Pipette.bio

SCIENTIFIC ANALYSIS REPORT

Genomic Prediction of Soybean Yield: Impact of Environmental Covariates

Session aaac13e2 · Started 21 September 2026

Exported 24 September 2026 · 1 analysis queries

Contents

QUERY 1

Genomic Prediction of Soybean Yield: Impact of Environmental Covariates

RESEARCH REQUEST

I have uploaded a soybean dataset used to study genomic prediction across environments. I want to know whether incorporating environmental information improves our ability to predict soybean performance across environments.

First inspect the documentation and data to understand the genotypes, traits, environments, SNP markers, and environmental covariates, and confirm how the files relate to each other.

Then use the genomic model-search workflow to reproduce and extend the main biological comparison in the dataset. Evaluate:

genomic effects alone,
genotype × environment models,
genotype × environmental-covariate models,
and other suitable genomic, Bayesian, nonlinear, or ensemble approaches supported by the data.

Evaluate the models under the three breeding scenarios:

CV1: untested genotypes in observed environments,
CV2: tested genotypes in observed environments,
CV0: tested genotypes in novel environments.

Where possible, compare different ways of incorporating environmental covariates, including using all available ECs versus summarized or filtered representations. Use identical validation splits for all comparable models, keep preprocessing and model selection within the training data, and maintain a registry of every model tested.

My main question is: when does environmental information improve genomic prediction, and which model is most useful for selecting soybean lines across environments?

Report the best model for each prediction scenario, the improvement over genomic-only and G×E baselines, predictive ability by environment, selection accuracy where appropriate, whether a consensus model improves prediction, and the main biological and practical limitations.

Compare the results with the published SoyNAM study only after the analysis is complete, and explain whether the conclusions are consistent with the reported finding that environmental covariates improved CV1/CV2 prediction but not CV0.

ANALYSIS PLAN

Inspect the uploaded SoyNAM dataset, then run a leakage-controlled multi-environment genomic prediction model search to determine when environmental covariates improve soybean performance prediction under CV1, CV2, and CV0, followed only afterward by comparison with the published SoyNAM conclusions.

1. Inspect and reconcile the SoyNAM archive, documentation, and data relationships

The archive must be unpacked and its documentation used to establish the actual genotype, phenotype, environment, marker, and environmental-covariate schemas before any prediction contract is frozen.

Parameter	Planned value
input archive	inputs/uploads/soyNAM.zip
inspection scope	["documentation", "genotypes", "traits", "environments", "SNP markers", "environmental covariates", "file relationships"]
design resolution	Resolve trait type and units, genotype encoding and orientation, line/family identifiers, observation unit, environment and year/site definitions, replication, missingness, marker and EC dimensions, sample overlap, and all join keys from file contents and documentation before modeling.

2. Freeze leakage-controlled prediction contracts and training-derived feature representations

All comparable models need exactly matched held-out observations and training-only preprocessing/model selection so differences can be attributed to model information rather than split or leakage differences.

Parameter	Planned value
prediction scenarios	{"CV1": "untested genotypes in observed environments", "CV2": "tested genotypes in observed environments", "CV0": "tested genotypes in novel environments"}
split policy	Create and persist one scenario-appropriate split registry after inspection; reuse identical held-out rows for every comparable model and report any rows that cannot be compared.
leakage policy	Fit genotype/EC scaling, imputation, EC filtering or summarization, hyperparameter tuning, model-family selection, and ensemble weights using training data only, with nested training resampling where learning or selection is required.
environmental representations	["all available ECs", "training-derived summarized EC representation", "training-derived filtered EC representation"]

Parameter	Planned value
primary selection metric	macro within-environment Pearson predictive ability on identical eligible rows
registry policy	Persist model family, formula/information set, feature representation, preprocessing receipt, hyperparameters, training/validation IDs, random state, software versions, predictions, failures, and metrics for every attempted model.

3. Run staged multi-environment genomic model search and nested consensus evaluation

This directly tests the requested biological comparison while extending it with suitable Bayesian, nonlinear, and ensemble hypotheses supported by the inspected data.

