

Confidence intervals for validation statistics with data truncation in genomic prediction

M. Bermann, A. Legarra, A. Alvarez Munera,
I. Misztal, D. Lourenco

09/2024



UNIVERSITY OF
GEORGIA

College of Agricultural &
Environmental Sciences

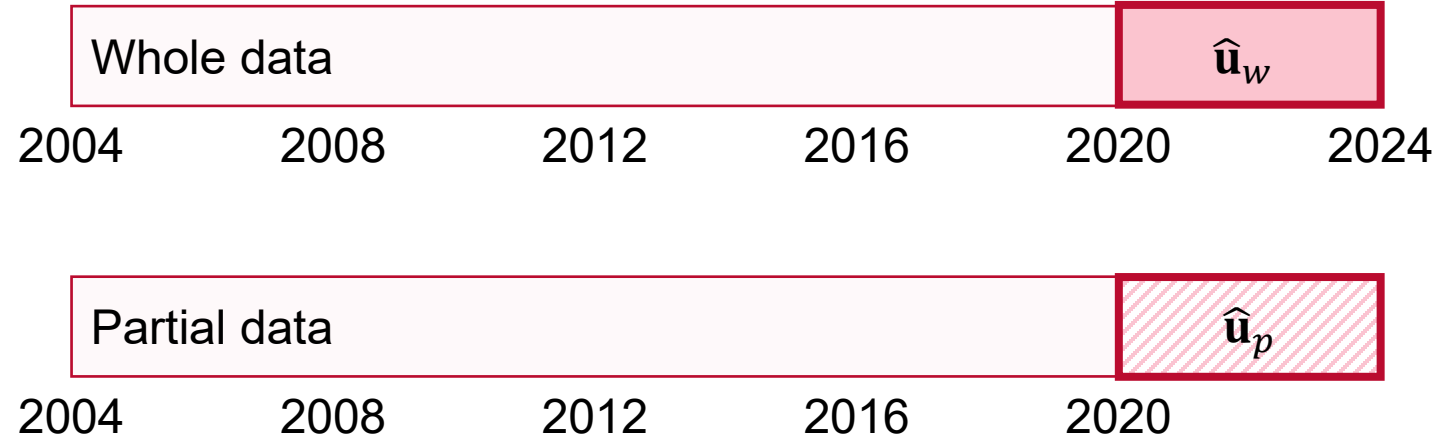
*Animal Breeding and
Genetics Group*

Introduction

- Forward validation
 - Model selection/evaluation
 - Predict genetic progress
- Validation statistics
 - Random variable
 - No analytical confidence intervals (CI)
 - Bootstrap is an option
- **Objectives**
 - **Derive analytical CI for forward validation statistics**
 - **Compare them vs. bootstrap CI**



Forward validation



- Comparison between $\hat{u}_w$ and $\hat{u}_p$
- Comparison with adjusted phenotypes (y^*)
- LR method and predictivity

LR method and predictivity

Bias

$$\mu_{wp} = \text{mean}(\hat{\mathbf{u}}_p - \hat{\mathbf{u}}_w)$$

0 unbiased



LR method and predictivity

Bias

$$\mu_{wp} = \text{mean}(\hat{\mathbf{u}}_p - \hat{\mathbf{u}}_w)$$

0 unbiased

Dispersion

$$b_{wp} = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\text{var}(\hat{\mathbf{u}}_p)}$$

1 no over/under dispersion

LR method and predictivity

Bias

$$\mu_{wp} = \text{mean}(\hat{\mathbf{u}}_p - \hat{\mathbf{u}}_w)$$

0 unbiased

Dispersion

$$b_{wp} = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\text{var}(\hat{\mathbf{u}}_p)}$$

1 no over/under dispersion

Reliability

$$\rho_{cov_{wp}}^2 = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\sigma_{g_i}^2}$$

0 - 1

LR method and predictivity

Bias

$$\mu_{wp} = \text{mean}(\hat{\mathbf{u}}_p - \hat{\mathbf{u}}_w)$$

0 unbiased

Dispersion

$$b_{wp} = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\text{var}(\hat{\mathbf{u}}_p)}$$

