

Large-Scale Genomic Predictions with Sequence Data

Daniela Lourenco
Shogo Tsuruta, Ignacy Misztal
January 15, 2023

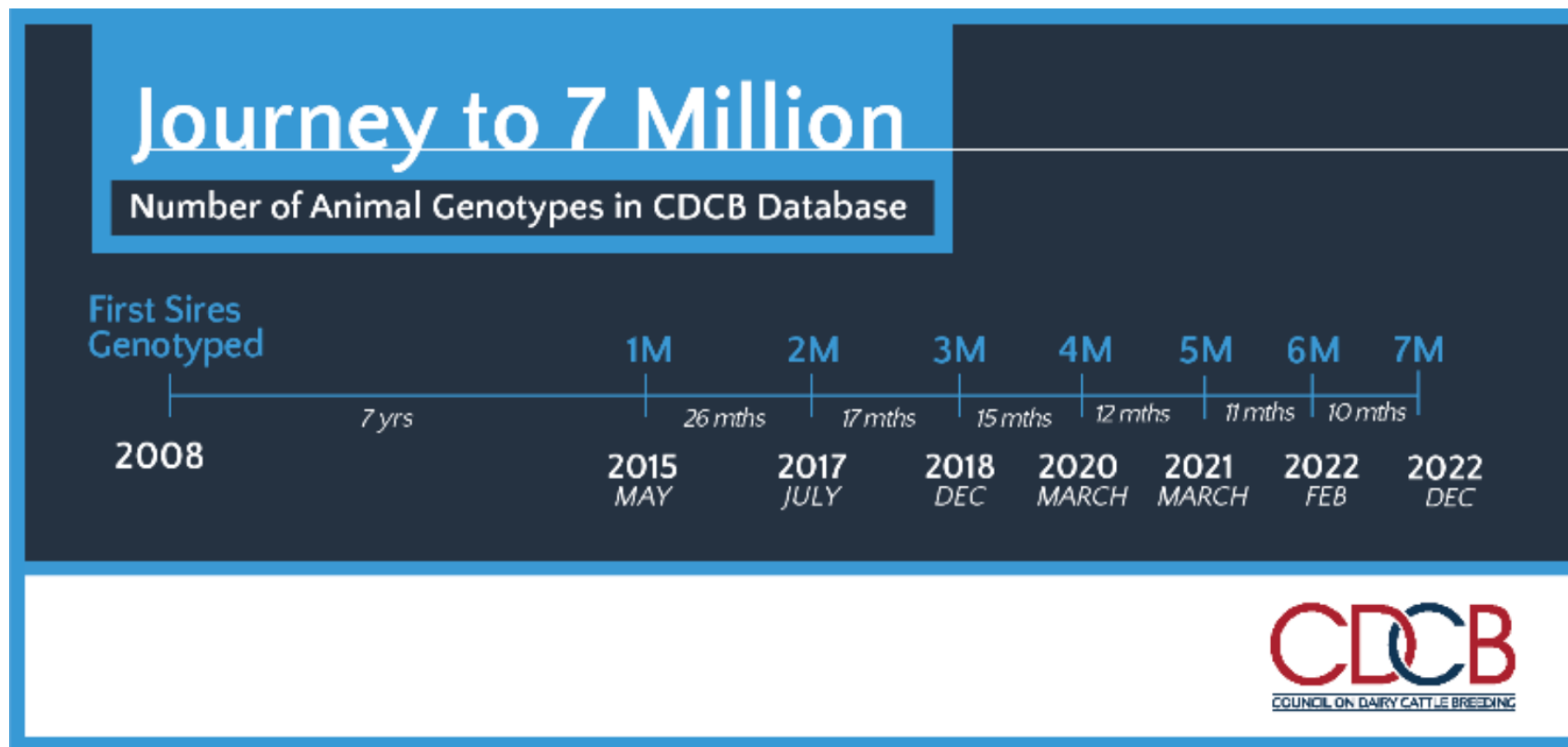


**UNIVERSITY OF
GEORGIA**

**College of Agricultural &
Environmental Sciences**



Massive uptake of genomics in 15 years





Alberto Cesarani



J. Dairy Sci. 105:5141–5152

<https://doi.org/10.3168/jds.2021-21505>

© 2022, The Authors. Published by Elsevier Inc. and Fass Inc. on behalf of the American Dairy Science Association®.
This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor

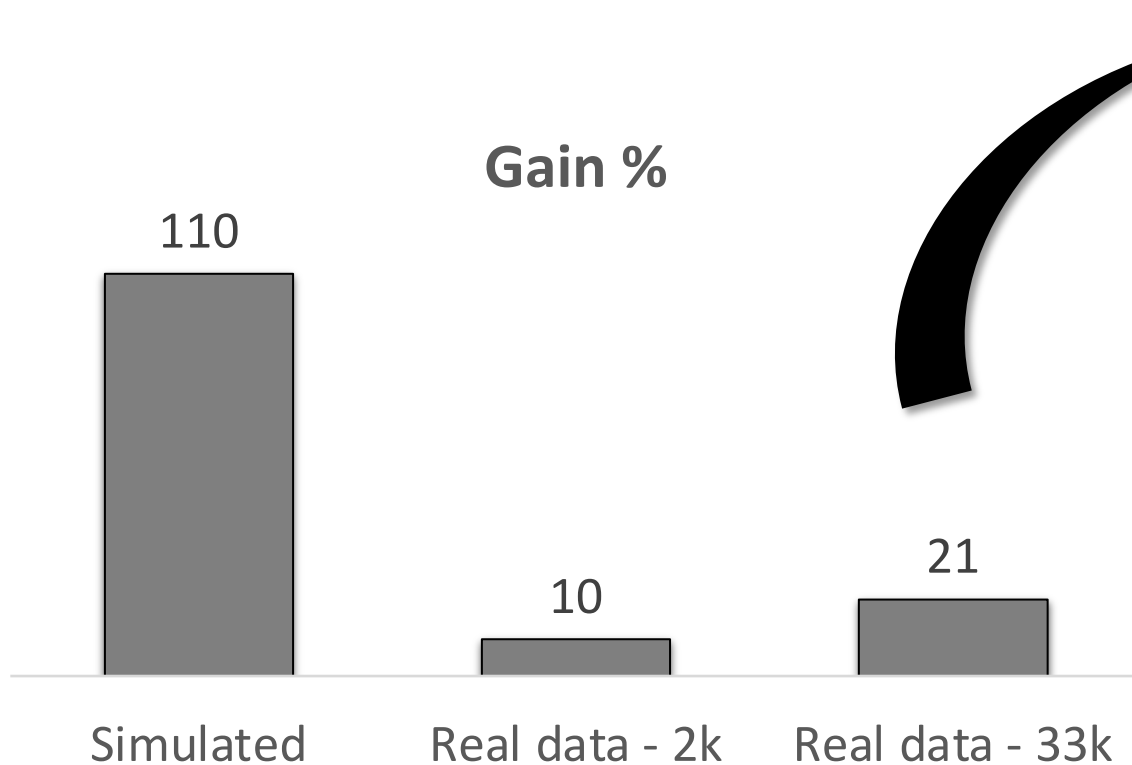
A. Cesarani,^{1*}  D. Lourenco,¹  S. Tsuruta,¹  A. Legarra,²  E. L. Nicolazzi,³  P. M. VanRaden,⁴ 
and I. Misztal¹ 

Why do we care about lots of data?

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

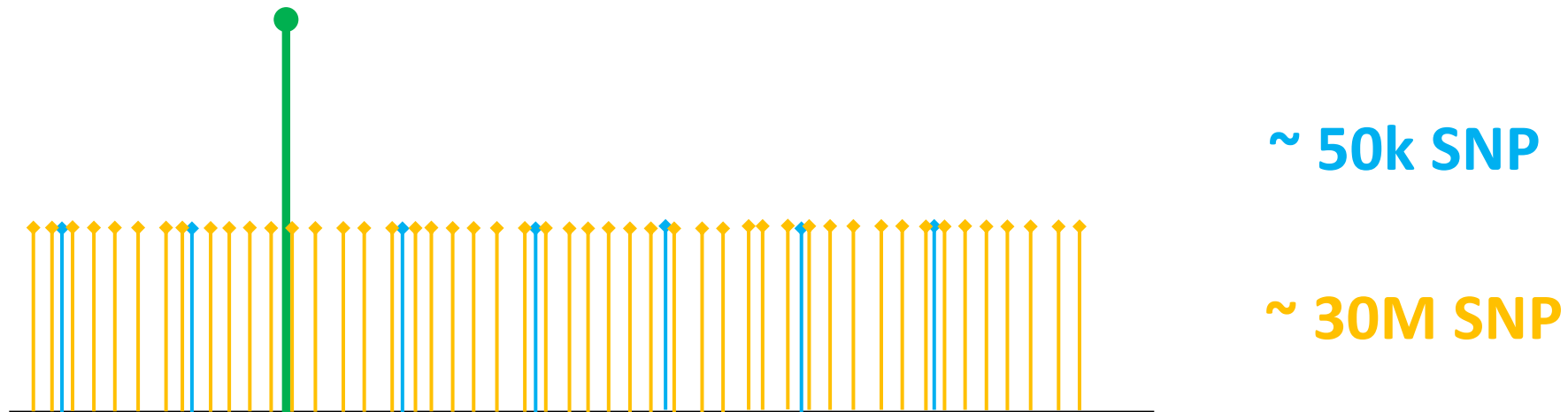
- Predicting things is very hard
 - Gather enough data
 - Use the right statistical tools

Gain with genomics



- 50k SNP may not be enough
- We should use sequence data

The idea behind sequence data



Sequencing is becoming affordable



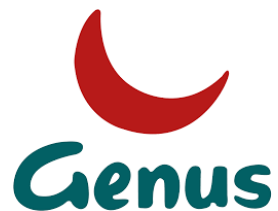
The screenshot shows the Ferrari Silicon Valley website. At the top, there are logos for Ferrari and Maserati, the company name "Ferrari Silicon Valley", and the address "2750 El Camino Real Redwood City, CA 94061". Contact information includes Sales: (888) 378-7586, Service: (888) 377-1063, Parts & Accessories: (888) 430-4670, and Body Shop: (866) 981-0953. A search bar is located in the top right corner. Below the header is a navigation menu with buttons for Inventory, Service, Parts & Accessories, Body Shop, News & Events, Racing Team, and About Us. The main content area features a "2013 Ferrari 458 Spider" listing. The car is shown in a showroom setting. To the right of the car image, there is a "Ferrari Videos" button and a "Share" button with social media icons. The price "\$398,000" is displayed in large black text, and a red "\$0.01" is overlaid on the bottom right of the price.

