Large-Scale Genomic Predictions with Sequence Data

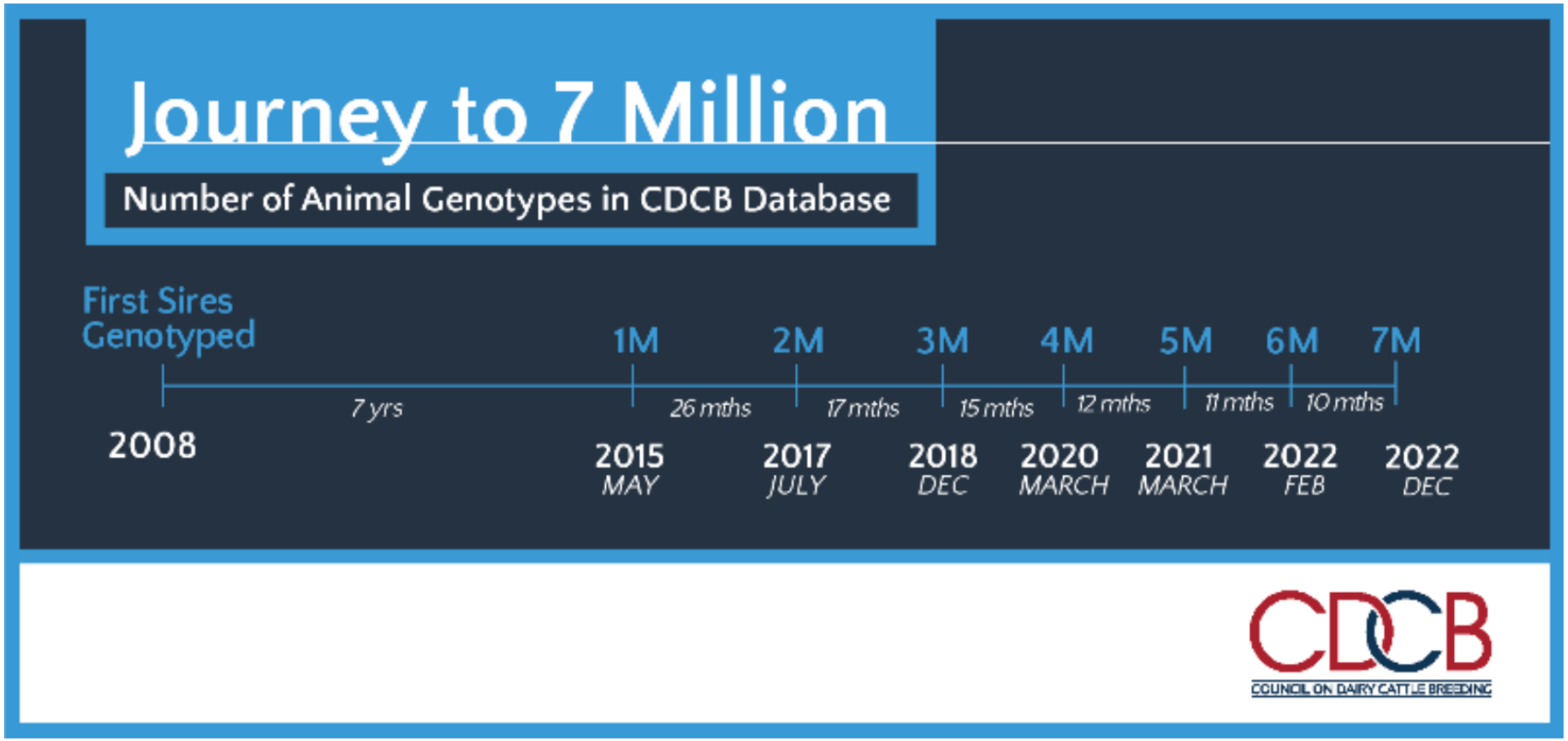
Daniela Lourenco

Shogo Tsuruta, Ignacy Misztal

January 15, 2023

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Massive uptake of genomics in 15 years



~ 4 million genotyped animals

Alberto Cesarani

J. Dairy Sci. 105:5141–5152
https://doi.org/10.3168/jds.2021-21505

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

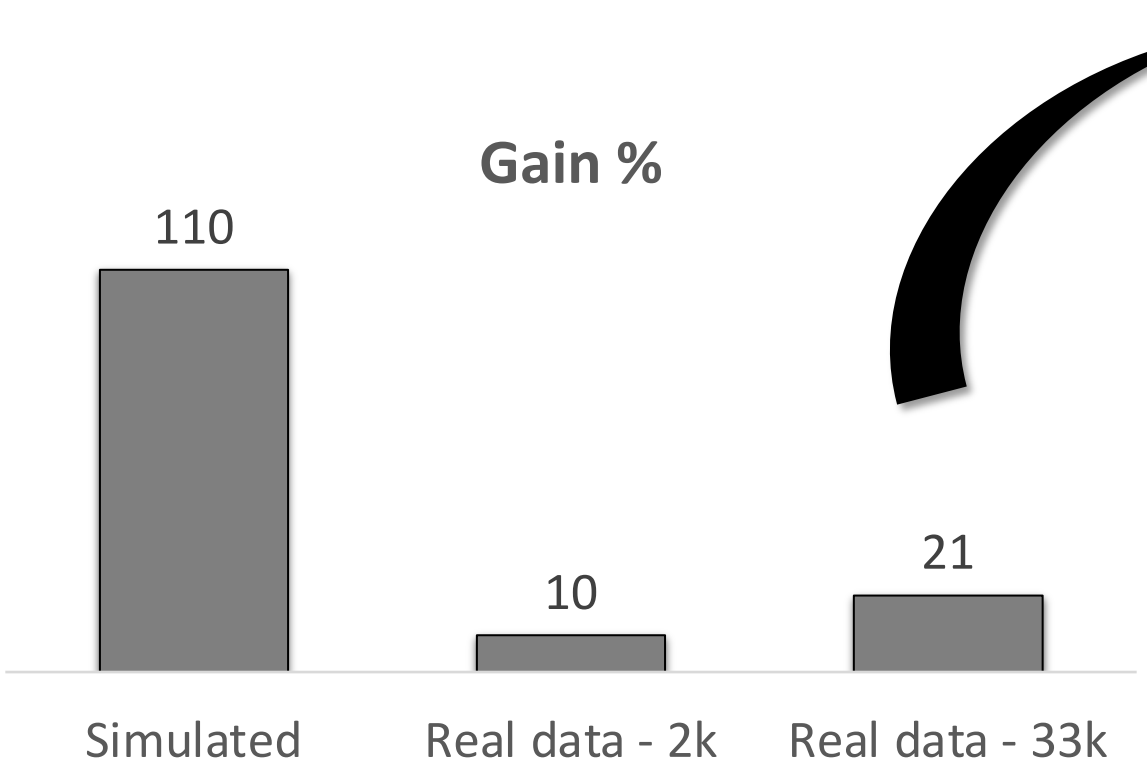
A. Cesarani,[1*] D. Lourenco,[1] S. Tsuruta,[1] A. Legarra,[2] E. L. Nicolazzi,[3] P. M. VanRaden,[4] and I. Misztal[1]

Why do we care about lots of data?

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

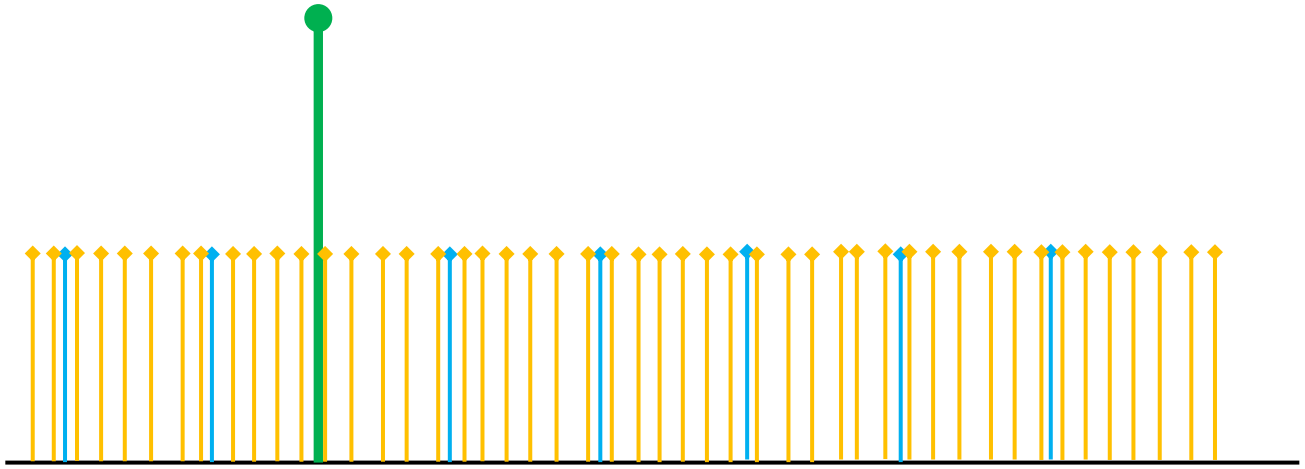
- Predicting things is very hard
  - Gather enough data
  - Use the right statistical tools

Gain with genomics



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

The idea behind sequence data



~ 50k SNP

~ 30M SNP

Sequencing is becoming affordable

Euan Ashley (2022 @Nanopore Conference)

… and fast

Euan Ashley (2022 @AGBT Conference)