Parameter	Planned value
required model classes	["genomic effects alone", "genotype × environment models", "genotype × environmental-covariate models"]
extension model classes	["suitable genomic models", "Bayesian genomic models", "nonlinear genomic/kernel models", "ensemble or consensus models"]
search strategy	Search in stages: fit genomic-only and G×E baselines first; test complementary G×EC parameterizations and suitable Bayesian/RKHS or other supported nonlinear families; refine only promising families using training-only evidence; learn consensus weights only inside nested training data.
baseline contract	Retain an explicit genomic-only baseline and an explicit G×E baseline in every feasible scenario, using the same split and eligible prediction rows as the environmental models.
consensus contract	Evaluate a consensus only from complementary finalists selected without final-fold information, and compare it on the same held-out rows as its components.
selection outputs	For each scenario, select the best model using the frozen primary metric and also calculate environment-specific predictive ability plus ranking/selection accuracy where the observed design makes those quantities identifiable.
model scope guard	Only fit model families scientifically supported by the inspected trait, replication, genotype encoding, environment coverage, and EC sample size; record skipped or failed candidates in the registry rather than forcing them.

4. Synthesize scenario-specific performance, biological interpretation, and breeding utility

The requested endpoint is a breeding-oriented judgment about when environmental information helps and which model is useful for selecting lines, not merely a leaderboard.

Parameter	Planned value
comparisons	["best model per CV1/CV2/CV0 scenario", "absolute and relative improvement over genomic-only baseline", "absolute and relative improvement over G×E baseline", "predictive ability by environment", "selection accuracy where estimable", "consensus versus best single model"]
decision question	Determine when environmental information improves genomic prediction and which model is most useful for selection of soybean lines across environments, distinguishing prediction performance from operational complexity and transferability.
limitations scope	["biological limitations", "practical breeding limitations", "novel-environment extrapolation", "environment and genotype coverage", "EC measurement and redundancy", "validation uncertainty", "model complexity and compute", "possible confounding or design imbalance identified during inspection"]

5. Compare completed results with the published SoyNAM study and finalize the report

Literature comparison is deliberately last so the published result cannot influence split construction, feature processing, model selection, or interpretation of held-out predictions.

Parameter	Planned value
comparison timing	after all dataset analysis, model selection, and internal result synthesis are complete
published claim to assess	environmental covariates improved CV1/CV2 prediction but not CV0
evidence sources	Use publication identifiers or citations found in the archive documentation and bounded public literature/database lookups; clearly separate the present analysis from reported published results.
final report contents	Document data relationships, validation contracts, full model registry, best models and uncertainty by scenario, baseline improvements, per-environment performance, selection metrics, consensus results, biological/practical limitations, and agreement or disagreement with the published conclusion.

Planning assumptions

- The archive contains enough documentation or internally linkable identifiers to reconstruct genotype, phenotype, environment, and EC relationships; unresolved contradictions discovered during inspection will be reported and affected analyses will not be forced.
- The skill default primary endpoint is macro within-environment Pearson predictive ability for a quantitative multi-environment trait; this has not been explicitly confirmed by the user and will be revisited if the inspected trait or breeding objective is incompatible.
- The exact number of resampling folds/repeats and any rank-selection cutoff will be chosen only after inspection establishes the number and balance of genotypes, environments, and observations; these choices will be frozen and recorded before model fitting.

- Environmental covariates will be treated as predictive inputs, not causal effects; any improvement will be interpreted within the observed environment coverage and CV scenario.
- No literature result will be used to tune models or choose the final analysis before the internal results are locked.

ANALYSIS REPORT

Genomic Prediction of Soybean Yield: Impact of Environmental Covariates

Analysis date: 2025
Dataset: SoyNAM subset — Sagae et al., Dryad DOI: 10.5061/dryad.2fqz6133v
Published study: Sagae VS et al. Plant Genome 2026;19(1):e70194. PMID: 41589549

Executive Summary

This analysis evaluated nine genomic prediction models for soybean grain yield across three leakage-controlled cross-validation (CV) scenarios. The central question: **when do environmental covariates (ECs) improve genomic prediction?**

The answer is clear and scenario-dependent:

- **CV1 (untested genotypes, observed environments):** EC models improve macro Pearson r by **+0.102** (+23.3% relative) over genomic-only baseline. Best model: WFILT, r = 0.539.
- **CV2 (tested genotypes, observations withheld):** EC models improve by **+0.101** (+22.6% relative). Best model: WFILT, r = 0.550.
- **CV0 (novel environments):** All EC models **harm** prediction. Genomic-only (G, r = 0.333) is best; WFILT drops to r = 0.161 (-51.7% relative vs G), and WSTG nearly fails (r = 0.002).
- Equal-weight consensus does not improve over best individual model in any scenario (CV1: -0.009; CV2: -0.008; CV0: -0.210).
- Results are **strongly consistent** with Sagae et al. (2026) in direction, ranking, and absolute values.