... and fast



The screenshot shows the Ferrari Silicon Valley website. The header includes the Ferrari and Maserati logos, the company name "Ferrari Silicon Valley", and contact information: "2750 El Camino Real Redwood City, CA 94061", "Sales: (888) 378-7586", "Service: (888) 377-1063", "Parts & Accessories: (888) 430-4670", and "Body Shop: (866) 981-0953". A search bar is located in the top right. The main navigation menu includes "Inventory", "Service", "Parts & Accessories", "Body Shop", "News & Events", "Racing Team", and "About Us". The featured car is a "2013 Ferrari 458 Spider". To the right of the car image is a "Ferrari Videos" button and a "Share" button with social media icons. The text next to the car reads: "Top speed: ~~199mph~~", "New top speed: 32,000,000 mph", and "New comparator vehicle required". Below this text is an image of the Star Trek Enterprise ship flying through a blue warp speed tunnel. A footnote at the bottom of the image reads: "* Nerd footnote: 14 million meters/sec is 1/20th warp speed (~the speed of light)".

Largest pig sequence data



Line	Genotyped individuals	Sequenced individuals	Sequenced/Imputed
ML1	76k	1,365	76k
ML2	67k	1,491	67k
TL1	60k	731	60k
TL2	42k	760	42k
TL3	105k	1,865	105k
TL4	29k	381	29k

Total = 379k

Terminal lines



Jang et al.
(under review)

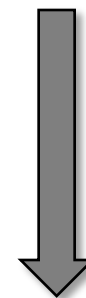
Lines	ADG	BF	ADGX	BFX	Animals in pedigree	Sequenced/ Imputed
TL1	0.36M	0.34M	150k	149k	1.13M	60k
TL2	0.30M	0.30M	158k	156k	0.84M	42k
TL3	0.94M	0.86M	299k	247k	3.14M	105k
Multi	1.60M	1.50M	578k	525k	> 5M	207k

Sequence Variants

15M to 20M variants



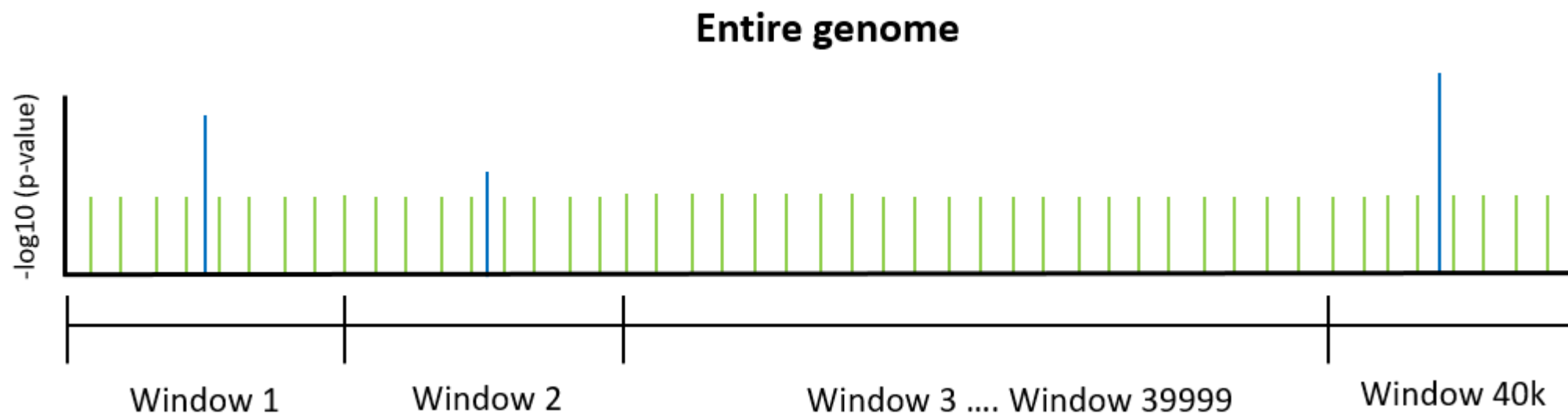
~ 10M segregated across lines



Should we use all 10M?

SNP preselection based on GWAS - I

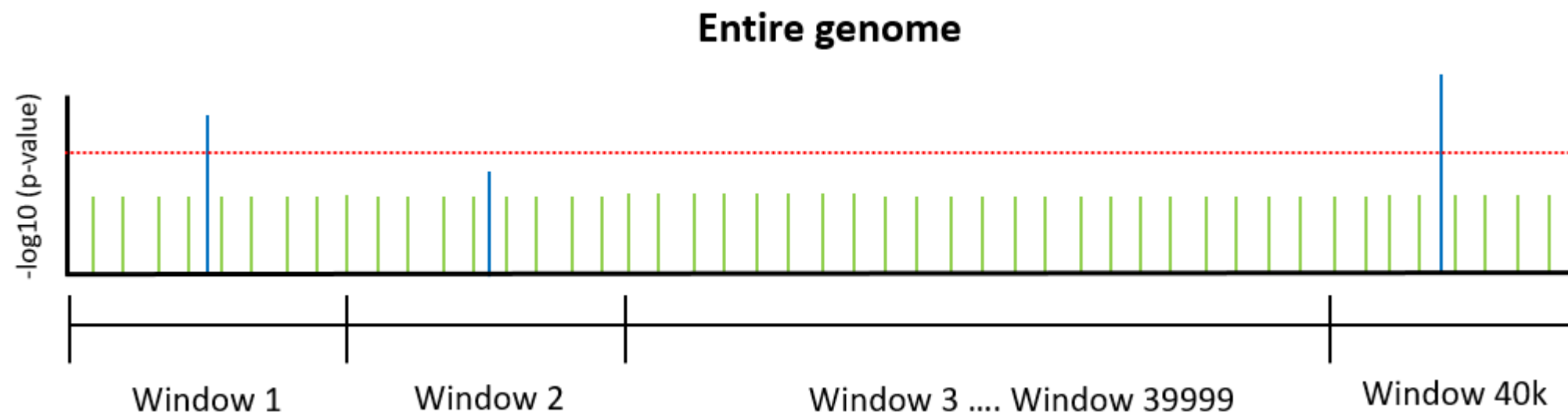
- Top 40k



Extracting only 40k SNP: Similar number as the regular SNP chip (~40k)

SNP preselection based on GWAS - II

- **Chip+Sign**



Extracting only significant ones + 40k SNP chip

Steps

1) Accuracy of GEBV with SNP preselected from sequence data

- Many animals with sequence

2) Single-line and multi-line ssGBLUP evaluations

3) Compare ssGBLUP with BayesR from Roslin

Ros-Freixedes et al.
Genetics Selection Evolution (2022) 54:65
<https://doi.org/10.1186/s12711-022-00756-0>



RESEARCH ARTICLE

Open Access



Genomic prediction with whole-genome
sequence data in intensely selected pig lines

Roger Ros-Freixedes^{1,2*}, Martin Johnsson^{1,3}, Andrew Whalen¹, Ching-Yi Chen⁴, Bruno D. Valente⁴,
William O. Herring⁴, Gregor Gorjanc¹ and John M. Hickey¹

Step 1 – Accuracy with preselected variants

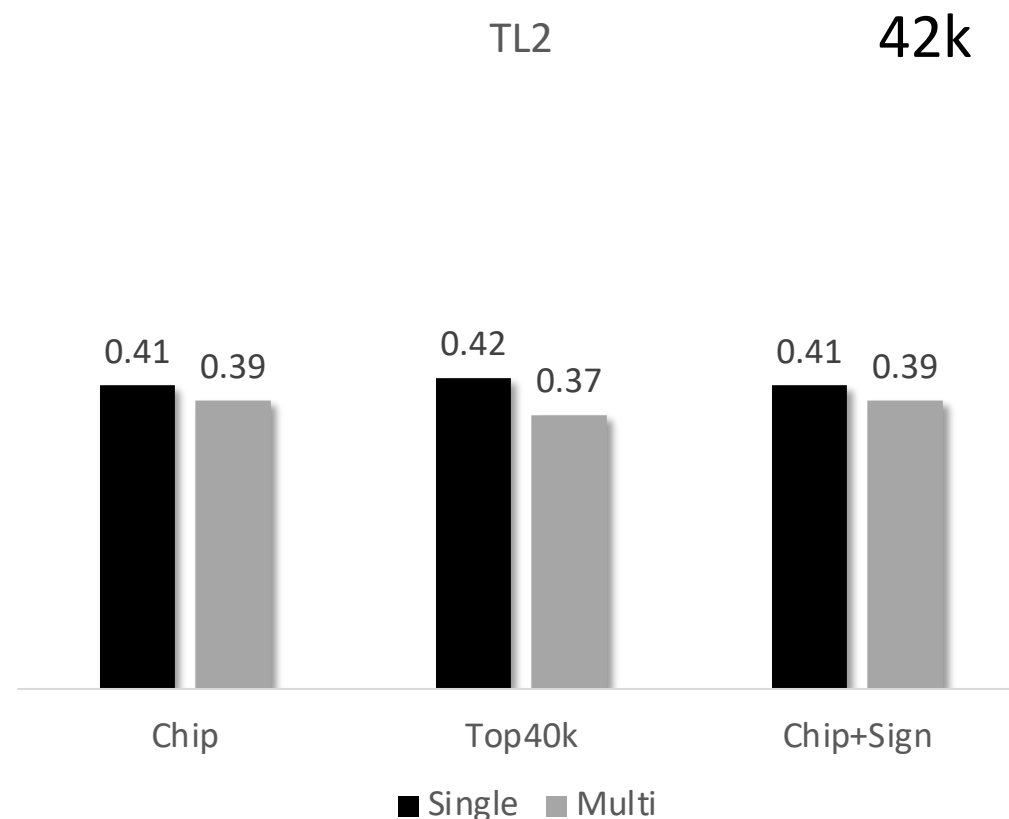
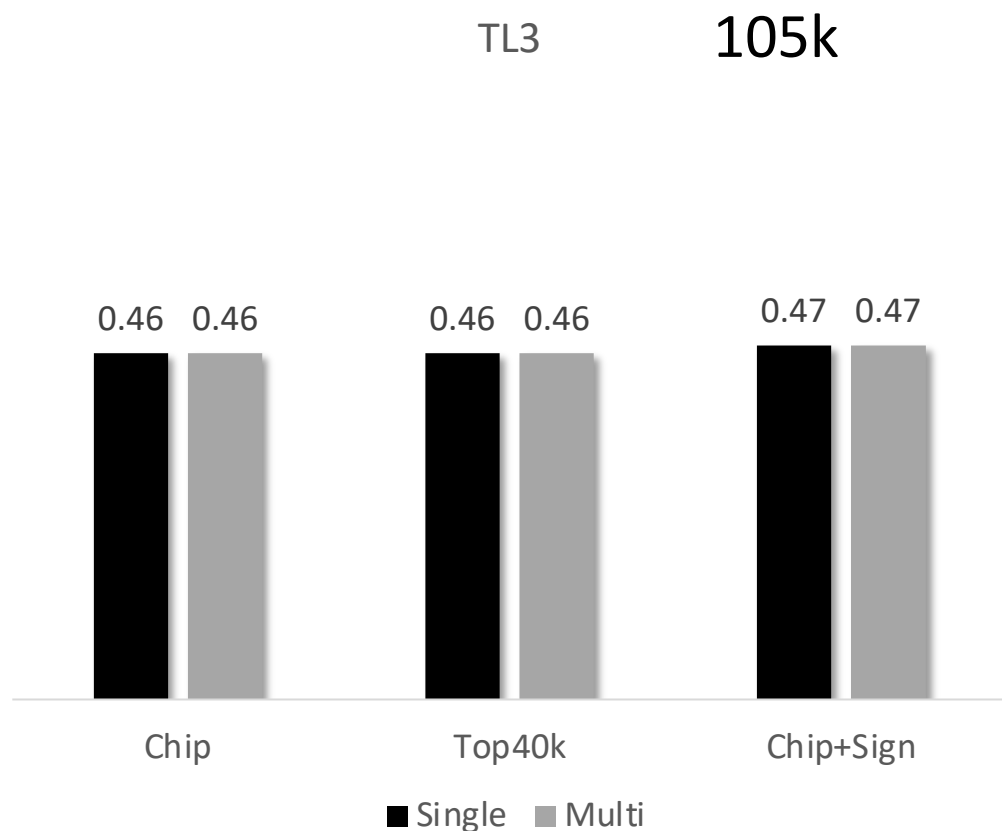
- Prediction accuracy = $\text{cor}(\text{DEBV}, \text{GEBV})$



