

University of São Paulo
"Luiz de Queiroz" College of Agriculture

Influence of multi-trait modeling, dominance, and population structure in
genomic prediction of maize hybrids

Danilo Hottis Lyra

Thesis presented to obtain the degree of Doctor in
Science. Area: Genetics and Plant Breeding

Piracicaba
2017

Danilo Hottis Lyra
Bachelor in Biological Sciences

Influence of multi-trait modeling, dominance, and population structure in genomic prediction of maize hybrids

Advisor:
Prof. Dr. **ROBERTO FRITSCHÉ-NETO**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2017

RESUMO

Influência da modelagem *multi-trait*, dominância, e estruturação populacional na predição genômica em híbridos de milho

Predição genômica de híbridos simples é uma promissora ferramenta no melhoramento de milho, pois permite aumentar os ganhos genéticos por unidade de tempo, principalmente por reduzir o tempo de seleção. Uma estratégia que pode aumentar a acurácia das predições genômicas é realizar esta para múltiplos caracteres considerando os mesmos simultaneamente, ou utilizar índices de seleção, os quais captam a performance dos genótipos tanto em condições ótimas como em condições de estresse. Além disso, fatores como dominância, variantes estruturais, e estruturação populacional podem influenciar a acurácia de estimativas dos valores genéticos genômicos (VGG). Portanto, os objetivos foram aplicar predição genômica em híbridos de milho (i) incluindo modelos *multi-trait*, (ii) incorporando desvios de dominância e efeitos da variação no número de cópias, e (iii) controlando a estruturação populacional. Para isto, dois conjuntos de milho (HELIX e USP) foram utilizados, consistindo de 452 e 906 híbridos simples. Os caracteres avaliados foram produtividade de grãos, altura de planta e espiga, senescência, e quatro índices de seleção. A partir das análises *multi-trait* dos modelos GBLUP e GK, pôde-se concluir que a combinação dos índices é uma alternativa viável, aumentando a acurácia seletiva. Além disso, os resultados sugerem que o melhor método é a predição de híbridos incluindo desvios de dominância, principalmente para caracteres complexos. Observou-se também que incluir efeitos relacionados a variação no número de cópias indica ser adequado, devido ao aumento da acurácia e redução do viés nos modelos de predição genômica. Por outro lado, a acurácia de predição não aumentou quando se adicionou quatro diferentes conjuntos de estruturação como covariáveis fixas no modelo GBLUP. No entanto, usando o escalonamento multidimensional não métrico e o agrupamento do fineSTRUCTURE aumentaram a confiabilidade de estimação do VGG para produtividade de grãos e altura de plantas, respectivamente.

Palavras-chave: Milho tropical; Efeitos não-aditivos; Kernel Gaussiano; Variação no número de cópia

ABSTRACT

Influence of multi-trait modeling, dominance, and population structure in genomic prediction of maize hybrids

Genomic prediction of single-crosses is a promising tool in maize breeding, increasing genetics gains and reducing selection time. A strategy that can increase accuracy is applying multiple-trait genomic prediction using selection indices, which take into account the performance under optimal and stress conditions. Moreover, factors such as dominance, structural variants, and population structure can influence the accuracy of estimates of genomic breeding values (GEBV). Therefore, the objectives were to apply genomic prediction (*i*) including multi-trait models, (*ii*) incorporating dominance deviation and copy number variation effects, and (*iii*) controlling population structure in maize hybrids. Hence, we used two maize datasets (HELIX and USP), consisting of 452 and 906 maize single-crosses. The traits evaluated were grain yield, plant and ear height, stay green, and four selection indices. From multi-trait GBLUP and GK, using the combination of selection indices in MTGP is a viable alternative, increasing the selective accuracy. Furthermore, our results suggest that the best approach is predicting hybrids including dominance deviation, mainly for complex traits. We also observed including copy number variation effects seems to be suitable, due to the increase of prediction accuracies and reduction of model bias. On the other hand, adding four different sets of population structure as fixed covariates to GBLUP did not improve the prediction accuracy for grain yield and plant height. However, using nonmetric multidimensional scaling dimensions and fineSTRUCTURE group clustering increased reliability of the GEBV for GY and PH, respectively.