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

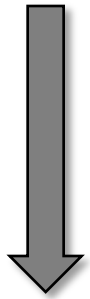
Terminal lines

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

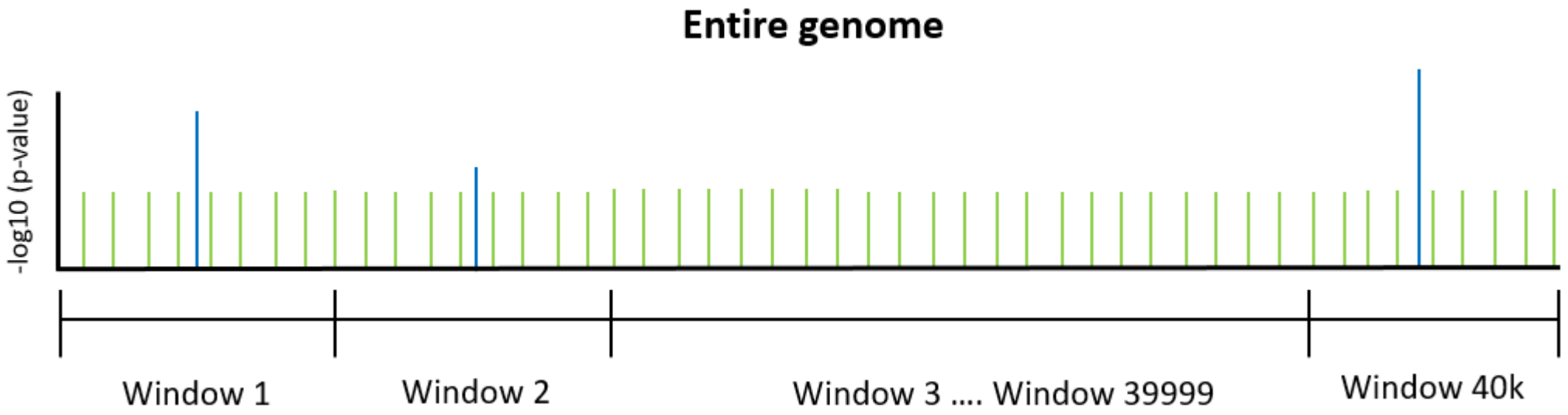
15M to 20M variants  ~ 10M segregated across lines



Should we use all 10M?

SNP preselection based on GWAS - I

- **Top 40k**

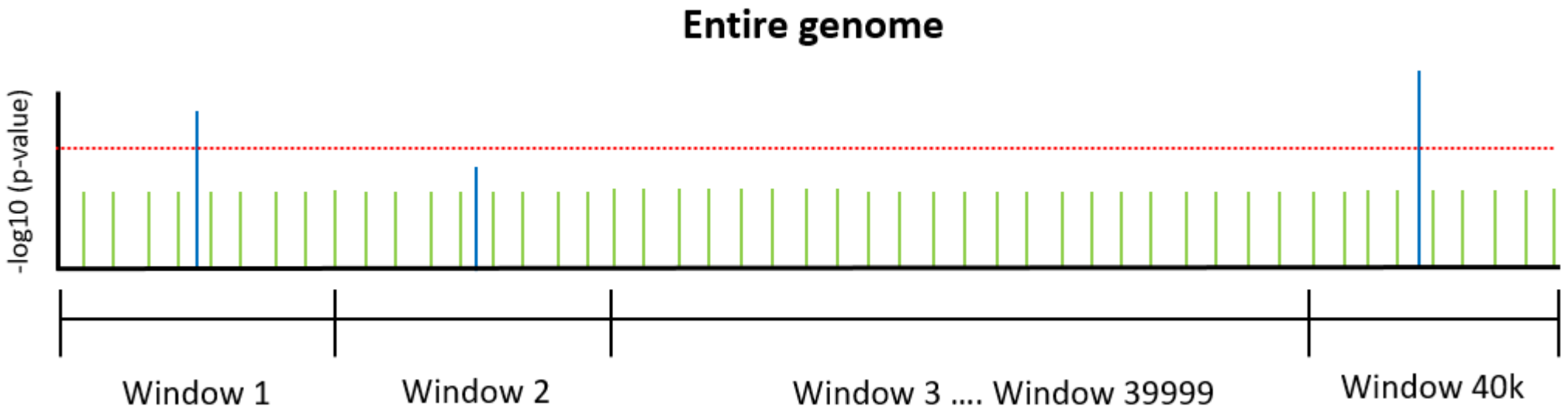


Extracting only 40k SNP: Similar number as the regular SNP chip (~40k)

Ros-Freixedes et al., 2022

SNP preselection based on GWAS - II

- **Chip+Sign**



Extracting only significant ones + 40k SNP chip

Ros-Freixedes et al., 2022

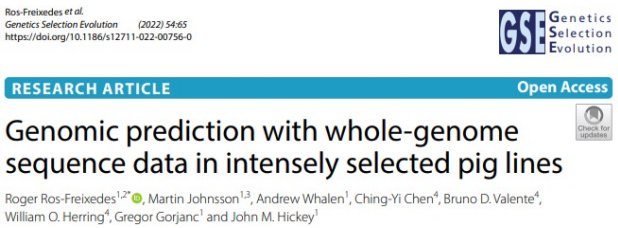
Steps

1) Accuracy of GEBV with SNP preselected from sequence data

- Many animals with sequence

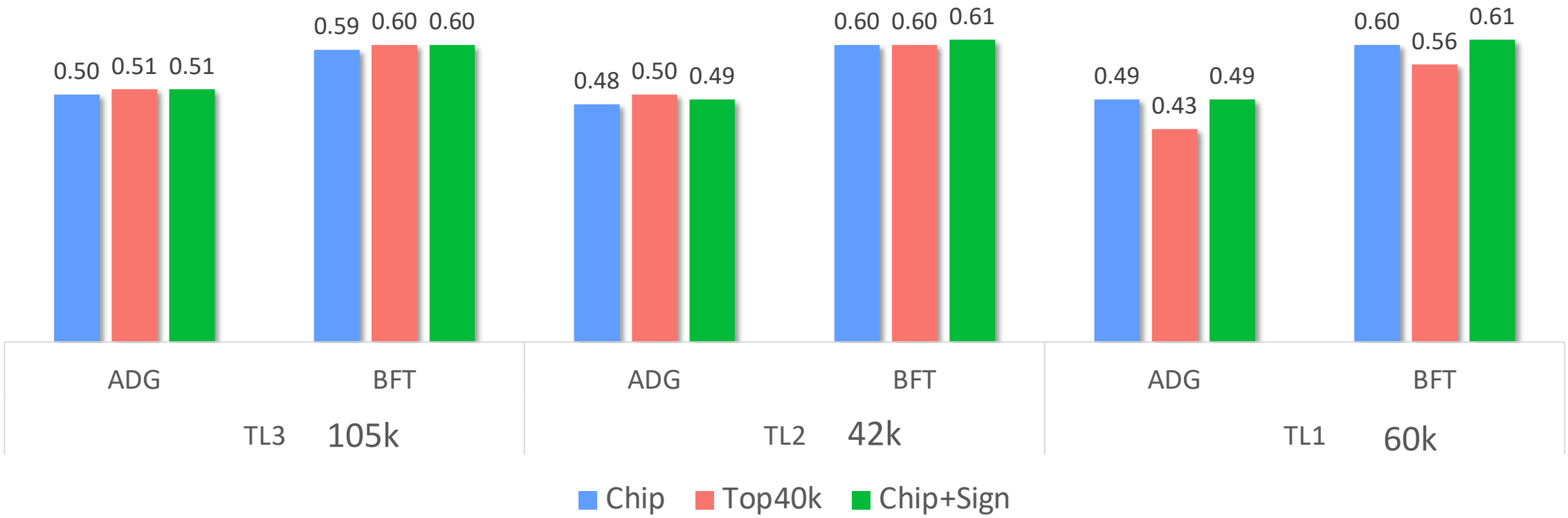
2) Single-line and multi-line ssGBLUP evaluations

3) Compare ssGBLUP with BayesR from Roslin



Step 1 – Accuracy with preselected variants

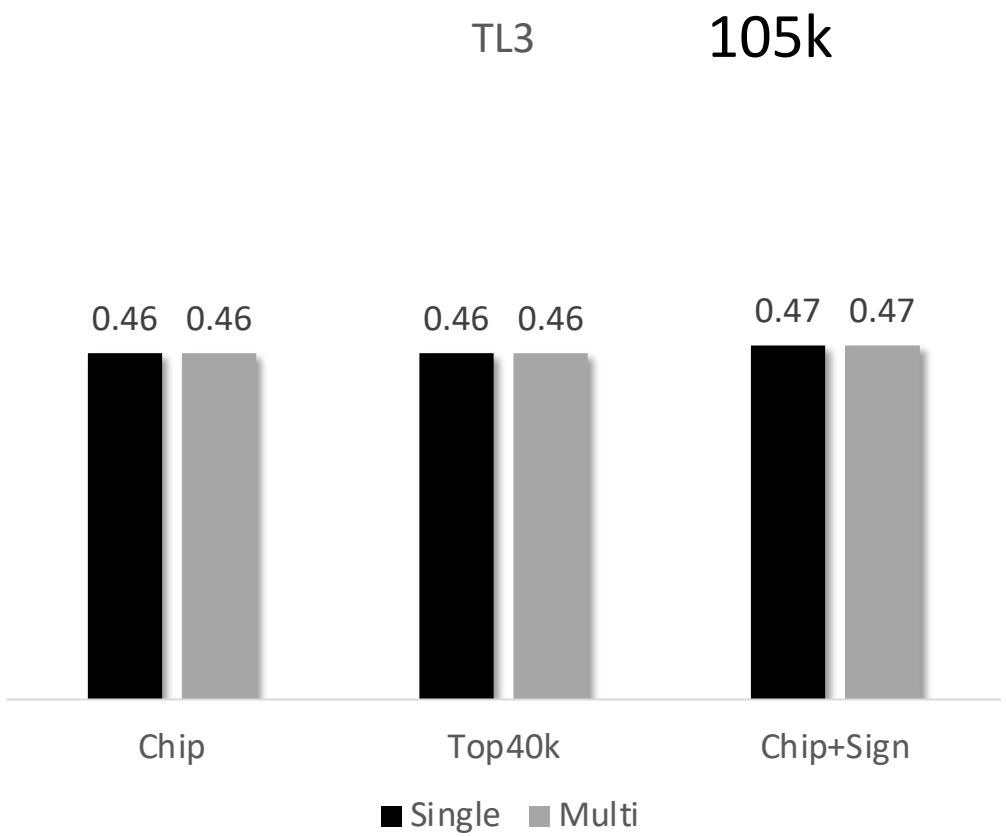
- Prediction accuracy = cor(DEBV, GEBV)

