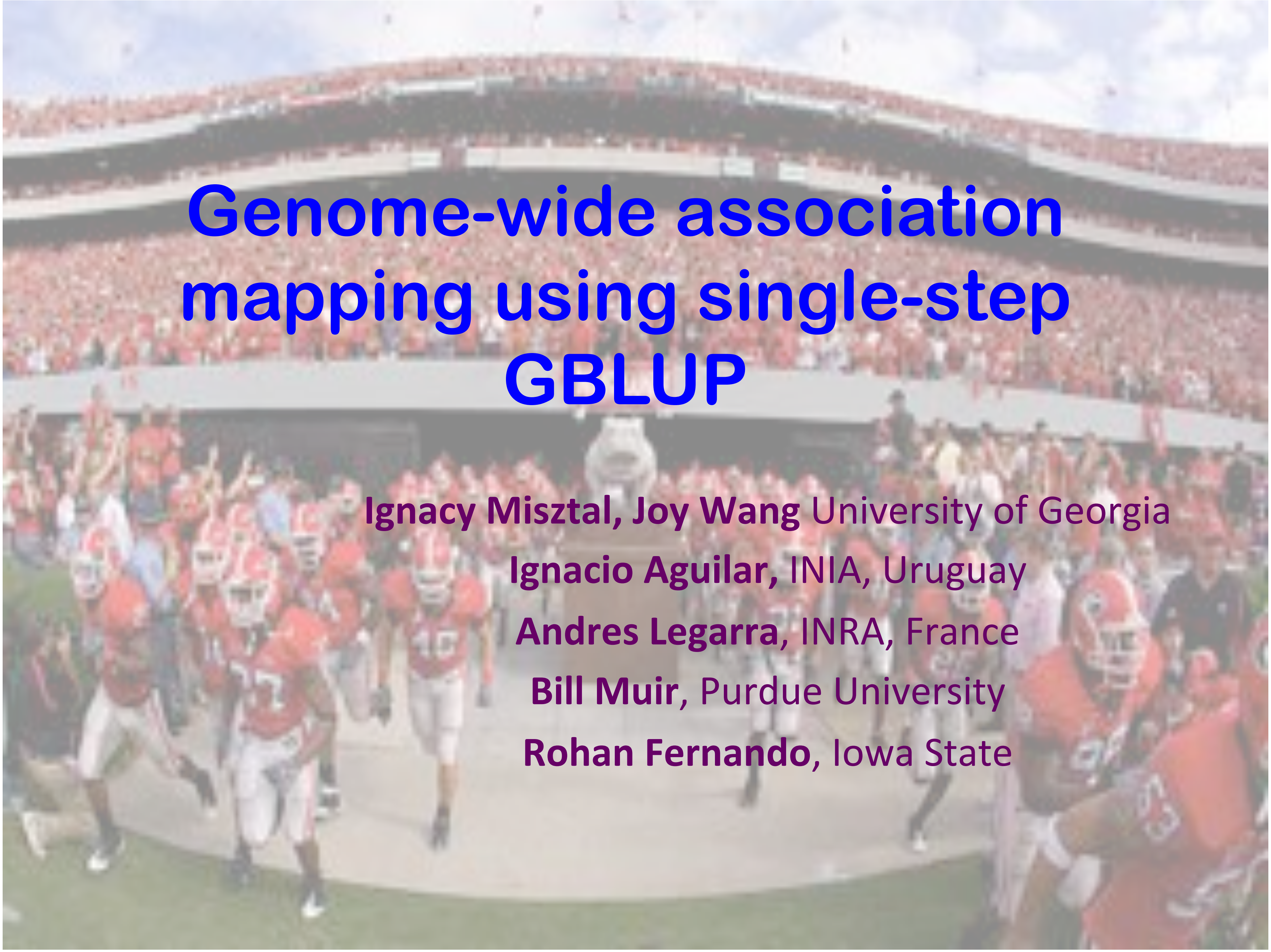


Genome-wide association mapping using single-step GBLUP

Ignacy Misztal, Joy Wang University of Georgia

Ignacio Aguilar, INIA, Uruguay

Andres Legarra, INRA, France

Bill Muir, Purdue University

Rohan Fernando, Iowa State

Genome Wide Association Studies

Large research interest in GWAS

Current methods

- Classical – single SNP analyses (e.g., Tassel, Wombat)
- BayesX - joint SNPs analysis (e.g., Gensel)

Limitations in current methods

- Simple models
- Single trait
- Slow if not optimized
- Complicated if not all animals genotyped

Single-step GBLUP

- **BLUP with combined pedigree-genomic relationship matrix** (Aguilar *et al.*, 2010; Christiansen et al., 2010)

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

- **Works with any model, any number of traits, and combination of genotyped and ungenotyped animals**
- **Can ssGBLUP be adapted for GWAS?**

Useful formulas

Conversion from SNP effects
To GEBV

GEBV

SNP
effects

$$\hat{a}_g = Z\hat{u}$$

Genomic relationship matrix

$$G = ZDZ'q$$

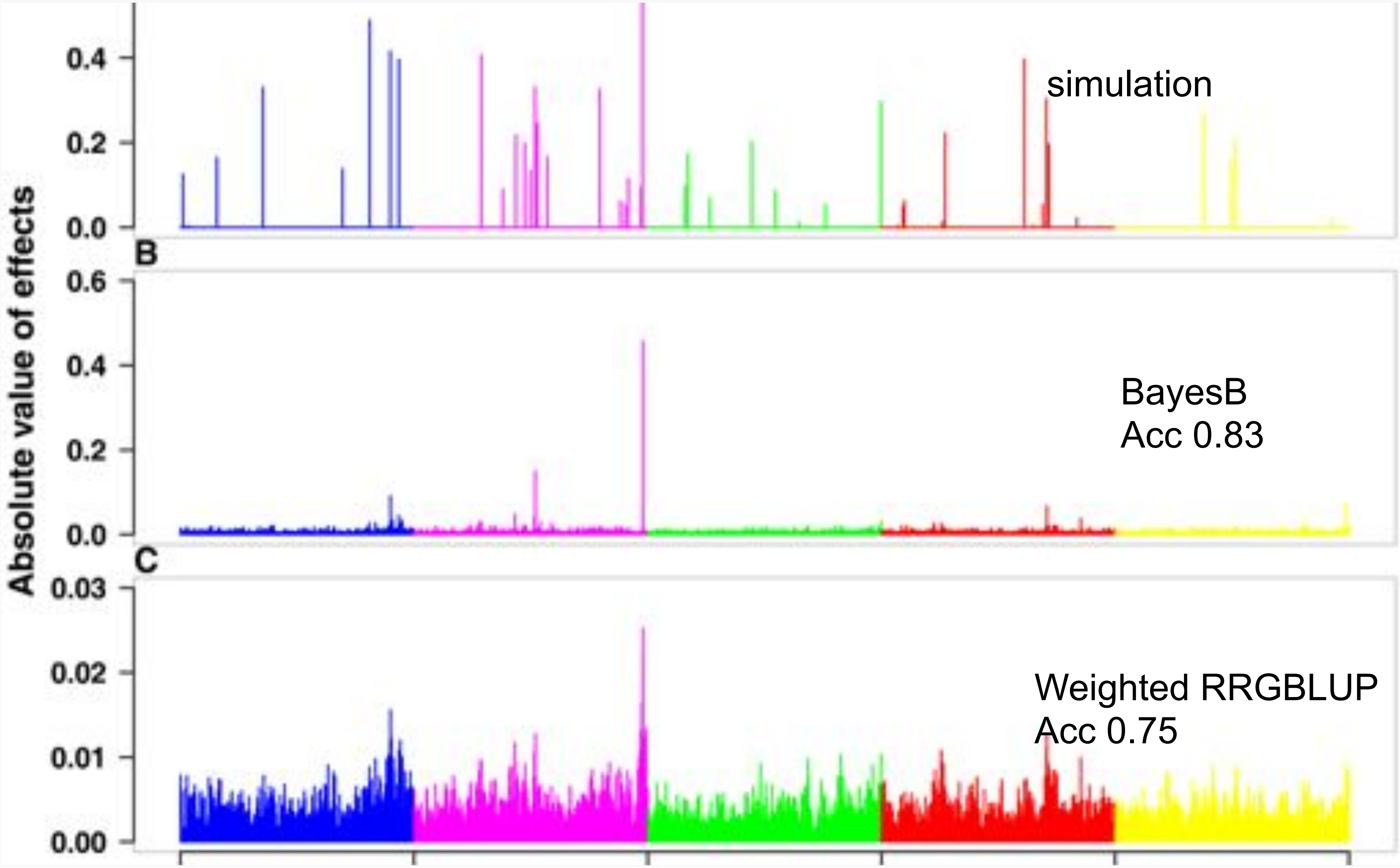
Estimate of SNP variance
(Zhang et al., 2010)

$$\hat{\sigma}^2_{u,i} = \hat{u}_i^2 2p_i(1-p_i)$$

Conversion from GEBV to
SNP effects (VanRaden, 2008;
Stranden and Garrick, 2010)

$$\hat{\mathbf{u}} = q\mathbf{DZ'G}^{*-1}\hat{\mathbf{a}}_g = \mathbf{DZ'[ZDZ']}^{-1}\hat{\mathbf{a}}_{\mathbf{g}}$$

Plots and accuracies in Zhang et al. (2010)

