Present and future of genomic selection at the commercial level

Ignacy Misztal

University of Georgia

Genomic selection

- Many studies– simulation and field data
- Different results by methods and species
- Contradictions

- Studies at UGA with many data sets
  - Dairy (up to 75k genotypes), Chicken (up to 14k genotypes), Pigs (up to 5k genotypes), Sheep (up to 5k genotypes)
  -
  - Single-step, GBLUP and Bayesian Regression (BayesA, B, etc.; Bayesian Lasso);
  - weighted single-step/GBLUP: “poor man BayesB”: assign more weight to SNPs of large effect

- Is consistent picture emerging?
- If so, what next?

Experiences in dairy

- High accuracy if many genotypes
- Little improvement from adding female genotypes
- Little improvement with high density chip
- Little predictivity across breeds
  - Predictivity if mixed reference populations
- Smaller accuracy for animals with few ancestors genotyped
- Foreign genotypes may help or not

- Nonlinear/Bayesian Regression/weightedGBLUP help if major genes (Fat & Protein). Otherwise prone to errors

Experiences – other species

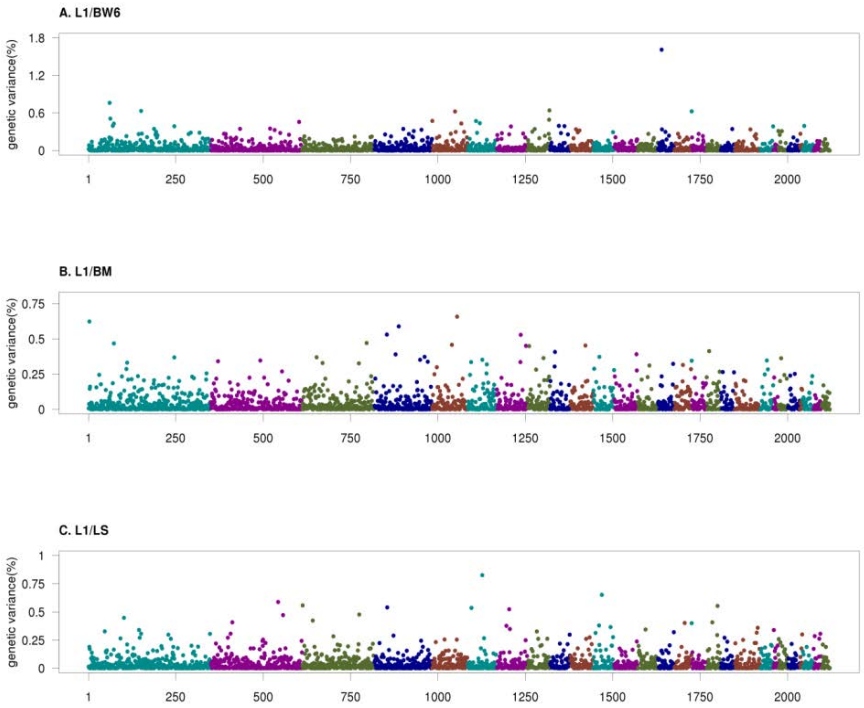
- Pigs
  - Moderate increases in accuracy with many genotypes
  - Outliers
- Chicken
  - Moderate increases with many genotypes
  - Males and females contribute
- Sheep
  - Moderate increases in dairy sheep
  - Small or no increases in meat with many genotypes

GBLUP or BayesX/nonlinear/… ?

