Extension of single-step ssGBLUP to many genotyped individuals

Ignacy Misztal

University of Georgia

Genomic selection and single-step

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

Aguilar et al., 2010
Christensen and Lund, 2010

- Simplicity
  - No DYD or DP
  - No index
  - No complexity

- Accuracy
  - Avoids double counting
  - Avoids fixed index
  - Accounts for preselection bias

Current implementation of SS

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

- G and A_{22} created explicitly
- Quadratic memory and cubic computations
- Cost per 100k genotypes - 1.5 hr (Aguilar et al.,2014)

Number of genotypes and impending problem

> 2 M for Holsteins

> 400k for Angus

Genomic pre-selection issue (Patry and Ducrocq, 2011; VanRaden et al., 2013)

- BLUP increasingly biased
- Need all data on preselection included

Unsymmetric equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{H}\mathbf{Z}'\mathbf{X} & \mathbf{H}\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{H}\mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Misztal et al., 2009

No convergence without good preconditioner

No convergence with large H or A

No G or A_{22} inverse model

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}_2'\mathbf{W}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{W}_1'\mathbf{W}_1 + \alpha_u \mathbf{A}^{11} & \alpha_u \mathbf{A}^{12} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_2 & \alpha_u \mathbf{A}^{12} & \mathbf{W}_2'\mathbf{W}_2 + \alpha_u \mathbf{A}^{22} & \alpha_u \mathbf{I} & -\alpha_u \mathbf{I} \\ \mathbf{0} & \mathbf{0} & \alpha_u \mathbf{I} & \alpha_u \mathbf{A}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \alpha_u \mathbf{I} & \mathbf{0} & \alpha_u \mathbf{G} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ -\hat{\varphi} \\ -\hat{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_1'\mathbf{y}_1 \\ \mathbf{W}_2'\mathbf{y}_2 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

Legarra and Ducrocq (2011)

Slow convergence with few genotypes

Divergence with many genotypes

SNP model for genotyped animals

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

Legarra and Ducrocq, 2011

No successful programming

SNP model for genotyped animals

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_p & \mathbf{X}'\mathbf{W}_1 & \mathbf{X}'\mathbf{W}_2 & \mathbf{0} \\ \mathbf{Z}_p'\mathbf{X} & \mathbf{Z}_p'\mathbf{Z}_p + \mathbf{I}\delta & \mathbf{Z}_p'\mathbf{W}_1 & \mathbf{Z}_p'\mathbf{W}_2 & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X} & \mathbf{W}_1'\mathbf{Z}_p & \mathbf{W}_1'\mathbf{W}_1 + \lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{12} & \mathbf{0} \\ \mathbf{W}_2'\mathbf{X} & \mathbf{W}_2'\mathbf{Z}_p & \lambda\mathbf{A}^{21} & \mathbf{W}_2'\mathbf{W}_2 + \lambda\left(\mathbf{A}^{22} + \left(\frac{1}{k} - 1\right)\mathbf{A}_{22}^{-1}\right) & -\frac{1}{k}\lambda\mathbf{A}_{22}^{-1}\mathbf{Z} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\frac{1}{k}\lambda\mathbf{Z}'\mathbf{A}_{22}^{-1} & \lambda\left(\mathbf{B}^{-1} + \frac{1}{k}\mathbf{Z}'\mathbf{A}_{22}^{-1}\mathbf{Z}\right) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{p}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_p'\mathbf{y} \\ \mathbf{W}_1'\mathbf{y} \\ \mathbf{W}_2'\mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

Liu et al, 2014

SNP effects for all animals (Fernando et al., 2014)

imputed genotypes

centered genotypes

$$\hat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* \\ \mathbf{X}_2^* \end{bmatrix} \boldsymbol{\beta}^* + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{M}}_1 \boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \mathbf{M}_2 \boldsymbol{\alpha} \end{bmatrix} + \mathbf{e}$$

Cost of imputation

Requires new type of programming

Extension to complex models unclear

Can regular ssGBLUP be made more efficient?

Scaling up A_{22}^{-1}

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

- $\mathbf{A}_{22}^{-1}$ dense (Faux et al., 2014)
- For PCG iteration (Stranden et al., 2014)

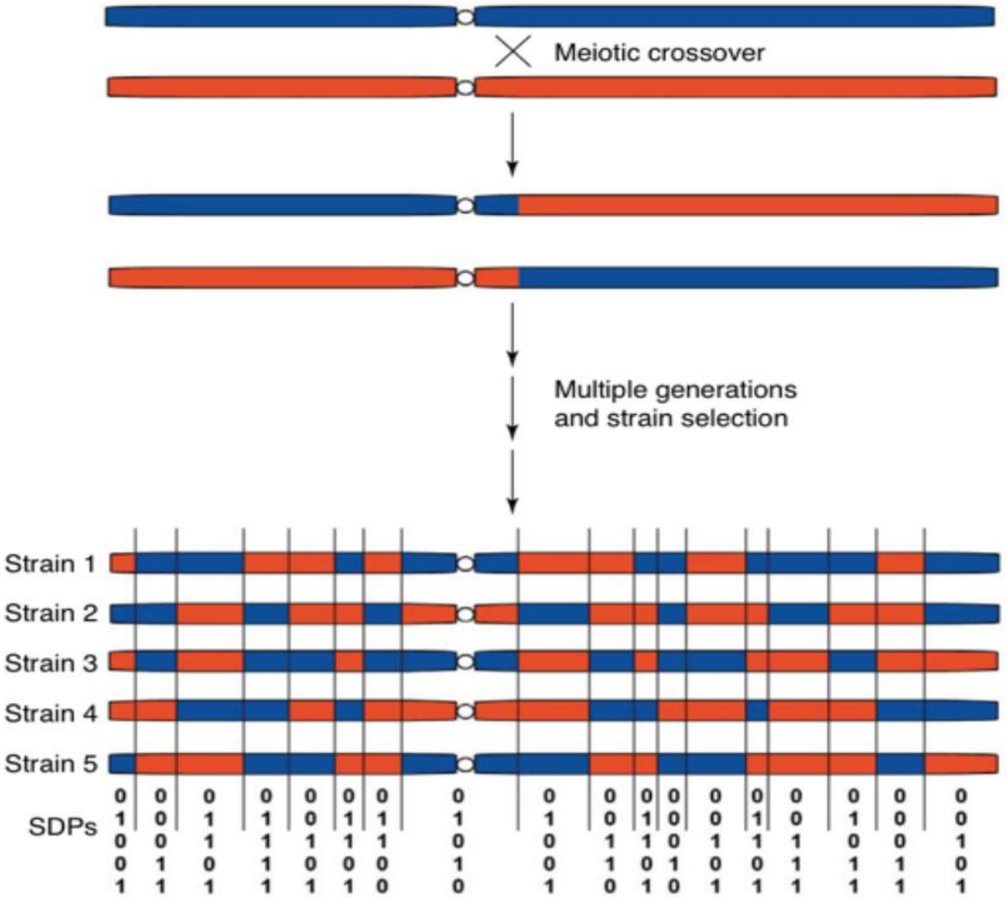
$$\mathbf{A}_{22}^{-1}\mathbf{q} = \mathbf{A}^{22}\mathbf{q} - \left\{\mathbf{A}^{21}\left[\left(\mathbf{A}^{22}\right)^{-1}\left(\mathbf{A}^{12}\mathbf{q}\ \right)\right]\right\}$$

