A Unified Approach To Utilize Phenotypic, Full Pedigree, And Genomic Information For Genetic Evaluation

Ignacy Misztal, Ignacio Aguilar, Shogo Tsuruta, Ching-Yi Chen

University of Georgia

Andres Legarra, INRA Toulouse

Dave Johnson, LIC, New Zealand

Tom Lawlor, Holstein Association, USA

Selma Forni, PIC

Introduction

- Emphasis on genomic selection
- Seemingly new needs on software

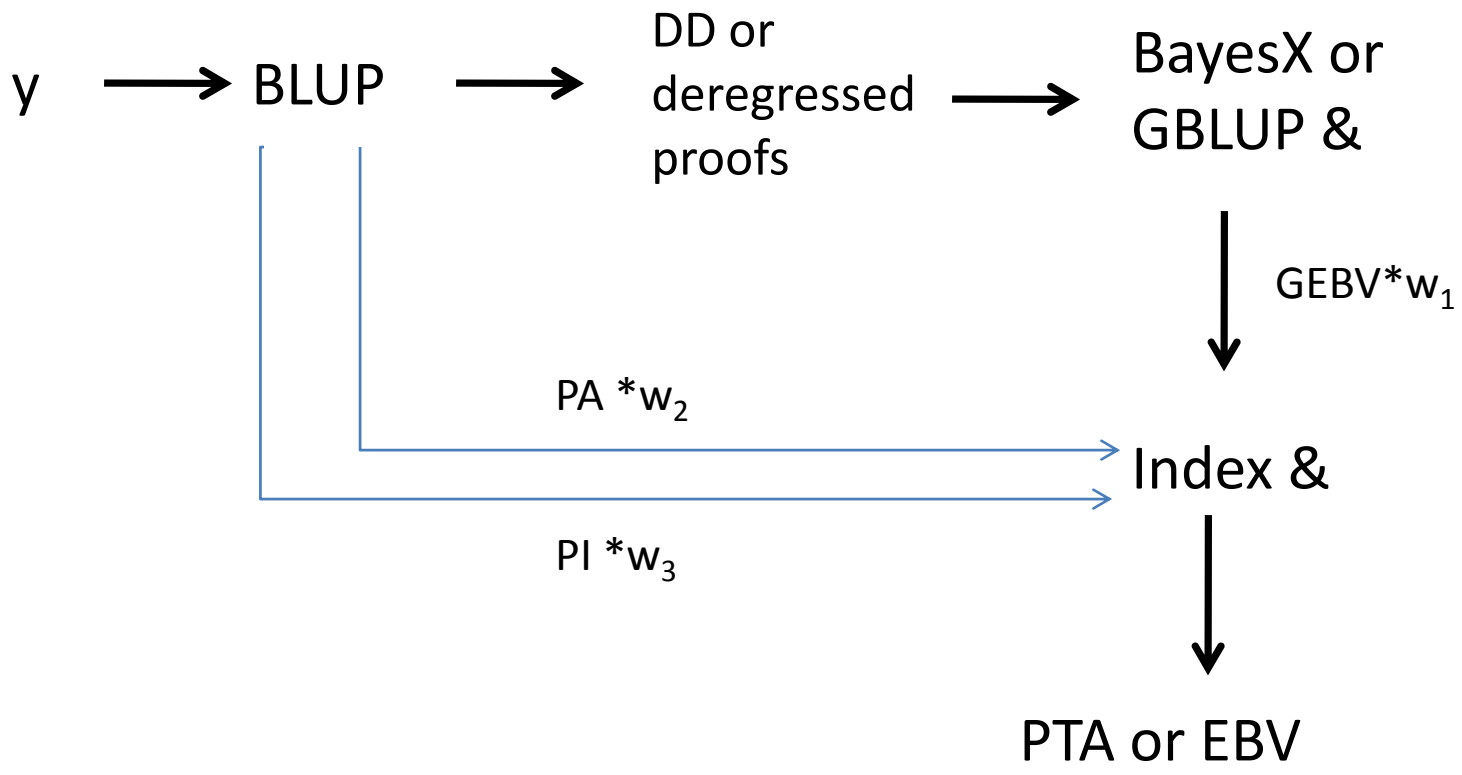
- New software or reuse old?

