Selective Genotyping and Cross-Validation Strategies in Genomic Selection

Guilherme J. M. Rosa

Department of Animal Sciences

Department of Biostatistics & Medical Informatics

Development of Efficient Design and Statistical Analysis Strategies for Genome-wide Association Studies in Livestock

Hatch Project No: WIS01433, Accession No: 0218979
Period: 10/01/2009-09/30/2013, Report: Year 2 of 4

⇨ Specific Questions

- Assess alternative cross-validation designs
- Effect of selective genotyping in genomic selection

Accuracy of genome enabled prediction in dairy cattle and wheat populations using different cross-validation designs

M. Angeles Pérez-Cabal[1], Ana I. Vazquez[2], Daniel Gianola[3], Guilherme J. M. Rosa[3], and Kent A. Weigel[3]

[1] University of Madrid, and Polytechnic University of Madrid, Spain
[2] University of Alabama-Birmingham, AL, USA
[3] University of Wisconsin-Madison, WI, USA

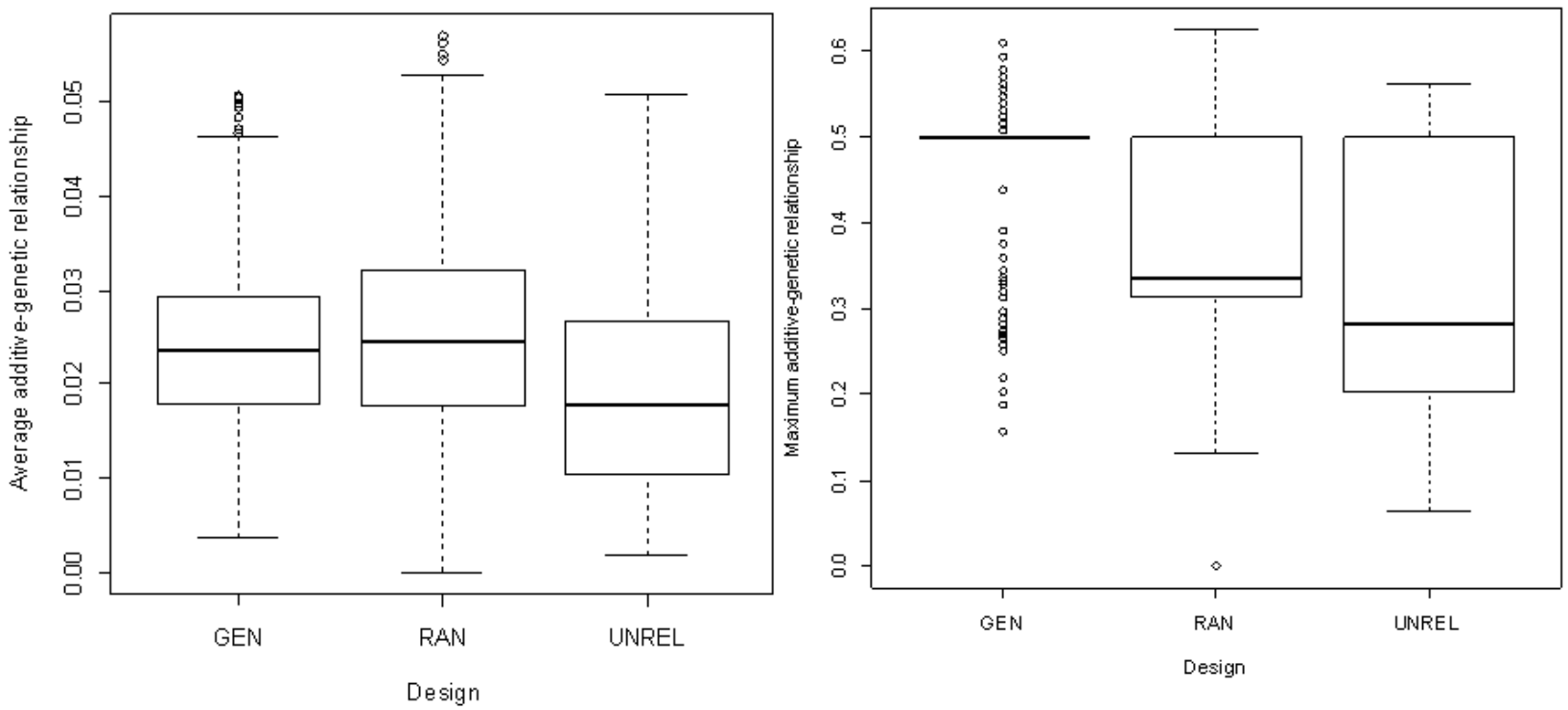
Introduction

- Genome-enabled prediction: Estimation of BV (breeding), or producing ability (management), or yet-to-be phenotypes (personalized medicine)
