Status of single-step and utility for Interbull / MACE

Ignacy Misztal, Breno Fragomeni, Yutaka Masuda, Daniela Lourenco, Shogo Tsuruta

The University of Georgia

Andres Legarra,

INRA SCIENCE & IMPACT

INIA Instituto Nacional de Investigación Agropecuaria URUGUAY

Ignacio Aguilar

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

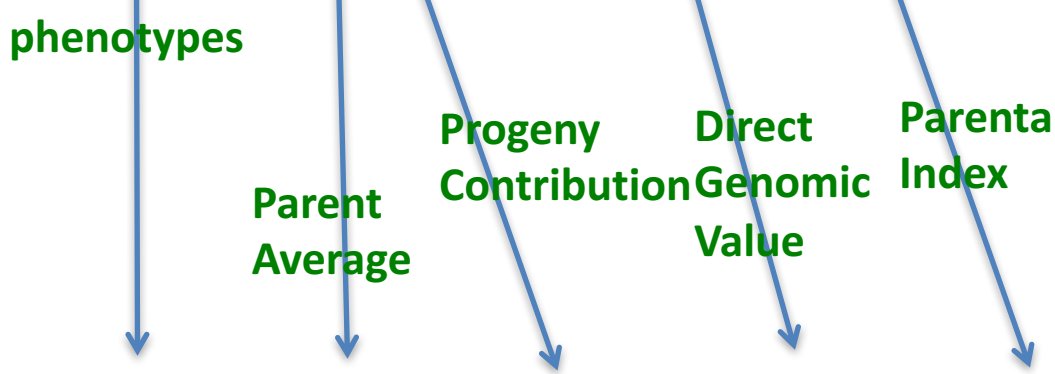
Tom Lawlor

Topics

- Decomposition of GEBV
- Convergence, costs and UPGs
- Is APY algorithm for inversion of GRM sound?
- SNP selection and accuracy
  - Causative SNPs
- Validation, etc.

Decomposition of GEBV in Single-step

$$\left\{ Z'MZ + \alpha A^{-1} + \alpha \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix} \right\} \hat{u} = Z'My$$



$$GEBV = w_1 CD + w_2 PA + w_3 PC + w_4 DGV - w_5 PI$$

GEBV for young animals

$$GEBV = w_2 PA + w_4 DGV - w_5 PI$$

PI=0 if genotyped animals unrelated
PI=PA if all animals genotyped
PI≈PA if parents genotyped

If genotyped but unrelated

$$GEBV = w_2 PA + w_4 DGV$$

PI=0 if genotyped unrelated

If genotype and parents genotyped

$$GEBV \approx DGV$$

PA and PI cancel out

GEBV

$$GEBV = w_1 CD + w_2 PA + w_3 PC + w_4 DGV - w_5 PI$$

For proven animals $GEBV = PC$ Genomics does not matter

If no genotype
No phenotype
No progeny

$GEBV = PA$ Little improvement with genomics if animal not genotyped

Output from single-step for MACE:

For bulls: PC (?)
For cows: CD (?)

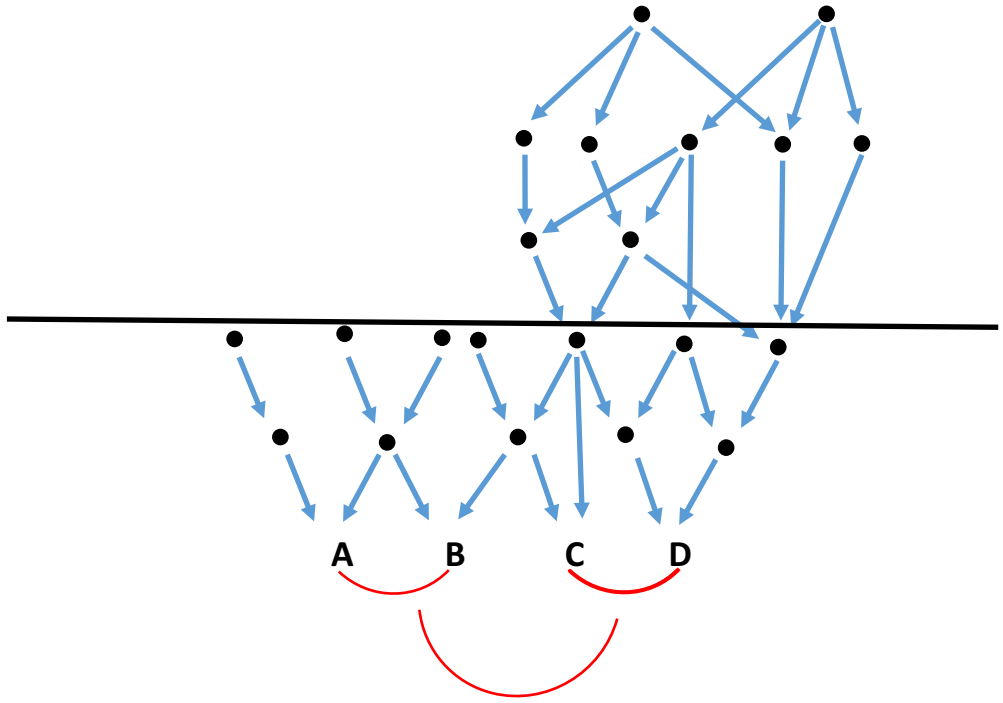
Extraction of components easy

Convergence problems in single-step

- No problem with some groups of animals (e.g., broilers with 3 generations of data/pedigree)
- Problems with other species
  - Smaller after cutting pedigrees
  - Larger with UPG
- One solution:

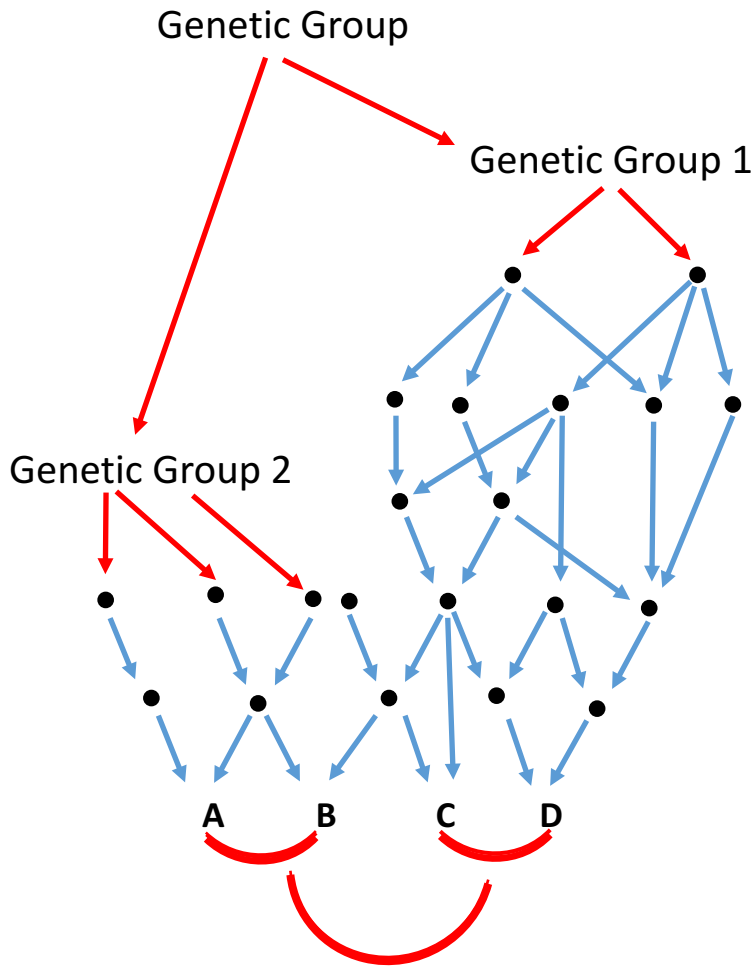
$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{0.7A_{22}^{-1}} \end{bmatrix}$$

Compatibility of G and A_{22}



- Pedigree same basis

Compatibility of G and A_{22}



- Genetic group / common founder
- Pedigree same basis
- Inbreeding

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

Why problems and solutions

- Incompatibility between **G** and $\mathbf{A}_{22}$
  - Inbreeding in $\mathbf{A}_{22}$ but not in **A**
  - Relationships in **A** function of missing pedigree
  - Modifications for UPG not fully included in **H**

- Solutions
  - Metafounders as generalized UPGs (Legarra et al., 2015)
  - Truncated data/pedigree and include UPGs in **H**

Unknown parent groups in A and H

$$\mathbf{A}^{*} = \begin{bmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{Q} \\ -\mathbf{Q}'\mathbf{A}^{-1} & \mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q} \end{bmatrix}$$

Quaas, 1988

Journal of Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Unknown-parent groups in single-step genomic evaluation

