

KANG2P enables interpretable genotype-to-phenotype prediction through genetically anchored molecular-state learning

Lei Ling^{1,†}, Yang Liu^{1,†}, Pingzhou Wang^{1,†}, Fengxin Li¹,
Hang Ning⁴, Zhimin Zhang¹, Cheng Yang^{2,3}, Liang Qiao^{5,6}, Si Qin⁷,
Alexander Schönhuth⁸, Taihong Wu^{1,*}, Yuansheng Liu^{2,3,*}, Xiao Luo^{1,*}

¹ Hunan Research Center of the Basic Discipline for Cell Signaling, College of Biology, Hunan University, Changsha, China

² College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

³ Yuelushan Laboratory, Changsha, China

⁴ College of Forestry and Horticulture, Hubei Minzu University, Enshi, China

⁵ Taichu (Wuxi) Electronic Technology Co., Ltd., Wuxi, China

⁶ Department of Computer Science and Technology, Tsinghua University, Beijing, China

⁷ Laboratory of Food Function and Nutrigenomics, College of Food Science and Technology, Hunan Agricultural University, Changsha, China

⁸ Faculty of Technology, Bielefeld University, Bielefeld, Germany

[†]These authors contributed equally to the work.

^{*}To whom correspondence should be addressed.

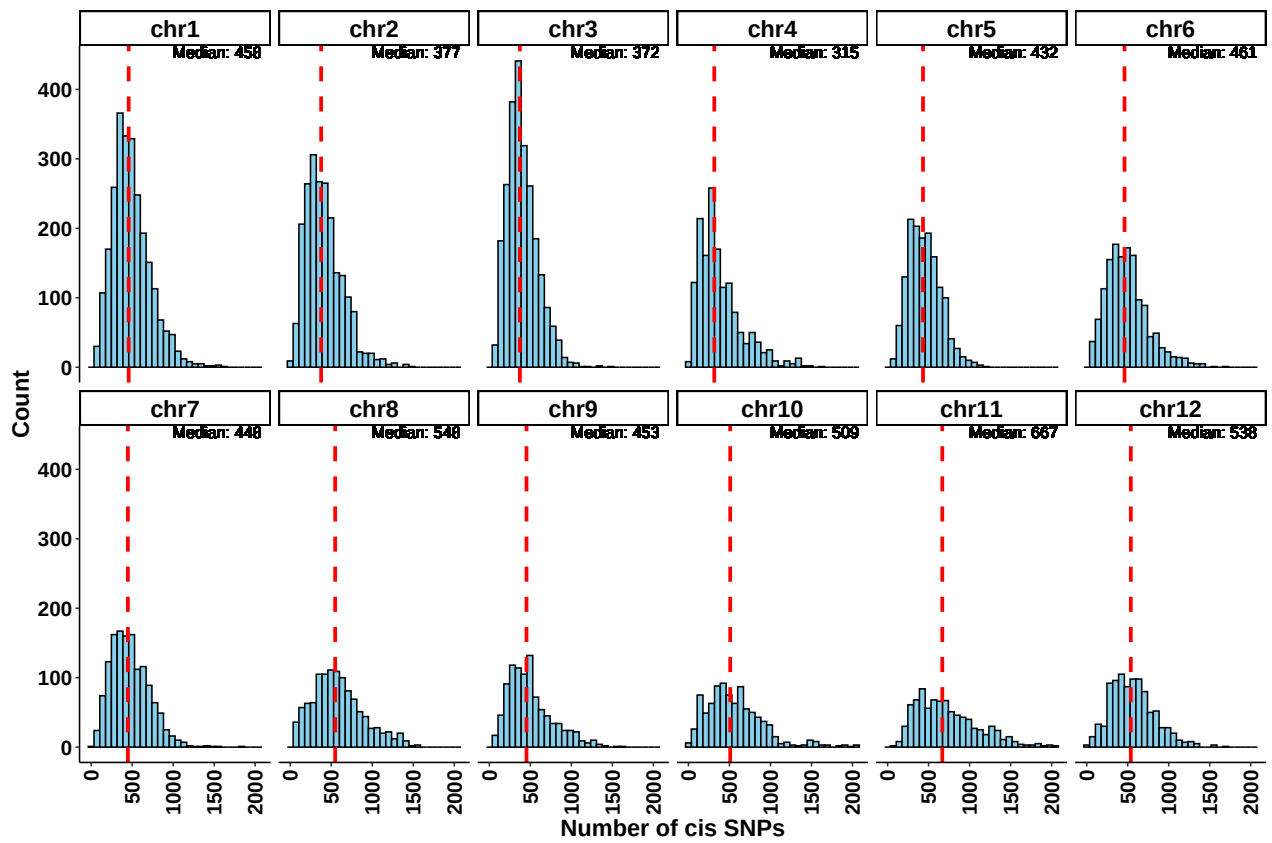
Email: taihongwu@hnu.edu.cn

Email: yuanshengliu@hnu.edu.cn

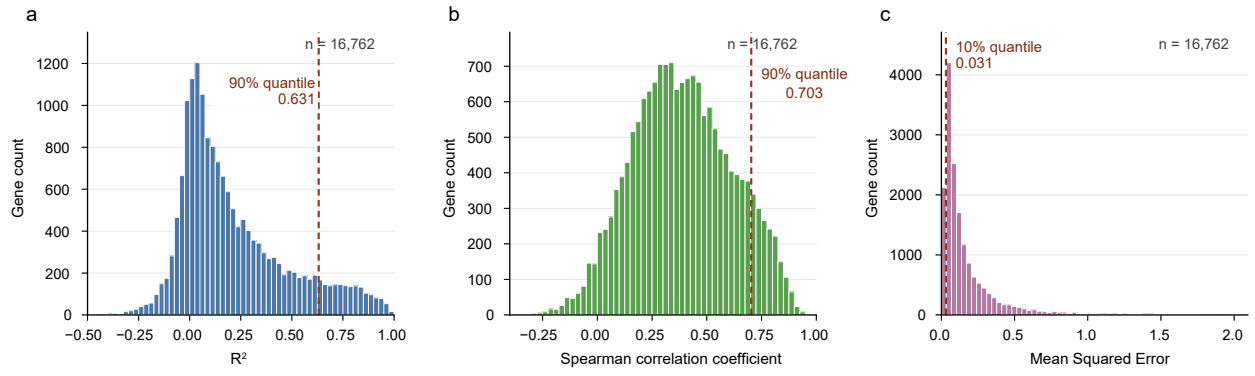
Email: xluo@hnu.edu.cn

Supplementary Figures and Tables

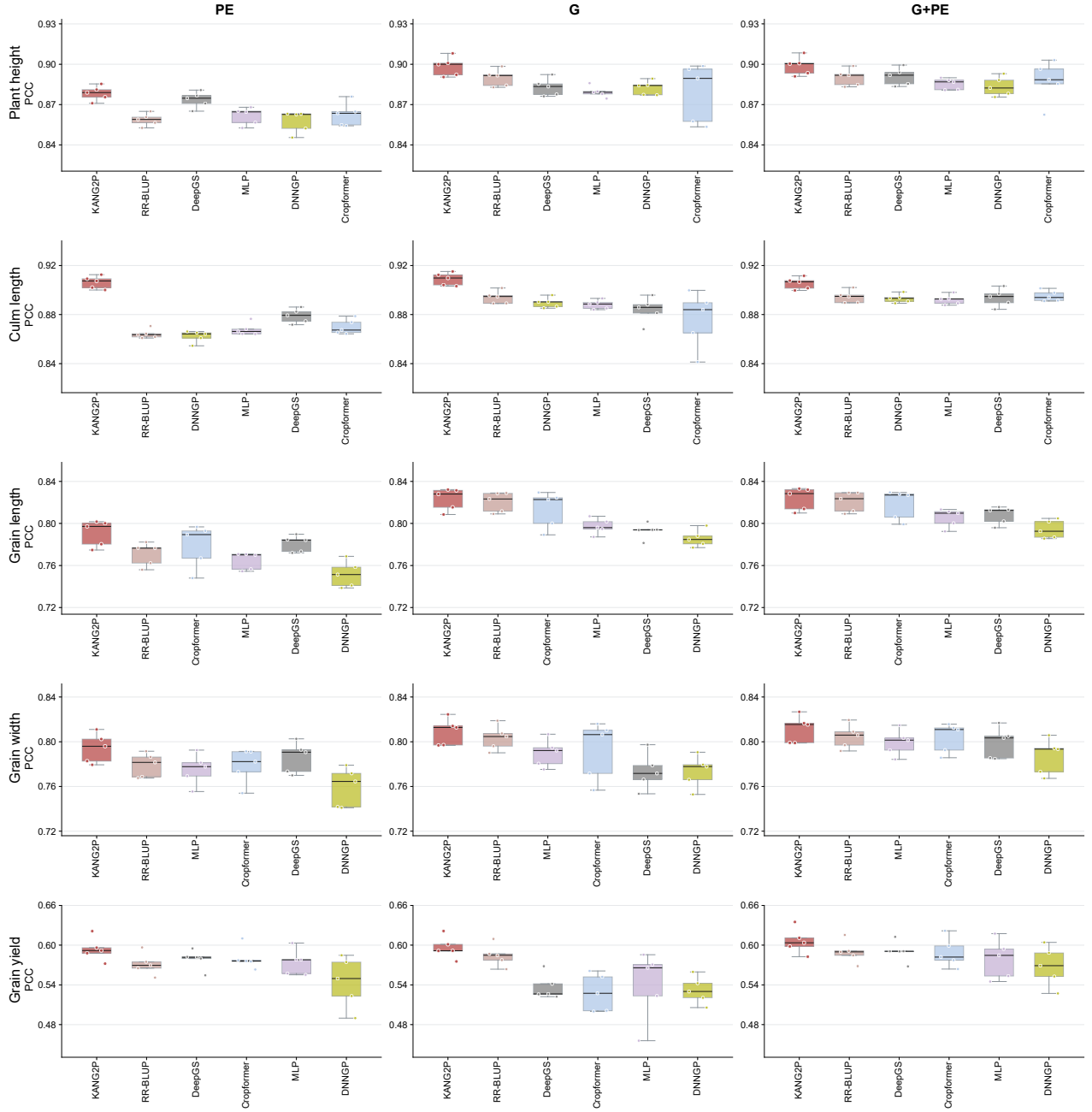
Supplemental Figures



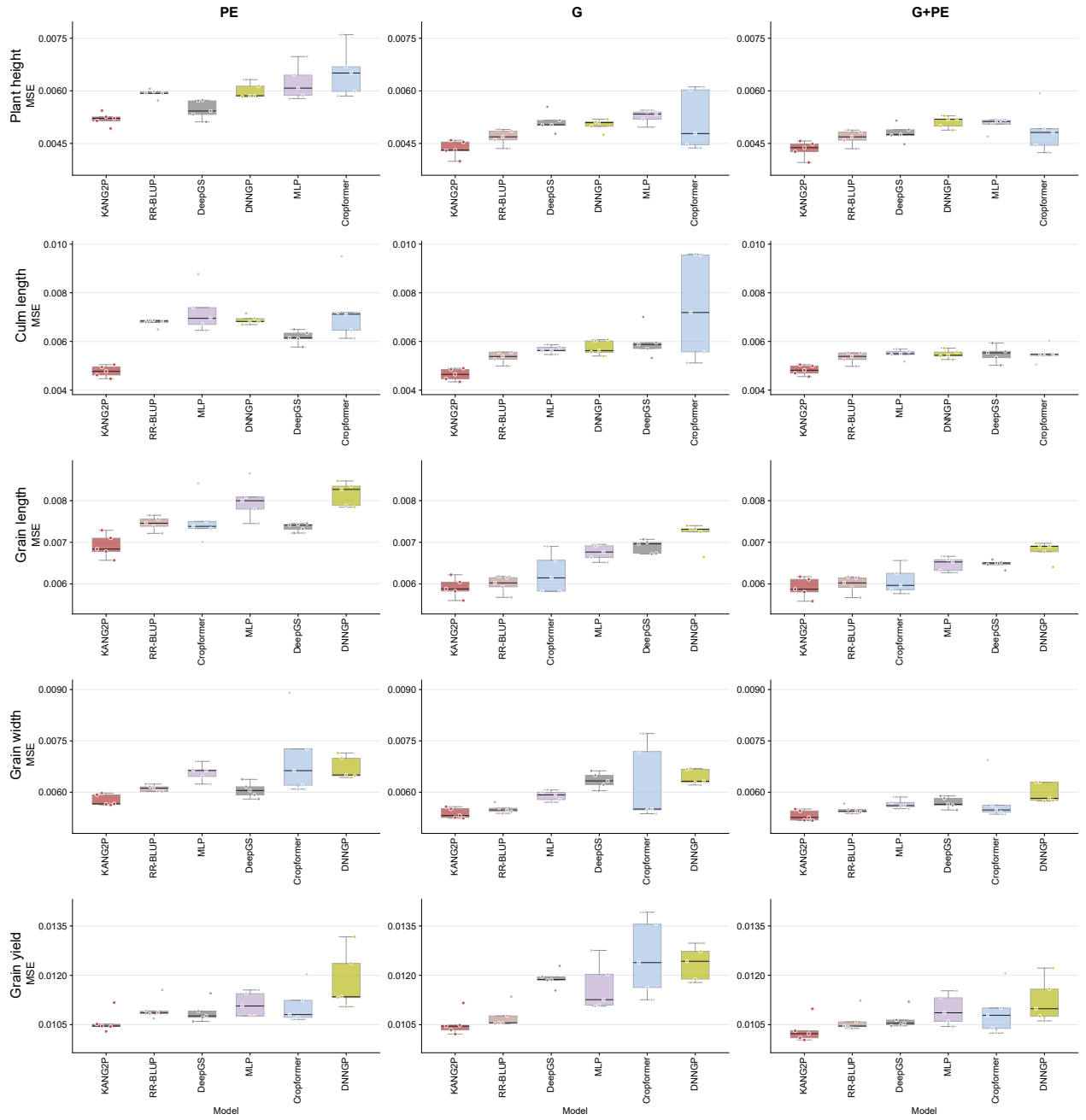
Supp. Fig. 1. Distribution for the number of *cis* SNPs in genes expression levels prediction models on the Rice277 reference population.