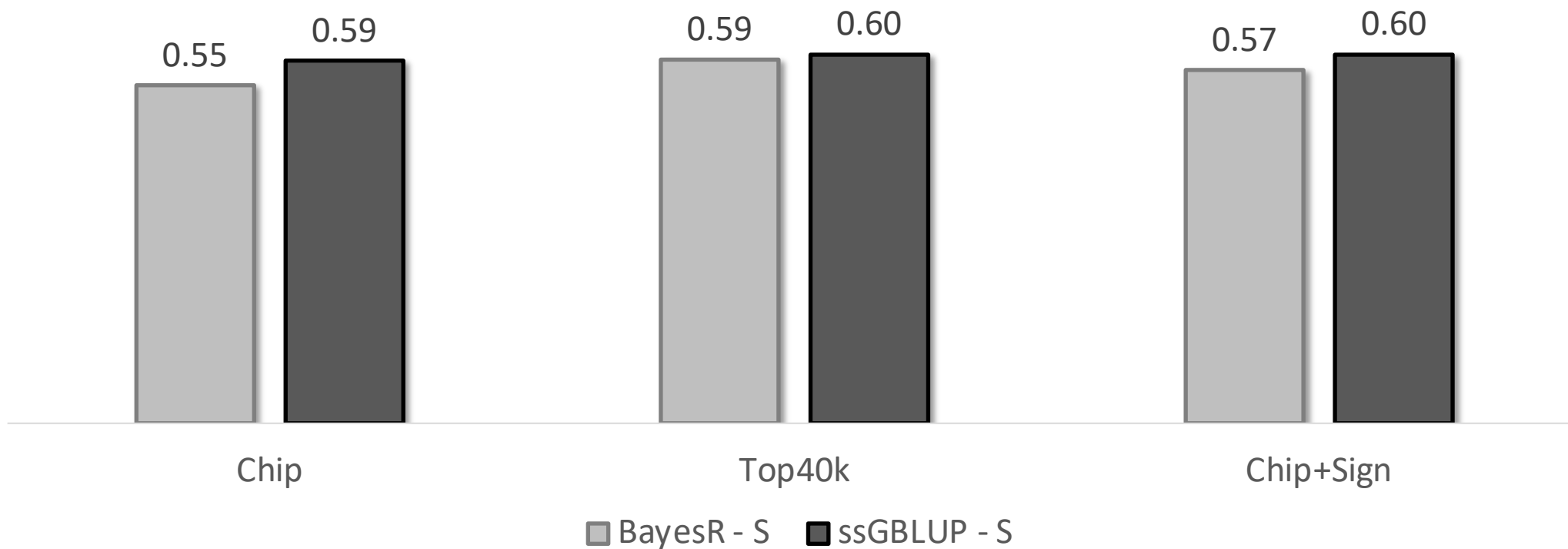
Step 2 – Single vs. Multi-line all traits

- Prediction accuracy = $\text{cor}(\text{DEBV}, \text{GEBV})$

Multi-line GWAS and predictions dominated by TL3



Step 3 - ssGBLUP vs. BayesR



ssGBLUP vs. BayesA in Dairy

VanRaden et al. *Genet Sel Evol* (2017) 49:32
DOI 10.1186/s12711-017-0307-4



RESEARCH ARTICLE

Open Access



Selecting sequence variants to improve genomic predictions for dairy cattle

Paul M. VanRaden^{1*}, Melvin E. Tooker¹, Jeffrey R. O'Connell², John B. Cole¹ and Derek M. Bickhart¹



J. Dairy Sci. 102:10012–10019
<https://doi.org/10.3168/jds.2019-16262>
© American Dairy Science Association[®], 2019.

Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants

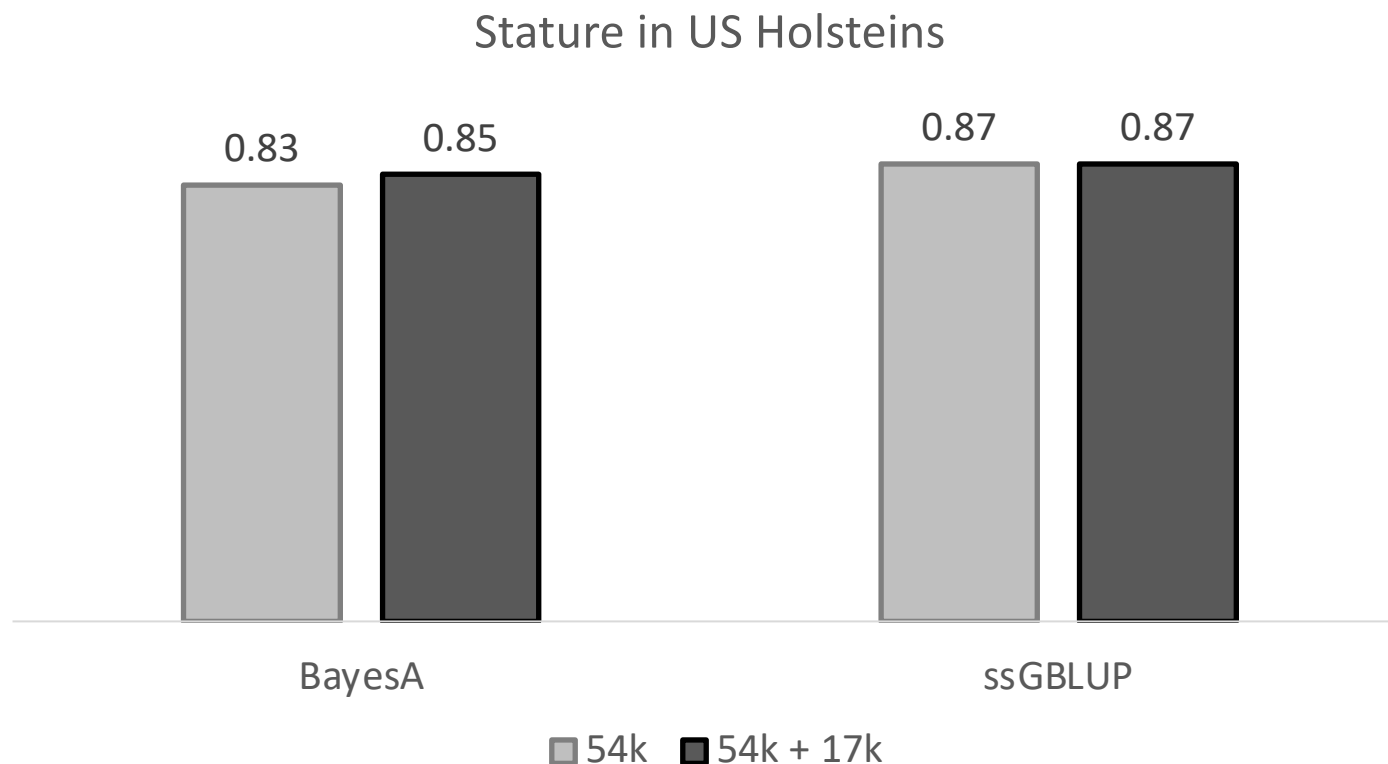
B. O. Fragomeni,^{1*} D. A. L. Lourenco,² A. Legarra,³ P. M. VanRaden,⁴ and I. Misztal²

