Supplemental material

Optimal breeding value prediction using a Sparse Selection Index

Marco Lopez-Cruz[*] & Gustavo de los Campos[†, ‡, §, 1]

[*] Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, 48824, USA.

[†] Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, 48824, USA.

[‡] Department of Statistics and Probability, Michigan State University, East Lansing, MI, 48824, USA.

[§] Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI, 48824, USA

[1] Corresponding author. Institute for Quantitative Health Science and Engineering, 775 Woodlot Dr, Office 1311, East Lansing, MI, 48824, USA. Tel: +1(517)884-7607. E-mail: gustavoc@msu.edu

FIGURES

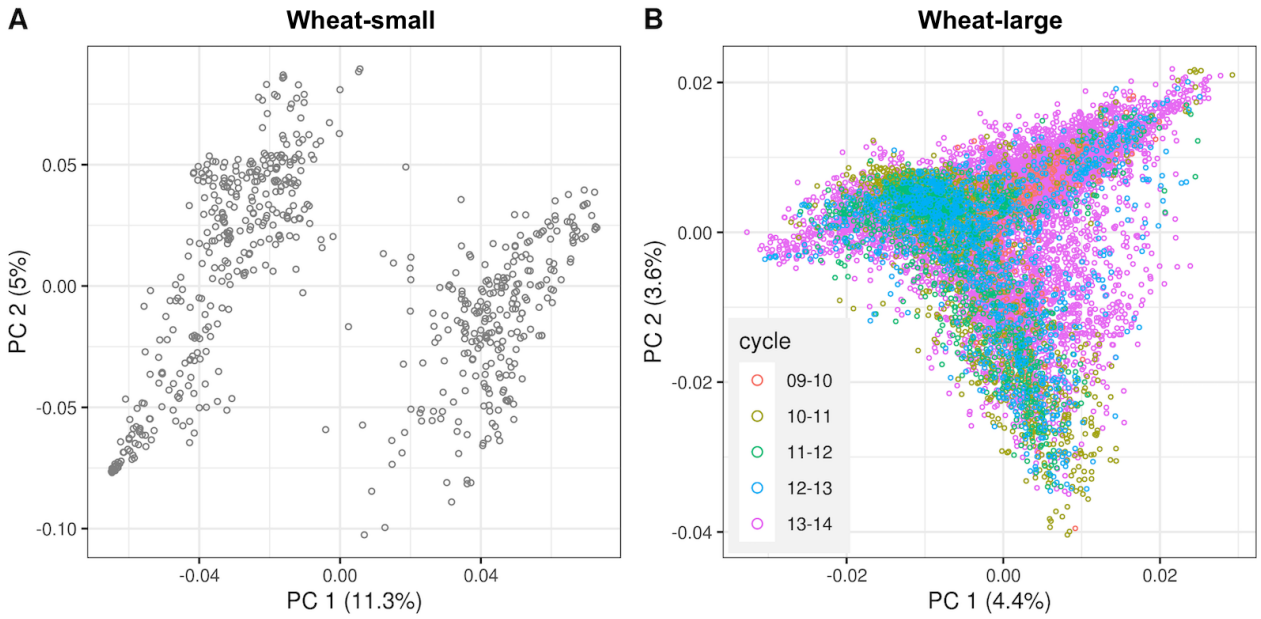

Figure S1. Top two principal components of the genomic relationship matrix, ***G***, for each data set. Each point represents individuals. (A) Wheat-small data set. (B) Wheat-large data set. Individuals are color-grouped by the cycle (sowing-harvest year).