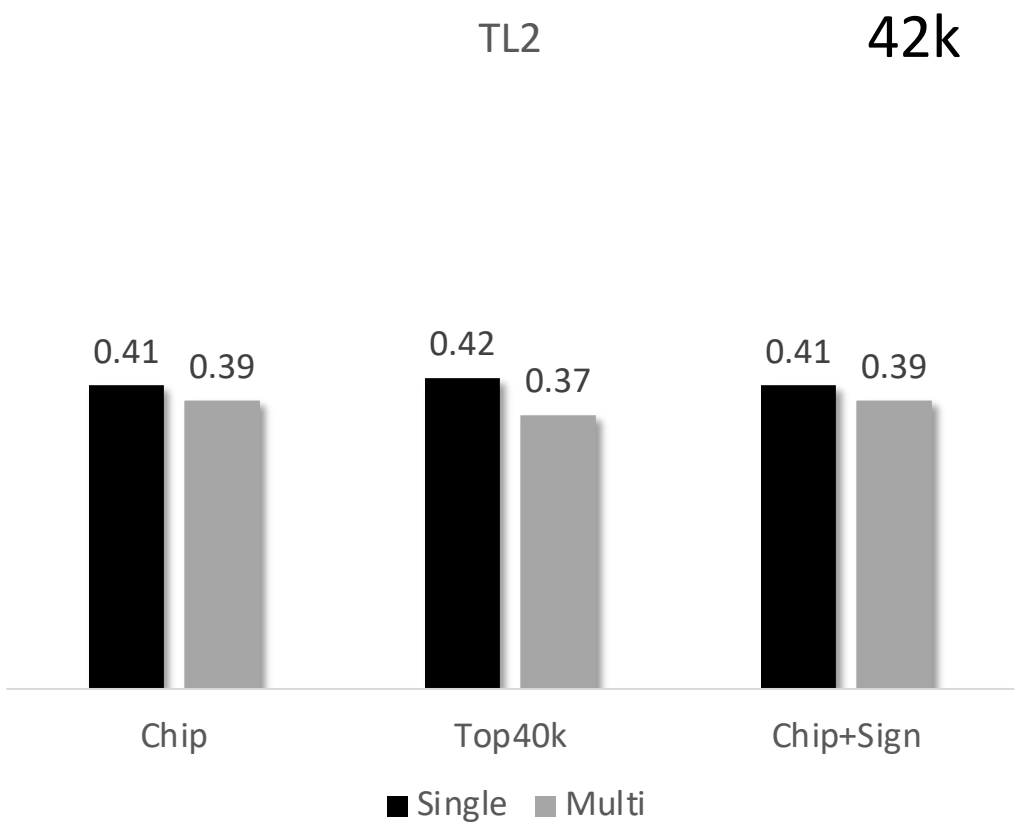


Step 2 – Single vs. Multi-line all traits

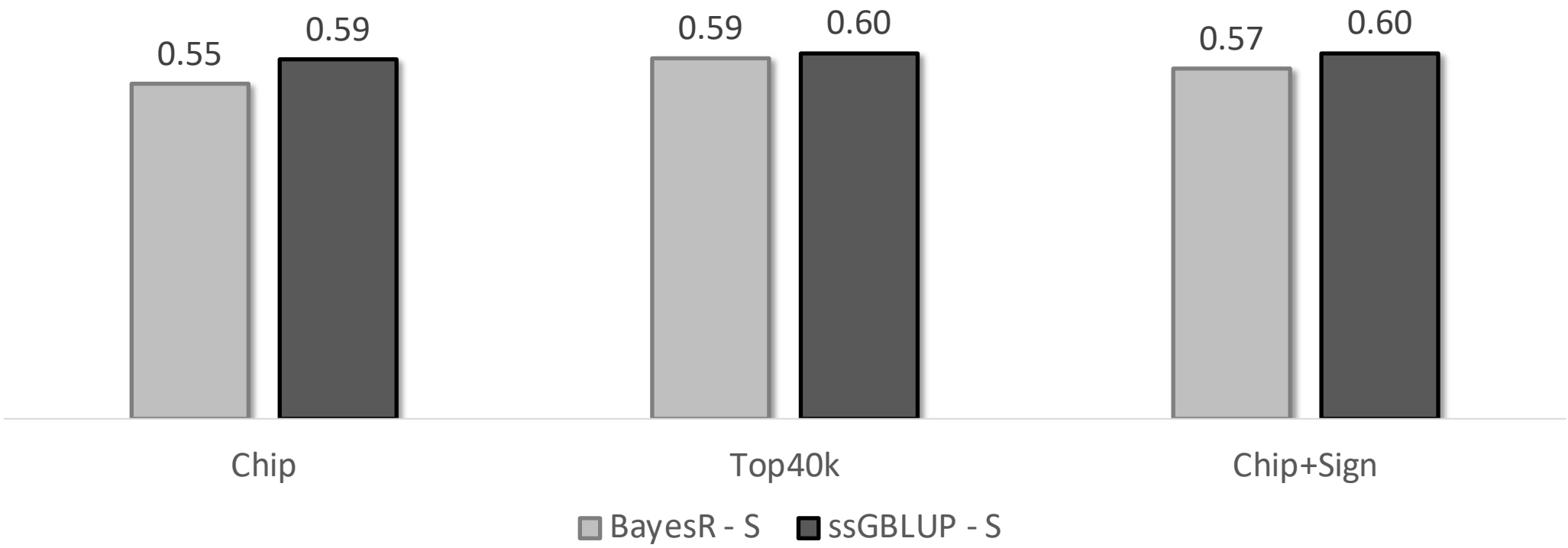
- Prediction accuracy = cor(DEBV, GEBV)

Multi-line GWAS and predictions dominated by TL3





Step 3 - ssGBLUP vs. BayesR



BayesR from Ros-Freixedes et al., 2022

ssGBLUP vs. BayesA in Dairy

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

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

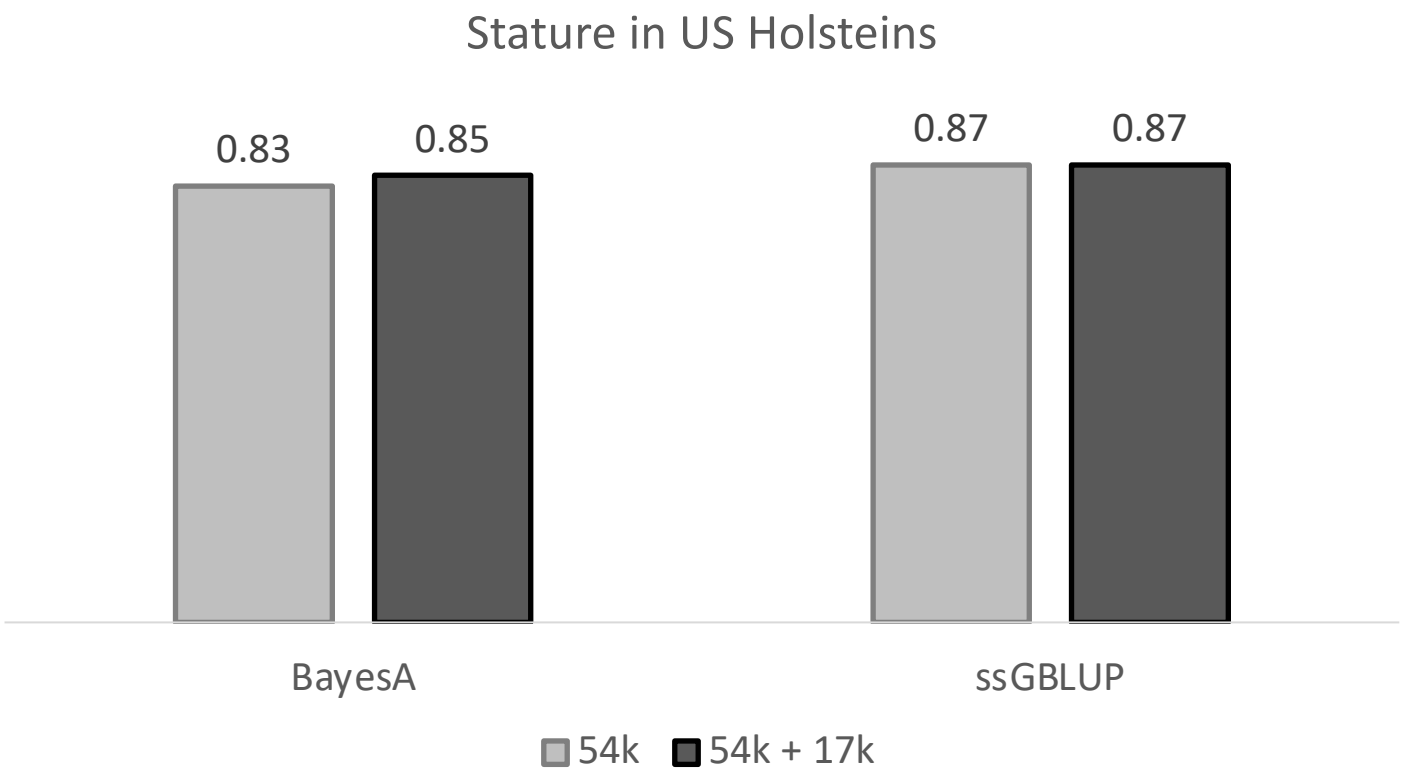
Selecting sequence variants to improve genomic predictions for dairy cattle

Paul M. VanRaden[1*], Melvin E. Tooker[1], Jeffrey R. O'Connell[2], John B. Cole[1] and Derek M. Bickhart[1]

