Single-step GWAS with p-values for large genotyped populations

Natalia Leite, Matias Bermann, Shogo Tsuruta,
Ignacy Misztal, **Daniela Lourenco**

June 28, 2023

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

ADSA Annual Meeting
June 25-28, 2023
Ottawa, Ontario, Canada

Equivalence ssGBLUP – ssSNPBLUP

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

Aguilar et al. (2010)

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X}_g'\mathbf{Z}_g\mathbf{M}_g & \mathbf{X}_n'\mathbf{Z}_n \\ \mathbf{M}_g'\mathbf{Z}_g'\mathbf{X}_g & \mathbf{Q} & \mathbf{M}_g'\mathbf{A}^{gn}\frac{\sigma_e^2}{\sigma_g^2} \\ \mathbf{Z}_n'\mathbf{X}_n & \mathbf{A}^{ng}\mathbf{M}_g\frac{\sigma_e^2}{\sigma_g^2} & \mathbf{Z}_n'\mathbf{Z}_n + \mathbf{A}^{nn}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\mathbf{u}}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{M}_g'\mathbf{Z}_g'\mathbf{y}_g \\ \mathbf{Z}_n'\mathbf{y}_n \end{bmatrix}$$

Fernando et al. (2016)

- Equivalent models under same assumptions and data
  - $\hat{\mathbf{u}} = \mathbf{Z}\hat{\boldsymbol{a}}$
  - $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$
  - $\mathrm{Var}(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}'\mathbf{G}^{-1}(\mathbf{G} - \mathbf{C}^{\mathbf{u_2u_2}})\mathbf{G}^{-1}\mathbf{Z}k$

VanRaden (2008)
Stranden and Garrick (2009)

Guladron-Duarte et al. (2014)

- $\text{p-value}_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$ → ssGWAS


Aguilar et al. Genet Sel Evol (2019) 51:28
https://doi.org/10.1186/s12711-019-0469-3

GSE Genetics Selection Evolution

SHORT COMMUNICATION Open Access

Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle

Ignacio Aguilar[1], Andres Legarra[2*], Fernando Cardoso[3,4], Yutaka Masuda[5], Daniela Lourenco[5] and Ignacy Misztal[5]

Equivalence enables ssGWAS

- Why ssGWAS?
- Assumption of single-marker GWAS: Genotyped individuals have phenotypes
- Animal populations: genotypes and phenotypes may not be on the same individuals
  - Deregressed EBV: are biased

- ssGWAS
  - All data on genotyped and non-genotyped individuals
  - Multi-trait models to accommodate correlations

- Negative aspect of ssGWAS
  - Heavy computations for p-values -> same limitation as REML

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

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

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

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

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

Masuda et al. (2016)

Equivalence APY ssGBLUP – ssSNPBLUP

JOURNAL ARTICLE

Indirect predictions with a large number of genotyped animals using the algorithm for proven and young

Andre L S Garcia, Yutaka Masuda, Shogo Tsuruta, Stephen Miller, Ignacy Misztal, Daniela Lourenco

Journal of Animal Science, Volume 98, Issue 6, June 2020, skaa154,
https://doi.org/10.1093/jas/skaa154

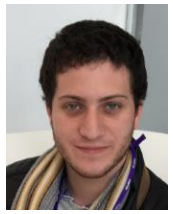
Garcia *et al. Genetics Selection Evolution* (2022) 54:66
https://doi.org/10.1186/s12711-022-00752-4

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Theoretical accuracy for indirect predictions based on SNP effects from single-step GBLUP

Andre Garcia[1*], Ignacio Aguilar[2], Andres Legarra[3], Shogo Tsuruta[1], Ignacy Misztal[1] and Daniela Lourenco[1]



Bermann *et al. Genetics Selection Evolution* (2022) 54:52
https://doi.org/10.1186/s12711-022-00741-7

