CRISTIANO FERREIRA DE OLIVEIRA

ANÁLISE DE FATORES PARA REDUÇÃO DE DIMENSIONALIDADE EM ESTUDOS DE PREDIÇÃO GENÔMICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientador: Cosme Damião Cruz

Coorientador: Moysés Nascimento

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
O48a
2022

Oliveira, Cristiano Ferreira de, 1990-
Análise de fatores para redução de dimensionalidade em estudos de predição genômica / Cristiano Ferreira de Oliveira. – Viçosa, MG, 2022.
1 tese eletrônica (63 f.): il. (algumas color.).

Texto em português e inglês.
Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.026
Modo de acesso: World Wide Web.

1. Soja - Melhoramento genético. 2. Polimorfismos de nucleotídeo único. 3. Marcadores genéticos - Seleção. 4. Análise fatorial. 5. Aprendizado do computador. 6. Haplótipos. I. Cruz, Cosme Damião, 1958-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 633.342

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

CRISTIANO FERREIRA DE OLIVEIRA

ANÁLISE DE FATORES PARA REDUÇÃO DE DIMENSIONALIDADE EM ESTUDOS DE PREDIÇÃO GENÔMICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae.*

APROVADA: 12 de dezembro de 2022.

Assentimento:

gov.br
Documento assinado digitalmente
CRISTIANO FERREIRA DE OLIVEIRA
Data: 01/03/2023 21:05:11-0300
Verifique em https://verificador.iti.br

Cristiano Ferreira de Oliveira
Autor

Cosme Damião Cruz
Orientador

Aos meus pais Ana e Luiz,

A meus irmãos, sobrinhos e amigos.

AGRADECIMENTOS

Agradeço a Deus por ter me dado força para vencer mais uma importante etapa da minha vida. Aos meus pais, Luiz e Ana e irmãos Luiz e Fernando por estarem sempre ao meu lado. Ao meu orientador e amigo Cosme Damião Cruz pelos conselhos, pela disponibilidade, incentivo e confiança depositada na execução deste trabalho. Ao professor, co-orientadora e amigo Moysés Nascimento pelas sugestões e apoio. Ao professor Luiz Alexandre Peternelli pela oportunidade de atuar no LAPEA, pelos ensinamentos e amizade. Aos membros da banca examinadora, Prof. Doutor Hélcio Duarte Pereira, Prof. Doutor Vinícius Quintao Carneiro, Profa. Doutora Gabi Nunes Silva pela disponibilidade e pelas valiosas sugestões para o enriquecimento deste trabalho. Aos meus amigos do Laboratório de Bioinformática e LAPEA pelos ótimos momentos e pela valiosa amizade. Aos amigos de Viçosa pelos bons momentos e parceria. À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade e aos professores e funcionários do Departamento de Estatística da UFV, pela competência profissional e por todo apoio dado ao longo das minhas atividades acadêmicas. À CAPES, pela concessão da bolsa de estudos. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

OLIVEIRA, Cristiano Ferreira, D.Sc., Universidade Federal de Viçosa, dezembro de 2022. **Análise de fatores para redução de dimensionalidade em estudos de predição genômica**. Orientador: Cosme Damião Cruz. Coorientador: Moysés Nascimento.

O conceito de seleção genômica tem como base o desequilíbrio de ligação (LD) entre locos de características quantitativas (QTLs) e marcadores. Uma variação genética que se relaciona com a forma que o fenótipo é expresso conduz a múltiplas associações estatísticas em marcadores próximos em termos de ligação fatorial ou de desequilíbrio, podendo estas associações ser ou não de causa e efeito. Assim ao construir modelos preditivos, em geral não é conhecido quais SNPs possuem de fato associação de causa e efeito com o fenótipo de interesse, consequentemente o modelo é construído utilizando todas as informações genotípicas. Com o intuito de aumentar a acurácia dos modelos de predição, diferentes abordagens de seleção de marcadores foram propostas. São estratégias utilizadas para isto selecionar SNPs relatados anteriormente em estudos de associação para a característica de interesse, estimar a significância dos SNPs no conjunto de dados para cada característica utilizando um modelo preditivo e o efeito dos marcadores estimados pelo modelo, ou a seleção subconjuntos dos marcadores uniformemente espaçados ao longo do genoma. Dentre as abordagens citadas anteriormente, a seleção uniformemente espaçada ao longo do genoma é a mais versátil, uma vez que um painel de baixa densidade formado por meio dela pode ser utilizado em estudos de predição de valores genéticos de qualquer característica, diferentemente das outras abordagens citadas. Porém esta seleção está sujeita a possibilidade de excluir por completo blocos de haplótipos em LD relacionados com o fenótipo de interesse. Este trabalho foi desenvolvido com o objetivo de propor uma abordagem de seleção de marcadores espaçados dentro de blocos de haplótipos construídos utilizando Análise de Fatores (AF). Mostramos, utilizando dados simulados que a Análise de Fatores pode ser utilizada para construir os blocos de haplótipos, sendo ela capaz de sintetizar a relação linear entre marcadores e criar fatores comuns que podem ser interpretados como blocos de LD. Em seguida utilizamos em um conjunto de dados de soja, contendo 41985 marcadores do tipo SNPs com informação de 20087 acessos de soja, esta abordagem para construir os blocos e então foi feito a seleção espaçada dentro dos blocos formados a partir da AF. Três painéis de SNPs foram considerados, contendo 1%, 5% e 100% dos marcadores. Para avaliar o êxito desta abordagem, foi considerado a acurácia em uma tarefa

de predição do valor fenotípico dos indivíduos utilizando os painéis reduzidos e o painel completo. Os resultados mostram que ao utilizar os painéis reduzidos não há diferença significativa de acurácia seletiva comparado a acurácia obtida utilizando o painel completo e para uma das características avaliadas também não foi encontrada diferença significativa para acurácia preditiva.

