

How to estimate genetic parameters in genomic era

Ignacy Misztal



UNIVERSITY OF
GEORGIA

Why parameter estimation

- Input for genetic evaluation
- Estimate possible genetic progress
- Calculate correlated responses
- Troubleshoot data recording and models
- See changes of parameters over time

Parameters and models

- Usually estimates of additive variance + environmental variances
- Estimation of dominance and epistatic variances requires large data - nonadditive effects small and hard to use
- Real data
 - Additive model good approximation in short time
 - Epistatic changes over time
 - Change of trait definitions due to improved management
- Simulated data – ok for few generations but useless over many

Old times for parameter estimation

- Pick a subset
- Estimate variance components
- Use a fact that mixed-model equations sparse

- AIREML
 - 100k-1M animals
 - Cubic costs with multiple traits, crashes if too many
 - Often fast convergence to final value

- Gibbs sampler
 - Over 1 M animals
 - Linear costs with traits
 - Not sure when to finish samples



[Full Access](#)

Reliable computing in estimation of variance components

[Correction\(s\) for this article](#) 

I. Misztal

What changed with genomic selection?

- Possibly lower standard errors compared to nongenomic analyses
 - Forni et al. (2011); Veerkamp et al., 2011
- Biases due to genomic preselection
 - Evaluation unbiased if all data used for selection included in analyses (Henderson, 1984).
- Long or impossible computing
- Inflation/deflation if incorrect scale of genomic information
- Accelerated “aging” of the additive model

Computing issues

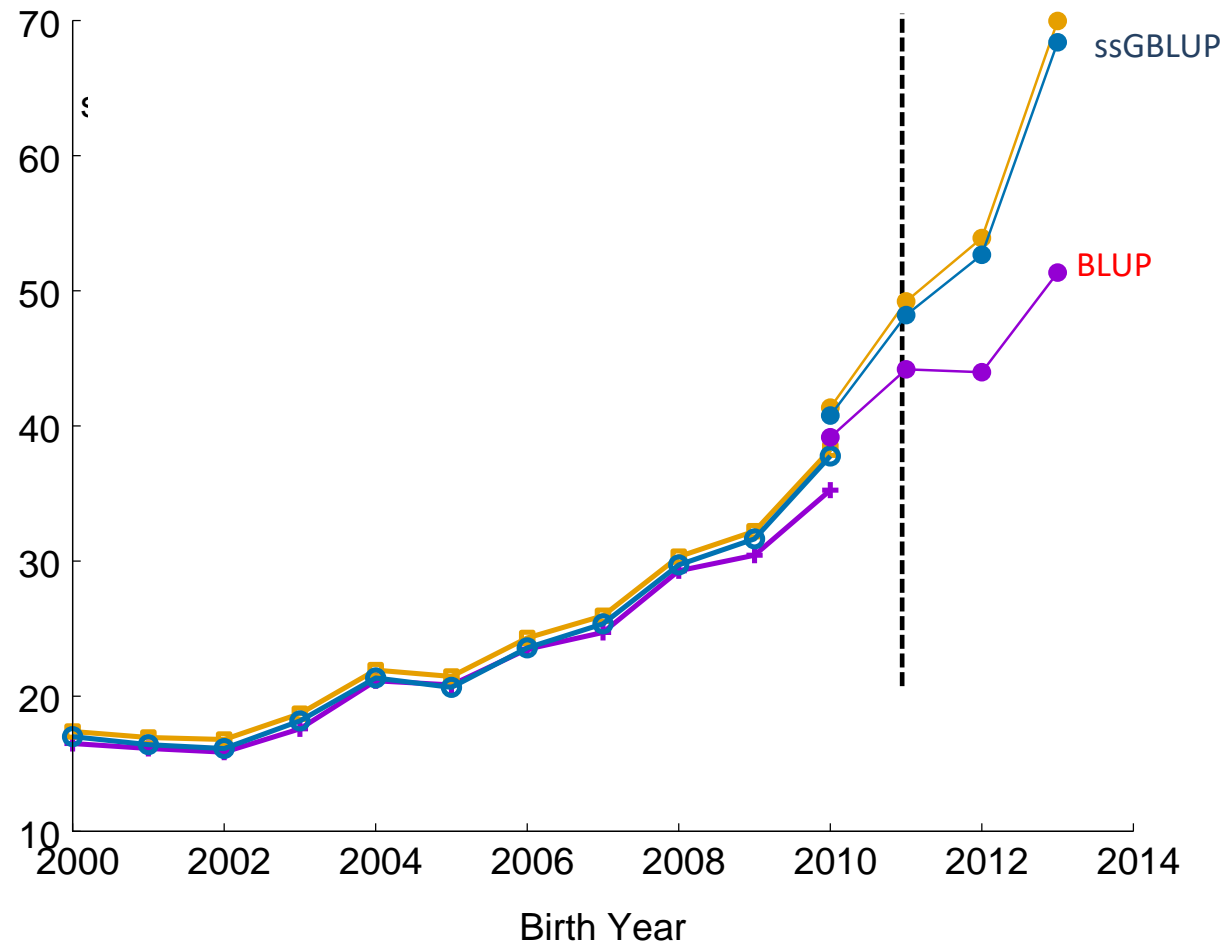
- Mixed models stored as sparse matrix
 - If 4 byte integer, number of elements $< 2^{31} = 2 \text{ G}$ elements
- Equations no longer sparse – dense blocks
 - Limit of about 50k genotypes for 1 trait, 30k for two traits
- REML: Sparse matrix factorization/inversion inefficient with dense blocks
- YAMS (Masuda et al., 2015) – sparse matrix package that recognizes dense blocks
 - Up to 100 time faster than older FSPAK
 - About 10 hrs (AIREMLF90) for 200k animal (15k genotyped) 4 trait model

