UNIVERSITY OF GEORGIA

How large-scale genomic evaluations are possible

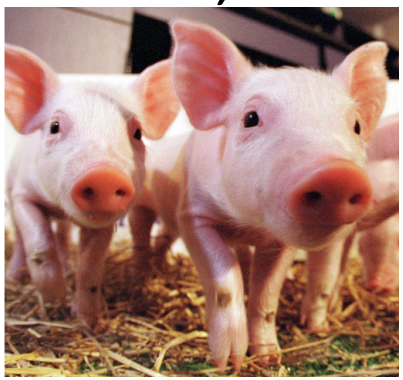
Daniela Lourenco

05-24-2018

How big is your genomic data?

15 Gb

250,000



http://sesenfarm.com/raising-pigs/

17 Gb

255,000

https://www.usjersey.com/AJCA-NAJ-JMS/AJCA/AnimalIdentificationServices/HerdRegister.aspx

26 Gb

500,000

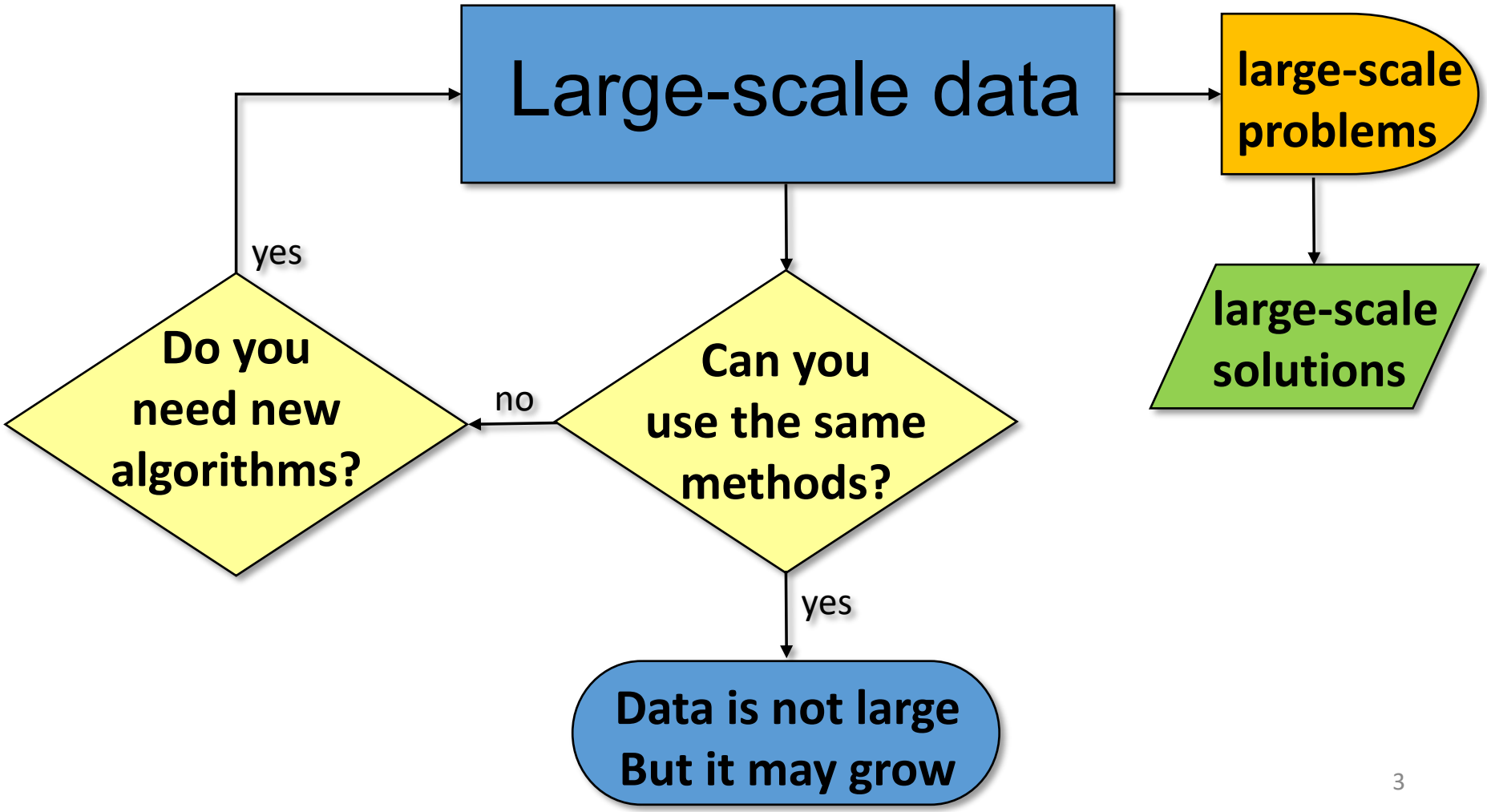
http://www.angus.org/AGI/default.aspx

130 Gb

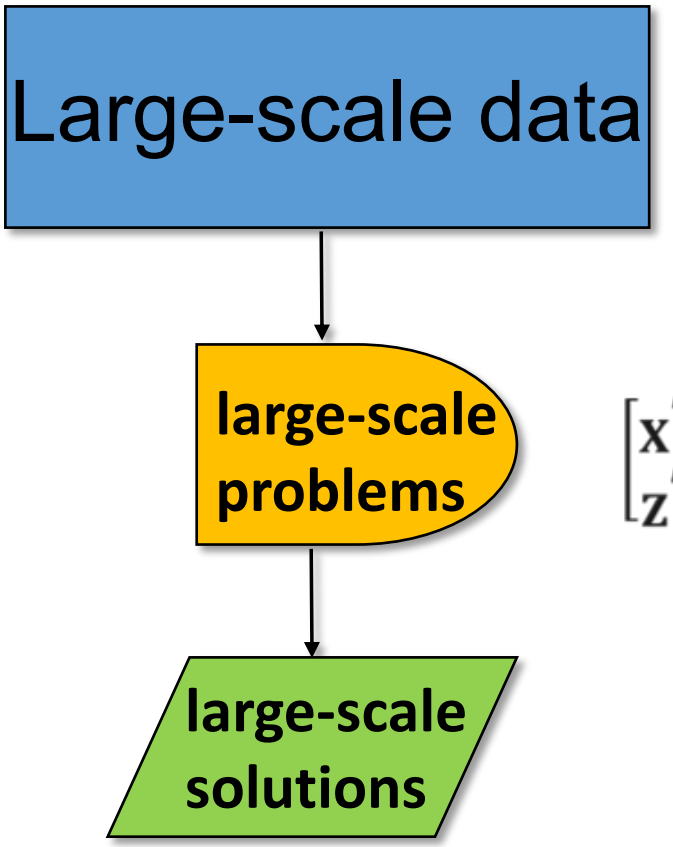
2,000,000

http://www.holsteinusa.net/programs_services/backgrounds.html

Do you have big data?



How large is your genomic data?



- 150,000
- >2 hours
- > 700Gb RAM

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

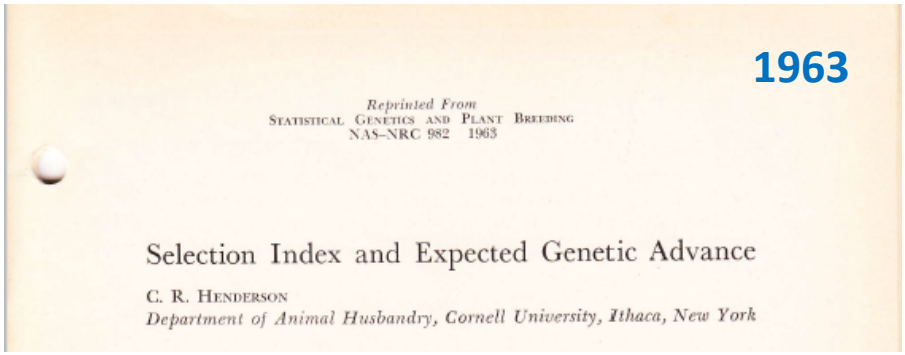
ssGBLUP

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

$$G=\frac{MDM'}{2\sum p_i(1-p_i)}$$

Solution for large-scale evaluations

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$



1963

Reprinted From
STATISTICAL GENETICS AND PLANT BREEDING
NAS–NRC 982 1963

Selection Index and Expected Genetic Advance

C. R. HENDERSON
Department of Animal Husbandry, Cornell University, Ithaca, New York

Problem to invert A back in 1963

BLUP MME was "sleeping"