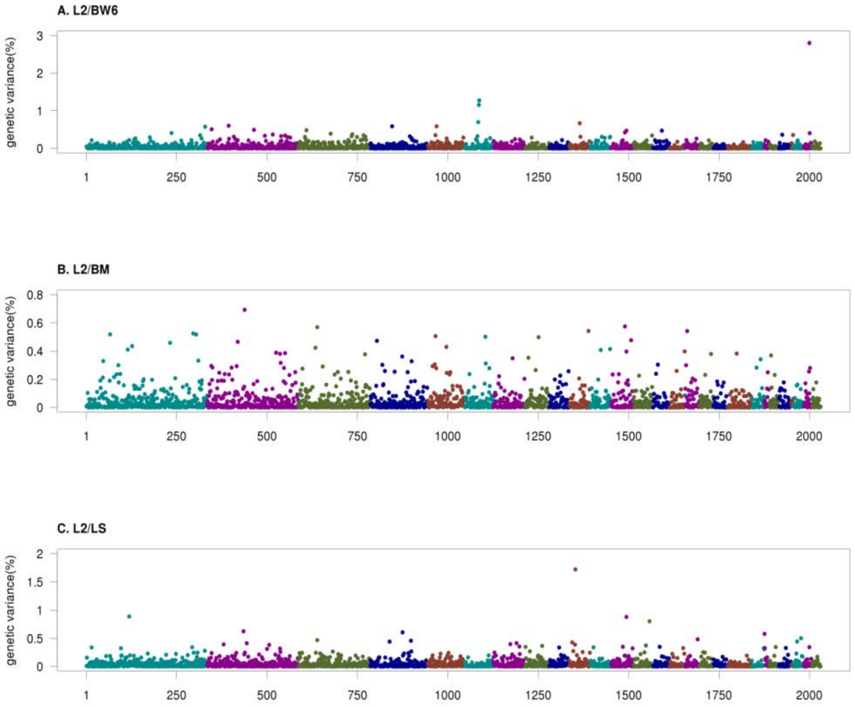
- Are major genes important?
  - If so ➔ predictivity across lines/breeds

Manhattan plots for 2 lines in chicken

Line 1



Line 2



Wang et al., 2013

No overlap across lines!

GBLUP/BayesB/.. – Questions?

- Can major genes remain in selected populations?
- Are major genes important in index selection?
- Advantage of Bayesian Regression greater if multigenerational genotypes – retained haplotypes?
- If SNP of large effect, can use W(eighted) GBLUP (fastBayesA; Sun et al. 2012) and W(eighted)ssGBLUP (Wang et al., 2012) – not based on sampling

Genomic accuracy in daughter equivalents

No contributions from other lines (except by LD)

Low contribution from low accuracy animals

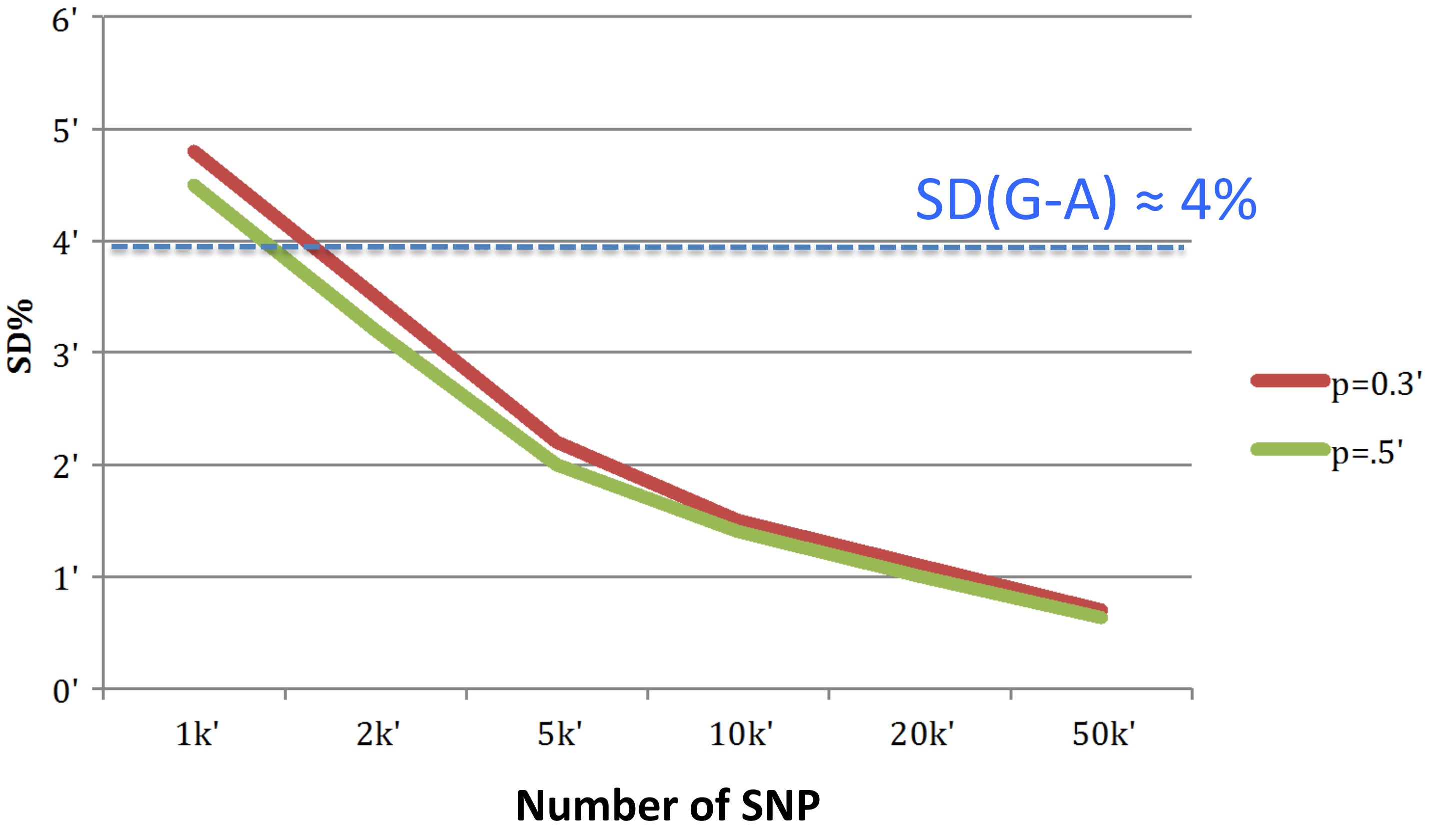
$$DE_i \sim \sum_{j,j\neq i} \left[(g_{ij} - a_{22,ij})^2 acc_j^2 \right]$$