J. Dairy Sci. 102:10012–10019
https://doi.org/10.3168/jds.2019-16262

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. Misztal[2]
[1]Department of Animal Science, University of Connecticut, Storrs-Mansfield 06269
[2]Department of Animal and Dairy Science, University of Georgia, Athens 30602
[3]Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326
[4]Animal 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)

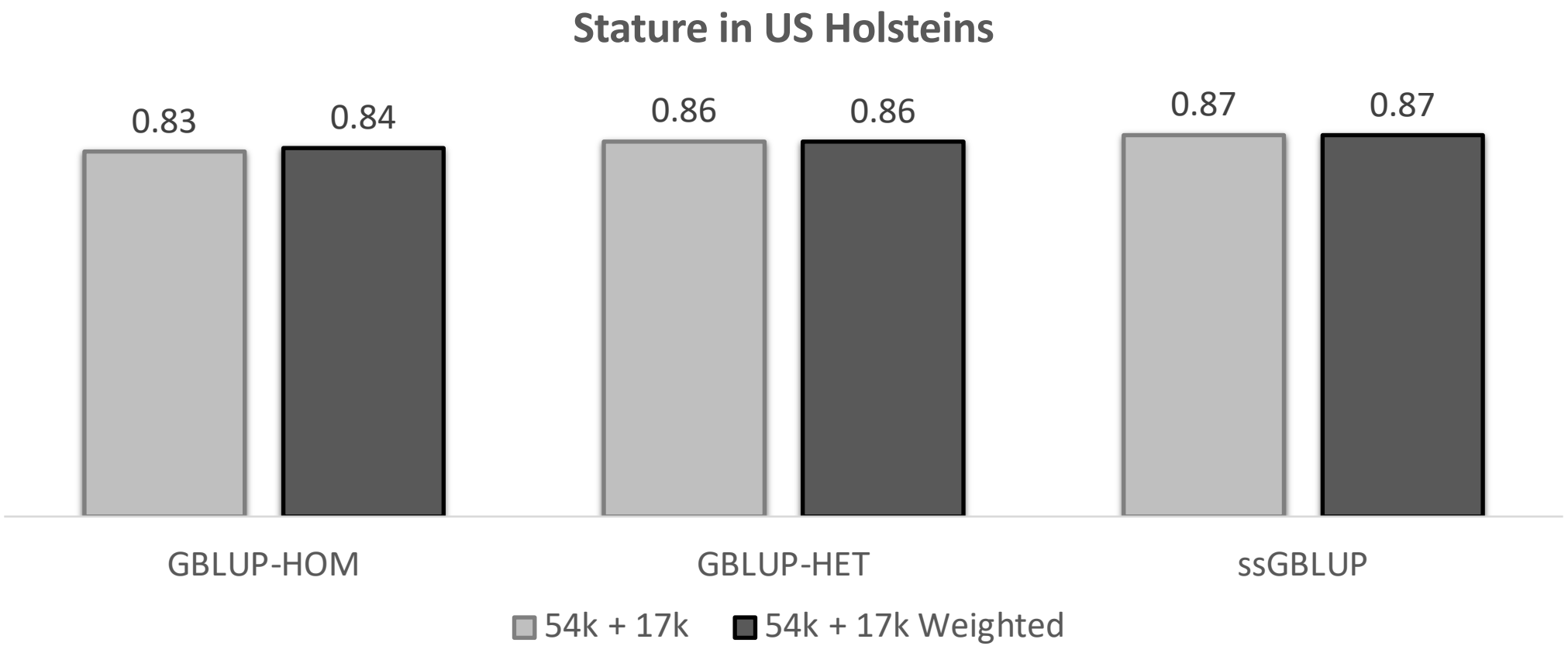
$$\mathbf{G} = \mathbf{ZDZ}'q$$

Different weights for SNP

J. Dairy Sci. 102:10012–10019
https://doi.org/10.3168/jds.2019-16262

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. Misztal[2]
[1]Department of Animal Science, University of Connecticut, Storrs-Mansfield 06269
[2]Department of Animal and Dairy Science, University of Georgia, Athens 30602
[3]Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326
[4]Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705



How many SNP do we need?

- How many SNP?
- How many genotyped individuals?

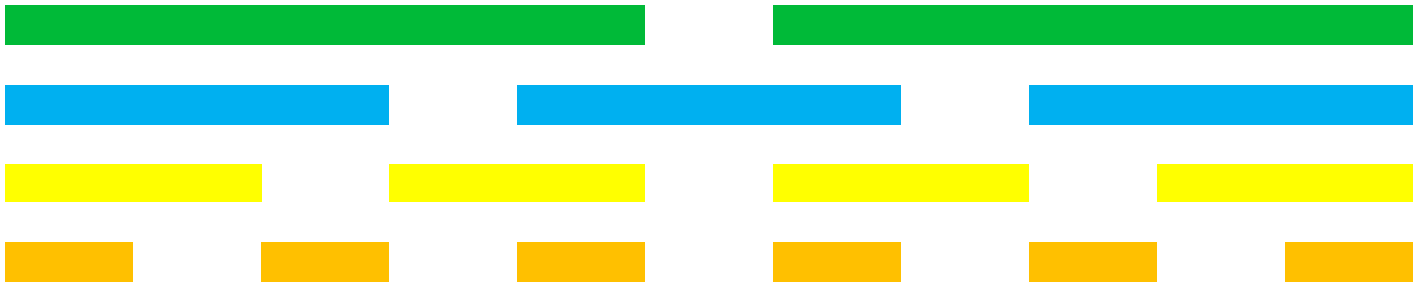
→ Dimensionality of Genomic information

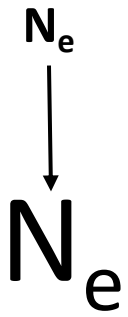
Theory of junctions Fisher (1949)

$E(Me) = 4N_eL$ Stam (1980)

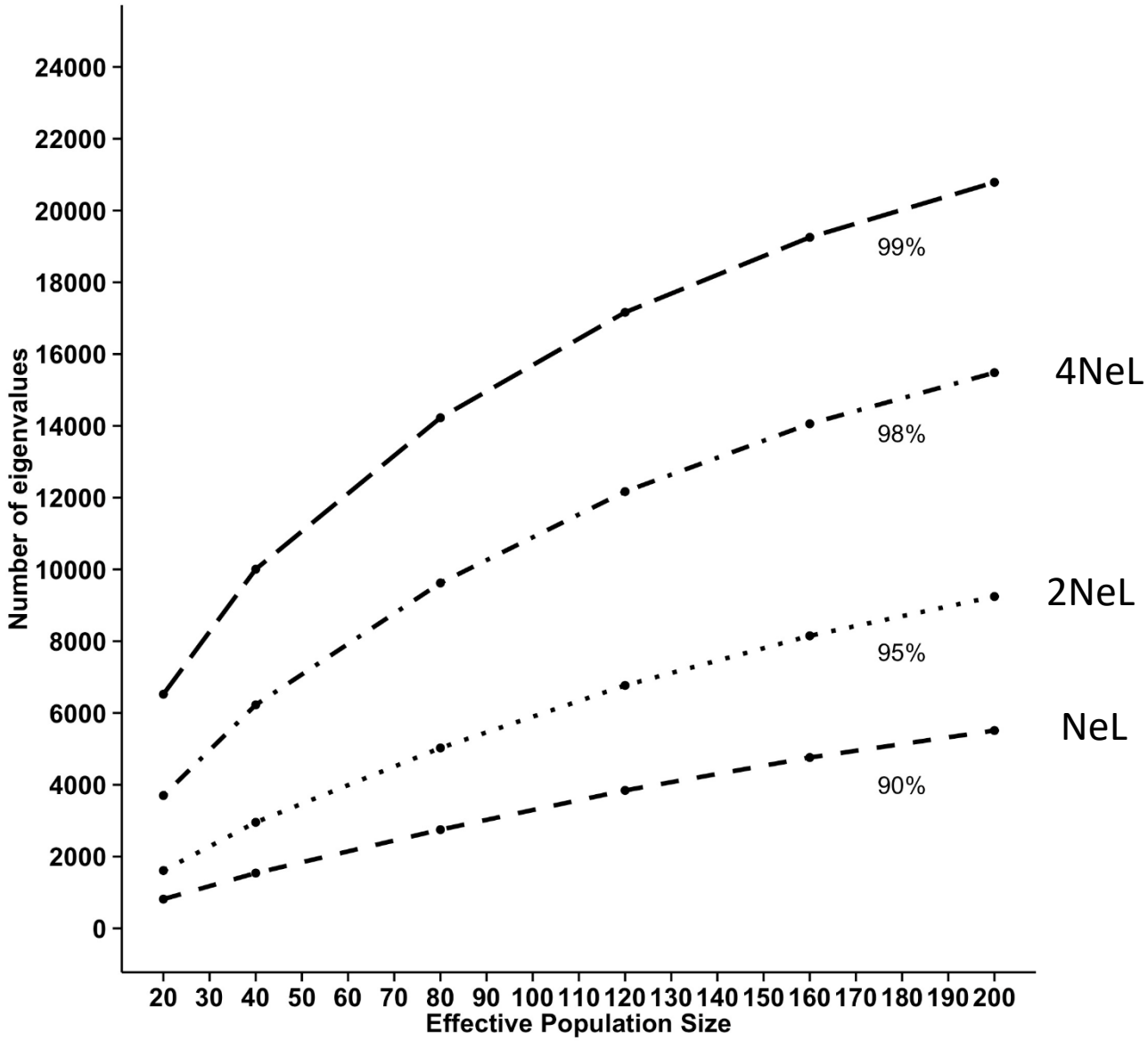
Points where the founder chromosome of origin changes

Me – Independent chromosome segments
N_e – Effective population size
L – Length of genome in Morgans





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'Elevage, F-31326 Castanet-Tolosan, France