- Seconds for 500k animals with good programming (Masuda et al., 2017)

Is dimensionality of genomic information limited?

- Regular G not positive definite past ~5k
  - Blending with A (VanRaden, 2008)
- Dimensionality of SNP BLUP small (Maciotta et al., 2013)
- Success of imputation
- Manhattan plots noisy until averaged by 300k-10Mb (depending on species)

Origin of Haplotype blocks



	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SDPs	1	0	1	1	0	1	1	1	1	0	1	0	0	1	0	1	0	0
	0	0	1	1	1	1	1	0	0	1	1	0	1	1	1	0	0	1
	0	1	0	1	0	0	0	1	0	1	0	1	0	0	1	1	1	0
	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1

Cuppen, 2005

Heterogenetic and homogenic tracts in genome (Stam, 1980)

E(#tracts)=4NeL (Stam, 1980)

Ne – effective population size

L –length of genome in Morgans

Holsteins: Ne ≈100 L=30

Me=12,000

Inversion via SVD/eigenvalue decomposition

Assume 1 million animals genotyped with 60k chip

$\mathbf{G} = \mathbf{ZZ}' = \mathbf{UDU}'$ Eigenvalue decomposition (1M x 1M)

$\mathbf{G}^{-} = \mathbf{UD}^{-}\mathbf{U}'$ Generalized inverse (1M x 1M)

$\mathbf{Z} = \mathbf{USV} = \mathbf{UD}^{0.5}\mathbf{V}$ - SVD decomposition (1M x 60k)
10h for 720k animals (Masuda, 2017)

t - index for non-negligible eigenvalues, say 10k

$$\mathbf{G}^{-} = \mathbf{U}_t\ \mathbf{D}_t^{-1}\mathbf{U}_t' = \mathbf{U}_t\ \mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{U}_t' = \mathbf{U}_*\ \mathbf{U}_*$$

For PCG iteration

$\mathbf{G}^{-1}\mathbf{q} = \mathbf{U}_*\ \left(\mathbf{U}_*\ \mathbf{q}\right)$ - only 1 M x 10k elements

Inverse by Woodbury formula

$\mathbf{G} = \mathbf{ZZ'} + \mathbf{I\varepsilon},$

$\mathbf{G}^{-1} = \frac{1}{\varepsilon}\mathbf{I} - \frac{1}{\varepsilon}\mathbf{Z}(\frac{1}{\varepsilon}\mathbf{Z'Z} + \mathbf{I})^{-1}\mathbf{Z'}\frac{1}{\varepsilon}$

Woodbury formula

Z’Z 60k x 60k

For PCG iteration:

Mantysaari et al., 2017

$$\mathbf{G}^{-1}\mathbf{q} = \frac{\mathbf{1}}{\boldsymbol{\varepsilon}}\{\mathbf{I} - \mathbf{Z}(\mathbf{UDU'})^{-1}\mathbf{Z'}\}\boldsymbol{q} = \frac{\mathbf{1}}{\boldsymbol{\varepsilon}}\{\mathbf{I} - \mathbf{SS'}\}\boldsymbol{q}$$

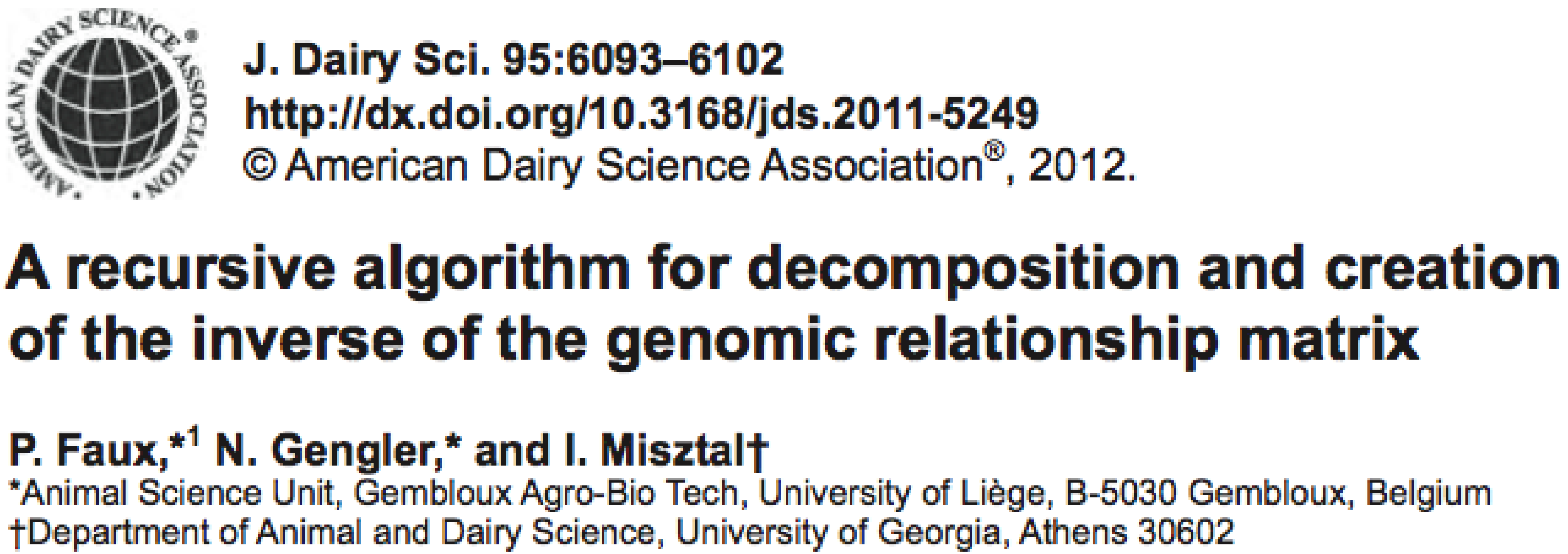
$$\mathbf{S} = \mathbf{ZU'D}^{-1/2}$$

With reduced rank $\mathbf{S} = \mathbf{ZU}_t{}'(\mathbf{D}_t)^{-\frac{1}{2}}$ (1M x 10k)

Ostersen et al., 2017

If G has limited dimensionality, can G^{-1} be sparse like A^{-1}?