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young

Matias Bermann[1*], Daniela Lourenco[1], Natalia S. Forneris[2,3], Andres Legarra[4] and Ignacy Misztal[1]

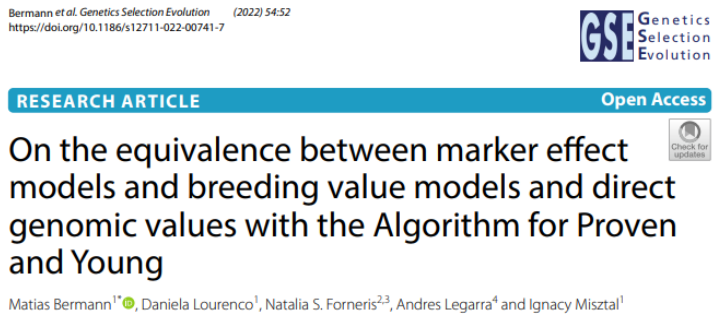
- If using APY in ssGBLUP
  - Numerical equivalence
  - $\hat{\mathbf{u}} = \mathbf{Z}\hat{\boldsymbol{a}}$
  - $\hat{\boldsymbol{a}}|\hat{\mathbf{u}} = k\mathbf{Z}'\mathbf{G}_{\text{APY}}^{-1}\hat{\mathbf{u}}$
  - $\text{Var}(\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}) = k\mathbf{Z}'\mathbf{G}_{\text{APY}}^{-1}(\mathbf{G} - \mathbf{C}^{\mathbf{u_2u_2}})\mathbf{G}_{\text{APY}}^{-1}\mathbf{Z}k$

↳ Function of all genotyped animals

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

↳ Function of CORE animals

Equivalence APY ssGBLUP – ssSNPBLUP


Bermann et al. Genetics Selection Evolution (2022) 54:52
https://doi.org/10.1186/s12711-022-00741-7

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young

Matias Bermann[1*], Daniela Lourenco[1], Natalia S. Forneris[2,3], Andres Legarra[4] and Ignacy Misztal[1]

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

?

- Exact - Inverse of the LHS of MME
- Approximating reliabilities of GEBV

GEBV are published with reliability

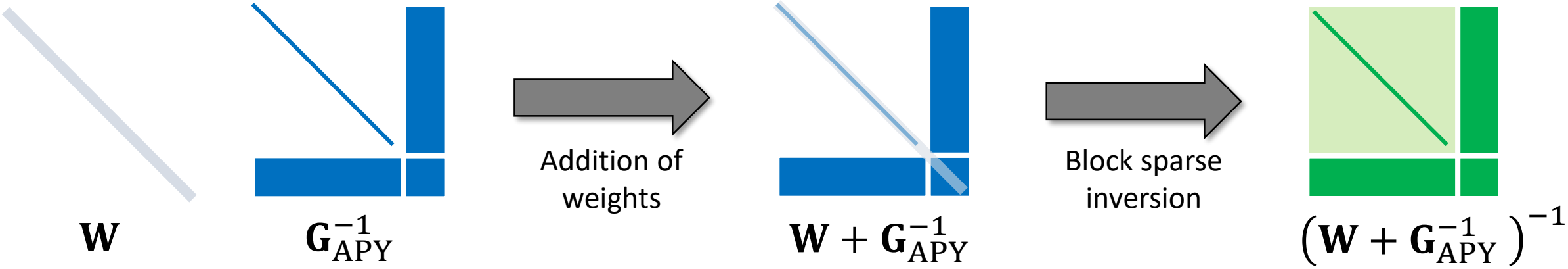
- Reliability based on PEV
  - Approximated for large populations
  - Weights based on approximations
  - Block sparse inversion with APY