Dimensionality of genomic information

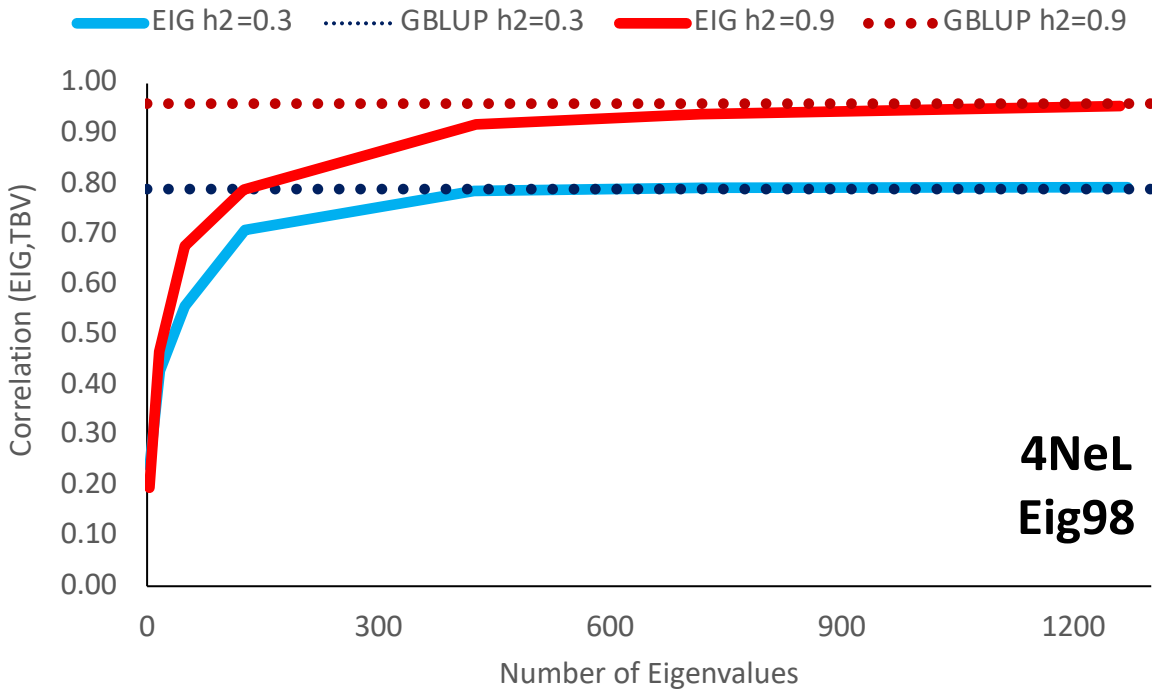
Pocrnic et al. (2016b)
Pocrnic et al. (2018)
Pocrnic et al. (2019)

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



GENETICS | GENOMIC SELECTION

The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

Ivan Pocrnic,*,† 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'Elevage, 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

GSE Genetics Selection Evolution

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

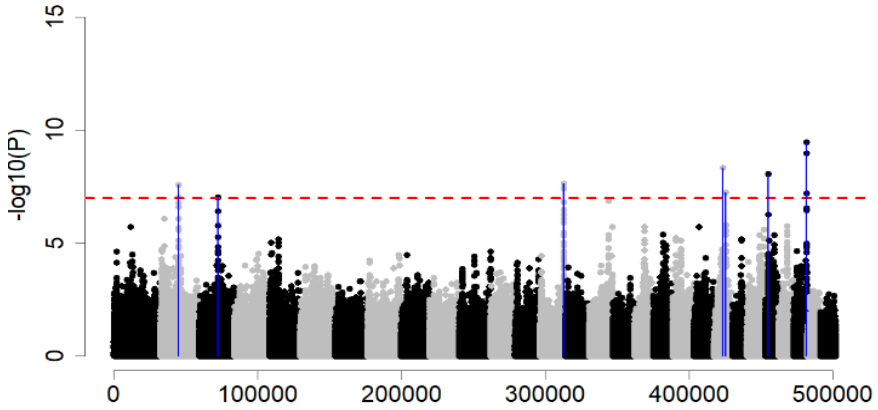
Ivan Pocrnic, Daniela A. L. Lourenco, Yutaka Masuda and Ignacy Misztal

Why small gains in accuracy with sequence?

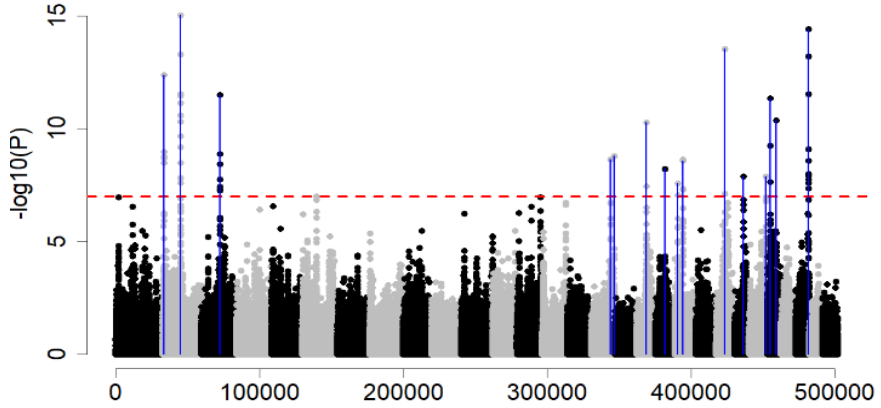
- Amount of information to identify causative variants

Ne=200 QTN=2000

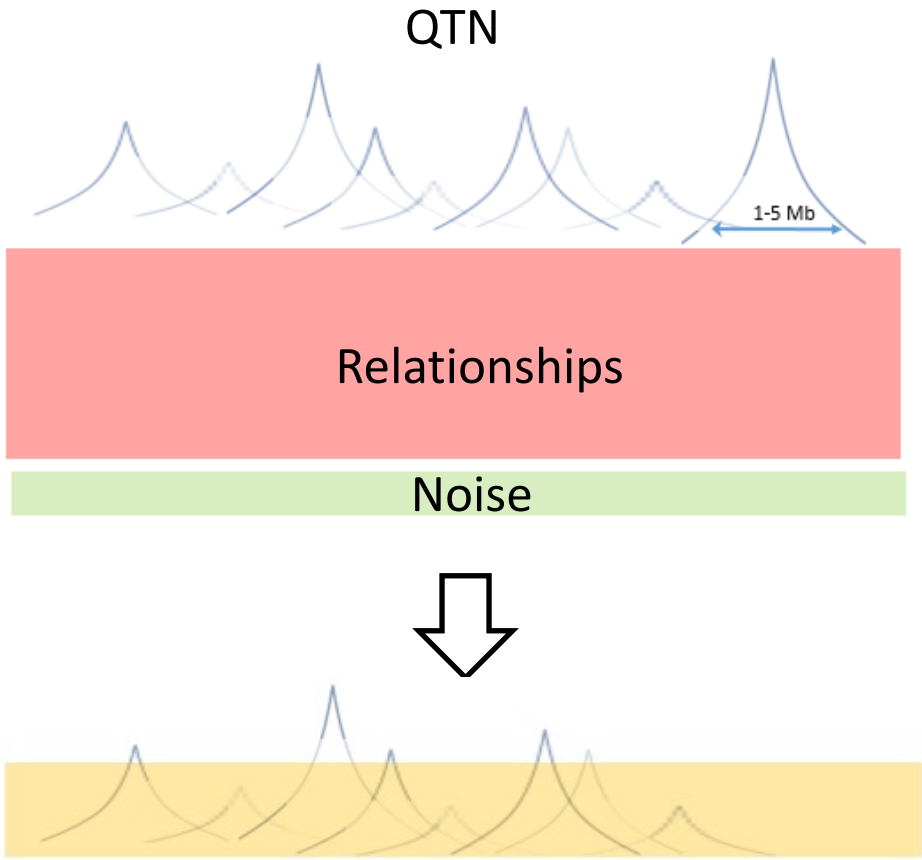
Eig98 N=15,200



All N=30,000



Jang et al.
(under review)

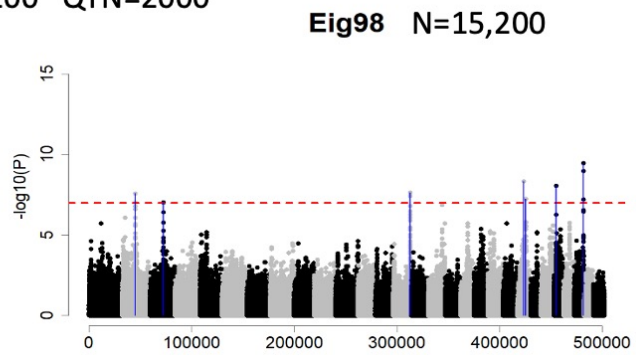


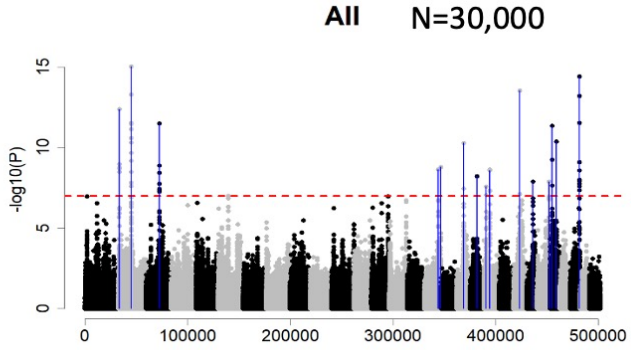
Why small gains in accuracy with sequence?

