Different approaches to genomic prediction model validation in soybean

Vuk Đorđević, Marina Ćeran, Jegor Miladinović, Svetlana Balešević-Tubić, Kristina Petrović, Predrag Ranđelović, Jelena Marinković
Institute of filed and vegetable crops, Novi Sad, Serbia

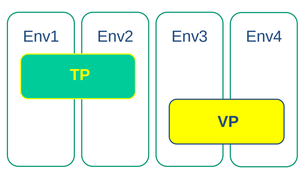
Introduction

When dealing with the improvement of quantitative inherited traits, such as yield, strategy to simultaneously use genome-wide molecular markers that able to capture all small effect loci influencing a trait, was promising strategy. The successful implementation of genomic prediction in the process of soybean breeding is determined by the ability of the developed model to predict or estimate the genetic potential of new breeding lines for a specific trait
Cross-validation is a commonly performed validation procedure. However, in cross-validation, both the training and validation sets are tested under the same environmental conditions which is not realistic scenario in applied breeding and can lead to an overestimation of the model performance. Therefore, besides evaluation of the effect of varying factors influencing prediction model performance, we shall examine the properties of the genomic prediction model in the external validation, analyzing the power of prediction when the genotypes of the validation set are not part of the training population, and when validation population is tested under different environmental conditions, simulating real breeding process.

Methods

The training population consisted of 227 diverse soybean lines that were used for genomic prediction model development. Training population was evaluated for yield at three consecutive years and SNP data obtained by Genotyping-by-sequencing protocol. Prediction ability was evaluated using six mathematical models, including parametric and non-parametric and were validated on two different levels: cross-validation (5-fold) and external validation (historical data).

Validation procedure



TP - training population model development

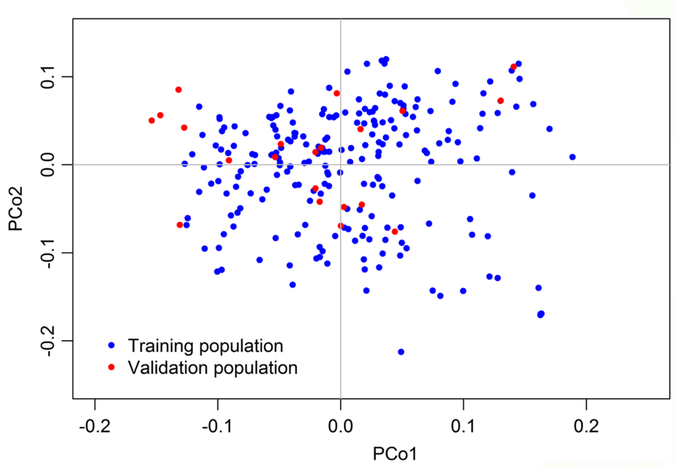
VP - external validation unknown genotypes in different environments

Mathematical models

Type	Model
Parametric models	RR-BLUP
	BayesA
	BayesB
	BayesCPi
Non-parametric models	Random Forest
	Support Vector Machine

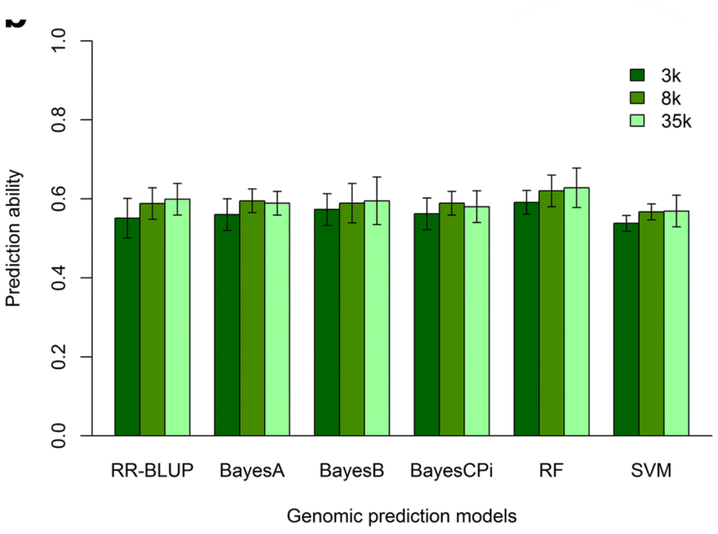
Results

Overall, genomic prediction ability for soybean yield was relatively high (0.60) and the results indicate a modest influence of mathematical model and marker number on the prediction ability using cross-validation and external validation. However, model had variable ability to predict phenotypic performance in separate environments, with especially high prediction ability in years not impacted by yield-limiting factors, when the genetic potential was fully achieved. Improvement of model performance in cross-validation and external validation was achieved by increasing the phenotyping intensity that must reflect the target environment variability.

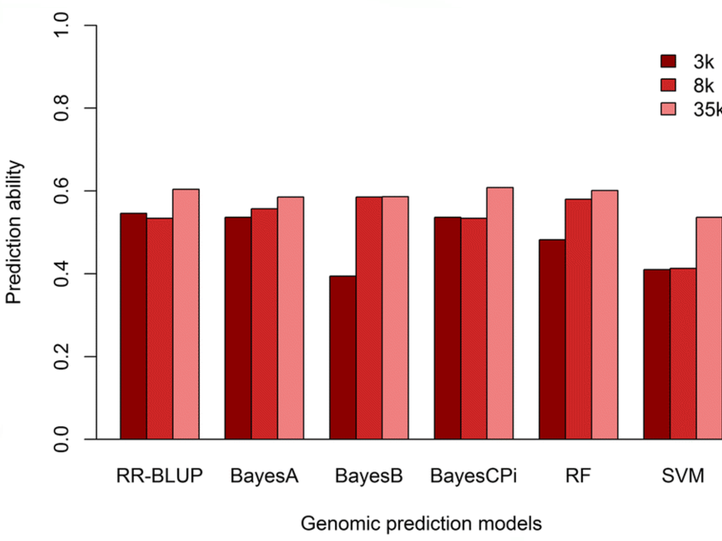


Population structure

Effect of SNP data set mathematical models on genomic prediction ability

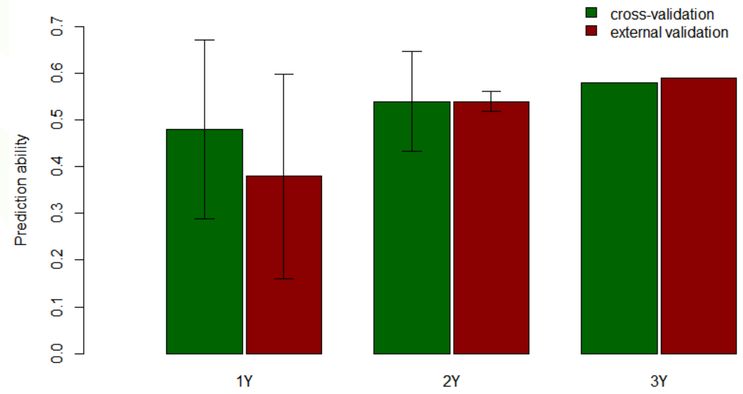


Cross validation



External validation

TP phenotypic data intensity (bars represents range of variation)



Single environment had low prediction ability and high prediction variability

Obtained results indicate that genomic prediction can be integrating part of breeding process as useful tool that can increase breeding efficiency and decrees breeding time. Particular implementations are diverse, from germplasm screening and parental choice to the forward breeding and direct section based on genomic prediction.

Vuk Djordjevic vuk.djordjevic@ifvcns.ns.ac.rs