1 no over/under dispersion

Reliability

$$\rho_{cov_{wp}}^2 = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\sigma_{g_i}^2}$$

0 - 1

Increase in accuracy

$$\rho_{wp} = \text{corr}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)$$

0 - 1

LR method and predictivity

Bias

$$\mu_{wp} = \text{mean}(\hat{\mathbf{u}}_p - \hat{\mathbf{u}}_w)$$

0 unbiased

Dispersion

$$b_{wp} = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\text{var}(\hat{\mathbf{u}}_p)}$$

1 no over/under dispersion

Reliability

$$\rho_{cov_{wp}}^2 = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\sigma_{g_i}^2}$$

0 - 1

Increase in accuracy

$$\rho_{wp} = \text{corr}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)$$

0 - 1

Predictive ability

$$\rho_{\mathbf{y}^*, \hat{\mathbf{u}}_p} = \frac{\text{corr}(\mathbf{y}^*, \hat{\mathbf{u}}_p)}{\sqrt{h^2}}$$

0 - 1

Confidence intervals

- Frequentist
 - $CI_{100(1-\alpha)}(\theta) = \theta \pm z_{1-\frac{\alpha}{2}} sd(\theta)$
- Bootstrap
 - For n samples
 - Subset $\hat{\mathbf{u}}_w$, $\hat{\mathbf{u}}_p$, and $\mathbf{y}^*$
 - Calculate validation statistics
 - Empirical CI
 - Non-parametric
- Bayesian

RESEARCH ARTICLE

Open Access

Confidence intervals for validation statistics with data truncation in genomic prediction



Matias Bermann^{1*} , Andres Legarra², Alejandra Alvarez Munera¹, Ignacy Misztal¹ and Daniela Lourenco¹

Frequentist CI for LR and predictivity

Bias

$$\mu_{wp} = \text{mean}(\hat{\mathbf{u}}_p - \hat{\mathbf{u}}_w)$$

$$\text{CI}(\mu_{wp}) = \mu_{wp} \pm z_{1-\frac{\alpha}{2}} \sqrt{n^{-2} \mathbf{1}'(\mathbf{C}_p^{22} - \mathbf{C}_w^{22}) \mathbf{1}}$$

Dispersion

$$b_{wp} = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\text{var}(\hat{\mathbf{u}}_p)}$$

$$\text{CI}(b_{wp}) = b_{wp} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{tr}(\mathbf{S}(\mathbf{C}_p^{22} - \mathbf{C}_w^{22})\mathbf{S}(\mathbf{G} - \mathbf{C}_p^{22}))}{2 \text{tr}(\mathbf{S}(\mathbf{G} - \mathbf{C}_p^{22})\mathbf{S}(\mathbf{G} - \mathbf{C}_p^{22})) + \text{tr}(\mathbf{S}(\mathbf{G} - \mathbf{C}_p^{22}))^2}}$$

Reliability

$$\rho_{cov_{wp}}^2 = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\sigma_{g_i}^2}$$

$$\text{CI}(\rho_{cov_{wp}}^2) = \rho_{cov_{wp}}^2 \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{tr}(\mathbf{S}(\mathbf{C}_p^{22} - \mathbf{C}_w^{22})\mathbf{S}(\mathbf{G} - \mathbf{C}_p^{22})) + 2 \text{tr}(\mathbf{S}(\mathbf{G} - \mathbf{C}_p^{22})\mathbf{S}(\mathbf{G} - \mathbf{C}_p^{22}))}{n \sigma_{g_i}^2}}$$

Increase in accuracy

$$\rho_{wp} = \text{corr}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)$$

$$\text{CI}(\rho_{wp}) = \tanh\left(\tanh^{-1}(\rho_{wp}) \pm z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}\right)$$

Predictive ability

$$\rho_{\mathbf{y}^*, \hat{\mathbf{u}}_p} = \frac{\text{corr}(\mathbf{y}^*, \hat{\mathbf{u}}_p)}{\sqrt{h^2}}$$

$$\text{CI}(\rho_{\mathbf{y}^*, \hat{\mathbf{u}}_p}) = \frac{1}{h} \tanh\left(h \tanh^{-1}(\rho_{\mathbf{y}^*, \hat{\mathbf{u}}_p}) \pm z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}\right)$$