1976

BIOMETRICS *32*, 69–83
March, 1976

A SIMPLE METHOD FOR COMPUTING THE INVERSE OF A NUMERATOR RELATIONSHIP MATRIX USED IN PREDICTION OF BREEDING VALUES

C. R. HENDERSON
Department of Animal Science, Cornell University, Ithaca, New York 14853, U.S.A.

1988

Research-Article *J. Dairy Sci.*

Additive Genetic Model with Groups and Relationships

R.L. Quaas [1]

Solution for large-scale Genomic evaluations

- Recursions for A^{-1}

$$u_i = 0.5(u_{s_i} + u_{d_i}) + \varphi_i \qquad \mathbf{u} = \mathbf{P}\mathbf{u} + \mathbf{\Phi} \qquad \mathbf{A}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P})$$

Henderson (1976); Quaas (1988)

- Recursions for G^{-1}
  - Split genotyped animals into **core** and **non-core**

$$u_i \mid u_1, u_2, \ldots, u_{i-1} = \sum_{j=\text{core}} p_{ij} u_j + \varepsilon_i \qquad \mathbf{u}_n = \mathbf{P}_{nc}\mathbf{u}_c + \mathbf{\Phi}_n$$

APY - Algorithm for Proven and Young

Misztal et al. (2014)
Misztal (2016)

Algorithm for Proven and Young (APY)

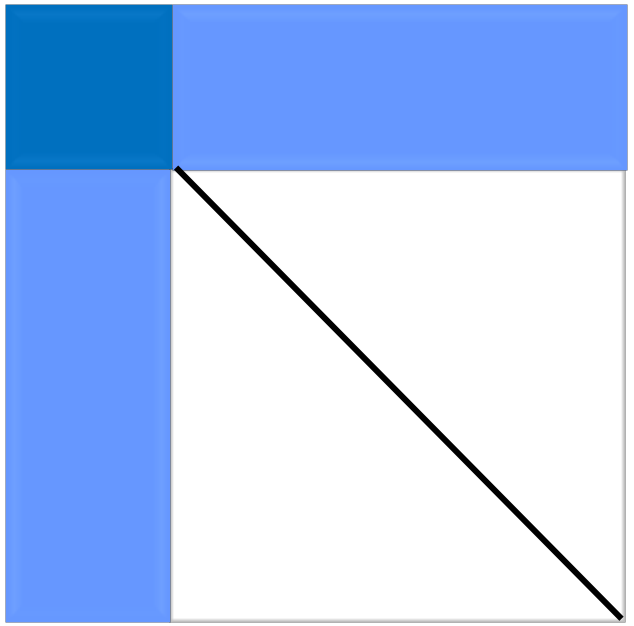
$$\mathbf{G}_{\text{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathrm{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & \mathrm{I} \end{bmatrix}$$

$$\mathrm{M}_{nn} = \mathrm{diag}\{\mathrm{g}_{ii} - \mathrm{g}_{ic}\mathbf{G}_{cc}^{-1}\mathrm{g}_{ci}\}$$

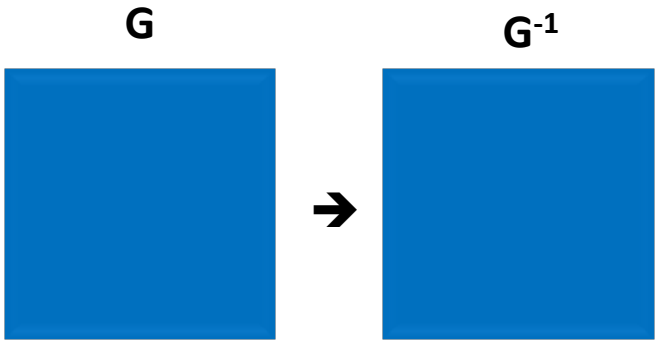
USA_287506
USA_287507
USA_287508
USA_287509
USA_287510
USA_287511
USA_287513
USA_287514
USA_287515
USA_287516
USA_287518
USA_287519
USA_287520
USA_287521
USA_287522
USA_287523
USA_287524
USA_287525
USA_287526
USA_287527
USA_287528
USA_287529
USA_287530

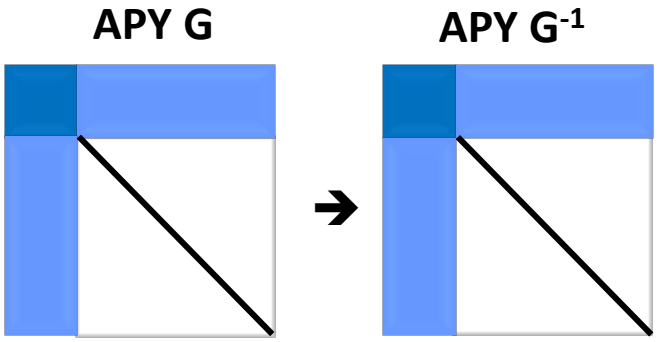
core

non-core



Algorithm for Proven and Young (APY)





- APY $\mathbf{G}^{-1}$ sparse
- Efficient computation
- Why does it work?

APY and dimension of G

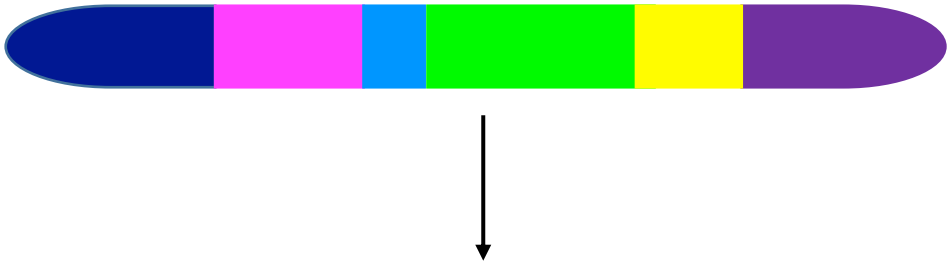
genotyped animals > # SNP

$G = \alpha G + (1-\alpha)A_{22}$ VanRaden (2008)

G has a limited dimensionality
- independent blocks
- Dependent blocks

Dimension of G = min (#animals, # independent SNP, *Me*)

APY and dimension of G



$E(Me) = 4\ Ne L$

Stam (1980)

Dimension of G

Number of core animals

Ne ?

How many core animals in APY?

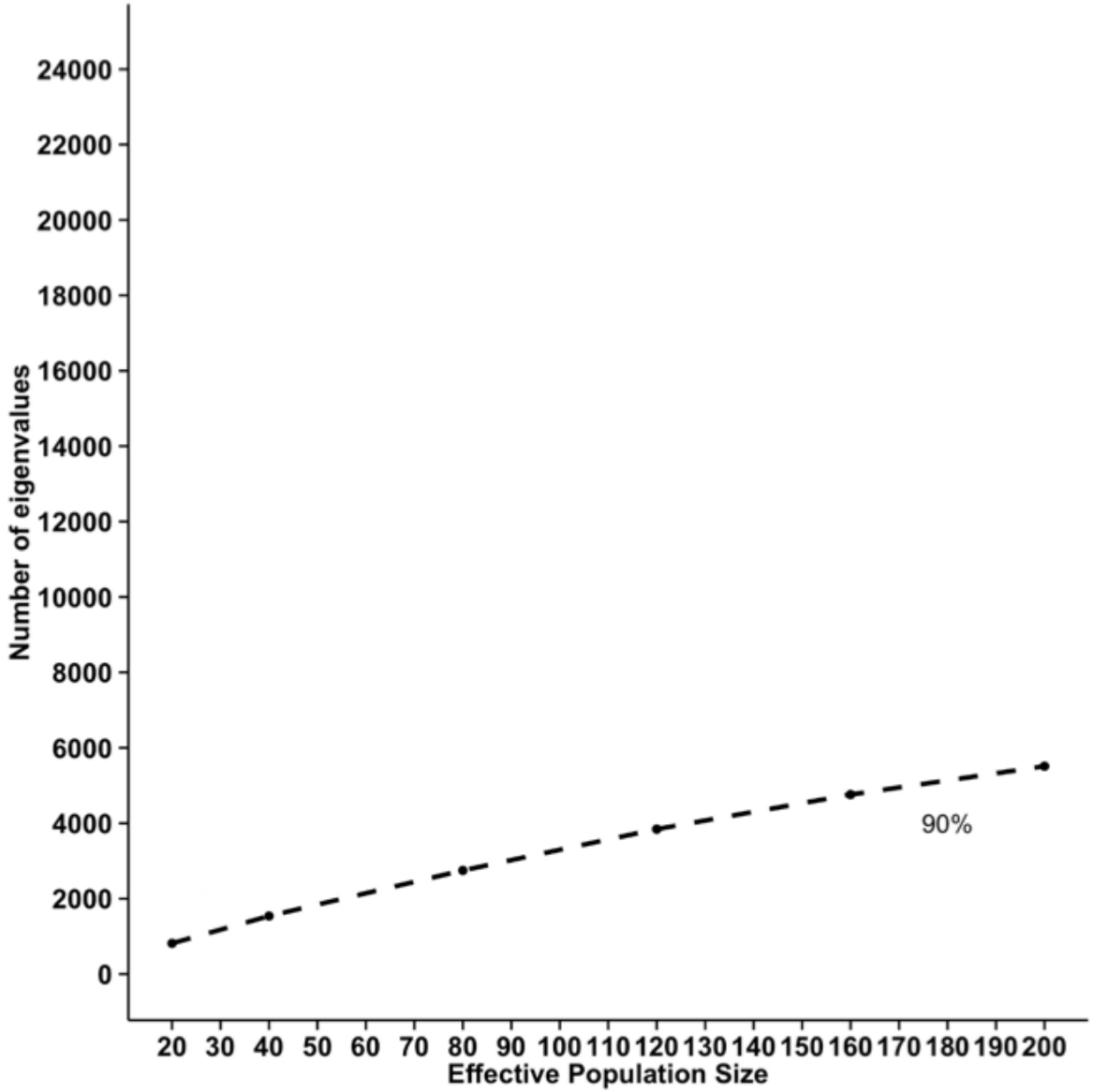
The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