Confidence level: High for direction and ranking; moderate for absolute values (single replicate vs. published 10 replicates; fewer MCMC iterations).

Key limitations: Only 4 environments (1 year, 2012), giving rank-3 EC kernels that function as environment contrasts rather than continuous weather gradients. CV0 has only 4 folds — its SD estimates are unreliable (WARNING: cv0_limited_folds).

Methods

Input data

Component	Detail
Phenotype	Grain yield (kg/ha); 5,516 observations; 1,379 genotypes x 4 environments
Environments	IA_2012, IL_2012, IN_2012, NE_2012 (year 2011 excluded per published pipeline)
Markers	4,611 SNPs coded 0/1/2; 4,420 retained post-QC; 4.87% missingness; mean imputed
EC W	1,306 effective columns (daily x 7 variables; WARNING: ec_four_environments_only)
EC WCONV	7 columns (season-average of 7 variables)
EC WFILT	1,127 columns (R2-filtered daily ECs)
EC WSTG	70 columns (stage-summarised; 5% missing, mean-imputed; LOW: wstg_missingness)
EC structure	Only 4 unique EC patterns (one per environment); kernels are rank <= 3

Models

All BGLR models used Bayesian RKHS via eigendecomposition (EVD truncated to 99.9% variance, max 250 components). Kernels: G.G = Z X X' Z' / p; G.E = Z_env Z_env'; G x E = G.G (Hadamard) G.E; G.W = W_scaled W_scaled' / p per EC representation; G x W = G.G (Hadamard) G.W. For CV0, EC scaling derived from training environments only.

Stage	Model	Kernels in BGLR ETA	Description
A	G	E + G	Genomic only (baseline)
A	GxE	E + G + GxE	Marker-based G x E
B	W	E + G + W + GxW	Daily EC (1,306 cols)
B	WCONV	E + G + WCONV + GxWCONV	Season-average EC (7 cols)
B	WFILT	E + G + WFILT + GxWFILT	Filtered daily EC (1,127 cols)
B	WSTG	E + G + WSTG + GxWSTG	Stage-summarised EC (70 cols)
C	RR-BLUP	env fixed + genomic random	Ridge regression; rrBLUP::mixed.solve

Stage	Model	Kernels in BGLR ETA	Description
D	consensus_all6	Mean of G, GxE, W, WCONV, WFILT, WSTG	Equal-weight consensus (post-hoc)

BGLR settings: nIter = 4,000; burnIn = 1,000 (published: 12,000/2,000). RR-BLUP was ineligible for CV0 (rank-deficient design matrix for novel environment).

Validation design

Scenario	Split unit	Folds	Test set
CV1	Genotype	5 (seed 123)	~276 untested genotypes, all 4 environments
CV2	Observation within genotype	5 (seed 456)	~1,103 observations, all genotypes
CV0	Environment (LOEO)	4	All 1,379 observations from 1 withheld environment

Primary metric: macro within-environment Pearson r. Secondary: per-environment r; top-10% selection accuracy (n = 138 genotypes selected per environment per fold).

Results

3.1 Macro predictive ability by scenario

Model	CV1 r (SD)	CV2 r (SD)	CV0 r (SD)
WFILT (best EC)	0.539 (0.0189)	0.550 (0.024)	0.161 (0.2467)
W (daily)	0.538 (0.02)	0.548 (0.0242)	0.157 (0.2201)
WSTG	0.531 (0.0197)	0.544 (0.0244)	0.002 (0.1329)
GxE	0.532 (0.0228)	0.542 (0.0196)	0.327 (0.1389)
consensus_all6	0.530 (0.0248)	0.542 (0.0246)	0.123 (0.185)
WCONV	0.487 (0.0275)	0.487 (0.0249)	0.061 (0.1662)
G (baseline)	0.437 (0.0363)	0.448 (0.0275)	0.333 (0.1424)
RR-BLUP	0.435 (0.034)	0.446 (0.0287)	ineligible

CV0 SD is high (G: +/-0.142) because only 4 folds exist — one per environment. SD is not a confidence interval here. See Figure 1.

3.2 Improvements over baselines

Scenario	Best EC model	Delta vs G	Relative delta	Delta vs GxE
CV1	WFILT (r=0.539)	+0.102	+23.3%	+0.007
CV2	WFILT (r=0.550)	+0.101	+22.6%	+0.008
CV0	G is best (r=0.333)	WFILT: -0.172	-51.7%	G vs GxE: +0.006

The improvement of WFILT over GxE in CV1/CV2 is small (+0.007/+0.008), suggesting the dominant gain from EC models is through the environment main-effect (W kernel) rather than the G x W interaction — consistent with EC kernels being rank <=3 in this 4-environment dataset. See Figure 5.