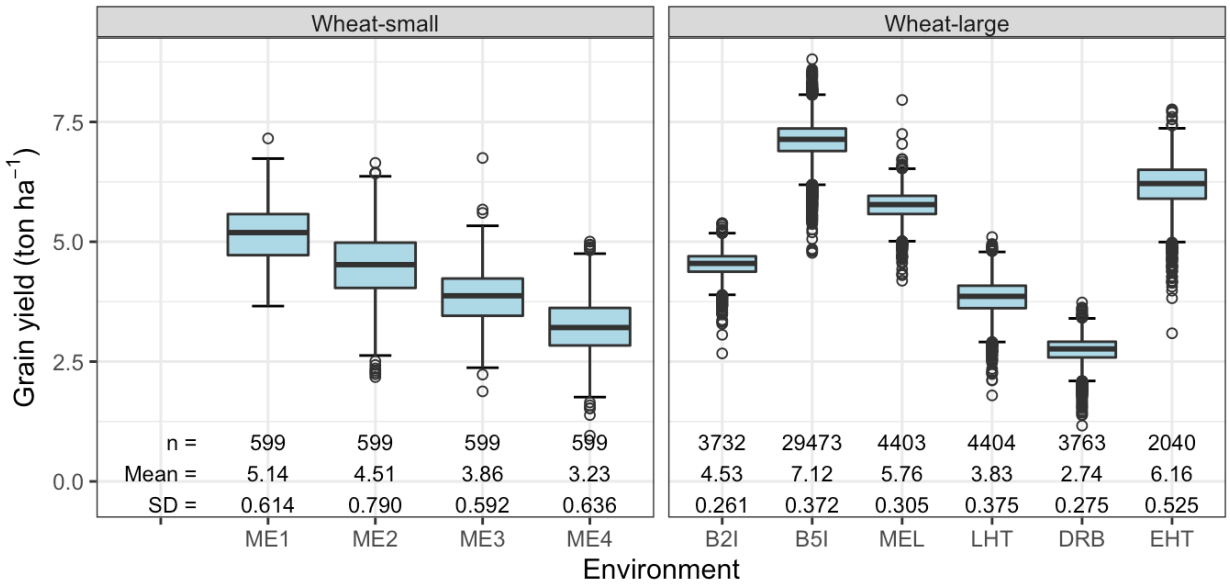

Figure S2. Boxplot of grain yield phenotypic records (in ton ha^{-1}) by environmental condition for both Wheat-small and Wheat-large data sets. SD standard deviation.

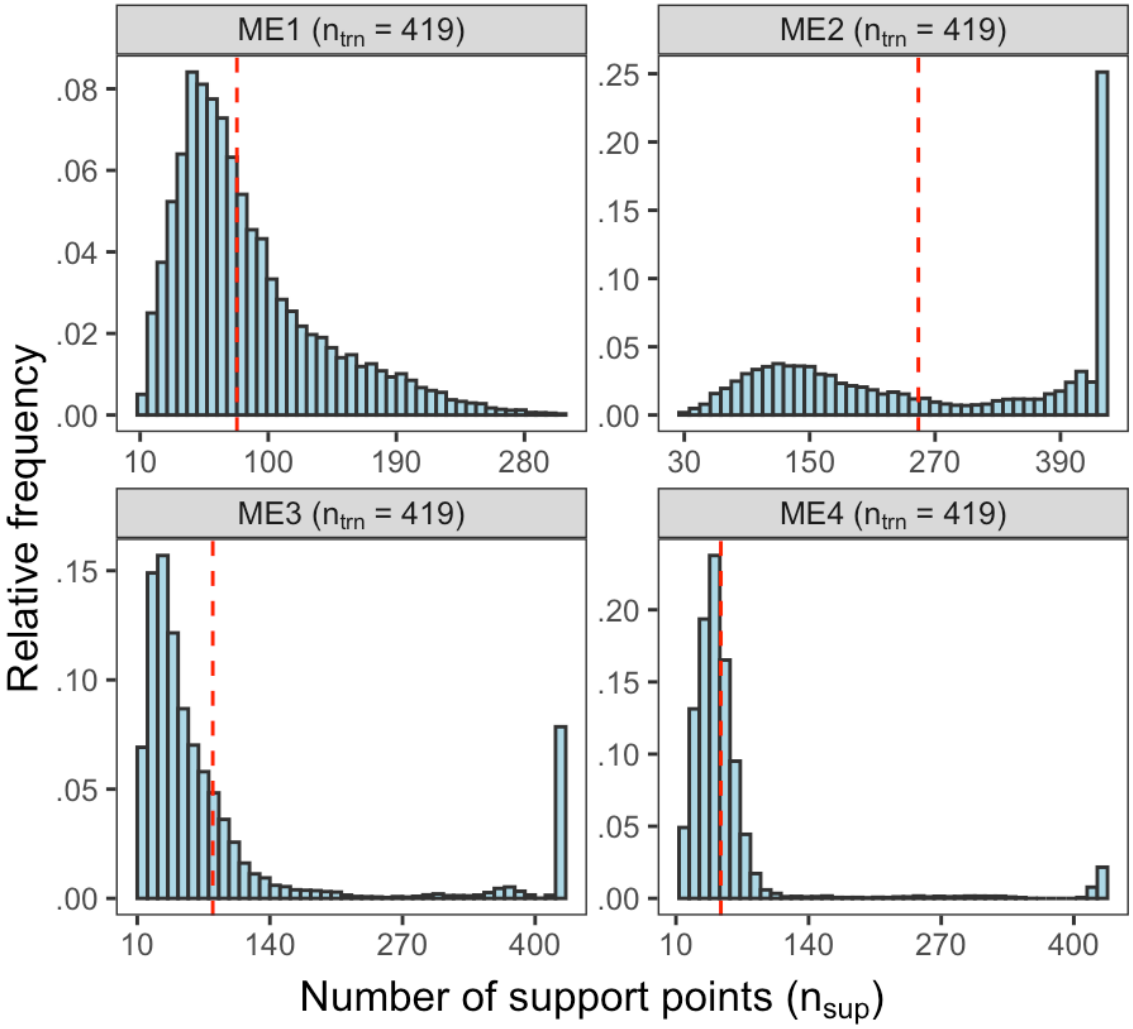

Figure S3. Distribution of the number of training support points (n_{sup}) in the optimal SSI for grain yield (results obtained over 100 trn-tst partitions; n_{trn}= size of the training data set), by environmental condition (ME: mega-environment), Wheat-small data set.