"It is a capital mistake to theorize before one has data.

Insensible one begins to twist facts to suit theories, instead of theories to suit facts"

Sherlock Holmes

Steps in genomic selection



Deregressed proofs an approximation

Uncorrelated residuals in BayesX an approximation

Index weights approximated

Equivalent "genomic" equations

(VanRaden, 2008, Hayes et al., 2009)

Pseudo-obs SNP effects

$$\tilde{y} = \mu + Za + e, \quad var(a) = D\sigma_a^2$$

breeding values

$$\tilde{y} = \mu + u + e, \quad var(u) = G\sigma_u^2, \quad G = ZDZ'/k$$

Z – centered design matrix

G – genomic relationship matrix

$$u = Za \qquad a = DZ'(ZDZ')^{-1}u$$

Stranden and Garrick, 2009

Genomic information ≈ genomic relationships

Easy conversion between BV and SNP effects

Typical result of assuming different SNP distributions

Method	*Average Correlation Across all traits*	
Bayes BLUP	*0.589*	**59**
Bayes A	*0.578*	58
Bayes C	*0.597*	60
LASSO	0.595	60
SVR	0.587	59
GBLUP	0.588	59
PLS	0.592	59

Verbyla et al, 2009

For traits with major genes – can use D from BayesX in GBLUP

Possibility of one-step evaluation

- Current BLUP evaluation

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \alpha \boldsymbol{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix}$$

- One-step genomic evaluation

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \alpha \boldsymbol{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix}$$

A – conventional numerator relationship matrix,
H - matrix modified to account for genomic relationships

Combined relationship matrix

Matrix H (Legarra ,2009)

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} [\mathbf{G} - \mathbf{A}_{22}] [\mathbf{I} \quad \mathbf{I}] \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

G –genomic relationship matrix

1 –ungenotyped animals

2-genotyped animals

Nonsymmetric equations

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{HZ'X} & \boldsymbol{HZ'Z} + \alpha \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{HZ'y} \end{bmatrix}$$

Equations for singular **H**: Harville (1978)
Computable after refinements (Misztal et al., 2009)

Inverse of H

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

Johnson; Aguilar et al., 2010

Christensen and Lund, 2010
Boemcke et al., 2010

Single-Step / Unified Method

- No need for pseudo-observations
- Problem of heterogeneous and correlated residuals nonexistent
- Weights determined automatically
- No model restrictions

- Little need for additional programming
  - Given new $\mathbf{H}^{-1}$, old programs should work

Questions with H

- Does it really work?
  - Accuracy, bias,…
- Is the choice of G critical?
- Costs of computing
- Effect on convergence in VCE and BLUP programs

Implementation at UGA

- Program PREGSf90 (Aguilar et al., 2010)
  - **G** with various options
  - $\mathbf{A}_{22}$ by Colleau (2002) algorithm
  - Inversion by optimized libraries
  - Minutes for 7k genotypes, hours for 30k genotypes

- Tweaks for:
  - renumbering program
  - “hash” matrix for variance component programs
  - iteration-on-data program (PCG iteration)

Predictions for US final scores in Holsteins (Aguilar et al., 2010)

Prediction in 2004	DD2009	
	R^2	δ (Regr)
Parent Avg	24	0.76
Multistep (VanRaden)	40	0.86
Single-step		
G^{-1}-A_{22}^{-1}	41	0.76
$1.5G^{-1}$-$0.9A_{22}^{-1}$	42	0.87
$1.5G^{-1}$-$0.6A_{22}^{-1}$	41	0.96

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

Timing

Creation of A_{22} , G, G^{-1}, and A_{22}^{-1} < 10 minutes

Evaluation – 2 hrs (similar to regular)