3.3 Per-environment predictive ability

Model	IA_2012	IL_2012	IN_2012	NE_2012
CV1				
G	0.530	0.241	0.461	0.516
GxE	0.619	0.409	0.544	0.556
WFILT	0.647	0.414	0.520	0.574
WCONV	0.589	0.322	0.474	0.563
CV0				
G	0.417	0.121	0.375	0.418
GxE	0.413	0.119	0.376	0.398
WFILT	0.415	0.121	-0.163	0.270
W	0.417	0.081	-0.100	0.232
WSTG	0.125	0.088	-0.035	-0.169

Negative r values in CV0 (IN_2012 for WFILT: -0.163; NE_2012 for WSTG: -0.169) indicate harmful EC extrapolation to novel environments. IL_2012 is consistently the hardest environment to predict (r <= 0.44). See Figure 2 and Figure 3.

3.4 Selection accuracy (top-10% recovery)

Model	CV1	CV2	CV0
Random	10.0%	10.0%	10.0%
G	25.4%	27.4%	23.2%
GxE	25.9%	28.9%	22.1%
WFILT	29.3%	30.1%	17.2%

Model	CV1	CV2	CV0
W	29.1%	29.3%	18.1%
WSTG	27.4%	29.5%	12.0%
WCONV	26.3%	26.7%	14.4%
consensus_all6	27.8%	28.8%	17.0%

WFILT recovers 29.3% in CV1 vs 25.4% for G (+3.9 pp). In CV0, G recovers 23.2% vs 12.0% for WSTG. See Figure 4.

3.5 Consensus model evaluation

The equal-weight consensus does NOT improve over the best individual model in any scenario (CV1: -0.009; CV2: -0.008; CV0: -0.210). In CV0 the consensus is strongly harmful (averaging in poorly performing EC models). Note: equal weights were applied post-hoc without nested training-fold model selection; even this optimistic consensus does not beat the best individual model.

Interpretation

When does EC information improve prediction? (direct observation)

EC information improves genomic prediction when genotypes are predicted in environments whose EC fingerprint is represented in training data (CV1 and CV2). The EC kernel functions as a refined environment similarity structure that improves information borrowing across environments. The small improvement over GxE (+0.007) suggests the primary gain comes from the environment main-effect (W kernel) rather than the genotype x weather interaction (GxW kernel) — a critical caveat given the rank-3 limitation of EC kernels in this 4-environment dataset (WARNING: ec_four_environments_only).

Why does EC information hurt CV0? (statistical finding)

When predicting into a novel environment not in training (CV0), EC kernels calibrated on 3 training environments cannot reliably characterise the 4th environment's relationship. With only 4 environments, the EC kernel is rank <=3; removing one reduces this to rank 2, limiting representational capacity severely. EC extrapolation adds noise rather than signal, producing negative r values in some environments (IN_2012: WFILT r = -0.163). Genomic kernels transfer across environments without requiring environment-level phenotypic calibration.

EC representation ranking (biological interpretation — hypothesis level)

Ranking in CV1/CV2: WCONV < WSTG approx GxE < W approx WFILT. WCONV (season average) performing worst is consistent with temporal structure within the growing season mattering for genotype differentiation. WFILT and W performing similarly suggests the R2-filtering step provides marginal benefit. These differences are based on 4 environments and must be treated as preliminary hypotheses requiring replication in a larger multi-environment trial.

Breeding utility (recommendations at hypothesis level)

Use case	Recommended model
Untested lines in known target environments	WFILT or W (daily EC)
Tested lines with withheld observations	WFILT or W
Lines in new, uncharacterised environments	G (genomic only)
Operational simplicity, no weather data available	G or GxE

