UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences

Big Data Genomic Analysis in Dairy Cattle

Daniela Lourenco

A. Cesarani, S. Tsuruta, A. Legarra, E. Nicolazzi, P. VanRaden, I. Misztal

January 14, 2023

The fact

3.9 million genotyped animals

Why is it important?

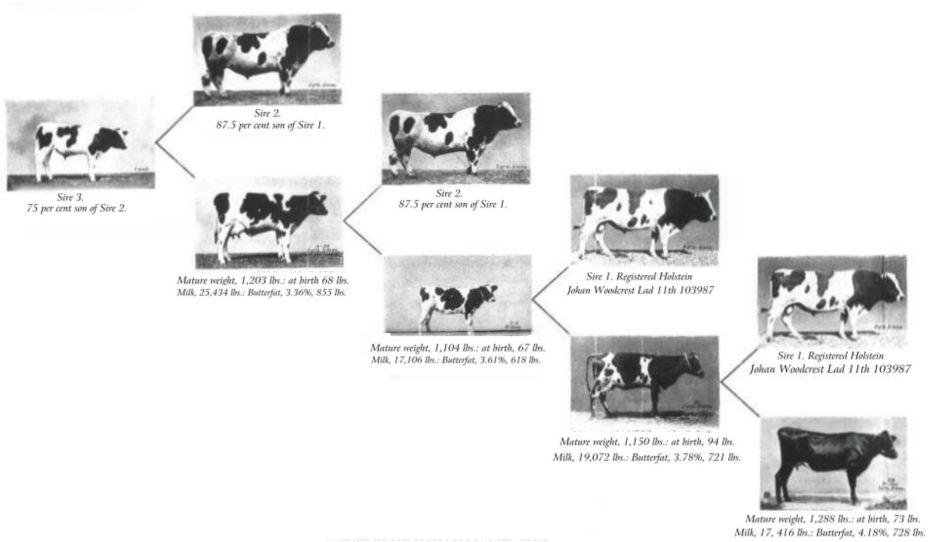
- Impossible in 2015
  - Methods
  - Computing resources vs. algorithms
  - Data availability

- Sends a message
  - Work to accommodate new data
  - Make the most out of available resources

Data

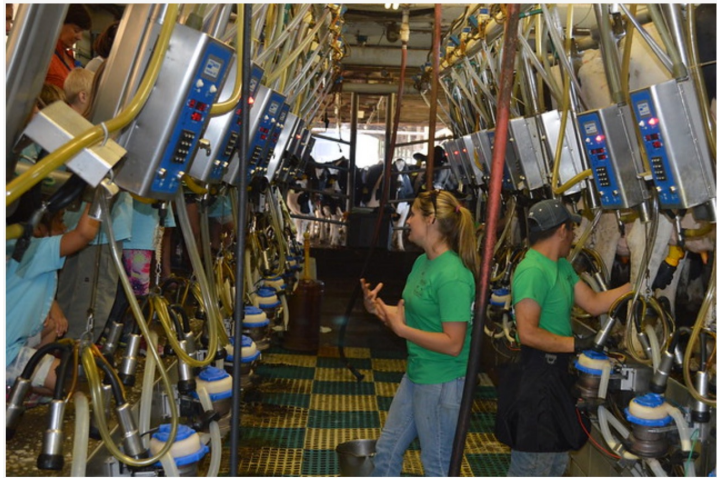
3.9 million

~ 30 million



VanRaden & Miller

45 million



https://www.usda.gov/media/blog/2020/06/18/data-saydairy-has-changed

UNIVERSITY OF GEORGIA

Data

	Phenotypes		Animals	
Breed	**N**	**Cows**	**Genotypes**	**Total**
All	**45M**	**19.4M**	**3.9M**	**29.5M**
Ayrshire	116k	47k	9.2k	94k
Brown Swiss	328k	138k	47k	292k
Guernsey	129k	58k	5k	100k
Holstein	40.3M	17.5M	3.4M	26.6M
Jersey	4.1M	1.7M	427k	2.5M