Options for parameter estimation with genomics

- Ignore genomic information
- Use only phenotypes of genotyped animals
 - SNP and GBLUP models
- Use information from genotyped and ungenotyped animals (single-step methods)

- Data reduction
 - fewer generations
 - fewer genotypes

Genetic trend for genotyped bulls (Masuda et al., 2019)



When parameter estimation biased? Simulation by Cesarani et al. (2018)

- 3 of 3k animals each
- 30% animals in each generation genotyped
- REML, GREML (genotyped animals only) and ssGREML
- Selection or random mating
- Random or best animals genotyped
- 1 to 3 generations of data

ORIGINAL ARTICLE

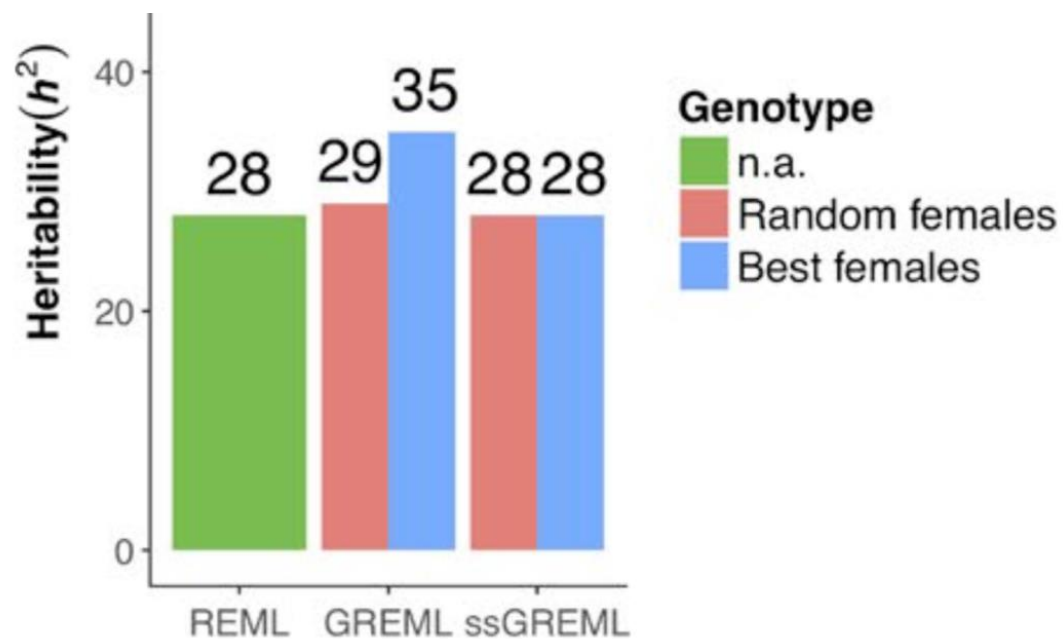
WILEY  Animal Breeding and Genetics

Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies

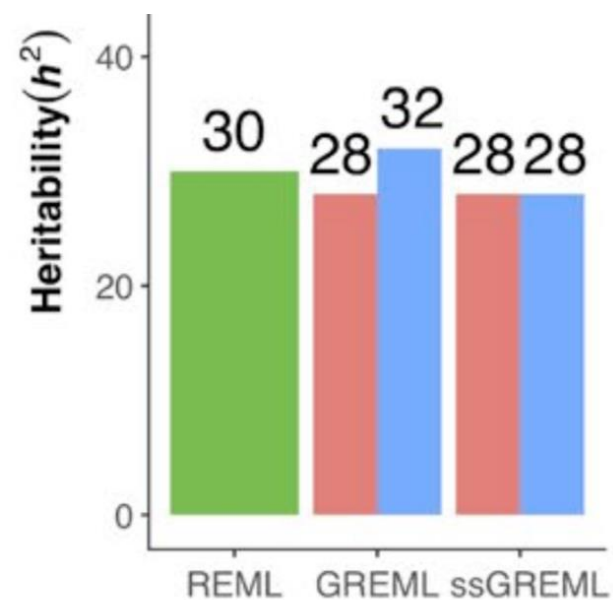
Alberto Cesarani^{1,2}  | Ivan Pocrnic¹  | Nicolò P. P. Macciotta²  |
Breno O. Fragomeni³  | Ignacy Misztal¹  | Daniela A. L. Lourenco¹ 

No selection ($h^2 = 30\%$)

One generation of data



Three generations of data

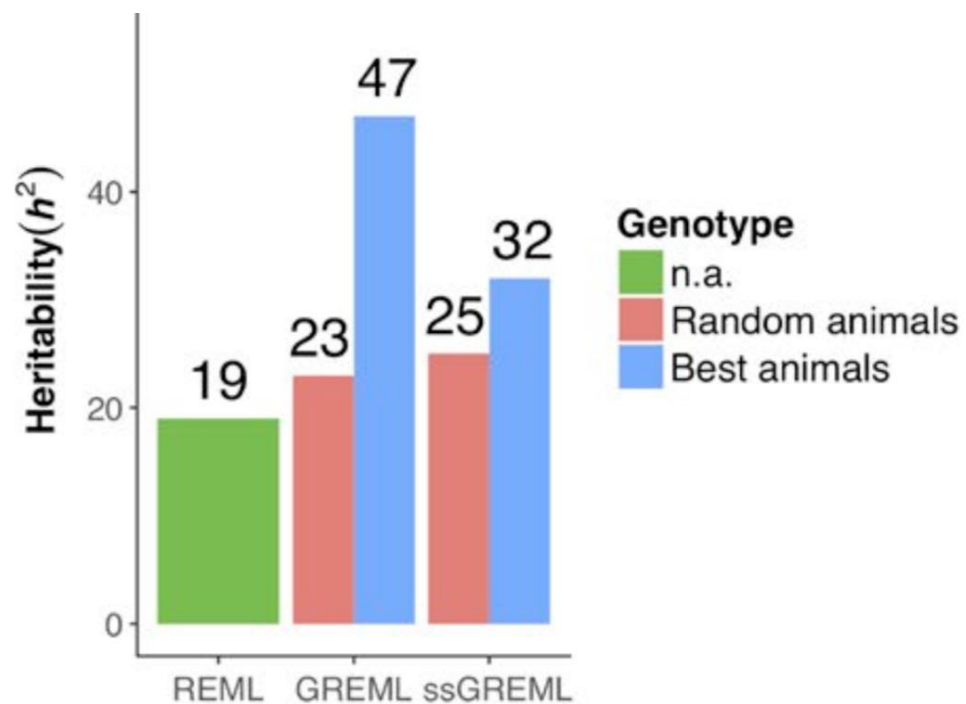


Mostly unbiased estimates without selection

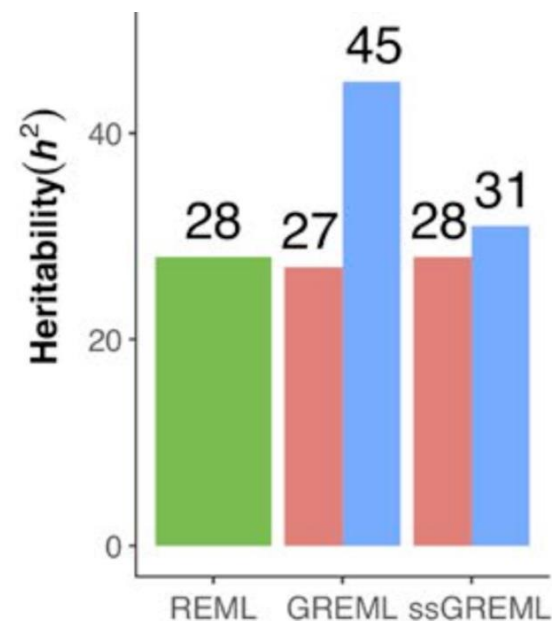
Some inflation with small data in GREML and selective genotyping

EBV selection ($h^2=30\%$)