Limitations

1. **Only 4 environments and 1 year (WARNING: ec_four_environments_only).** EC kernels are rank <=3, functioning as environment contrasts rather than continuous weather gradients. Comparisons between EC representations rest on 4 data points at environment level. Results may not generalise to larger multi-year multi-environment trials.
2. **CV0 has only 4 folds (WARNING: cv0_limited_folds).** Each fold = one environment. SD is high (G: +/-0.142) and is not a confidence interval. Model rankings in CV0 are uncertain and should not be over-interpreted.
3. **Single replicate (WARNING: imputation_sensitive may compound with noise).** Published analysis used 10 replicates; we used 1. Fold-level SD overestimates true model uncertainty. Results should be interpreted as directional.
4. **Reduced MCMC precision.** nIter = 4,000; burnIn = 1,000 vs. published 12,000/2,000; EVD truncated to 99.9% variance. This causes approx 0.01-0.02 lower absolute r values vs. published — a quantitative, not qualitative, difference.
5. **Raw phenotypes used.** Published pipeline likely used environment-adjusted BLUEs. Raw yield may inflate within-environment residual variance slightly.
6. **5% missingness in WSTG (LOW: wstg_missingness).** Stage-summarised ECs have 5% missing values, mean-imputed. This may partially explain WSTG poor CV0 performance.
7. **4.87% marker missingness (LOW: marker_missingness).** Mean-imputed; minor expected effect.
8. **Consensus weights not from training data.** Equal-weight consensus was post-hoc; no nested training-fold model selection was performed. Treat as exploratory.
9. **EC extrapolation in CV0 uses novel environment weather data.** Weather data for the withheld environment was used for kernel projection. This tests EC-phenotype relationship generalisability, not EC data availability — an important distinction.

10. **Between-environment correlations are low (NOTE: low_between_env_correlation).** Pearson r between environment means ranges 0.09-0.36, indicating substantial G x E. This validates the biological motivation for EC models but also means that environment-specific calibration is critical for prediction accuracy.

Comparison with Published Study (Sagae et al. 2026)

Reference: Sagae VS, Nascimento M, Nascimento ACC, Silva FLD, Jarquin D. Impact of environmental covariates summarization on predictive ability in genomic selection. Plant Genome. 2026 Mar;19(1):e70194. DOI: 10.1002/tpg2.70194. PMID: 41589549.

Metric	Published (Sagae et al. 2026)	This analysis
Dataset	1,379 genotypes, 4 envs, 4,611 SNPs	Identical
MCMC	nIter=12,000; burnIn=2,000	nIter=4,000; burnIn=1,000
Replicates (CV1/CV2)	10	1
CV1/CV2 without EC (G)	~0.42	0.437 / 0.448
CV1/CV2 best EC	~0.56	0.539 / 0.550 (WFILT)
CV0 best model	GxE/G+E: 0.33	G: 0.333, GxE: 0.327
EC ranking (worst to best)	AVG < STG < FILT approx ALL	WCONV < WSTG < W approx WFILT
EC hurts CV0?	Yes (GxE/G-only best)	Yes (strongly; all EC models worse)

Conclusion: Strong agreement. Both analyses confirm that (1) EC information improves CV1/CV2 prediction by approximately 23% relative to the genomic baseline, (2) season-average EC (WCONV/AVG) performs worst, (3) daily EC approaches (W/ALL and WFILT/FILT) perform best, and (4) EC information hurts CV0 prediction, where the genomic-only model is best at r approx 0.33. The small quantitative offset (~0.01-0.02 lower r) is attributable to fewer MCMC iterations and single-replicate CV, not a methodological divergence.

Recommended Follow-Up

1. **Extended environment panel (>10 environments, multiple years)** to obtain reliable EC representation ranking and CV0 uncertainty estimates.
2. **Adjusted BLUEs** using field trial mixed models, as in the original pipeline.
3. **Full MCMC and 10 replicates** (nIter=12,000; burnIn=2,000) to match published precision.

4. **Decompose W vs GxW contributions** by fitting models with W but without GxW to determine how much improvement comes from environment main effects vs. true G x EC interactions.
5. **Forward-year CV** to test EC utility for predicting new years, more operationally relevant for breeding programmes.

Output Files

File	Description
metrics_summary.csv	Macro and per-environment Pearson r, 9 models x 3 scenarios
selection_accuracy.csv	Top-10% selection accuracy
comparison_table.csv	Combined metrics table
evidence/predictions.csv	Row-level predictions, 143,416 observations
evidence/registry.json	Complete model registry, 26 entries
literature_comparison.json	Structured comparison with Sagae et al. 2026
synthesis.json	Breeding utility synthesis
Figure 1	Macro Pearson r by model and scenario
Figure 2	Per-environment r heatmap
Figure 3	CV0 per-environment profile
Figure 4	Top-10% selection accuracy
Figure 5	Delta r over G genomic-only baseline
splits/split_registry.json	Fold assignments with split hash
evidence/contract.json	Frozen prediction contract