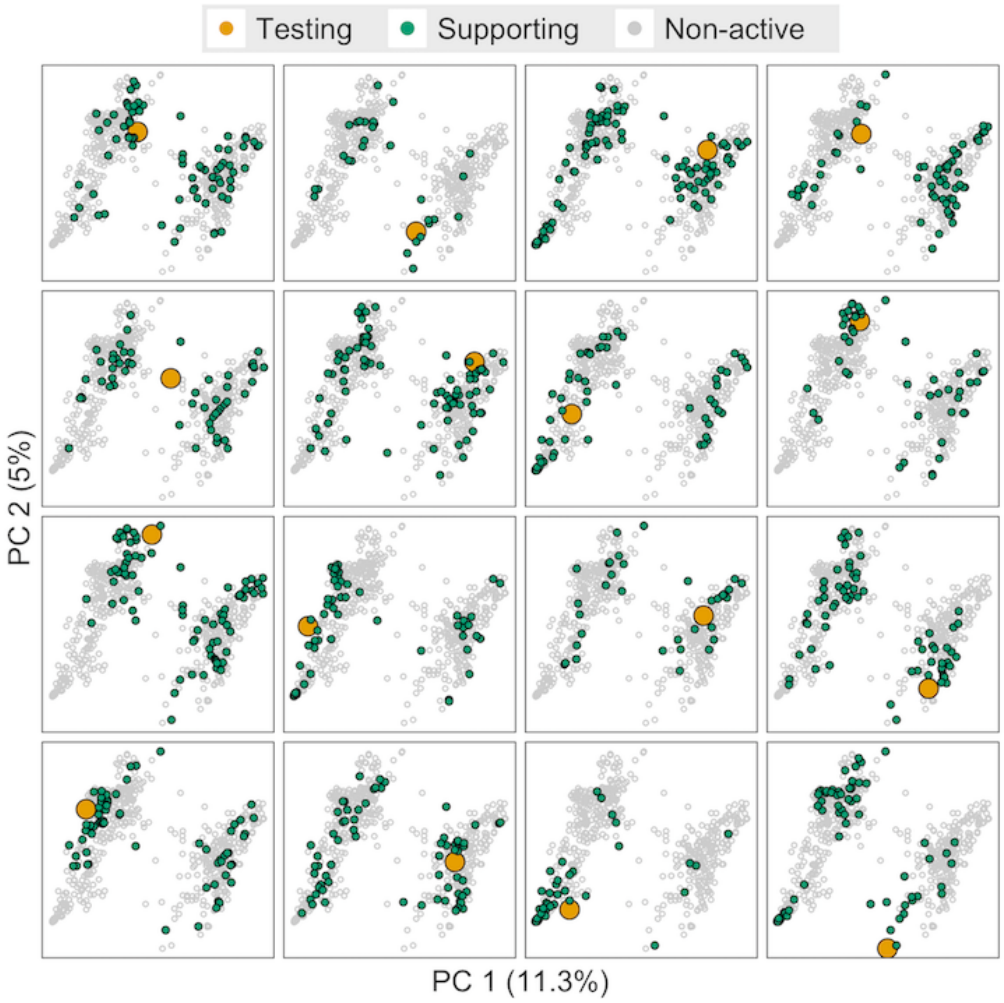

Figure S4. First two principal components coordinates for prediction points (yellow) and the corresponding support points (green). Grey points represent genotypes that did not contribute to the prediction of the genetic value of grain yield of the genotype in yellow. All panels represent solutions for the mega-environment ME1, Wheat-small data set.