Use of a la Henderson's rules?



J. Dairy Sci. 95:6093–6102
http://dx.doi.org/10.3168/jds.2011-5249
© American Dairy Science Association®, 2012.

A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix

P. Faux,*[1] N. Gengler,* and I. Misztal†
*Animal Science Unit, Gembloux Agro-Bio Tech, University of Liège, B-5030 Gembloux, Belgium
†Department of Animal and Dairy Science, University of Georgia, Athens 30602

Use of relatives for **G**$^{-1}$

Accuracies not good enough

Theory not clear

Assumption of limited dimensionality

s – n x 1 vector containing additive information of population (haplotypes, chromosome segments, LD blocks)?

Breeding value

Very small error

$$\mathbf{u} = \mathbf{T}\mathbf{s} + \mathbf{e}$$

If $\mathbf{u}_c$ contains n animals:

$$\mathbf{s} \approx \mathbf{T}_c^{-1}\mathbf{u}_c$$

Breeding values of any n animals contains all additive information

Choose core "**c**" and noncore "**n**" animals

$$\mathbf{u}_n = \mathbf{P}_{nc}\mathbf{u}_c + \varepsilon_n$$

$$\mathbf{u}_c = \mathbf{u}_c$$

$$\begin{bmatrix} \mathbf{u}_c \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{u}_c \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

$$\text{var}(\boldsymbol{\varepsilon}_n) = \mathbf{M_{nn}}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{0} \\ 0 & \mathbf{M}_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{P}_{cn} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

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

How to estimate **P** and inv(**G**)?

$$\mathrm{var}\left(\begin{bmatrix}\mathbf{u}_c\\ \mathbf{u}_n\end{bmatrix}\right)=\begin{bmatrix}\mathbf{G}_{cc} & \mathbf{G}_{cn}\\ \mathbf{G}_{nc} & \mathbf{G}_{nn}\end{bmatrix}\sigma_u^2$$

G is “true” relationship matrix

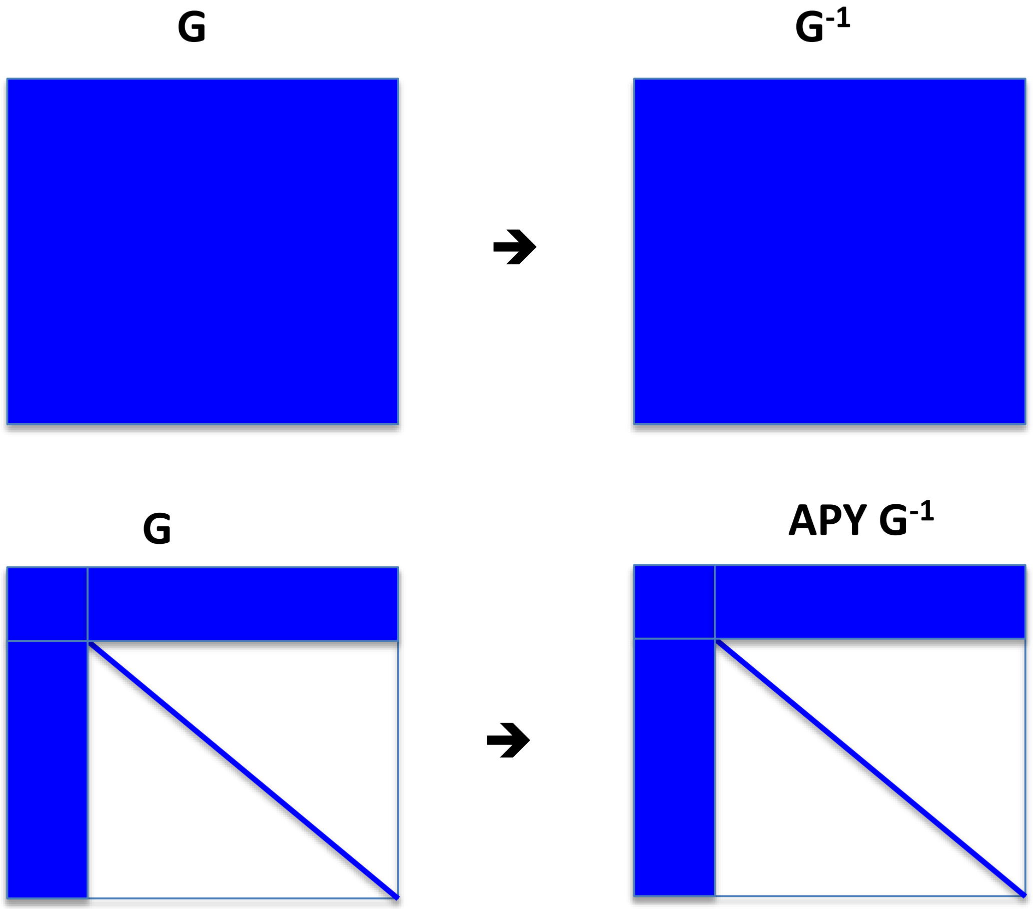
$$\mathbf{u}_n \mid \mathbf{u}_c = \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{u}_c, \quad \mathbf{P} = \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}$$