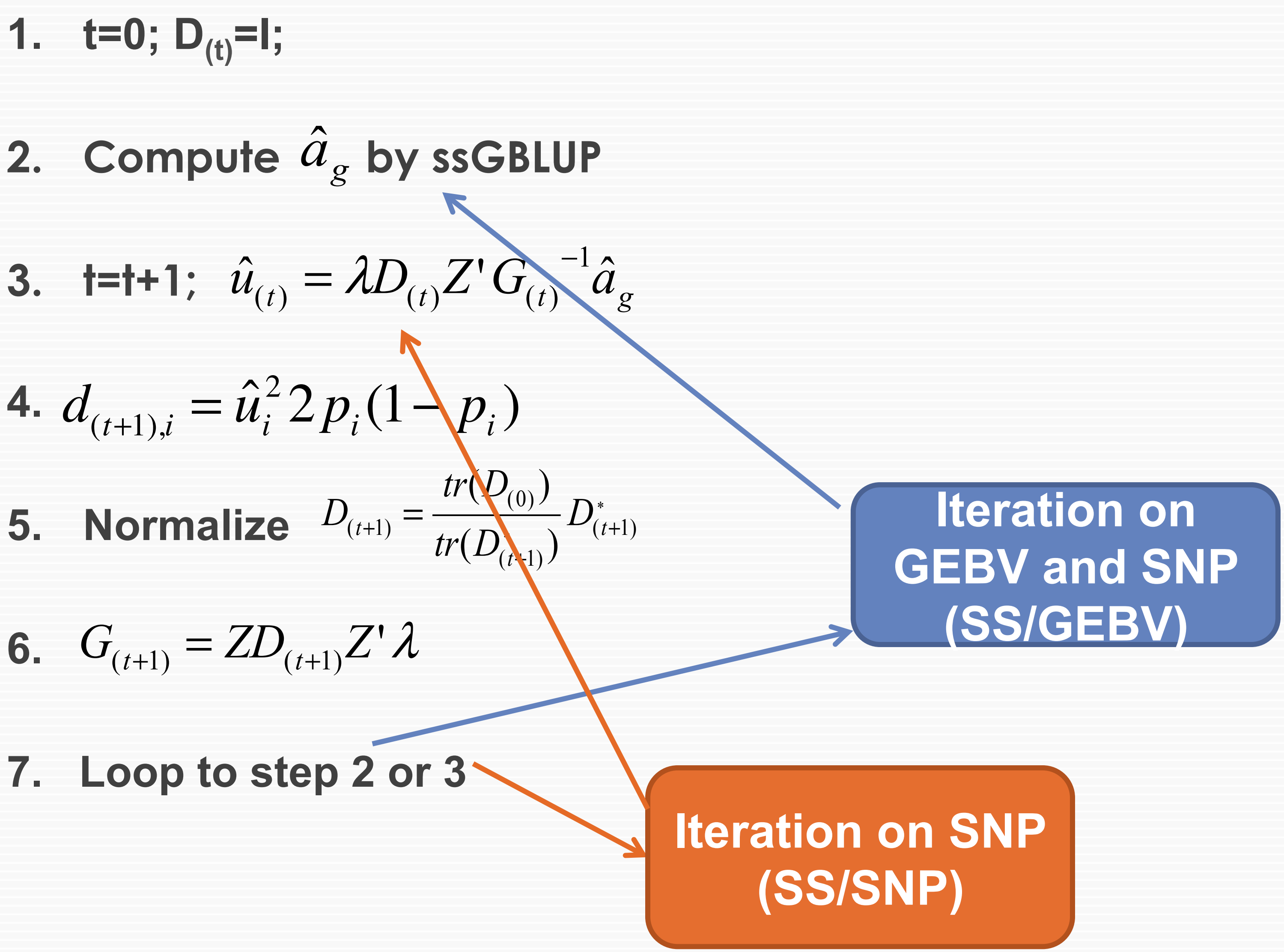


simulation

BayesB
Acc 0.83

Weighted RRGBLUP
Acc 0.75

GWAS under ssGBLUP



1. t=0; $D_{(t)}$=I;
2. Compute $\hat{a}_g$ by ssGBLUP
3. t=t+1; $\hat{u}_{(t)} = \lambda D_{(t)} Z' G_{(t)}^{-1} \hat{a}_g$
4. $d_{(t+1),i} = \hat{u}_i^2 2 p_i (1 - p_i)$
5. Normalize $D_{(t+1)} = \frac{tr(D_{(0)})}{tr(D^*_{(t+1)})} D^*_{(t+1)}$
6. $G_{(t+1)} = Z D_{(t+1)} Z' \lambda$
7. Loop to step 2 or 3

Iteration on GEBV and SNP (SS/GEBV)

Iteration on SNP (SS/SNP)

Genet. Res., Camb. (2012), **94**, *pp*. 73–83.
doi:10.1017/S0016672312000274

Genome-wide association mapping including phenotypes from relatives without genotypes

H. WANG[1]*, I. MISZTAL[1], I. AGUILAR[2], A. LEGARRA[3] AND W. M. MUIR[4]
[1] *Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602-2771, USA*
[2] *Instituto Nacional de Investigación Agropecuaria, INIA Las Brujas, 90200 Canelones, Uruguay*
[3] *INRA, UR631 Station d'Amélioration Génétique des Animaux (SAGA), BP 52627, 32326 Castanet-Tolosan, France*
[4] *Department of Animal Science, Purdue University, West Lafayette, IN 47907-1151, USA*

(*Received 19 September 2011; revised 8 December 2011, and 9 March 2012; accepted 13 March 2012*)

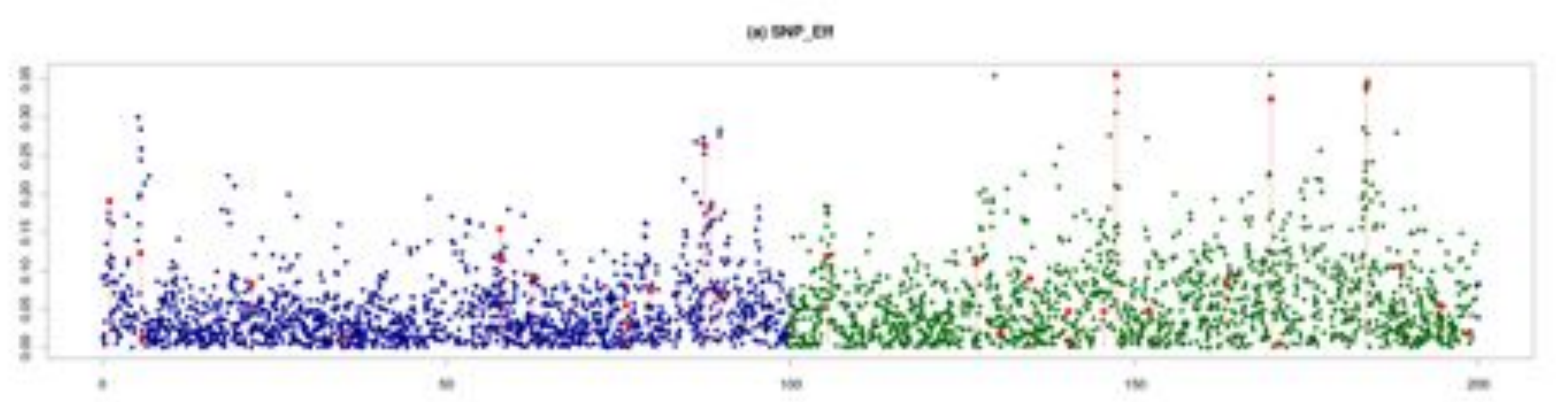
Comparisons by simulations
(Wang et al., 2012)

- **Data**
  - 15,800 individuals in 5 generations
  - 1500 genotyped
  - 3k SNP in 2 chromosomes

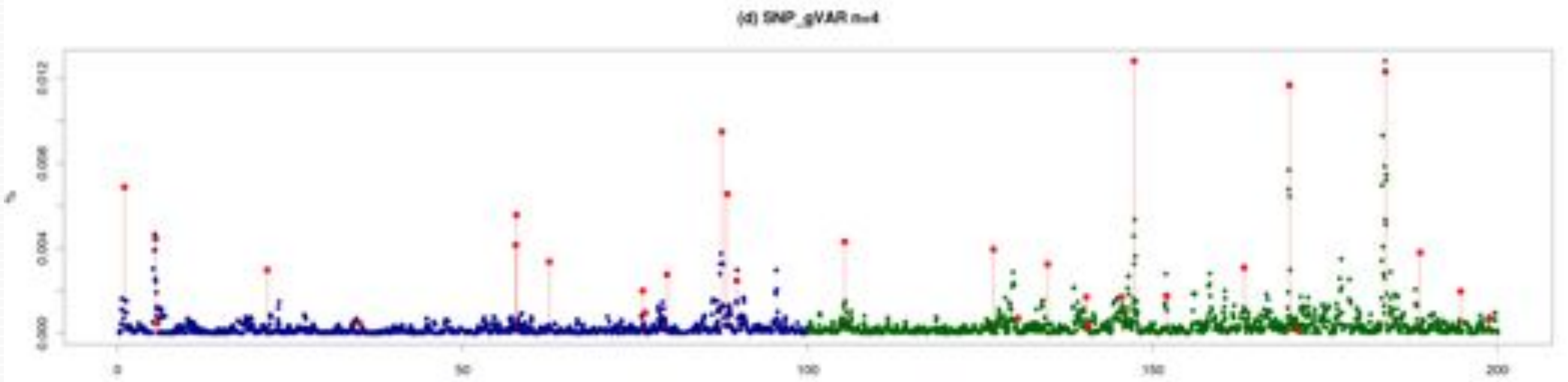
- **Methods**
  - Classical GWAS - Wombat (Meyer & Tier, 2012)
    - Degressed proofs
  - BayesB - GenSel (Habier *et al*., 2011)
    - Degressed proofs (c=0.1), 100k rounds
  - ssGBLUP - iterations on SNP and on GEBV