Keywords: Tropical maize; Non-additive effects; Gaussian kernel; Copy number variation

INTRODUCTION

Modern plant breeding comprises combinations of different approaches that include traditional methods, and the use of molecular markers as a “tool” for selecting plants with desirable traits. The last decade has seen tremendous advances in genome-scale data analysis, which was possible due to high-throughput DNA sequencing. In this way, single nucleotide polymorphism (SNPs), representing various regions of all chromosomes, are obtained to be applied in genomic studies (Guo et al. 2016).

Quantitative traits of agricultural importance in plants are influenced by many genomic regions. Thus, whole genome-enabled prediction, such as Genomic Prediction (GP) or Genomic Selection (GS), emerged as a statistical approach to overcome this biological complexity. The main objective of GP, proposed by Meuwissen et al. (2001), is to improve prediction of complex traits based on marker information, increasing precision of selection by generating a genomic estimated breeding value (GEBV) for selection candidates. The accuracies of GP models are most often evaluated by applying validations (independent validation, fold-validation, or jackknife), where all genotypes are randomly divided into training and validation sets (TS, VS). The TS is used to train the prediction model and estimate the marker effects, and by a correlation test using the predicted with the observed values in the validation set, it is possible to obtain the prediction accuracy (PA). The procedure is repeated several times to obtain robust estimates (Zhao et al. 2015).

The most commonly used methods in GP is the genomic best linear unbiased prediction (GBLUP), which utilizes a genomic relationship matrix (GRM) to estimate the genetic merit of an individual. The matrix defines the covariance between individuals based on observed similarity at the genomic level, rather than on expected similarity based on pedigree. Morota and Gianola (2014) reviewed whole-genome regression models using kernel methods to capture non-additive effects, either parametrically (GBLUP) or non-parametrically (Gaussian kernel, GK).

Genomic prediction is superior to phenotypic selection for increasing genetic gains per unit time and shortening the length of the breeding cycle (Heffner et al. 2010). According to Bernardo (2016), GP became a bandwagon in plant breeding in the late 2000s and has been implemented in major seed companies routinely, especially in maize and soybean. In addition, the author argued that similarly with phenotypic selection, applying GP routinely might work reasonably well on average. However, GP still faces challenges in predicting phenotypes of highly polygenic traits due to the complex biological processes, which several factors could influence the estimation of GEBV, such as non-additive effects, population structure, and structural variations (copy number variation). Therefore, the objectives were to apply genomic prediction in maize hybrids (*i*) including multi-trait

models, (ii) incorporating dominance deviation and CNV effects, and (iii) controlling population structure.

CONCLUSION

Multi-trait genomic prediction using the combination of selection indices is an excellent strategy to increase accuracy of selection under abiotic stress conditions.

Including dominance deviation and copy number variation effects in single-crosses prediction for complex traits seems to be suitable, due to the increase of accuracies and reduction of model bias.

Controlling population structure in genomic prediction models may increase reliability and precision of estimation of genomic breeding values in maize hybrids.

REFERENCES

- Bernardo, R. (2016). Bandwagons I, too, have known. *Theoretical and Applied Genetics*, 129: 2323-2332.
- Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z. and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics*, 129(12): 2413-2427.
- Heffner, E. L., Lorenz, A. J., Jannink, J. L. and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Science*, 50: 1681-1690.
- Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157: 1819-1829.
- Morota, G. and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in genetics*, 5: 363.
- Zhao, Y., Mette, M. F. and Reif, J. C. (2015). Genomic selection in hybrid breeding. *Plant breeding*, 134(1): 1-10.