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

APY algorithm

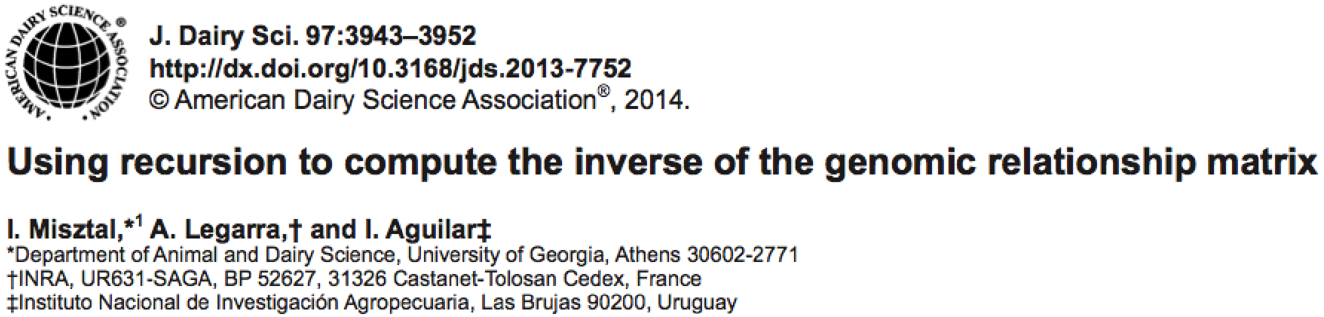
(Algorithm for Proven and Young)

Properties of APY algorithm



Cost:
Quadratic memory and cubic computations

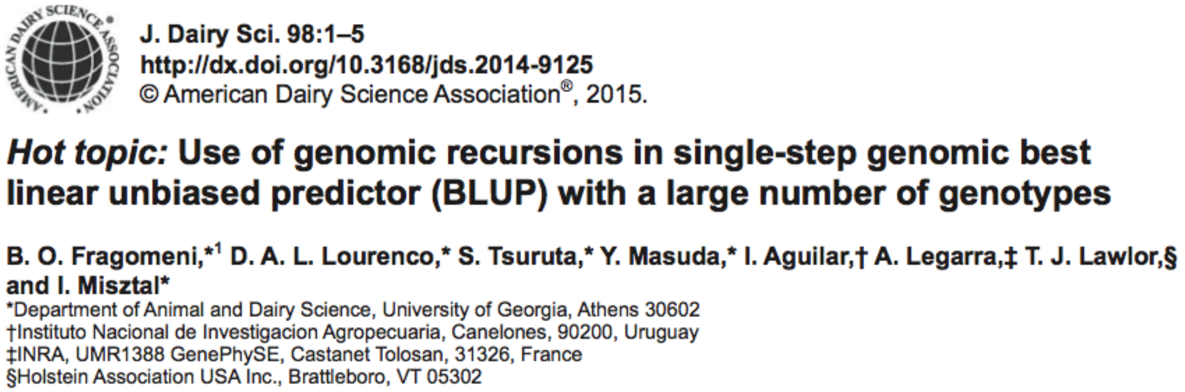
Cost:
Almost linear memory and computations



J. Dairy Sci. 97:3943–3952
http://dx.doi.org/10.3168/jds.2013-7752
© American Dairy Science Association®, 2014.

Using recursion to compute the inverse of the genomic relationship matrix

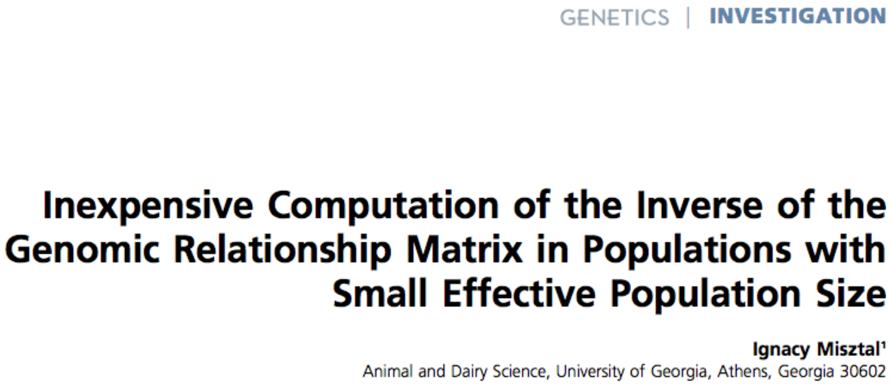
I. Misztal,*[1] A. Legarra,† and I. Aguilar‡
*Department of Animal and Dairy Science, University of Georgia, Athens 30602-2771
†INRA, UR631-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
‡Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay



J. Dairy Sci. 98:1–5
http://dx.doi.org/10.3168/jds.2014-9125
© American Dairy Science Association®, 2015.

Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes

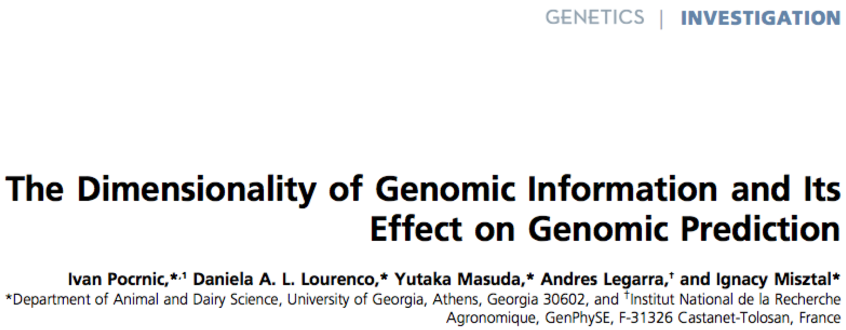
B. O. Fragomeni,*[1] D. A. L. Lourenco,* S. Tsuruta,* Y. Masuda,* I. Aguilar,† A. Legarra,‡ T. J. Lawlor,§ and I. Misztal*
*Department of Animal and Dairy Science, University of Georgia, Athens 30602
†Instituto Nacional de Investigacion Agropecuaria, Canelones, 90200, Uruguay
‡INRA, UMR1388 GenePhySE, Castanet Tolosan, 31326, France
§Holstein Association USA Inc., Brattleboro, VT 05302



GENETICS | INVESTIGATION

Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size

Ignacy Misztal[1]
Animal and Dairy Science, University of Georgia, Athens, Georgia 30602