MANHANTTAN PLOTS

Wombat

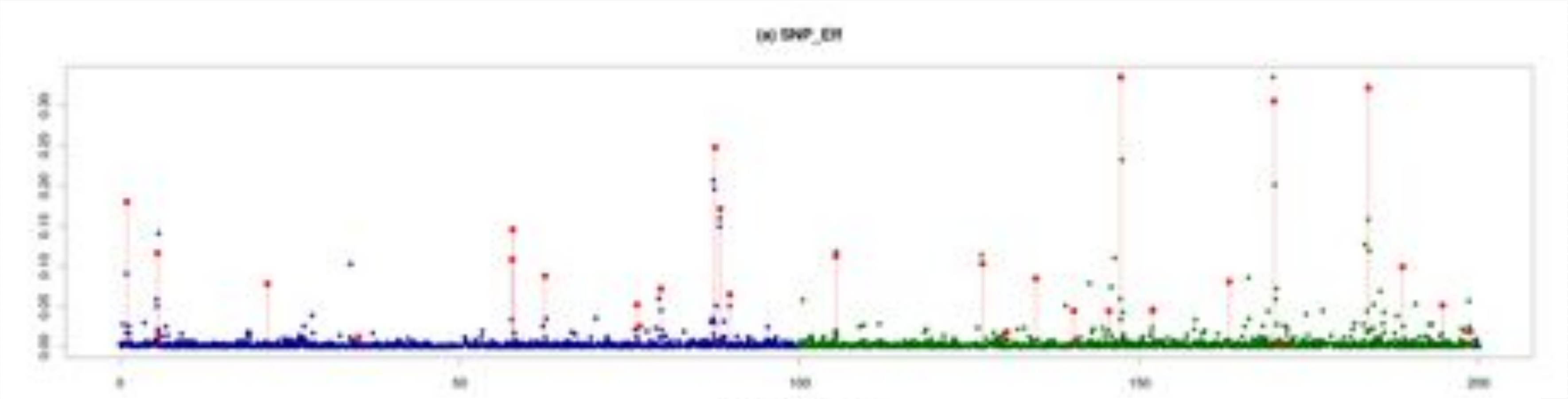


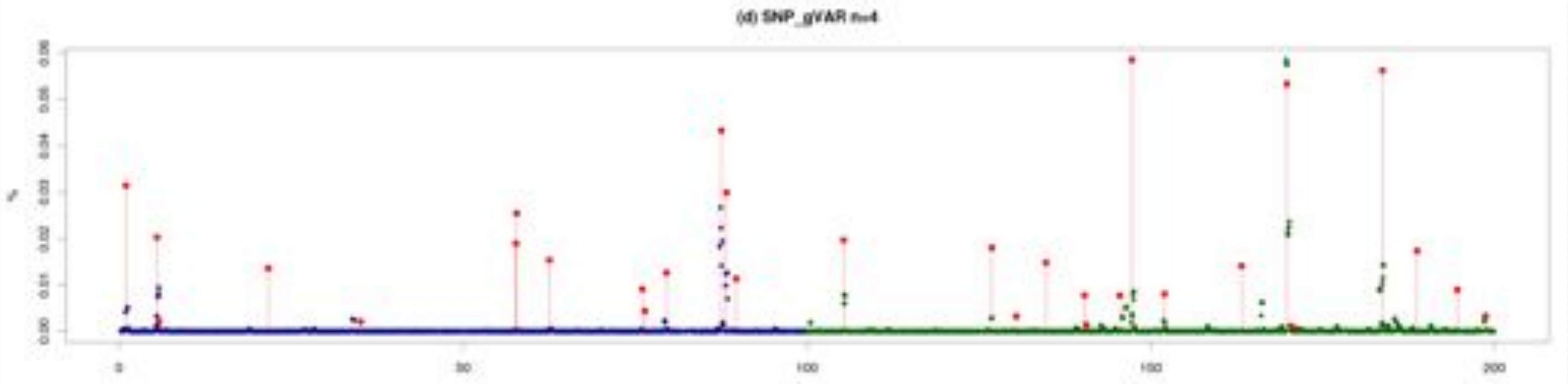
(a) SNP_ER



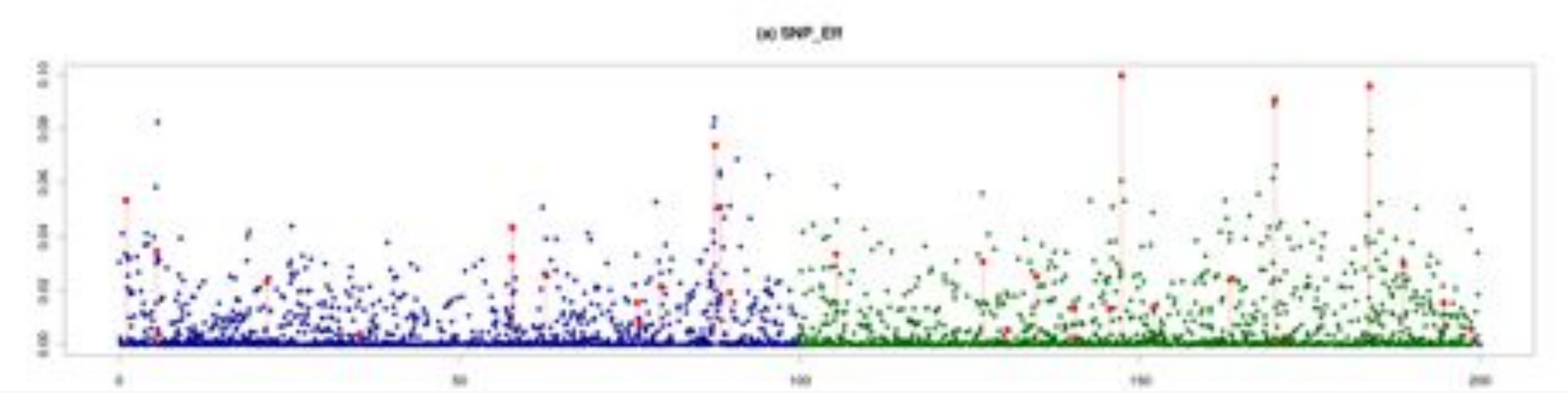
(d) SNP_gVAR n=4

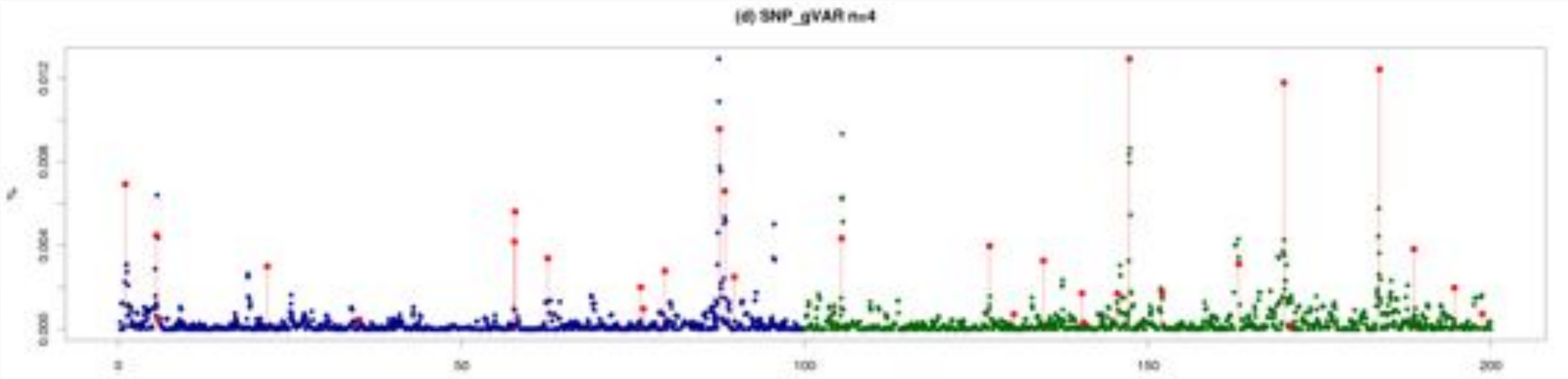
BayesB (weighted DP)



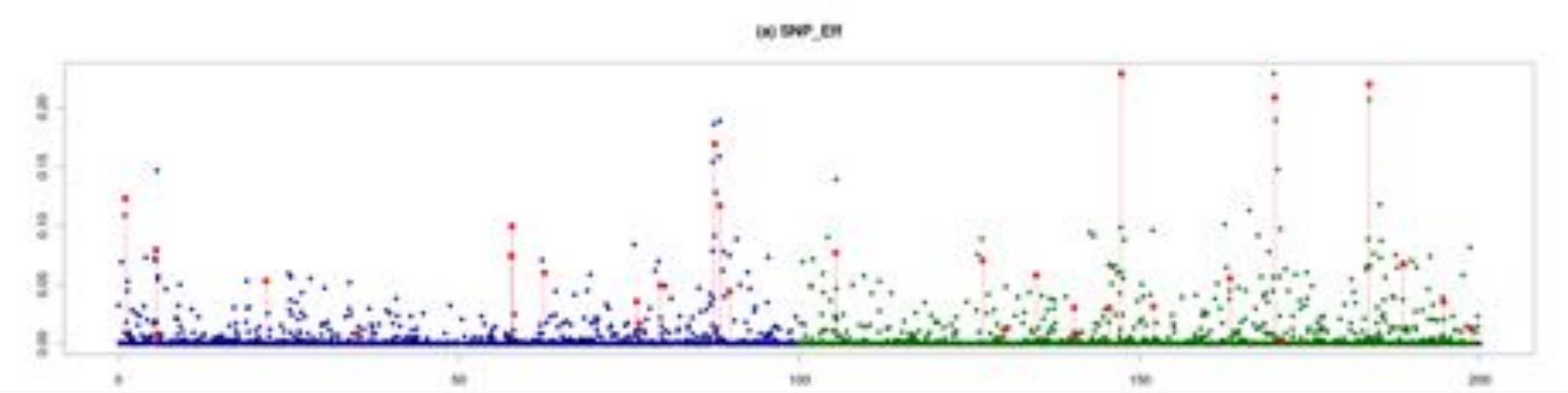


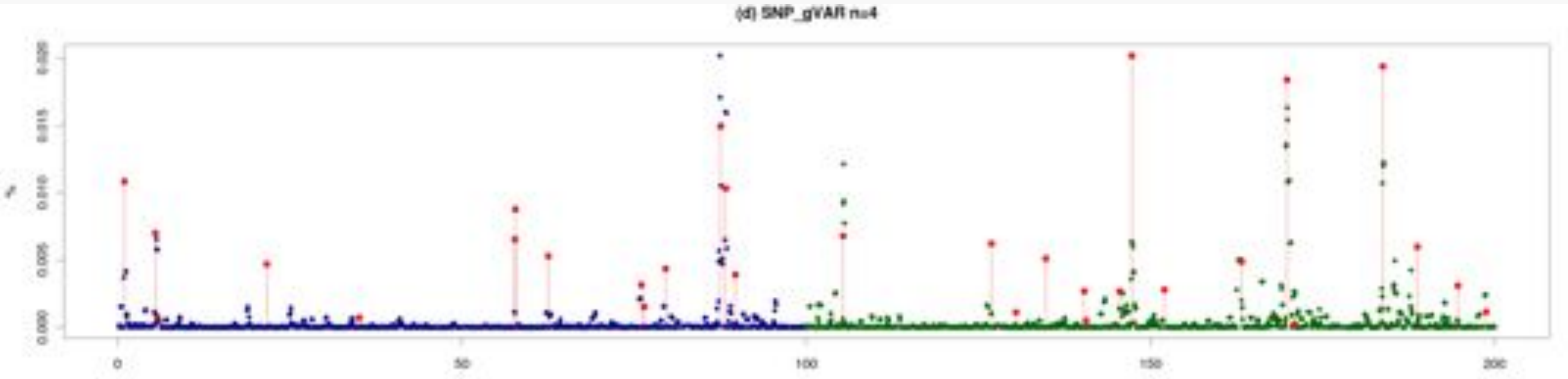
ssGBLUP / SNP (it3)


(a) SNP_ER


(d) SNP_gVAR n=4

ssGBLUP / GEBV (it3)





RESULTS (Simulated data : GEBVs)

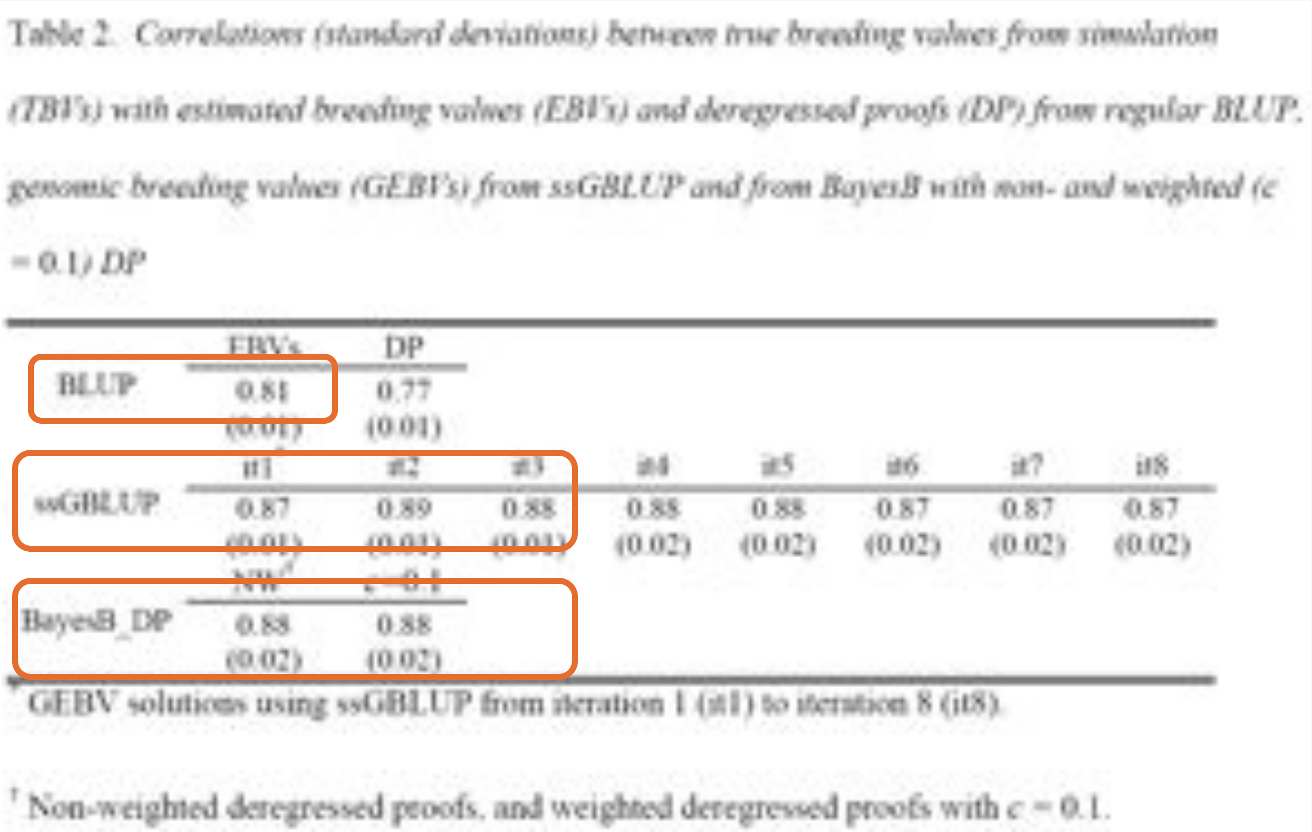


Table 2. *Correlations (standard deviations) between true breeding values from simulation (TBVs) with estimated breeding values (EBVs) and deregressed proofs (DP) from regular BLUP, genomic breeding values (GEBVs) from ssGBLUP and from BayesB with non- and weighted (c = 0.1) DP*

	EBVs	DP						
BLUP	0.81 (0.01)	0.77 (0.01)						
	it1[*]	it2	it3	it4	it5	it6	it7	it8
ssGBLUP	0.87 (0.01)	0.89 (0.01)	0.88 (0.01)	0.88 (0.02)	0.88 (0.02)	0.87 (0.02)	0.87 (0.02)	0.87 (0.02)
	NW[†]	c=0.1						
BayesB_DP	0.88 (0.02)	0.88 (0.02)						

[*] GEBV solutions using ssGBLUP from iteration 1 (it1) to iteration 8 (it8).