Objectives

Alberto Cesarani

- ssGBLUP multi-breed genomic evaluations for US dairy cattle
- Reliability of GEBV from multi-breed ≈ single-breed
- Improve software efficiency ⟶ BLUPF90

Single-step (ssGBLUP)

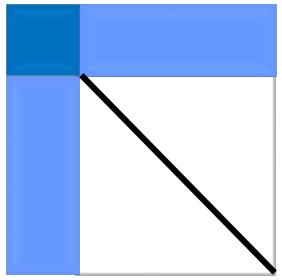
$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

Aguilar et al. (2010)
Christensen and Lund (2010)

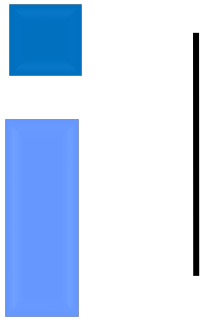
$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix}$$

$\mathbf{G}^{-1}$

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



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



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

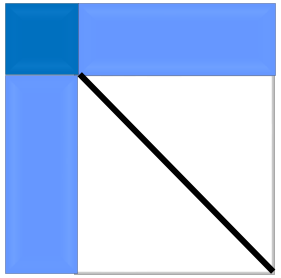
Misztal et al. (2014)

APY

$\mathbf{G}^{-1}$ 

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

Genotyped animals ➔ Core and Noncore

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

Sparse ➔ $u_i|u_{c1} + u_{c2} + u_{c3}, \ldots, u_{ci} = \sum_{j=1}^{c} p_{ij}u_j + \varepsilon_i$

$$\mathbf{G}^{-1}_{\text{APY}} = \begin{bmatrix} \mathbf{G}^{-1}_{cc} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}^{-1}_{cc}\mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}^{-1}_{nn} \begin{bmatrix} -\mathbf{G}_{nc}\mathbf{G}^{-1}_{cc} & \mathbf{I} \end{bmatrix}$$

Misztal et al. (2014)
Fragomeni et al. (2015)
Lourenco et al. (2015)
Masuda et al. (2016)