- Model fit: Bias-variance tradeoff (Cross-validation)
- Cross-validation: Training-testing partition (Legarra et al. 2008; Harbier 2007, 2010; Luan et al. 2009)
- Objective: To assess the importance of genetic relatedness on accuracy of genome-enabled predictions using two distinct populations, and to compare alternative cross-validation strategies

Material and Methods

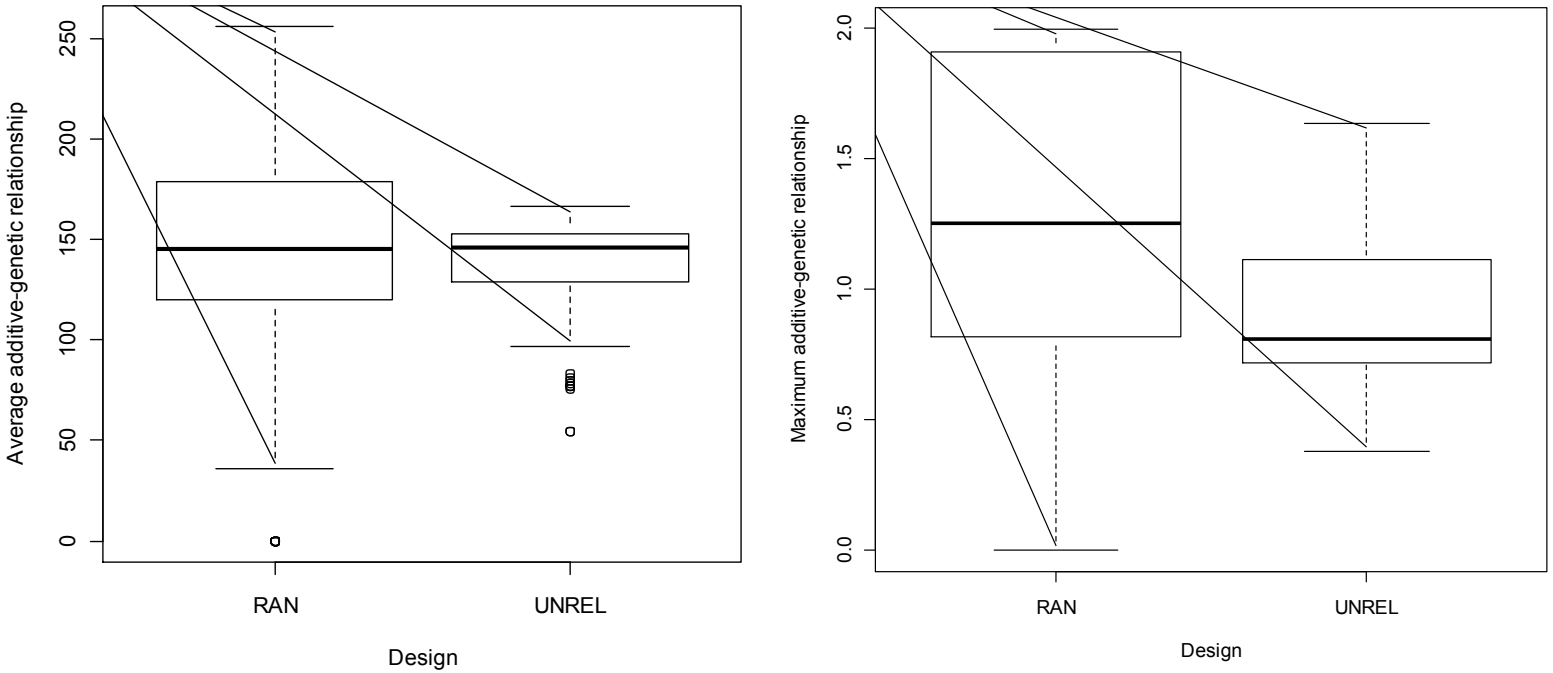
- Two populations:
  - Wheat: grain yield; 599 lines (CIMMYT) with 1,279 SNPs after editing
  - Dairy cattle: sire PTA for protein yield and somatic cell score; 4,703 sires with 32,518 SNPs after editing
- Three CV strategies:
  - RAN: random split for training and test sets
  - GEN: split by generation: older individuals in training
  - REL: two sets of less related animals
- Methods: Bayesian LASSO; prediction accuracy measured by correlation between genomic predictions and PTA (dairy cattle) or phenotype (wheat) in the testing sets

Results



Box plots of average and maximum additive-genetic relationships between training and testing set animals for the GEN, RAN, and UNREL designs with the cattle data.