[†] Non-weighted deregressed proofs, and weighted deregressed proofs with c = 0.1.

Correlations between QTLs and clusters of SNP effects -ssGBLUP

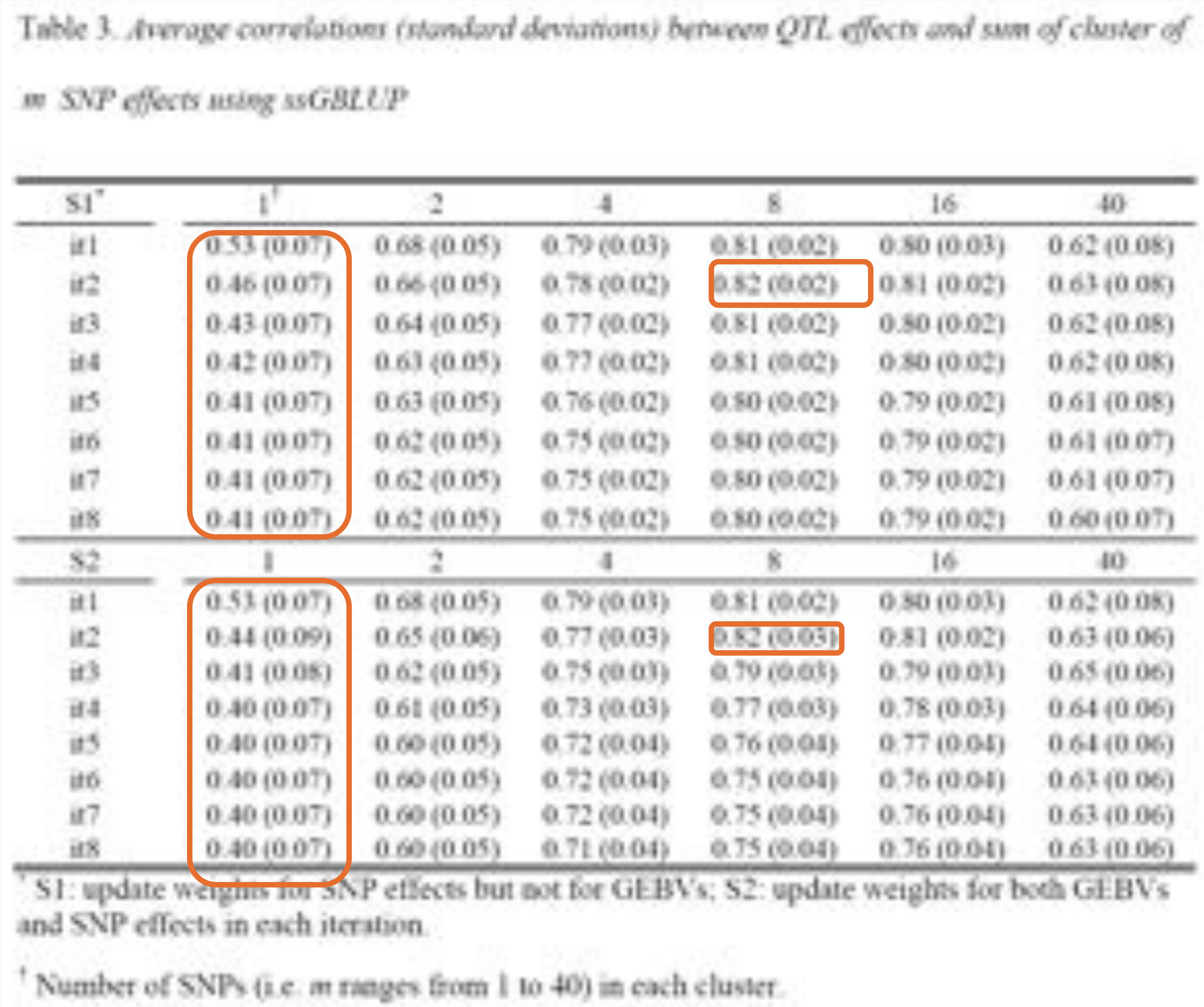


Table 3. *Average correlations (standard deviations) between QTL effects and sum of cluster of m SNP effects using ssGBLUP*

SS/SNP

S1[*]	1[†]	2	4	8	16	40
it1	0.53 (0.07)	0.68 (0.05)	0.79 (0.03)	0.81 (0.02)	0.80 (0.03)	0.62 (0.08)
it2	0.46 (0.07)	0.66 (0.05)	0.78 (0.02)	0.82 (0.02)	0.81 (0.02)	0.63 (0.08)
it3	0.43 (0.07)	0.64 (0.05)	0.77 (0.02)	0.81 (0.02)	0.80 (0.02)	0.62 (0.08)
it4	0.42 (0.07)	0.63 (0.05)	0.77 (0.02)	0.81 (0.02)	0.80 (0.02)	0.62 (0.08)
it5	0.41 (0.07)	0.63 (0.05)	0.76 (0.02)	0.80 (0.02)	0.79 (0.02)	0.61 (0.08)
it6	0.41 (0.07)	0.62 (0.05)	0.75 (0.02)	0.80 (0.02)	0.79 (0.02)	0.61 (0.07)
it7	0.41 (0.07)	0.62 (0.05)	0.75 (0.02)	0.80 (0.02)	0.79 (0.02)	0.61 (0.07)
it8	0.41 (0.07)	0.62 (0.05)	0.75 (0.02)	0.80 (0.02)	0.79 (0.02)	0.60 (0.07)

SS/GEBV

S2	1	2	4	8	16	40
it1	0.53 (0.07)	0.68 (0.05)	0.79 (0.03)	0.81 (0.02)	0.80 (0.03)	0.62 (0.08)
it2	0.44 (0.09)	0.65 (0.06)	0.77 (0.03)	0.82 (0.03)	0.81 (0.02)	0.63 (0.06)
it3	0.41 (0.08)	0.62 (0.05)	0.75 (0.03)	0.79 (0.03)	0.79 (0.03)	0.65 (0.06)
it4	0.40 (0.07)	0.61 (0.05)	0.73 (0.03)	0.77 (0.03)	0.78 (0.03)	0.64 (0.06)
it5	0.40 (0.07)	0.60 (0.05)	0.72 (0.04)	0.76 (0.04)	0.77 (0.04)	0.64 (0.06)
it6	0.40 (0.07)	0.60 (0.05)	0.72 (0.04)	0.75 (0.04)	0.76 (0.04)	0.63 (0.06)
it7	0.40 (0.07)	0.60 (0.05)	0.72 (0.04)	0.75 (0.04)	0.76 (0.04)	0.63 (0.06)
it8	0.40 (0.07)	0.60 (0.05)	0.71 (0.04)	0.75 (0.04)	0.76 (0.04)	0.63 (0.06)

[*] S1: update weights for SNP effects but not for GEBVs; S2: update weights for both GEBVs and SNP effects in each iteration.

[†] Number of SNPs (i.e. m ranges from 1 to 40) in each cluster.

Correlations between QTLs and clusters of SNP effects –BayesB & Wombat

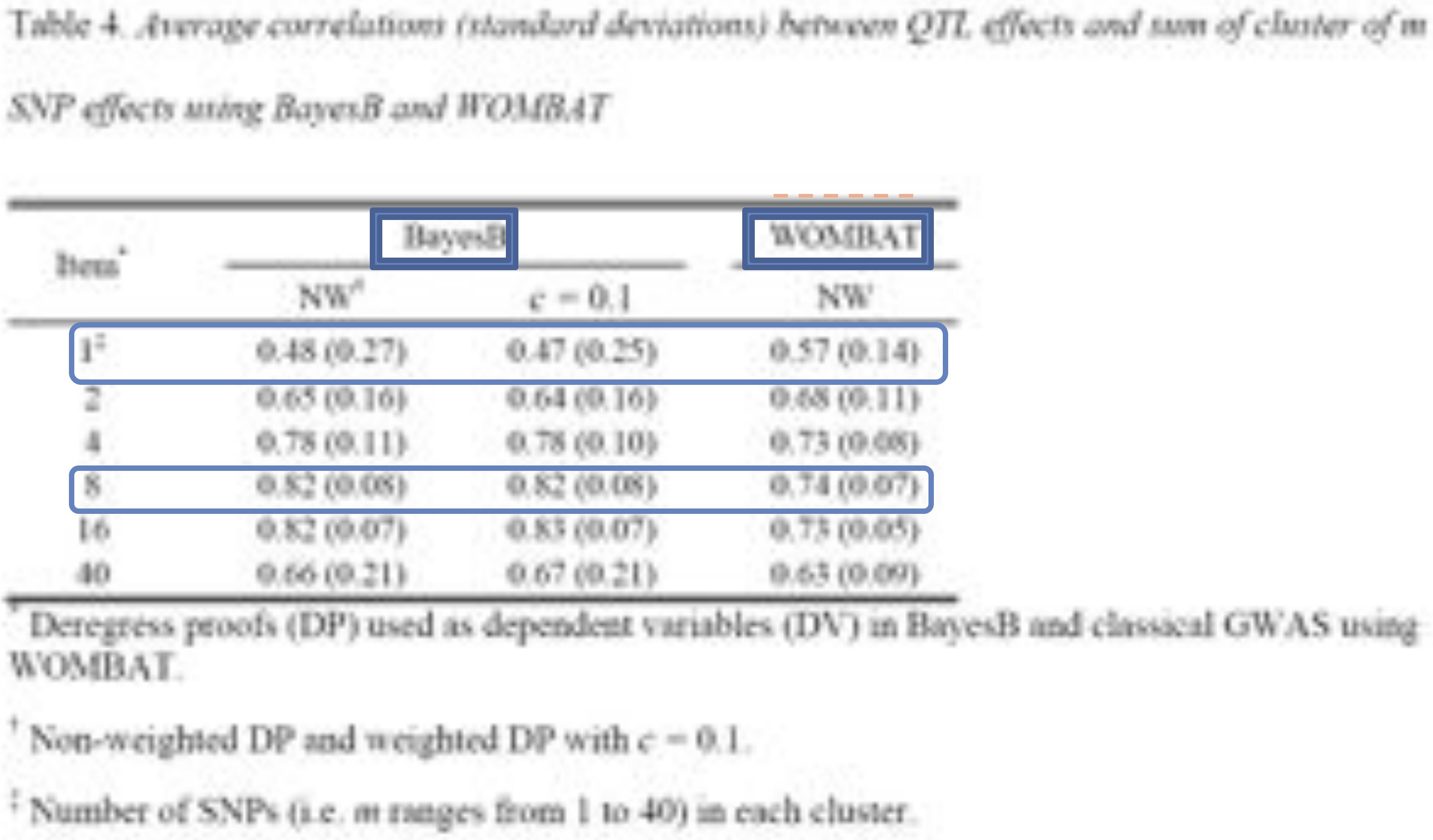


Table 4. *Average correlations (standard deviations) between QTL effects and sum of cluster of m SNP effects using BayesB and WOMBAT*

Item[*]	BayesB		WOMBAT
	NW[†]	c = 0.1	NW
1[‡]	0.48 (0.27)	0.47 (0.25)	0.57 (0.14)
2	0.65 (0.16)	0.64 (0.16)	0.68 (0.11)
4	0.78 (0.11)	0.78 (0.10)	0.73 (0.08)
8	0.82 (0.08)	0.82 (0.08)	0.74 (0.07)
16	0.82 (0.07)	0.83 (0.07)	0.73 (0.05)
40	0.66 (0.21)	0.67 (0.21)	0.63 (0.09)

[*] Deregress proofs (DP) used as dependent variables (DV) in BayesB and classical GWAS using WOMBAT.

[†] Non-weighted DP and weighted DP with c = 0.1.

