Efficient inversion of genomic relationship matrix by APY algorithm

Ignacy Misztal, Breno Fragomeni, Yutaka Masuda, Daniela Lourenco, Shogo Tsuruta
University of Georgia

Andres Legarra, INRA, France
Ignacio Aguilar, INIA, Uruguay
Tom Lawlor, Holstein Association

Advantages of single-step GBLUP

- Simplicity
  - No DYD or DP
  - No index
  - No complexity
- Accuracy
  - Avoids double counting
  - Avoids fixed index
  - Accounts for preselection bias

Current implementation of SS

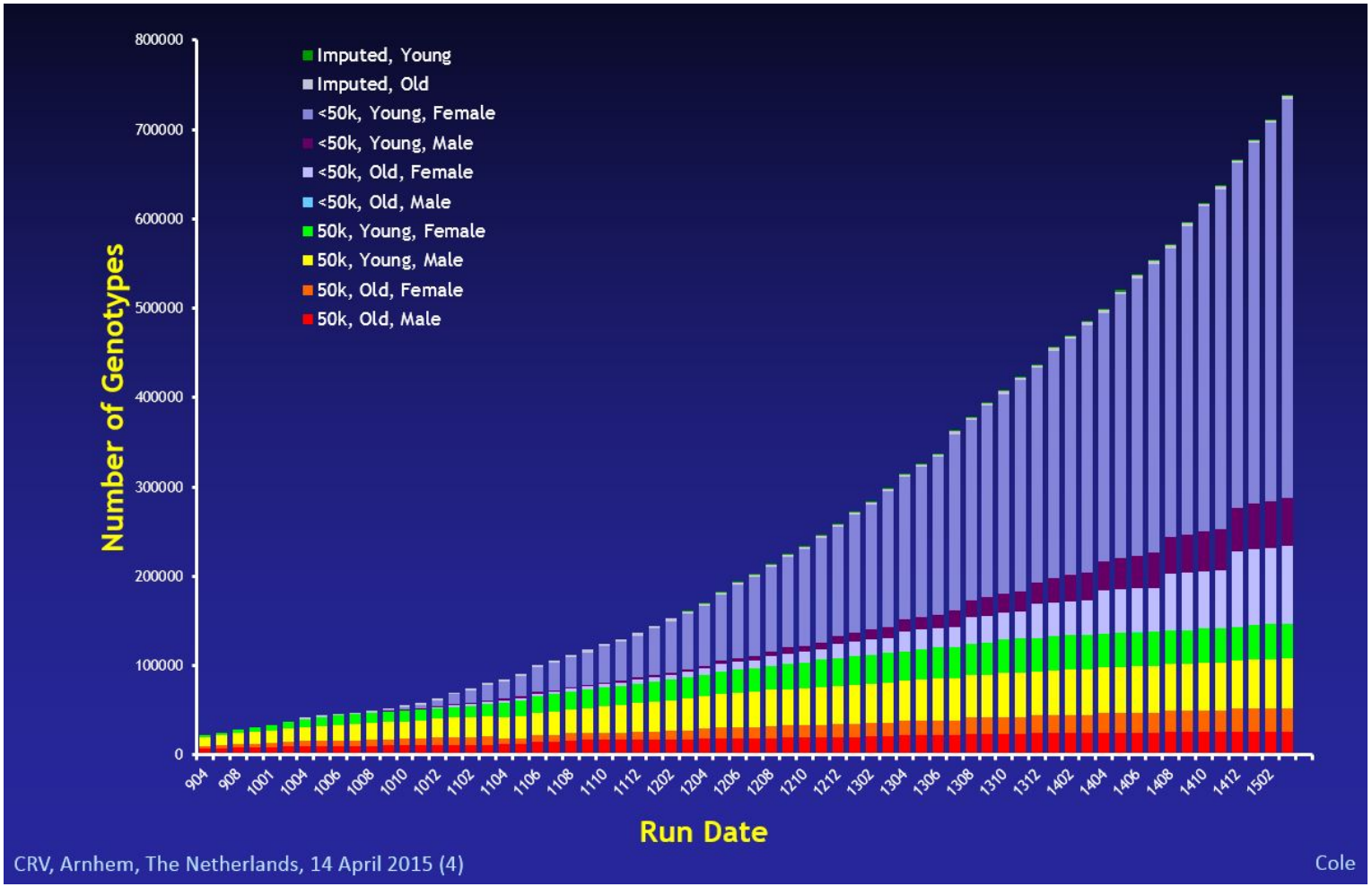
$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