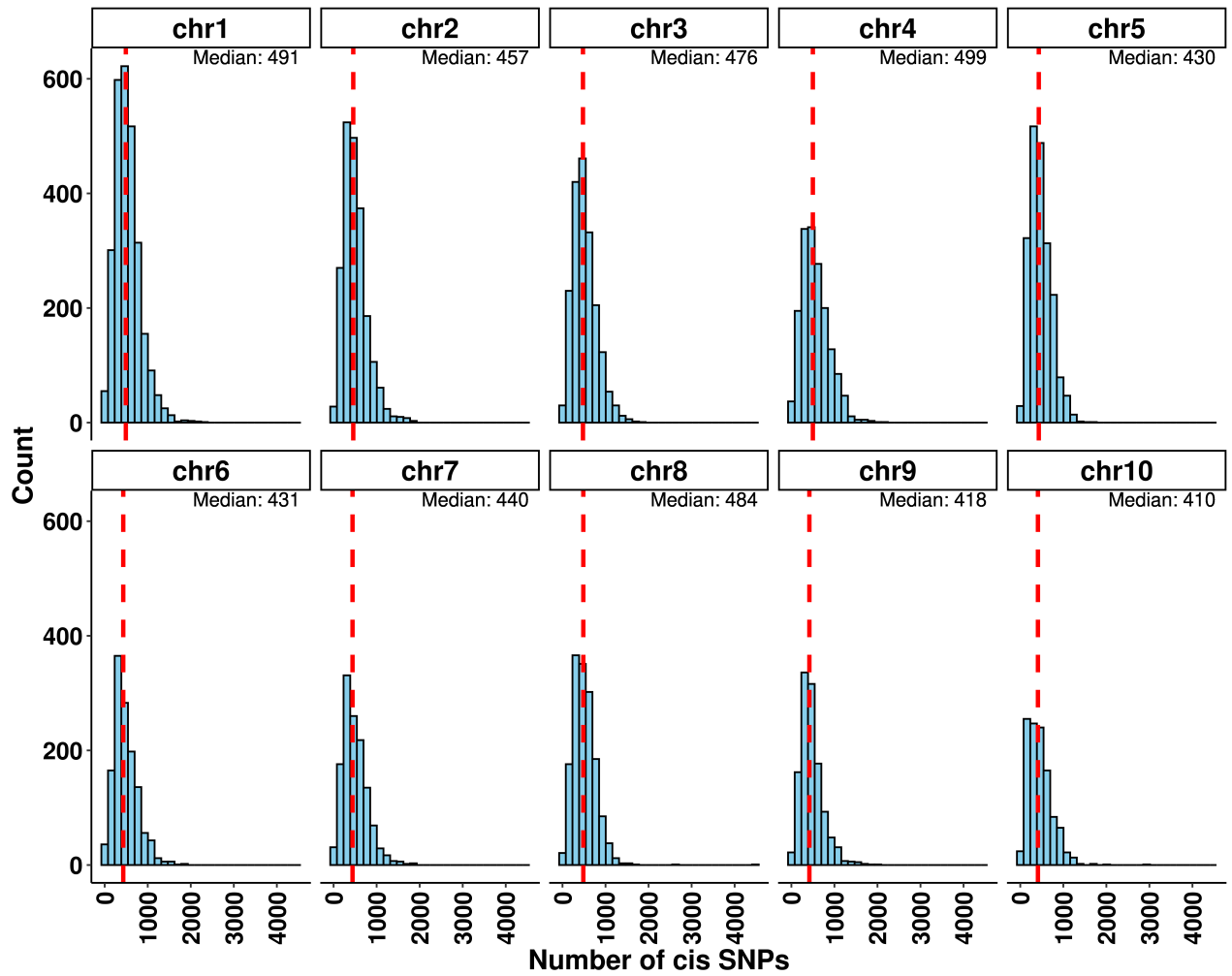
Supp. Fig. 2. Predictive performance of gene expression models in the Rice277 reference population. a–c, Frequency distributions of model performance across gene expression prediction models in the Rice277 reference population, quantified by the coefficient of determination, R^2 (a), Spearman’s correlation coefficient, SCC (b), and mean squared error, MSE (c). Each bar represents the number of genes falling within the corresponding performance interval.



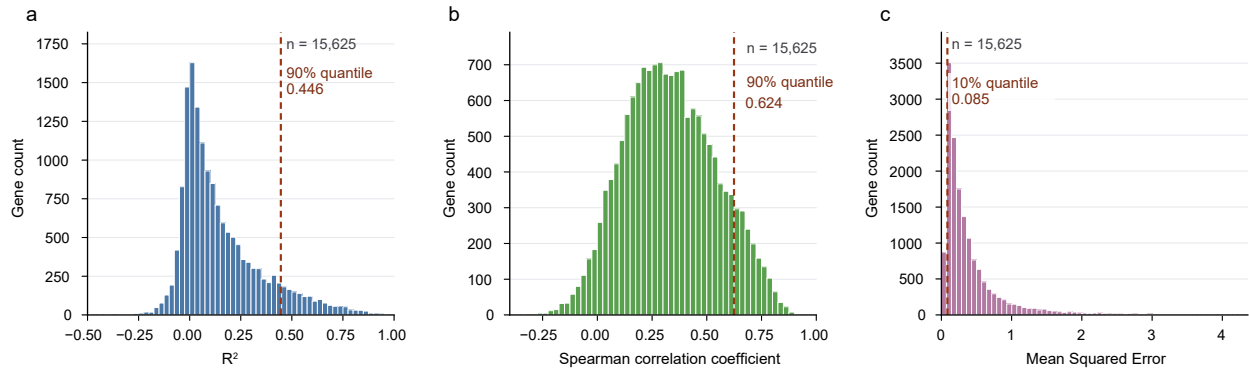
Supp. Fig. 3. Predictive performance comparison of six regression models across Rice18K agronomic traits. Box plots show outer-fold PCC values from nested cross-validation for six regression models across five Rice18K agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. MLP, multilayer perceptron.