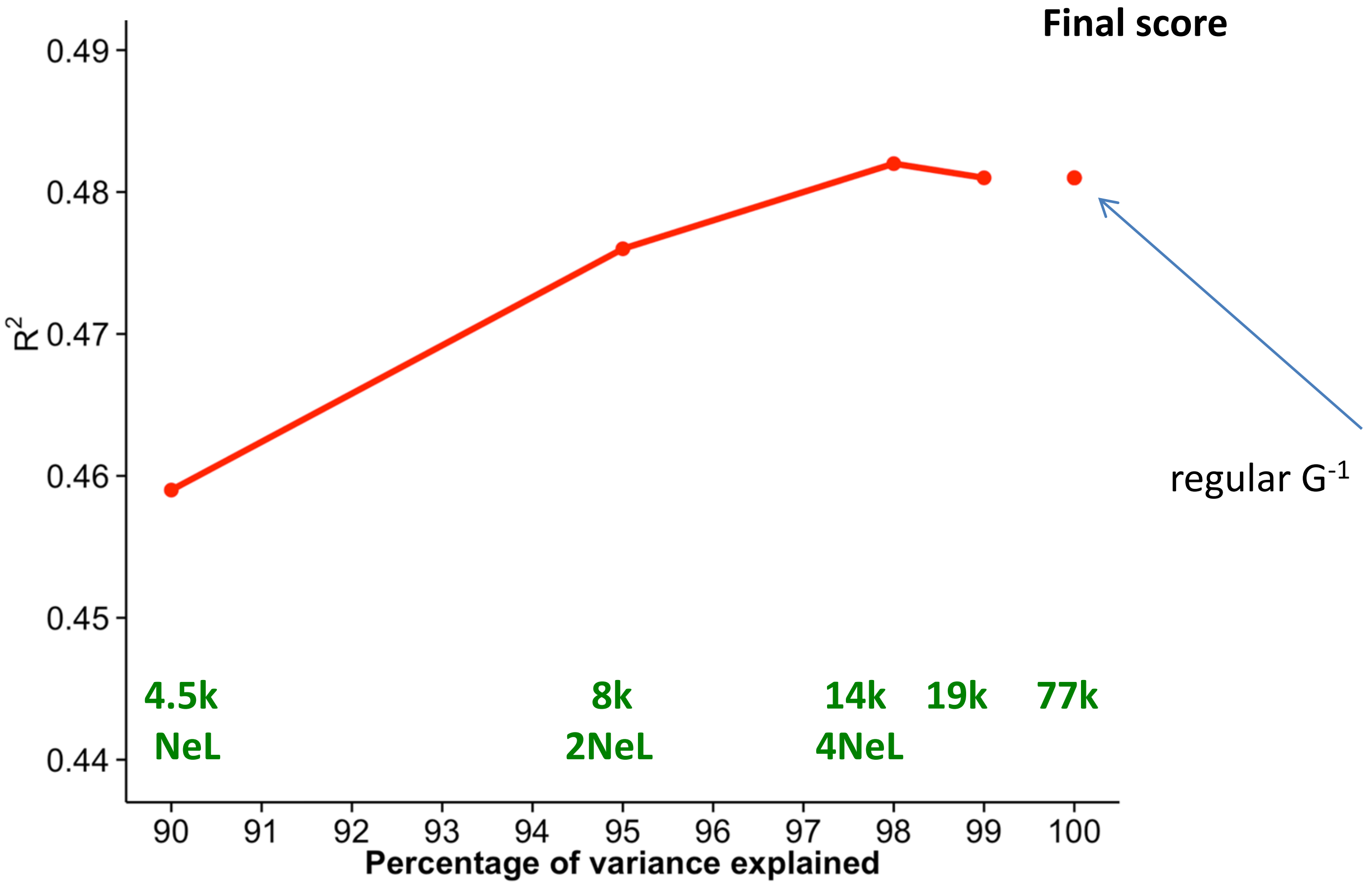


GENETICS | INVESTIGATION

The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

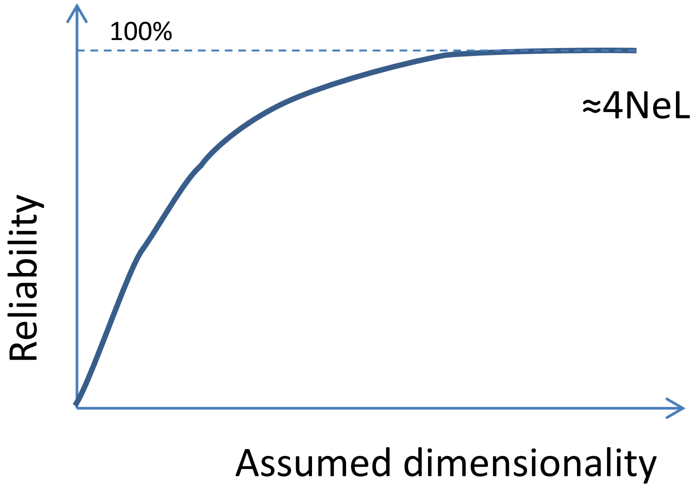
Ivan Pocrnic,*[1] Daniela A. L. Lourenco,* Yutaka Masuda,* Andres Legarra,† and Ignacy Misztal*
*Department of Animal and Dairy Science, University of Georgia, Athens, Georgia 30602, and †Institut National de la Recherche Agronomique, GenPhySE, F-31326 Castanet-Tolosan, France

Reliabilities – Holsteins (77k)



Pocrnic et al., 2016b

Distribution of segments/haplotypes/..



Costs with 720k genotyped animals

- 30 M Holsteins
- 50 M records
- 764k 60k genotypes

Item	BLUP	ssGBLUP
APY G	-	7 h
A22-1	-	10 min
rounds	402	464
Time/round	51 s	83 s
Total time	**6 h**	**17 h**