One generation of data



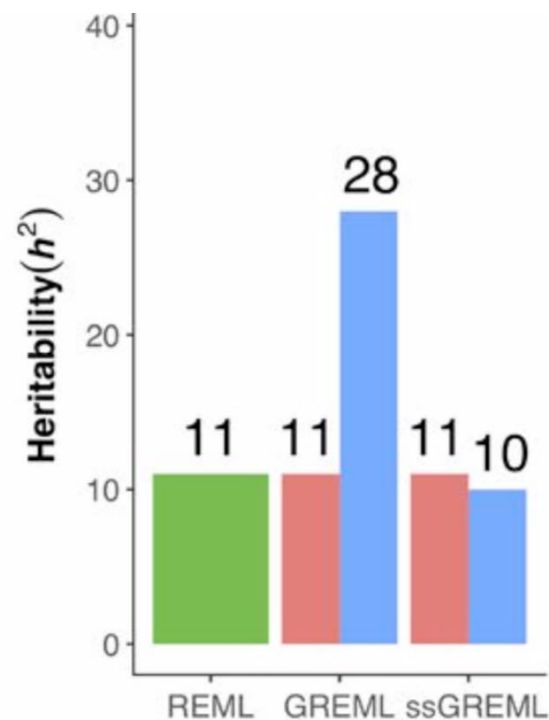
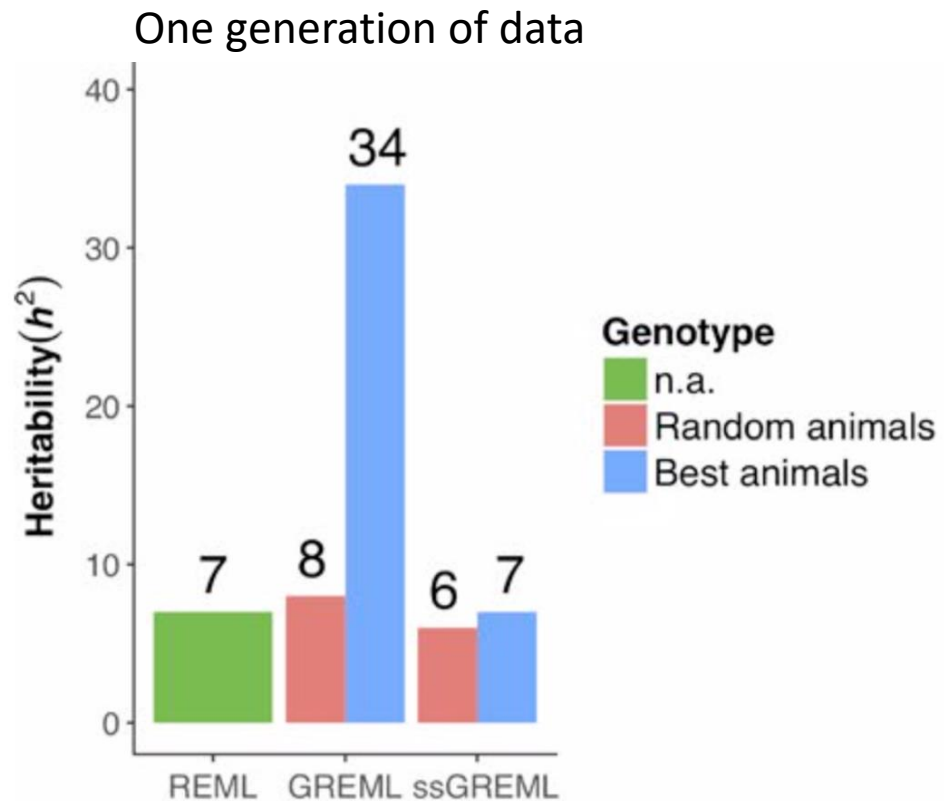
Three generations of data



Heritability biased down in REML with small data
Heritability inflated in GREML and selective genotyping

Selection ($h^2 = 10\%$)

Three generations of data



Greatly inflated estimates in GREML with selective genotyping

Genetic variances in ssGREML and REML with selective genotyping (Wang et al., 2020)

Analysis model	Analysis genotyping strategy and proportion (%)		$\hat{\sigma}_a^2$ (SE)	$\hat{\sigma}_e^2$ (SE)
H-AM ssGREML	Selective	10	13,358 (635) ^{ab}	22,358 (324) ^{ab}
		20	40,597 (729) ^{ab}	11,904 (243) ^{ab}
		30	55,051 (639) ^{ab}	8695 (136) ^{ab}
	Random	10	9265 (469)	24,231 (270) ^{ab}
		20	8967 (440)	24,382 (243)
		30	8873 (421)	24,461 (223)
A-AM GREML			11,475 (544) ^{ab}	23,148 (217) ⁵⁶

Wang et al. *Genet Sel Evol* (2020) 52:31
<https://doi.org/10.1186/s12711-020-00550-w>