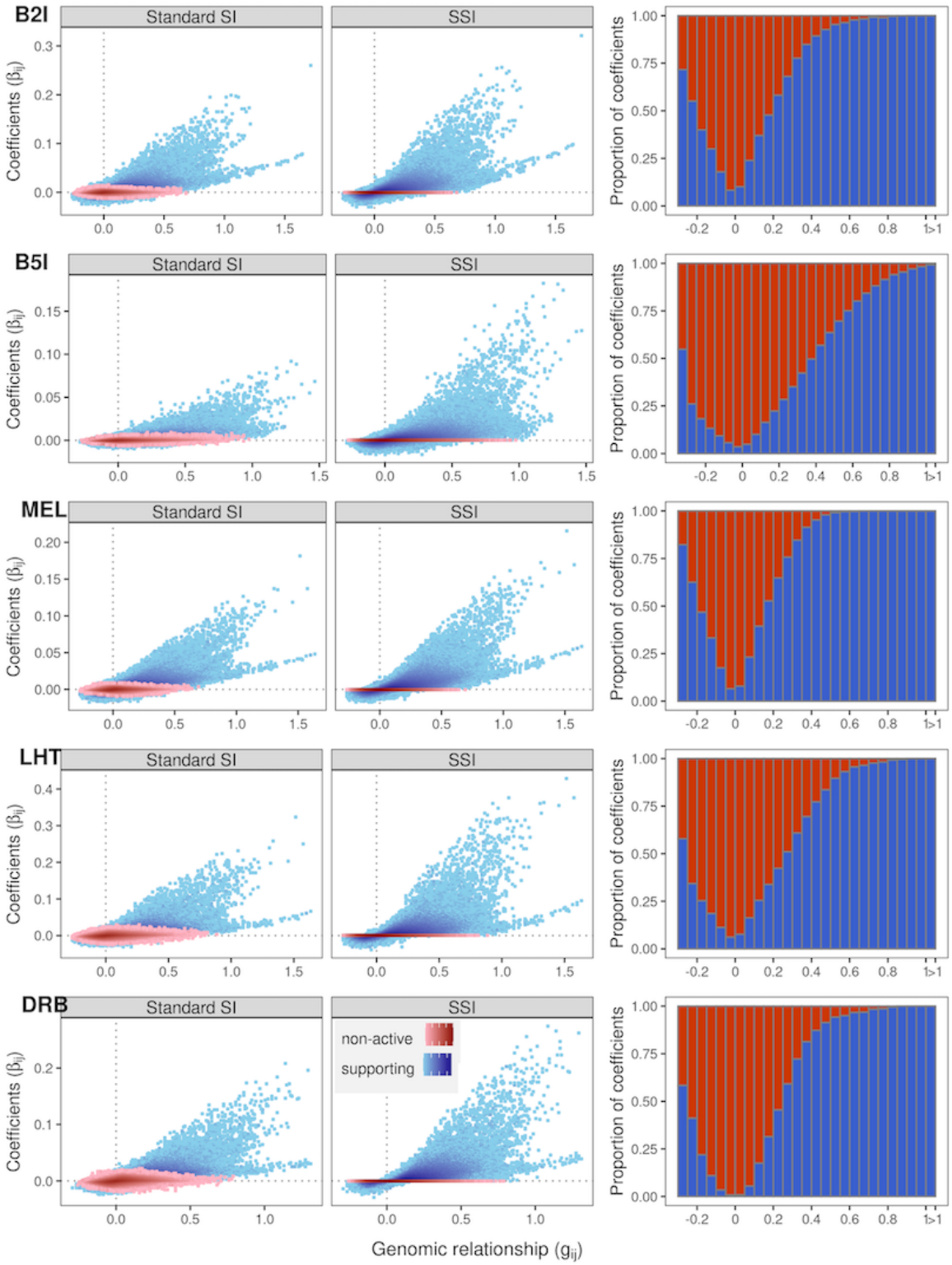

Figure S5. (left and center) Weights (β_{ij}) of a standard SI (G-BLUP) and of the optimal SSI versus the genomic relationship (g_{ij}), and (right) proportion of weights in the SSI that belonged to either the supporting or non-active sets, by genomic-relationship; by environment, Wheat-large data set.