I. Misztal[1], Z.G. Vitezica[2], A. Legarra[3], I. Aguilar[4] & A.A. Swan[5]

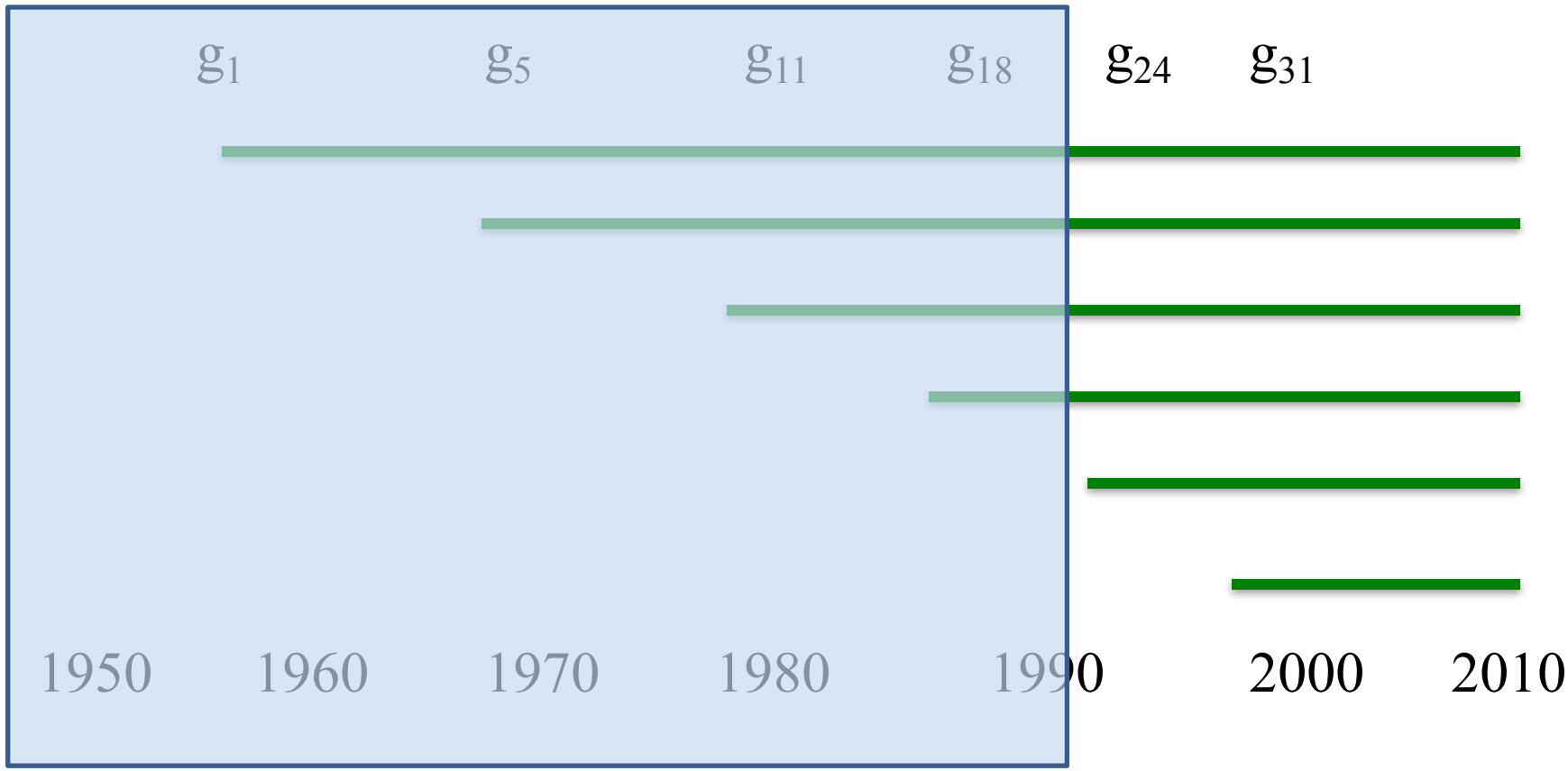
$$\mathbf{H}^{*} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right)\mathbf{Q}_2 \\ 0 & \mathbf{Q}_2'\left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) & \mathbf{Q}_2'\left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right)\mathbf{Q}_2 \end{bmatrix}$$

Seemed hard to implement

Not hard for Matailinen et al. (2016)

Low cost for Masuda et al.(2017)

Pedigree unifications via pedigree cut



Same or higher accuracy with cut data/pedigree (Lourenco et al., 2014)

Results of mods

- 18 trait model for type - Holsteins