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

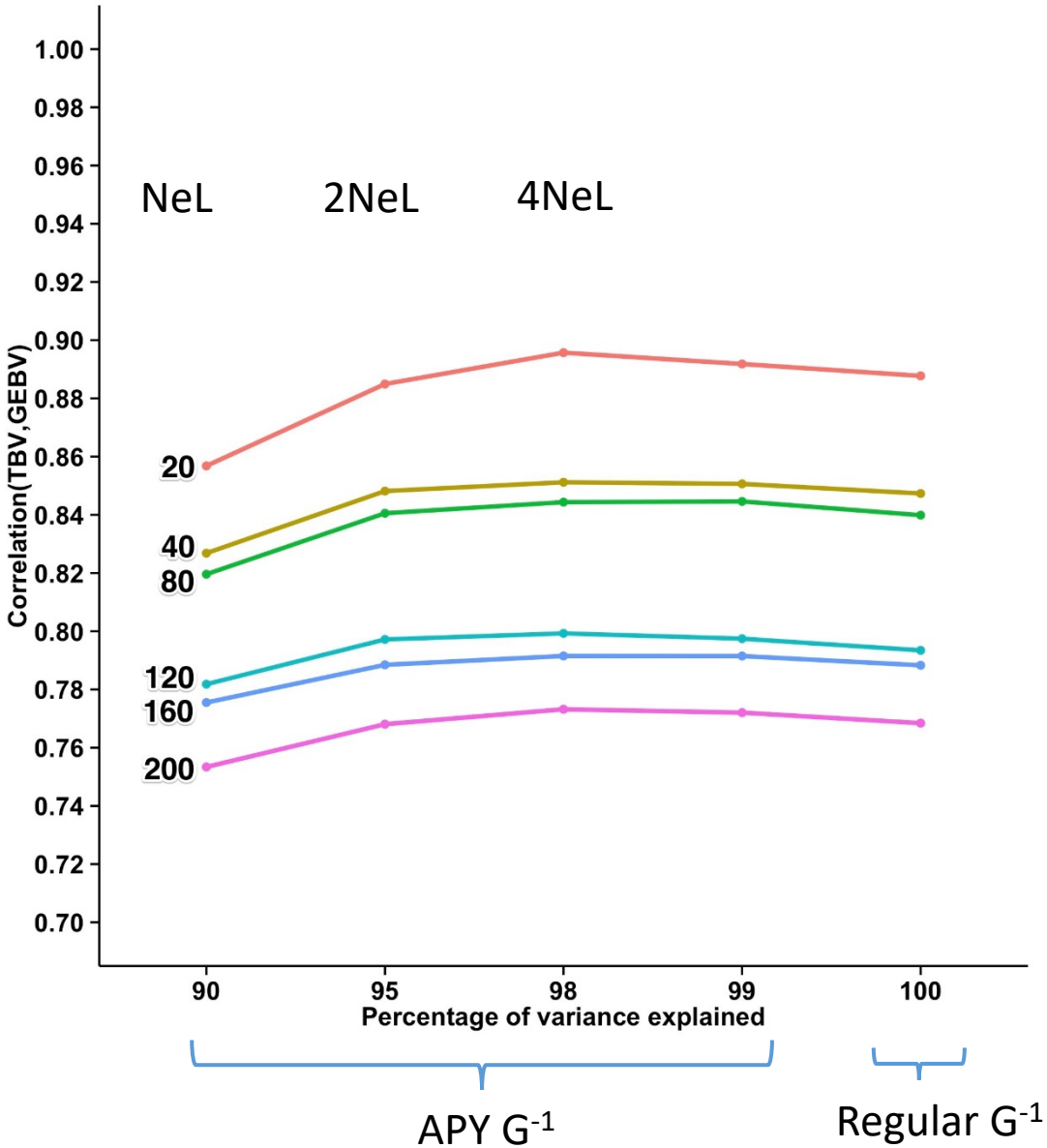
- Simulated populations
- Ne = 20, 40, 80, 120, 160, 200
- #genotyped animals = 75,000

- Dimensionality of G as number of largest eigenvalues of G
  - #eigenvalues vs. Ne
- #eigenvalues as #core animals in APY ssGBLUP
  - cor(TBV,GEBV_APY) vs. cor(TBV,GEBV)

How many core animals in APY?



How many core animals in APY?



How many core animals in APY?

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

Dimensionality of G depends on Ne

→ # largest eigenvalues 98%

↓

Number of core animals

→ ≥ accuracy as regular G^{-1}

How many core animals in APY?

Real Livestock Populations

Pocrnic et al. Genet Sel Evol (2016) 48:82
DOI 10.1186/s12711-016-0261-6

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species

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

77k gen
61k SNP
10M ped

75k gen
61k SNP
2.5M ped

16k gen
39k SNP
200k ped

81k gen
38k SNP
8M ped

23k gen
37k SNP
2.5M ped

How many core animals in APY?

largest eigenvalues of G explaining 98% ~ 99% variance

14k ~ 19k

11k ~ 16k

11k ~ 14k

4k ~ 6k

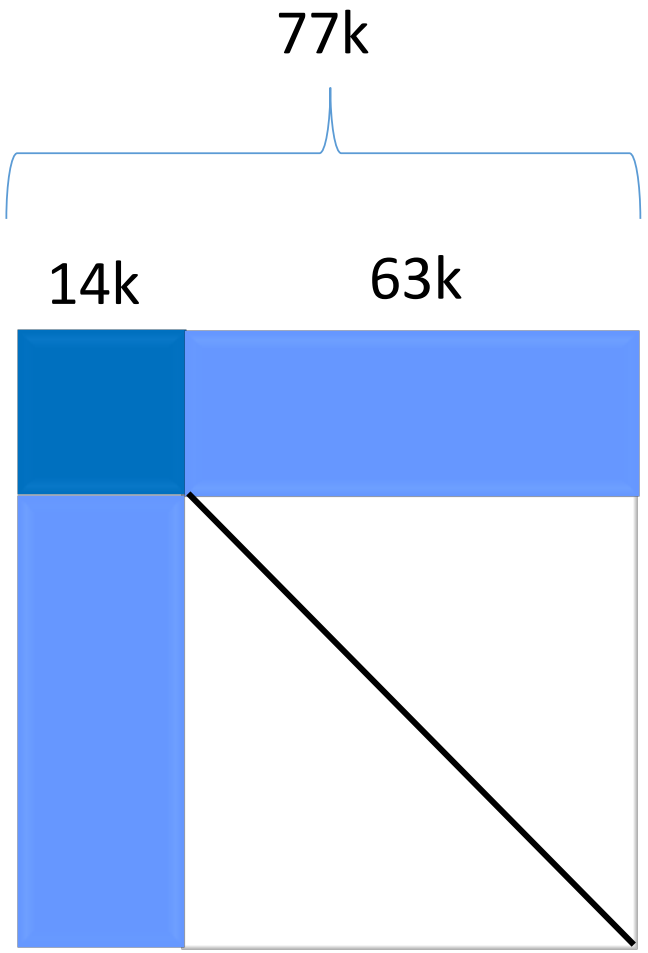
4k ~ 6k

How many core animals in APY?

