

Adapting single-step GBLUP for complex data, models, and sequence information

Progress report -2023

Ignacy Misztal, Daniela Lourenco Yutaka Masuda,
Tom Lawlor, and Andres Legarra



UNIVERSITY OF
GEORGIA

College of Agricultural &
Environmental Sciences

Goals

- Develop better ssGBLUP predictive models for large data sets with genotyped animals that have incomplete pedigrees and may be from different breeds
 - Single breed
 - **Multiple breeds**
 - Multiple breeds and crossbreds
- Establish a robust approximation of individual theoretical accuracy for very large genotyped populations using the APY algorithm
 - **Developed and implemented**
- Enable computations of p-values in ssGBLUP to select sequence variants for genomic prediction in very large genotyped populations
 - **Developed and implemented**
 - Features of GWAS in populations with small effective population size

Single-step GBLUP

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1} \frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

Single-step GBLUP
(ssGBLUP)

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



Pedigree



Genomic



Pedigree for
genotyped

Aguilar et al. (2010)
Christensen and Lund (2010)

Segments

Theory of junctions Fisher (1949)

$$E(Me) = 4N_e L$$

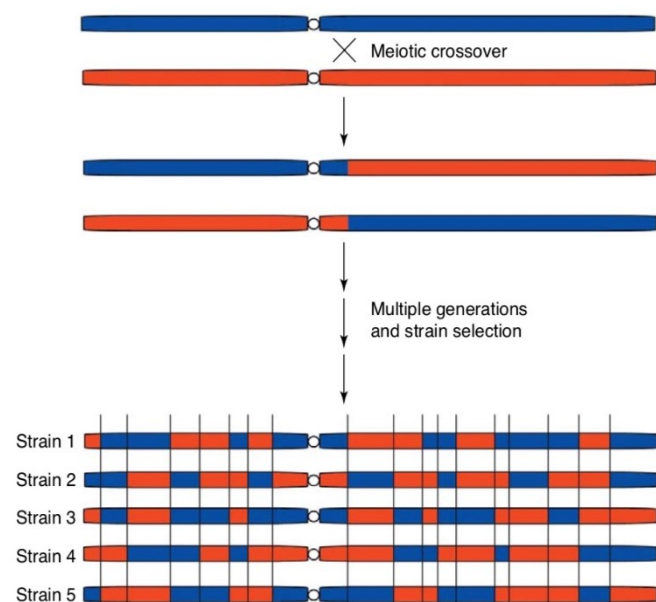
Stam (1980)

Points where the founder chromosome of origin changes

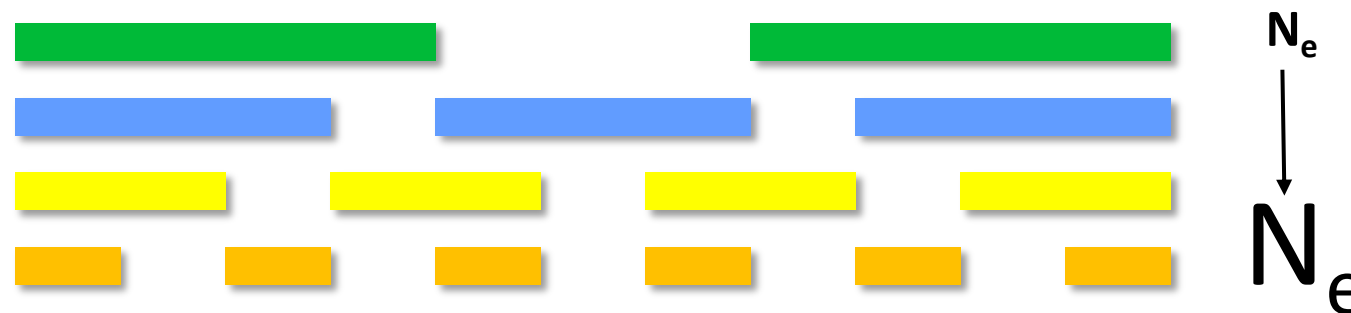
Me – Independent chromosome segments

N_e – Effective population size

L – Length of genome in Morgans



Cuppen (2005)



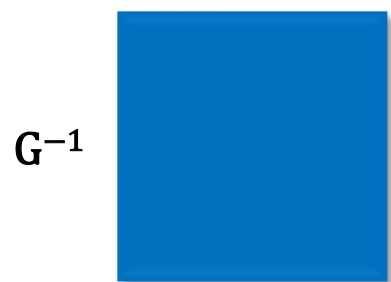
About 10-15k segments in cattle, 5 k in pigs and chickens