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



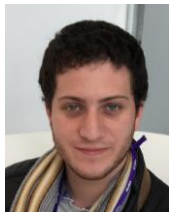
$\mathbf{W}$ $\quad \mathbf{G}_{\text{APY}}^{-1}$ $\quad \mathbf{W} + \mathbf{G}_{\text{APY}}^{-1}$ $\quad \left(\mathbf{W} + \mathbf{G}_{\text{APY}}^{-1}\right)^{-1}$

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

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

Single-step GWAS – Many genotypes

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

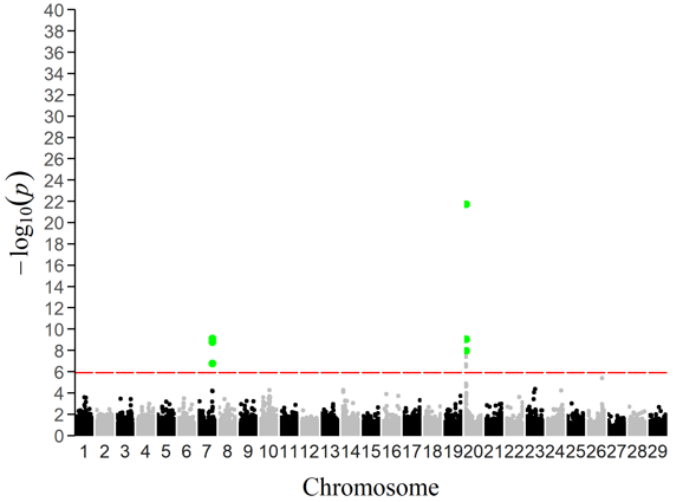
Leite, Bermann et al. (in preparation)

$$p\text{-}value_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$$

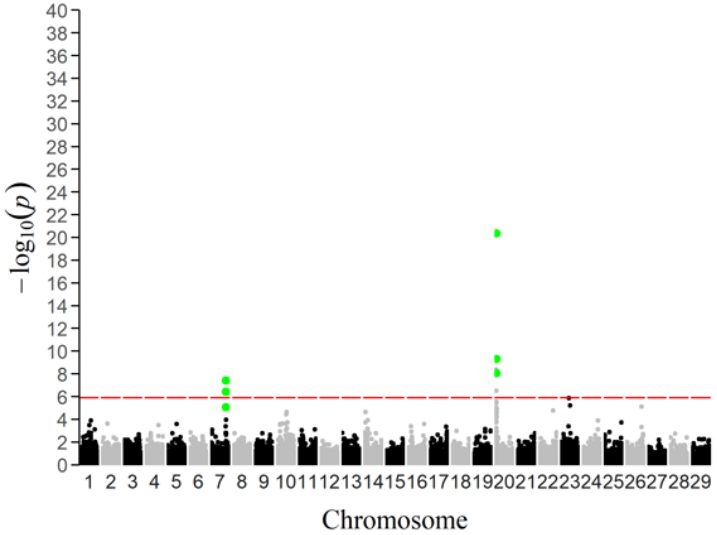
- Initial tests – AGI data
  - 845k phenotypes for post-weaning gain
  - 50k genotyped + 1.58M pedigree
  - 450k genotyped (13k core + 437k noncore) + 1.8M pedigree