eigen98%
=
14k

Random choice

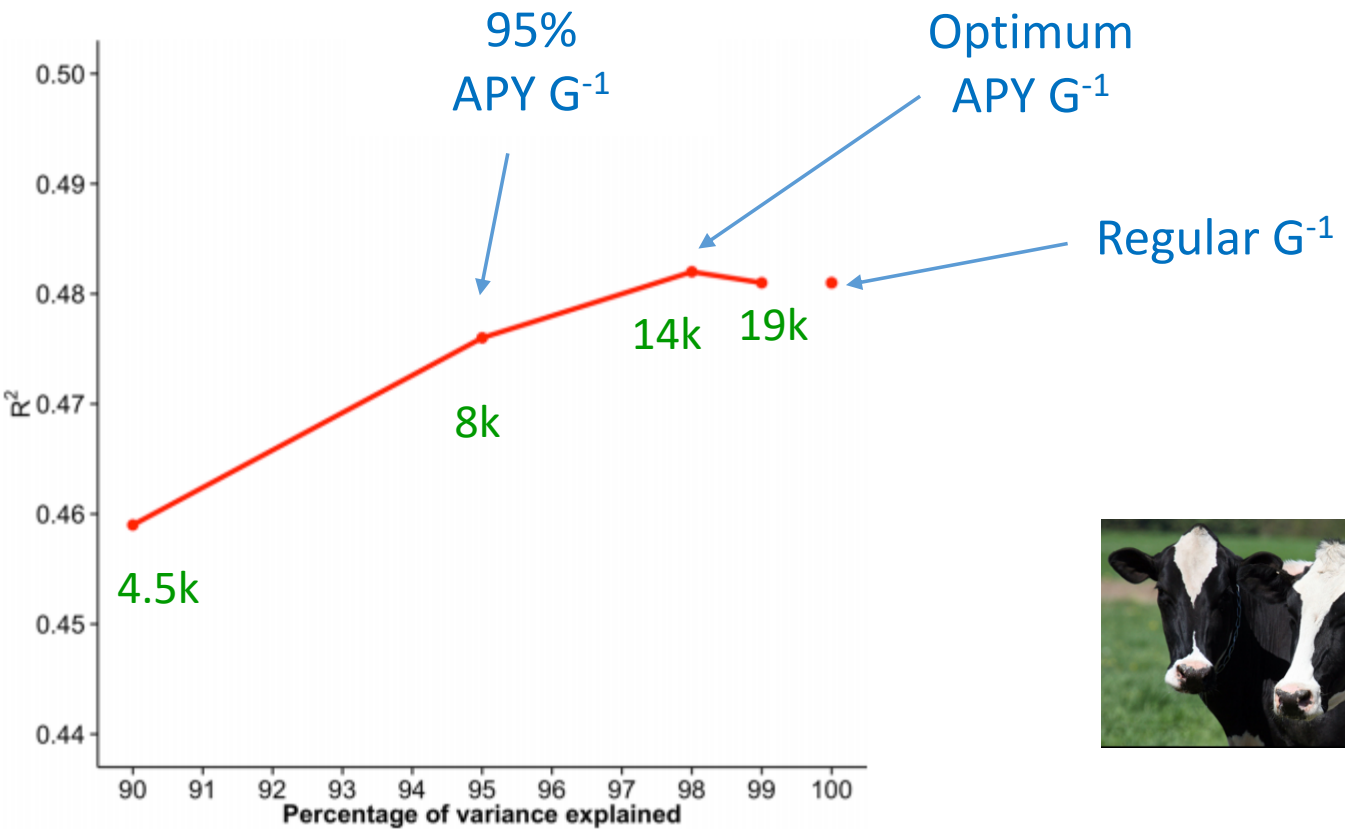
Cor(GEBV,GEBV_APY) > 0.99



How many core animals in APY?

What happens if more animals are genotyped?

Should we change the core number over time?



How many core animals in APY?

- Ostersen et al. (2016)
  - 21k genotyped pigs

Core	Cor (G^{-1}, G^{-1} APY) Genotyped
Random 10%	0.98
Oldest 10%	0.93
Youngest 10%	0.93

core animals < ideal number from Pocrnic et al. (2016)

weak links to recent population

core has no phenotypes

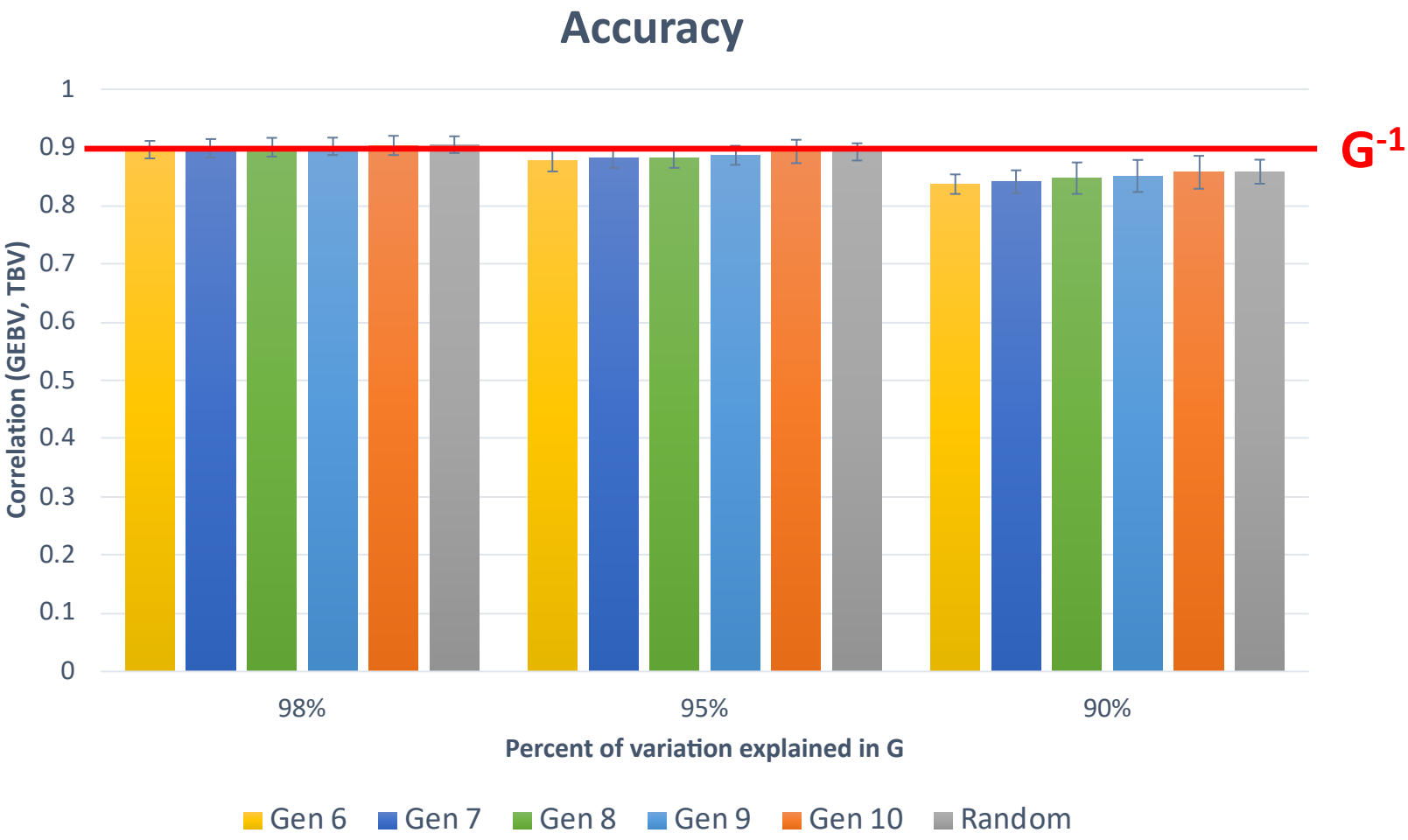
Which core animals in APY?