- G and A_{22} created explicitly and inverted
- Cubic cost

- Cost per 100k genotypes - 1.5 hr (Aguilar et al.,2014)

Number of genotypes for US Holsteins

- Total ~ 800k in 2015



Options

- Unsymmetric Single-Step (Legarra and Ducrocq, 2011)
  - Does not converge
- SS SNP model with imputation for ungenotyped animals (Fernando et al., 2014)
  - Very expensive and new unproven machinery
- SS with SNP effects for genotyped animals only (Legarra and Ducrocq, 2011; Liu et al., 2014)
  - Does not converge

Recursions, Inversion, A^{-1}

$$u_i \mid u_1, u_2, .., u_{i-1} = \mathbf{p}_i{}' \mathbf{u} + \varphi_i$$ Generic recursion based on Cholesky decomposition

$$\mathbf{u} = \mathbf{P}\mathbf{u} + \Phi, \quad \mathrm{var}(\Phi) = \mathbf{M}\sigma_a^2$$ M diagonal

$$\mathrm{var}(\mathbf{u}) = \mathbf{A}\sigma_a^2 \qquad \mathbf{A}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P})$$

Inverse by recursion
High cost if P dense

Henderson (1976): $u_i \mid u_1, u_2, .., u_{i-1} = u_i \mid u_{si}, u_{di}$

$$u_i = \frac{u_{si} + u_{di}}{2} + \varphi_i$$

P - 2 nonzero elements per row
Cost of computing $\mathbf{A}^{-1}$ by inversion of **A** huge
Cost of computing $\mathbf{A}^{-1}$ indirectly by recursion trivial

Genomic recursions

$$u_i = \sum_{i=1}^{i-1} p_{ij} u_j + \varepsilon_i$$

$$\mathbf{p}_{i,1:i-1} = \mathbf{g}_{i,1:i-1}' \left(\mathbf{G}_{1:i-1,1:i-1} \right)^{-1}, \quad \text{var}(\varepsilon_i) = g_{i,i} - \mathbf{p}' \mathbf{g}_{i,1:i-1} = e_i^g$$

$$E(\mathbf{a} \mid \mathbf{b}) = \text{cov}(\mathbf{a}, \mathbf{b}) \, \text{var}(\mathbf{b})^{-1}$$

$$\text{Var}(\mathbf{a} \mid \mathbf{b}) = \text{cov}(\mathbf{a}, \mathbf{b})' E(\mathbf{a} \mid \mathbf{b})$$

$$\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{P})' \mathbf{M}^{-1} (\mathbf{I} - \mathbf{P}) = \mathbf{T}' \mathbf{M}^{-1} \mathbf{T}$$

Cost low only if P sparse

Recursion on relatives (Faux et al., 2011)

Number of genotyped US Holsteins in 2015

- Total ~ 800k
  - 25k proven bulls
  - 30k eligible cows
  - remaining cows and bulls not eligible for regular evaluation

Algorithm for proven and young animals (APY)

For young animals

=0 in GBLUP

$$u_i \mid u_1, u_2, \ldots, u_{i-1} = \sum_{j=\text{"proven"}} p_{ij} u_j + \sum_{j=\text{"young"}} p_{ij} u_j + \varepsilon_i$$

Misztal et al., 2014

p=proven y=young; G=ZZ'

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

$\mathbf{Z_p}$ – genotypes for proven animals

$\mathbf{Z_y}$ – genotypes for young animals

$m_i = g_{ii} - \mathbf{z_i}'\mathbf{Z_p}'\mathbf{G}_{pp}^{-1}\mathbf{Z_p}\mathbf{z_i}$

Inversion for $\mathbf{G}_{pp}$ only!

Linear cost for young animals

Tests with US Holsteins

(Fragomeni et al., 2015)

- US Holstein final score (h^2=0.31)
- 10.3M animals
- 11.6M records from 7.1M cows

All Genotyped	100k out of 569k
Bulls	23k
Cows	27k
Young Animals	50k

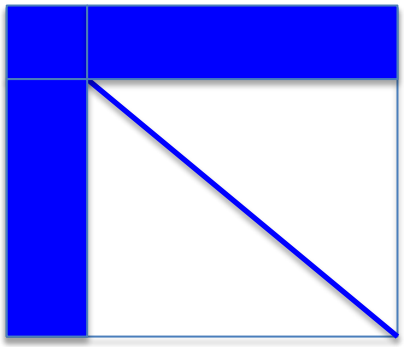
Correlations between GEBV with regular and APY G^{-1}