Masuda et al., 2017

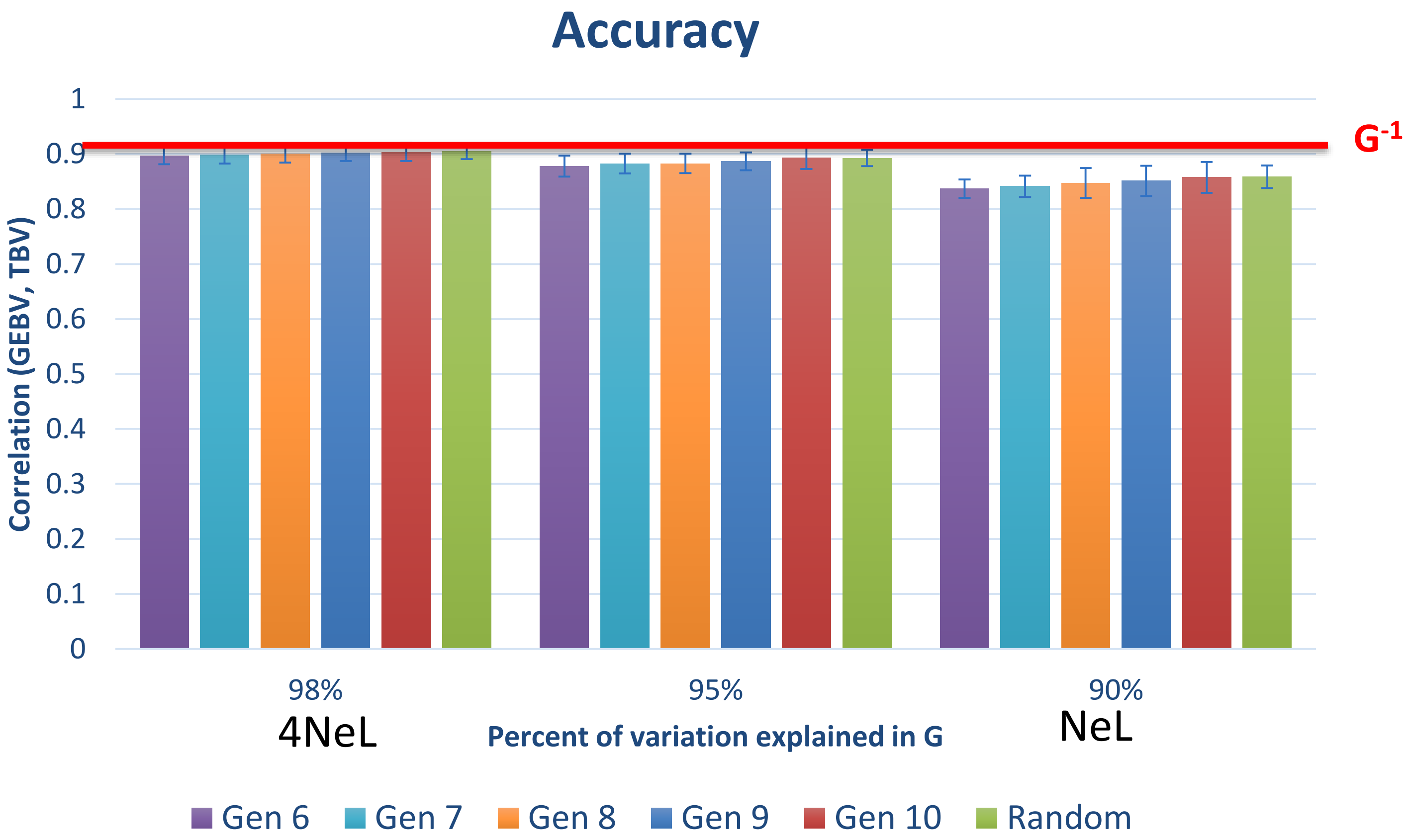
Which core animals in APY?

Bradford et al. (2017)

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

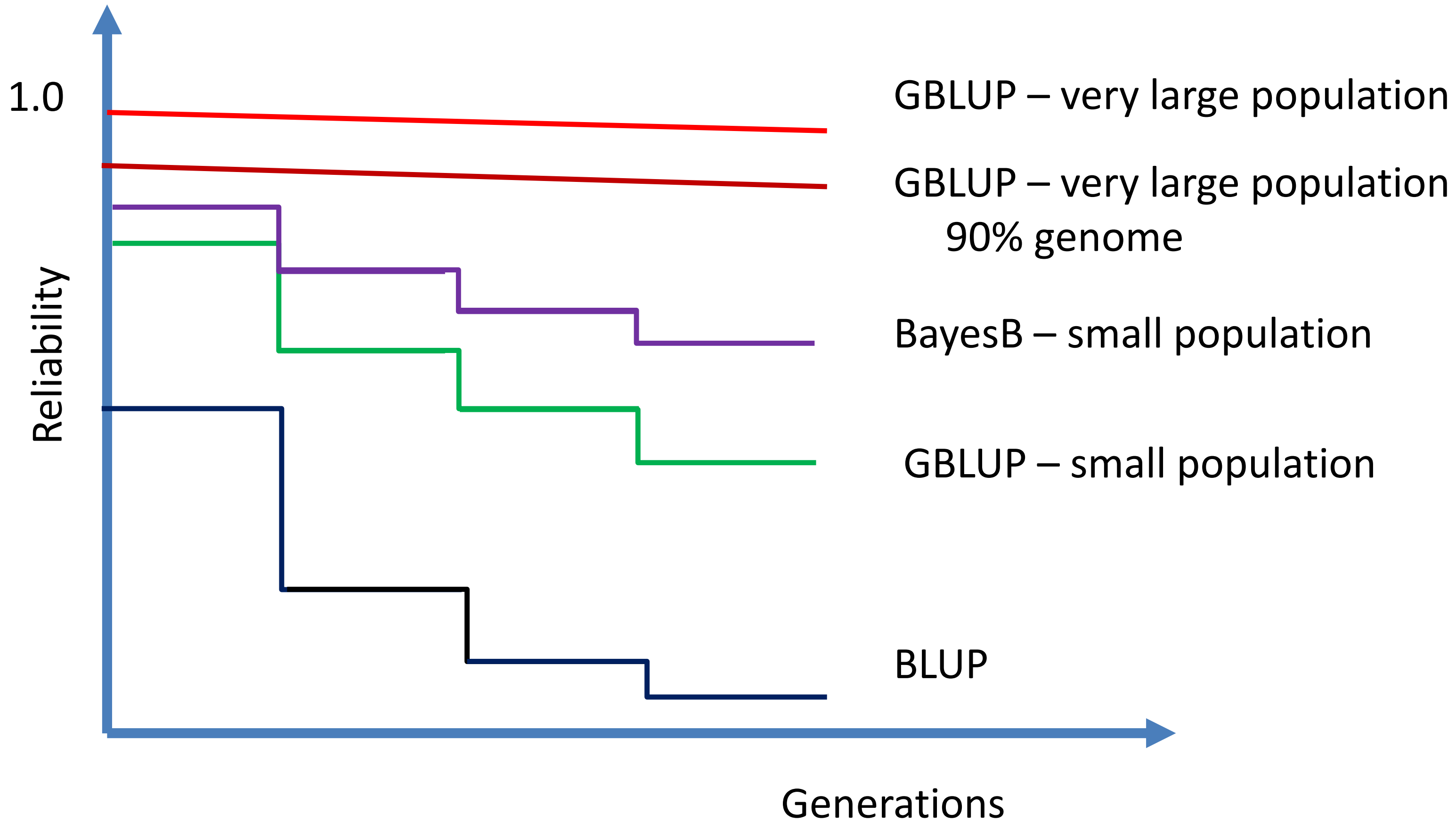
- Core animals:
  - Random gen 6 || gen 7 || gen8 || gen9 || gen 10 (y)
  - Random all generations

Which core animals in APY?