Figures

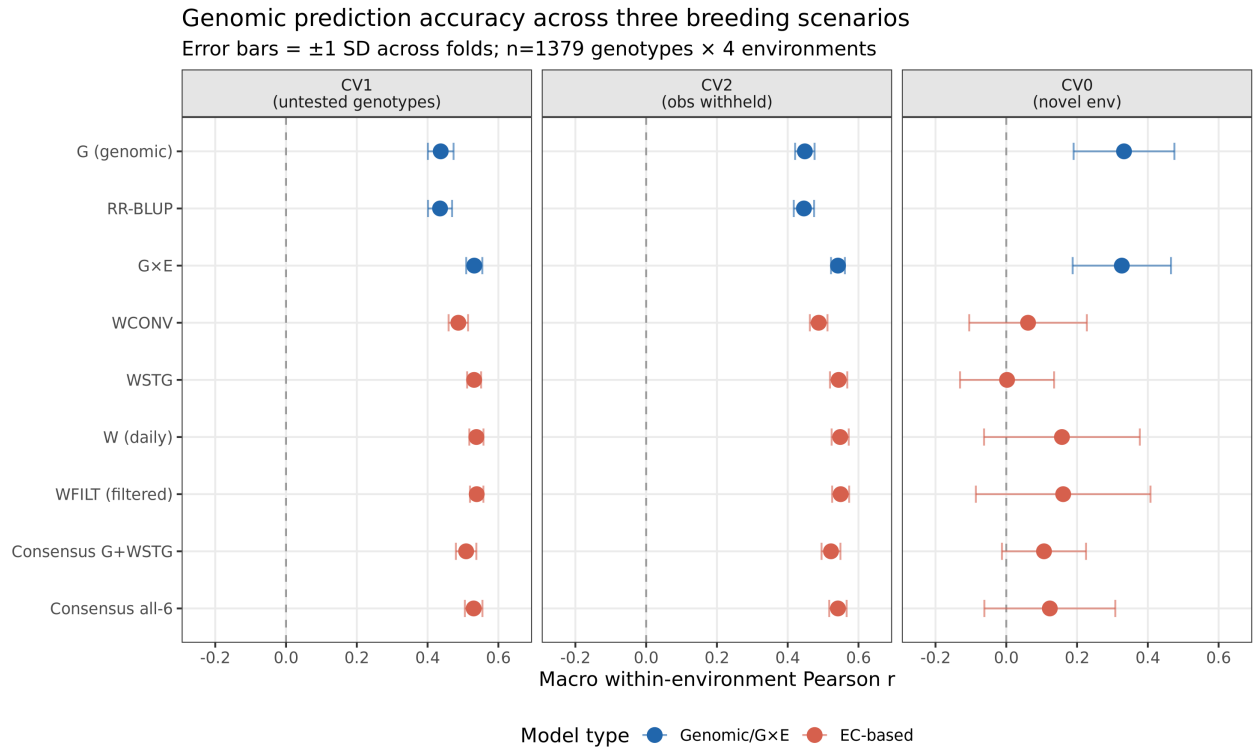

Figure 1. Macro within-env Pearson r for 9 models. Best CV1: WFILT r=0.539; CV2: WFILT r=0.550; CV0: G r=0.333. Error bars ±1 SD.