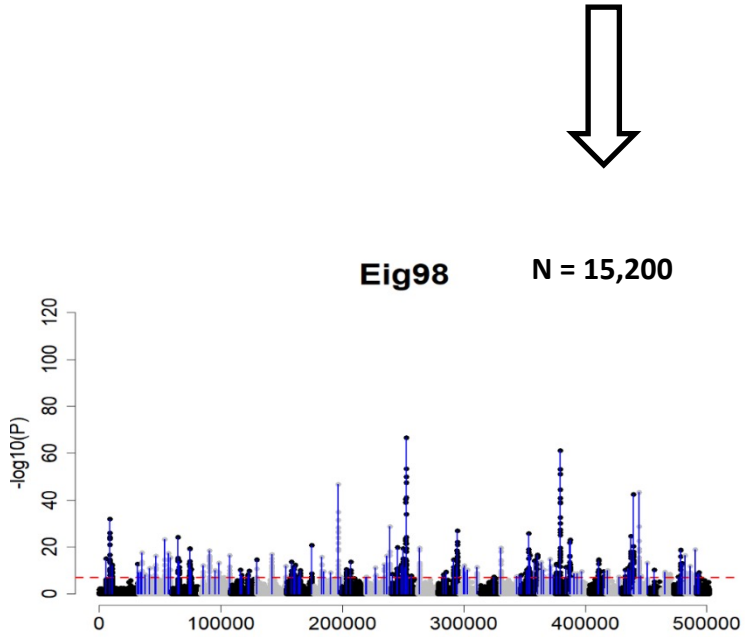
Jang et al.
(under review)

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

Ne=200 QTN=2000

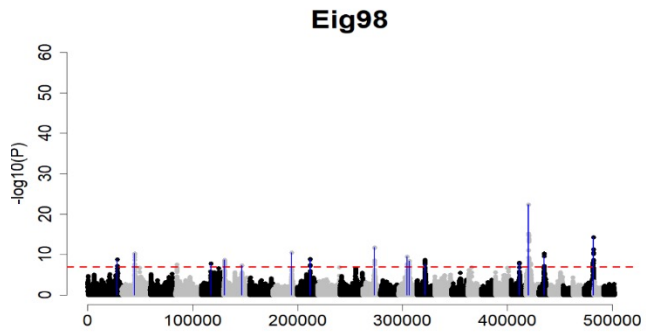






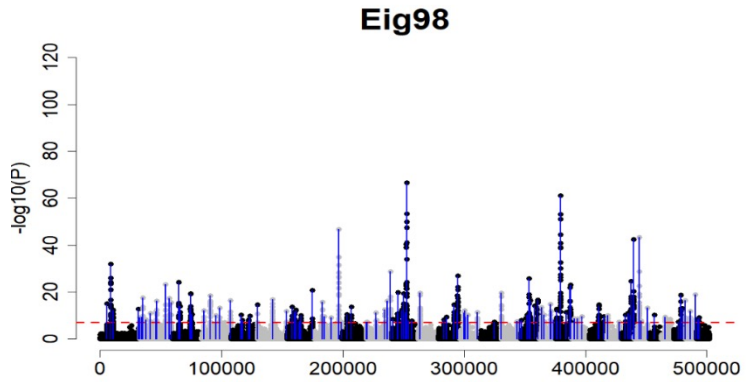
Lots of records for each genotyped animal

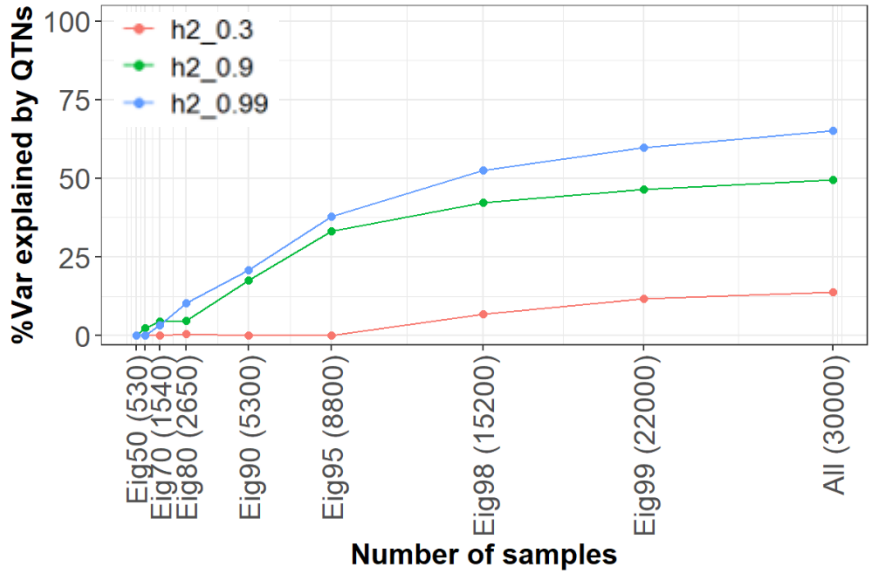
Ne = 20



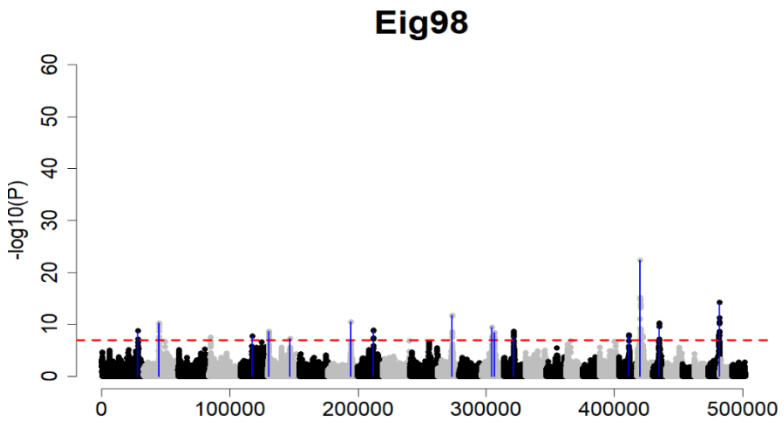
Amount of data in GWAS

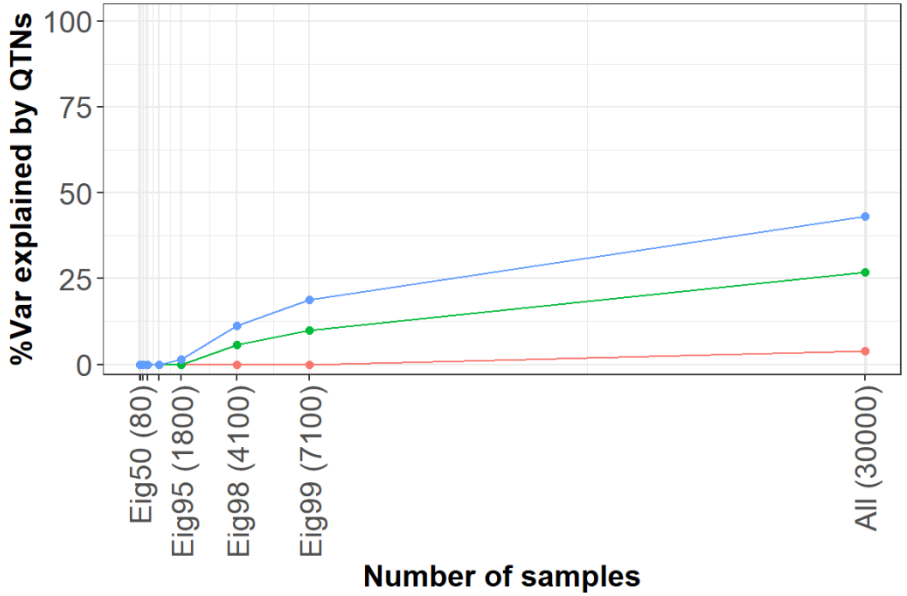
Ne = 200





Ne = 20

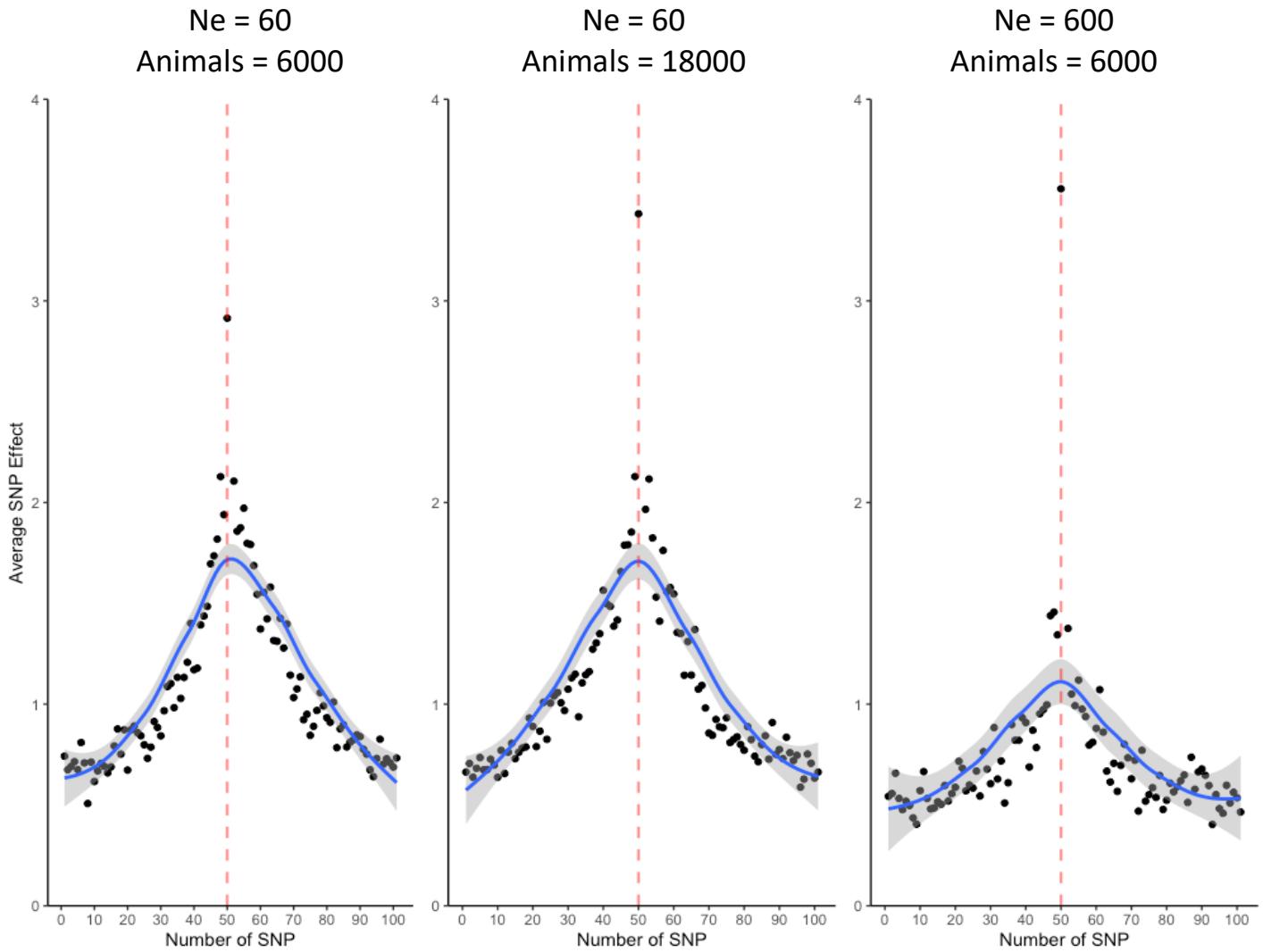




Amount of data in GWAS

- Simulated population (10 QTN per CHR)