Frequentist CI for LR and predictivity

- Complicated formulas
 - Prediction error (co)variances
- Computationally expensive for large datasets
- Approximation
 - Unrelated
 - Non-inbred



Approx. frequentist CI for LR and predictivity

Bias

$$\mu_{wp} = \text{mean}(\hat{\mathbf{u}}_p - \hat{\mathbf{u}}_w)$$

$$\text{CI}(\mu_{wp}) \approx \mu_{wp} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_g^2}{n} (\overline{rel}_w - \overline{rel}_p)}$$

Dispersion

$$b_{wp} = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\text{var}(\hat{\mathbf{u}}_p)}$$

$$\text{CI}(b_{wp}) \approx b_{wp} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{(c-1)(\text{Var}(\overline{rel}_p) + \overline{rel}_p^2)}{2(\text{Var}(\overline{rel}_p) + \overline{rel}_p^2) + n \overline{rel}_p^2}}$$

Reliability

$$\rho_{cov_{wp}}^2 = \frac{\text{cov}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)}{\sigma_{g_i}^2}$$

$$\text{CI}(\rho_{cov_{wp}}^2) \approx \rho_{cov_{wp}}^2 \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{(1+c)\sigma_g^4}{n \sigma_{g_i}^4} (\text{Var}(\overline{rel}_p) + \overline{rel}_p^2)}$$

Increase in accuracy

$$\rho_{wp} = \text{corr}(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)$$

$$\text{CI}(\rho_{wp}) = \tanh \left(\tanh^{-1}(\rho_{wp}) \pm z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}} \right)$$

Predictive ability

$$\rho_{\mathbf{y}^*, \hat{\mathbf{u}}_p} = \frac{\text{corr}(\mathbf{y}^*, \hat{\mathbf{u}}_p)}{\sqrt{h^2}}$$

$$\text{CI}(\rho_{\mathbf{y}^*, \hat{\mathbf{u}}_p}) = \frac{1}{h} \tanh \left(h \tanh^{-1}(\rho_{\mathbf{y}^*, \hat{\mathbf{u}}_p}) \pm z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}} \right)$$

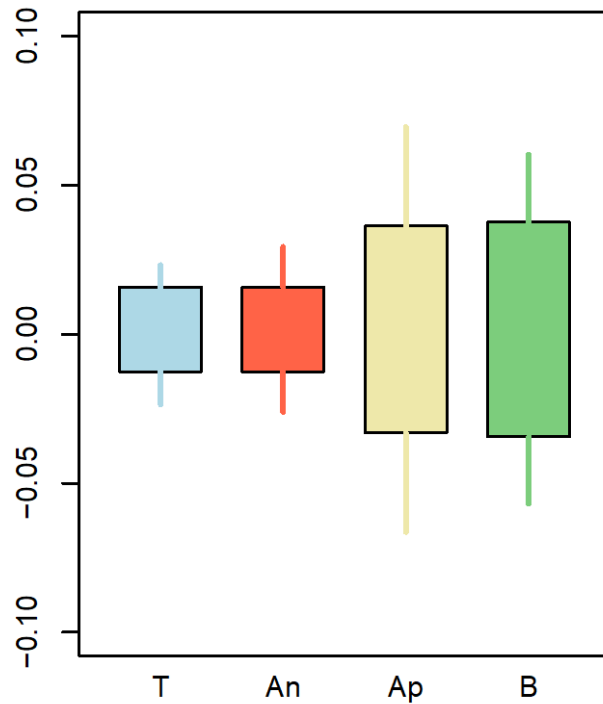
Approx. frequentist CI for LR and predictivity

- Easy to calculate
- Approximation of accuracies for focal animals
- Routine evaluations