¹Department of Animal Science, University of Connecticut, Storrs-Mansfield 06269

²Department of Animal and Dairy Science, University of Georgia, Athens 30602

³Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326

⁴Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705



Weights in ssGBLUP

- Bayesian Alphabet
 - Different weights for SNP

- Regular (ss)GBLUP
 - Same weights for SNP

$$\mathbf{G} = \mathbf{ZZ}' q$$

VanRaden, 2008

- Weighted (ss)GBLUP
 - Different weights for SNP
 - Wang et al. (2012)
 - Zhang et al. (2016)

$$\mathbf{G} = \mathbf{ZDZ}' q$$

Different weights for SNP



J. Dairy Sci. 102:10012–10019
<https://doi.org/10.3168/jds.2019-16262>
 © American Dairy Science Association[®], 2019.

Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants

B. O. Fragomeni,^{1*} D. A. L. Lourenco,² A. Legarra,³ P. M. VanRaden,⁴ and I. Misztal²

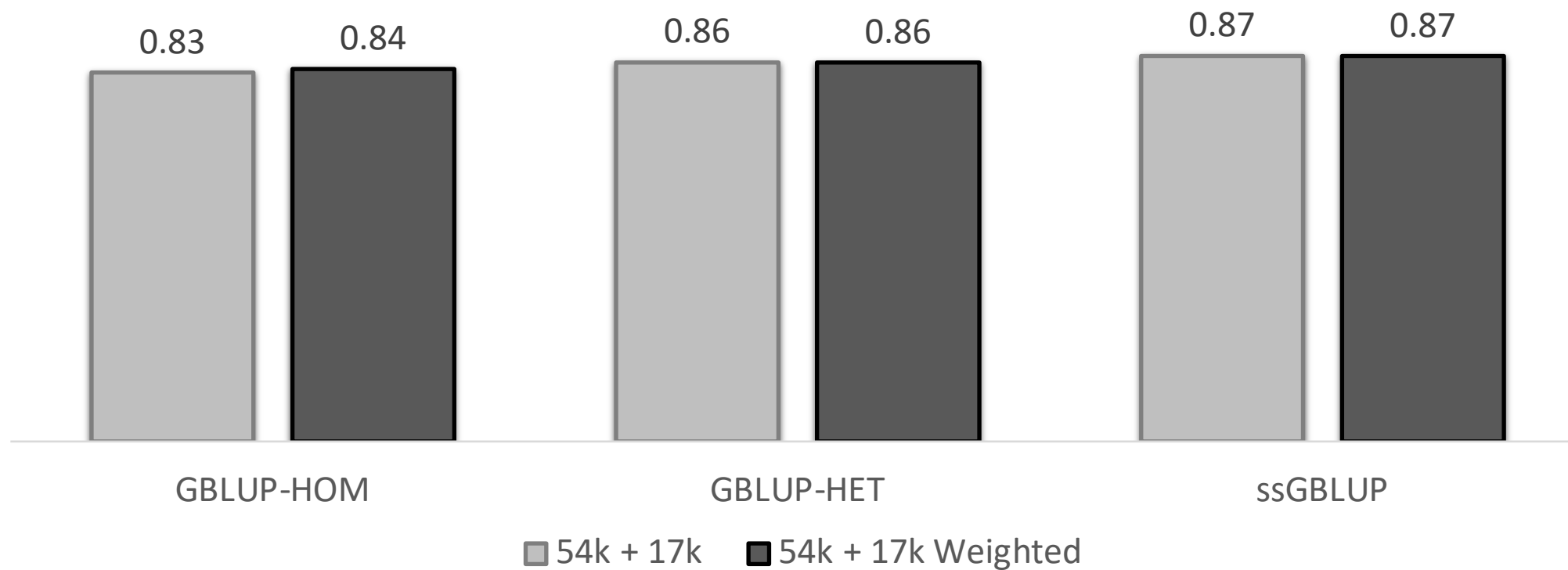
¹Department of Animal Science, University of Connecticut, Storrs-Mansfield 06269

²Department of Animal and Dairy Science, University of Georgia, Athens 30602

³Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326

⁴Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705

Stature in US Holsteins



How many SNP do we need?

- How many SNP?
- How many genotyped individuals?



Dimensionality of
Genomic information

Theory of junctions Fisher (1949)

$$E(Me) = 4N_eL$$

Stam (1980)

Points where the founder chromosome of origin changes

Me – Independent chromosome segments

N_e – Effective population size

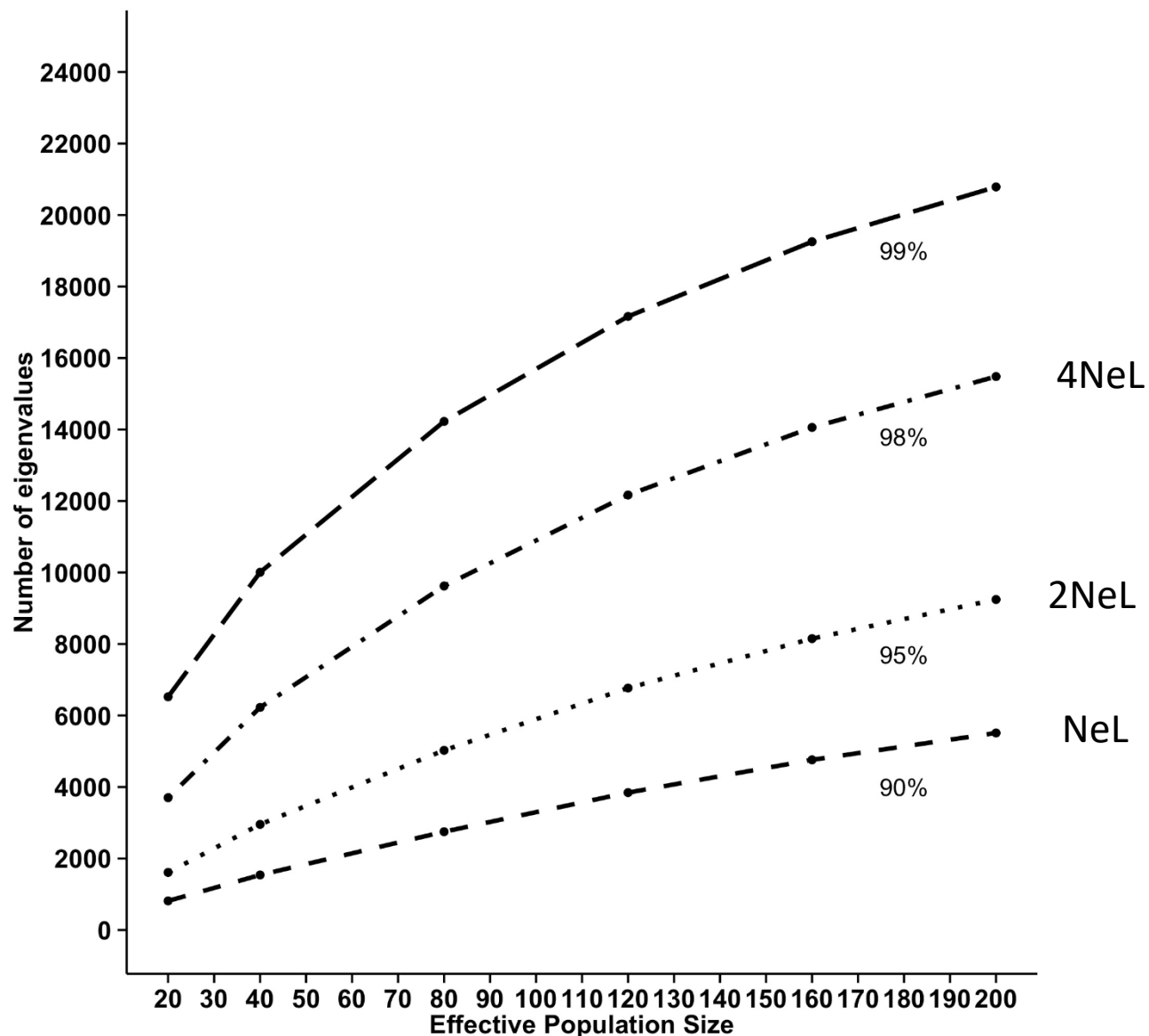
L – Length of genome in Morgans



N_e
↓
 N_e

Dimensionality of genomic information

- Dimensionality of **G**
 - $Me = 4NeL$
 - Eigen 98% of **G**



GENETICS | GENOMIC SELECTION

The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

Ivan Pocrnic,^{*1} Daniela A. L. Lourenco,^{*} Yutaka Masuda,^{*} Andres Legarra,[†] and Ignacy Misztal^{*}
^{*}Department of Animal and Dairy Science, University of Georgia, Athens, Georgia 30602, and [†]Institut National de la Recherche Agronomique, Génétique, Physiologie et Systèmes d'Élevage, F-31326 Castanet-Tolosan, France

Dimensionality of genomic information

Pocrnic et al. (2016b)

Pocrnic et al. (2018)

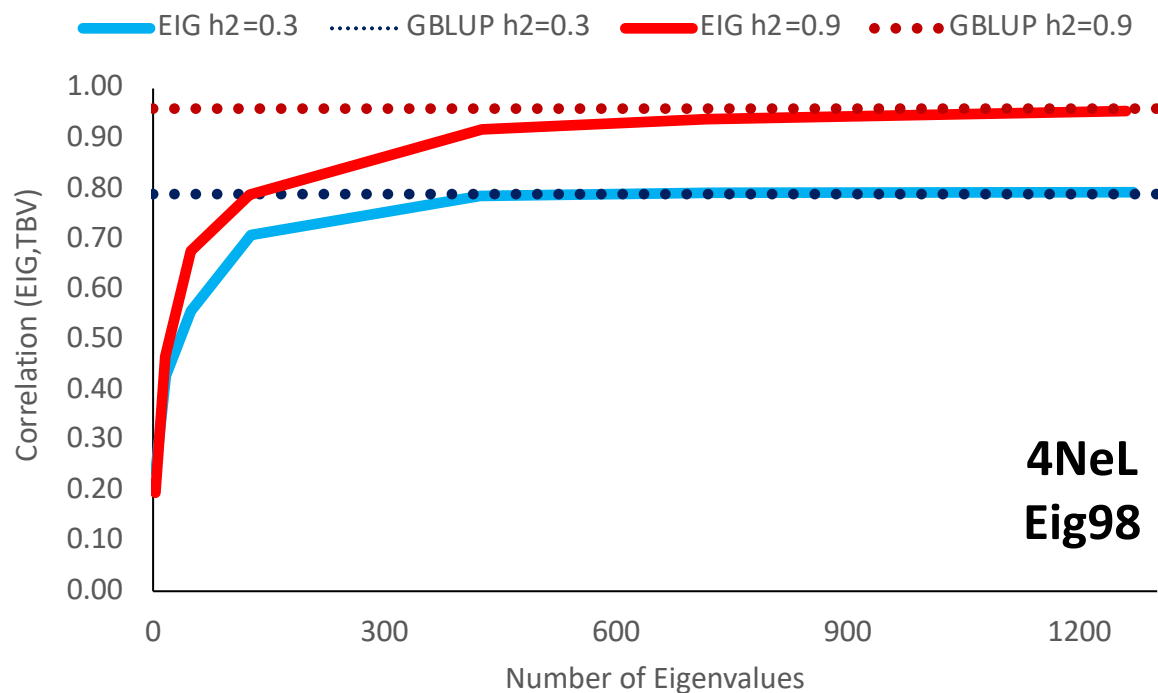
Pocrnic et al. (2019)

Population (Ne)	Genotyped	SNP	Dimensionality (# eigen 98%)	Optimal chip (#eigen 98% *12)
Pig (48)	23 k	37 k	4.1 k	49 k
Chicken(44)	16 k	39 k	4.2 k	50 k
Catfish (45)	7.5 k	55 k	4.5 k	55 k
Angus (113)	81 k	38 k	11 k	127 k
Holstein (149)	77 k	61 k	14 k	168 k

- Optimal chip: 12 times the number of segments (MacLeod et al., 2005)

Why small gains in accuracy with sequence?