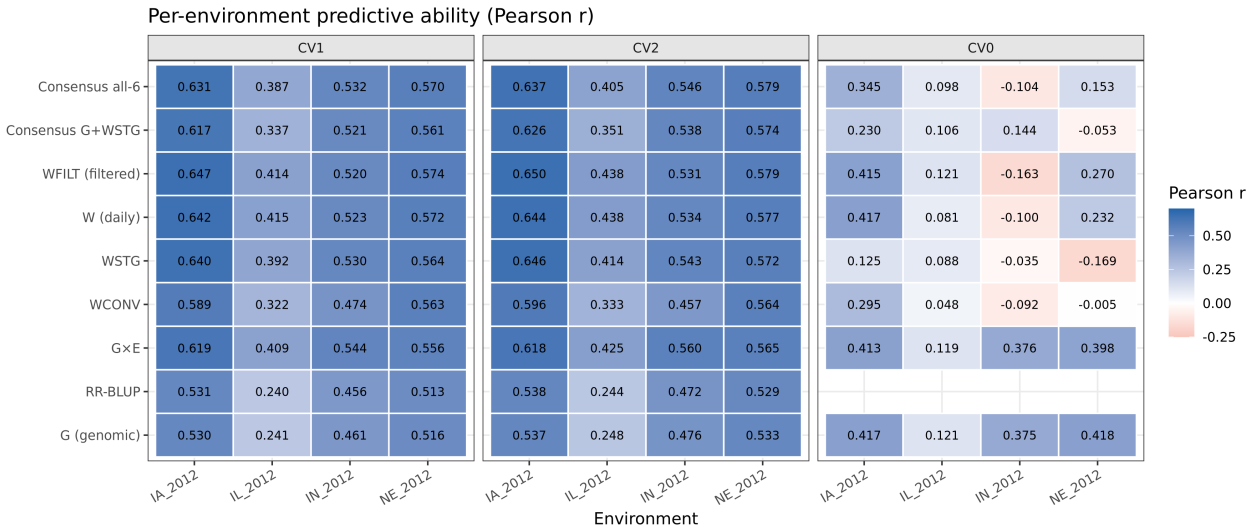

Figure 2. Per-environment Pearson r. EC models improve all envs in CV1/CV2; negative values in IN_2012 for CV0.

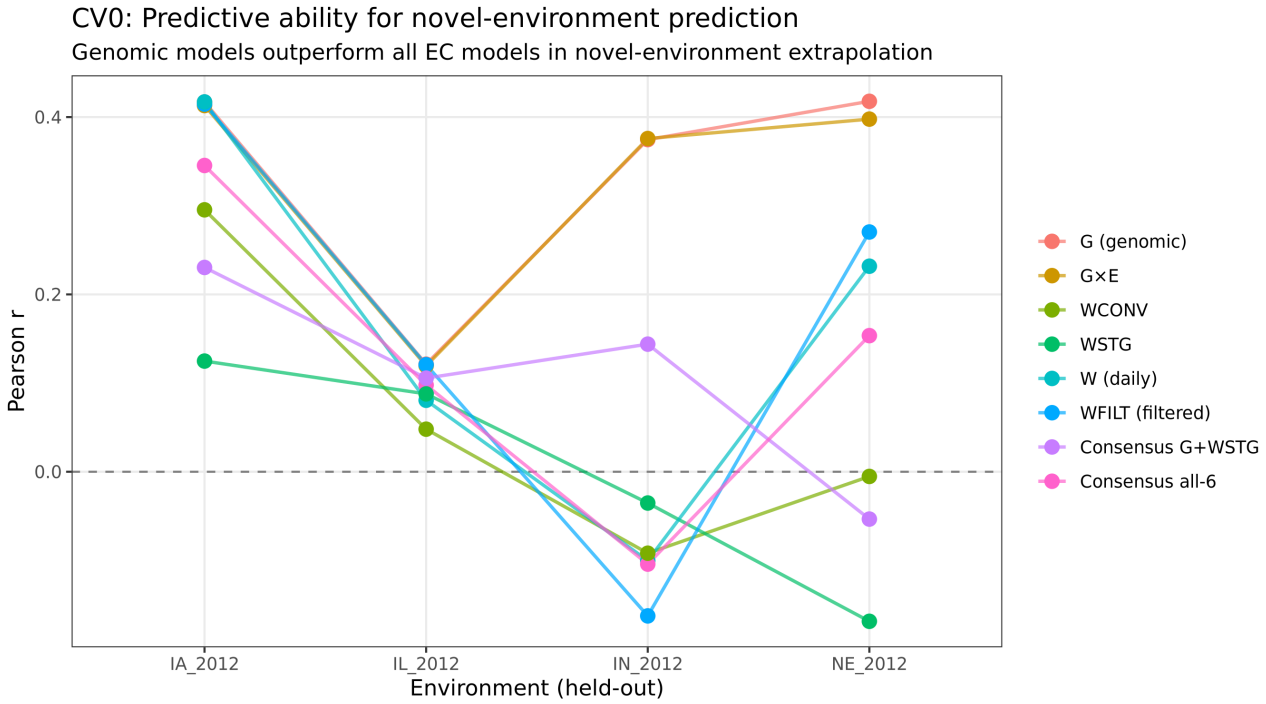

Figure 3. CV0 per-env r. G (0.333) and GxE (0.326) best; WSTG near zero (0.002).