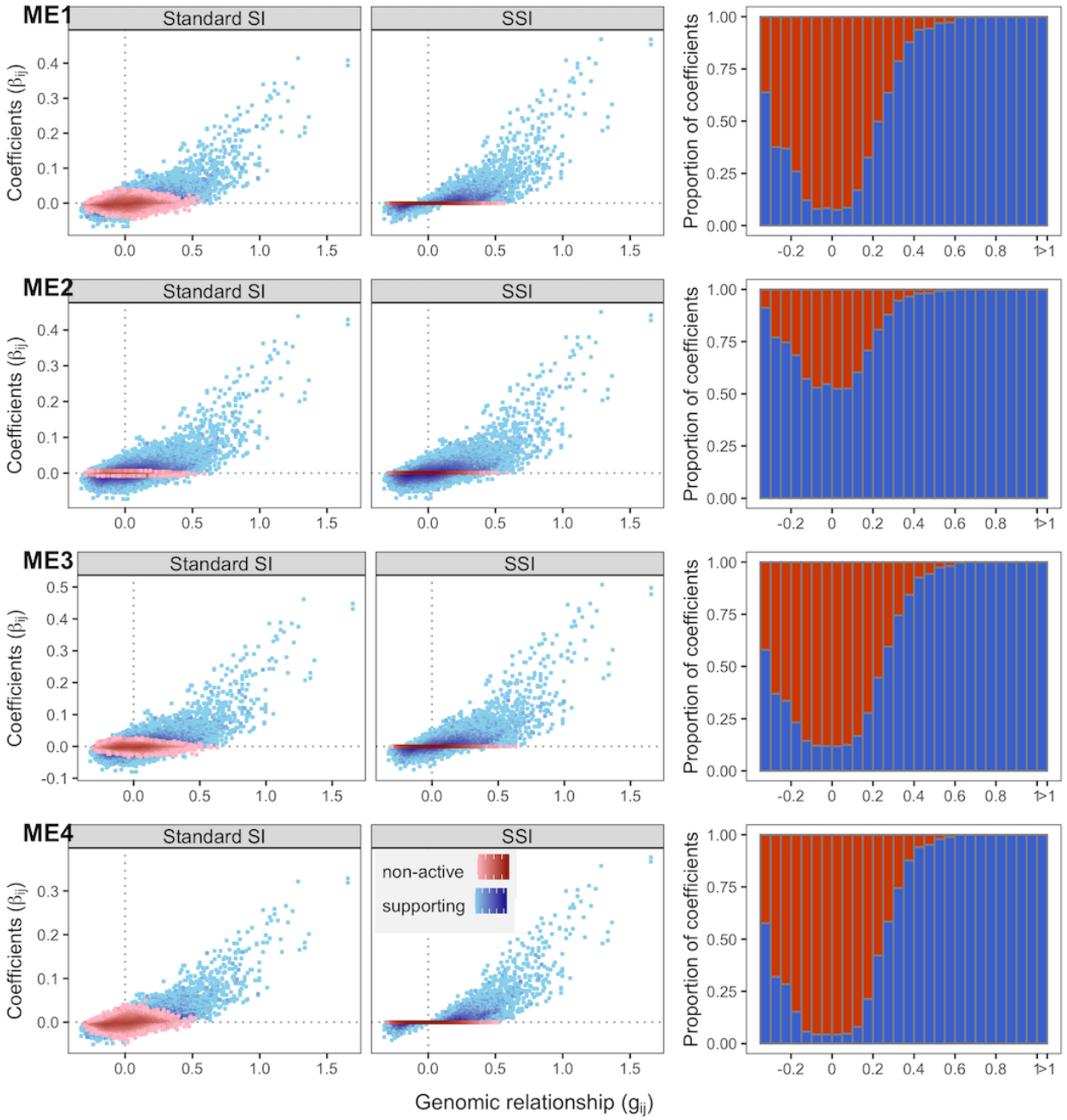

Figure S6. (left and center) Weights (β_{ij}) of a standard SI (G-BLUP) and of the optimal SSI versus the genomic relationship (g_{ij}), and (right) proportion of weights in the SSI that belonged to either the supporting or non-active sets, by genomic-relationship; by environment, Wheat-small data set.