Misztal (2016), Pocrnic et al., (2016)

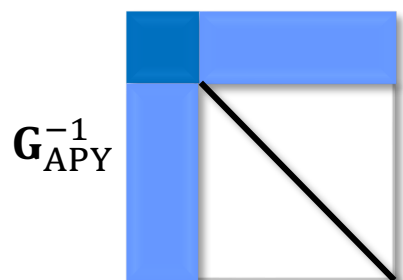
Algorithm for Proven and Young (APY)

- Realized relationship matrix in ssGBLUP

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



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



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

Condition on a set of features or animals = CORE animals

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



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



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

Masuda et al. (2016)

Multiple breed evaluation of dairy data

Breed	Phenotypes		Animals	
	N	Cows	Genotypes	Total
All	45M	19.4M	3.9M	29.5M
Ayrshire	116k	47k	9.2k	94k
Brown Swiss	328k	138k	47k	292k
Guernsey	129k	58k	5k	100k
Holstein	40.3M	17.5M	3.4M	26.6M
Jersey	4.1M	1.7M	427k	2.5M

ssGBLUP with 4 M genotyped animals

Core animals across breeds

AY = 32 (9.2k)

BS = 182 (47k)

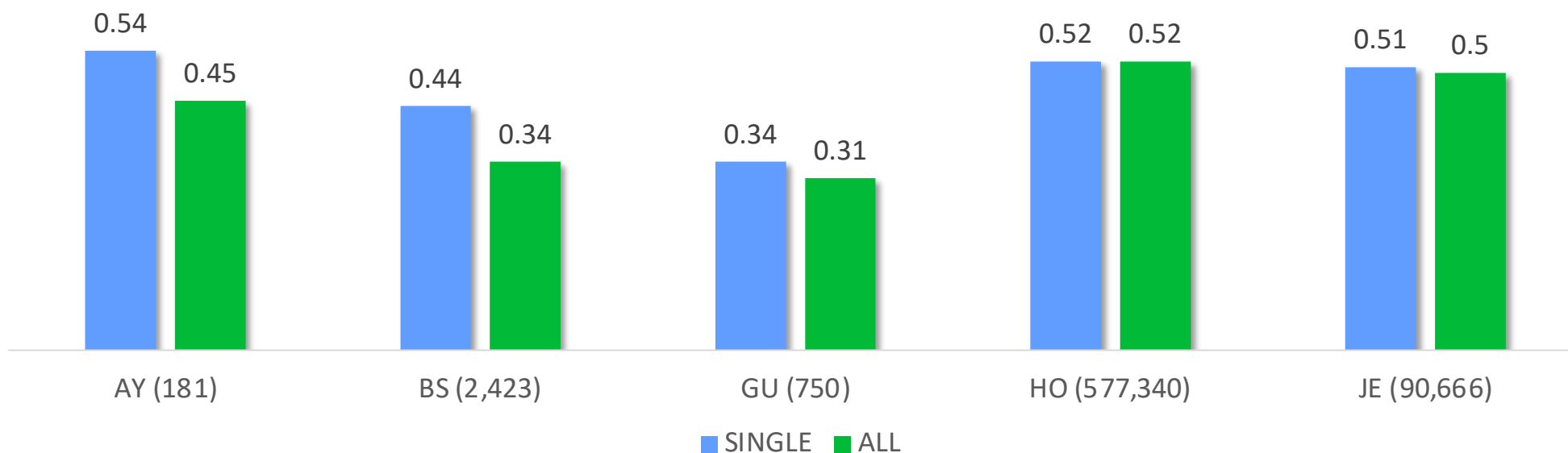
GU = 17 (5k)

HO = 13k (3.4M)

JE = 1.7k (427k)