Single-step GWAS – Scenarios

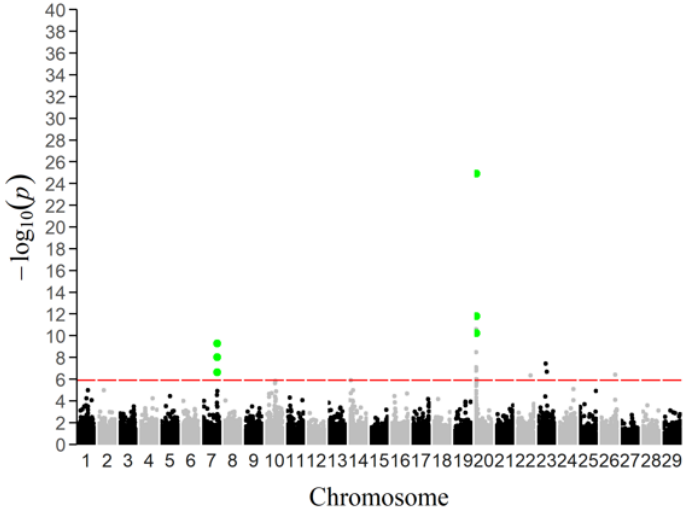
Exact - $\mathbf{G}^{-1}$ 50k



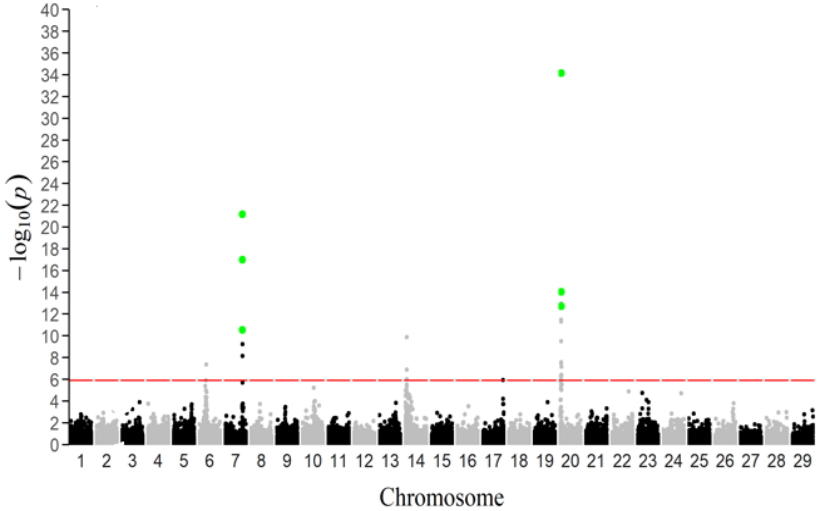
Exact - $\mathbf{G}^{-1}_{\mathrm{APY}}$ 50k



Approximation - $\mathbf{G}^{-1}_{\mathrm{APY}}$ 50k



Approximation - $\mathbf{G}^{-1}_{\mathrm{APY}}$ 450k



Single-step GWAS – Computations

Method	Elapsed time, h:min*	Peak memory, GB*
Exact - $\mathbf{G}^{-1}$ 50k	106:46	159.66
Exact - $\mathbf{G}_{\mathrm{APY}}^{-1}$ 50k	110:59	178.30
Approx - $\mathbf{G}_{\mathrm{APY}}^{-1}$ 50k	2:50	16.62
Approx - $\mathbf{G}_{\mathrm{APY}}^{-1}$ 450k	24:28	87.64

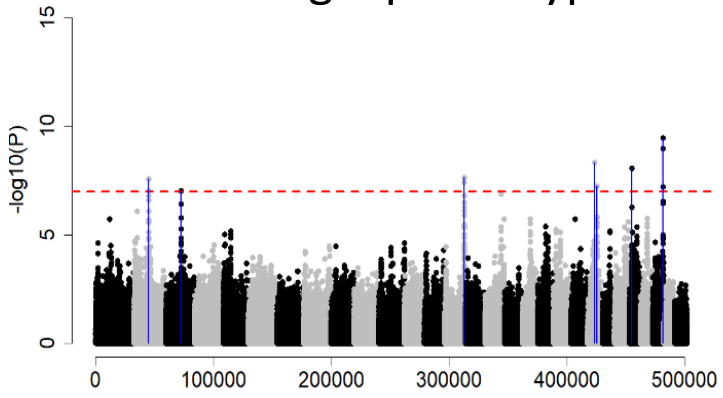
- Computing cost still high
- Eliminates the limitation on the amount of data for ssGWAS