SD≈0.04

summation over genotyped animals
DE – daughter equivalents
g_{ii} – genomic relationships
$a_{22,ii}$ – pedigree relationship
acc – accuracy

Misztal et al., 2012

Approximate SD of G with limited number of SNP



Small improvement beyond 20k

Specifics of species

- Dairy – large progeny groups
  - Holsteins –almost single population worldwide but some composites
- Pigs – Smaller progeny groups, open populations
- Chicken – small progeny groups, closed populations
- Sheep - single breed multibreed, heterogeneous data structure

Case studies at UGA

- Single-Step vs. multistep (deregressed proofs +Bayesian Regression)
  - Similar accuracies in dairy with many genotypes and high acc bulls
    - Pseudo-observations ≈ EBV, index with PA unimportant
  - SS better otherwise
    - No approximations due to pseudo-obs or index
  - SS worse if model deficiency

- Recent cases at UGA
  - Best acc for BLUP ➔ wrong validation
  - Large biases for one trait ➔ add one fixed effect
  - Low PA/GEN acc for specific traits➔ cut old data
  - Nearly zero GEN accuracies ➔bad imputation

All modeling or quality control issues

Future

- Genotyping costs decrease / possibly millions genotyped
- High computing cost at little marginal benefit? Or loss…
- Need tools for multi-trait genomic selection that unify all information:
  - Simple
  - Fool proof
  - Allow for model refinement and account for changes over time
- Single step a la Aguilar et. al. (2010), Meuwissen et. al. (2011),…,?

Computationally realistic?

Example: move from CC (Contemporary Comparison) to BLUP

Problem	Solution
Inversion of A expensive	Algorithm to create A^{-1} directly (Henderson, 1975)
Creation of MME expensive	Iteration on data (Schaeffer and Kennedy, 1986; Misztal et al., 1997)
Programming for complex models hard, poor convergence	PCG algorithm (Berger et al., 1987; Lidauer et al, 1999; Tsuruta et al., 2001)
Slow computing if many effects and traits: **LHS q = (W'W)q**	Sequential multiplication: **LHS=(W(W'q))** (Stranden, 1999)
Computing with GS expensive, cost~#genotypes$^{2\text{-}3}$	**Simple algorithm to account for all genomic info (?????, 2014?)**

Conclusions

- Better understanding of genomic selection
- Genomic selection for commercial use closer to maturity
  - Attention to detail of utmost importance !
- Breakthrough(s) welcome

Acknowledgements

- Discussions with countless individuals including Andres Legarra, Shogo Tsuruta, Ignacio Aguilar, Bill Muir, Zulma Vitezica, Selma Forni, Romdhane Rekaya,…