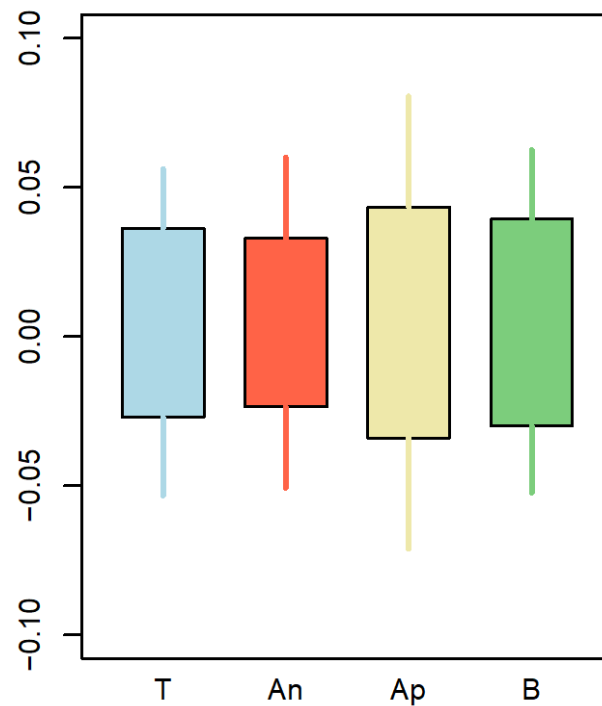
Simulated example

- Single-trait animal model
- ~5000 animals
- Grid of h^2 and proportion of animals with phenotypes (p)
- Phenotypes simulated from $N(\mathbf{Xb}, \mathbf{ZAZ}' \sigma_g^2 + \mathbf{I} \sigma_e^2)$
- CI
 - True
 - Frequentist
 - Approximated
 - Bootstrap