Results



Box plots of average and maximum additive-genetic relationships between training and testing set lines for the RAN, and UNREL designs with the wheat data.

Results

Prediction accuracy for protein yield (dairy cattle) and grain yield (wheat) for the GEN, RAN, and UNREL training-testing designs.

Population	GEN	RAN	REL
Dairy cattle (Protein yield)	0.71	0.82	0.81
Wheat (Grain yield)	--	0.46	0.38

Concluding Remarks

- Different CV strategies resulted in somewhat different predictive abilities
- Overall, slightly higher accuracy levels with higher genetic relationships between training and testing sets, especially for low heritability traits.
- General advice: CV should mimic the manner in which genomic predictions will be used
- Alternative, multiple CV layouts

Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection

A. A. Boligon[1], N. Long[2], L. G. Albuquerque[1], K. A. Weigel[3], D. Gianola[3] and G. J. M. Rosa[3]

[1] Sao Paulo State University (UNESP), SP, Brazil
[2] Duke University, Durham, NC, USA
[3] University of Wisconsin-Madison, WI, USA

Introduction

- Genomic selection: Genome-enabled prediction of breeding values
- Genotyping cost: Genotype subset of animals
- Effect of prediction accuracy?
- Objective: To evaluate the quality of GEBV for candidates to selection based on different strategies of selective genotyping of a population undergoing selection, with different selection intensities

Material and Methods

Population/Generations

5,000 generations
(mutation-drift equilibrium)

t_1: 100 animals
(50 males + 50 females)

⇩

⋮ Random mating

⇩

t_{5000}: 100 animals

⇩ Factorial mating

$G_{0:}$ 2,500 animals

⇩ Directional selection

G_1: 2,500 animals

Selection intensities	
%	#
2	50
6	150
10	250
14	350
20	500
26	650
34	850
100	2,500

Material and Methods

Markers and Genetics Effects

- Genome: 10 chromosomes with 100 cM each
- Loci: 302 bialellic loci (202 markers + 100 QTL) in each chromosome

$$M_1 - M_2 - Q_1 - M_3 - M_4 - \ldots - M_{199} - M_{200} - Q_{100} - M_{201} - M_{202}$$