Bradford et al. (2017)

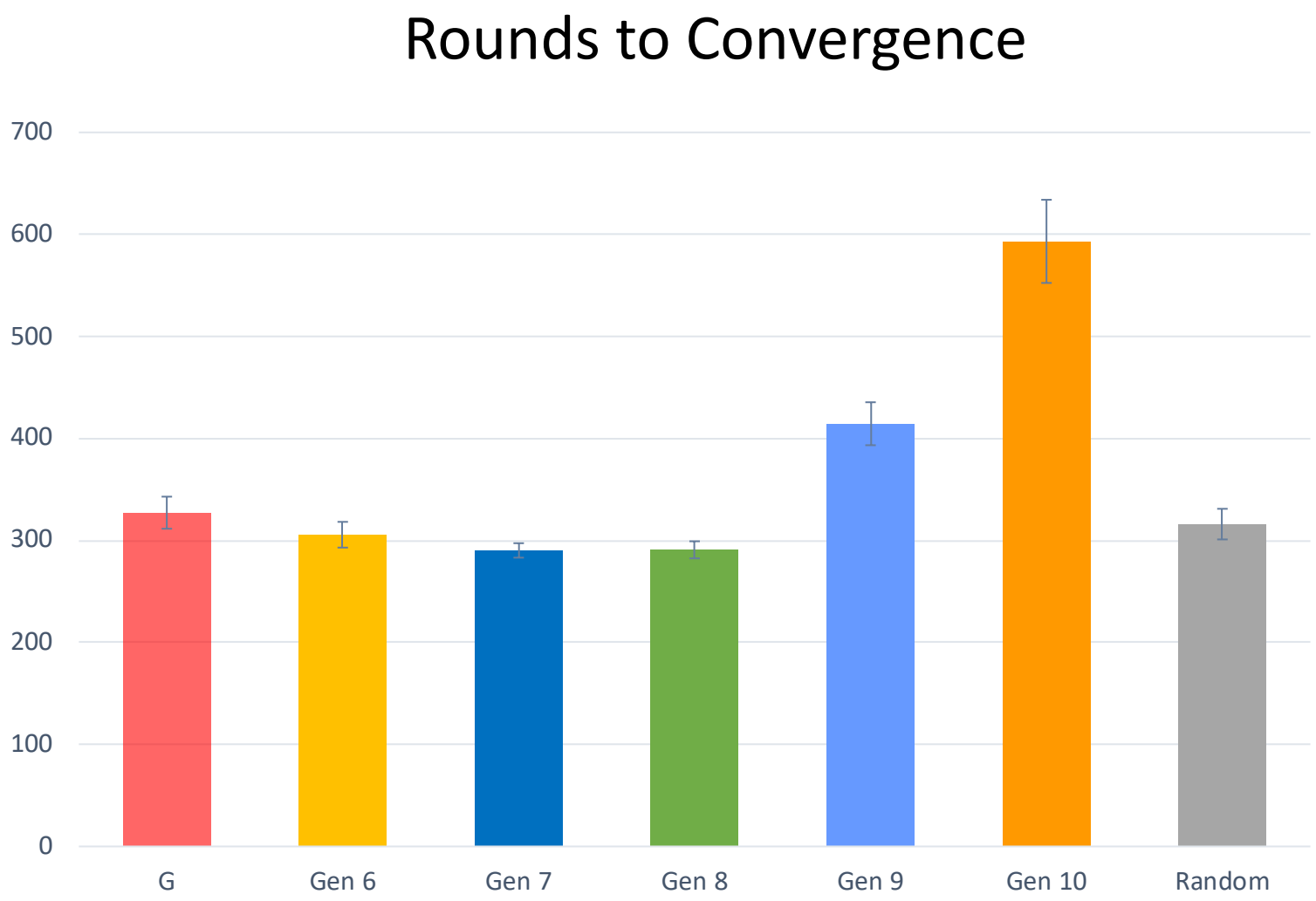
- Simulated populations (QMSim; Sargolzaei and Schenkel, 2009)
- Ne = 40
- #genotyped animals = 50,000

- Core animals:
  - Random gen 6 || gen 7 || gen8 || gen9 || gen 10 (y)
  - Random all generations
  - Incomplete pedigree
  - Genotypes in gen 9 and 10 imputed with 98% accuracy

Which core animals in APY?

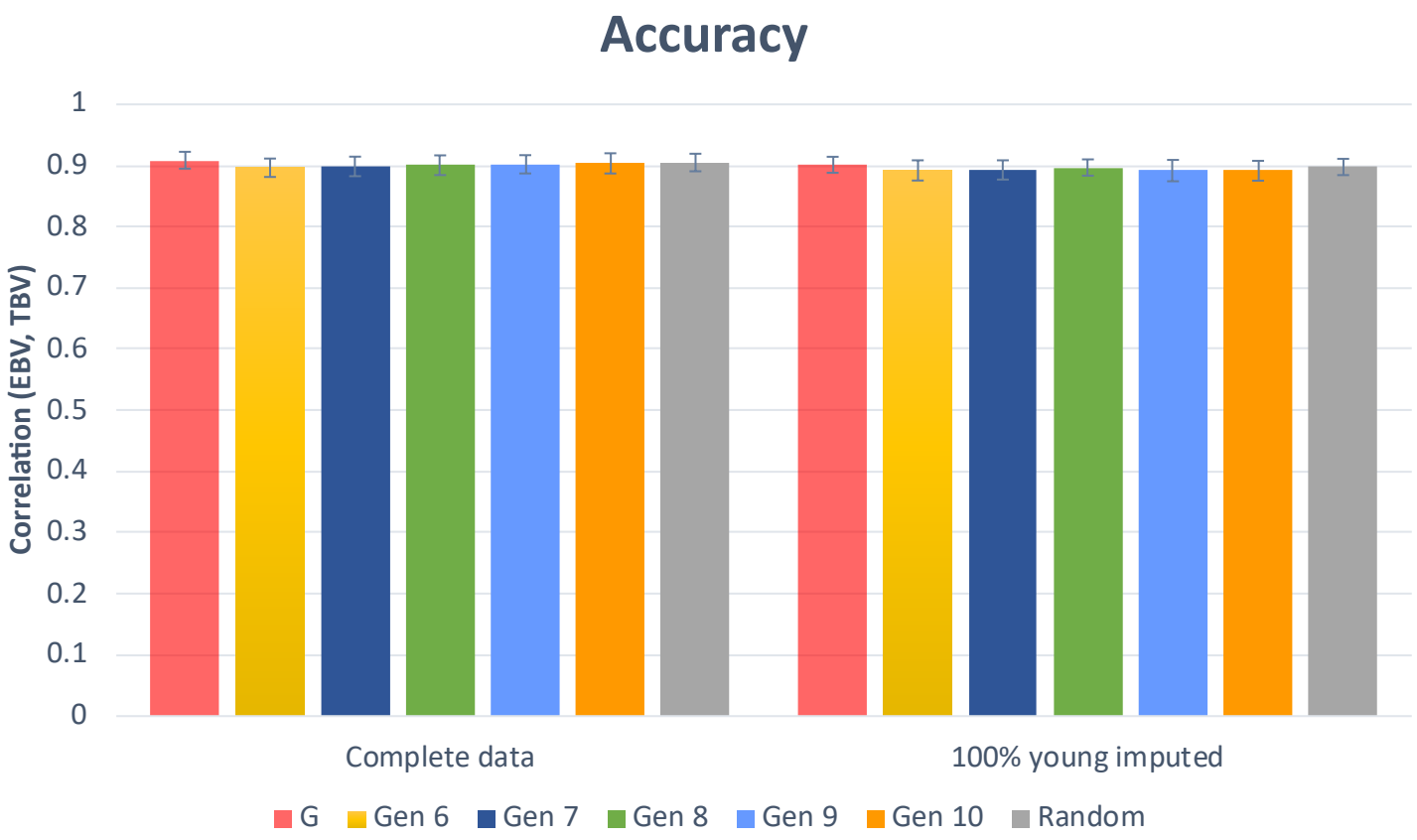


Which core animals in APY?