  - Before mods: ~ 4,000 rounds
  - After UPG mod ~ 550 rounds, like BLUP
  - After cutting ~ 500 rounds, same REL

- Time per round < 2 x BLUP
  - single trait: 20s/round
  - 18 traits: 60s/round
  - 18 traits cut data: 45s/round

Dimensionality of genomic infromation

BV — SNP effects

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

$$\mathbf{Z} = \mathbf{U}\ \ \boldsymbol{\Delta}\ \ \mathbf{V}$$

Singular value decomposition
U'U=I, V'V=I, $\boldsymbol{\Delta}$

$$\mathbf{G} = \mathbf{U}\boldsymbol{\Delta}\boldsymbol{\Delta}\mathbf{U}' = \mathbf{UDU}'$$

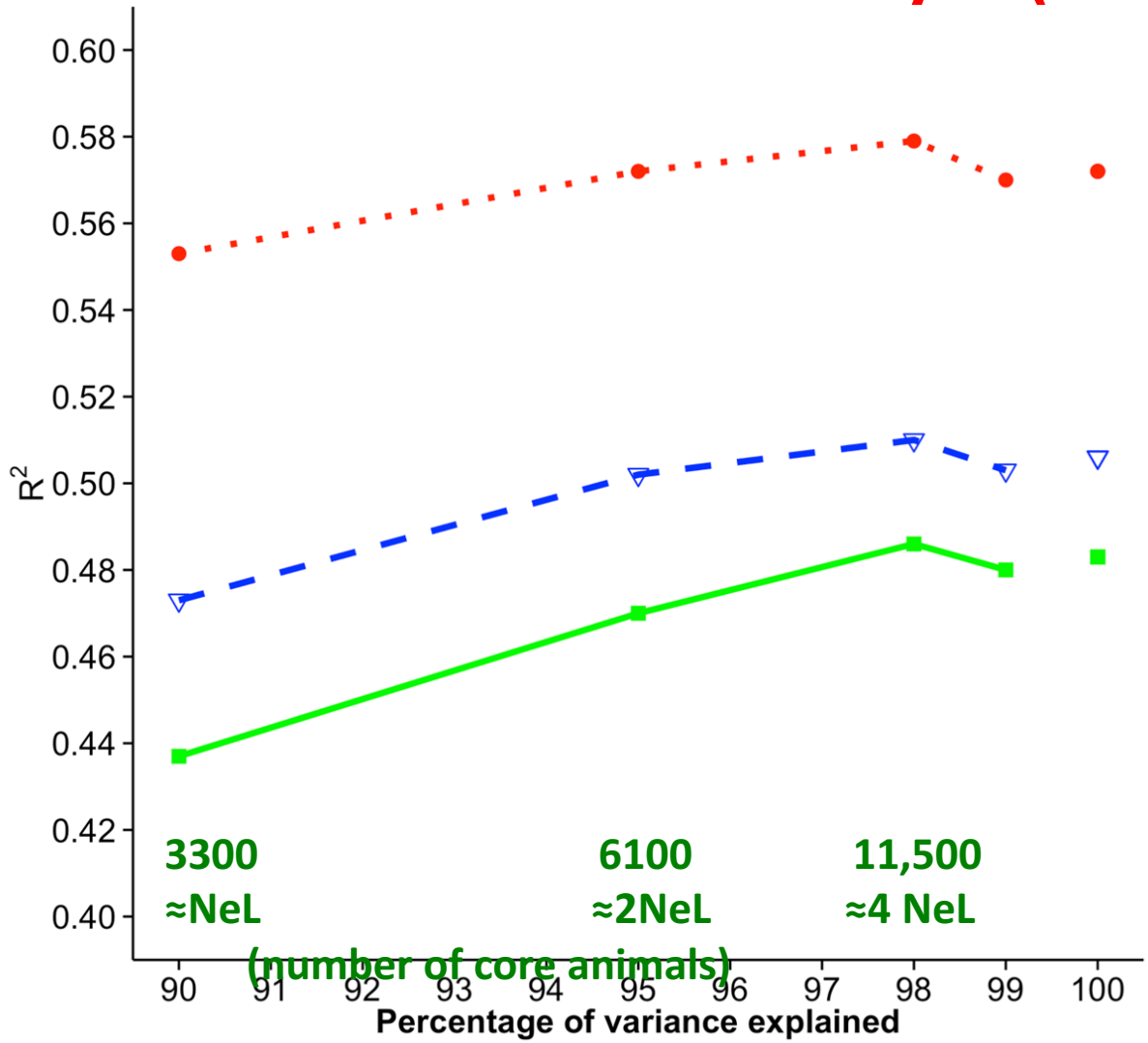
Genomic relationship matrix
Rank(G) ≤ min(#SNP,#anim)