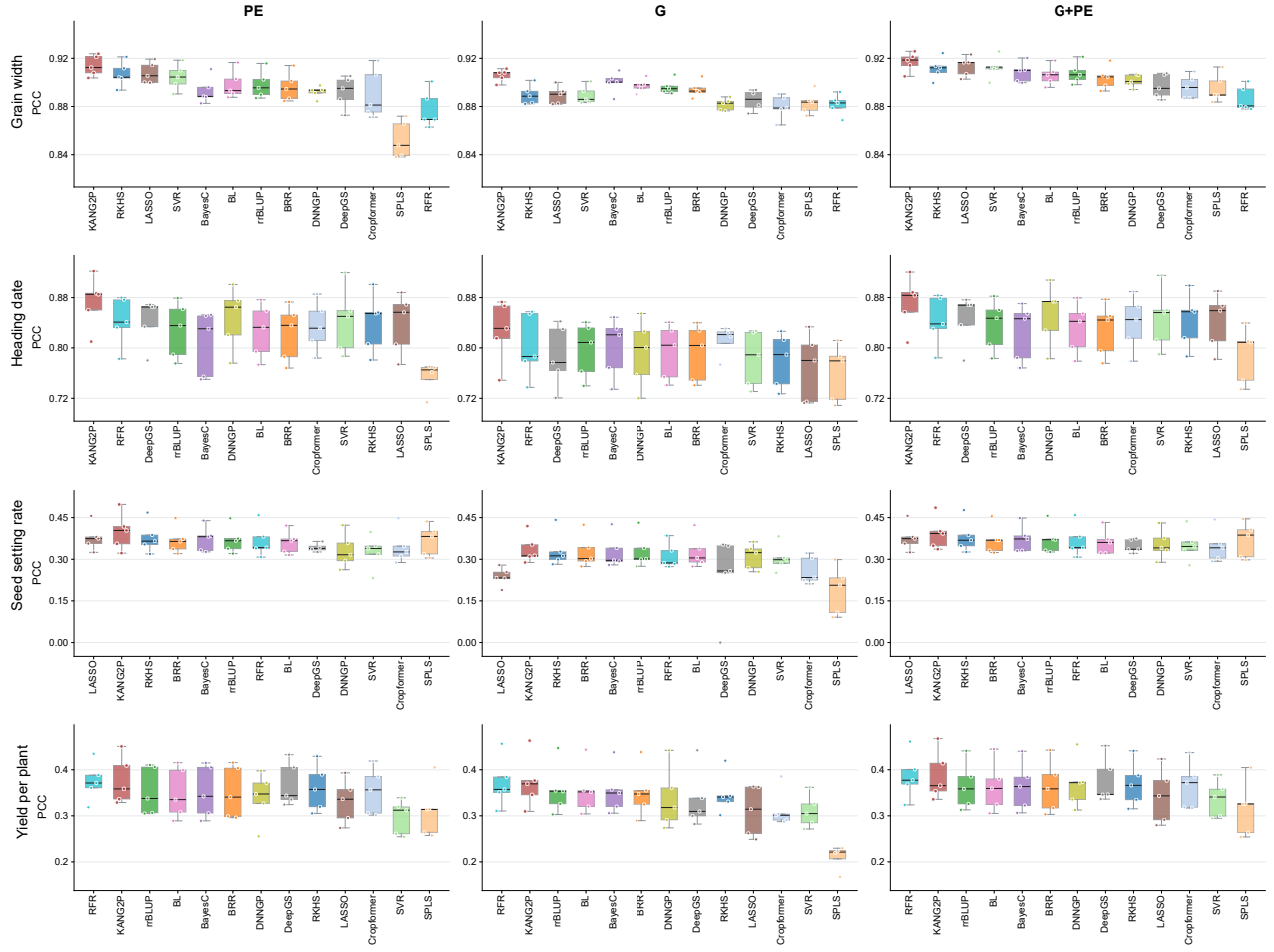
Supp. Fig. 4. Predictive performance comparison of six regression models across Rice18K agronomic traits. Box plots show outer-fold MSE values from nested cross-validation for six regression models across five Rice18K agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. MLP, multilayer perceptron.



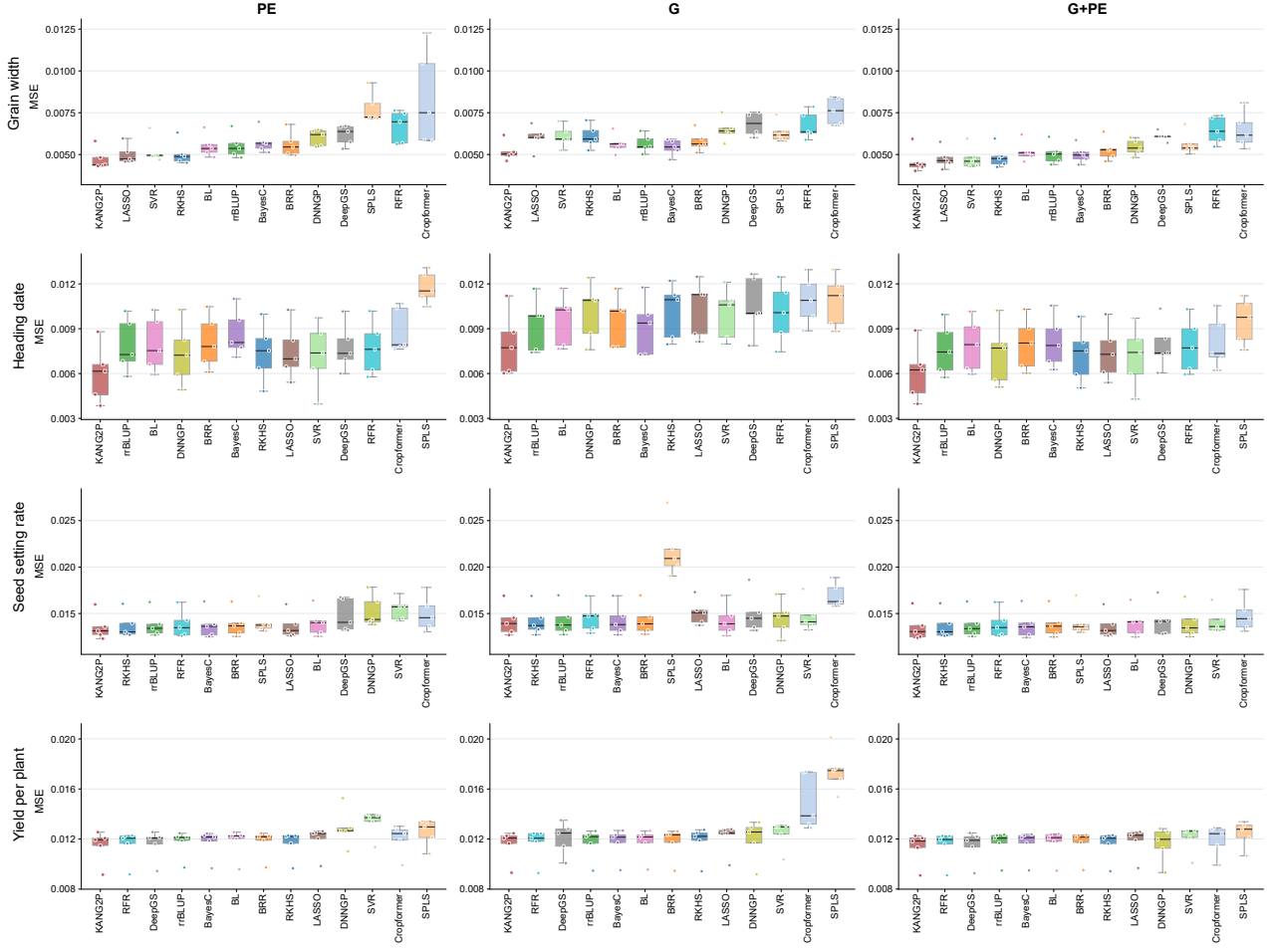
Supp. Fig. 5. Distribution for the number of *cis* SNPs in gene expression levels prediction models on the Maize CUBIC population.