Adding more data in GWAS

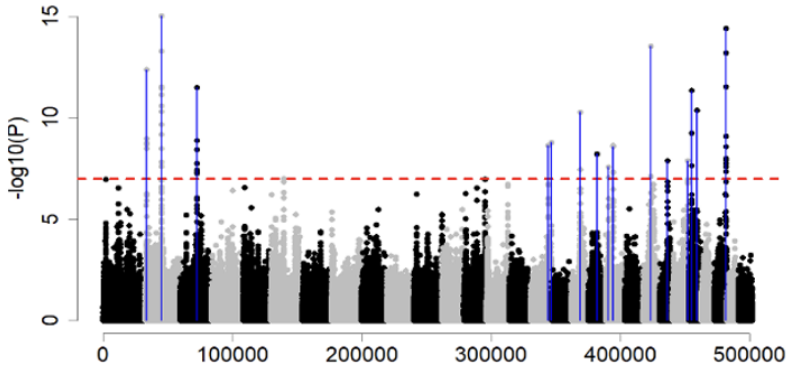
- Adding more genotypes and phenotypes
  - Less noise
  - Much higher resolution

Jang et al. (accepted)

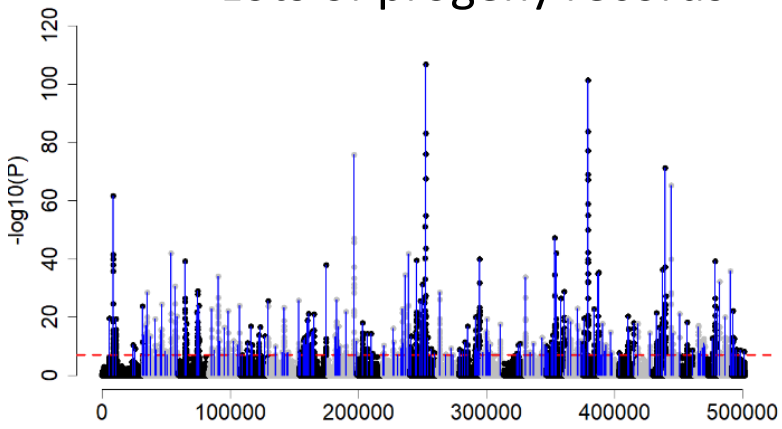
15,800 genotyped animals
+
Single phenotypes



30,000 genotyped animals
+
Single phenotypes



30,000 genotyped animals
+
Lots of progeny records



Take home messages

- ssGWAS allows for using all data
  - Phenotypes
  - Pedigree
  - Genotypes

- Virtually any number of genotyped animals
  - Improvement in computing time

- Possible because of the limited dimensionality of genomic information
  - Depends on the quality of approximated reliabilities
  - Already implemented in BLUPF90

UNIVERSITY OF GEORGIA

UGA AB&G team

USDA

NIFA

ANGUS THE BUSINESS BREED
AMERICAN ANGUS ASSOCIATION
3201 Frederick Avenue • St. Joseph, MO 64506 • (816) 383-5100 • Fax (816) 233-9703 • E-mail: angus@angus.org

AGI

PIC

USDA United States Department of Agriculture
Agricultural Research Service
Warmwater Aquaculture Research Unit

zoetis

Smithfield

USDA Agricultural Research Service
U.S. DEPARTMENT OF AGRICULTURE
Cool and Cold Water Aquaculture Research

Cobb

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

DNA GENETICS
Greater gains. Improved performance.

ABCZ

Rural Development Administration
National Institute of Animal Science

WAGYU
AMERICAN WAGYU ASSOCIATION

BAYER Crop Science

U.S. Registered Holsteins
FOR MAXIMUM PROFIT
HOLSTEIN ASSOCIATION USA, INC.

THE MASCHHOFFS
Progressive Farming. Family Style.

ABS

SEMEX
Genetics for Life

PEAK

PFHBiPM

PLEMDAT

Topigs Norsvin

methodgenetics

ANCP