Haplotype blocks for genomic prediction: A comparative evaluation in multiple crop datasets

Sven Weber, Rod Snowdon

Department of Plant Breeding, Justus Liebig University Giessen, Germany

Breeding value pattern recognition

Introduction

Genomic selection has greatly improved plant and animal breeding. However, high-throughput SNP genotyping tools frequently have high redundancy caused by markers with strong linkage disequilibrium (LD). Because of LD, genomic information tends to be inherited in a block-like structure. Recently, interest has grown in use of haplotype blocks for genomic selection, with the additional advantage that haplotype blocks theoretically capture local epistatic patterns and are in higher LD with QTL. Here we test the usefulness of haplotype blocks considering the four commonly used genomic prediction models GBLUP, Bayessian LASSO, EGBLUP and RKHS regression. Blocks were constructed based on LD, adjacency, physical distance and „Haploview". The methods were tested in four major crops: Canola, maize, wheat and soybean.

Material and methods

Here, we test and compare haplotype blocks based on LD, physical distance, number of adjacent markers and „Haploview" (Gabriel et al., Four Gamete Rule and Solid Spine of LD) along a wide parameter space with regard to their predictive ability for agronomically important traits. As baseline, standard SNPs were used as predictors. Predictive ability was assessed with three commonly used parametric genomic prediction models (Genomic Best Linear Unbiased Prediction [GBLUP], Extended GBLUP [EGBLUP] and Bayesian LASSO) and the semiparametric Reproducing Kernel Hilbert Space Regression (RKHS). We utilized data from four published datasets: Canola (Jan et al. 2016), maize (Lehermeier et al 2014), wheat (Voss-Fels et al. 2019) and soybean (Bandillo et al. 2015). The formula describing the relationship is between haplotypes and phenotype is:

$$y = X\beta + Ma + e$$

where y is a n*1 vector of adjusted entry means, X is the design matrix for the fixed non-genetic effects ß, while M is a matrix of haplotypes/SNPs relating the random haplotype/SNP effects a to the individuals and e is the random error term.

Results and Discussion

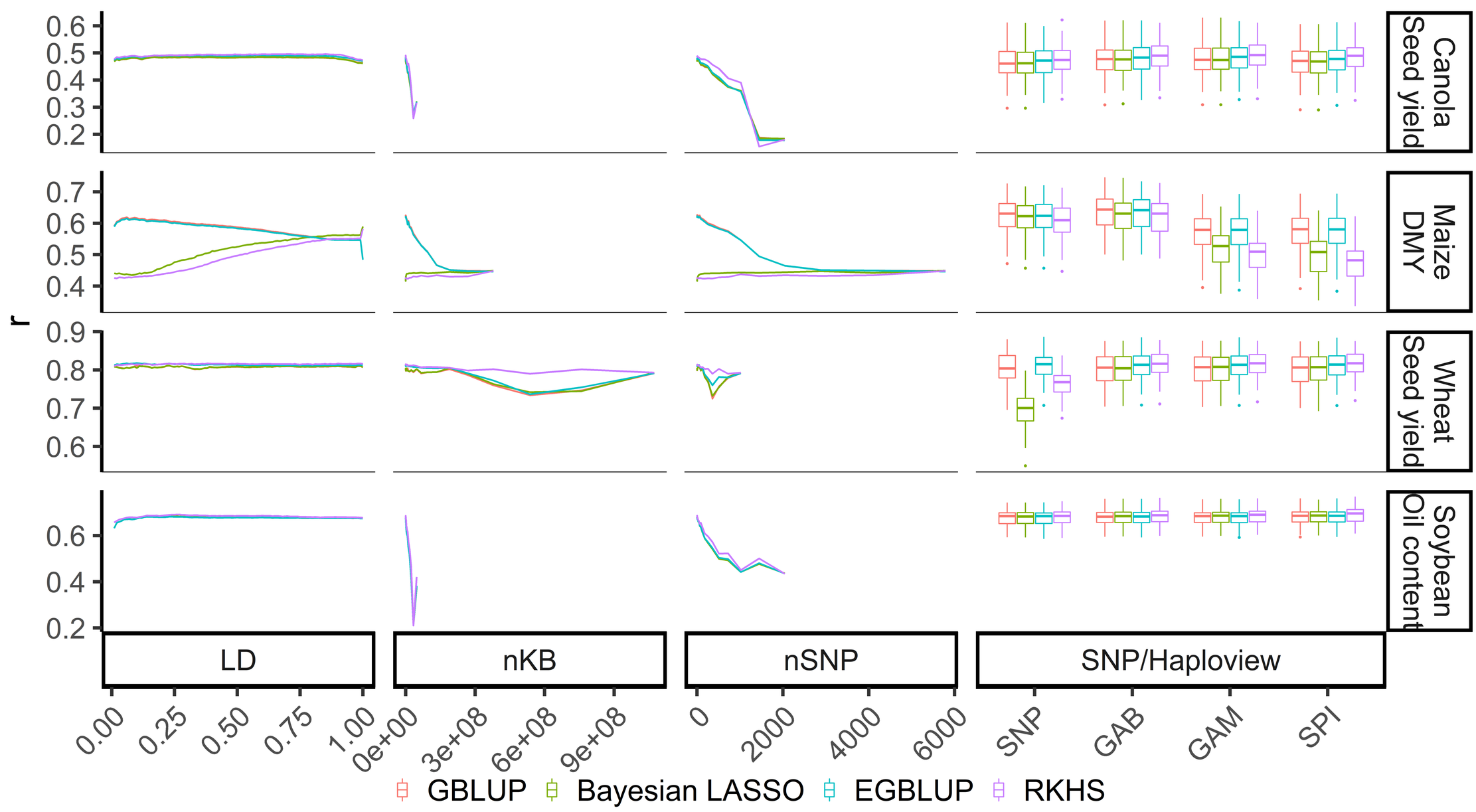

Fig. 1 Prediction accuracy with haplotypes based on varying haplotype block strategies for four example traits in canola, maize, wheat and soybean

Haplotype blocks introduce many additional variants that could be in LD with QTL. However, most of the Haplotypes are rare, which is problematic in standard genomic selection pipelines with medium sized populations. Prediction accuracy was not improved notable with any algorithm and parameter combination. For excessive parameters, especially with fixed window blocks, a dramatic decrease in prediction accuracy could be observed in each dataset for all of the traits. In some scenarios, haplotype blocks were able to close the prediction accuracy gap of underperforming models to other models. Furthermore, there were some differences between prediction models, however no general advantage of any model could be observed. We hypothesize that the lack of accuracy improvement with haplotype blocks is due to inaccurate haplotyping. This could be a consequence of a insuffivient population size, a to low marker density or a consequence of genotyping errors. Summarizing, haplotype blocks seem to have no benefit for genomic prediction in the examined populations.

SPONSORED BY THE

Federal Ministry of Education and Research

This work was performed within the public-private research consortium "*BreedPatH: Breeding Value Pattern Recognition in Hybrid Crops*" with funding from the German Federal Ministry of Education and Research (BMBF grant 031B0890)

References: Bandillo et al. (2015) *The Plant Genome* 8: 04.0024; Jan et al. (2016) *PLOS ONE* 11; Lehermeier et al. (2014) *Genetics* 198: 3-16; Voss-Fels et al. (2019) *Nature Plants* 5: 706-714;

Contact: sven.e.weber@agrar.uni-giessen.de