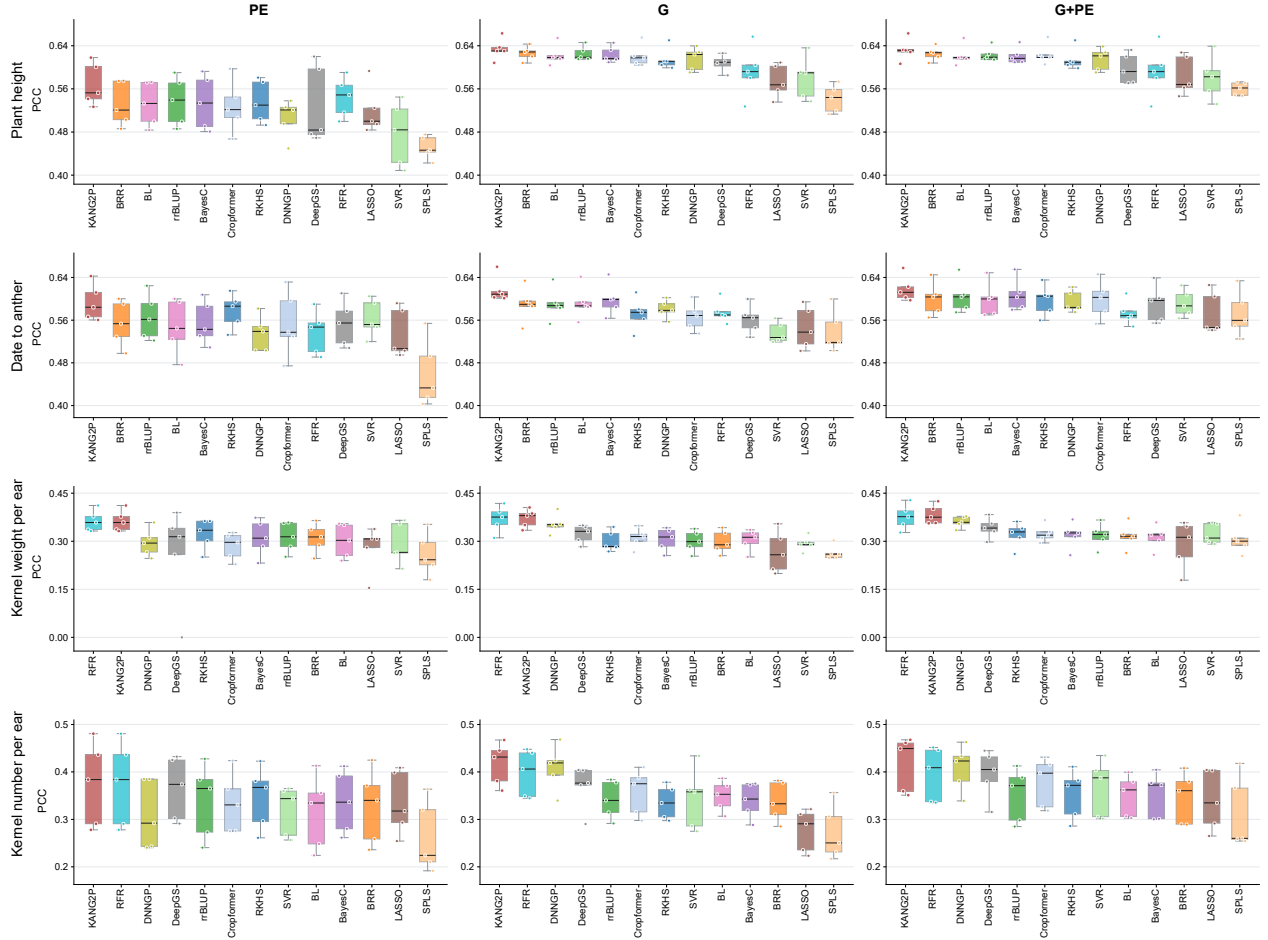
Supp. Fig. 6. Predictive performance of gene expression models in the Maize CUBIC reference population. a–c, Frequency distributions of model performance across gene expression prediction models in the CUBIC reference population, quantified by the coefficient of determination, R^2 (a), Spearman’s correlation coefficient, SCC (b), and mean squared error, MSE (c). Each bar represents the number of genes falling within the corresponding performance interval.