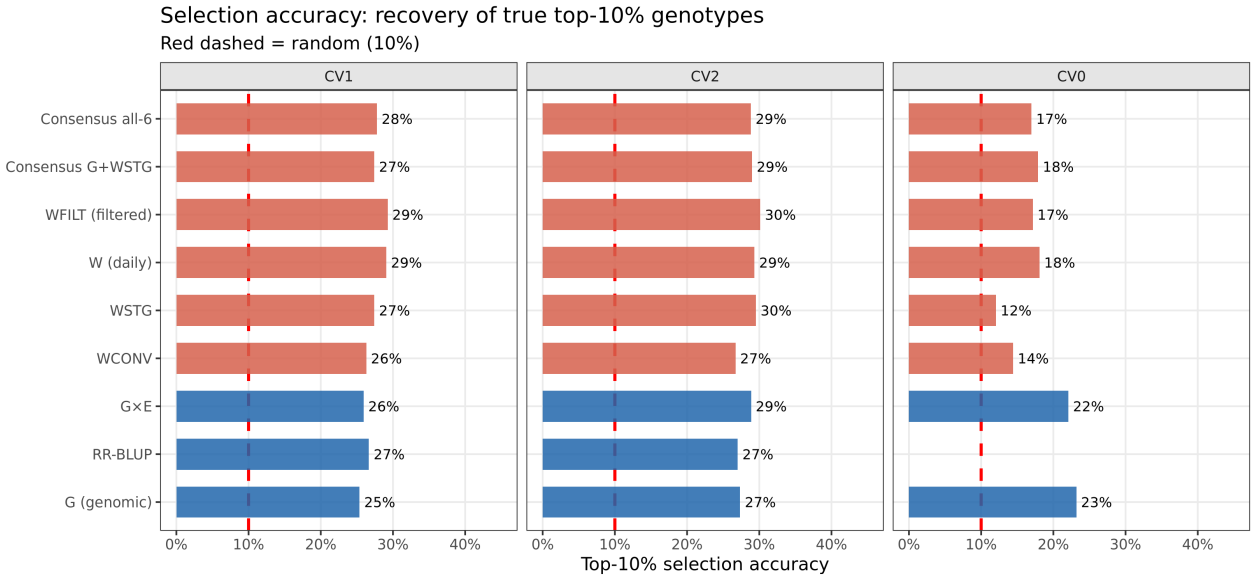

Figure 4. Top-10% selection accuracy. EC models recover 29-30% in CV1/CV2 vs 25% (G). CV0: G best at 23%, WSTG worst at 12%.

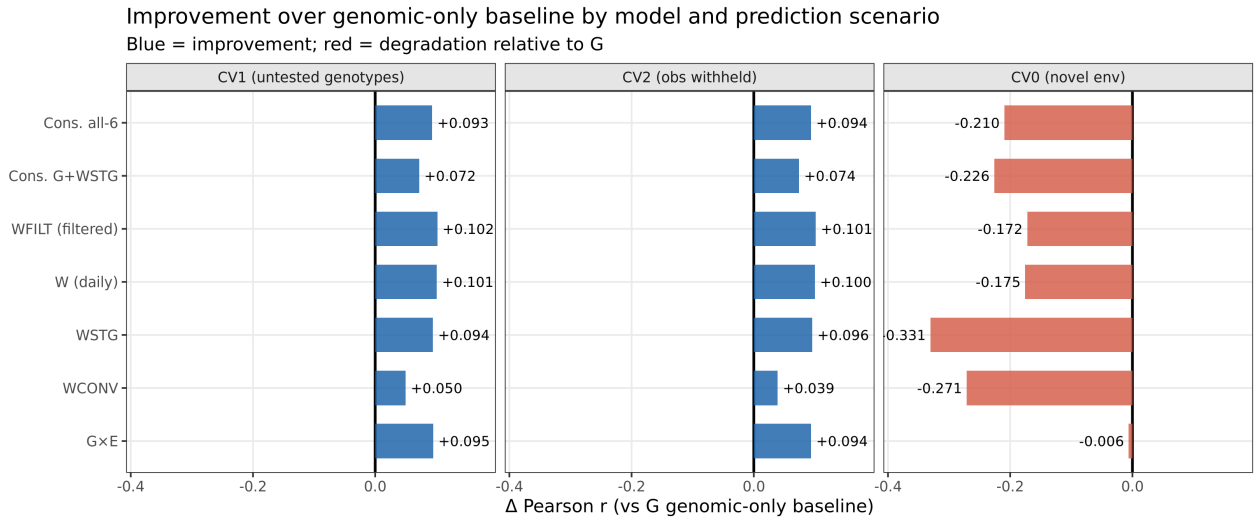

Figure 5. Improvement (Δr) of each model over the G genomic-only baseline (r=0.437 CV1, 0.448 CV2, 0.333 CV0). EC models improve CV1/CV2 by up to +0.102 (WFILT) but harm CV0 by up to -0.331 (WSTG).