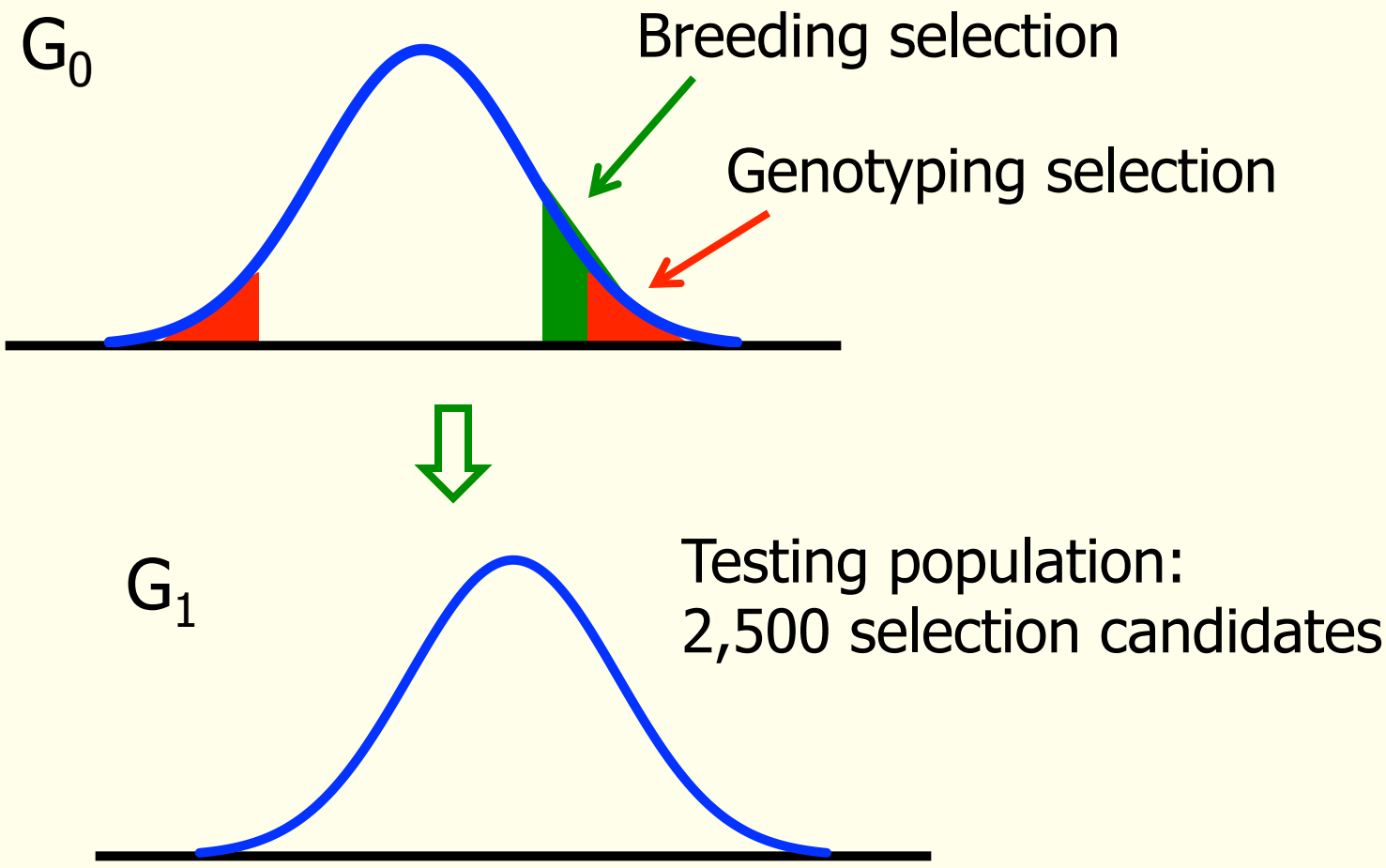
- Mutation rates: QTL 2.5×10^{-5}, Markers 2.5×10^{-3}
- QTL effects: Normally distributed
- Heritability: $h^2 = 0.10$, 0.25 and 0.50