core animals = # eigenvalues of **G** explaining ≥ 98% variance

- Cattle 10k to 15k
- Pigs and chicken 4k to 8k

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

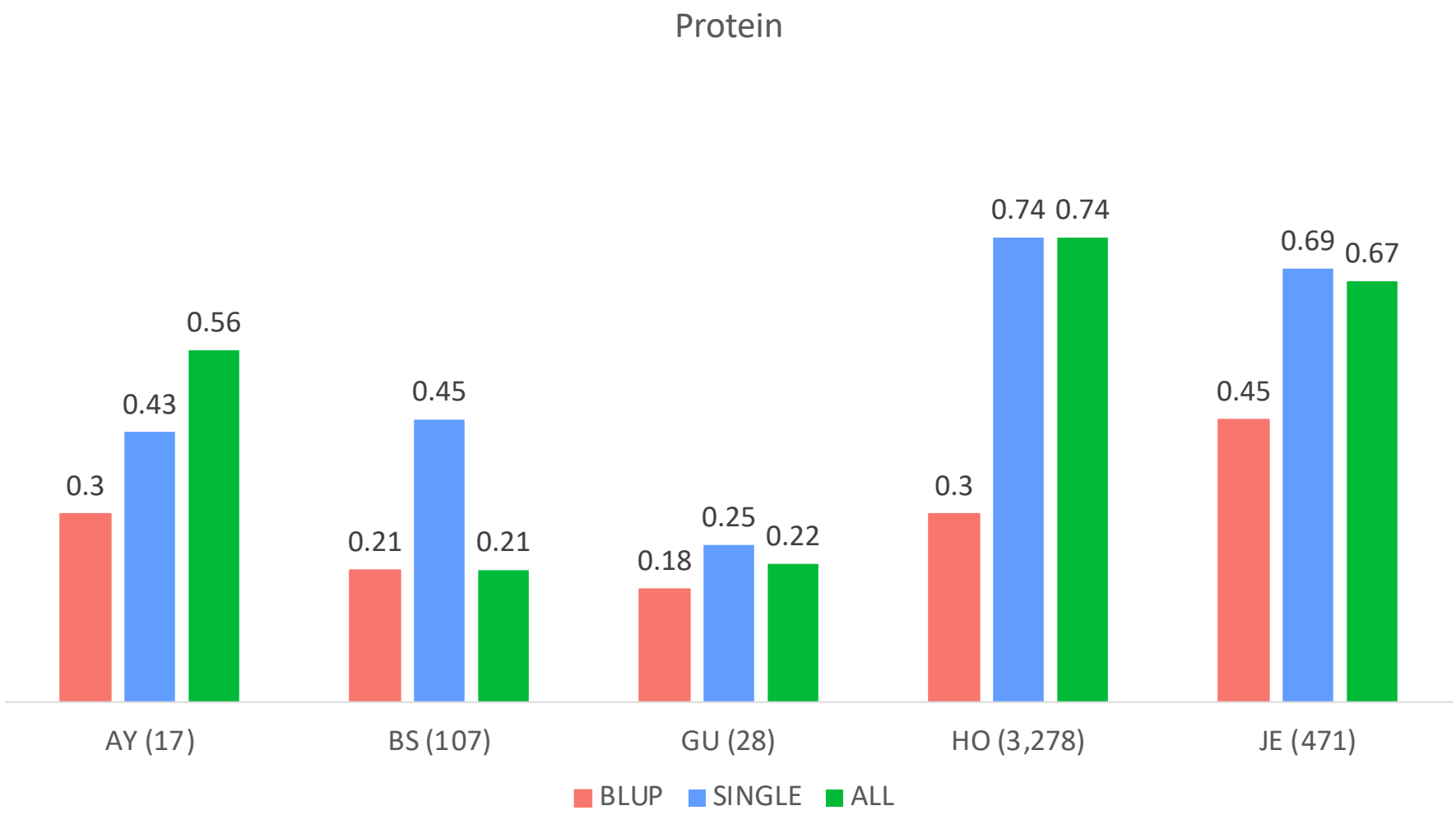
Evaluations

- **SINGLE BREED**
  - each breed separately
  - HO and JE: APY ssGBLUP with 15k random core animals

- **ALL BREEDS**
  - five breeds together
  - breed-specific effects
  - 15k random core animals in APY

Milk (MY), fat (FY), and protein (PY) yields recorded from January 2000 to June 2020