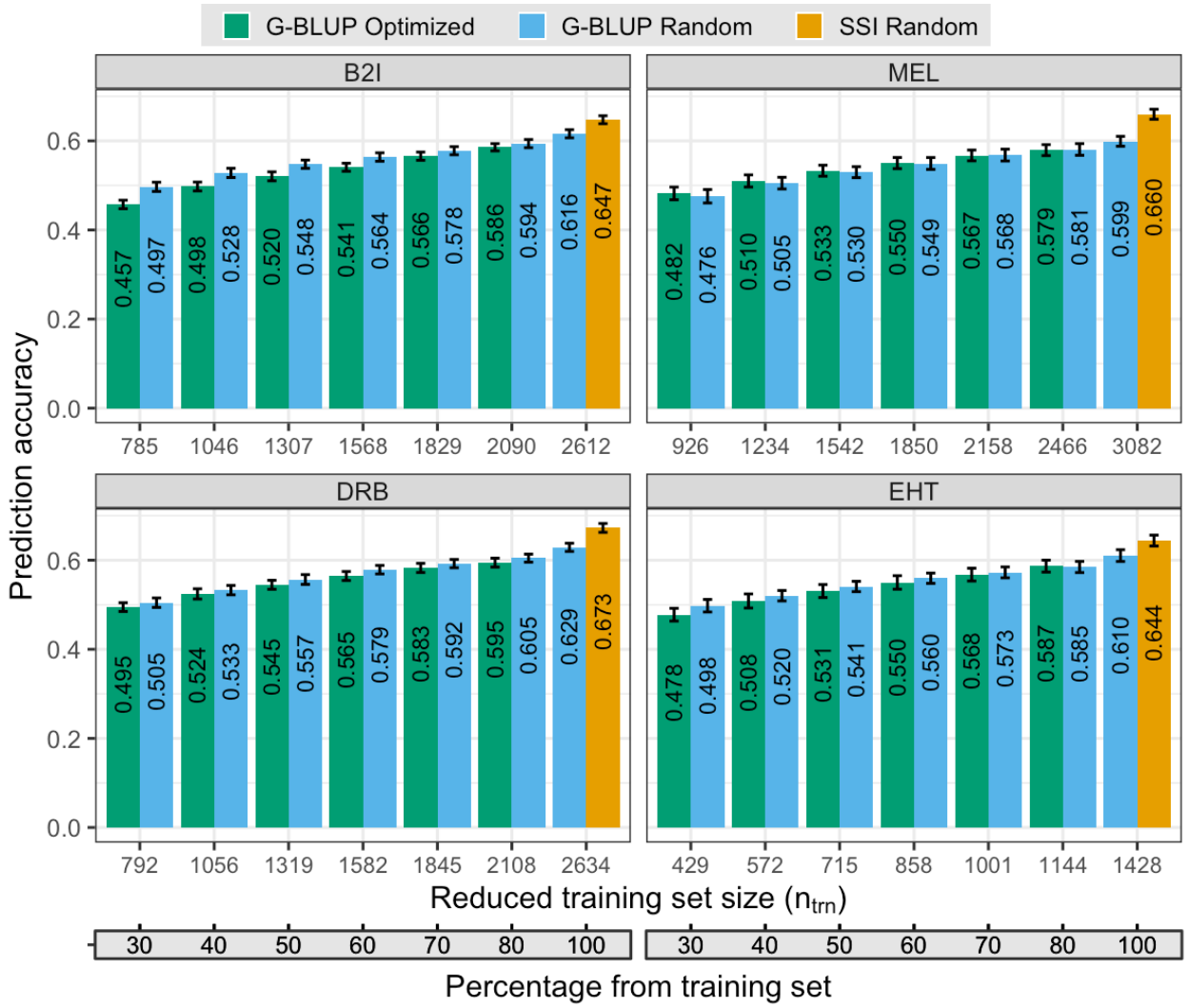

Figure S7. Prediction accuracy for grain yield (average over 50 partitions) of the optimal SSI and of the G-BLUP. For each training-testing (trn-tst) partition, the testing set (30%) was predicted using the (random) training set (70%) with the SSI and G-BLUP. Then, the same testing set was predicted with the initial training set reduced to 80%, 70%, …, 30% of its initial size by: (i) randomly sampling (G-BLUP Random) and (ii) selection using the expected reliability (CDmean) optimization criteria (G-BLUP Optimized, as described in Rincent *et al.* 2012 and implemented in the STPGA R-package, Akdemir *et al.*, 2015). By environmental condition (B2I: bed planting + 2 irrigations, MEL: flat planting + 5 irrigations, EHT: early planting date, DRB: bed planting + drip irrigation), Wheat-large data set.