ALL = 15k core

Predictability for cows - Protein



ssGBLUP with 4 M genotyped animals

Dimensionality within each breed

AY = 5k

BS = 5k

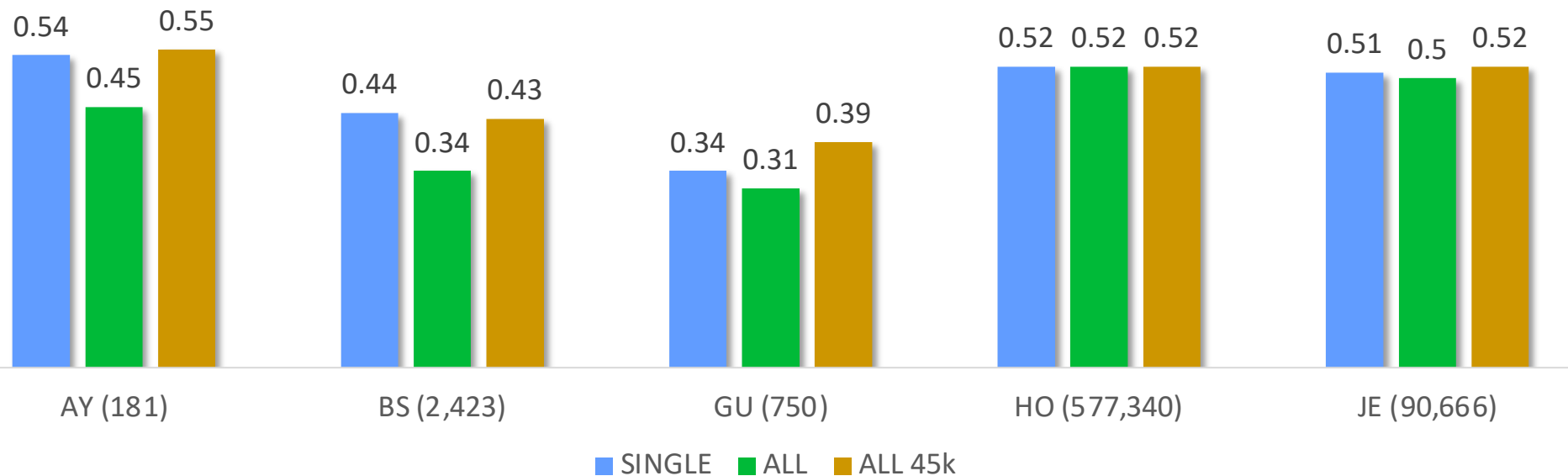
GU = 5k

HO = 15k

JE = 15k

Accuracy for cows - Protein

ALL 45k = 45k core




Accuracy approximations

- Weights based on phenotypes and pedigrees
- Block sparse inversion with APY inverse