Treated as proven	Correlations	Rounds to conv
23k bulls	>0.99	432
23k bulls + 27k cows	>0.99	466
27k cows	>0.99	797
Random 20k animals	>0.99	~420
Random 10k animals	>0.98	~395
Random 5k animals	>0.97	~360

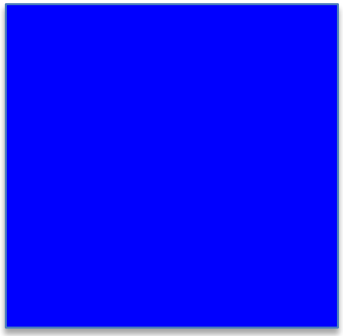
Random 5 samples

Results of APY

- High accuracy when > 10k animals in recursion



As accurate or more than



- Choice of animals not very important
- Best convergence with random samples

- "Proven" → Base "Young" → Nonbase

Everything should be made as simple as possible, but not simpler

Einstein

Theory for APY

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

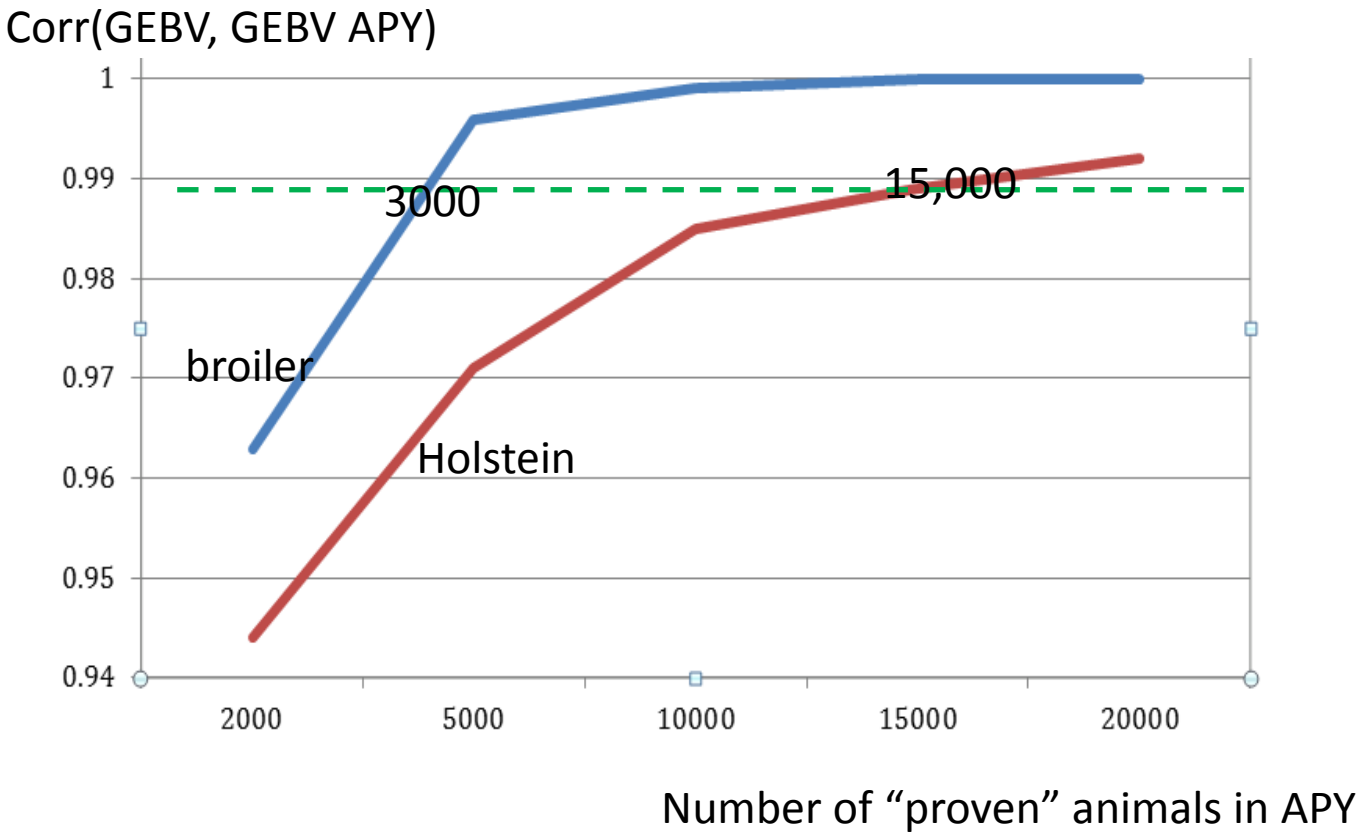
- Me=4 Ne L (Stam, 1980)