$$\mathbf{Z}'\mathbf{Z} = \mathbf{V}'\boldsymbol{\Delta}\boldsymbol{\Delta}\mathbf{V}$$

SNP BLUP design matrix
Rank(Z'Z) ≤ min(#SNP,#anim)

Same dimensionality of gene content, GRM, and SNP BLUP design matrix

Reliabilities – Jerseys (75k animals)



Pocrnic et al., 2016b

Milk

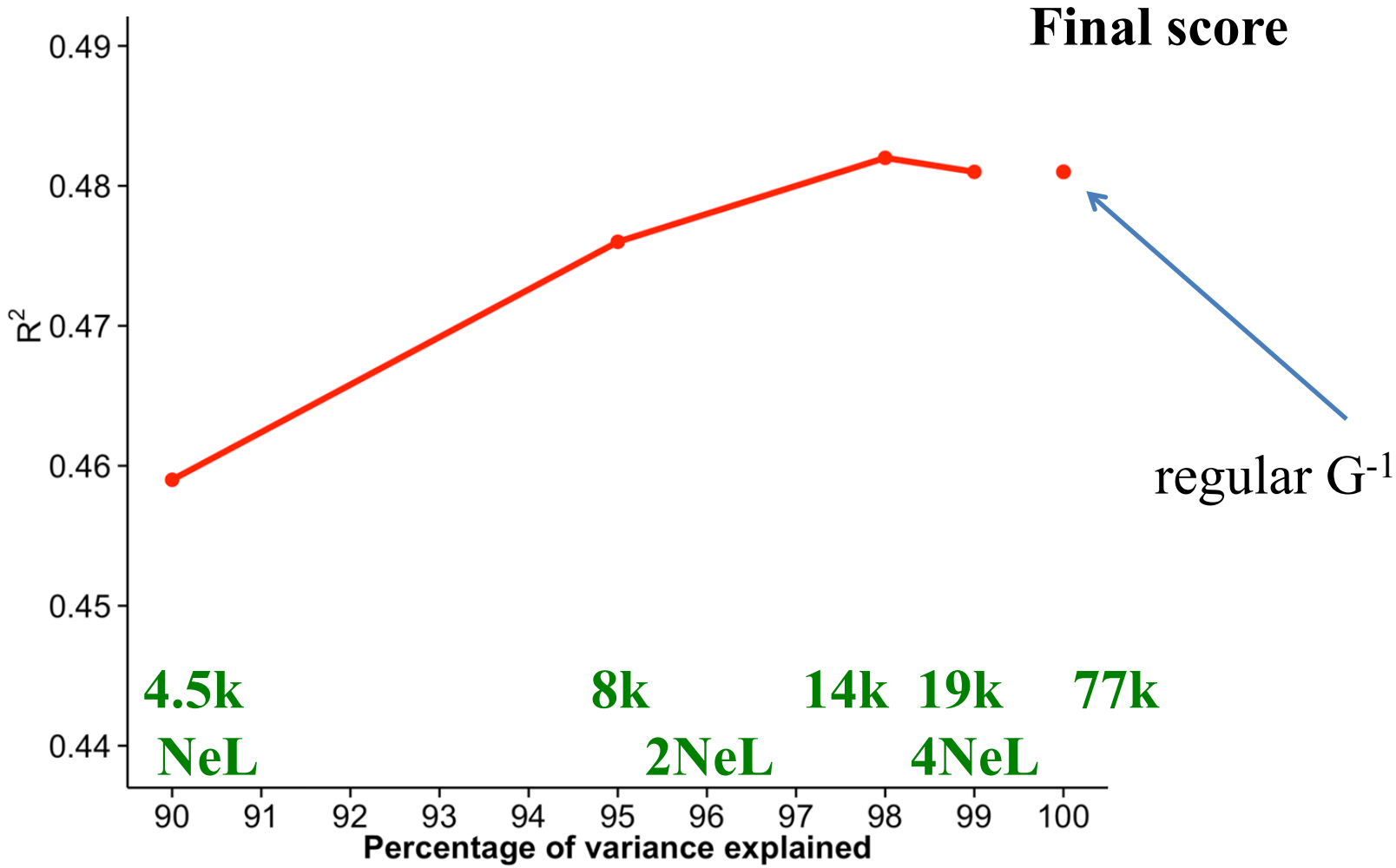
Protein

Fat

assumed dimensionality

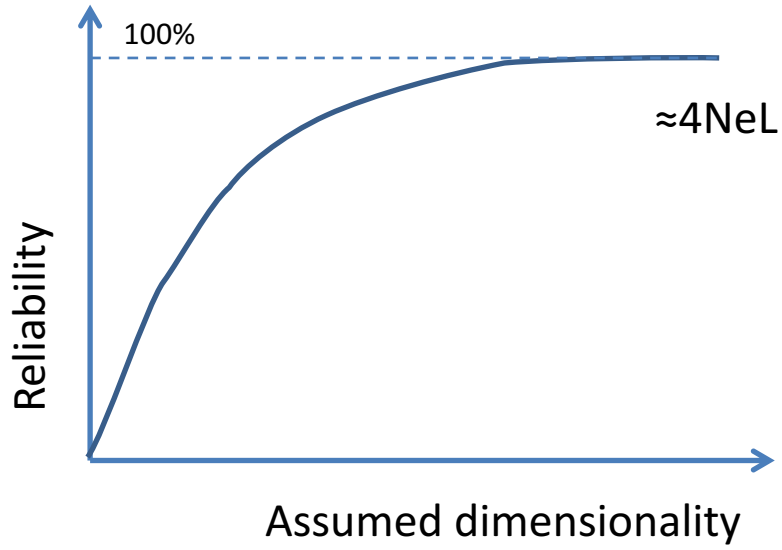
100% = full inverse ➔ lower accuracy

Reliabilities – Holsteins (77k)



Pocrnic et al., 2016b

Distribution of segments



Is inverse of GRM by APY sound?

s – n x 1 vector containing all additive information of population

Breeding value

Very small error

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

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

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

u of any n animals contain all additive information

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

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

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

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

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

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

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

$$\operatorname{var}\left(\begin{bmatrix}\mathbf{u}_p\\ \mathbf{u}_y\end{bmatrix}\right)=\begin{bmatrix}\mathbf{G}_{pp} & \mathbf{G}_{py}\\ \mathbf{G}_{yp} & \mathbf{G}_{yy}\end{bmatrix}\sigma_u^2$$

G is "true" relationship matrix

$$\mathbf{u}_y \mid \mathbf{u}_p = \mathbf{G}_{yp}\mathbf{G}_{pp}^{-1}\mathbf{u}_p, \quad \mathbf{P}=\mathbf{G}_{yp}\mathbf{G}_{pp}^{-1}$$

$$\mathbf{G}^{-1}=\begin{bmatrix}\mathbf{G}_{pp}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}+\begin{bmatrix}\mathbf{G}_{pp}^{-1}\mathbf{G}_{py}\\ \mathbf{I}\end{bmatrix}\mathbf{M}^{-1}\begin{bmatrix}\mathbf{G}_{yp}'\mathbf{G}_{pp}^{-1} & \mathbf{I}\end{bmatrix}$$

How to account for genetic architecture?

- Create SNP BLUP
  - Include regular SNP
  - Include causative SNP from sequence analysis
  - Estimate variance of each SNP

$$\mathbf{y} = \ldots + \mathbf{Za} + \mathbf{e},$$

$$\mathrm{var}(\mathbf{a}) = \mathbf{D}\sigma_a^2$$

- Create Genomic relationship matrix

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

Estimated dimensionality, effective population size and optimal number of SNP

Specie	Range of Me (95-99%)	Effective population size (L=30M)	Number of SNP (12 x Me)
Holsteins	8k-14k	149	100-180k
Jerseys	6k-12k	101	70k-150k
Angus	6k-11k	113	70k-130k
Pigs	2k-6k	43 (L=20M)	24k-72k
Chicken	3k-6k	44	36K-72k

Pocrnic et al. (2016b)