[‡] Number of SNPs (i.e. m ranges from 1 to 40) in each cluster.

Field data set

Data

Body weight in broiler chicken at 6 weeks

N=275k ; N_g=4500, 40K SNP (after edits)

6 generations

Model for ssGBLUP:

- : fixed effects (sex, contemporary group)
- : maternal environment effects
- : animal effects

Models of Classical GWAS and BayesB:

Classical GWAS: $y = Xb* + Wp + Za + e$

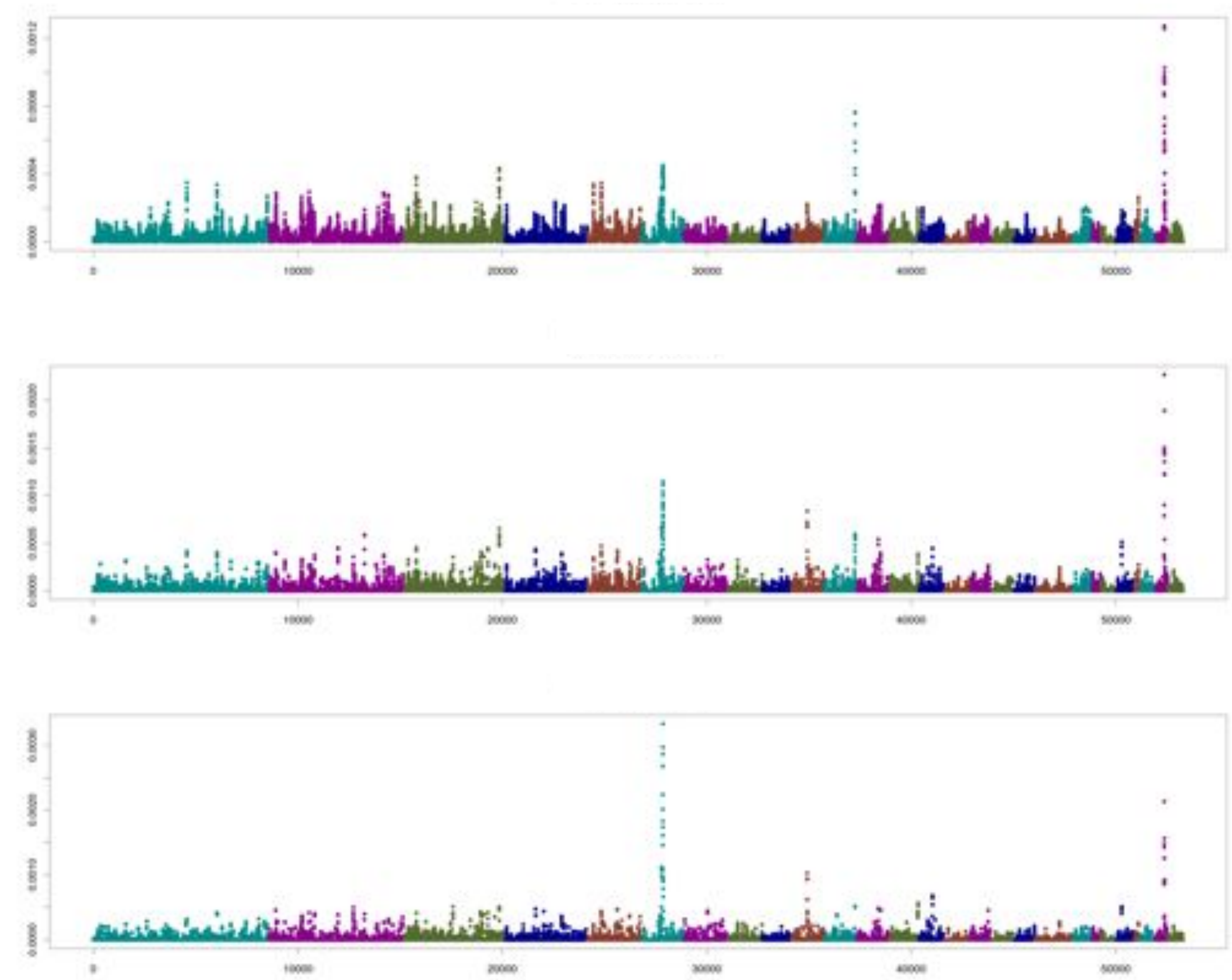
- Wombat (Meyer & Tier, 2012)
- ***y :*** phenotypic records
- $b*$ **:** sex, CG, and single snp marker

BayesB: $y = 1\mu + Zg + e$

- GenSel (Habier *et al*., 2011)
- g***:*** a vector of SNP markers
- ***y*** **:**
  - NDP: non-weighted degressed proofs
  - WDP: weighted degressed proofs (c=0.1)
- t = 51, 000 (first 1,000 as burn-in)
- $\pi = 0.9$

