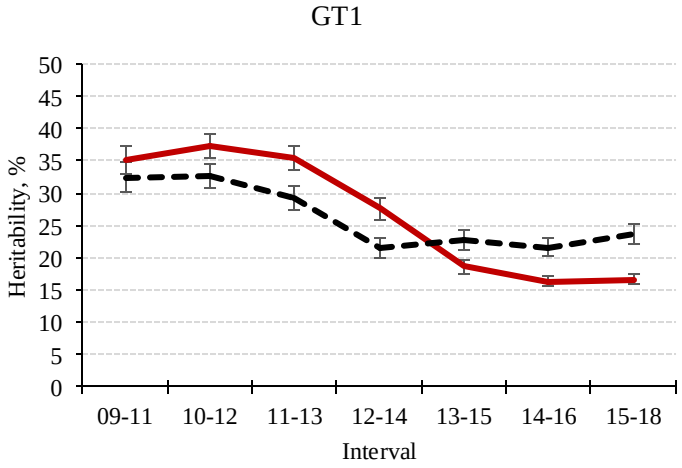
Fernando Bussiman

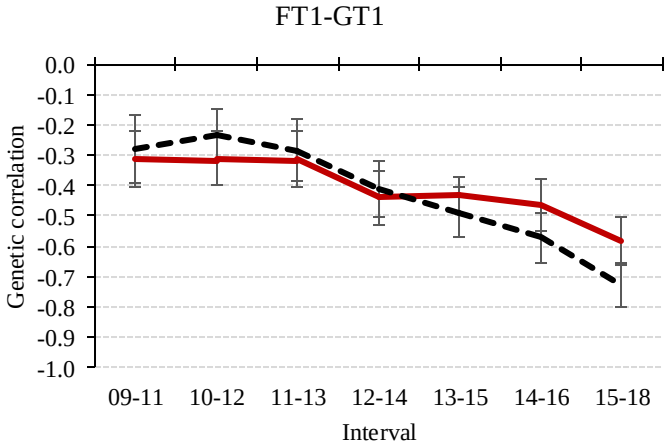
	CBLUP	CATEGF90	CBLUP	CATEGF90
	AC		TD	
CBLUP	1	1	1	1
CATEGF90		1		1
	MT		FN	
CBLUP	1	1	1	1
CATEGF90		1		1

CATEGF90 iter = 515
BLUP90IOD3 iter = 461,546
Wall-clock time (min) = 15,897

Limitations - VCE

- Faster changes with genomic selection
- Different genetic parameters with and without genomics





JOURNAL ARTICLE

Changes in genetic parameters for fitness and growth traits in pigs under genomic selection