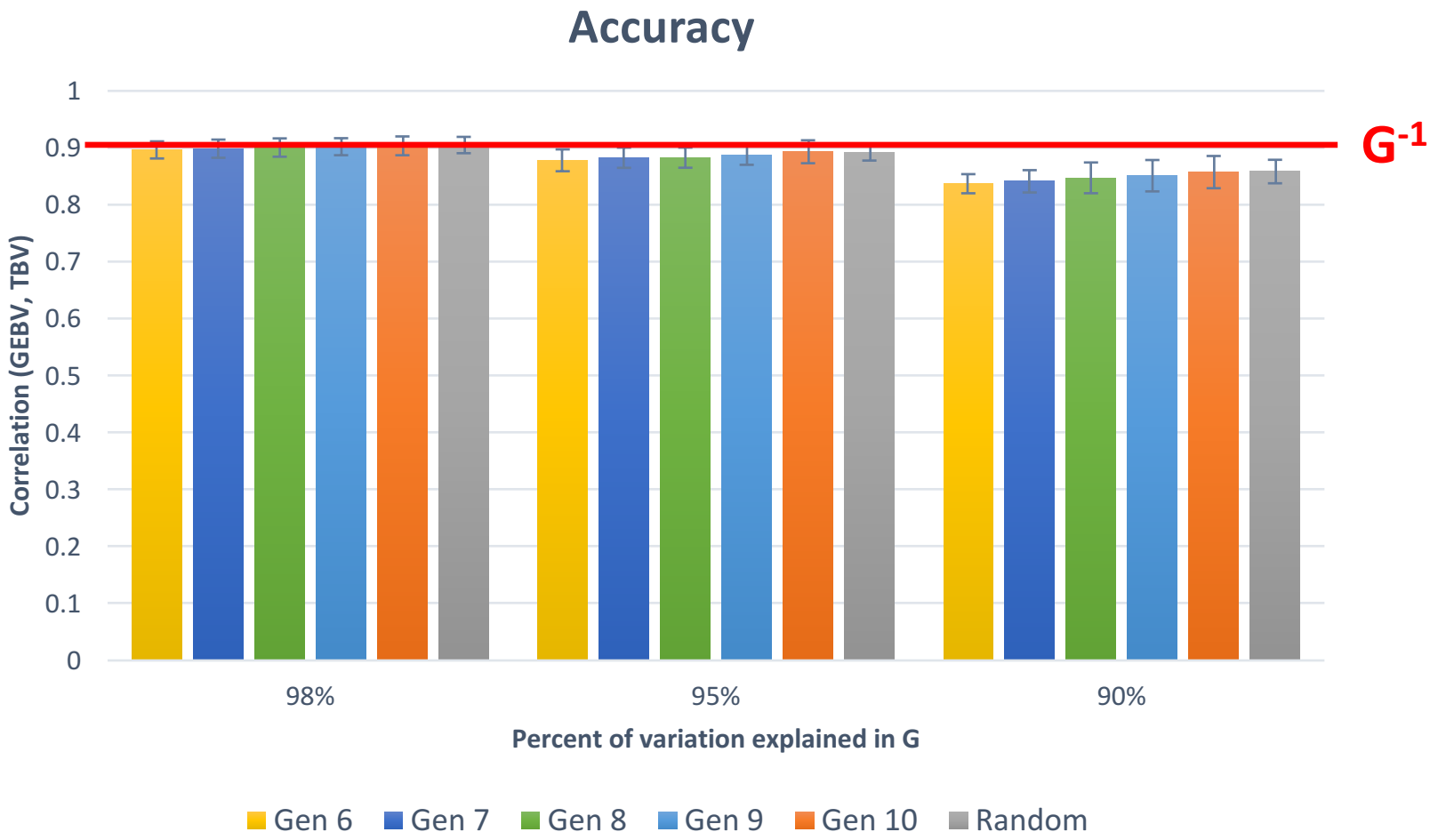
Which core animals in APY?

Bradford et al. (2017)

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

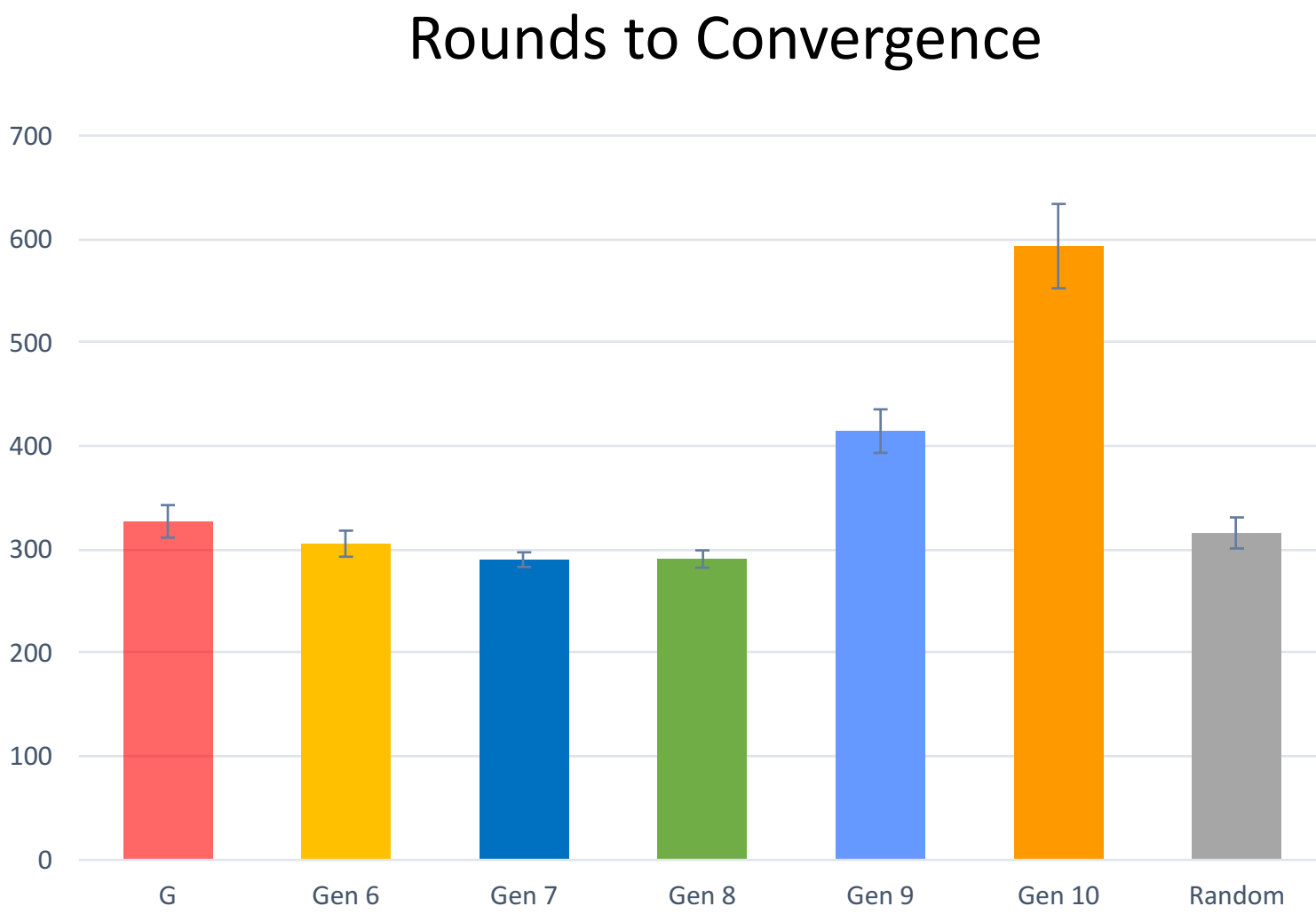
- Core animals:
  - Random gen 6 || gen 7 || gen8 || gen9 || gen 10 (y)
  - Random all generations
  - Incomplete pedigree
  - Genotypes in gen 9 and 10 imputed with 98% accuracy