ssGBLUP – iterations on SNP only

Sliding window n=10



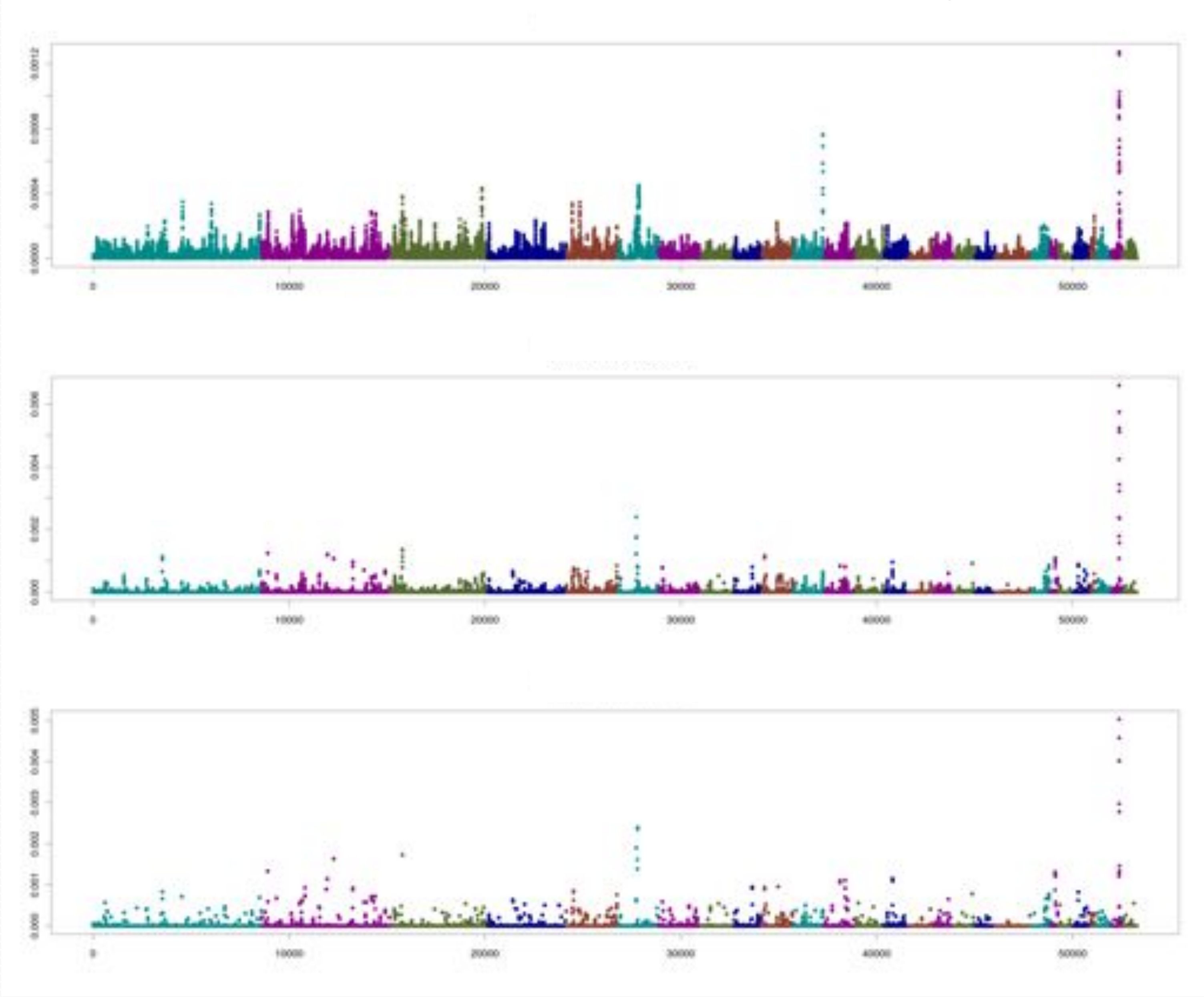
it1

it3

it5

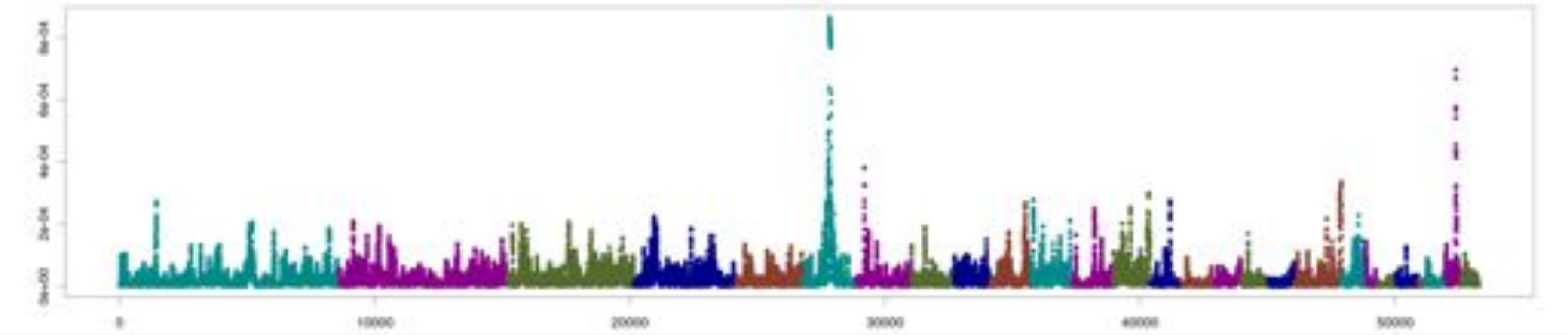
ssGBLUP – iterations on SNP and GEBV

Sliding window n=10

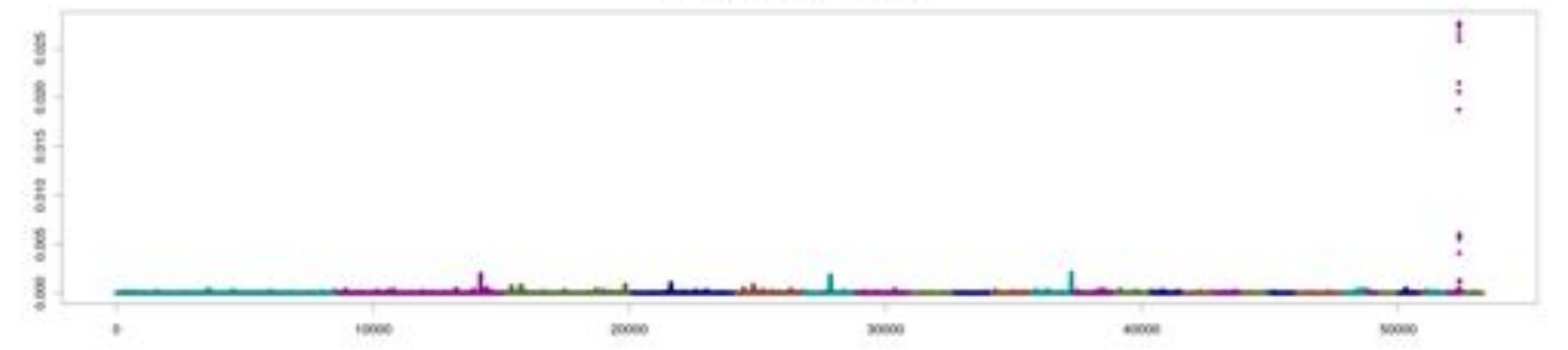


Sliding window n=10

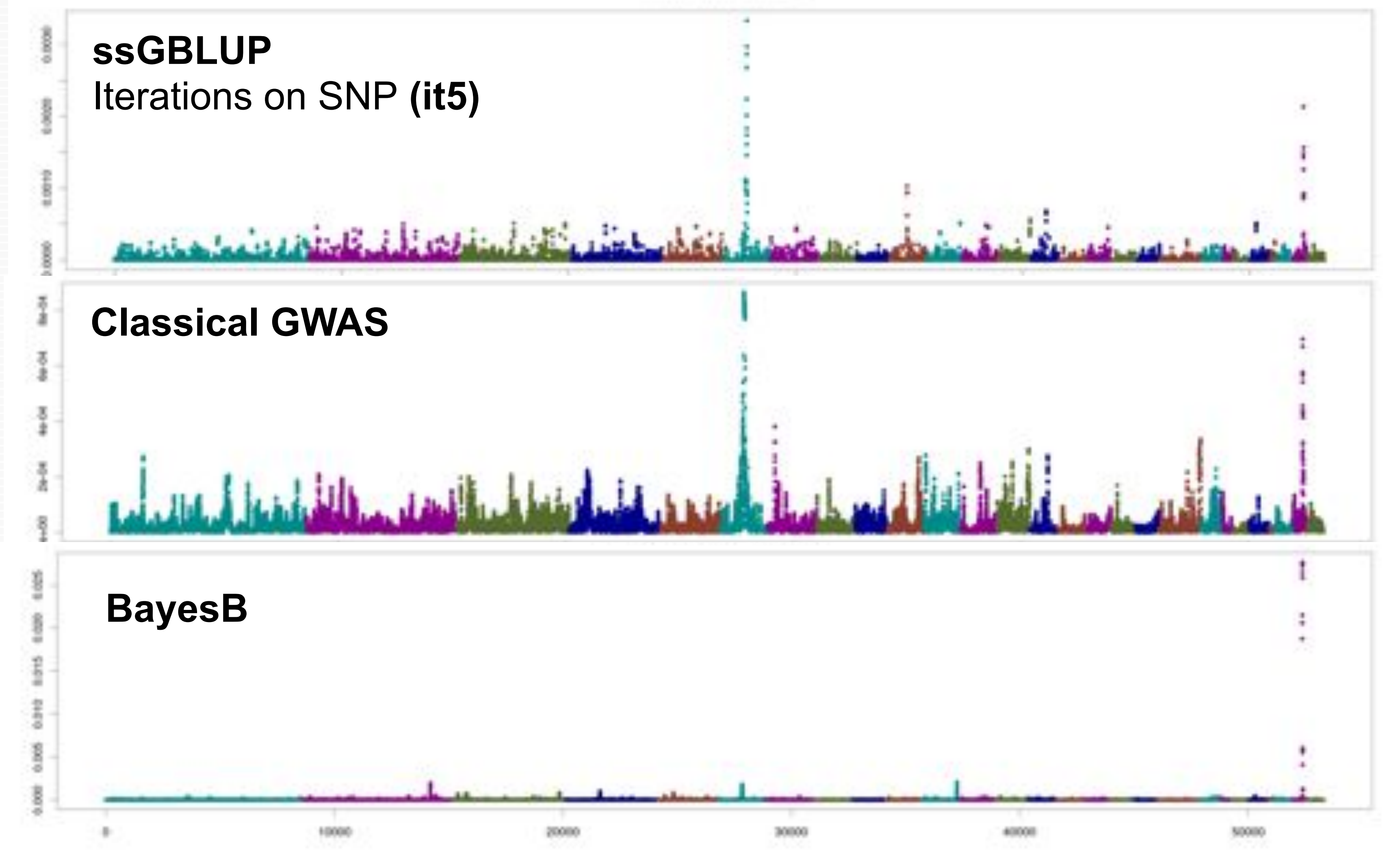
Classical GWAS



BayesB



Comparison of Three Methods:



Ranking of SNP regions in ssGBLUP during iteration

	SNP							SNP+GEBV						
it1	**it2**	**it3**	**it4**	**it5**	**it6**	**it7**	**it8**	**it2**	**it3**	**it4**	**it5**	**it6**	**it7**	**it8**
1	1	1	1	1	1	2	2	1	1	1	1	1	1	2
2	3	3	3	2	2	1	1	9	351	351	479	489	492	493
3	2	2	2	4	5	10	14	6	256	256	472	570	610	617
4	12	21	32	36	46	57	65	2	72	72	100	106	98	98
5	4	4	4	3	3	3	3	16	3	3	2	2	2	1
6	9	11	14	14	17	13	12	20	575	575	766	840	857	863

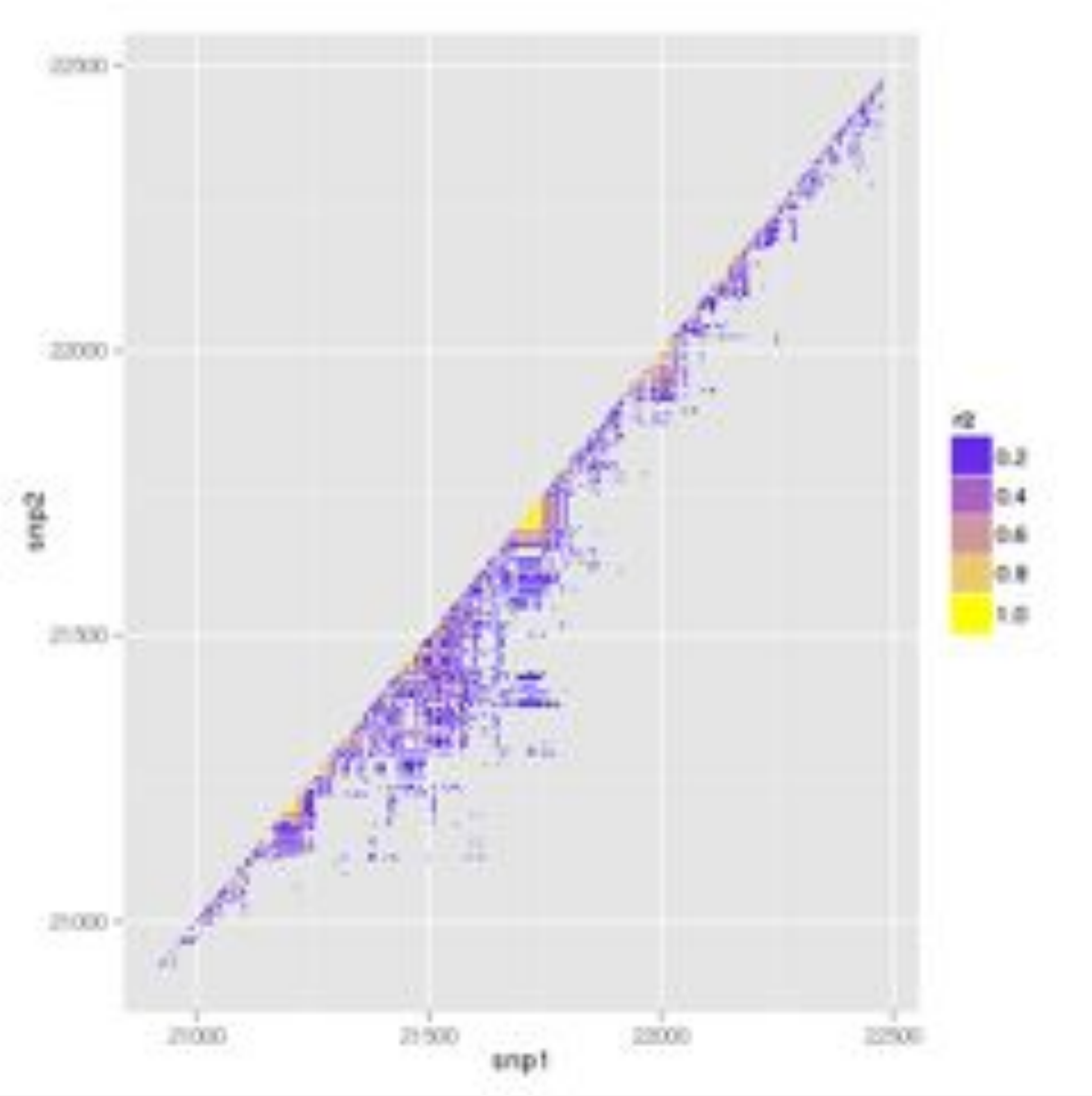
Regions of 20 SNP

SS/SNP(3)	chr	Var	SS/EBV(3)	wombat	BayesB
1	27	2.5%	1	6	1
2	6	1.3%	62	1	2
3	6	0.9%	110	2	3
4	6	0.8%	8	3	40
5	10	0.7%	54	59	93
6	5	0.6%	16	423	8
7	2	0.6%	57	32	9
8	1	0.5%	21	76	23
9	4	0.5%	105	450	7
10	12	0.5%	13	357	31

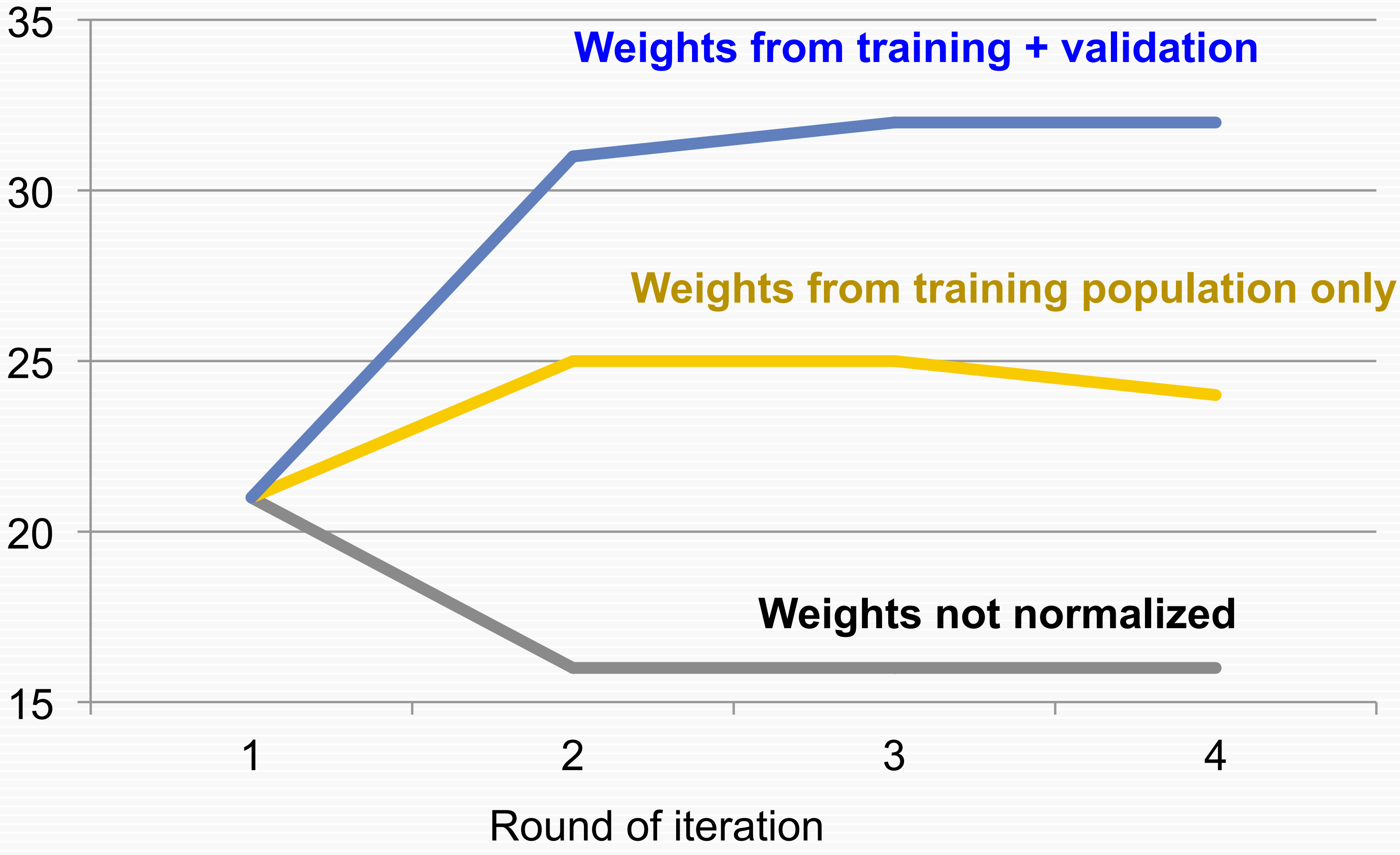
BayesB	chr	Var	SS/SNP(3)	SS/EBV(3)	wombat
1	27	23.1%	1	1	6
2	6	2.3%	2	62	1
3	6	1.9%	3	110	2
4	11	1.4%	15	31	279
5	2	1.0%	42	63	656
6	3	1.0%	144	166	11
7	4	0.7%	9	105	450
8	5	0.7%	6	16	423
9	2	0.6%	7	57	32
10	2	0.5%	264	119	53