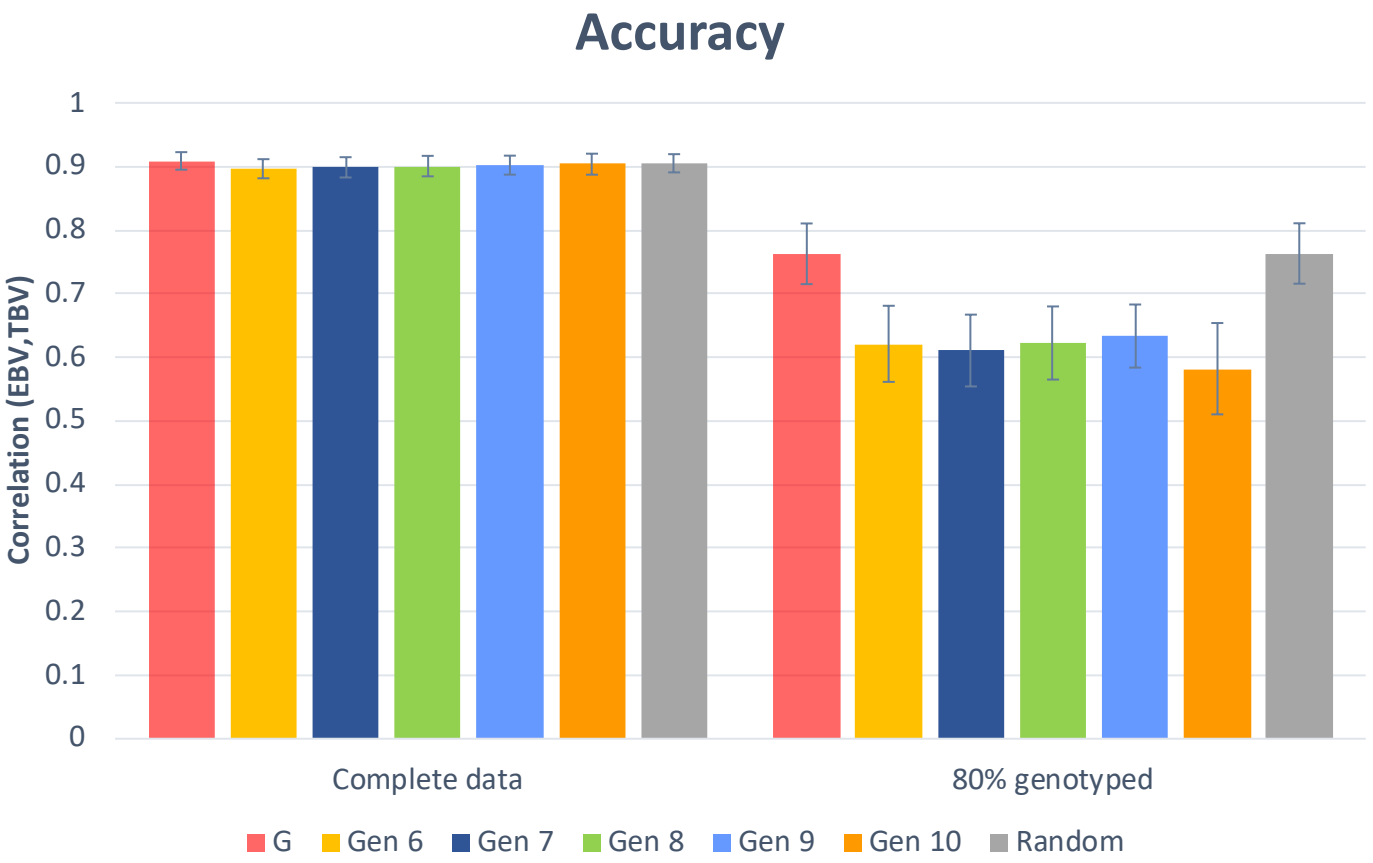
Which core animals in APY?

Imputation with 2% error



Which core animals in APY?

80% genotyped animals with missing pedigree



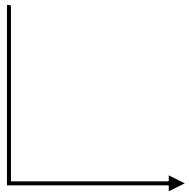
Which core animals in APY?

```
If (sire != 0 .and. dam != 0) then

   core = any definition

else

   core = random
```



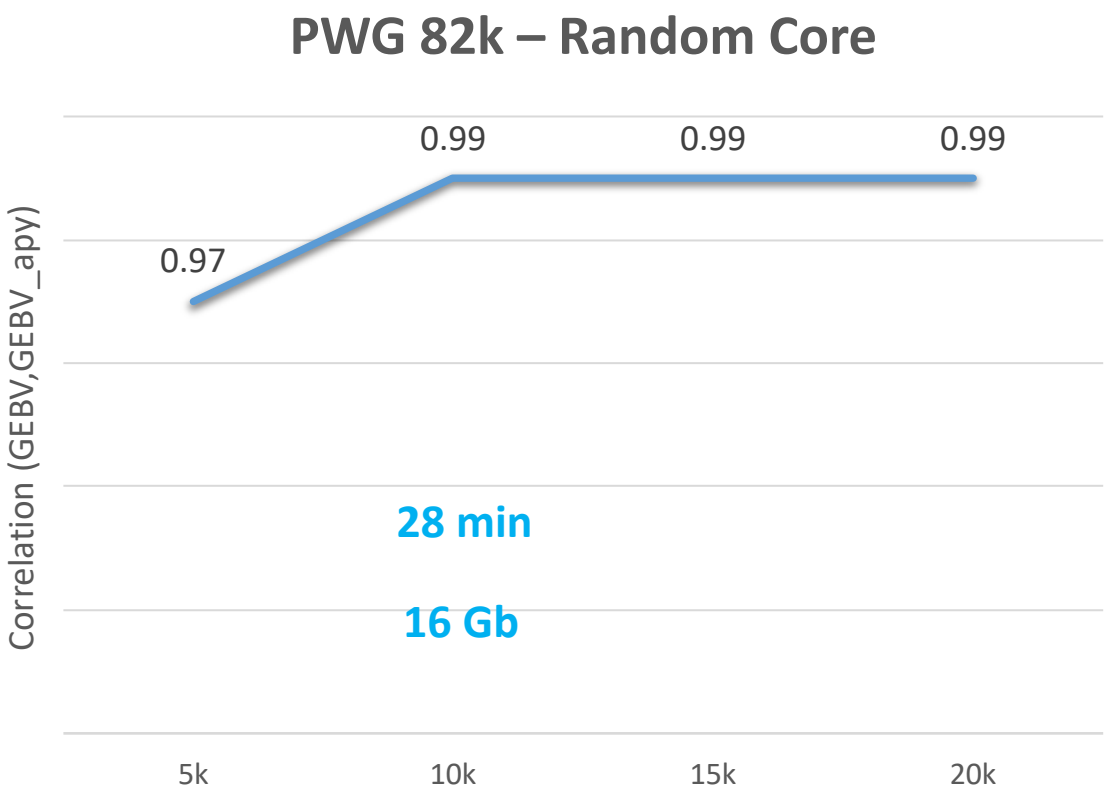
all generations represented

Do I need APY?

- APY is just an algorithm to construct G^{-1} when inverting G is **computationally not feasible**

```
If (number_genotyped > 50,000) then
   APY = True
else
   APY = False
```

How fast is APY?



Regular inversion = 213 min

230 Gb

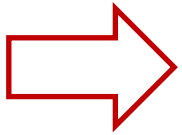
Lourenco et al., 2015

Largest evaluations with APY?

UGA & Collaborators

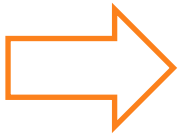
- American Angus 

  - 500k genotyped animals
  - 19k core
  - all traits
  - ~ 2 hour (G^{-1} APY)

 Spring/2017

- US Holsteins 

  - ~500k genotyped animals
  - several traits

 Fall/2016

Largest evaluations with APY?

- US Holsteins
  - 760k genotyped animals
  - 14k core
  - 23M pedigree
  - 37M phenotypes
  - M / F / P
  - ~ 74Gb RAM

Masuda et al. (2016)

APY with 2M genotyped animals?