- Genomic selection acts in segments and not individual SNP
 - $Me = 4NeL$ Stam (1980)
 - $Me = 4NeL = Eig98$



The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

Ivan Pocrnic,^{*1} Daniela A. L. Lourenco,^{*} Yutaka Masuda,^{*} Andres Legarra,¹ and Ignacy Misztal^{*}
^{*}Department of Animal and Dairy Science, University of Georgia, Athens, Georgia 30602, and ¹Institut National de la Recherche Agronomique, Génétique, Physiologie et Systèmes d’Élevage, F-31326 Castanet-Tolosan, France

Species/breed	Ne	Segments
Pigs	48	4.1k
Chicken	44	4.2k
Jersey	101	11.5k
Angus	113	10.6k
Holstein	149	14k

Pocrnic et al. *Genet Sel Evol* (2019) 51:75
<https://doi.org/10.1186/s12711-019-0516-0>



RESEARCH ARTICLE

Open Access

Accuracy of genomic BLUP when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study

Ivan Pocrnic¹, Daniela A. L. Lourenco, Yutaka Masuda and Ignacy Misztal

Why small gains in accuracy with sequence?

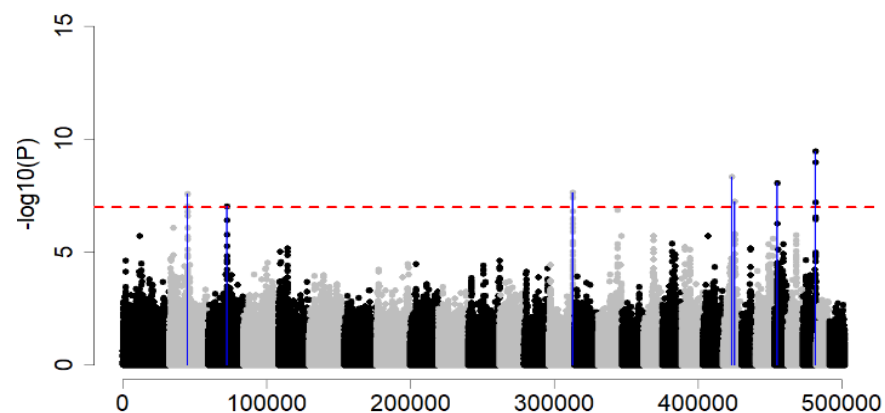
- Amount of information to identify causative variants



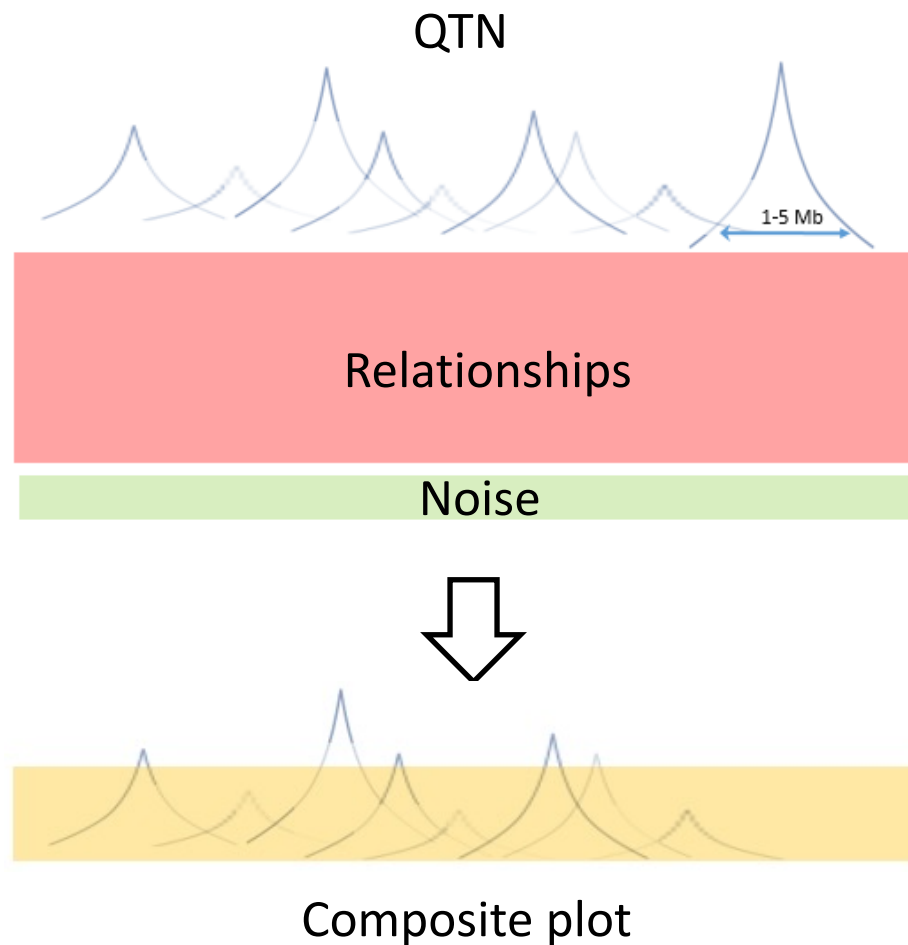
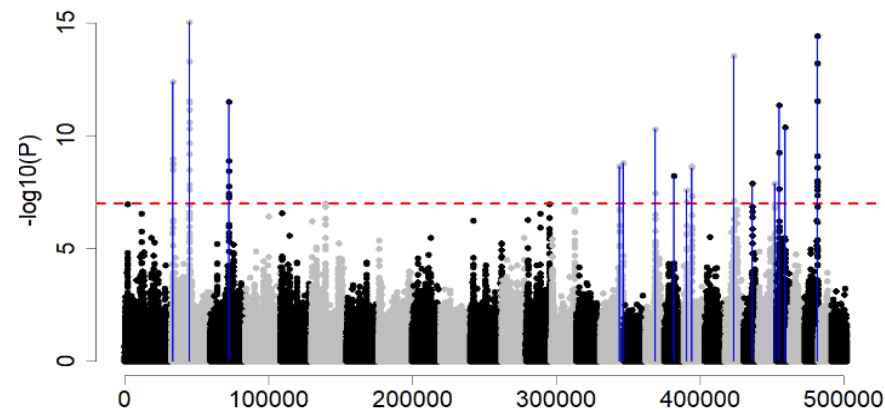
Jang et al.
(under review)

Ne=200 QTN=2000

Eig98 N=15,200



All N=30,000





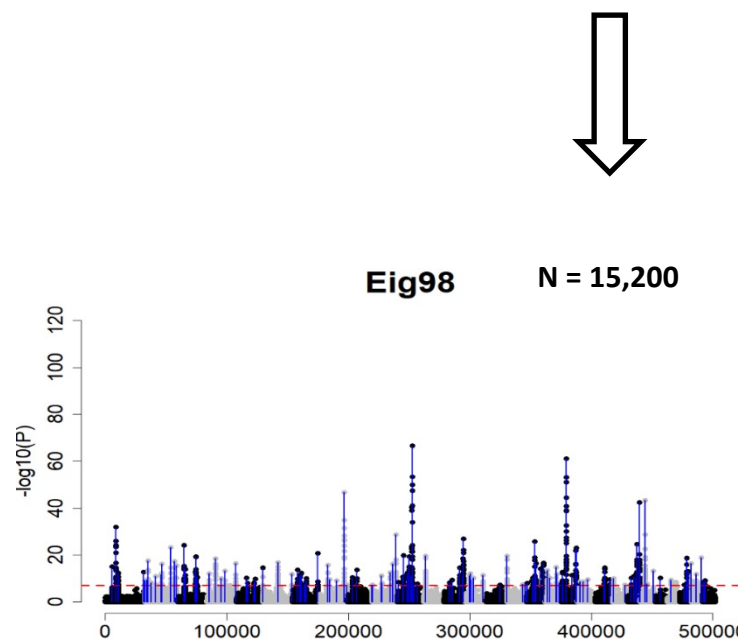
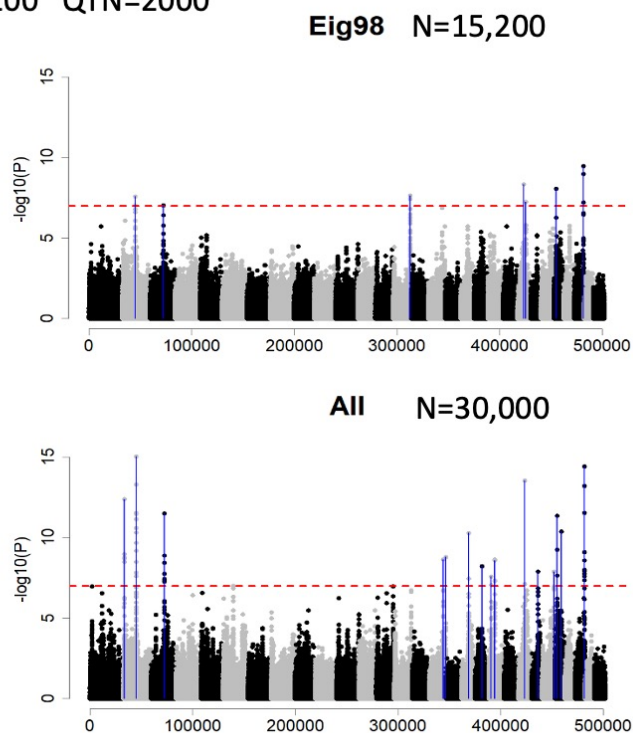
Why small gains in accuracy with sequence?