wombat	chr	Var	SS/SNP(3)	SS/EBV(3)	BayesB
1	6	3.1%	2	62	2
2	6	2.9%	3	110	3
3	6	1.3%	4	8	40
4	6	1.0%	360	810	322
5	6	0.8%	278	565	27
6	27	0.8%	1	1	1
7	6	0.6%	668	1216	1646
8	7	0.5%	314	927	99
9	12	0.5%	855	925	387
10	4	0.4%	274	903	173

HapMap of Chromosome 6

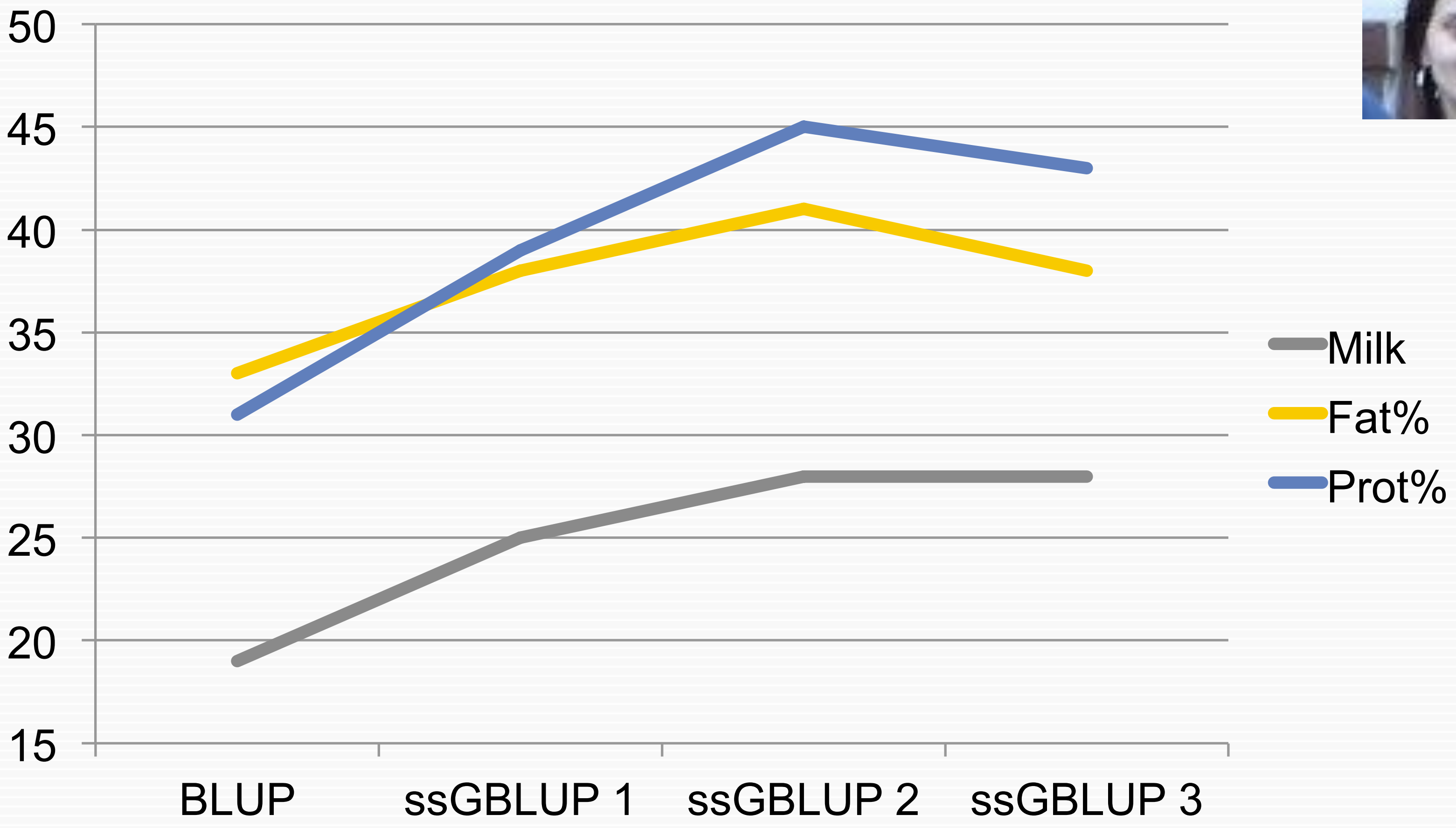


Realized accuracies of ssGBLUP/GEBV during iteration



R^2 in dairy – 1400 genotypes

(Lino et al., 2012)



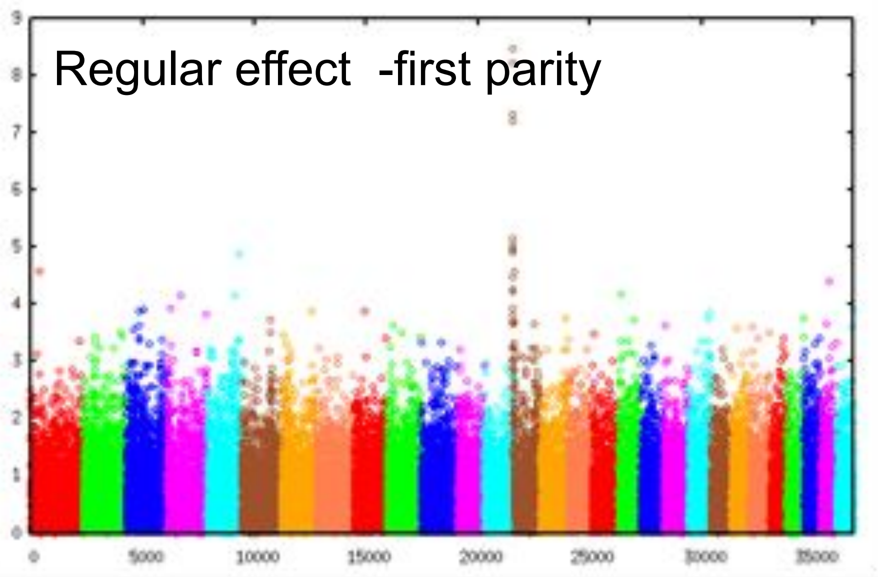
ssGBLUP/SNP for Heat Stress in Holsteins (Aguilar, 2011)

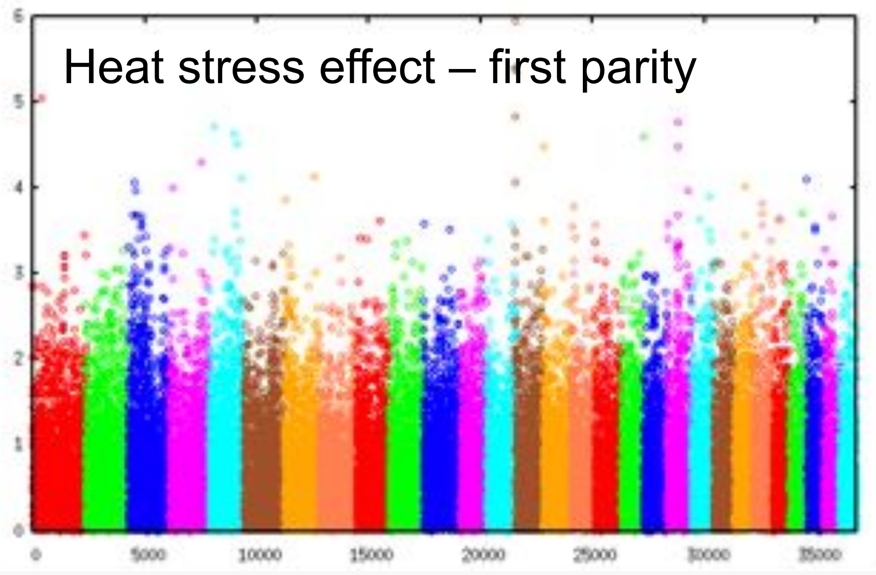
Multiple-Trait Test-Day model, heat stress as random regression