RESEARCH ARTICLE

Open Access

Effect of genomic selection and genotyping strategy on estimation of variance components in animal models using different relationship matrices

Lei Wang^{1*}, Luc L. Janss¹, Per Madsen¹, John Henshall², Chiyong-Huoy Huang², Danye Marois², Setegn Alemu¹, AC Sørensen¹ and Just Jensen¹



Predictive ability for broilers (Wang et al., 2020)

	Model	Male	
		Genotyped	Ungenotyped
BLUP	A-AM	0.15	0.21
ssGBLUP	H-AM	0.31	0.22
ssGBLUP correct parameters	H-AM (VC-A)	0.40	0.42

Compatibility of relationships in single-step analyses

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

1-ungenotyped animals
2-genotyped animals

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

Scales different

$$\mathbf{G} - \mathbf{A}_{22}$$

Inflation/ deflation

Levels different

$$\mathbf{G} - \mathbf{A}_{22}$$

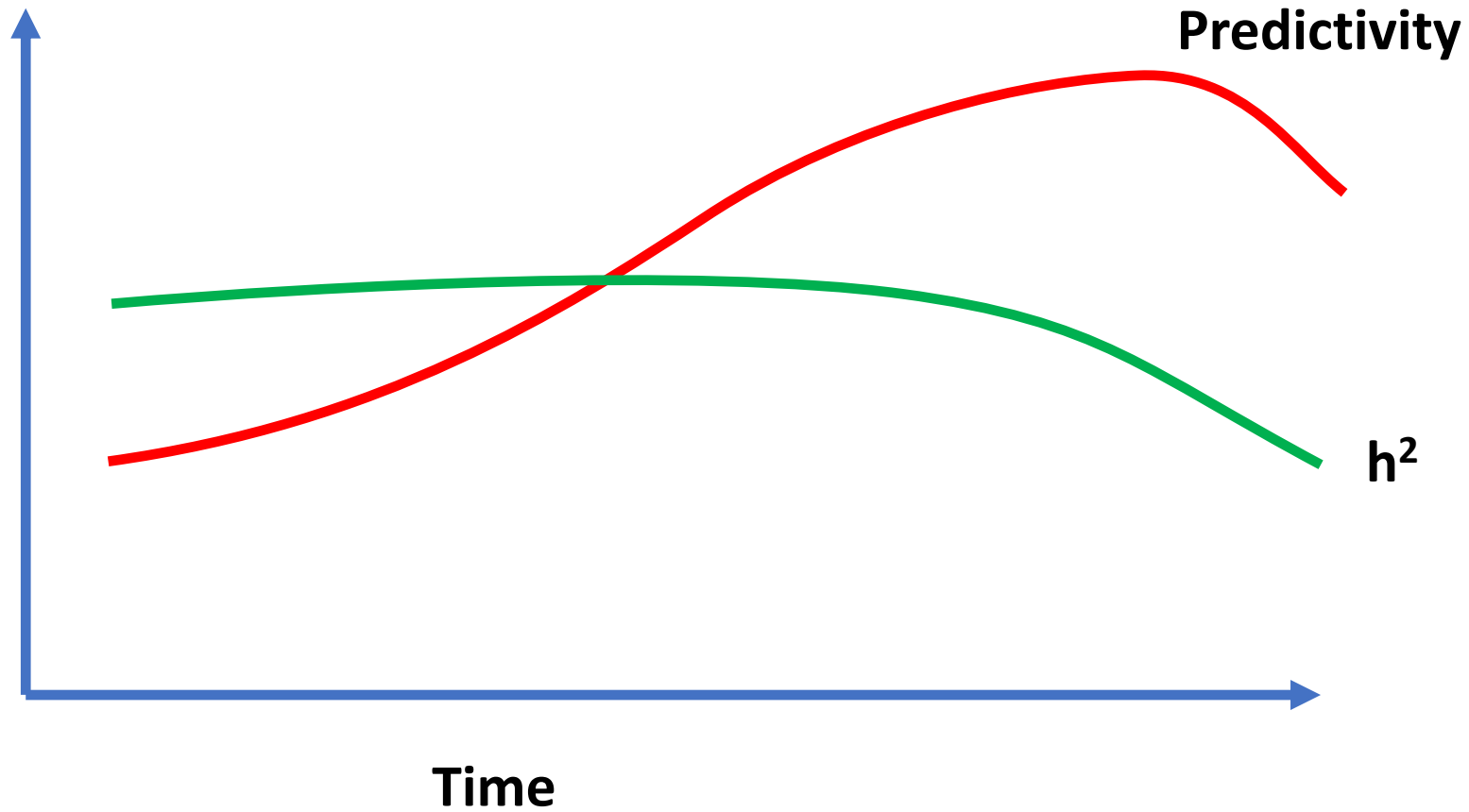
Bias

Incomplete pedigree

$$\mathbf{G} - \mathbf{A}_{22}$$

Divergence, inflation

Predictivity over time in pigs



Hollifield et al. (2020)