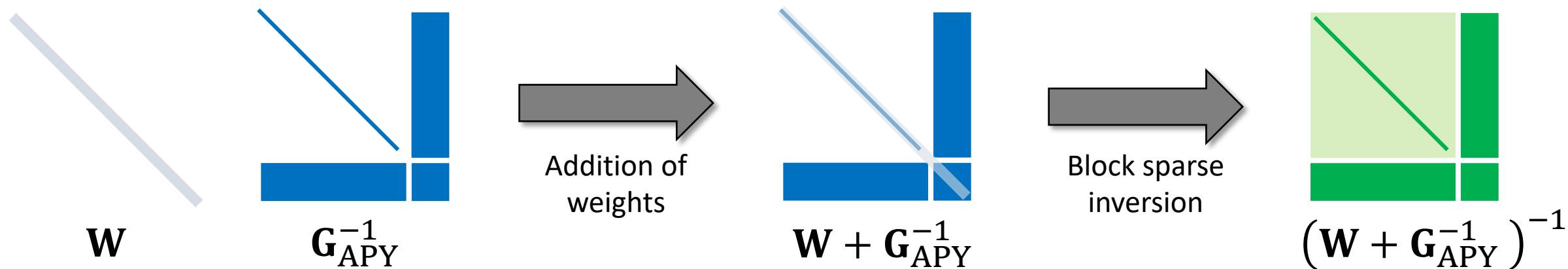


JOURNAL ARTICLE

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

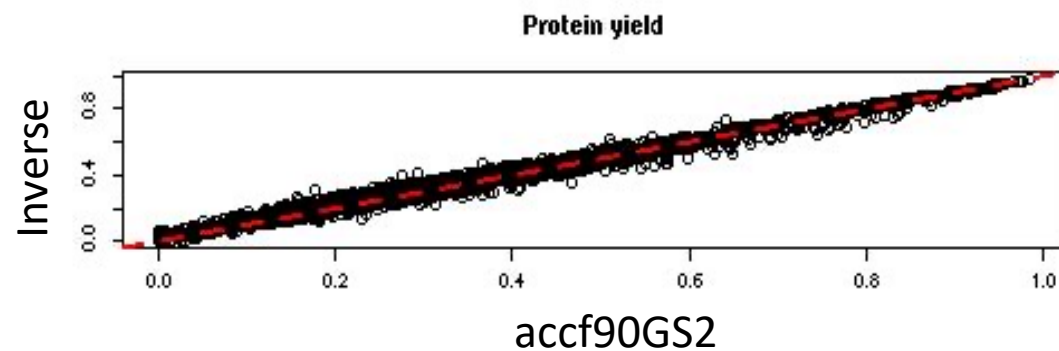
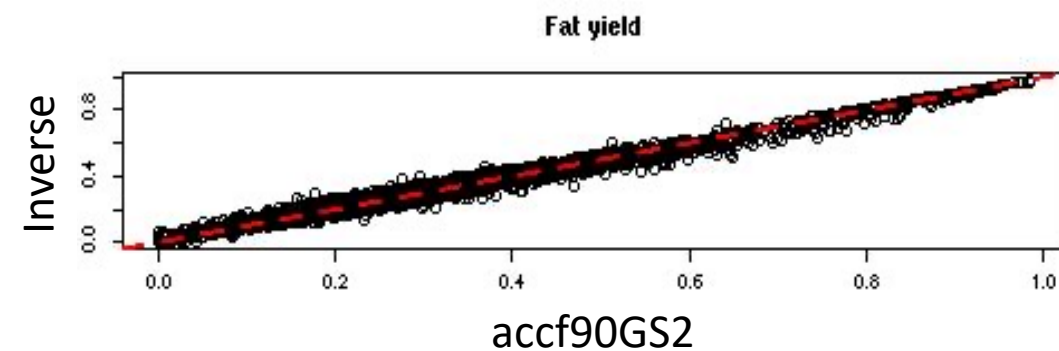
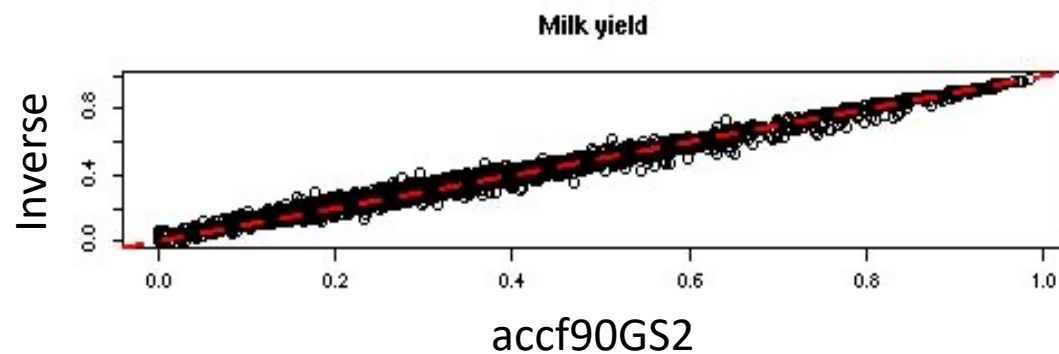
Matias Bermann , Daniela Lourenco, Ignacy Misztal

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



$$diag(W + G_{APY}^{-1})^{-1} = \frac{diag((W_{nn} + M_{nn}^{-1})^{-1} + (W_{nn} + M_{nn}^{-1})^{-1} G^{nc} (W_{cc} + G^{cc} - G^{cn} (W_{nn} + M_{nn}^{-1})^{-1} G^{nc})^{-1} G^{cn} (W_{nn} + M_{nn}^{-1})^{-1})}{diag((W_{cc} + G^{cc} - G^{cn} (W_{nn} + M_{nn}^{-1})^{-1} G^{nc})^{-1})}$$

GEBV are published with accuracy



Cesarani et al.
(unpublished)

P-values for GWAS in (ss)GBLUP

$$pval_i = 2 \left(1 - \Phi \left(\left| \frac{\widehat{snp}_i}{sd(\widehat{snp}_i)} \right| \right) \right) \quad (\text{Chen et al., 2017})$$

Algorithm

1. Calculate PEV for GEBV
2. Convert GEBV to SNP effects
3. Convert PEV/GEBV to PEV/SNP

If $sd(\widehat{snp}_i)$ approximately constant, Manhattan plots based on $|\widehat{snp}_i|$ and $pval_i$ similar

Large data – PEV from accuracy approximations (Bermann et al., 2021)