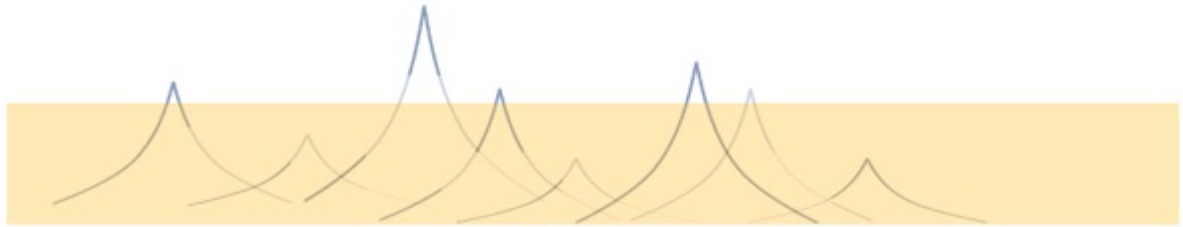
Pocrnic et al.
(in preparation)



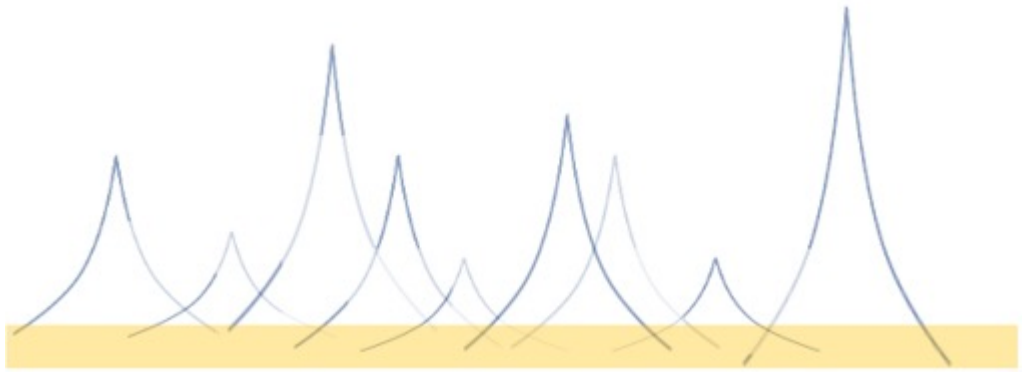
Amount of data in GWAS

Composite Manhattan plot

Pocrnic et al.
(in preparation)



Small populations



Large populations

Single-step GWAS

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

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W}+\mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix} \qquad \mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

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

$$\mathrm{Var}(\mathbf{u}) = \mathbf{H}\sigma_u^2$$

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

$\mathbf{u}$ = GEBV

$\mathbf{u} = \mathbf{Za}$

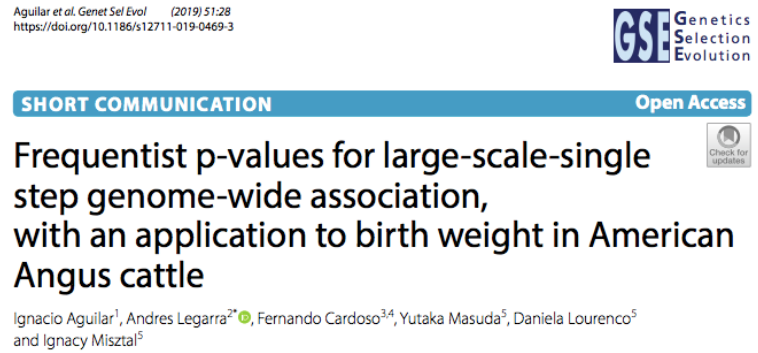
- $\mathbf{a}$ → Vector of SNP effects
- $\mathbf{Z}$ → Matrix of gene content

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

- Backsolve PEV of GEBV into PEV of SNP $\quad Var(\hat{a}_i) = k\mathbf{z}_i'\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{u_2u_2})\mathbf{G}^{-1}\mathbf{z}_i k$

Single-step GWAS

- P-values for SNP effects
- Heavy computations
  - Inverse of LHS of MME
  - limit of ~ 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

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]