Palavras-chave: SNP. GWS. Seleção de Marcadores. Análise Fatorial. Soja. Aprendizado de Máquina. Blocos de Haplótipos.

ABSTRACT

OLIVEIRA, Cristiano Ferreira, D.Sc., Universidade Federal de Viçosa, December, 2022. **Factor Analysis for dimensionality reduction in genomic prediction studies**. Adviser: Cosme Damião Cruz. Co-adviser: Moysés Nascimento.

The concept of genomic selection is based on the linkage disequilibrium (LD) between quantitative trait loci (QTLs) and markers. A genetic variation that is related to the way the phenotype is expressed leads to multiple statistical associations in close markers in terms of factorial linkage or disequilibrium, these associations may or may not be cause and effect. Thus, when building predictive models, it is generally not known which SNPs have a cause-and-effect association with the phenotype of interest, consequently, the model is built using all genotypic information. To increase the accuracy of prediction models, different marker selection approaches have been proposed. Strategies used for this are selecting previously reported SNPs in association studies for the trait of interest, estimating the significance of SNPs in the dataset for each trait using a predictive model and the effect of markers estimated by the model, or selecting subsets of markers evenly spaced throughout the genome. Among the approaches mentioned above, selection evenly spaced throughout the genome is the most versatile, since a low-density panel formed through it can be used in studies to predict the genetic values of any trait, unlike the other approaches mentioned. However, this selection is subject to the possibility of completely excluding blocks of haplotypes in LD related to the phenotype of interest. This work was developed with the aim of proposing an approach for selecting spaced markers within blocks of haplotypes constructed using Factor Analysis (FA). We show, using simulated data, that Factor Analysis can be used to build blocks of haplotypes, being able to synthesize the linear relationship between markers and create common factors that can be interpreted as blocks of LD. Next, we used this approach to build the blocks in a soybean dataset, containing 41985 SNPs type markers with information from 20087 soybean accessions, and then the spaced selection was made within the blocks formed from the FA. Three panels of SNPs were considered, containing 1%, 5% and 100% of the markers. To assess the success of this approach, the accuracy of a task to predict the phenotypic value of individuals using the reduced panels and the full panel was considered. The results show that when using the reduced panels there is no significant difference in selective accuracy compared to the accuracy obtained using

the full panel and for one of the evaluated characteristics no significant difference was found for predictive accuracy.

Keywords: Dissertation. SNP. GWS. Marker Selection. Factor Analysis. Soybean. Machine Learning. Haplotype Blocks.

SUMÁRIO

INTRODUÇÃO GERAL

A soja (*Glycine max*) é uma importante leguminosa que devido a sua versatilidade e importância nutricional é cultivada em diversos ambientes ao redor do mundo, para diversos fins (SHEA; M. SINGER; ZHANG, 2020). A demanda mundial pelo óleo de soja e outros produtos derivados dela tem aumentado. Além do desenvolvimento e evolução de técnicas de manejo o melhoramento genético da cultura é fundamental para aumentar sua produção.

Os avanços nas técnicas de genética molecular tornaram possível, aos programas de melhoramento de plantas e animais, genotipar indivíduos a baixo custo em relação a muitos milhares de marcadores polimórficos de nucleotídeo único (SNPs) espalhados por todo o genoma. A fim de apoiar a comunidade de pesquisa da soja um repositório central (https://soybase.org) de dados genéticos e genômicos foi criado. Neste repositório estão disponíveis diversas informações fenotípicas de milhares de plantas bem como um painel de SNPs de alta densidade contendo milhares de informações moleculares destas plantas.

O uso de informações genotípicas possibilita reduzir o tempo de geração de seleção e estimar, com maior precisão, o valor genético de indivíduos que não possuem registro fenotípico próprio e não tem descendência (MEUWISSEN; HAYES; GODDARD, 2001).

Com uma população de referência suficientemente grande contendo indivíduos com fenótipos e genótipos conhecidos, os efeitos do SNP podem ser estimados. Posteriormente, estas estimativas podem ser utilizadas para predizer valores genéticos de novos indivíduos que possuam apenas informações genotípicas. A seleção com base nesses valores genômicos de reprodução é chamada seleção genômica.

A seleção genômica tem como base o desequilíbrio de ligação (LD) entre locos de características quantitativas (QTLs) e marcadores (LI et al., 2018; REICH et al., 2001). Diferentes propostas para medir a extensão do desequilíbrio podem ser encontradas (PRITCHARD et al., (2001), CARNEIRO et al., (2002), MCRAE et al., (2002) e LI et al., (2016)), muitas delas tem como base o coeficiente de correlação de Pearson.

Padrões de LD podem ser utilizados para estudar a estrutura dos blocos de haplótipos da variação do SNP no DNA. Neste contexto, a análise fatorial (AF) pode ser considerada. Este procedimento estatístico pode ser utilizado para reduzir a complexidade do problema original, agrupando as p variáveis aleatórias, informações moleculares, X1, . . ., Xp, em grupos formados por variáveis fortemente correlacionadas.

Apesar de a AF ser utilizada mais comumente em dados de estrutura na qual o número de amostras é superior ao de variáveis (COSTELLO; OSBORNE, 2005) (HAIR et al., 2014) formulou-se a hipótese de que a técnica de análise de fatores seria capaz de identificar subgrupos de marcadores que refletissem grupos de ligação fatorial ou grupos de desequilíbrio de ligação (blocos de haplótipos) que pudessem ser utilizados para um processo de seleção de variáveis orientada.

Na literatura destacam-se três abordagens de seleção de marcadores, a seleção de marcadores baseada em estudos a priori de associação entre os SNPs e a característica de interesse, seleção baseada em modelos preditivos e seleção espaçada ao longo do genoma.