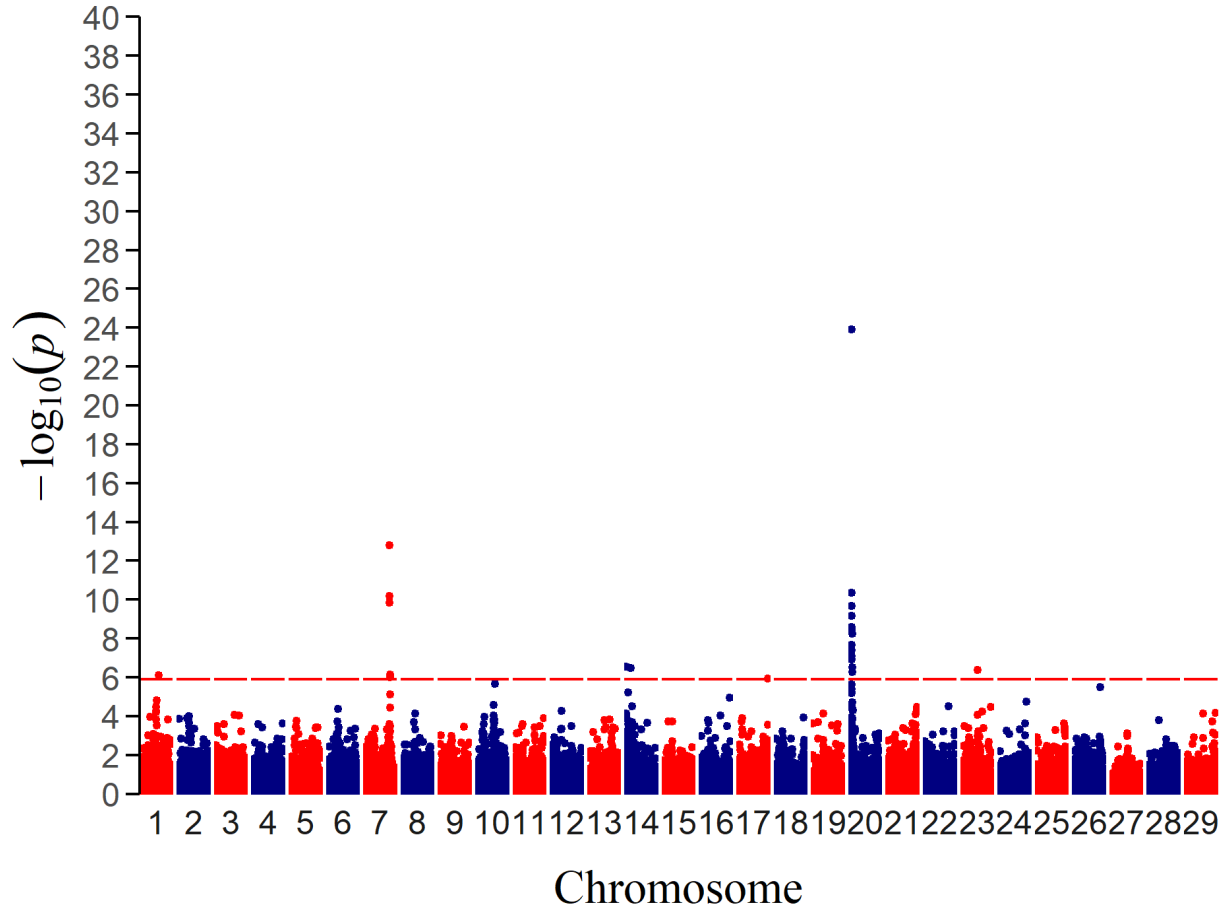
Application example

- Post-weaning gain in American Angus
- 845,000 phenotypes
- 450,000 genotypes
- 1,570,000 animals in the pedigree
- ssGBLUP (50k genotyped animals) vs. APY-ssGBLUP (450k genotyped animals)
- We expect:
 - Higher power
 - Less noise
 - Less false-positives