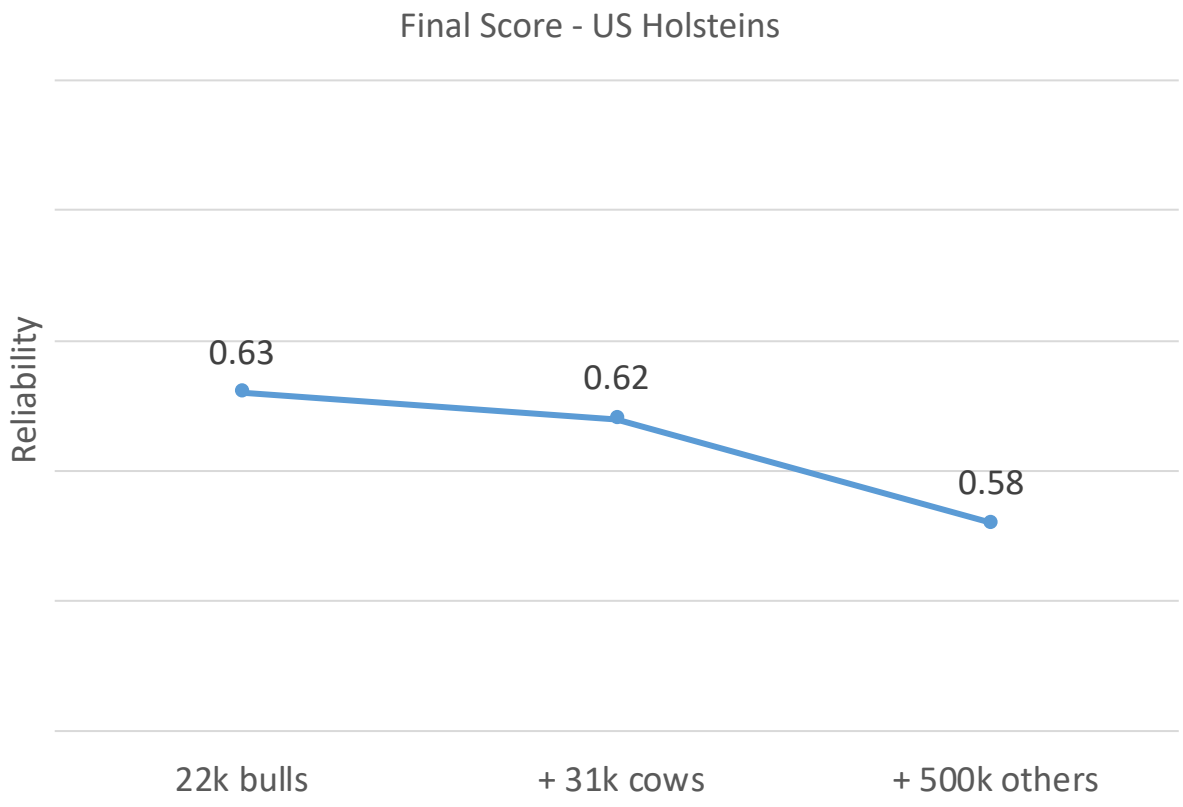
- Is it feasible?
  - 14k core
  - G^{-1} = 29 Tb vs. APY G^{-1} = 208 Gb

- Should we include all genotyped animals?
  - US Holsteins 2M
  - > 75% female LD + imputation + missing ped?

Should we include all genotyped animals?

Cooper et al. (2015)

Masuda et al. (unpublished)



Match
G and A22

What happens with A_{22} in APY?

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

- Fragomeni et al. (2015; unpublished): APY does not work for A_{22}
- Masuda et al. (2017): Rules for inversion of a partitioned matrix

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

Technical note: Avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient[1]

Y. Masuda,*[2] I. Misztal,* A. Legarra,† S. Tsuruta,* D. A. L. Lourenco,* B. O. Fragomeni,* and I. Aguilar‡

J. Anim. Sci. 2017.95:49–52
doi:10.2527/jas2016.0699

Other options for big data

- Angus data
- 500,000 genotyped animals
- 54,000 SNP
  - $\mathbf{G}^{-1}$ is a $500{,}000 \times 500{,}000$ matrix
  - $\mathbf{A}_{22}^{-1}$ is a $500{,}000 \times 500{,}000$ matrix
  - SNP-BLUP $\mathbf{Z}'\mathbf{Z}$ is a $54{,}000 \times 54{,}000$

- Indirect representations of **G**
- Sherman-Woodbury inversions

$$\mathbf{G} = \frac{1}{\varepsilon}\mathbf{I} + \mathbf{Z}\mathbf{Z}' \quad \text{and} \quad \mathbf{G}^{-1} = \frac{1}{\varepsilon}\mathbf{I} - \left(\frac{1}{\varepsilon}\mathbf{Z}\left(\frac{1}{\varepsilon}\mathbf{Z}'\mathbf{Z} + \mathbf{I}\right)^{-1}\mathbf{Z}'\frac{1}{\varepsilon}\right)$$

ssGTBLUP – Mantysaari et al. (2017)

Single-step GTBLUP

- ssGTBLUP is based on **Woodbury** matrix identity:

  If $\mathbf{G}_C = \mathbf{G}_0 + \mathbf{C} = \mathbf{Z}\mathbf{Z}' + \mathbf{C}$ then $\mathbf{G}_C^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{C}^{-1}\mathbf{Z} + \mathbf{I})^{-1}\mathbf{Z}'\mathbf{C}^{-1}$

  usually $\mathbf{G}_C = \mathbf{G}_0 + \varepsilon\mathbf{I}$ then $\mathbf{G}_\varepsilon^{-1} = \frac{1}{\varepsilon}\mathbf{I} - \frac{1}{\varepsilon}\mathbf{Z}\left(\frac{1}{\varepsilon}\mathbf{Z}'\mathbf{Z} + \mathbf{I}\right)^{-1}\mathbf{Z}'\frac{1}{\varepsilon}$

- This gives us an alternative form of the inverse:

$$\mathbf{G}_\varepsilon^{-1} = \frac{1}{\varepsilon}\mathbf{I} - \mathbf{T}_\varepsilon'\mathbf{T}_\varepsilon$$

  - This is no approximation, but instead exact
  - It gives significant computational savings when $n_{anim.} > > > n_{SNP}$
  - In case of 1 milj animals computing time ssGBLUP ~ 10 * ssGTBLUP

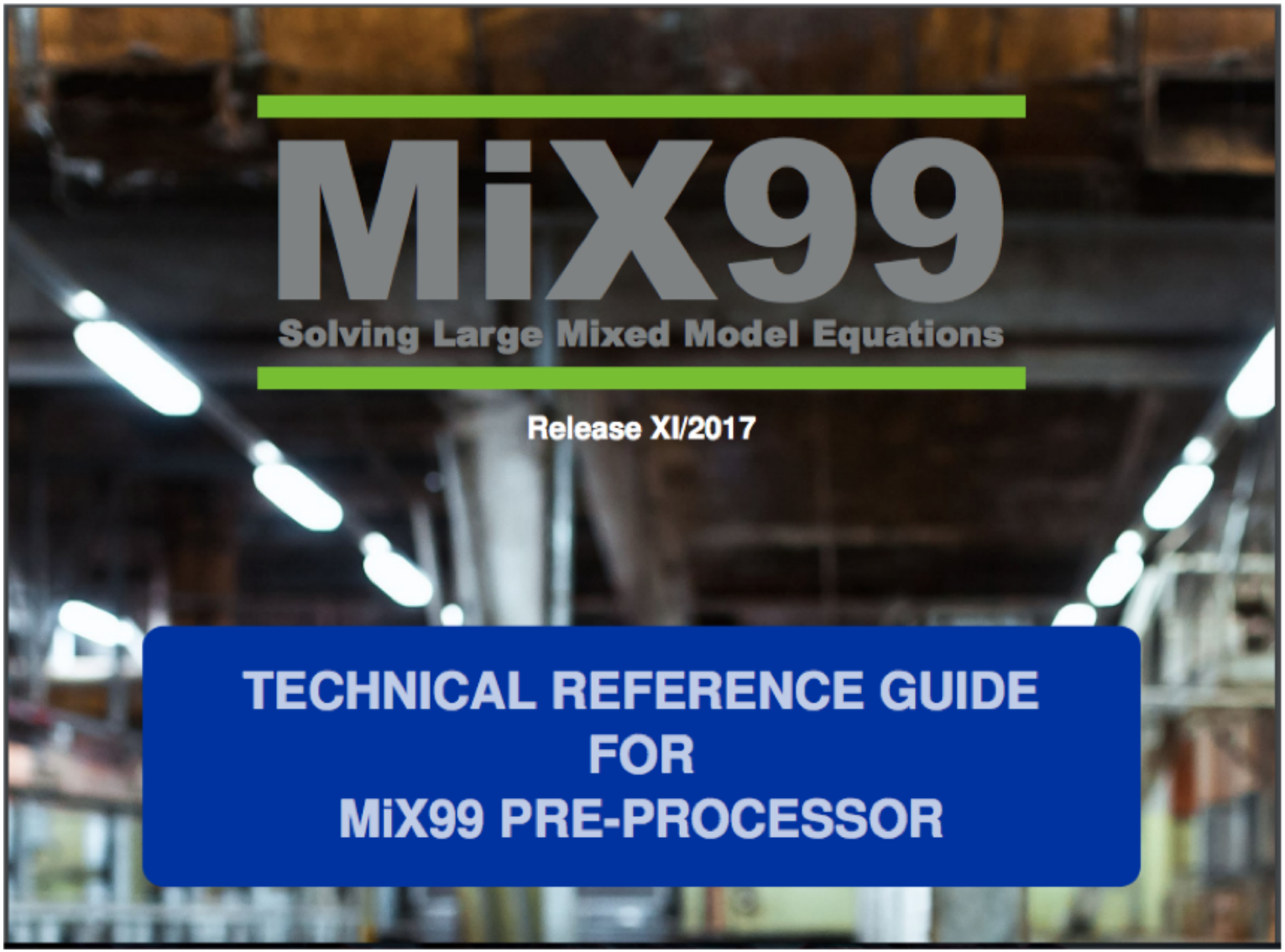
$\mathbf{T}'_\varepsilon$ has a size n x m