- Amount of information to identify causative variants
 - Animal with lots of information
 - GEV accuracy ~ 0.99
 - GEV backsolved to SNP effects
 - GWAS resolution with sample size = Me = Eig98 animals with almost perfect accuracy

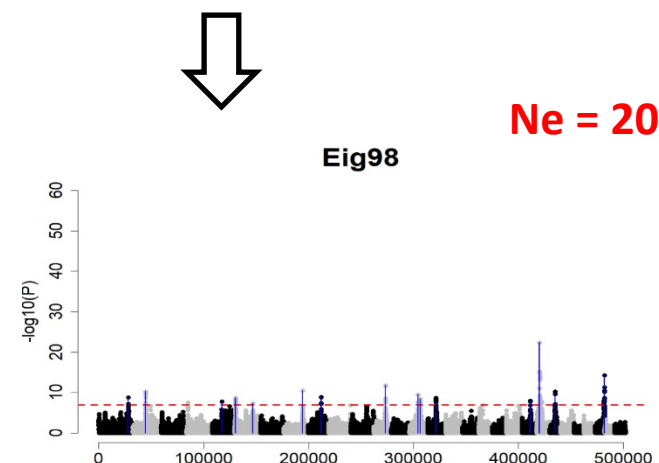


Jang et al. (under review)

$N_e=200$ $QTN=2000$



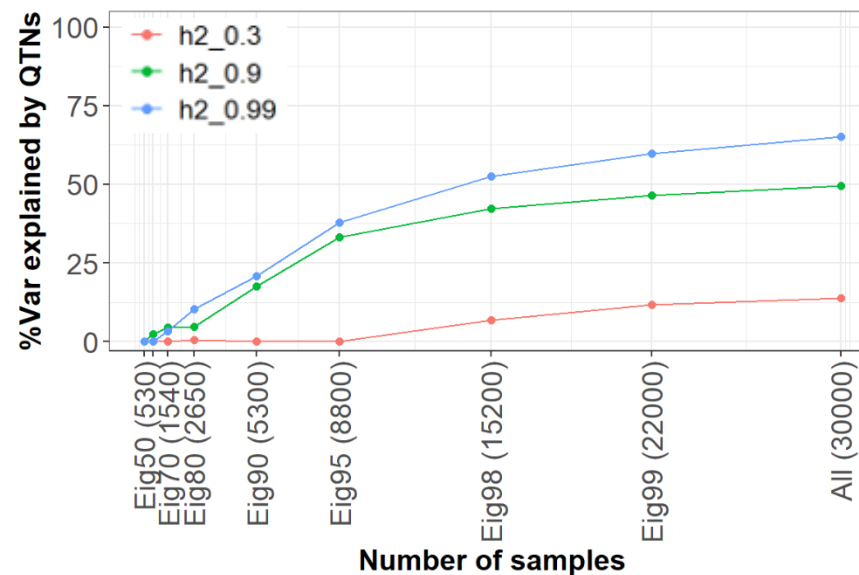
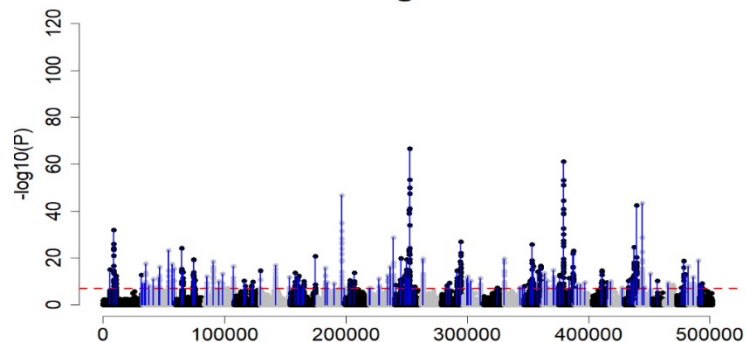
Lots of records for each genotyped animal



Amount of data in GWAS

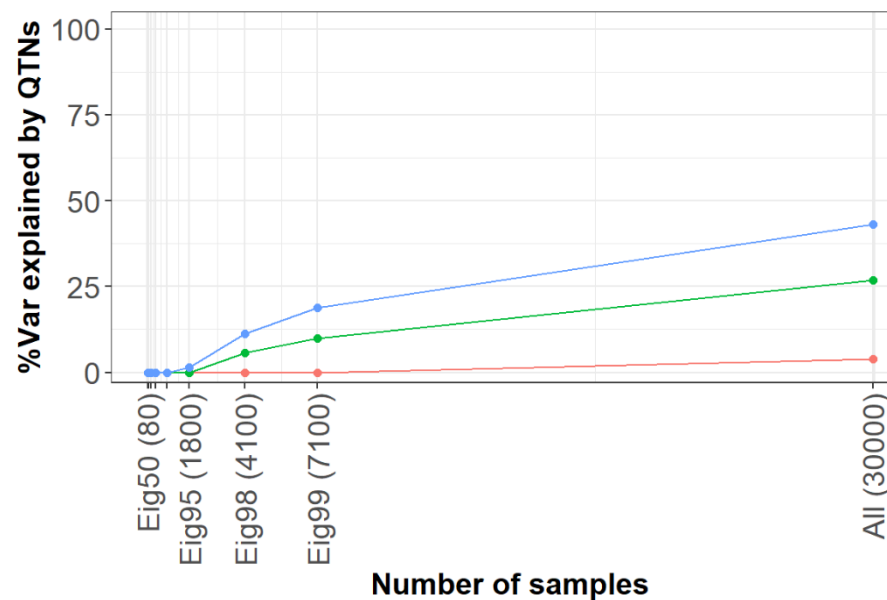
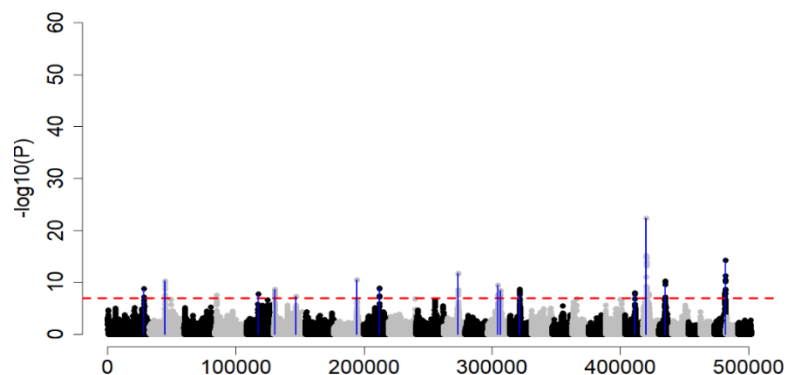
Ne = 200

Eig98



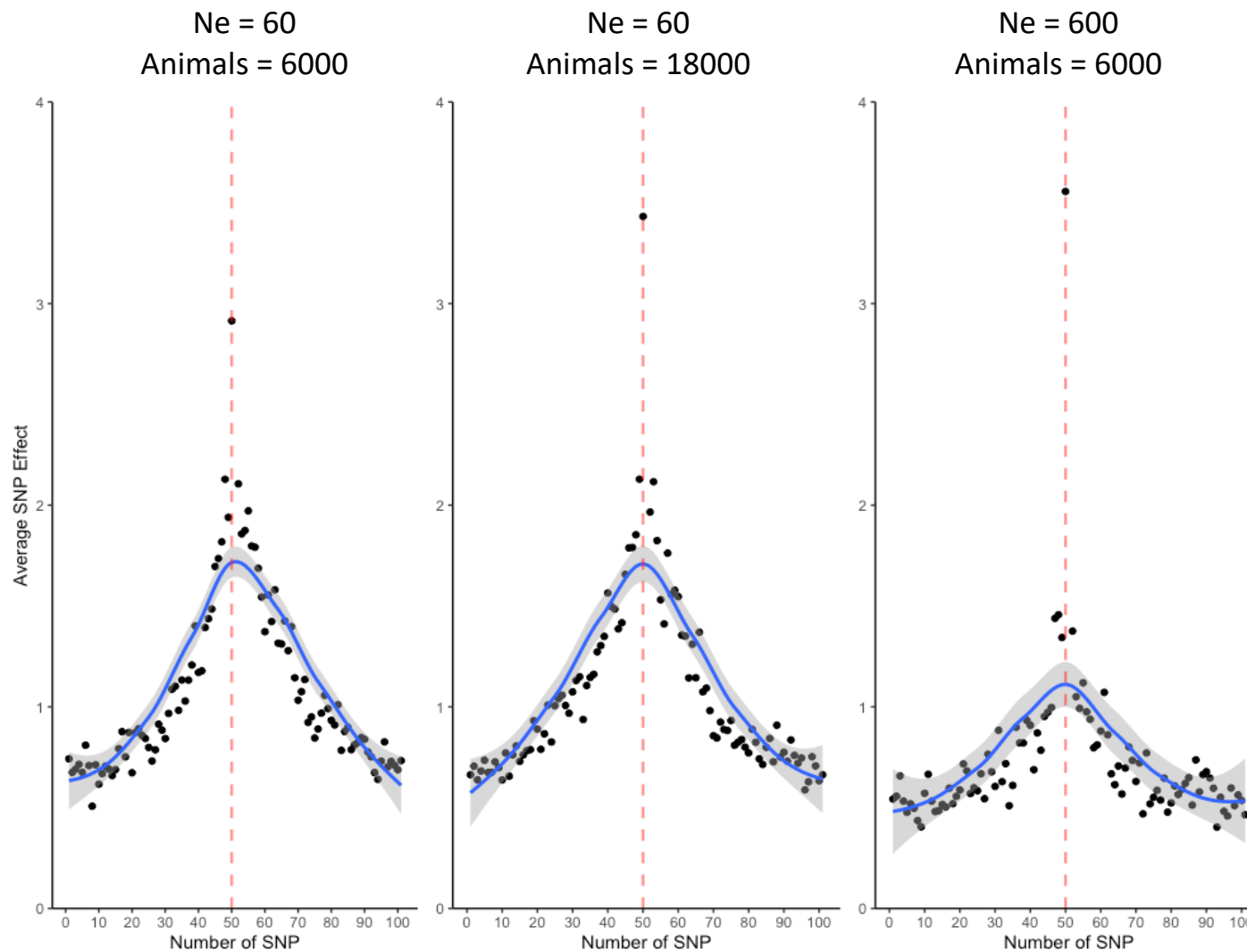
Ne = 20

Eig98



Amount of data in GWAS

- Simulated population (10 QTN per CHR)

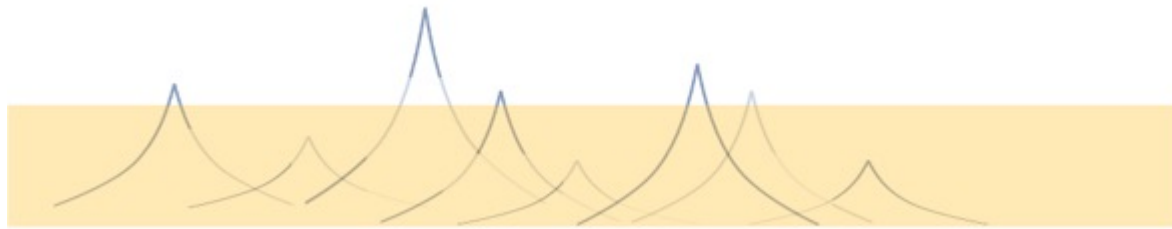


Pocrnic et al.
(in preparation)