- Grants from Holsteins Assoc., Angus Assoc., Cobb Vantress

- AFRI grants 2009-65205-05665 and 2010-65205-20366 from the USDA NIFA

- BARD Project IS-4394-11 R

EBV for young animals

$$u_i = \frac{u_s + u_d + \sum_{j,j\neq i} (-\tau g^{ij} + \omega a_{22}^{ij}) u_j}{2 + \tau g^{ii} - \omega a_{22}^{ii}} =$$

$$w_1 PA + w_2 GEBV - w_3 GPI$$

PA = Parent Average

GEBV = Genomic EBV

GPI = Parental Index for genotyped animals

Large genomic information

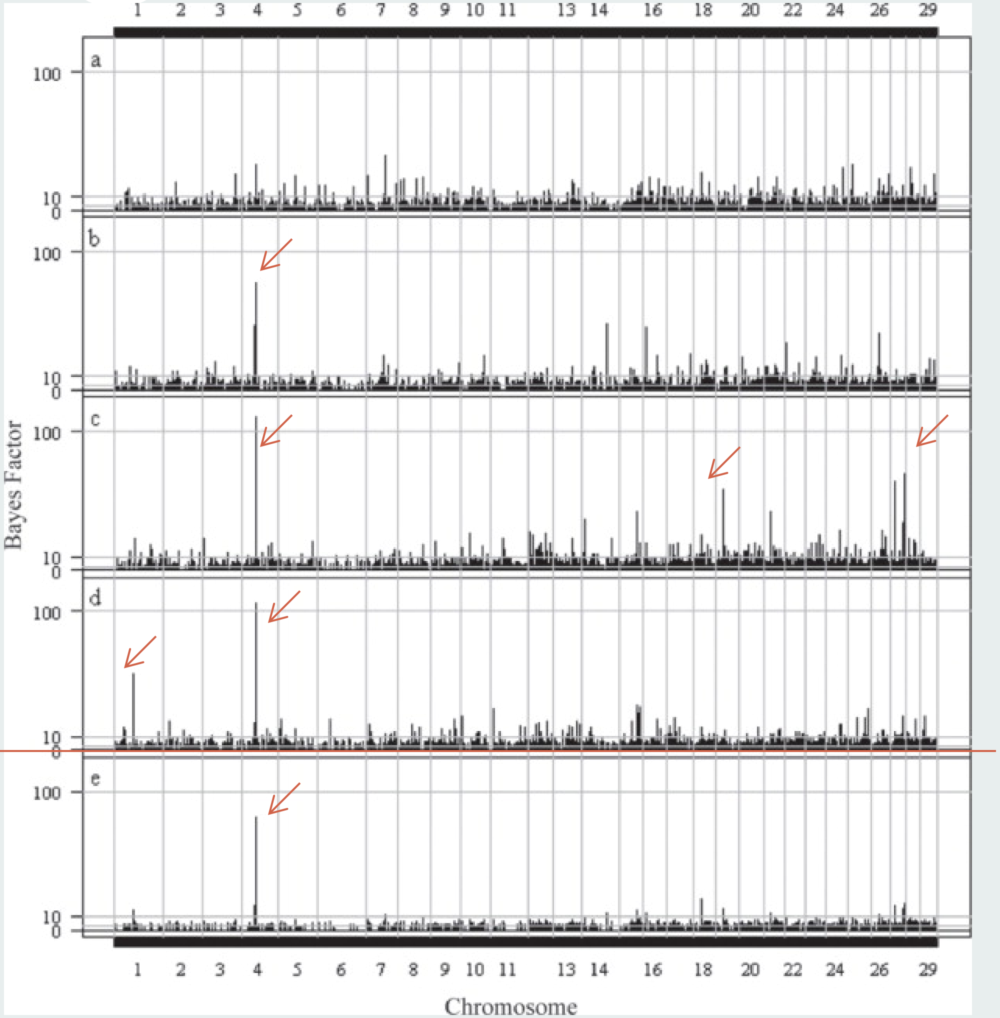
In dairy for popular bulls: g^{ii} □ 2, g^{ii} □ a_{22}^{ii}

$$u_i \approx \frac{\sum_{j,j\neq i}(-\tau g^{ij})u_j}{\tau g^{ii}} = \frac{\sum_{j,j\neq i}(-g^{ij})u_j}{g^{ii}} = \mathrm{GEBV}$$

Scaling factors cancel out

Prior influence: BayesX

- BayesB Chain 1
- Prior:
- Chain 2
-
- Chain 3
-
-
- Chain 4
- Average



van Hulzen et al. (2012) J. Dairy Sci. 95:2740-2748