A seleção de marcadores baseada em modelos preditivos estima a importância dos marcadores baseada em sua influência na característica estudada (BREIMAN, 2001; CHEN; ISHWARAN, 2012). A regressão por *stepwise* também pode ser utilizada para seleção de atributos, onde a escolha das marcas mais importantes é realizada por um procedimento automático no qual em cada etapa, um marcador é considerada para adição ou subtração do conjunto de variáveis explicativas (SANT'ANNA et al., 2020). Alguns modelos como GBLASSO (*Generalized Bayesian Lasso*) permitem selecionar marcadores considerando aqueles que possuem os maiores valores absolutos de coeficiente de regressão (SANT'ANNA et al., 2020; SOUSA et al., 2021).

A seleção de marcas com base em estudos a priori de GWAS consiste em incluir no painel de baixa densidade apenas SNPs que foram identificados através de associação genômica em estudos anteriores. VALLEJO et al., 2018 utilizaram um painel com 70 loci de caractere quantitativas (QTL) identificados a priori para predizer resistência à doença bacteriana da água fria (BCWD) em trutas arco-íris.

As duas abordagens citadas acima estão ligadas à característica a ser avaliada, ou seja, um painel de SNPs selecionados por estas abordagens tem sua eficiência ligada a característica de interesse. Se é de interesse predizer outra característica, então outro painel de SNPs deve ser construído para este propósito. Já a seleção espaçada ao longo do genoma independe de qual fenótipo deseja-se predizer. Um painel construído por meio desta abordagem pode ser utilizado para criar modelos preditivos de qualquer característica da população (OGAWA et al., 2014; VALLEJO et al., 2017, 2018). Apesar desta vantagem em relação as demais abordagens a seleção espaçada está sujeita a possibilidade de excluir blocos inteiros contendo marcadores em LD com QTL importantes para a característica em estudo.

Desta forma, o trabalho foi realizado com o objetivo de propor uma seleção espaçada de marcadores orientada por blocos de haplótipos formados via análise de fatores.

REFERÊNCIAS BIBLIOGRÁFICAS

BORÉM, A.; CAIXETA, E. **Marcadores Moleculares**. Viçosa: UFV, 2016. v. 1

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Outubro 2001.

CARNEIRO, M. S.; VIEIRA, M. L. C. Mapas genéticos em plantas. **Bragantia**, v. 61, n. 2, p. 89–100, ago. 2002.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, v. 99, n. 6, p. 323–329, jun. 2012.

DEON VILELA DE RESENDE, M.; FONSECA E SILVA, F.; FERREIRA AZEVEDO, C. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão, aleatória, seleção genômica, QTL, GWAS, estatística espacial e temporal, competição, sobrevivência.** Viçosa: UFV, 2014. v. 1

GRANT, D. et al. SoyBase, the USDA-ARS soybean genetics and genomics database. **Nucleic Acids Research**, v. 38, n. suppl_1, p. D843–D846, jan. 2010.

LI, B. et al. Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontiers in Genetics**, v. 9, p. 237, 2018.

MCRAE, A. F. et al. Linkage Disequilibrium in Domestic Sheep. **Genetics**, v. 160, n. 3, p. 1113–1122, 1 mar. 2002.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, v. 157, n. 4, p. 1819–1829, abr. 2001.

NADEEM, M. A. et al. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. **Biotechnology & Biotechnological Equipment**, v. 32, n. 2, p. 261–285, 4 mar. 2018.

OGAWA, S. et al. Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. **BMC Genetics**, v. 15, n. 1, p. 15, 2014.

PRITCHARD, J. K.; PRZEWORSKI, M. Linkage Disequilibrium in Humans: Models and Data. **The American Journal of Human Genetics**, v. 69, n. 1, p. 1–14, jul. 2001.

REICH, D. E. et al. Linkage disequilibrium in the human genome. **Nature**, v. 411, n. 6834, p. 199–204, maio 2001.

SANT'ANNA, I. DE C. et al. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. **Acta Scientiarum. Agronomy**, v. 43, p. e46307, 17 ago. 2020.

SHEA, Z.; M. SINGER, W.; ZHANG, B. Soybean Production, Versatility, and Improvement. Em: **Legume Crops [Working Title]**. [s.l.] IntechOpen, 2020.

SOUSA, I. C. DE et al. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, v. 78, n. 4, p. e20200021, 2021.

VALLEJO, R. L. et al. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. **Genetics Selection Evolution**, v. 49, n. 1, p. 17, dez. 2017.

VALLEJO, R. L. et al. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: Evidence that long-range LD is a major contributing factor. **Journal of Animal Breeding and Genetics**, v. 135, n. 4, p. 263–274, 2018.

CAPÍTULO 1

Identificação de padrões relacionados a grupos de ligação ou desequilíbrio por análise de fatores

ABSTRACT

Published as original paper to Ciência Rural
Reference: OLIVEIRA, C. F. de, TEIXEIRA, G., TEMOTEO, A. da S., et al. **Identifica-tion of patterns related to linkage groups or disequilibrium by factor analysis**, Ciência Rural, v. 51, n. 5, p. e20190984, 2021. DOI: 10.1590/0103-8478cr20190984.

Empirical patterns of linkage disequilibrium (LD) can be used to increase the statistical power of genetic mapping. This study was carried out with the objective of verifying the efficacy of factor analysis (FA) applied to data sets of molecular markers of the SNP type, in order to identify linkage groups and haplotypes blocks. The SNPs data set used was derived from a simulation process of an F2 population, containing 2000 marks with information of 500 individuals. The estimation of the factorial loadings of FA was made in two ways, considering the matrix of distances between the markers (A) and considering the correlation matrix (R). The number of factors (k) to be used was established based on the graph scree-plot and based on the proportion of the total variance explained. Results indicated that matrices A and R lead to similar results. Based on the scree-plot we considered k equal to 10 and the factors interpreted as being representative of the bonding groups. The second criterion led to a number of factors equal to 50, and the factors interpreted as being representative of the haplotype's blocks. This showed the potential of the technique, making it possible to obtain results applicable to any type of population, helping or corroborating the interpretation of genomic studies. The study demonstrated that FA was able to identify patterns of association between markers, identifying subgroups of markers that reflect factor binding groups and also linkage disequilibrium groups.