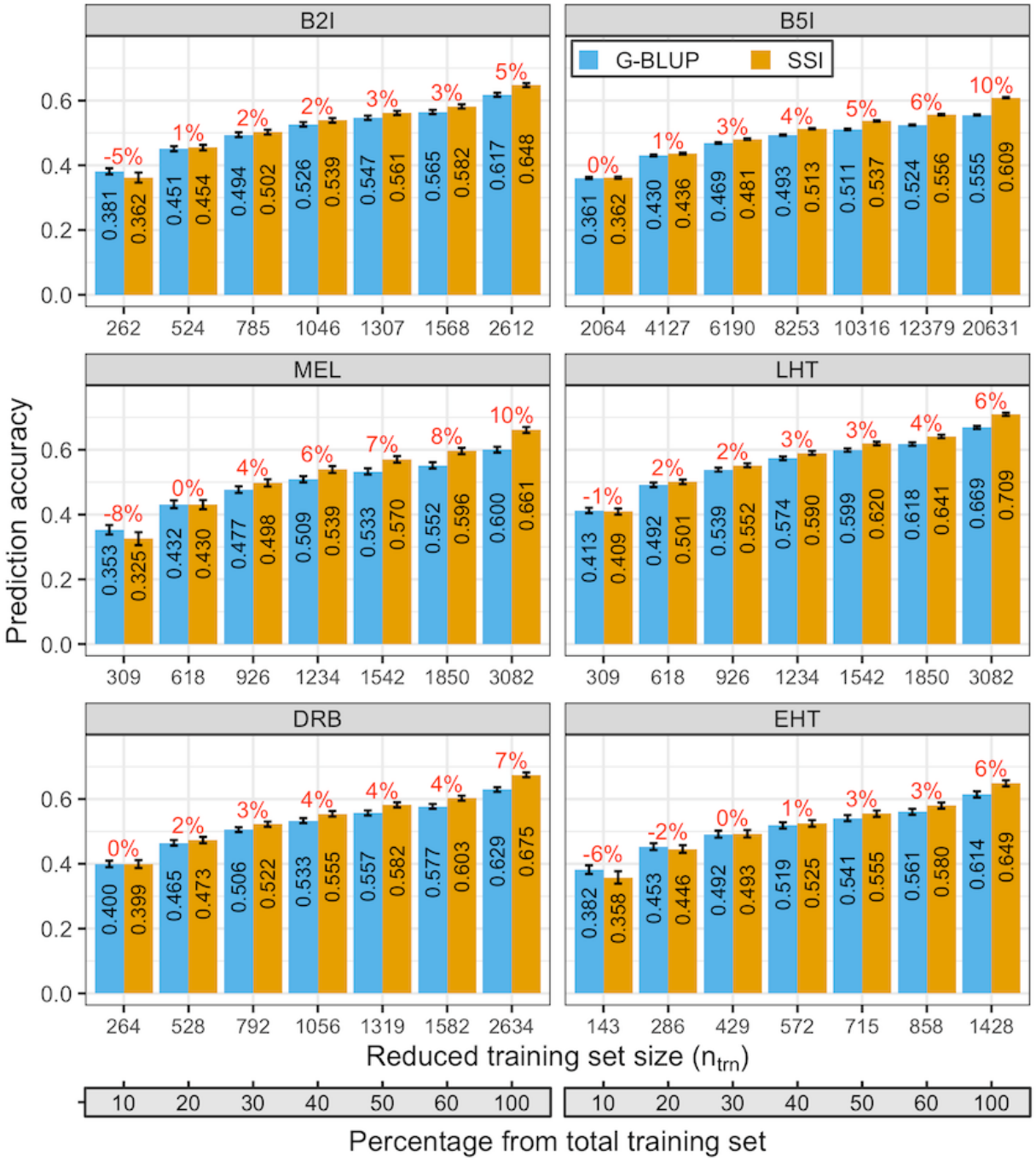

Figure S8. Prediction accuracy for grain yield (average over 100 partitions) of the optimal SSI and of the G-BLUP. For each training-testing (trn-tst) partition, the testing set (30%) was predicted using the training set (70%). Then, the same testing set was predicted with the initial training set reduced to 60%, 50%, …, 10% of its initial size by randomly sampling down. The numbers on top of the bars represent the gain (in percentage) in accuracy of the SSI over the G-BLUP. By environmental condition (B2I: bed planting + 2 irrigations, MEL: flat planting + 5 irrigations, EHT: early planting date, DRB: bed planting + drip irrigation), Wheat-large data set.

TABLES

Table S1. Number of available observations, average grain yield, and heritability by environmental condition for the Wheat-large data set

Planting conditions		Number of irrigations	Name	n	Average (SD) Yield	Heritability (SD)[a]
Date	System					
Optimum	Bed	2	B2I	3,732	4.53 (0.261)	0.41 (0.029)
Optimum	Bed	5	B5I	29,473	7.12 (0.372)	0.57 (0.025)
Optimum	Flat	5	MEL	4,403	5.76 (0.305)	0.23 (0.025)
Late	Bed	5	LHT	4,404	3.83 (0.375)	0.51 (0.025)
Optimum	Bed	Minimal	DRB	3,763	2.74 (0.275)	0.38 (0.029)
Early	Bed	5	EHT	2,040	6.16 (0.525)	0.41 (0.038)

SD. Standard deviation. [a]Posterior mean and SD obtained across 10,000 Monte Carlo replicates using Gibbs sampling.

Table S2. Number of available observations, average grain yield, and heritability by environmental condition for the Wheat-small data set

Moisture regime	Temperature	Name	n	Average (SD) Yield	Heritability (SD)[a]
Optimal irrigation, low rainfall	Optimal	ME1	599	5.14 (0.614)	0.50 (0.054)
High rainfall	Optimal	ME2	599	4.51 (0.790)	0.46 (0.056)
Low rainfall	High drought	ME3	599	3.86 (0.592)	0.43 (0.062)
Irrigation or rainfall	Hot, low humidity	ME4	599	3.23 (0.636)	0.44 (0.061)

SD. Standard deviation. [a]Posterior mean and SD obtained across 10,000 Monte Carlo replicates using Gibbs sampling.