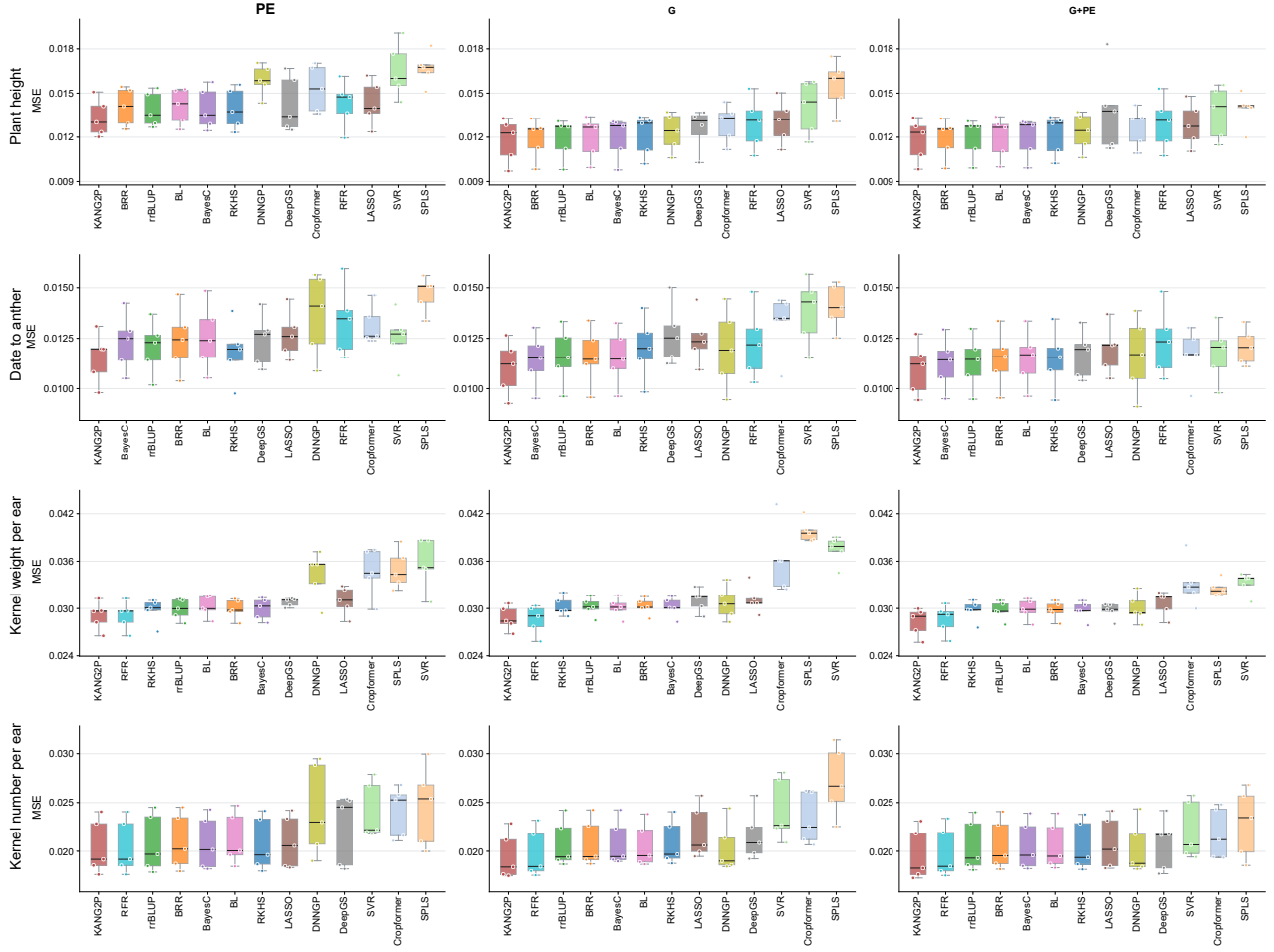
Supp. Fig. 7. Predictive performance comparison of 13 regression models across Rice1495 agronomic traits. Box plots show outer-fold PCC values from nested cross-validation for six regression models across four agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. RFR, Random Forest regressor. BL, Bayes LASSO. BRR, Bayes ridge regression. RKHS, Reproducing kernel Hilbert space. SPLS, sparse partial least squares. SVR, support vector machine regressor. RFR, Random Forest regressor.



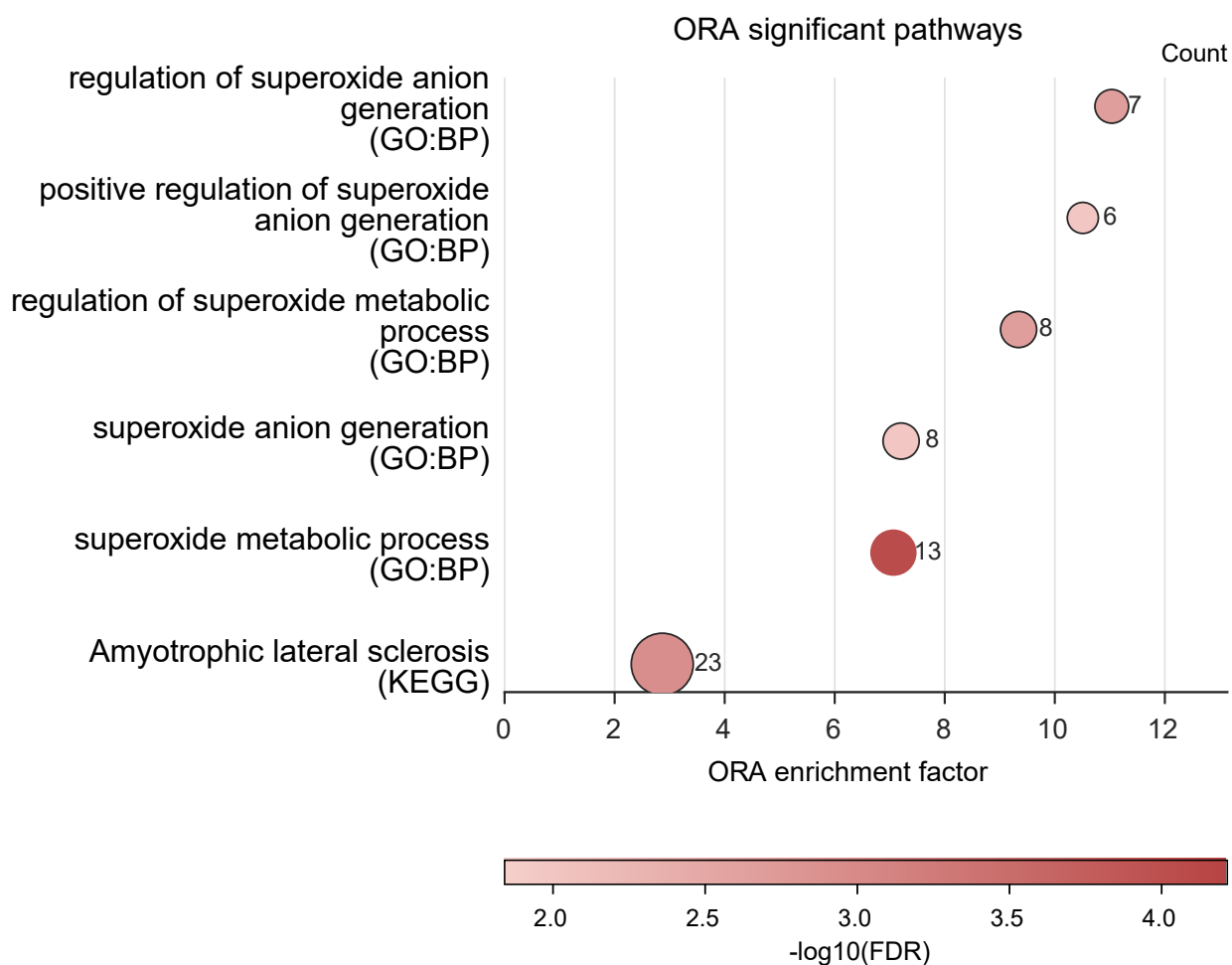
Supp. Fig. 8. Predictive performance comparison of 13 regression models across Rice1495 agronomic traits. Box plots show outer-fold MSE values from nested cross-validation for six regression models across four agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. RFR, Random Forest regressor. BL, Bayes LASSO. BRR, Bayes ridge regression. RKHS, Reproducing kernel Hilbert space. SPLS, sparse partial least squares. SVR, support vector machine regressor. RFR, Random Forest regressor.