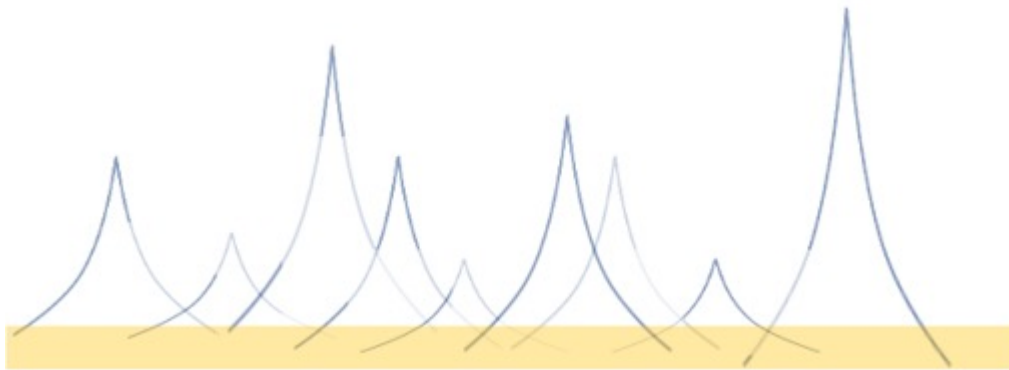
Amount of data in GWAS

Composite Manhattan plot

Pocrnic et al.
(in preparation)



Small populations



Large populations

Single-step GWAS

- Single-step GBLUP (ssGBLUP): genotyped and non-genotyped animals

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

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

Aguilar et al. (2010); Christensen & Lund (2010)

$$\text{Var}(\mathbf{u}) = \mathbf{H}\sigma_{\mathbf{u}}^2$$

- ssGBLUP vs. ssSNP-BLUP equivalent models
- GEBV vs. SNP effects

$$\mathbf{u} = \text{GEBV}$$

$$\mathbf{u} = \mathbf{Z}\mathbf{a}$$

↳ Vector of SNP effects
 ↳ Matrix of gene content

- Backsolve GEBV into SNP effects $\hat{\mathbf{a}}|\hat{\mathbf{u}} = k\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$

- Backsolve PEV of GEBV into PEV of SNP

$$\text{Var}(\hat{a}_i) = k\mathbf{z}_i'\mathbf{G}^{-1}(\mathbf{G}\sigma_{\mathbf{u}}^2 - \mathbf{C}^{u_2u_2})\mathbf{G}^{-1}\mathbf{z}_ik$$

Single-step GWAS

- P-values for SNP effects
- Heavy computations
 - Inverse of LHS of MME
 - limit of $\sim 50k$ genotyped animals
 - 1 M pedigree
 - 1 trait

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

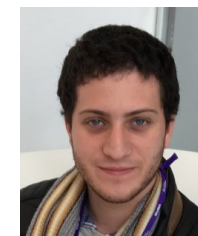
Ignacio Aguilar¹, Andres Legarra^{2*}, Fernando Cardoso^{3,4}, Yutaka Masuda⁵, Daniela Lourenco⁵ and Ignacy Misztal⁵



Accounting for all information in GWAS



Natalia Leite

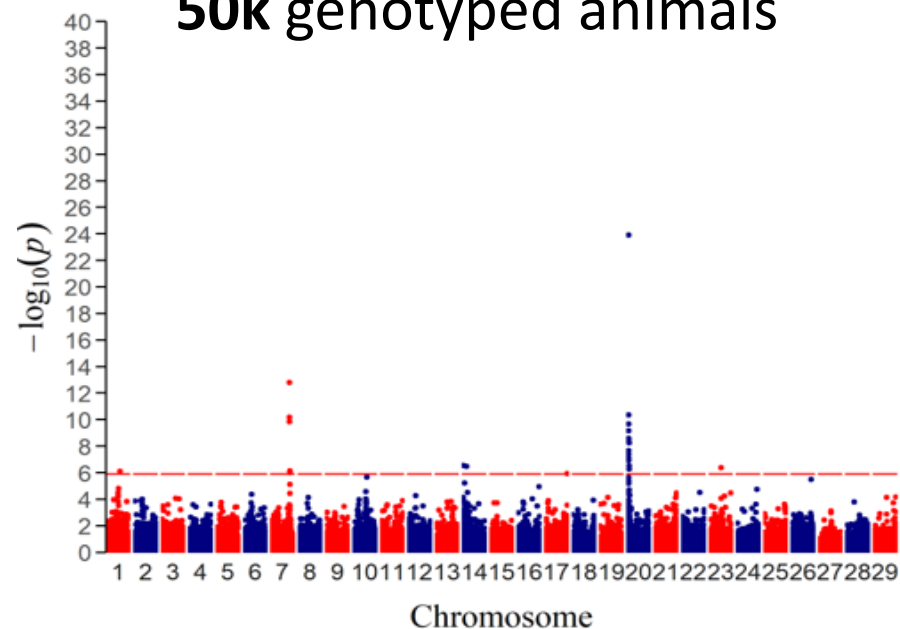


Matias Bermann

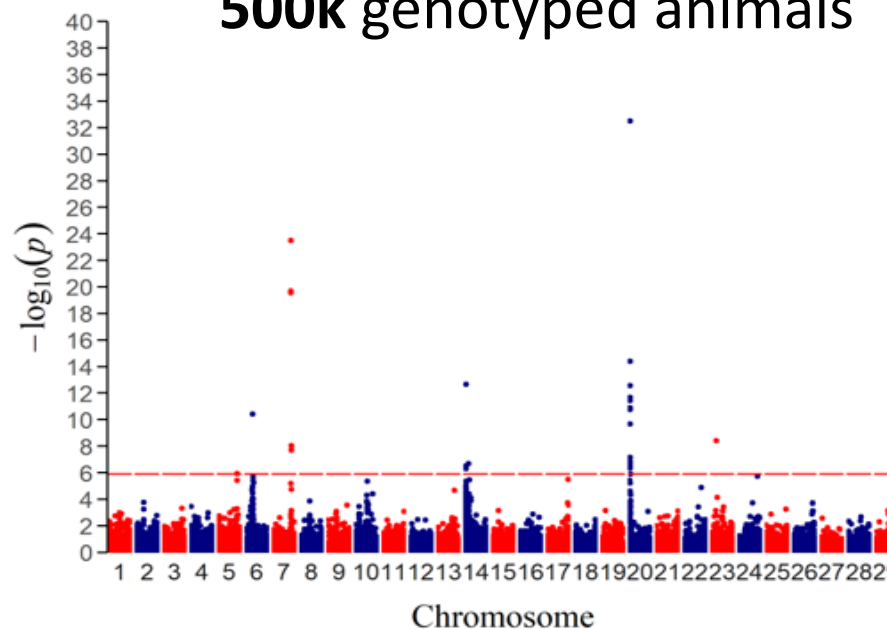
Regular ssGWAS – limitation ~ 50k

ssGWAS based on dimensionality

50k genotyped animals



500k genotyped animals



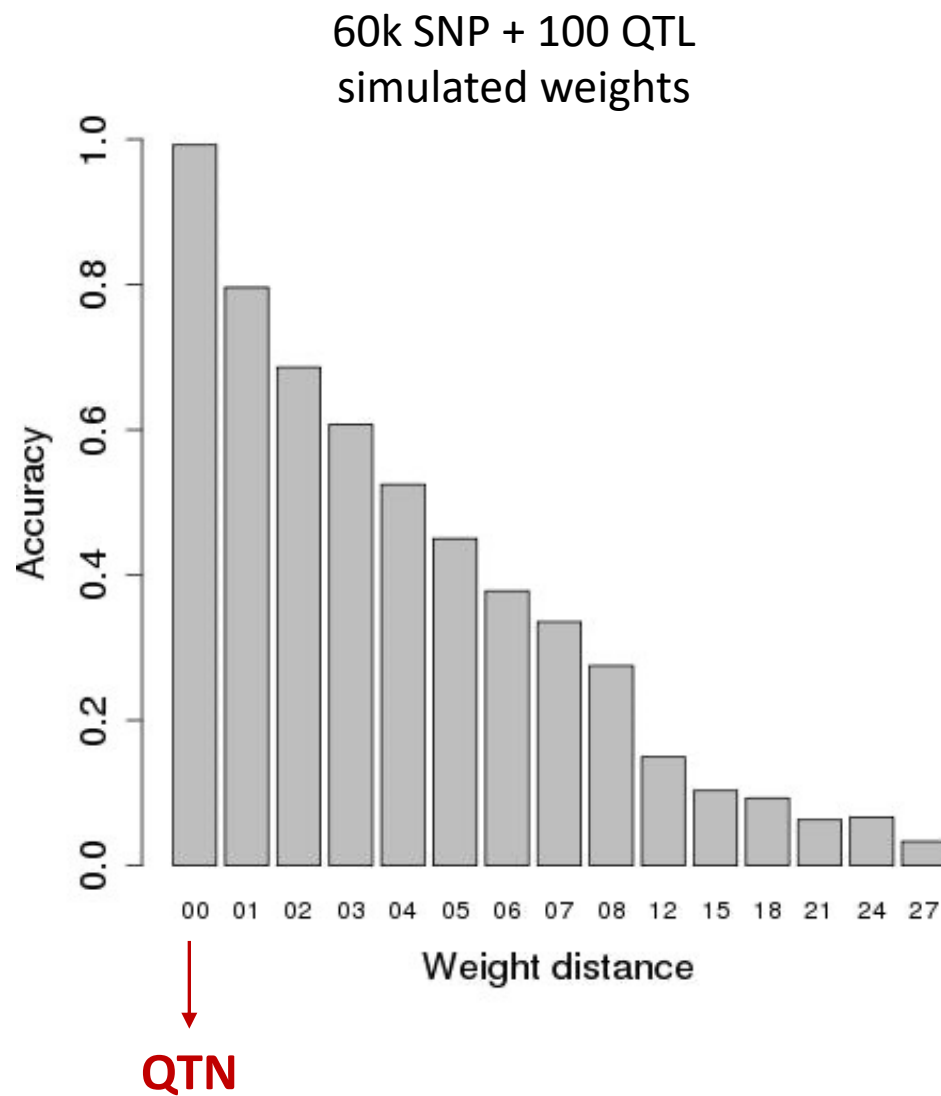
Identifying causative variants

Ability to identify
true causative variants
providing independent information



Drive the increase in
accuracy of genomic predictions

Causative variants – simulated data



Fragomeni et al. *Genet Sel Evol* (2017) 49:59
DOI 10.1186/s12711-017-0335-0

GSE Genetics
Selection
Evolution

RESEARCH ARTICLE

Open Access

Incorporation of causative quantitative
trait nucleotides in single-step GBLUP



Breno O. Fragomeni^{1*}, Daniela A. L. Lourenco¹, Yutaka Masuda¹, Andres Legarra² and Ignacy Misztal¹

What is next?