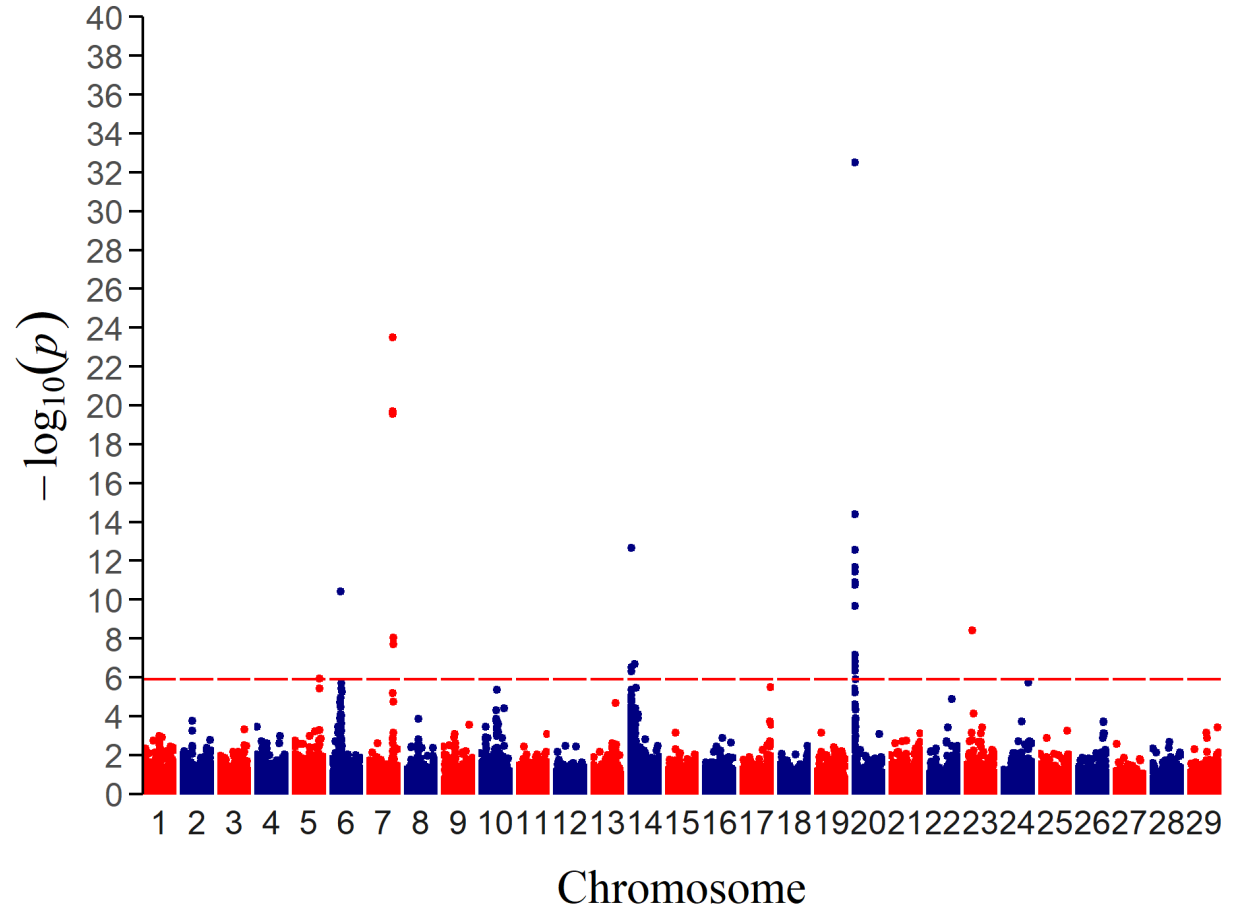


Leite et al.
(in progress)