Weights on G^{-1} and A_{22}^{-1}

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

$$\mathrm{Var}(\mathbf{u}_2) = \left[\tau\mathbf{G}^{-1} + \left(1-\omega\right)\mathbf{A}_{22}^{-1} \right]^{-1}.$$

$$\mathbf{u}_2 \mid \mathbf{A}_{22}, \mathbf{G} \quad \sim \quad \mathrm{N}\left(0, \frac{\mathbf{G}}{\tau}\right) \mathrm{N}\left(0, \frac{\mathbf{A}_{22}}{1\text{-}\omega}\right)$$

τ – scaling factor

ω– fraction of information from genomics

Multitrait national genomic evaluation for type (Tsuruta et al., 2010)

- US Holsteins
- 18 traits

- Convergence rate dependent on G and τ
- Double time per round
- 2 weeks time for about 10 million Holsteins

Genomic evaluations of broiler chicken (Chen et al., 2010)

- 180k broiler chicken
- 3 k genotyped with SNP60k chip

- 3 methods
  - BLUP- full data
  - BayesA – genotyped subset
  - Single step – subset and full data set

Accuracies for the validation population

Trait	Accuracy*100			
	BLUP	BayesA Subset	Single-step Subset	Single-step Full
Body Weight	56	60	67	68
Breast Meat	35	36	35	41
Leg Score	29	9	6	36

Next cycle of selection

Multiple trait

Body Weight	38		51	60	=
Breast Meat	39		49	65	68
Leg Score	28		7	34	=

Which genomic relationship matrix

- Options for G
  - Assumed gene frequencies
  - Minimum minor allele frequencies
  - Scaling

- Study by Selma Forni et al. (2010)
  - 300,000 litter sizes
  - 2000 genotypes
  - Multiple G

Estimates of additive variance

Relationship matrix	Full data set	Genotyped subset
Pedigree (A)	1.26 ±0.03	2.3 ±0.5
Genomic (G)		
Equal gene freq	similar	3.5 ±0.6
Average min allele freq	similar	3.5 ±0.6
Normalized	similar	2.3 ±0.3

Corr (GBV, EBV)=0.78-0.79 => same ranking

Accuracies by inversion inflated with 1-2

Simulation (Vitezica et al., 2010)

- 10 generations
- Mass or EBV selection
- Fraction of all generations genotyped

	Corr (EBV,TBV) h^2=0.3	
Method	Mass selection	EBV selection
BLUP	20	23
2-step with DYD	46	52
Single-step	54	47
Single-step calibrated G	**55**	**60**

Strong bias

Little Bias

Practical limits of Single-Step

- Current limit 50k-100k genotypes
  - Cost → cubic with dense matrices
  - Possibly no limit if $\mathbf{H}^{-1}$ created directly