- Whole-genome sequence data
- Functional annotation
- Omics data
- Sensors
- Cameras
- Enviromics

Selecting sequence variants to improve genomic predictions for dairy cattle

Paul M. VanRaden , Melvin E. Tooker, Jeffrey R. O'Connell, John B. Cole & Derek M. Bickhart

Genetics Selection Evolution **49**, Article number: 32 (2017) | [Cite this article](#)


Incorporation of causative quantitative trait nucleotides in single-step GBLUP

Breno O. Fragomeni , Daniela A. L. Lourenco, Yutaka Masuda, Andres Legarra & Ignacy Misztal

Genetics Selection Evolution **49**, Article number: 59 (2017) | [Cite this article](#)


Front. Anim. Sci. 11 February 2021 | <https://doi.org/10.3389/fanim.2021.650324>

Grand Challenge in Precision Livestock Farming

 Guilherme J. M. Rosa*

Department of Animal and Dairy Sciences, Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States

Predicting Growth and Carcass Traits in Swine Using Microbiome Data and Machine Learning Algorithms

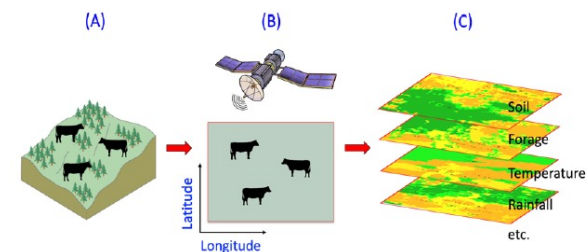
Christian Maltecca , Duc Lu, Constantino Schillebeeckx, Nathan P. McNulty, Clint Schwab, Caleb Shull & Francesco Tiezzi 

Scientific Reports **9**, Article number: 6574 (2019) | [Cite this article](#)

Genetic evaluation including intermediate omics features

Ole F Christensen , Vinzent Börner, Luis Varona, Andres Legarra

Genetics, Volume 219, Issue 2, October 2021, iyab130,
<https://doi.org/10.1093/genetics/iyab130>

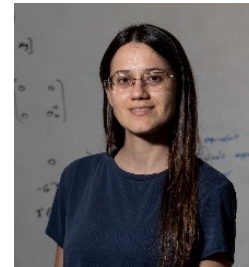
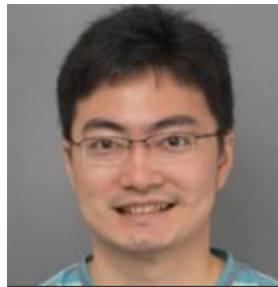
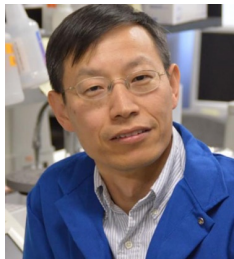
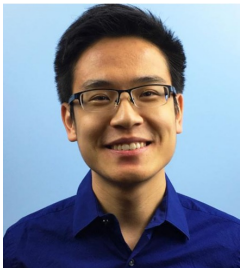


Enviromics-enabled precision breeding for adapted cattle:

Rosa, Lourenco et al. (2022)

What is next?

- Statistical methods
 - Intermediate omics data
 - Function annotation in genomic predictions
- Application to commercial pig data

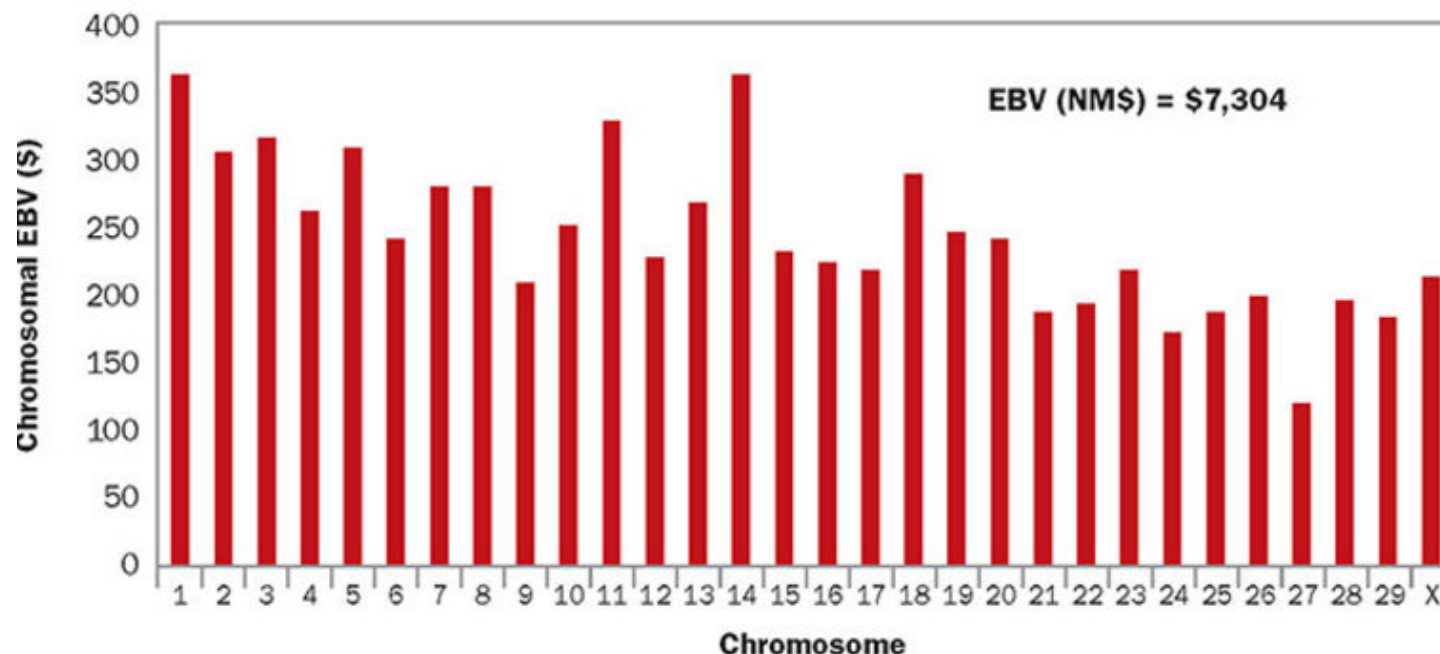


Have we reached the limit of GS?

Best chromosomes in the Us Holstein population

We want to get the best DNA together in one animal

John Cole
(2019)



Sum of the effects of SNP
in each chromosome
for each animal

- Hypothetical animal based on chromosomal EBV: NM\$ 7,304
- The top bull available for sale in 08/2022: NM\$ 1,285 (born in 2/21)

Take home message

- Using sequence variants for genomic predictions
 - Limited benefit in single and multi-breed populations
 - 0 – 5% gain in accuracy
 - Small or large populations
- Identifying true causative variants is the key
 - Poor job because of the amount of data
 - Identified SNP may be redundant with the SNP chip
 - Extra sources of information may help
- Several other uses for sequence data
 - Genomic predictions: flexibility - virtual SNP panel at any time
 - Genetic architecture of traits
 - Etc, ...

UGA AB&G team



<http://nce.ads.uga.edu>

ANIMAL BREEDING AND GENETICS GROUP

UNIVERSITY OF GEORGIA

[HOME](#) [NEWS](#) [RESEARCH](#) [PUBLICATIONS](#) [SOFTWARE](#) [EDUCATION](#) [PEOPLE](#) [ABOUT](#)

<http://nce.ads.uga.edu/wiki>



The screenshot shows the BLUPF90 wiki page. At the top left is the BLUPF90 logo, which consists of a pencil and a paper with the letters 'BLUPF90' written on it. To the right of the logo is the title 'BLUPF90'. In the top right corner, there is a search bar and two links: 'Media Manager' and 'Sitemap'. Below the search bar, there is a 'Trace' section with a link to 'start'. The main content area is titled 'BLUPF90 Family of Programs' and includes a sub-section 'Now with support for genomic selection'. Below this, there is a paragraph of text describing the software and its purpose. To the right of the main content, there is a 'Table of Contents' section with a list of links: 'BLUPF90 Family of Programs', 'Headline', and 'Courses'. At the bottom of the page, there is a 'Headline' section with a list of links: 'History', 'Modules', 'Condition of use', 'Distribution / Download', 'Documentation / Manual / Tutorial', 'Application program details', 'Support', 'FAQ', 'Tricks / Tips', 'To Do', 'Sample data', and 'Undocumented options'.

BLUPF90

Search

Media Manager Sitemap

Trace: [start](#)

BLUPF90 Family of Programs

Now with support for genomic selection

Ignacy Misztal and collaborators, University of Georgia

BLUPF90 family of programs is a collection of software in Fortran 90/95 for mixed model computations in animal breeding. The goal of the software is to be as simple as with a matrix package and as efficient as in a programming language. For general description, see a [paper](#) from the CCB'99 workshop or see a [paper](#) on BGF90 at 7th WCGALP.

For variance component estimation, the family offers choices for simple and complicated models; see paper ["Reliable computing in estimation of variance components"](#). From 2009 the programs are successively modified for genomic selection using a [Single-step](#) approach (or ssGBLUP) by Ignacio Aguilar and Shogo Tsuruta.

For support, join [Blupf90 Discussion Group](#) at Groups.io. We moved from Yahoo Groups to Groups.io on November 7, 2019, mainly because of the unavailability of key features in Yahoo Groups. We no longer maintain the old group.

Please visit [our main web-site](#) for details in research and publication.

Troubleshooting

! If the software crashes with segmentation fault, please change settings in your operating system. See [FAQ: Segmentation fault](#) for details. Also, The [FAQ pages](#) provide useful suggestions and solutions.

Headline

- [History](#)
- [Modules](#)
- [Condition of use](#)
- [Distribution / Download](#)
- [Documentation / Manual / Tutorial](#)
- [Application program details](#)
- [Support](#)
- [FAQ](#)
- [Tricks / Tips](#)
- [To Do](#)
- [Sample data](#)
- [Undocumented options](#)

Table of Contents

- [BLUPF90 Family of Programs](#)
- [Headline](#)
- [Courses](#)