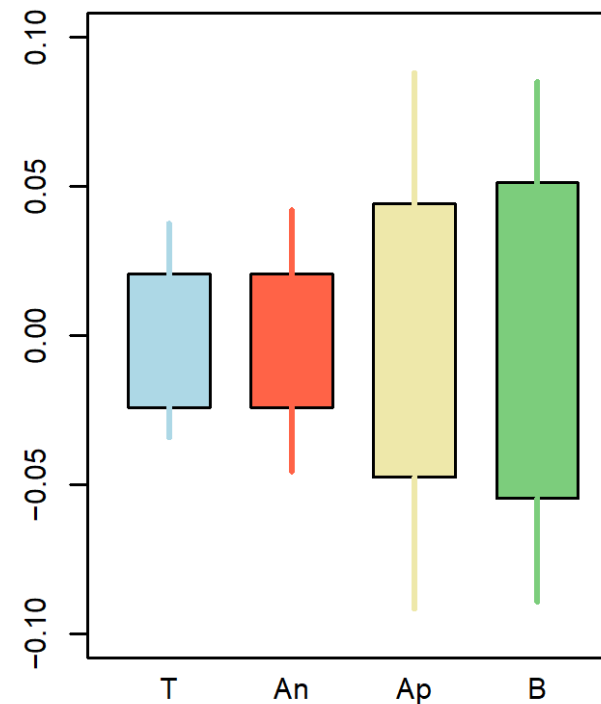
Results: bias



High h^2
High p

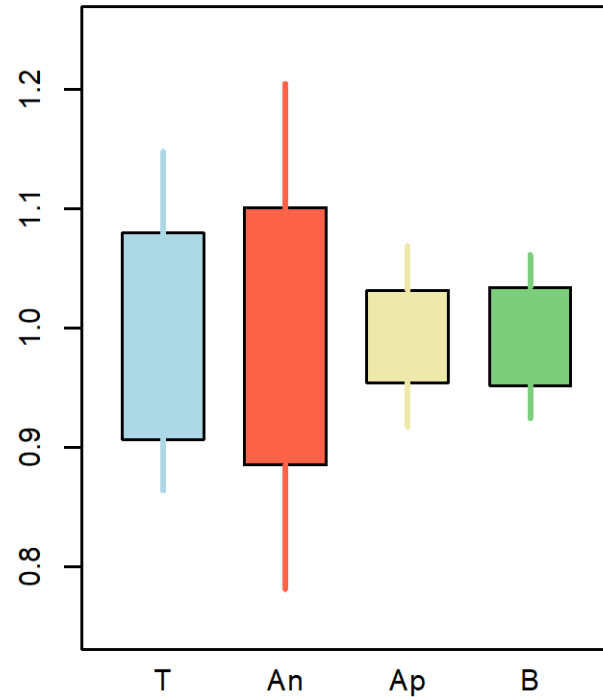


Moderate h^2
Moderate p

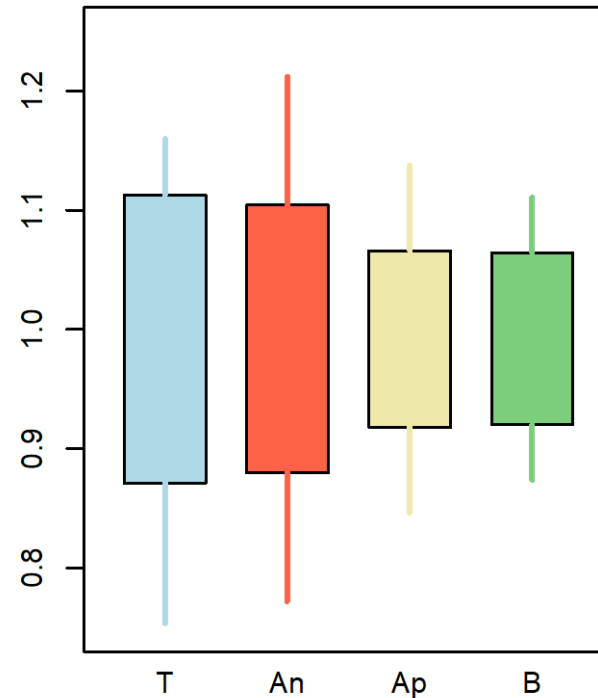


Low h^2
Low p

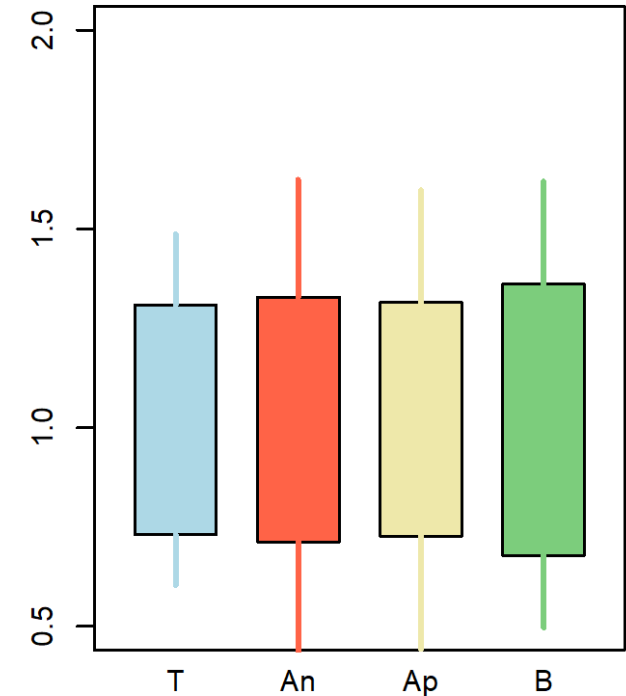
Results: dispersion



Low h^2
High p

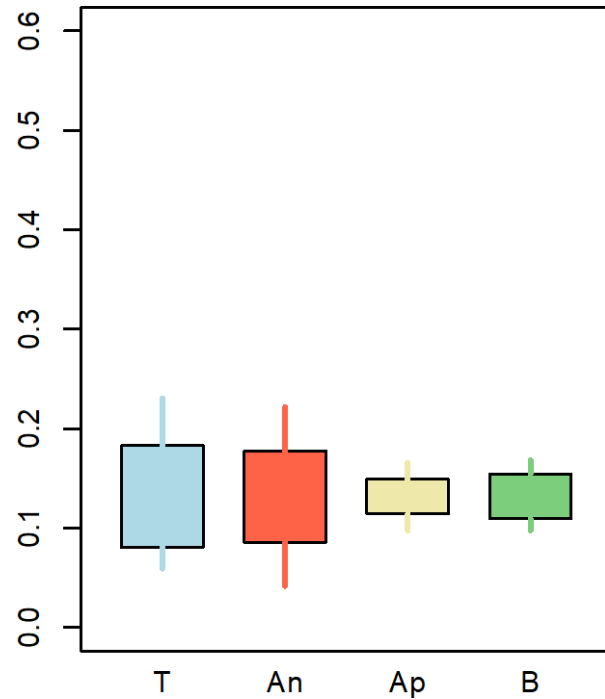


Moderate h^2
Moderate p

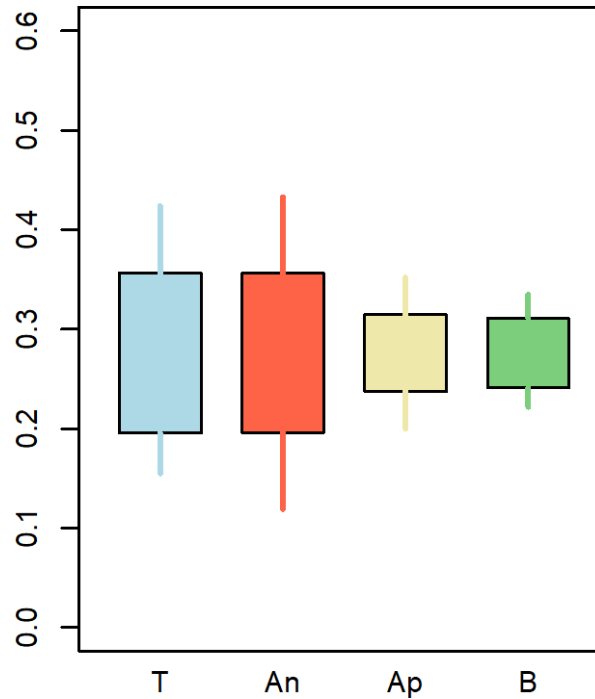


High h^2
Low p

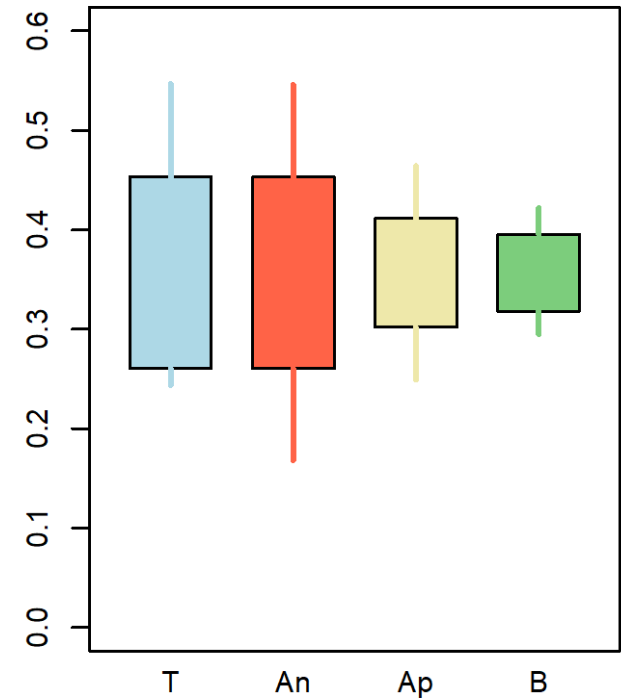
Results: reliability



Low h^2
Moderate p



Moderate h^2
Moderate p



High h^2
Moderate p

Implementation

- Statistics are easy, CI not
- Validationf90
 - Solutions with blupf90+
 - All statistics
 - Three options for CI

readme.validationf90

VALIDATIONF90

Summary

This program processes solutions obtained with blupf90+ or gibbsf90+ to obtain validation statistics presented in [Legarra et al. 2008](#) and [Legarra et al. 2018](#). Confidence intervals for validation statistics are calculated as in [Bermann et al. 2024](#).

The program works by obtaining solutions with a training (partial) dataset and the complete (whole) dataset. Then, validation statistics are obtained from those solutions and a list of validation individuals.