- ~ 90 millions records, ~ 9 millions pedigrees
- ~ 3,800 genotyped bulls

Computing time

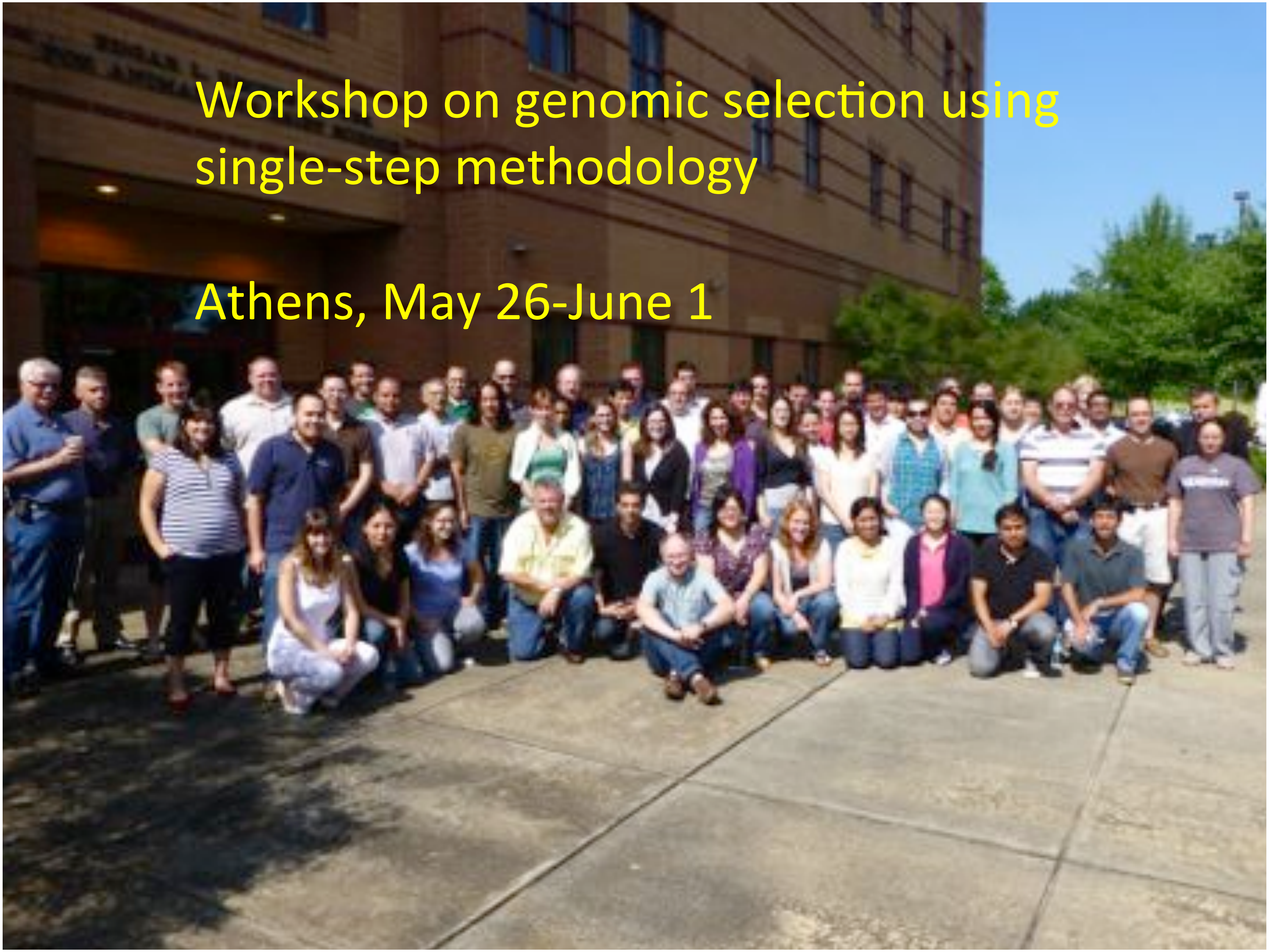
- Complete evaluation ~ 16 h

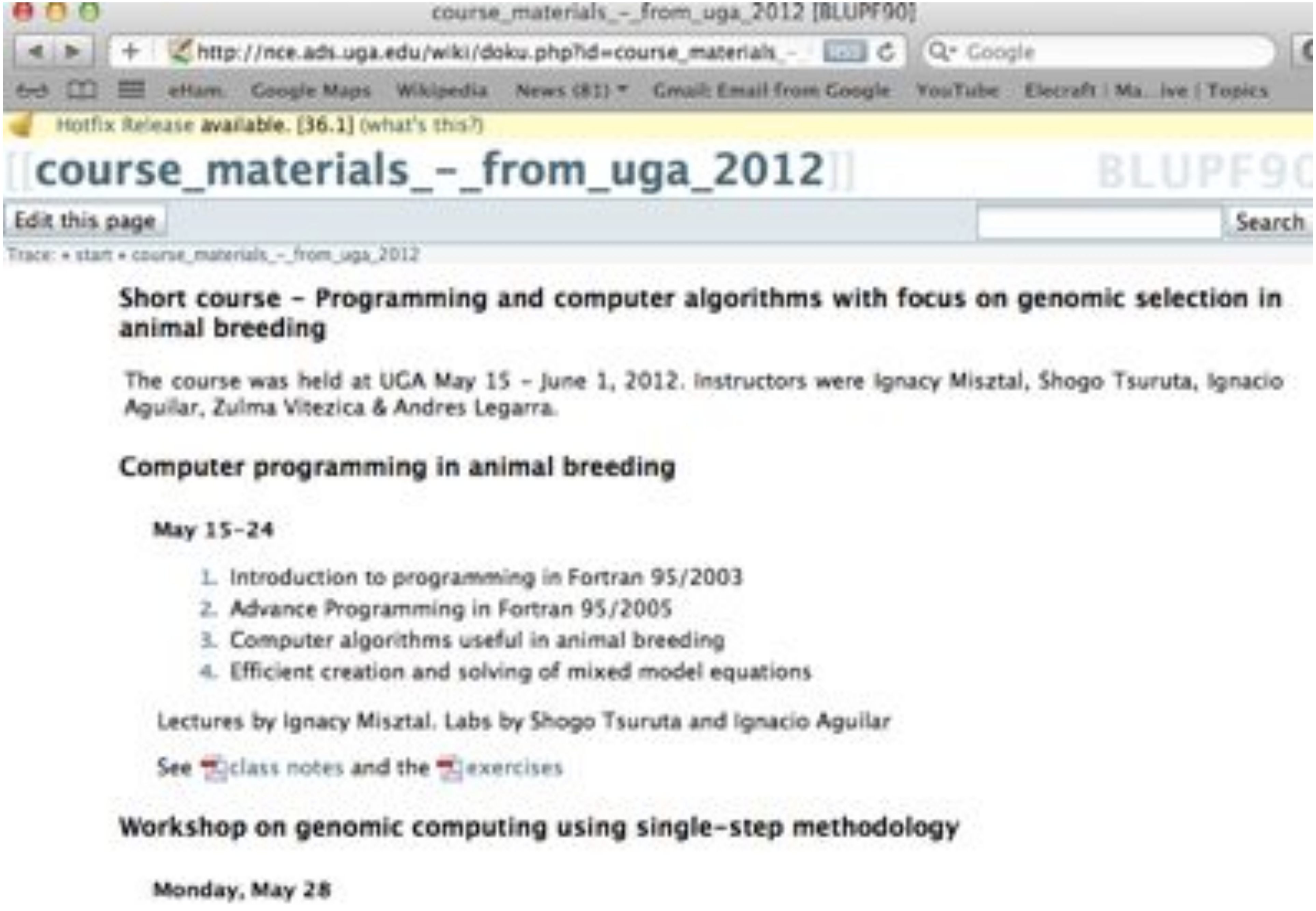

Regular effect -first parity


Heat stress effect – first parity

Workshop on genomic selection using single-step methodology

Athens, May 26-June 1





[[course_materials_-_from_uga_2012]]

BLUPF90

Short course - Programming and computer algorithms with focus on genomic selection in animal breeding

The course was held at UGA May 15 - June 1, 2012. Instructors were Ignacy Misztal, Shogo Tsuruta, Ignacio Aguilar, Zulma Vitezica & Andres Legarra.

Computer programming in animal breeding

May 15-24

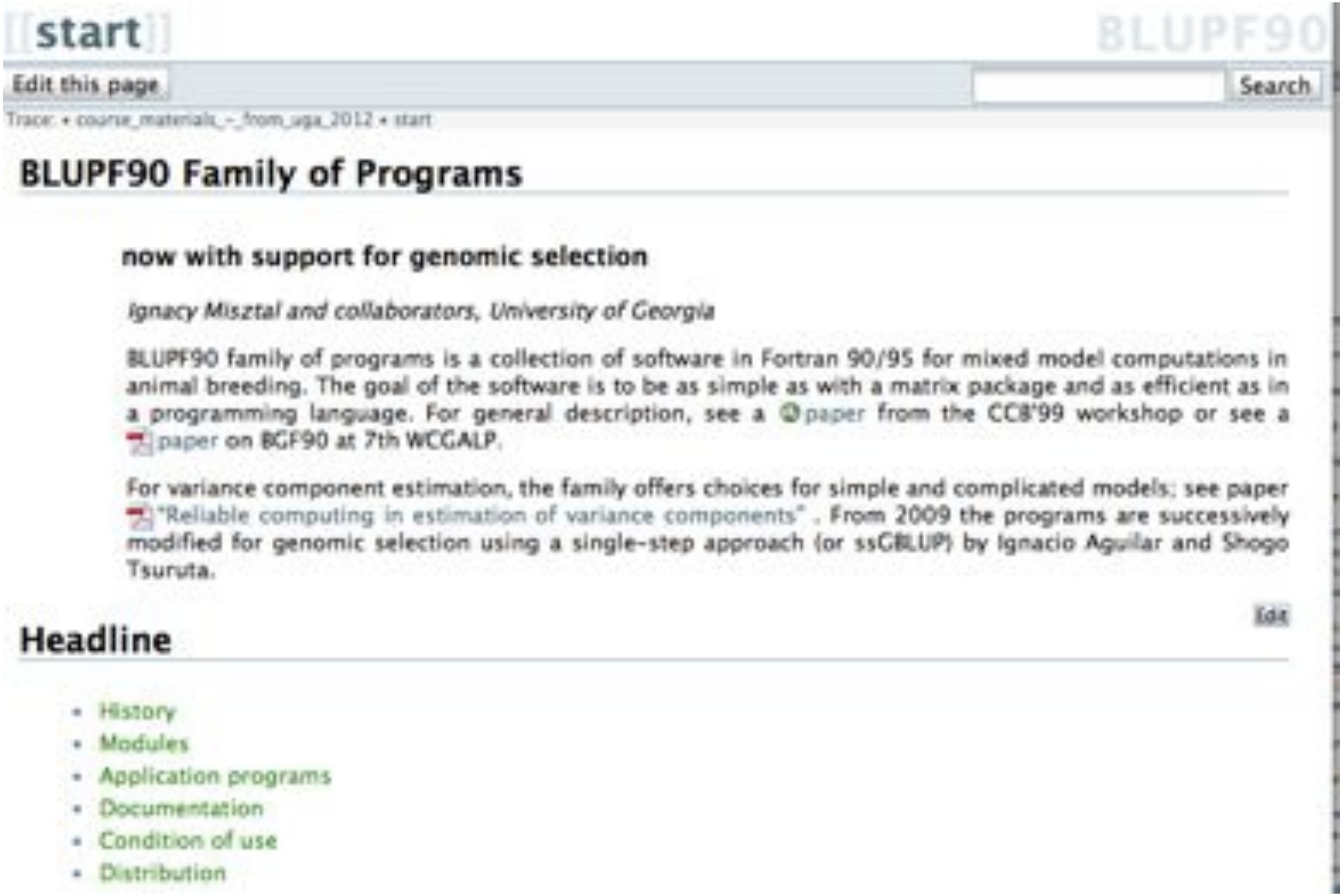
1. Introduction to programming in Fortran 95/2003
2. Advance Programming in Fortran 95/2005
3. Computer algorithms useful in animal breeding
4. Efficient creation and solving of mixed model equations

Lectures by Ignacy Misztal. Labs by Shogo Tsuruta and Ignacio Aguilar

See class notes and the exercises

Workshop on genomic computing using single-step methodology

Monday, May 28



BLUPF90 Family of Programs

now with support for genomic selection

Ignacy Misztal and collaborators, University of Georgia

BLUPF90 family of programs is a collection of software in Fortran 90/95 for mixed model computations in animal breeding. The goal of the software is to be as simple as with a matrix package and as efficient as in a programming language. For general description, see a paper from the CC8'99 workshop or see a paper on BGF90 at 7th WCGALP.

For variance component estimation, the family offers choices for simple and complicated models; see paper "Reliable computing in estimation of variance components" . From 2009 the programs are successively modified for genomic selection using a single-step approach (or ssGBLUP) by Ignacio Aguilar and Shogo Tsuruta.

Headline

- History
- Modules
- Application programs
- Documentation
- Condition of use
- Distribution

Renumbering
RENUMF90

Computing of extra matrices
PreGSF90

BLUP in memory
BLUPF90

BLUP – iteration on data
BLUP90IODF
CBLUP90IOD

Variance component estimation
REMLF90 AIREMLF90
GIBBS2F90 THRGIBBS2F90

Sample analysis
POSTGIBBSF90

Approximate accuracies
ACCF90

GEBV to SNP conversions
GWAS
PostGSF90

Predictions via SNP
PredGSF90

Issues

- Alternative sampling of SNP variances (Sun et al., 2011)
- Significance testing
- Multiple trait models with large QTL/regions for some traits
- Maximum number of genotypes

CONCLUSIONS

- **ssGBLUP for GWAS:**
  - Simple and Fast
  - Applicable to any model

- **ssGBLUP/SNP**
  - Optimal if no large SNP effects
  - Applicable to multiple traits

- **ssGBLUP/GEBV**
  - 1-2 rounds enough
  - useful for more accurate GEBV if major SNP

- **Large potential for research**

Acknowledgements

Cobb Vantress

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

Daniela Lino for testing

Dorian Garrick (GenSel) and Karin Meyer (Wombat)