Bradford et al. (2016)

Persistence over generations

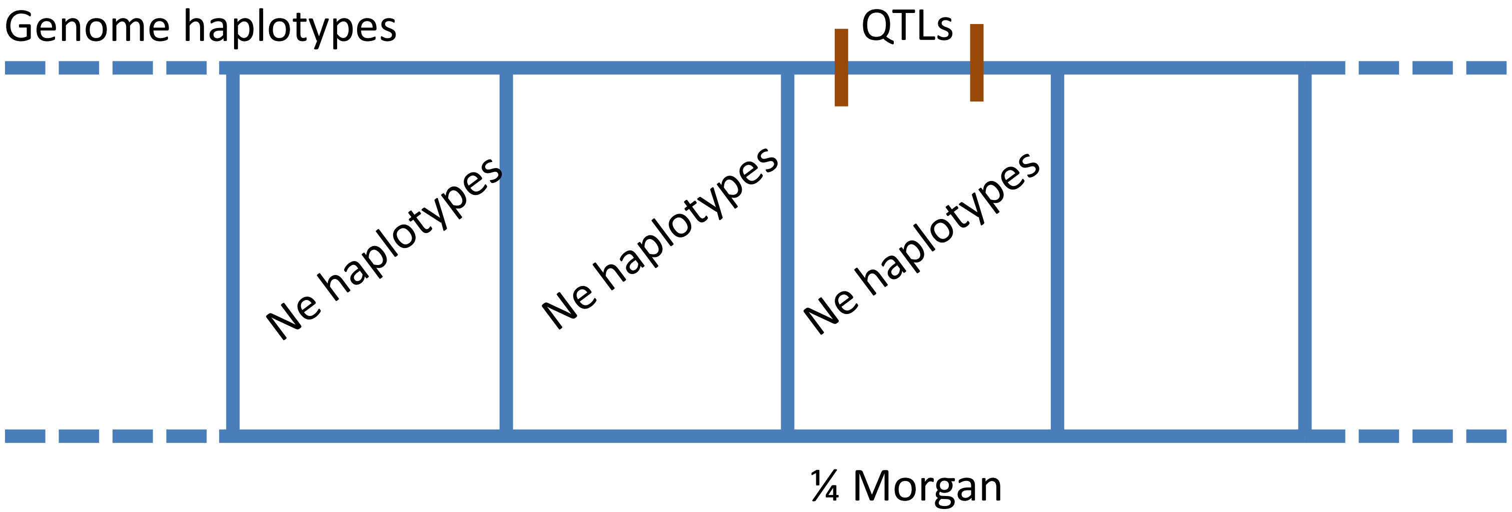


Very large – equivalent to 4NeL animals with 99% accuracy
Are SNP effects from Holstein national populations converging