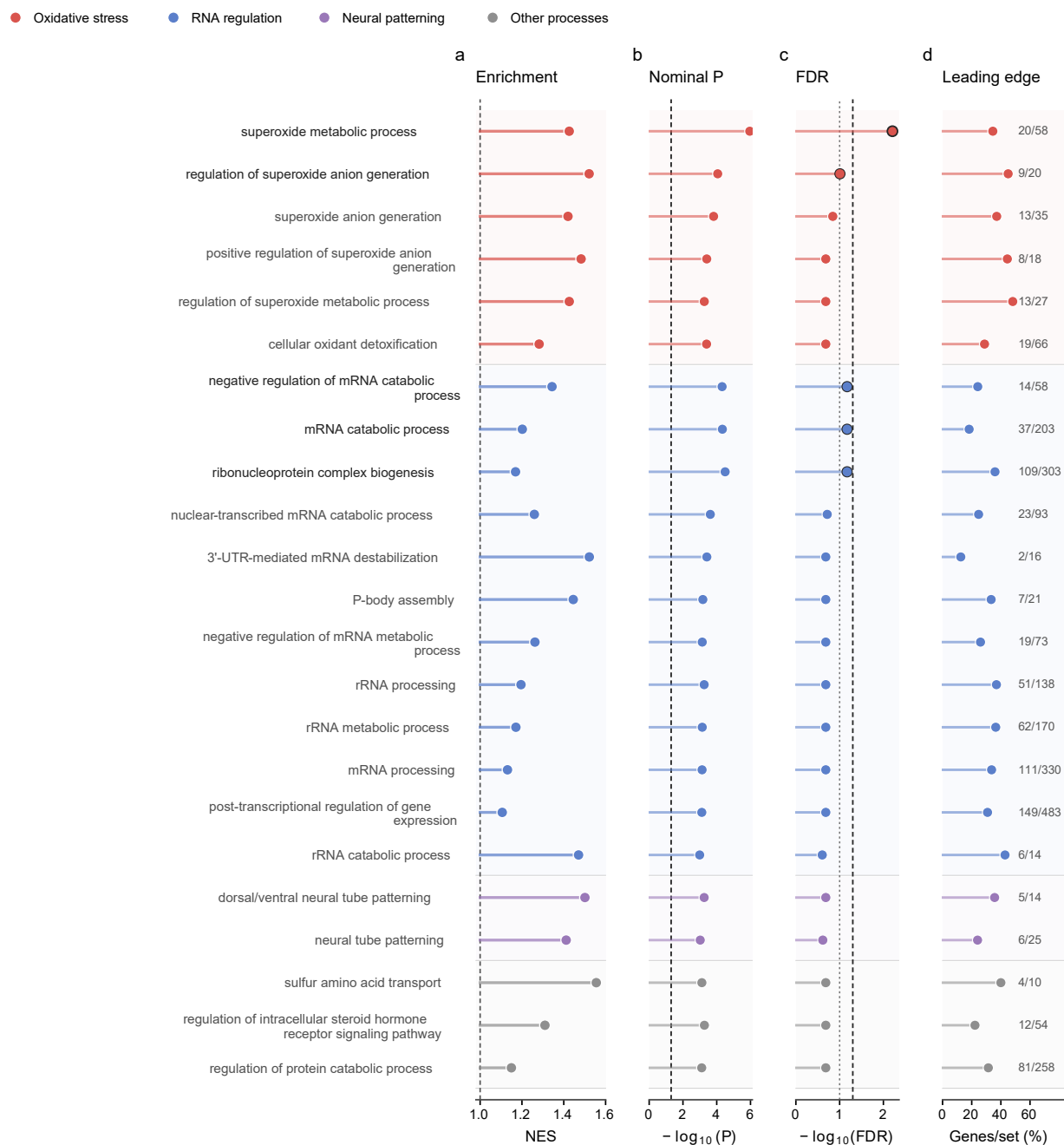
Supp. Fig. 9. Predictive performance comparison of 13 regression models across Maize1404 agronomic traits. Box plots show outer-fold PCC values from nested cross-validation for six regression models across four agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. RFR, Random Forest regressor. BL, Bayes LASSO. BRR, Bayes ridge regression. RKHS, Reproducing kernel Hilbert space. SPLS, sparse partial least squares. SVR, support vector machine regressor. RFR, Random Forest regressor.



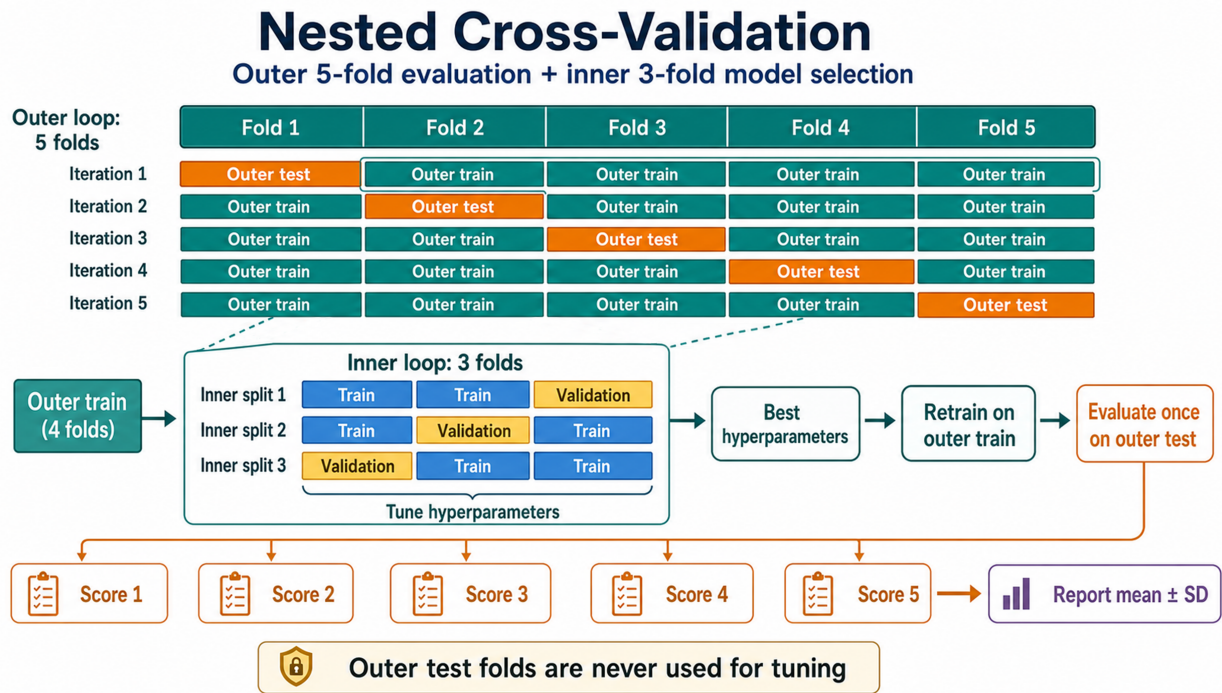
Supp. Fig. 10. Predictive performance comparison of 13 regression models across Maize1404 agronomic traits. Box plots show outer-fold MSE values from nested cross-validation for six regression models across four agronomic traits. Rows indicate traits and columns indicate input feature settings: genotype only (G), predicted expression profile only (PE) and the integrated combinations (G+PE). Each point represents one outer-fold estimate ($n = 5$). For each box, the centre line denotes the median, box bounds indicate the 25th and 75th percentiles, and whiskers extend to 1.5 times the interquartile range. RFR, Random Forest regressor. BL, Bayes LASSO. BRR, Bayes ridge regression. RKHS, Reproducing kernel Hilbert space. SPLS, sparse partial least squares. SVR, support vector machine regressor. RFR, Random Forest regressor.