Keywords: Linkage Disequilibrium. Factorial Analysis. SNP. Haplotype Blocks. QTL.

INTRODUCTION

Genetic markers of the SNP type (Single Nucleotide Polymorphism) are based on the occurrence of polymorphism resulting from the alteration of a single genome base. Besides being the most abundant form of polymorphism reported in the genome, the SNPs stand out when compared to other types of molecular markers at low cost, low mutation rate, and genotyping ease (BORÉM; CAIXETA, 2016; DEON VILELA DE RESENDE; FONSECA E SILVA; FERREIRA AZEVEDO, 2014; NADEEM et al., 2018). The use of molecular information in the process of genetic breeding brings great benefits; phenotypic information, combined with genotypic information, provides greater precision to predict its genetic value (MEUWISSEN; HAYES; GODDARD, 2016; SPINDEL et al., 2015).

Many important agricultural characteristics, such as grain yield and fruits and their primary components, are of quantitative gene control ruled by many genes with complex actions and interactions. The regions within genomes that contain genes associated with a quantitative trait are known as quantitative character loci. Identification of QTLs (Quantitative Trait Loci) based only on conventional phenotypic evaluation is not possible, but with the use of DNA markers, it is possible to establish gene binding maps and subsequently detect, map, and quantify its effects on and importance to genetic variability (COLLARD et al., 2005). These maps are used to identify chromosome regions that contain genes linked to quantitative characteristics (KUMAR et al., 2015).

The high-density single nucleotide polymorphism maps make it possible to map genes efficiently, exploring the linkage disequilibrium between genes of interest and adjacent markers (MCRAE et al., 2002). The linkage disequilibrium (LD), a measure of dependence or not of alleles of two or more loci, is crucial for the detection of QTL, for the selection aided by markers, and for the prediction by genome wide selection (CAETANO, 2009; GHOLAMI et al., 2015).

LD refers to the non-independence of alleles in different locations. Considering a locus with alleles A and a with frequencies pA and (1 - pA), respectively, and a second with alleles B and b with frequencies pB and (1 - pB). If the loci are independent, the frequency expected for the AB haplotype will be pA pB. If the population frequency of the AB haplotype is greater or less than the expected value, the specific alleles tend to be observed together, and then we say that the two loci are in LD.

In the literature we reported several proposals to measure the extent of disequilibrium (CARNEIRO; VIEIRA, 2002; PRITCHARD; PRZEWORSKI, 2001). McRAE et al. (2002), studied the extension of LD in two domestic sheep populations and LI, G. et al. (2016), in order to identify new resistance genes to leaf rust in the core of wheat germplasm using GWAS, calculated the LD for all comparisons between pairs of SNPs.

Empirical patterns of LD can be used to study the structure of the block of haplotypes of the variation of the SNP in DNA. The structure of haplotypes blocks can be used to increase the statistical power of genetic mapping (GREENSPAN; GEIGER, 2004). There are many studies that have proposed to identify linkage disequilibrium (LD) patterns, which resemble blocks, across the genome (DALY et al., 2001; GABRIEL et al., 2002; PATIL et al., 2001; REICH et al., 2001; WANG et al., 2002).

SHIFMANet al., 2003 measured the LD between pairs of SNPs using the absolute value of Lewontin's D′ (|D′|) and r statistics. They found that measuring LD with r or r2 has several advantages over D' exhibiting more reliable sampling properties.

There are different ways of defining haplotype blocks (REICH et al., 2001; DALY et al., 2001; PATIL, 2001; WANG et al., 2002). Haplotype blocks are understood as groups of highly correlated markers, probably in LD.

In this context, factor analysis has the potential to be used due to its statistical procedure, that allows us to reduce the complexity of the original problem, grouping p random variables, which are understood in this research as representatives of molecular information, X 1 , ..., X p , in groups formed by strongly correlated variables.

The factor analysis (AF) is used more commonly in data whose number of observations is greater than the number of variables (COSTELLO; OSBORNE, 2005; HAIR, 2010) A great challenge encountered by researchers, who reported in the area of molecular genetics, is to manipulate and analyze large data sets containing large numbers of variables such as matrices from SNPs chips, containing thousands of variables.

Thus, the hypothesis that the technique of factor analysis would be able to identify, in a large set of molecular information, subgroups of markers that reflected factorial binding groups or groups of linkage disequilibrium (blocks of haplotypes) to orient future dimensionality reduction or structural simplification was formulated. Thus, the study was carried out with the objective of verifying the efficacy of AF applied to data sets with high dimensionality and data from molecular markers of the SNP type, aiming to identify groups of linkage and blocks of haplotypes.

MATERIAL AND METHODS

Molecular data

The SNPs data set used was derived from a simulation process made with the computational application, Genes (CRUZ, 2016). Initially, the basic genome, matrix G, was generated containing ten linkage groups, with 200 marks per linkage group. G was used to generate the genotype information for the genitors P1 and P2, and these were contrasting homozygous parents.

From the genotypic information of P1 and P2, F1 and its random mating were simulated, giving rise to the genotypic information of an F2 population was simulated, containing 2000 marks with information of 500 individuals, generating the matrix of SNPs, with the columns ordered according to the chromosome to which the marker belonged and its position on the chromosome. Considering that the size of these data allows verifying, without loss of generalization, the efficiency of factor analysis in a scenario in which the number of p variables is greater than the number of observations, the technique was applied to identify and group the SNPs.

Methodology

In order to establish a way of recognition in the set of subsets of markers representing certain linkage groups or groups of disequilibrium, it is recommended that once these groups are used, it is possible to establish samples within each group and to continue the prediction study with a set of information of lower dimensionality. The strategies used were:

Establishment of genetic maps

It is a strategy applicable only in populations derived from controlled crossings from parent homozygous contrast, such as F2 ... Fn, RILs (Recombinant Inbred Lines), double-haploid, and backcrosses or exogamic populations, such as half-siblings or full-sib families. It requires a preliminary study proving the Mendelian segregation of each brand measured, followed by the calculation of the distance between markers, grouping, and ordering.