Theory of limited dimensionality

Number of haplotypes: 4 Ne L

Ne within each ¼ Morgan segment



Dimensionality of ¼ Morgan case: Ne

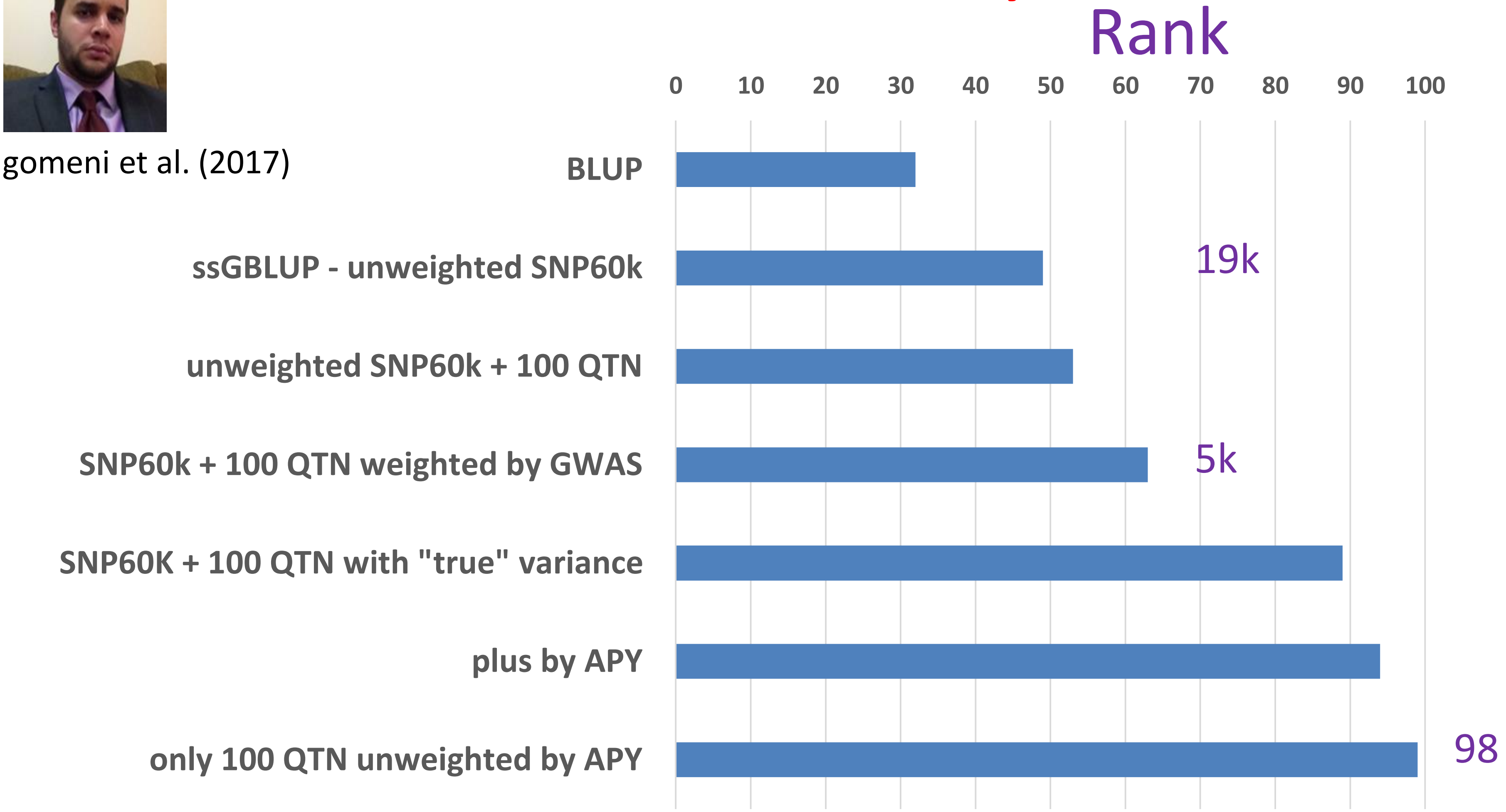
or number of identified QTLs

➔Reduced dimensionality with weighted GRM

Fragomeni et al., 2018

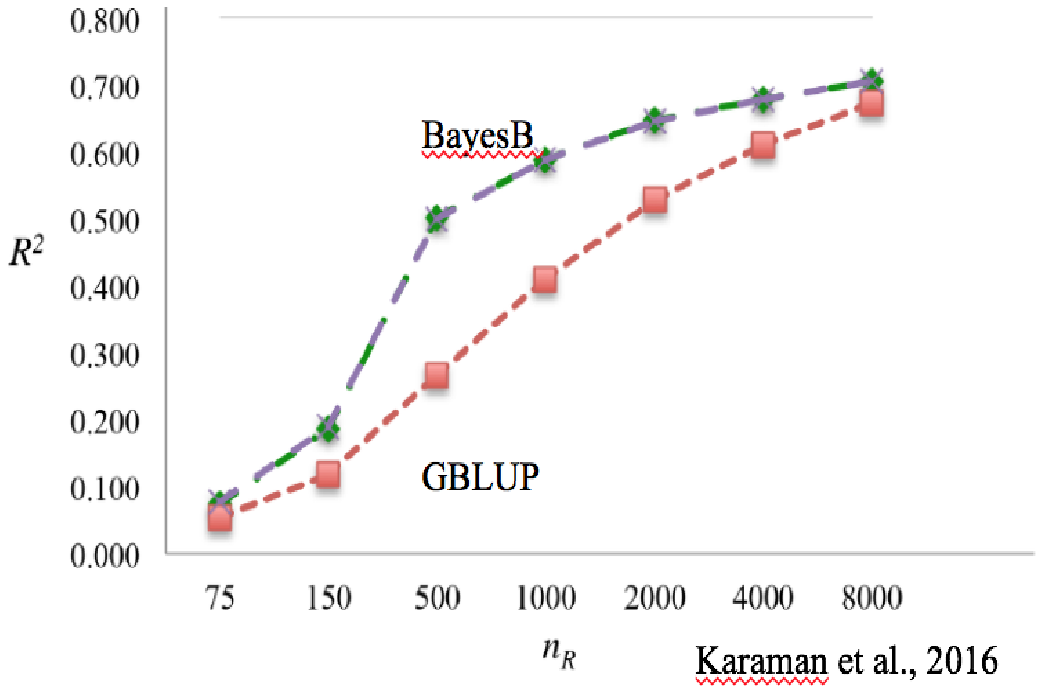
ssGBLUP accuracies using SNP60K and 100 QTNs – simulation study

Fragomeni et al. (2017)



Multitrait ssGBLUP or SNP selection?

- SNP selection/weighting (BayesB, etc.)
  - Large impact with few genotypes
  - Little or no impact with many



Karaman et al., 2016

Variance components

- Based on SNP
  - limitations
- REML based on relationships
  - Equations no longer sparse
  - YAMS sparse matrix package –up to 100 times speedup (Masuda et al., 2017)
  - APY for REML
- Method R (Legarra and Reverter, 2017)

Extra topics

- Matching pedigrees and genomic relationships
- Missing pedigrees
- Crossbreeding
- Causative SNP

- Haplotypes for crossbreds (Christensen et al., 2016)
- Metafounders (Legarra et al., 2016)
- Approximation of reliabilities

Conclusions

- Limited dimensionality of genomic information due to limited effective population size

- ssGBLUP suitable for any data set and model

- With large data sets for Holsteins:
  - Good persistence of predictions
  - Convergence of predictions from different countries

United States Department of Agriculture

National Institute of Food and Agriculture

Acknowledgements

Tom Lawlor, Holstein Assoc
Paul VanRaden, AGIL USDA

Shogo Tsuruta

Ignacio Aguilar

Breno Fragomeni

Ivan Pocrnic

Daniela Lourenco

Yutaka Masuda

Andres Legarra

Heather Bradford

Theory for APY

- Breeding values of core animals linear functions of:
  - Independent chromosome segments (Me)
  - Independent effective SNP

- E(Me)=4 Ne L (Stam, 1980; VanRaden, 2008)

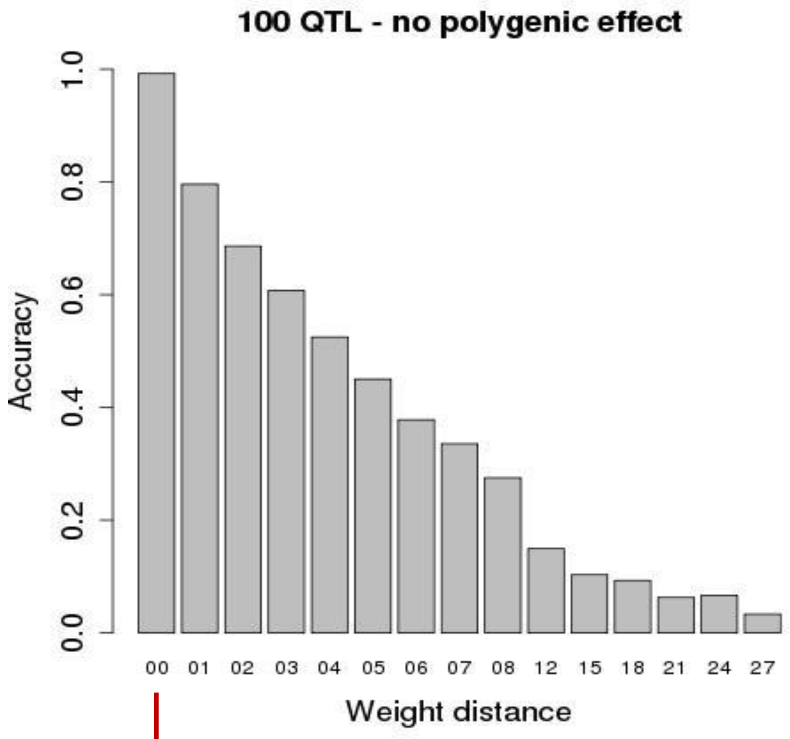
  Ne –effective population size

  L – length of genome in Morgans

  Me = 4 (Ne=100) (L=30) =12,000

Accuracy and distance from markers to QTL

Fragomeni et al. (2017)



QTL