Which core animals in APY?



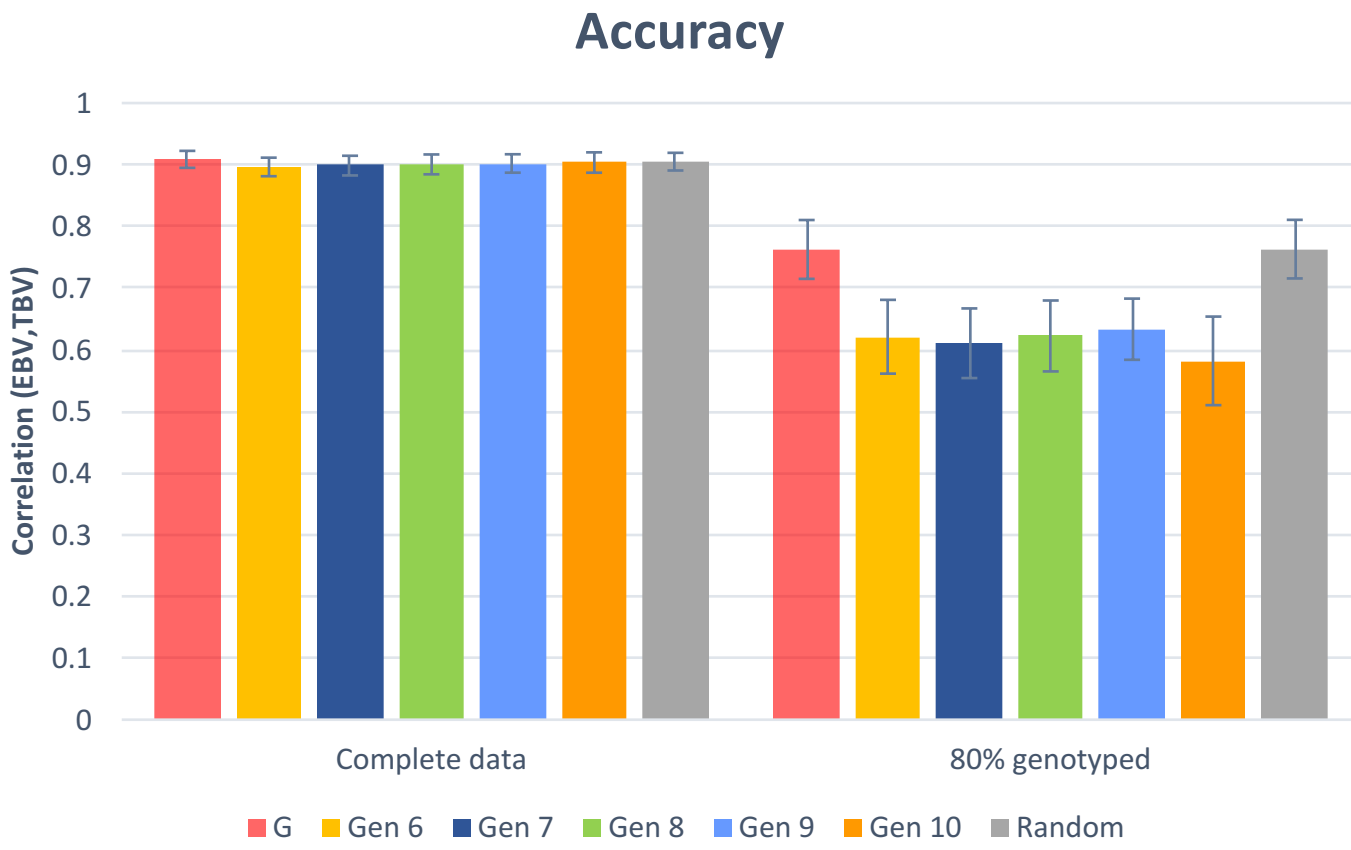
Bradford et al. (2016)

Which core animals in APY?