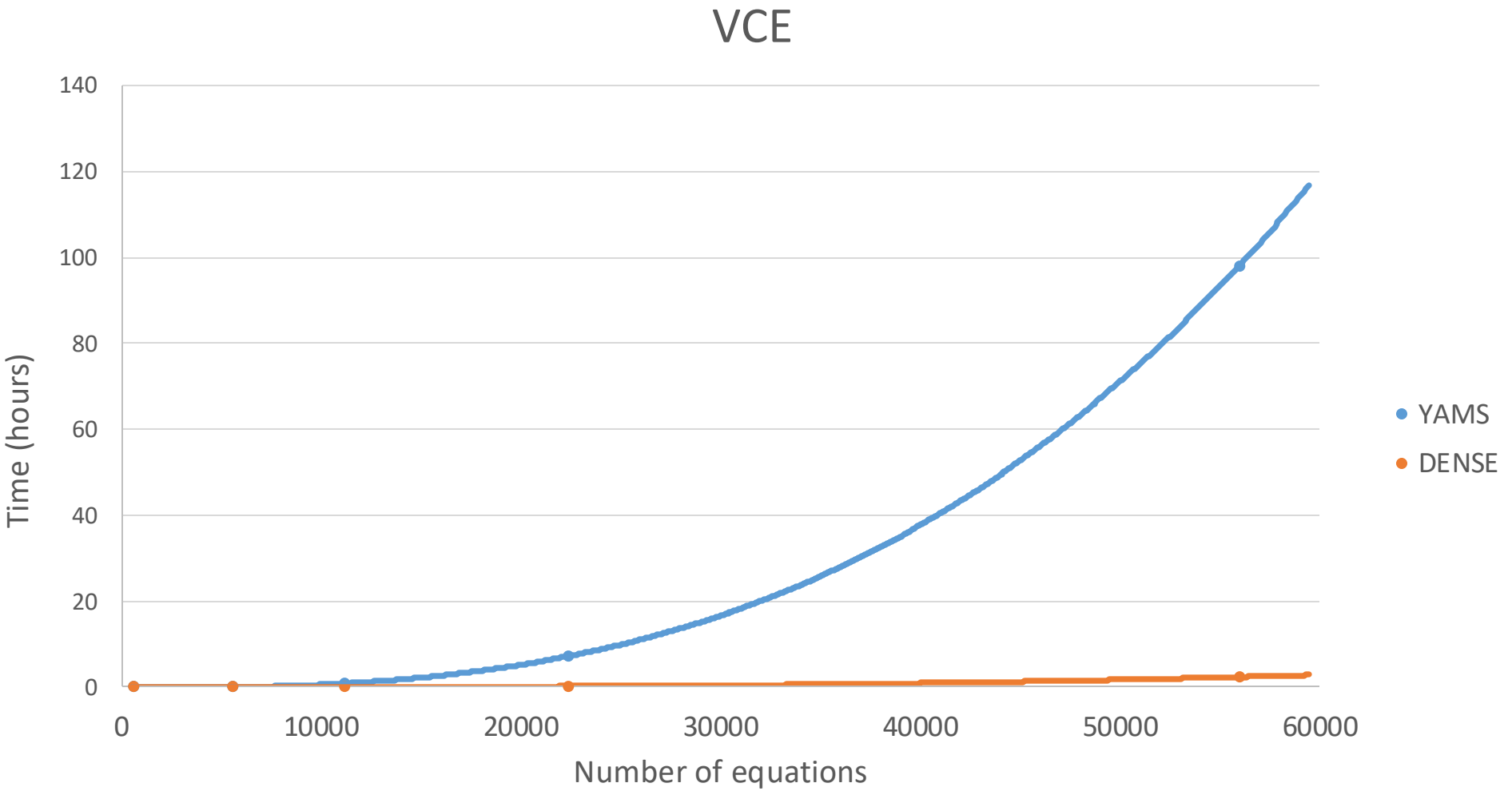
Jorge Hidalgo, Shogo Tsuruta, Daniela Lourenco, Yutaka Masuda, Yijian Huang, Kent A Gray, Ignacy Misztal

Journal of Animal Science, Volume 98, Issue 2, February 2020, skaa032,

- Hard to estimate VC with many genotyped individuals
  - Software optimization
  - New methods

Tsuruta et al. (in progress)
Gowane et al. (in progress)

Efficient VCE – software optimization



Bermann et al.
(unpublished)

Efficient VCE – New methods

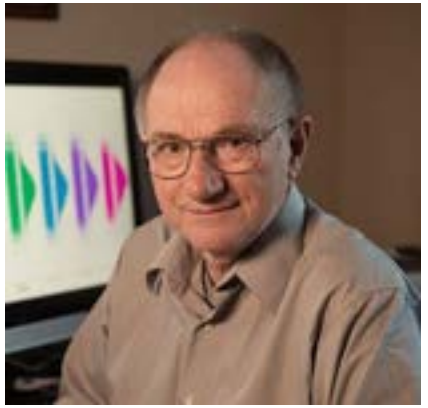
Formulas for estimating heritability

$$\widehat{h^2}: \sqrt{\frac{Nh^2}{Nh^2 + M_e}} = corr\,(y - Xb, \hat{u})/h$$

$$\widehat{h^2} = \frac{c^2 + \sqrt{c^4 + 4c^2 M_e/N}}{2}, c = corr(y - Xb, \hat{u})$$

$$SE(\widehat{h^2}) \approx \frac{1}{\sqrt{N_{val}}}\left[c + \frac{2c^2 + \frac{4M_e}{N}}{\sqrt{c^2 + \frac{4M_e}{N}}}\right] \widehat{h^2}) \approx \frac{3c}{\sqrt{N_{val}}}$$

N – # animals in reference N_{val} – number of animals in validation



How to estimate genetic correlations?

Predictivity for trait i $corr(y_i - Xb_i, \hat{u}_i) = acc_i\ h_i$

What is predictivity from trait i to trait j? $corr(y_i - Xb_i, \hat{u}_j) = ?$

.....

.....

$$corr(y_i - Xb_i, \hat{u}_j) = acc_j\ corr_{ij}\ h_i$$

$$corr_{ij} = \frac{corr(y_i - Xb_i, \hat{u}_j)}{h_i\ acc_j}$$

$$SD(corr_{ij}) \approx \frac{1}{h_i\ acc_j \sqrt{N_{val}}}$$

UGA AB&G team