Supp. Fig. 11. Over-representation analysis (ORA) results based on top 500 genes for pathways passing $\text{FDR} < 0.05$. Pathways are ordered by ORA enrichment factor, calculated from the observed gene ratio relative to the background ratio. Point size indicates the number of overlapping genes.



Supp. Fig. 12. Gene-set enrichment landscape of the ALS gene-priority ranking. Gene-set enrichment analysis (GSEA) results are shown for 23 positively enriched Gene Ontology biological-process terms. **a**, Normalized enrichment scores (NES), with positive values indicating enrichment towards the high-priority end of the ranked gene list. **b**, Nominal GSEA P values shown as $-\log_{10}(P)$. The black dashed line denotes the nominal significance threshold of $P = 0.05$. **c**, False discovery rates shown as $-\log_{10}(FDR)$. The black dashed and grey dotted lines denote $FDR = 0.05$ and $FDR = 0.10$, respectively; terms with $FDR < 0.10$ are outlined. **d**, Leading-edge genes expressed as a percentage of each GO gene set. Labels indicate the number of leading-edge genes and the total gene-set size. Pathways are grouped by biological theme and ordered within each group by increasing FDR and decreasing NES. Red, blue, purple and grey denote oxidative-stress, RNA-regulatory, neural-patterning and other biological processes, respectively. The superoxide metabolic process was the only term reaching $FDR < 0.05$, whereas RNA-regulatory and neural-patterning pathways showed suggestive or nominal enrichment.



Supp. Fig. 13. Nested cross-validation was used for model selection and performance estimation. The dataset was split into five outer folds; in each iteration, one fold was held out as the test set and the remaining four folds were used for model development. Within each outer-training set, threefold inner cross-validation was performed to tune hyperparameters and select the optimal model configuration. The selected model was then retrained on the full outer-training set and evaluated once on the corresponding held-out test fold. Final performance was summarized as the mean $\pm$ s.d. across the five outer-test scores, with the held-out test samples kept separate from all model-selection procedures.

Supplemental Tables

Tissue	No. genes	No. differentially expressed genes
Brain Substantia nigra	2422	294
Brain Spinal cord cervical c-1	3101	360
Brain Putamen basal ganglia	4224	482
Brain Nucleus accumbens basal ganglia	4619	559
Brain Hypothalamus	3459	392
Brain Hippocampus	3496	447
Brain Frontal Cortex BA9	4354	459
Brain Cortex	5252	585
Brain Cerebellum	6492	766
Brain Cerebellar Hemisphere	5481	651
Brain Caudate basal ganglia	4778	563
Brain Anterior cingulate cortex BA24	3358	412
Brain Amygdala	2661	327

Supplementary Table S1. The number of differentially expressed genes predicted from 13 brain tissues in the ALS training cohort (with a Wilcox test P value < 0.05).

Supplementary Table S2. Top five candidate drug-target pairs ranked by DeepDTAGen for each of the two ALS drugs, together with the corresponding AutoDock Vina results.

Drug	Rank	Target	UniProt	DTA score ^a	Best Vina score ^b (kcal mol ⁻¹)
Riluzole	1	<i>CALM3</i>	P0DP25	7.167	-4.858
Riluzole	2	<i>OR56A1</i>	Q8NGH5	6.920	-6.083
Riluzole	3	<i>GPR26</i>	Q8NDV2	6.900	-4.816
Riluzole	4	<i>SLC1A1</i>	P43005	6.853	-6.066
Riluzole	5	<i>SLC49A3</i>	Q6UXD7	6.803	-6.377
Edaravone	1	<i>OR56A1</i>	Q8NGH5	6.169	-6.032
Edaravone	2	<i>SLC1A1</i>	P43005	6.165	-6.440
Edaravone	3	<i>CALM3</i>	P0DP25	6.133	-5.066
Edaravone	4	<i>GDF7</i>	Q7Z4P5	6.122	-5.673
Edaravone	5	<i>SLC49A3</i>	Q6UXD7	6.078	-5.386

^a DeepDTAGen scores were generated using the BindingDB-pretrained model and were ranked in descending order separately for each drug. These values represent model outputs and should not be interpreted as experimentally measured binding affinities.

^b The best Vina score corresponds to the lowest score among the returned docking poses. More-negative values indicate more favourable scoring under the AutoDock Vina scoring function but do not represent experimentally measured binding energies. DrugBank accessions were DB00740 for riluzole and DB12243 for edaravone.

Strain	Genotype	Source
N2 Bristol	Wild-type	CGC
tm1060	<i>amph-1(tm1060)</i> IV	NBRP
tm4762	<i>anoh-1(tm4762)</i> III	NBRP
OS11091	<i>glt-1(ok206)</i> X	CGC
ZB1102	<i>glt-3(bz34)</i> IV	CGC
ZB1098	<i>glt-4(bz69)</i> II	CGC
ZB1099	<i>glt-5(bz70)</i> II	CGC
tm1316	<i>glt-6(tm1316)</i> IV	NBRP
tm1641	<i>glt-7(tm1641)</i> IV	NBRP
HA3299	<i>sod-1 rt451[sod-1(G85RC)]</i> II	CGC
VC913	<i>nex-3(gk385)</i> III	CGC
CL6049	<i>dvIs62 [snb-1p::hTDP-43/3' long UTR + mtl-2p::GFP]</i> X	CGC

Supplementary Table S3. *C. elegans* strains used in this study.