  Ne –effective population size

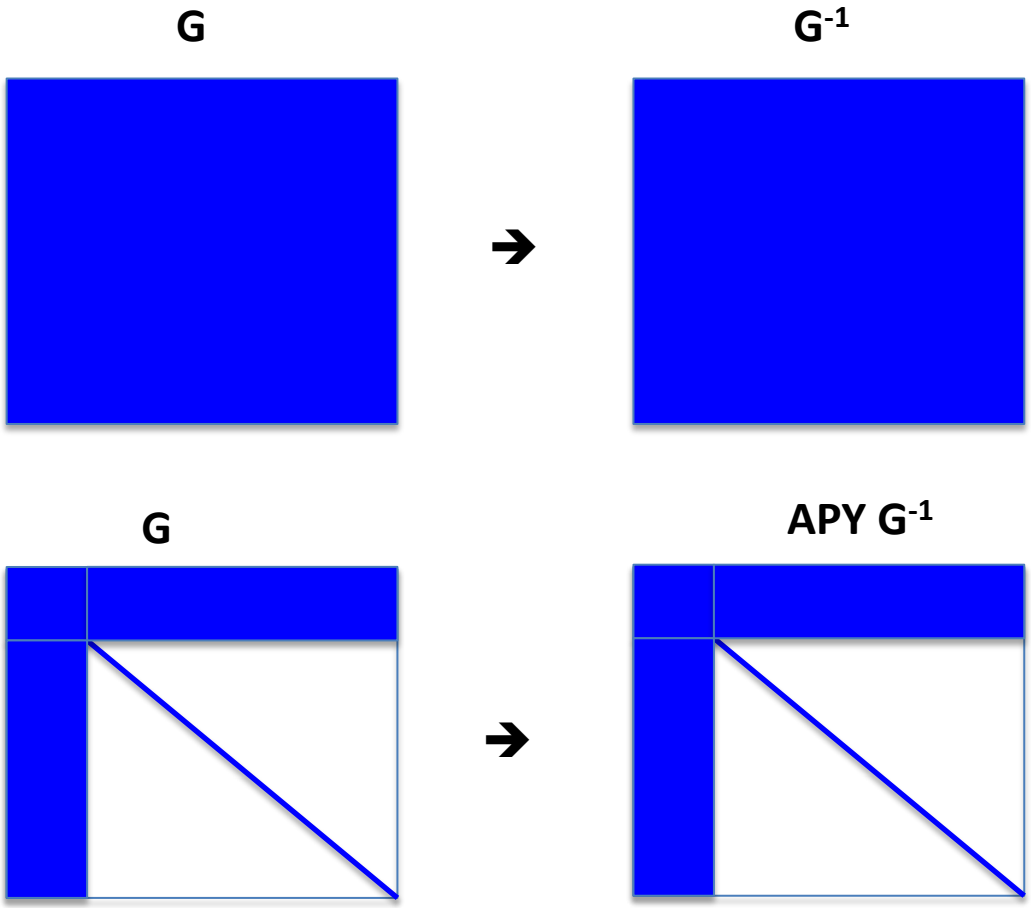
  L – length of genome in Morgans

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

Me in Holsteins and chicken



Why efficiency and accuracy of APY



Less **G** needed – less work and smaller noise from sampling variance

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

Henderson's (1976) algorithm for the genomic age

Questions and issues

- APY and major SNP or QTL
- Use of sequence data
  - Causative SNP with possible priors
- Number of independent chromosome segments, SNP density and GWAS resolution
- APY with Multibreed data
- MACE and External (G)EBV (Vanderplas et al., 2015)
- Use all genotypes in G^{-1} or indirect prediction for lower quality genotypes?

Conclusions

- Size limitations from single-step removed
- APY inverse applicable to SNP weights and causative SNP
- Potential to better address:
  - Multibreed evaluation
  - Comprehensive MACE

Acknowledgements

- Grants from Holsteins Assoc., Angus Assoc., Cobb-Vantress, Zoetis, Smithfield, PIC,…

- AFRI grant 2015-67015-22936 from USDA NIFA