R^2 for bulls



Core animals in ssGBLUP ALL

AY = 32

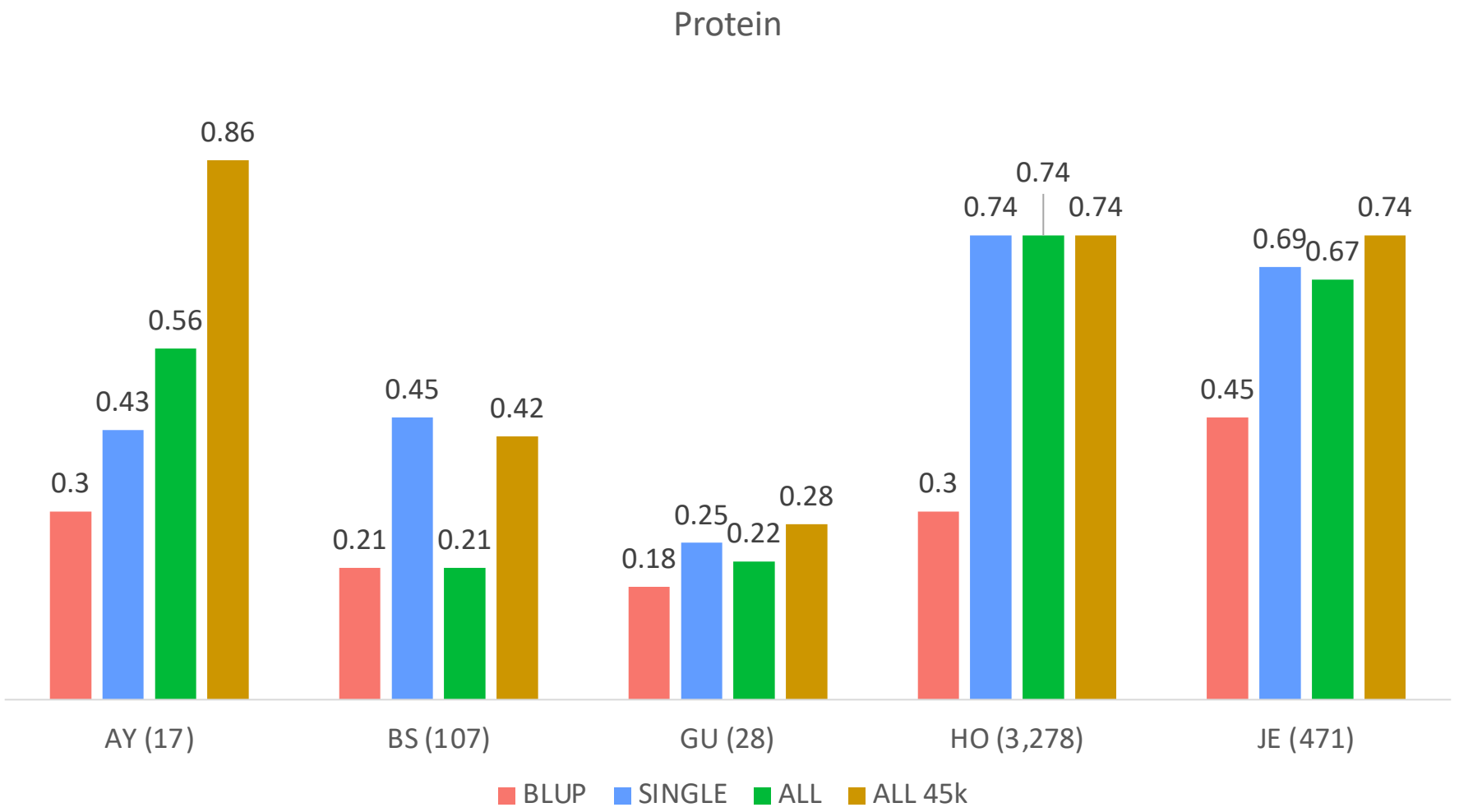
BS = 182

GU = 17

HO = 13k

JE = 1.7k

R^2 for bulls



ALL 45k

AY = 5k

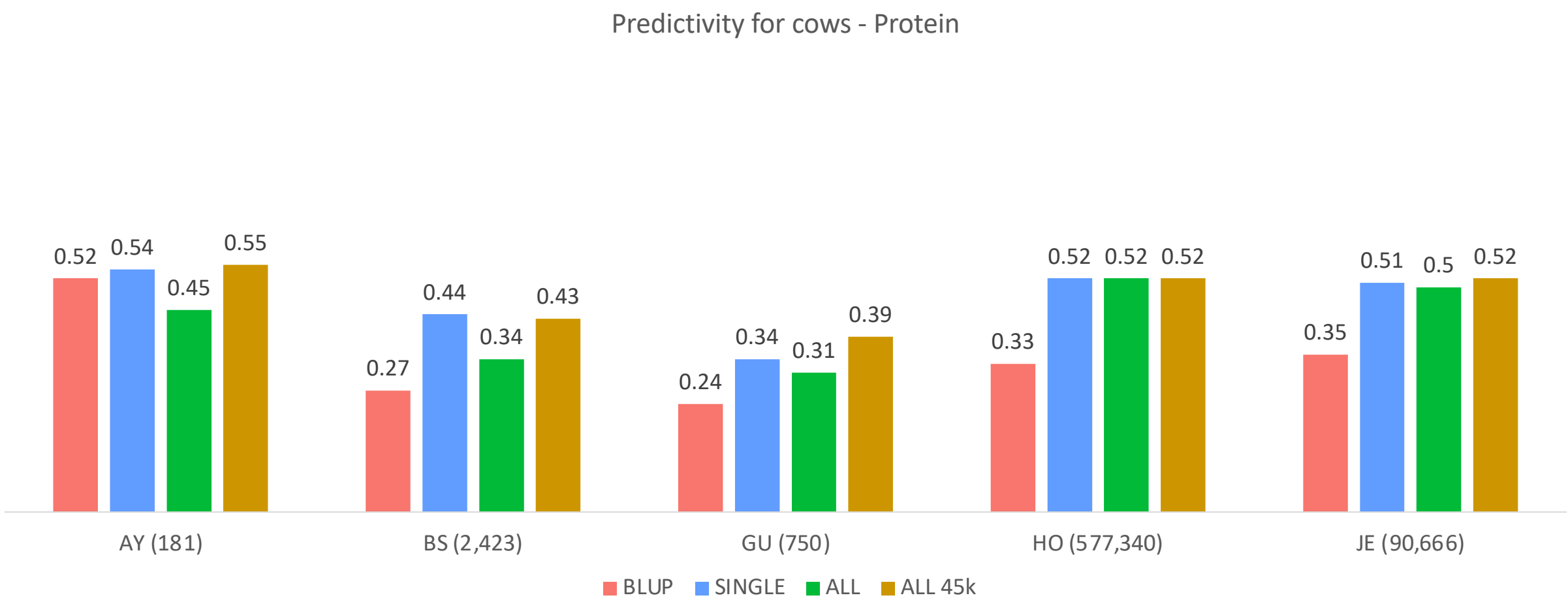
BS = 5k

GU = 5k

HO = 15k

JE = 15k

Predictivity for cows



Computing time

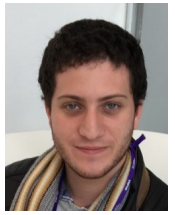
	Rounds		Sec / round		Time	
	BLUP	ssGBLUP	BLUP	ssGBLUP	BLUP	ssGBLUP
AY	504	863	0.08	0.08	< 1 min	~ 1 min
BS	364	867	0.18	0.45	1 min	~ 6 min
GU	345	757	0.07	0.07	< 1 min	< 1 min
HO	457	473	21.25	56.31	2.7 h	7.4 h
JE	586	432	2.00	5.58	~ 20 min	~ 40 min
ALL	643	1,142	27.01	64.84	4.8 h	~ 20 h
ALL_45k		1,763		130.68		~ 64 h

- 2.5 days for solutions
- 5 days for computing $\mathbf{G}_{\text{APY}}^{-1}$ and $\mathbf{A}_{22}^{-1}$ in ALL_45k