Material and Methods

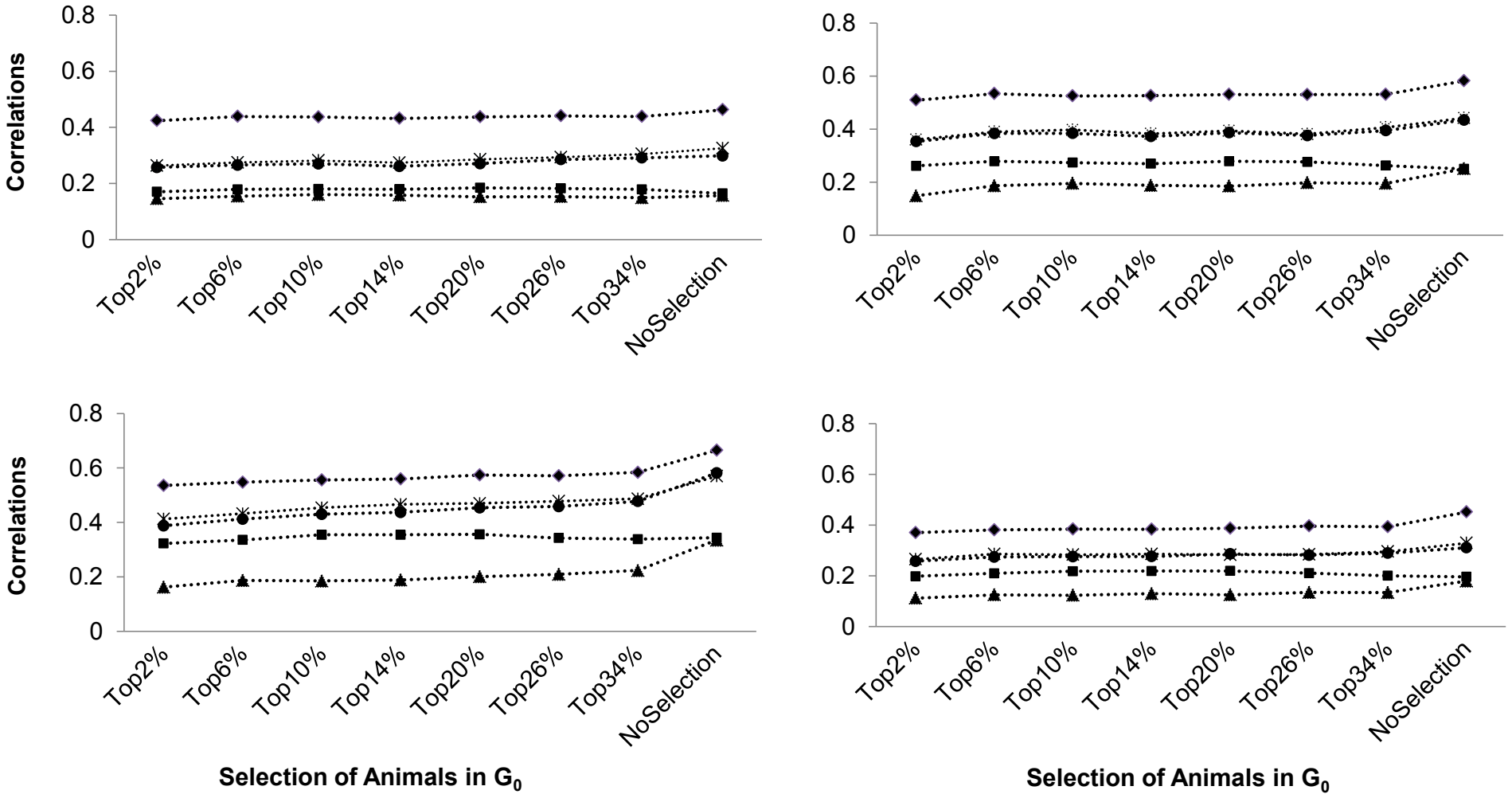
Analysis

- Training population: 500 genotyped animals in G_0
- Selective genotyping strategies:
  Random, Top, Bottom, Extreme, Less related
- Testing population: Generation G_1
- Model: Bayesian LASSO
- Performance: Correlations between GEBV and TBV (accuracy), and
  Predictive mean square error