- Can use reduced **Z** by single value decomposition

ssGTBLUP – Mantysaari et al. (2017)

- Computing time with 160 000 genotyped animals:
  - ssGBLUP with full **G** inverse **24 times** longer than with AM
  - ssGTBLUP with no approximation **18 times**
  - APY50K and ssGTBLUP(98) **12 times** longer than with AM
  - APY30K 8 times longer than with AM

ssGTBLUP – Mantysaari et al. (2017)



ssGBLUP with marker effects

- Legarra & Ducrocq 2012: ssGBLUP model on SNP effects ($\mathbf{g}$) and BV ($\boldsymbol{u}$)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}_2'\mathbf{W}_2\mathbf{Z} & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{W}_1'\mathbf{W}_1 + \alpha_u\mathbf{A}^{11} & \alpha_u\mathbf{A}^{12}\mathbf{Z} & \mathbf{0} \\ \mathbf{Z}'\mathbf{W}_2'\mathbf{X}_2 & \alpha_u\mathbf{Z}'\mathbf{A}^{12} & \mathbf{Z}'\mathbf{W}_2'\mathbf{W}_2\mathbf{Z} + \alpha_u\mathbf{Z}'\mathbf{A}^{22}\mathbf{Z} + \mathbf{D}^{-1}\sigma_e^2 & \alpha_u\mathbf{Z}' \\ \mathbf{0} & \mathbf{0} & \alpha_u\mathbf{Z} & \alpha_u\mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{g}} \\ -\hat{\boldsymbol{\varphi}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_1'\mathbf{y}_1 \\ \mathbf{Z}'\mathbf{W}_2'\mathbf{y}_2 \\ \mathbf{0} \end{bmatrix}$$

Non genotyped animals

Marker effects

Matrices **ZW** in this model get very complicated for complex models because they involve very intense products

Single-step with marker effects

- Rediscovered by Fernando et al. (2016)
- Super hybrid model

$$\begin{bmatrix} \mathbf{y}_n \\ \mathbf{y}_g \end{bmatrix} = \begin{bmatrix} \mathbf{X}_n \\ \mathbf{X}_g \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{0} & \mathbf{Z}_n \\ \mathbf{Z}_g\mathbf{M}_g & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{u}_n \end{bmatrix} + \mathbf{e}$$

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

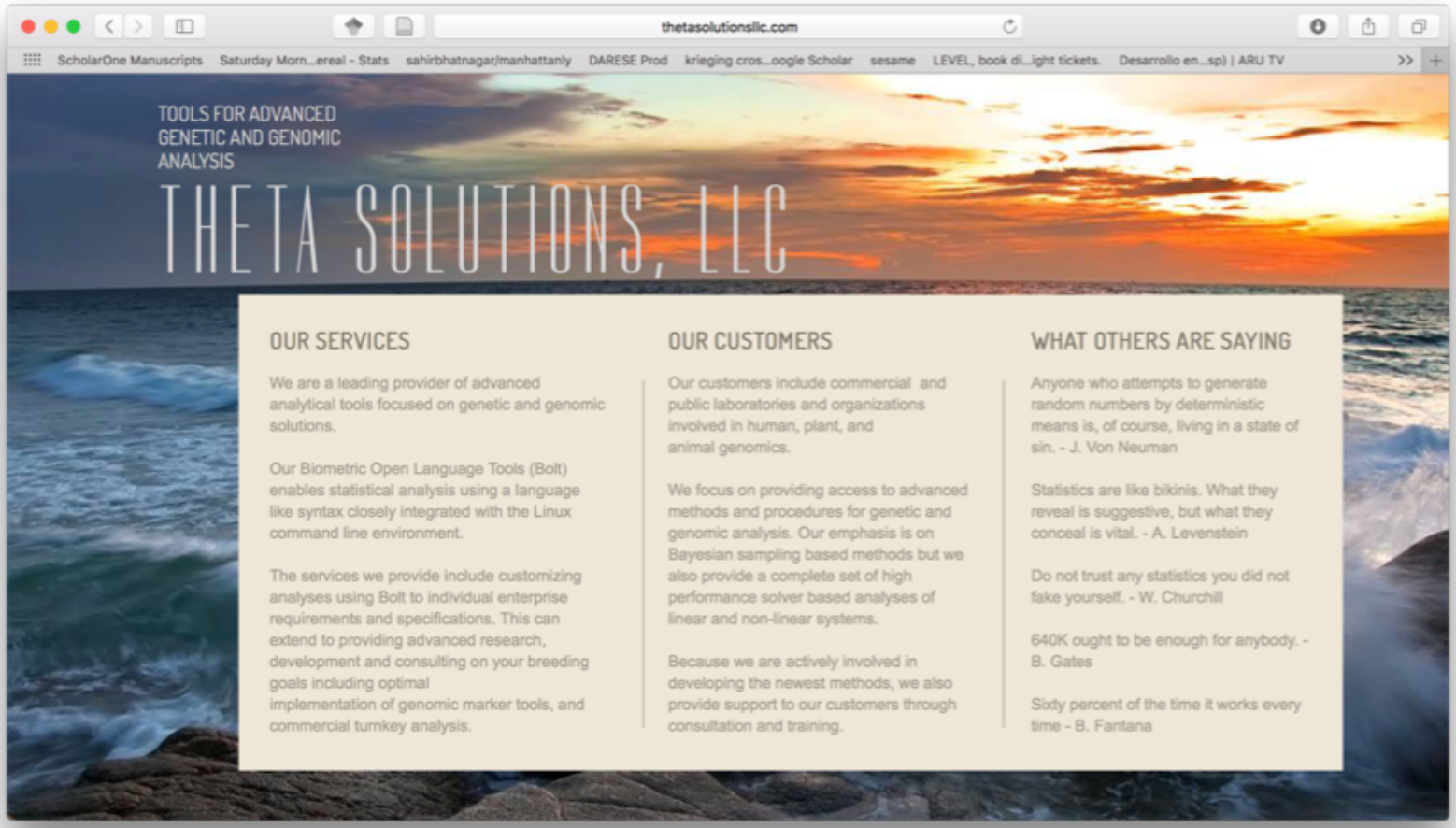
Marker effects

Non genotyped animals

$$\mathbf{Q} = \mathbf{M}'_g\mathbf{Z}'_g\mathbf{Z}_g\mathbf{M}_g + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} + \mathbf{M}'_n\mathbf{A}^{nn}\mathbf{M}_n\frac{\sigma_e^2}{\sigma_g^2} \qquad \mathbf{M}_n = \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{M}_g = -(\mathbf{A}^{nn})^{-1}\mathbf{A}^{ng}\mathbf{M}_g$$

Single-step with marker effects

Bolt



Large-scale genomic evaluations?

- Limited dimensionality of genomic information
- APY ssGBLUP
  - $u_i \mid \sum_{j=\text{core}} p_{ij} u_j + \varepsilon_i$
    - Number of core depends on Ne
    - # eigenvalues 98% = #core
    - Computing cost greatly reduced
    - Used for commercial large-scale genomic evaluation

Large-scale genomic evaluations?

- Large-scale genomic evaluations
  - Problem only for 1% of the users
- Currently, at least 3 different solutions
  - Blupf90
  - MiX99
  - Bolt
- The exact strategy may depend on the problem
- Maybe in 10 years all animals are genotyped
  - Old data is forgotten