Linux server (x86_64) with Intel Xeon E5-2683 2.10 GHz processor with 32 cores

Updates in $\mathbf{A}_{22}$ for blending

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



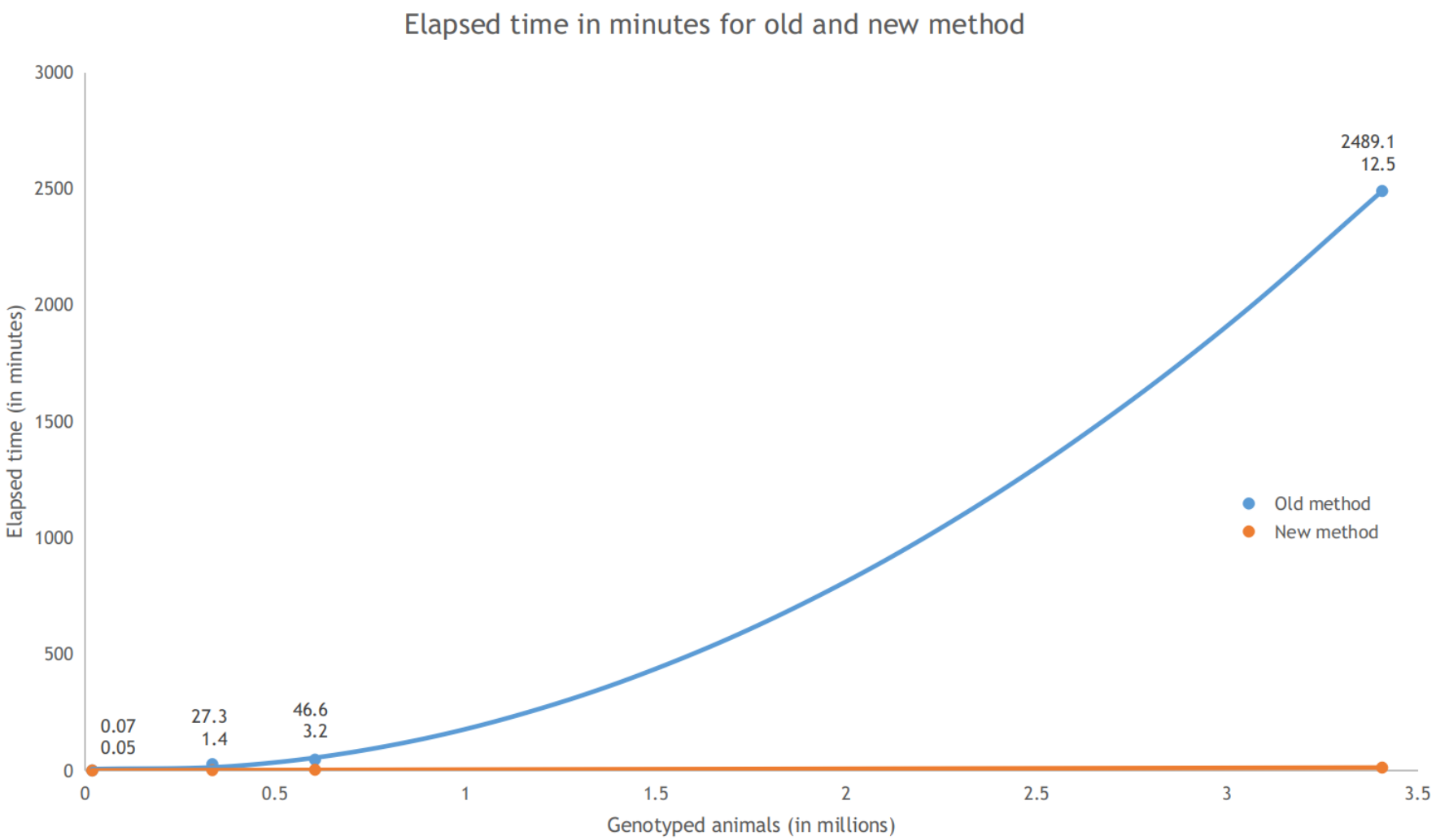
Matias Bermann

$\mathbf{A}_{22}^{-1}$ components: $\mathbf{A}_{22}^{-1}=\mathbf{A}^{22}-\mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}$

APY $\mathbf{G}^{-1}$: $\mathbf{G}_{\text{APY}}^{-1}=\begin{bmatrix}\mathbf{G}_{\text{cc}}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}+\begin{bmatrix}-\mathbf{G}_{\text{cc}}^{-1}\mathbf{G}_{\text{cn}}\\ \mathbf{I}\end{bmatrix}\mathbf{M}_{\text{nn}}^{-1}\begin{bmatrix}-\mathbf{G}_{\text{nc}}\mathbf{G}_{\text{cc}}^{-1} & \mathbf{I}\end{bmatrix}$

- Blending: $\mathbf{G}$ = 0.95 $\mathbf{G}^{*}$ + 0.05 $\mathbf{A}_{22}$
  - Colleau (2002)
  - 3.4M genotyped Holsteins 15k core: 41 hours
  - Rearranging Colleau for core and noncore: 12.5 minutes

Updates in $\mathbf{A}_{22}$ for blending



Alberto Cesarani

- From 5 days to 8 hours to compute $\mathbf{G}_{\mathrm{APY}}^{-1}$ and $\mathbf{A}_{22}^{-1}$ in ALL_45k

Next challenge

Take home message

Large-scale single-step genomic evaluations are feasible for dairy cattle

- All species
- Reasonable computing time
- Single- or multi-breed populations
    - Respect the dimensionality within each bread

Large-scale genomic evaluations are easier said than done

- Always challenged by the increasing amount of data
- Efficient algorithms and computing power

Acknowledgments

Grant no. 2020-67015-31030

Dairy producers who supplied data through their participation in the Dairy Herd Improvement program and Dairy Records Processing Centers that edited and relayed information on to the Council of Dairy Cattle Breeding

http://nce.ads.uga.edu danilino@uga.edu