How to estimate changes over time

- Random regression model on year of birth (Tsuruta et al., 2004)
 - Need to use all relevant genotypes
 - Multiple animal effects
 - Very expensive
- Computing by data slices
 - Small slices - less computing, more bias

Estimating changing parameters in pigs (Hidalgo et al., 2019)

- Growth ($h^2=0.3$) and fertility ($h^2=0.07$)
- 40k genotypes over 10 years
- 400k animals

- Random regression model ran for months

- Data slices for 3 years
 - Elimination of noncontributing genotypes
 - gibbs3f90

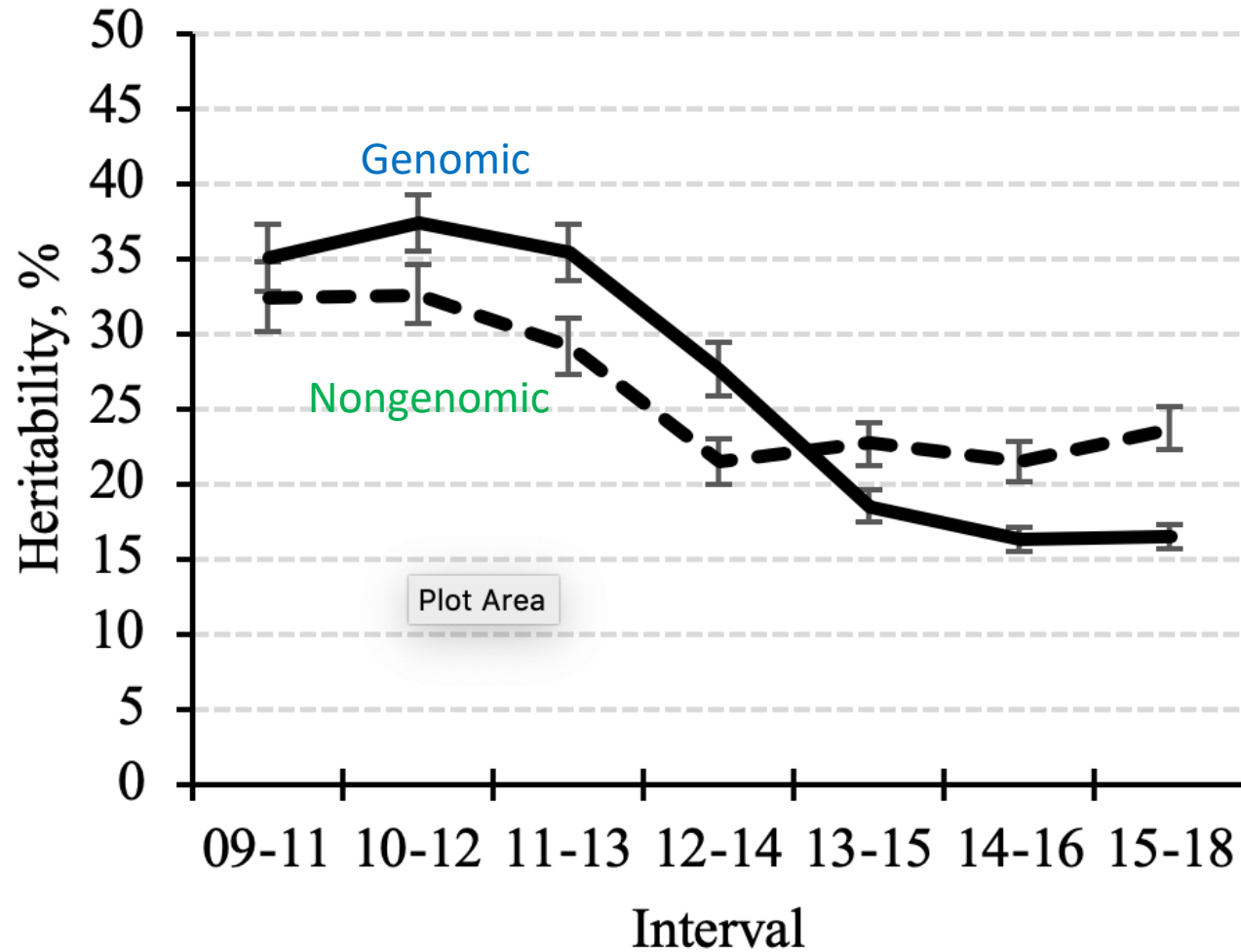
ANIMAL GENETICS AND GENOMICS

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

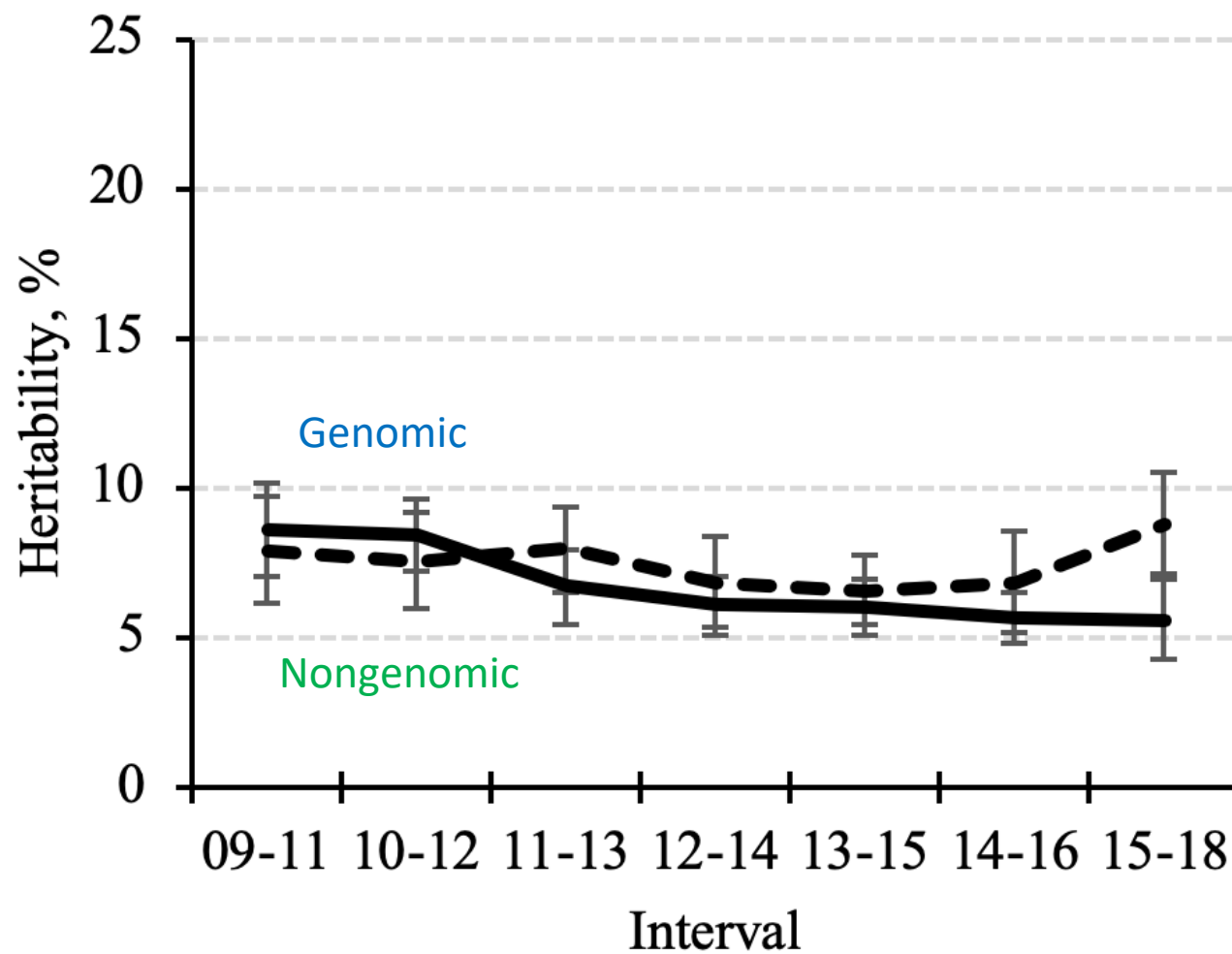
Jorge Hidalgo,^{†,1} Shogo Tsuruta,[†] Daniela Lourenco,[†] Yutaka Masuda,[†] Yijian Huang,[‡] Kent A. Gray,[‡] and Ignacy Misztal[†]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, [‡]Smithfield Premium Pigs, Smithfield, NC 27870