Table S3. Prediction accuracy for grain yield (average across 50 partitions) achieved by the SSI for different values of the parameter α of an Elastic-Net-type SSI, by environmental condition for the Wheat-large data set

Environment	λ_0[a]	α	λ_{opt}[b]	n_{sup}[c]	Accuracy (SD)
B2I	1.5320	1.00	0.0141	395	0.649 (0.032)
	0.7660	0.25	0.0667	320	0.663 (0.032)
		0.50	0.0320	330	0.664 (0.032)
		0.75	0.0233	290	0.664 (0.032)
		1.00	0.0155	338	0.664 (0.032)
B5I	1.8412	1.00	0.0119	1,226	0.610 (0.009)
	0.9215	0.25	0.0460	1,187	0.630 (0.009)
		0.50	0.0223	1,203	0.631 (0.009)
		0.75	0.0164	1,044	0.631 (0.009)
		1.00	0.0132	943	0.631 (0.009)
MEL	3.7934	1.00	0.0116	561	0.665 (0.046)
	1.8967	0.25	0.0705	338	0.685 (0.045)
		0.50	0.0406	270	0.686 (0.045)
		0.75	0.0294	236	0.687 (0.045)
		1.00	0.0195	282	0.687 (0.045)
LHT	0.9841	1.00	0.0218	237	0.712 (0.026)
	0.4921	0.25	0.0854	248	0.727 (0.025)
		0.50	0.0491	194	0.729 (0.025)
		0.75	0.0295	223	0.729 (0.025)
		1.00	0.0235	202	0.730 (0.025)
DRB	1.7555	1.00	0.0344	103	0.679 (0.038)
	0.8778	0.25	0.1461	102	0.694 (0.039)
		0.50	0.0823	79	0.696 (0.039)
		0.75	0.0588	69	0.697 (0.040)
		1.00	0.0386	85	0.697 (0.040)
EHT	1.5514	1.00	0.0284	120	0.657 (0.046)
	0.7757	0.25	0.0970	159	0.670 (0.048)
		0.50	0.0554	126	0.672 (0.048)
		0.75	0.0399	110	0.672 (0.048)
		1.00	0.0264	133	0.673 (0.048)

SD: Standard deviation across the 50 training-testing partitions. [a]Shrinkage factor involved in the standard SI (Equation 2). Within environment, in the top row a value of λ_0 was used as in the G-BLUP and in rows below, λ_0 was reduced to half. [b]Optimal value of λ (average across partitions) estimated by cross-validating the training set. [c]Average number of support points in the SSIs.

Table S4. Prediction accuracy for grain yield (average across 50 partitions) achieved by the SSI for different values of the parameter α of an Elastic-Net-type SSI, by environmental condition for the Wheat-small data set

Environment	λ_0[a]	α	λ_{opt}[b]	n_{sup}[c]	Accuracy (SD)
ME1	1.2101	1.00	0.0314	84	0.769 (0.062)
	0.5061	0.25	0.1042	99	0.772 (0.063)
		0.50	0.0492	101	0.773 (0.063)
		0.75	0.0296	110	0.773 (0.063)
		1.00	0.0236	103	0.773 (0.063)
ME2	1.3034	1.00	0.0175	151	0.708 (0.085)
	0.6517	0.25	0.0686	147	0.711 (0.086)
		0.50	0.0397	126	0.710 (0.086)
		0.75	0.0240	136	0.710 (0.086)
		1.00	0.0192	129	0.710 (0.086)
ME3	1.4084	1.00	0.0514	50	0.609 (0.090)
	0.7042	0.25	0.2213	48	0.611 (0.089)
		0.50	0.1017	48	0.610 (0.090)
		0.75	0.0601	54	0.609 (0.091)
		1.00	0.0474	50	0.609 (0.091)
ME4	1.4380	1.00	0.0615	40	0.722 (0.073)
	0.7190	0.25	0.2689	39	0.727 (0.074)
		0.50	0.1483	31	0.727 (0.074)
		0.75	0.0870	35	0.728 (0.075)
		1.00	0.0681	32	0.728 (0.075)

SD: Standard deviation across the 50 training-testing partitions. [a]Shrinkage factor involved in the standard SI (Equation 2). Within environment, in the top row a value of λ_0 was used as in the G-BLUP and in rows below, λ_0 was reduced to half. [b]Optimal value of λ (average across partitions) estimated by cross-validating the training set. [c]Average number of support points in the SSIs.

REFERENCES

Akdemir D., J. I. Sanchez, and J.-L. Jannink, 2015 Optimization of genomic selection training populations with a genetic algorithm. Genet. Sel. Evol. 47: 1–10.

Rincent R., S. Nicolas, T. Altmann, D. Brunel, P. Revilla, *et al.*, 2012 Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics 192: 715–728.