Table of Contents

- ♦ VALIDATIONF90
 - ♦ Summary
 - ♦ Options

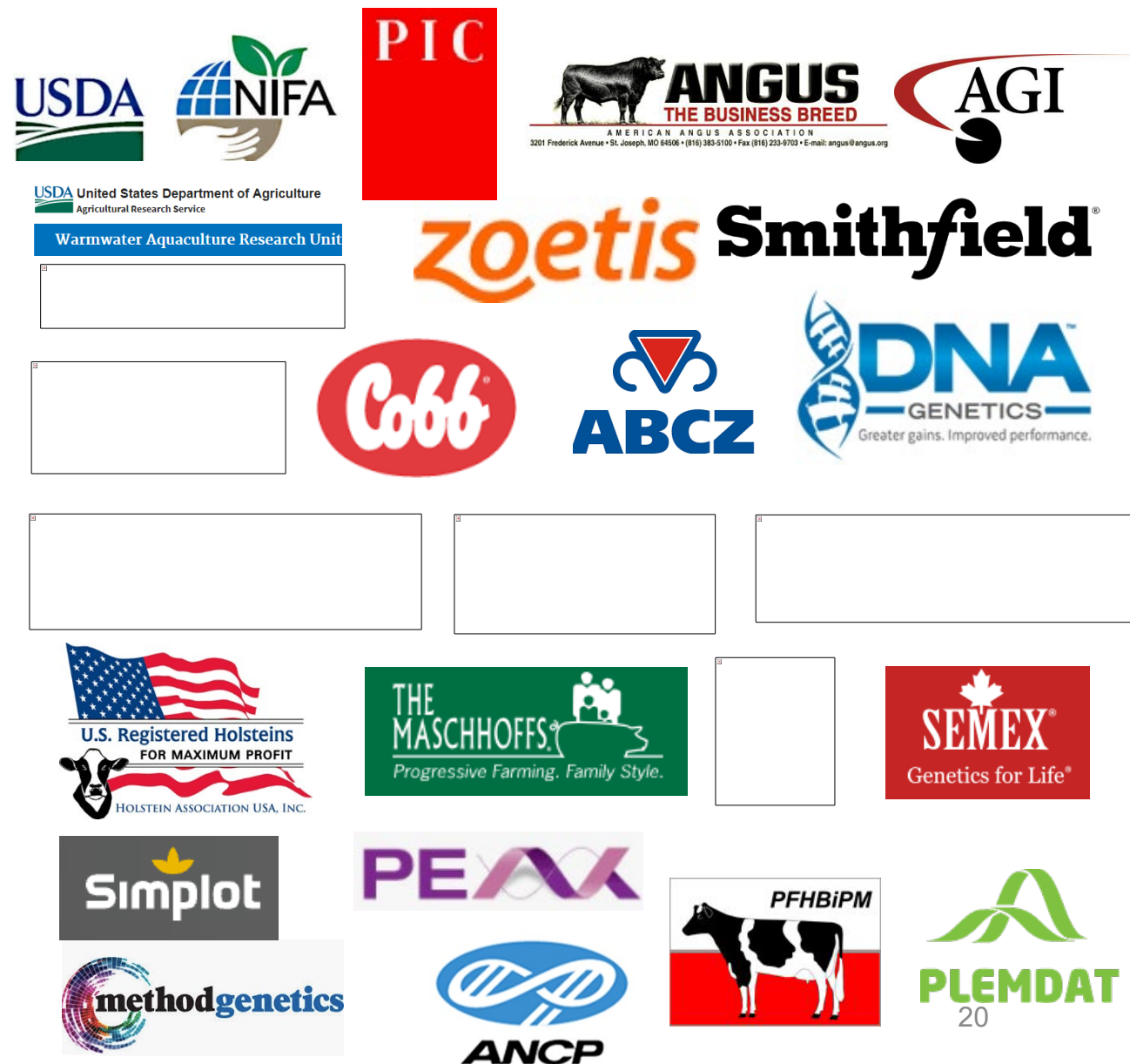
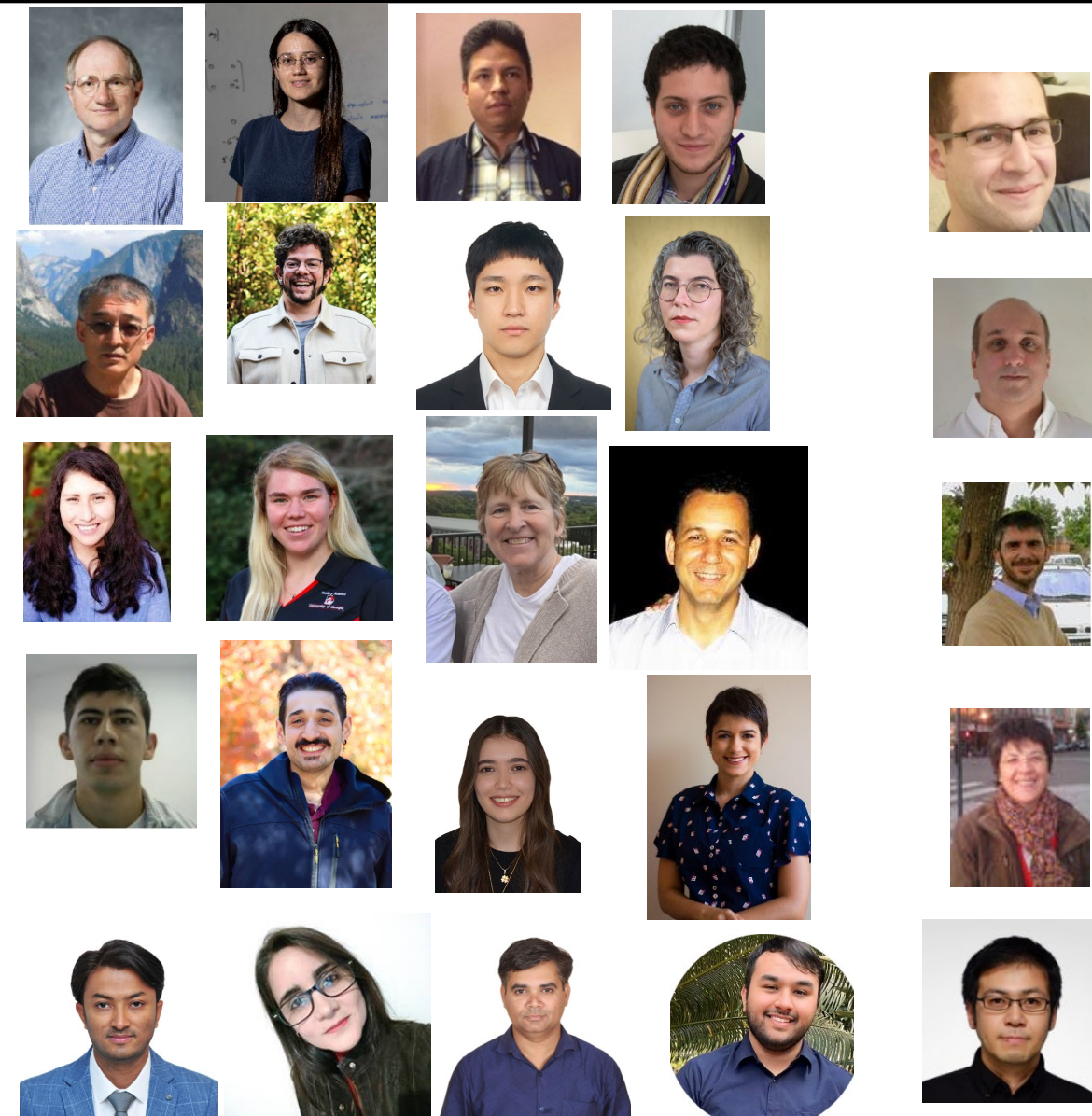
Cross-validation statistics for trait: 1

	Value	SD	l_95%_CI	h_95%_CI
Absolute bias:	-0.047	0.008	-0.064	-0.031
Bias in genetic sd:	-0.009	0.002	-0.012	-0.006
Dispersion:	1.003	0.003	0.996	1.009
Ratio of accuracies:	0.748	0.002	0.745	0.751
Reliability :	0.237	0.002	0.234	0.240
Predictive ability:	0.559	0.005	0.550	0.568

Conclusions

- Frequentist CI are better than bootstrap CI
- Approximated CI are like bootstrap CI
- Bootstrap seems to not account for (co)variance structure
- Formulas are complicated to apply
- validationf90 includes all possible ways for proper CI

UGA AB&G team



Questions?

mbermann@uga.edu