50k genotyped animals



500k genotyped animals

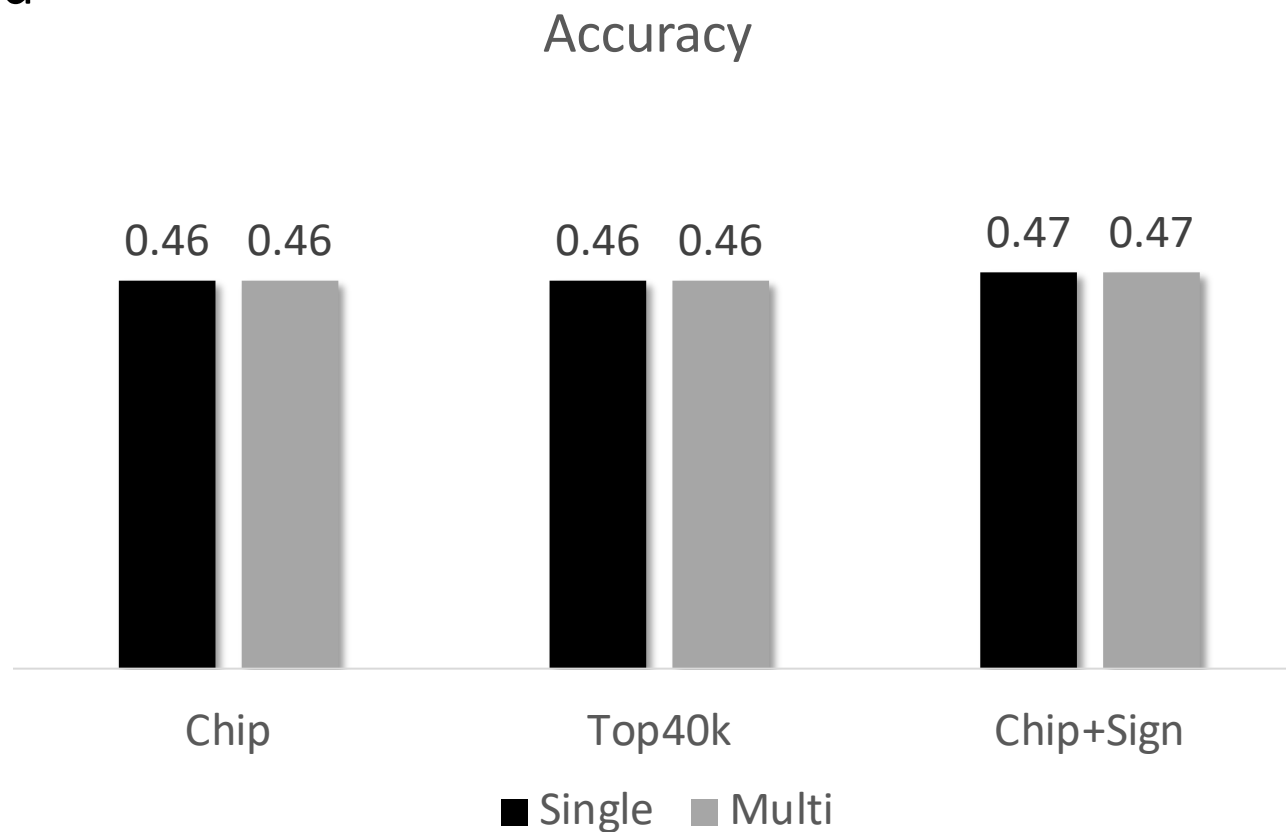


Sequence data

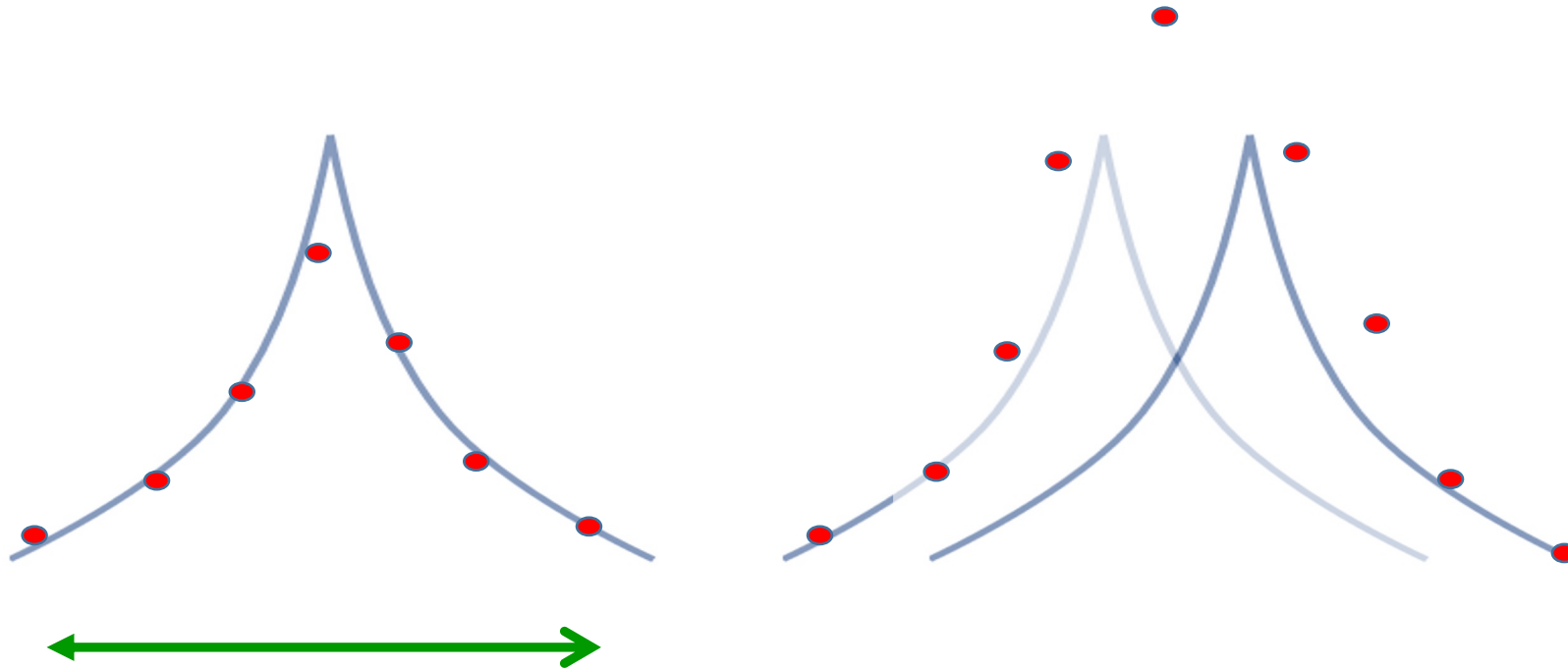
- 207k pigs with sequence
- 5M pedigree
- 1.5M records
- Single and multi-breed