Because it was an F2 population, it was possible to establish genetic maps identifying the bonding groups from which marker samples could be established for further prediction studies.

Establishing correlation maps

Correlations between pairs of markers were obtained and represented graphically, seeking to identify the intensity of the disequilibrium between the pairs of brands considered. However, it should be kept in mind that the objective is to identify representative sets of linkage group or groups of disequilibrium, which is complex, because there is not always prior information on the ordering of brands as in the considered set of data.

The graph heat map was used to illustrate the sample correlation matrix (R). The heat map is commonly used when we have a data set with many variables and we want to graphically visualize the intensity of the relationship between them. In this graph it is possible to identify the strength and signal of the correlations between the variables based on the colors.

Structuring the data set into groups of characters established through factor analysis

A more general strategy, which does not depend on the type of population or previous knowledge of the ordering of markers, is also presented and detailed in this work, referring to the analysis of factors consisting of the structural simplification of the matrix some common factors.

Establishment of common factors for marker information

The variables in the factorial model, in this case the molecular markers, are represented as a linear function of variables or common factors, not observable and by random error, which is specific to each marker.

The factorial analysis is a method used to investigate whether the number of variables of interest, X1, X2, ..., XP, is linearly related to a smaller number of the non-observable factors, F1, F2, ..., Fk. Considering, a p-dimensional random vector with an averages vector ${}_p\mu_1$ and covariance matrix ${}_p\Sigma_p$, the factorial model can be written as:

$$Y - \mu = \Gamma \mathrm{F} + \epsilon$$

such that ${}_p\Gamma_k = [Y_{ij}]$, which is a matrix of coefficients denominated by factorial loads and has rank $k \leq p$. ${}_m\mathrm{F}_1$ is a random vector of non-observable latent common factors, and ${}_p\epsilon_1$ is the vector of random errors.

Expanding in the form of a system of equations we would have:

$$\begin{array}{lcccccccccccc}
Y_1 - \mu_1 & = & \gamma_{11}F_1 & + & \ldots & + & \gamma_{1\mathrm{j}}F_{\mathrm{j}} & + & \ldots & + & \gamma_{1\mathrm{m}}F_{\mathrm{m}} & + & \epsilon_1 \\
\vdots & & & & & \vdots & & & & & \vdots & & \\
Y_i - \mu_i & = & \gamma_{\mathrm{i}1}F_1 & + & \ldots & + & \gamma_{\mathrm{ij}}F_{\mathrm{j}} & + & \ldots & + & \gamma_{\mathrm{im}}F_{\mathrm{m}} & + & \epsilon_i \\
\vdots & & & & & \vdots & & & & & \vdots & & \\
Y_p - \mu_p & = & \gamma_{\mathrm{p}1}F_1 & + & \ldots & + & \gamma_{\mathrm{pj}}F_{\mathrm{j}} & + & \ldots & + & \gamma_{\mathrm{pm}}F_{\mathrm{m}} & + & \epsilon_p
\end{array}$$

If $Cov(Y) = \text{IK}$ or, in other words, if the factors are not correlated, we call the model an orthogonal factorial.

Also, considering the assumptions that $E(Y) = \mu$, $E(F) = E(\epsilon) = 0$, $Cov(Y) = \Sigma$, $Cov(\epsilon) = \psi$, and $Cov(F, \epsilon) = 0$, such that,

$$\psi = \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p \end{bmatrix}$$

we can show that $\Sigma = \Gamma\Gamma^t + \psi$. The elements of the main diagonal of the $\Gamma\Gamma^t$ are called communalities or common variances and are defined by

$$h_i^2 = \sum_{j=1}^{k} Y_{ij}^2 .$$

To measure data's adequacy, the LEDOIT & WOLF, (2002) LEDOIT sphericity test, indicated for situations in $n < p$, was used (FERREIRA, 2011). The number of factors was defined through procedures based on the ordering of the eigenvalues of matrix R to verify their importance.

The first criterion is based on the scree-plot chart. In this chart the eigenvalues are sorted in descending order, and a point is sought, from which there is a decrease of importance in relation to the total variance (CATTELL, 1966).

The second criterion is based on the analysis of the proportion of the total explained variance, where the *k* number of factors was defined so that the proportion of the total variance explained up to *the k-th* factor was approximately 85%.

These two criteria consider only the numerical magnitude of the eigenvalues. The appropriate choice of the k value should take into account the interpretability of the factors and the principle of the model's parsimony (MINGOTI, 2005). For the estimation of factor loadings of factor analysis, the correlation matrix and the principal components method were used.

The formation of the groups was performed and established through an iterative process, with the criterion that variables whose higher factorial load was given to the *i-th* factor would be allocated in group *i* as illustrated in figure 1.

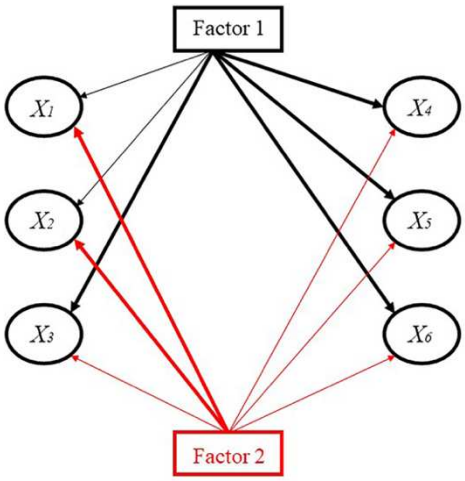


Figure 1: Grouping scheme using Analysis of Factors, with six variables and two factors. Thicker lines indicate higher factorial load. Group 1 consists of X3, X4, X5, and X6. Group 2 consists of X1 and X2.