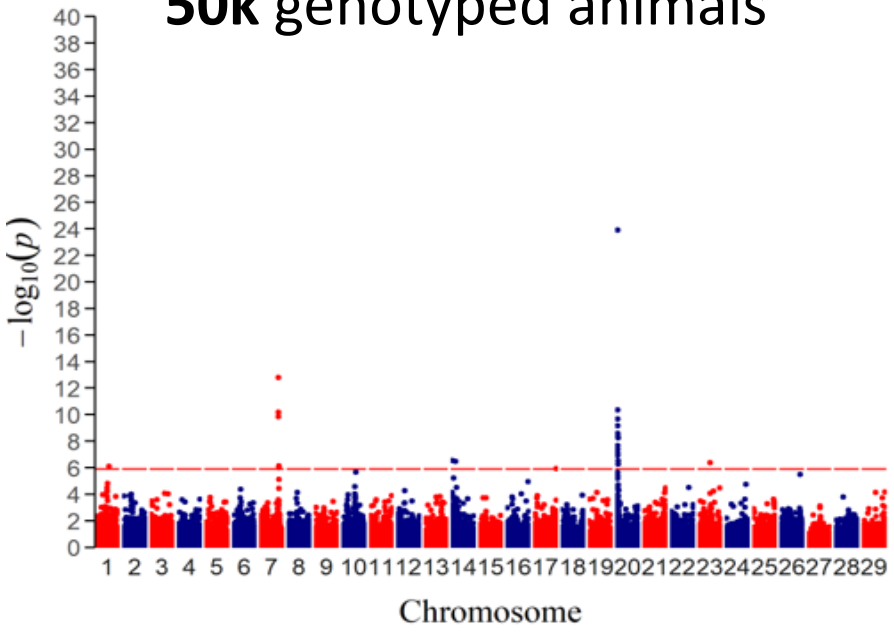
Accounting for all information in GWAS

Natalia Leite

Matias Bermann

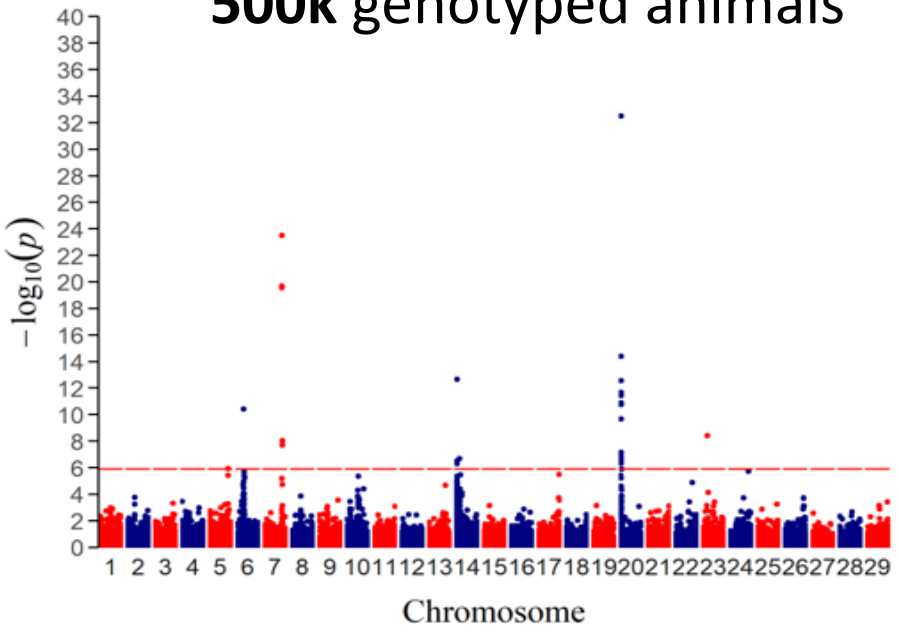
Regular ssGWAS – limitation ~ 50k

50k genotyped animals



ssGWAS based on dimensionality

500k genotyped animals



Identifying causative variants

Ability to identify

true causative variants

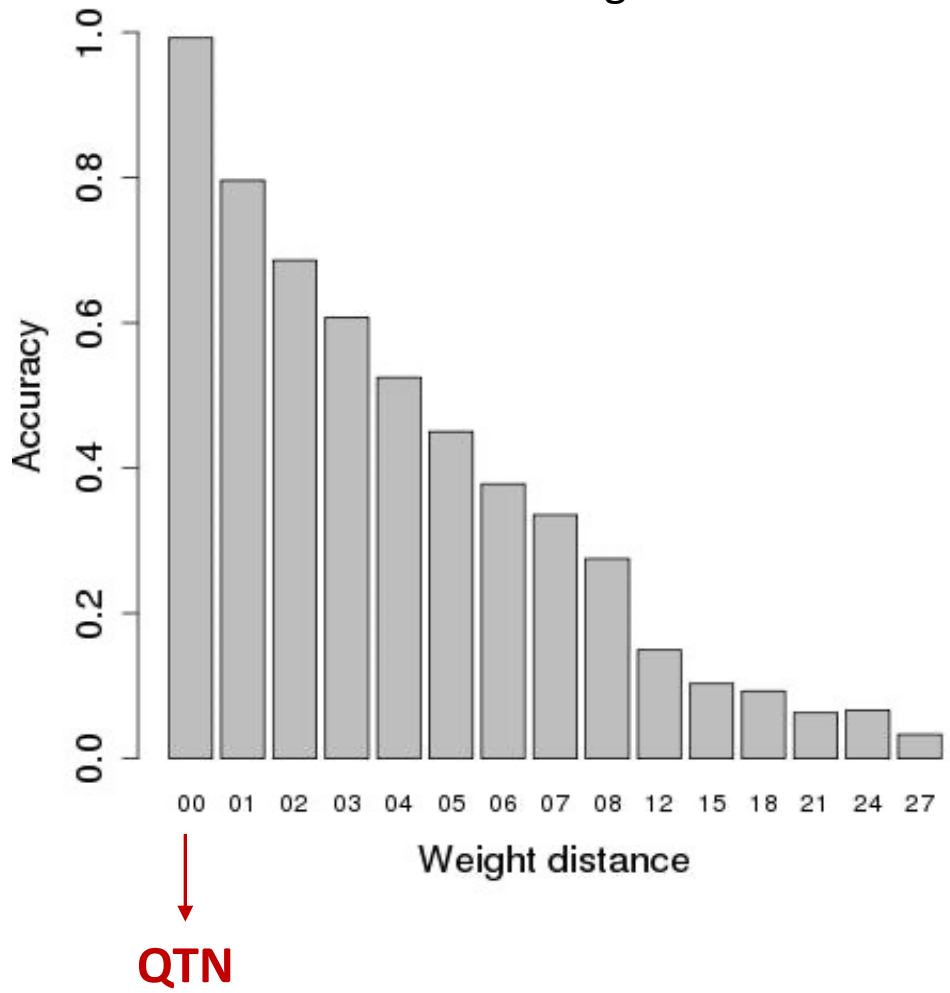
providing independent information

Drive the increase in

accuracy of genomic predictions

Causative variants – simulated data

60k SNP + 100 QTL
simulated weights



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

Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Incorporation of causative quantitative trait nucleotides in single-step GBLUP

Breno O. Fragomeni[1*], Daniela A. L. Lourenco[1], Yutaka Masuda[1], Andres Legarra[2] and Ignacy Misztal[1]

What is next?

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

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

Selecting sequence variants to improve genomic predictions for dairy cattle

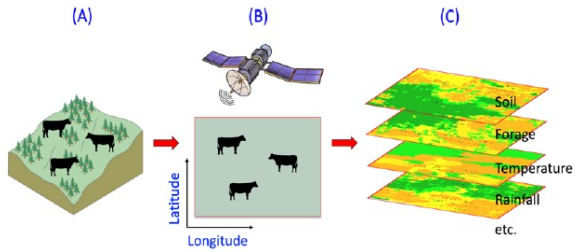
Paul M. VanRaden, Melvin E. Tooker, Jeffrey R. O'Connell, John B. Cole & Derek M. Bickhart

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

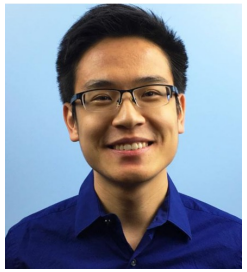
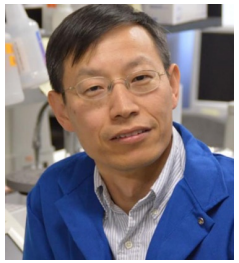
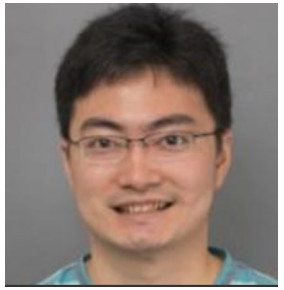
Genetics Selection Evolution **49**, Article number: 59 (2017) | Cite this article

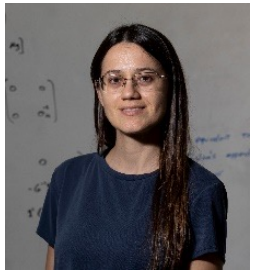


Enviromics-enabled precision breeding for adapted cattle:
Rosa, Lourenco et al. (2022)

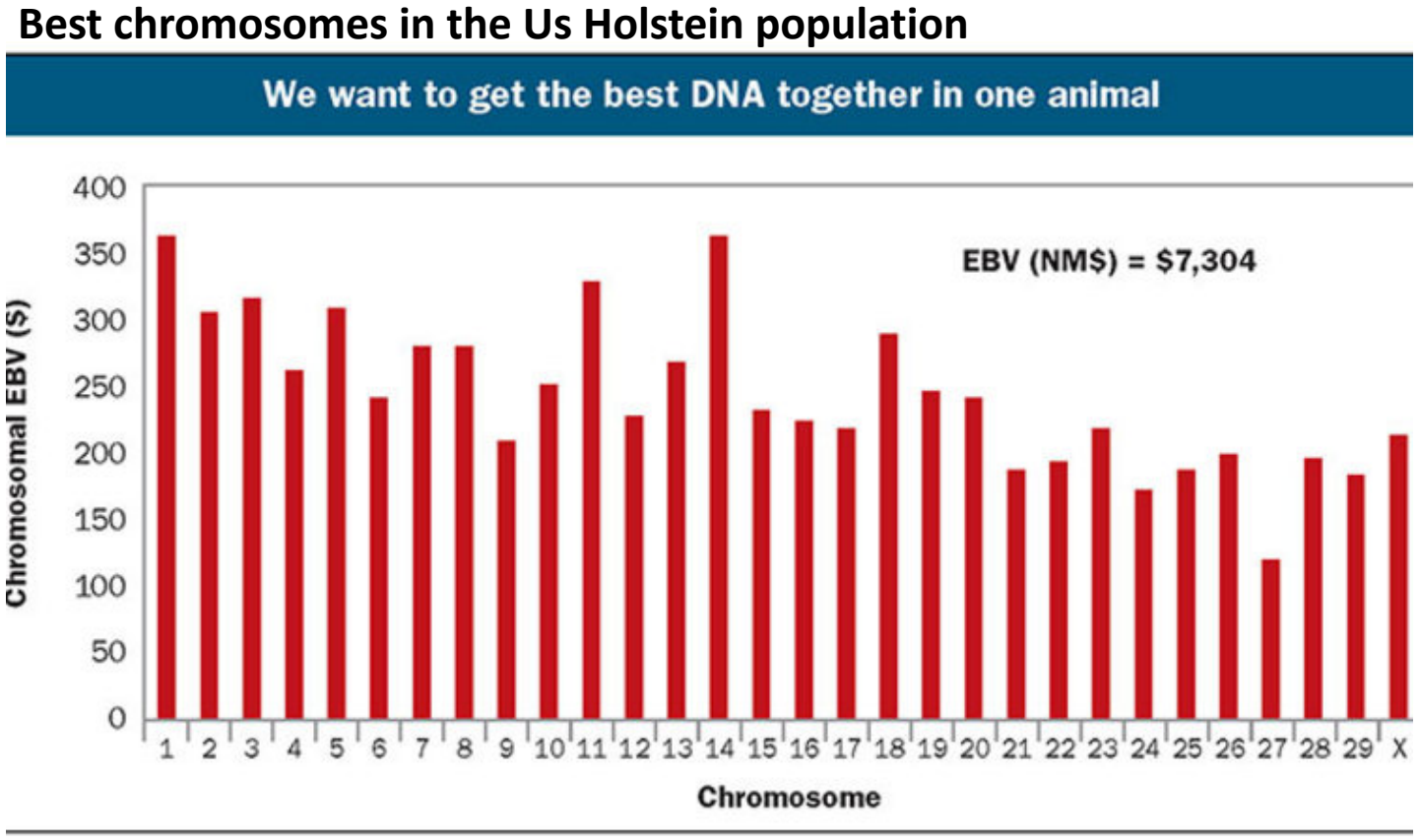
What is next?

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

Have we reached the limit of GS?



John Cole
(2019)

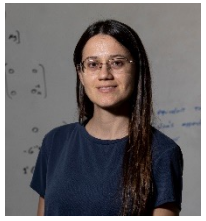
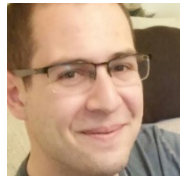
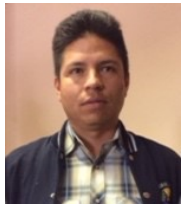
Sum of the effects of SNP
in each chromosome
for each animal

- 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)

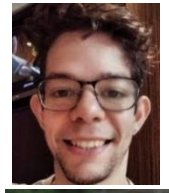
Take home message

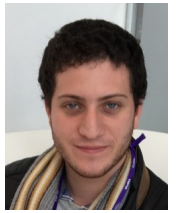
- Using sequence variants for genomic predictions
  - Limited benefit in single and multi-breed populations
  - 0 – 5% gain in accuracy
  - Small or large populations
- Identifying true causative variants is the key
  - Poor job because of the amount of data
  - Identified SNP may be redundant with the 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

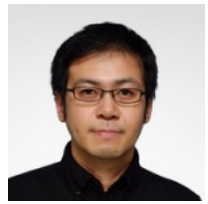













USDA United States Department of Agriculture
Agricultural Research Service
Warmwater Aquaculture Research Unit

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

Rural Development Administration
National Institute of Animal Science

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

UNIVERSITY OF GEORGIA

danilino@uga.edu

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