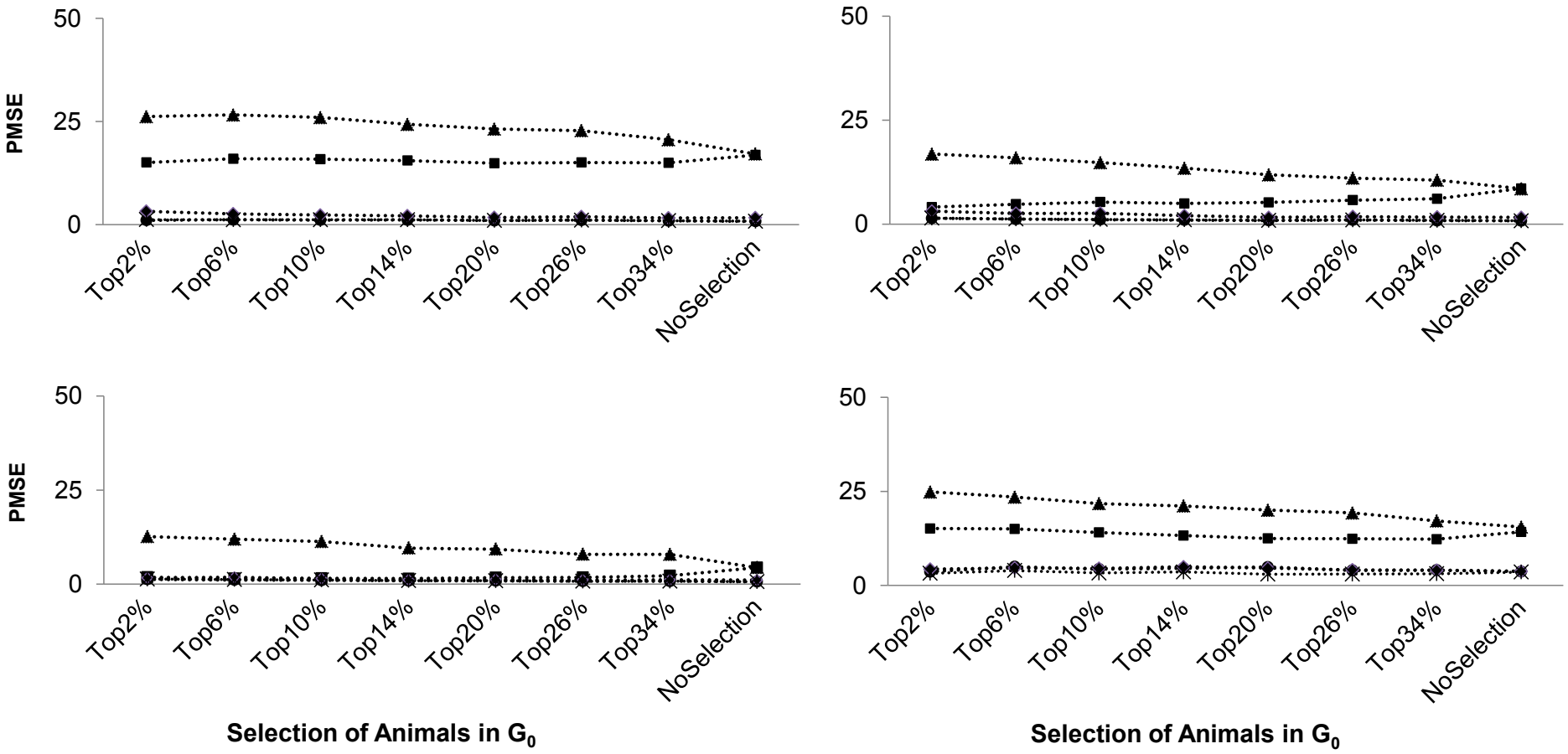
Material and Methods



Prediction accuracies (correlations between GEBV and TBV)



Predictive mean squared error (PMSE)



Number and percentage of coincidence animals

	No selection	Top2%	Top6%	Top10%	Top14%	Top20%	Top26%	Top34%
10% best animals (250 animals)								
Random	28% (71)	25% (62)	24% (61)	26% (64)	24% (61)	25% (63)	24% (60)	25% (62)
Top	21% (52)	24% (59)	22% (55)	22% (56)	22% (55)	22% (55)	22% (56)	22% (54)
Bottom	20% (49)	13% (33)	14% (36)	16% (39)	15% (38)	15% (37)	15% (37)	16% (39)
Extreme	38% (94)	33% (82)	34% (84)	34% (86)	34% (84)	34% (85)	35% (87)	38% (95)
Less Related	29% (73)	22% (54)	26% (64)	26% (64)	25% (62)	26% (65)	26% (65)	27% (67)
25% best animals (625 animals)								
Random	45% (279)	37% (234)	41% (258)	42% (263)	41% (256)	42% (262)	41% (259)	42% (262)
Top	36% (228)	36% (226)	39% (242)	38% (235)	37% (230)	38% (238)	38% (236)	38% (235)
Bottom	34% (215)	31% (191)	31% (194)	32% (202)	32% (200)	31% (196)	32% (199)	32% (201)
Extreme	51% (317)	49% (304)	50% (310)	49% (306)	49% (309)	50% (311)	50% (313)	50% (312)
Less Related	45% (279)	39% (246)	42% (260)	42% (264)	41% (257)	42% (265)	42% (261)	43% (267)
50% best animals (1250 animals)								
Random	64% (803)	60% (750)	62% (774)	63% (785)	62% (777)	63% (785)	62% (775)	62% (780)
Top	57% (718)	58% (726)	59% (741)	59% (734)	58% (729)	59% (742)	59% (733)	58% (730)
Bottom	58% (726)	55% (686)	56% (695)	56% (701)	53% (669)	56% (699)	56% (702)	56% (699)
Extreme	70% (871)	67% (839)	68% (847)	68% (845)	68% (846)	68% (851)	68% (846)	68% (847)
Less Related	64% (804)	61% (769)	58% (723)	63% (789)	62% (780)	63% (785)	62% (778)	62% (777)

Concluding Remarks

- Lowest accuracies with the Bottom strategy
- Random, Extreme and Less Related strategies: accuracies improved with lowest selection intensity
- These three strategies were better than the Top approach
- Extreme, Random and Less Related strategies showed lower prediction mean squared errors (PMSE), followed by the Top and then by the Bottom methods
- Overall, the Extreme genotyping strategy led to the best predictive ability