Because it was an F2 population, it was possible to obtain a matrix D, whose elements represented the distances between pairs of markers, expressed in centimorgan (cM) and ranging from 0 to 0.5. Thus, the analyses described above were also made using an input matrix A, originating from the distance matrix between markers. The array A was defined as:

$$A = 1 - 2D$$

In some cases, being able to identify which variables belong to which factor and interpret the original factors may not be an easy task due to the occurrence of coefficients with similar numerical quantities in several factors. In this case, the partition on k factors is unclear and we may be violating the assumption of orthogonality of the factors. To work around this problem, the orthogonal rotation of the original factors was used.

Orthogonal rotation alters the factorial loads but conserves the perpendicularity between the factors, as illustrated in figure 2, and maintains its statistical properties as the communalities and specific variances. The type of orthogonal varimax rotation is the most used (COSTELLO; OSBORNE, 2005; LIU et al., 2009; PALLANT, 2016), and this method seeks to minimize the

number of variables that present high loads in each factor (LOEHLIN, 2004). Seeing this, we opted for this orthogonal rotation.

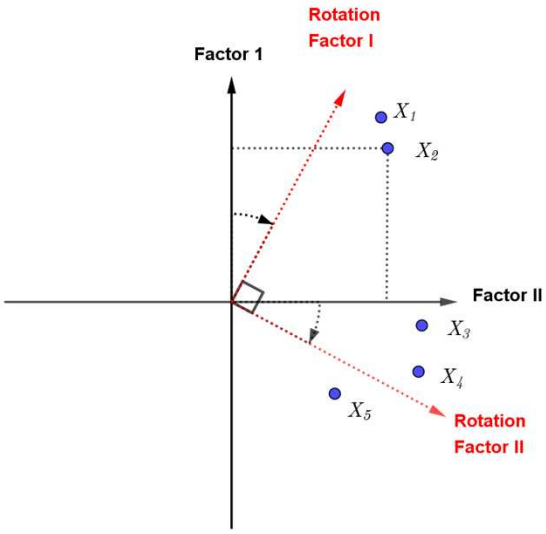


Figure 2: Illustration representing an orthogonal factorial rotation.

All analyses were performed in the free software R (R CORE TEAM, 2019).

RESULTS AND DISCUSSION

Recognition of groups of factorial linkage and linkage groups through genomic studies

The F2 population dataset underwent genomic analysis and provided the genetic map illustrated in figure 3. It can be observed that the 2000 markers genotyped were grouped and ordered, in order to reflect the basic number of chromosomes of the hypothetical species studied. This information is useful for assisted selection purposes and, in this context, to indicate that a sampling within each binding group would be an efficient procedure for reducing dimensionality, so that the smaller set of markers (independent variables) would also exploit the binding disequilibrium contained in the original group.

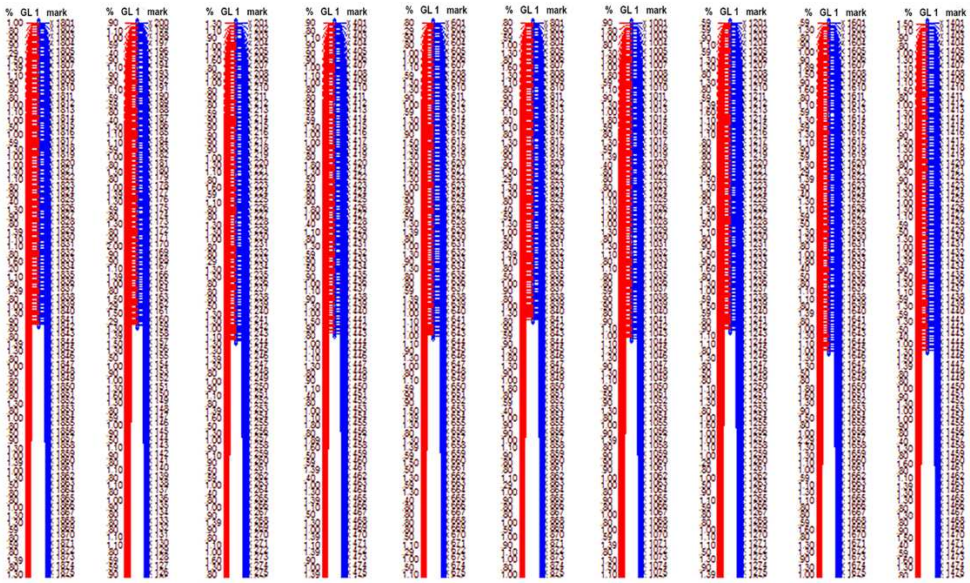


Figure 3: Genetic map established for a population F2 considering information of 2000 molecular markers of the SNP type.

QTL analysis has been used to identify molecular markers responsible for the variation in an observed trait (DHARIWAL et al., 2018; LI et al., 2016). In our study on genomic analysis, the detection of QTLs associated with a characteristic of interest is represented in Figure 4, and was established using the simple interval method (CRUZ, 2016; TERRA et al., 2016).

Two examples of the linkage group where the presence of QTL was detected and not detected were presented in figure 4a) and 4b), respectively, in view of the values of LOD (log of odds ratio) obtained in the analyses of each interval in each linkage group. For the purpose of dimensionality reduction, the researcher may choose samplings prioritizing markers that represents the regions with higher concentration of putative controlling loci of the characteristics included in the analyses.