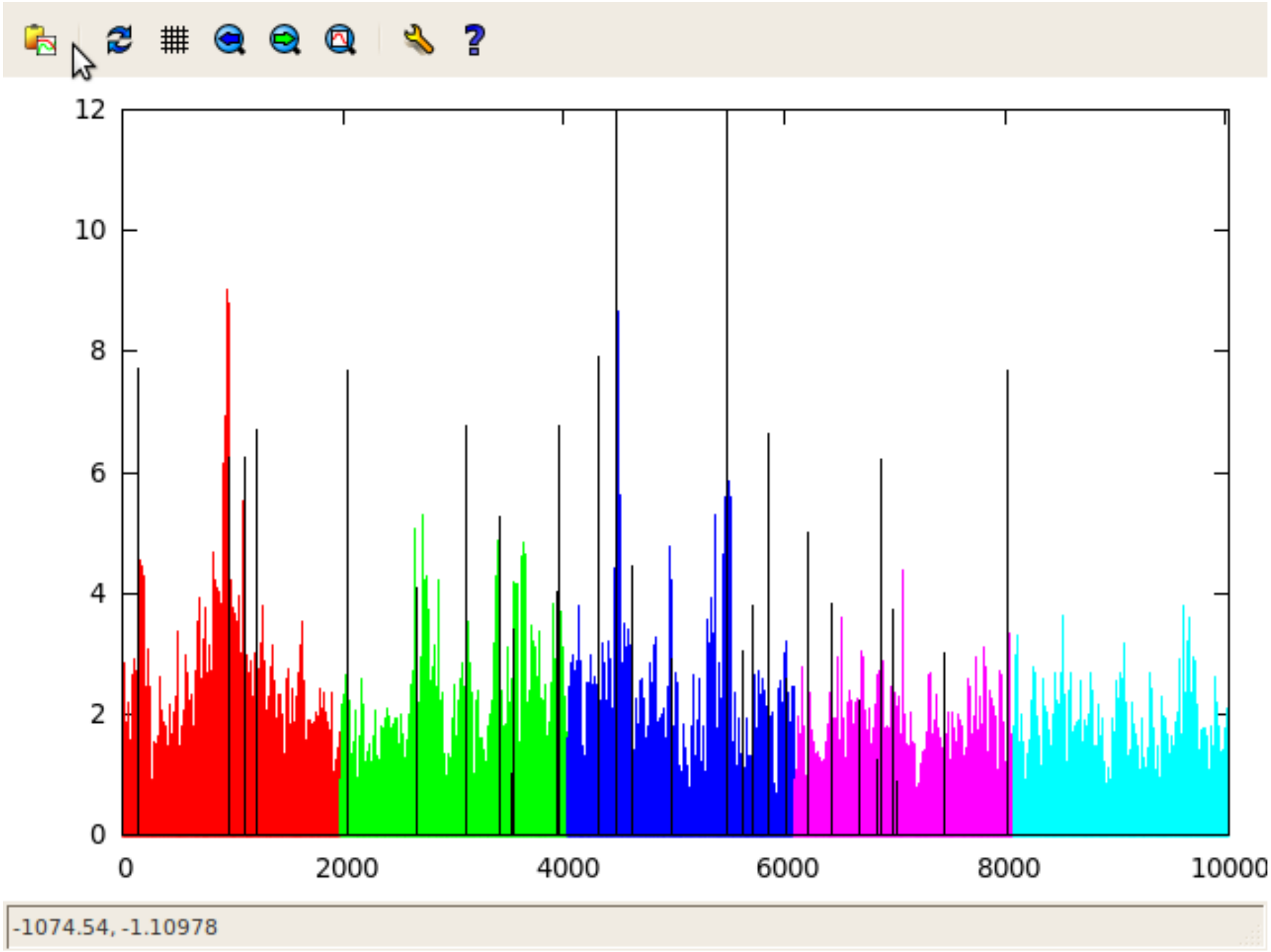
- If large data sets, two stage prediction
  - Main analysis with high accuracy animals
  - Indirect prediction for young animals
    - $\mathbf{u}_{new}=cov(\mathbf{u}_{new},\mathbf{u}_{old})\ var(\mathbf{u}_{old})^{-1}\ \mathbf{u}_{old}$
    - SNP effects