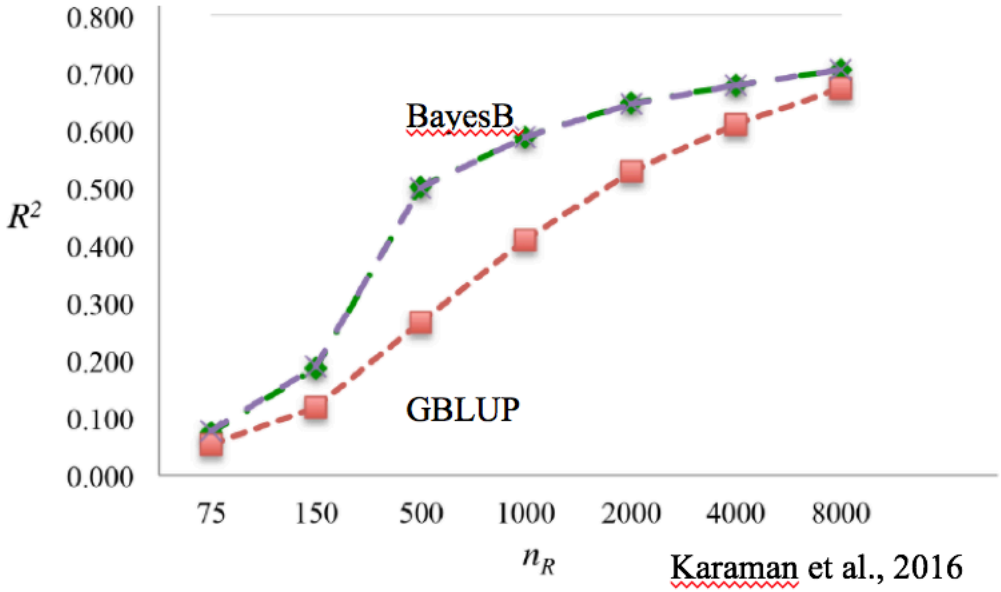
Which core animals in APY?

80% genotyped animals with missing pedigree



Multitrait ssGBLUP: Is SNP selection important?

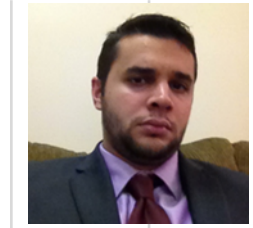
- SNP selection/weighting (BayesB, etc.)
  - Large impact with few genotypes
  - Little or no impact with many



Karaman et al., 2016

GBLUP accounts for QTLs when
genotypes ≳ chromosome segments?

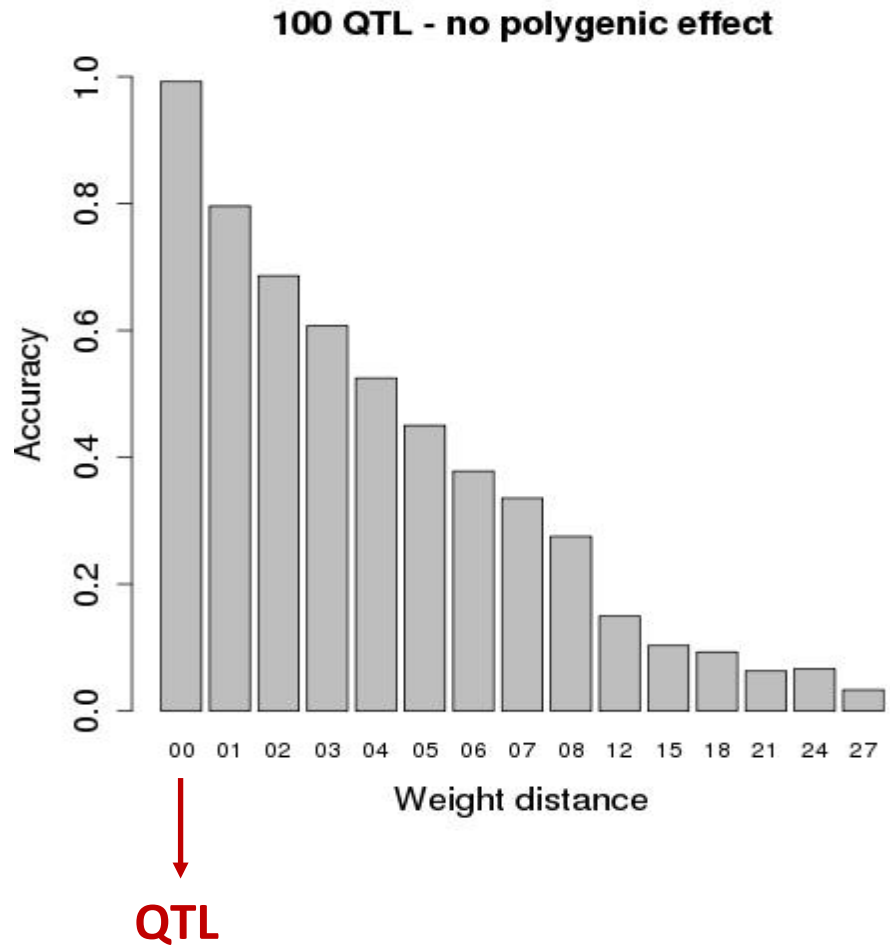
ssGBLUP accuracies using causative SNP



Fragomeni et al. (2017)

Accuracy and distance from markers to QTL

Fragomeni et al. (2017)



Extra issues

- Cross-validation by PEV
- Dimensionality and decay of genomic info
- Dimensionality 15,000: Is Eurogenomics == US data?

Conclusions

- Components of GEBV in single-step easily computed
- Single-step becoming computation viable
- APY algorithm sound
- Causative SNPs applicable to single-step – details
- Perhaps SNP selection not too important with many genotypes

Acknowledgements

- Grants from Holsteins Assoc., Angus Assoc., Cobb-Vantress, Zoetis, Smithfield, PIC,…
- AFRI grant 2015-67015-22936 from USDA NIFA
- Collaborators


Shogo Tsuruta


Ignacio Aguilar


Breno Fragomeni


Ivan Pocrnic


Daniela Lourenco


Yutaka Masuda


Andres Legarra


Heather Bradford