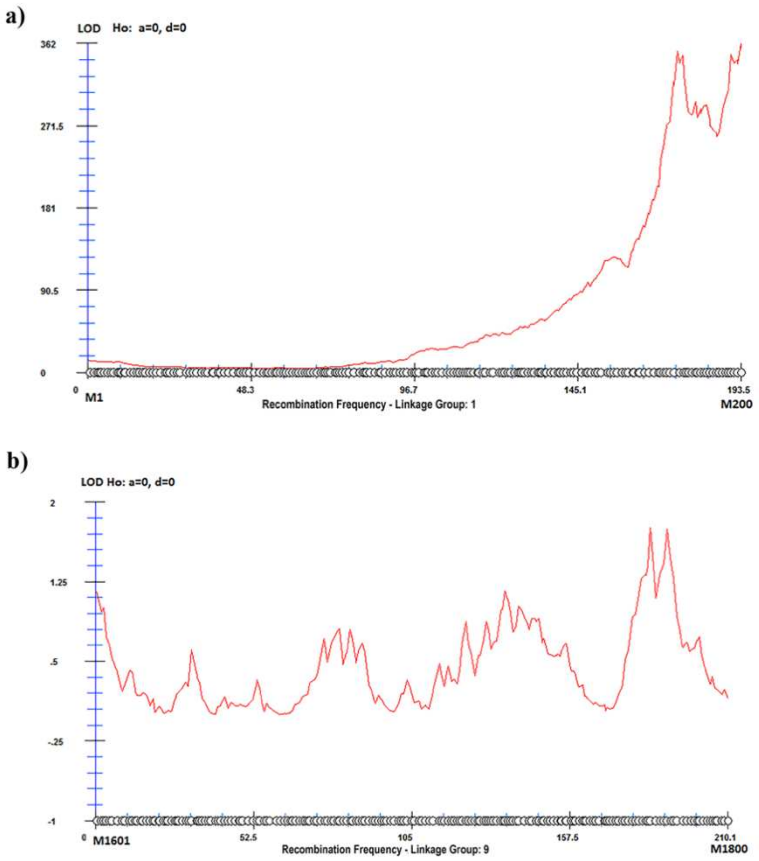


Figure 4: Genetic map and QTL detection map for a quantitative characteristic measured in a F2 population, evidencing the situations: a) Detection of QTLs (peak above LOD referential equal to 3; b) Absence of QTL for the characteristic of interest.

Recognition of factorial binding groups and linkage groups through pattern analysis in correlation matrix

In general r^2 decreases as the distance between markers increases, even between pairs of SNPs that can be defined as belonging to the same haplotype block (SHIFMAN, 2003). High values of r^2 can be found between markers belonging to the same block as verified by GABRIEL et al. (2002) and SHIFMAN et al. (2003).

In the heatmap, shown in figure 5, this same pattern can be observed. Highly correlated groups, in a structure in which there is high correlation within groups and low between groups. There were ten groups that were highlighted, formed by markers that are on the same chromosome, thus emphasizing the formation of the bonding groups. However, it should be highlighted that the explicit grouping pattern in figure 5 of them is the result of a previous

organization of the data set submitted to the analysis, where the ordering of the brands established by simulation was already known.

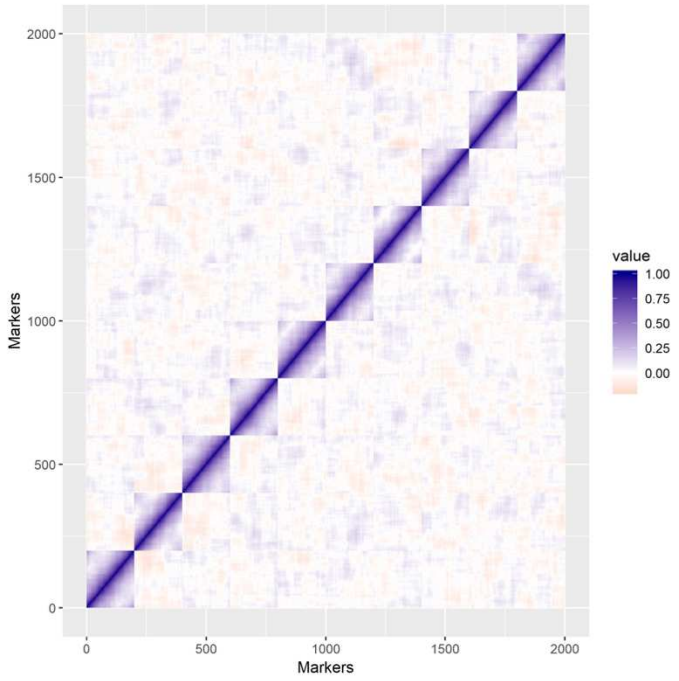


Figure 5: Correlation map between pairs of markers highlighting the intensity of the associations within the disequilibrium blocks.

Studies aiming to identify linkage disequilibrium patterns in populations in which mapping, grouping, and ordering of markers cannot be established, the heat map would only highlight the existence of the disequilibrium. In this context, the establishment of blocks would not be possible and therefore, the heatmap wouldn't have usefulness for targeted sampling orientation aiming at dimensionality reduction.

Recognition of factorial linkage groups and linkage groups by factor analysis

In this procedure, we seek to present a generalist method capable of subdividing the original set of markers in k subgroups without the need for genomic analyses and restricted to certain types of populations, and without previous knowledge of grouping and planning.

According to the sphericity test proposed by LEDOIT et al. (2002), which presented significance statistics *(P <0.01)*, it was reported that the data are adequate for factor analysis.

Definition of the number of factors

In figure 6 is presented the result of the first criterion defined with the objective of establishing the number of factors (k) that would enable the efficient simplification structure of the initial set of markers. Figure 6 a) ensures that the point of jump that would be representing a decrease of importance is between the 100 first self-values. Thus, for a better visualization, in figure 6 b) are presented the estimates the 100 first self-values ordered, and the occurrence of an expressive leap point in the tenth self-value indicates that the number of factors suitable for structural simplification should be equal to 10.