Jang et al.
(under review)



With large data (ss)GBLUP accounts for QTN



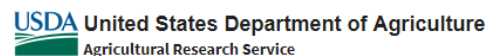
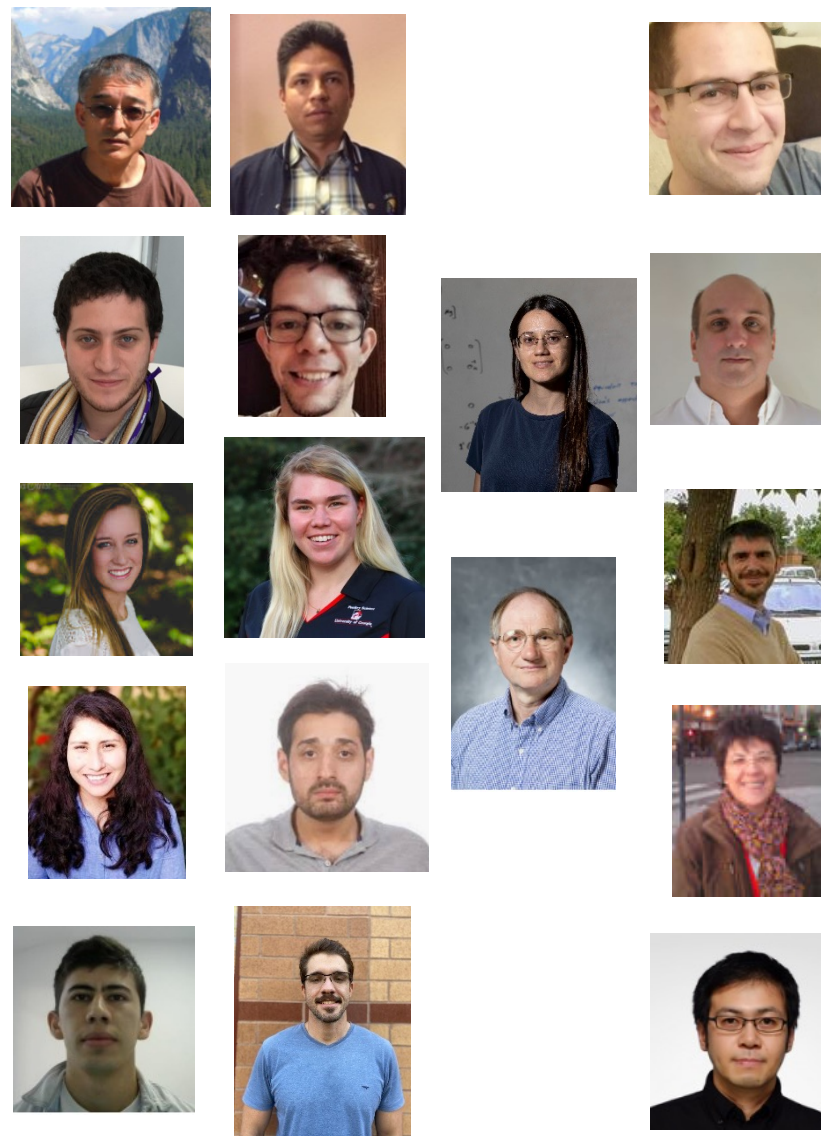
~ 2Mb in cattle

~ 6Mb in pigs and chickens

Conclusions

- Genomic information has limited dimensionality in farm animals
 - About 15,000 in cattle
- Limits can be exploited without affecting accuracy for:
 - Genomic evaluation of any size
 - Accuracy approximation of any size
 - GWAS of any size
- Few associations found with large data

UGA AB&G team



Warmwater Aquaculture Research Unit



Cool and Cold Water Aquaculture Research