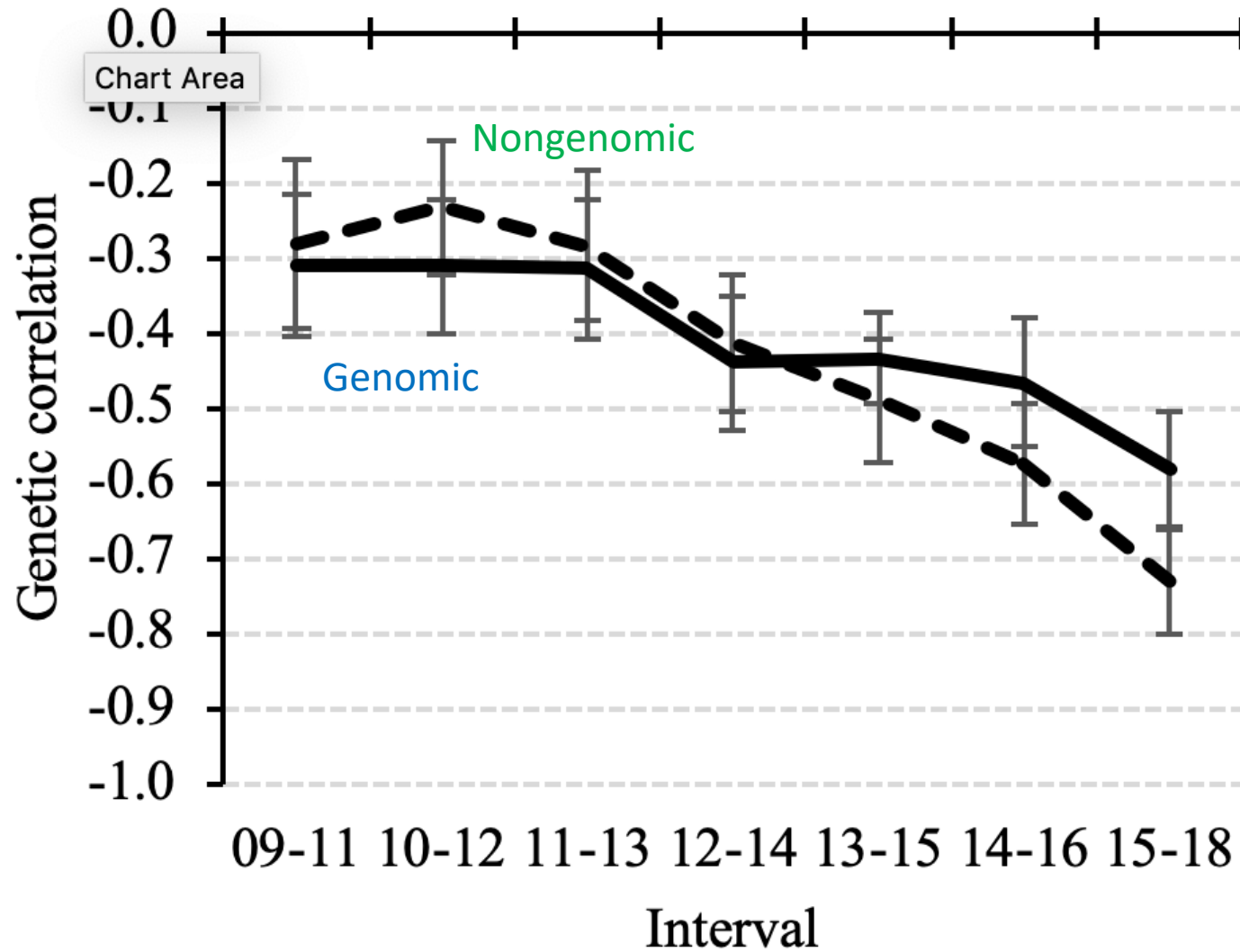
Heritability for growth



Heritability for fertility



Genetic correlations



Packages for parameter estimation

- General based on relationships – extensive modeling capabilities
 - BLUPF90
 - DMU
 - Wombat etc.
 - ASREML
- Packages estimating SNP effects – usually limited modeling
 - GVCBLUP (UMN)
 - GS3 (INRA)
 - GENSEL (IASTATE)
 - Golden Helix
 -

Other options with large data sets

- No limits
 - SNP model (50k)
 - GBLUP with inversion of G by APY algorithm
 - Perhaps ssGBLUP with APY and pedigree truncation
- Potential other methods
 - REML Monte Carlo (Matialainen, 2014)
 - Method R

Recommendation for parameter estimation

- Any methods works if unselected data
 - Genomic info reduces SE
- If selection, use at least 2-3 generations
- REML may be OK with many generations
- Costs with ssGBLUP reduced if data pruned
 - Phenotypes for 2-3 generations
 - Pedigrees 1-2 generations behind phenotypes
 - Only contributing genotypes
- Beware of very low and very high estimates – data selection or poor model

Conclusions

- Parameter estimates biased by preselection if genotypes not considered
- Estimation with genomic data very expensive
- Correct modeling important
- With genomic selection accelerated changes of parameters over time

Acknowledgements