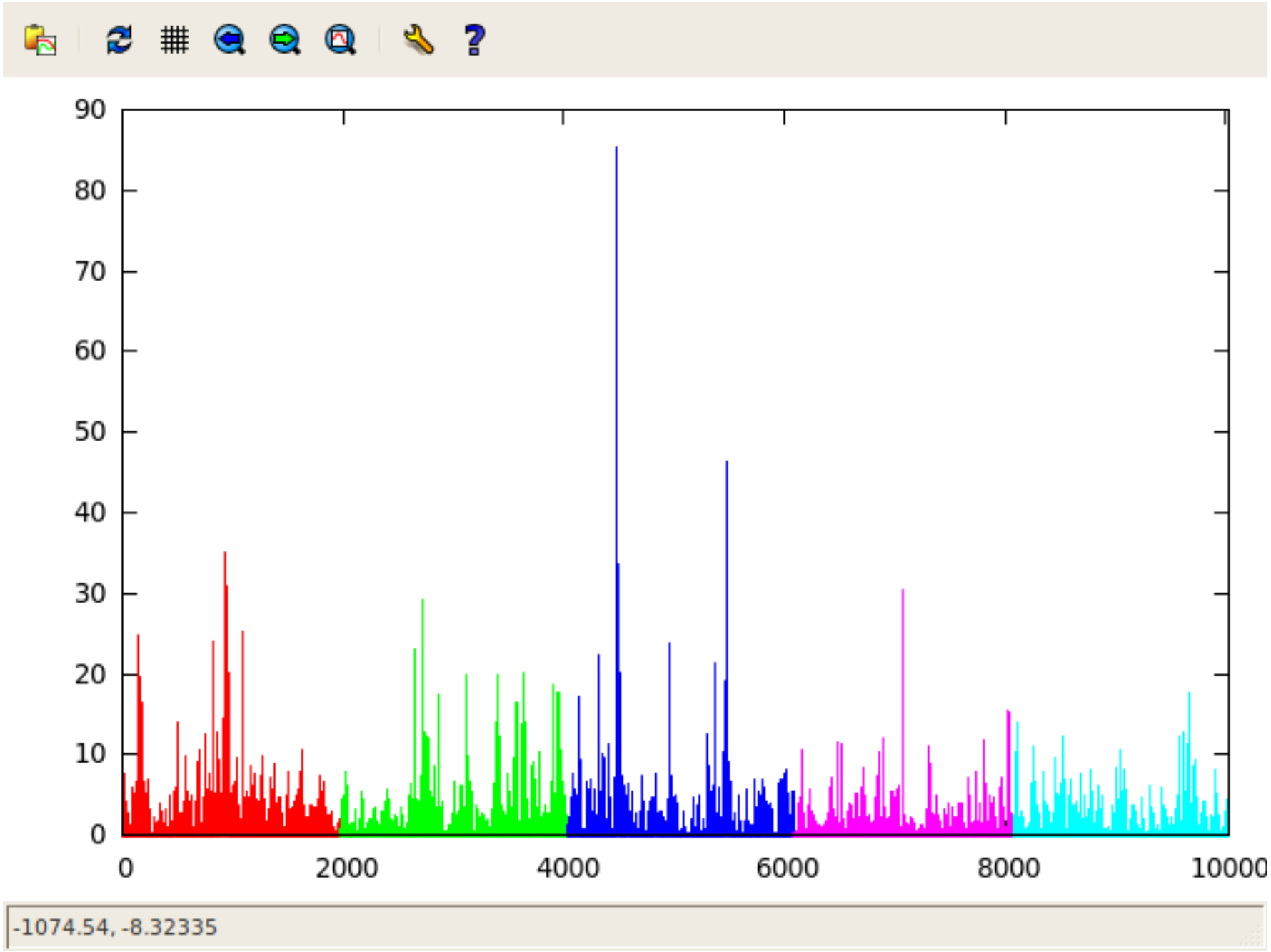
Future work / software

- Indirect predictions for young animals
- Approximation of accuracy
- Modeling different SNP chip sizes
- Direct creation of G^{-1} and H^{-1}
- Nonadditive effects
  - D=f(G), AA=G#G,….
- Genome wide association studies (GWAS)

Genome wide association studies

- Interest in finding major genes for research purposes (e.g., medical)
- Many spurious associations if heterogeneity of subjects ignored (Helgason et al., 2006)
- Associations better estimated if relationships considered (Kang et al., 2009, Visscher, 2010)
- Can single-step be an accurate tool for GWAS?


12
10
8
6
4
2
0
0
2000
4000
6000
8000
10000
-1074.54, -1.10978


90
80
70
60
50
40
30
20
10
0
0
2000
4000
6000
8000
10000
-1074.54, -8.32335

Conclusions

- Little extra software needed to implement genomic evaluation by single-step procedure
- Correct scale of G critical for some but not all applications
- Quantitative issues like modeling and selection still present with genomics (and even more important)
- Lots of exciting research to do

Acknowledgements

Paul VanRaden
Rohan Fernando
Bill Muir
George Wiggans
Curt VanTassel
Bruce Tier
Guilherme Rosa

NIFA grants
Holstein Association
Beef Cattle Consortium
Cobb-Vantress
Smithfield Premium Genetics
PIC