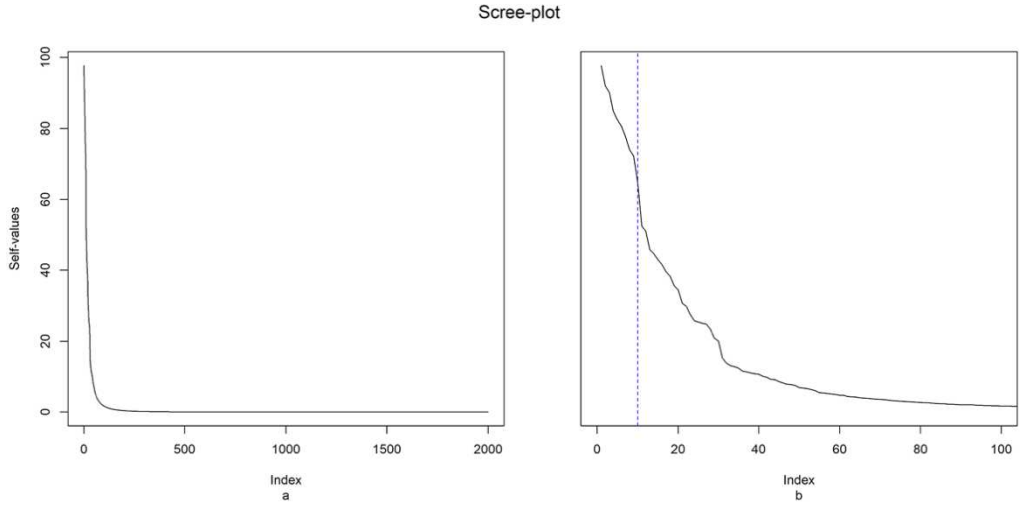


Figure 6: a) Estimates of self-values in descending order, obtained from the correlation matrix between pairs of markers and b) Highlight of the 100 first estimates of the self-values in descending order.

In contrast, when using the criterion of the explained proportion of accumulated variance, obtaining 85% of the variance would require 50 factors. According to these results, the analysis was performed considering two scenarios: 10 factors and 50 factors. In each scenario, the study was performed using the correlation matrix (R) and also the matrix A, established based on the genetic distance matrix between pairs of markers (D).

Interpretation of the factors

For the purposes of better interpretation, the analysis of factors was performed using the varimax rotation. The goal of rotation is to simplify the structure of the data, recognizing the markers of higher factorial loads so that, for the analysis with 10 factors, the most important

marks are those shown in figure 7. For the analysis with 50 factors, the results are represented in figures 8.

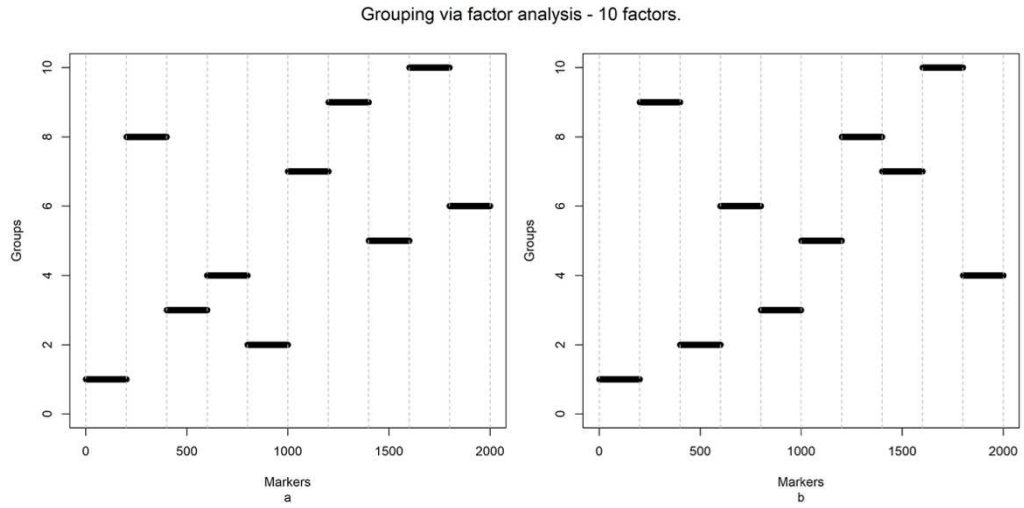


Figure 7: Groups of factorial complexes obtained through the analysis of factors with varimax rotation, considering 10 factors. a - Using the correlation matrix and b - Using the matrix A established as a function of distances genetics.

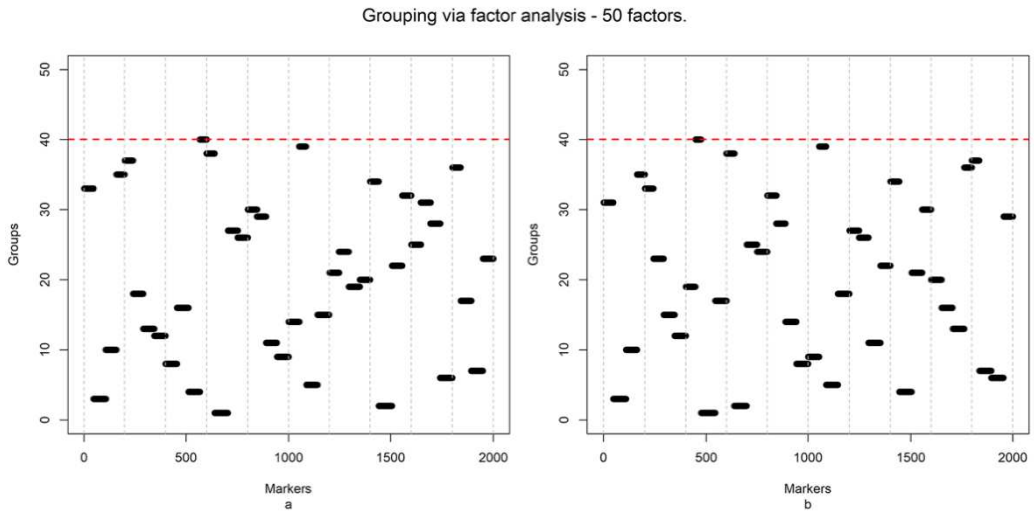


Figure 8: Groups of factorial complexes obtained through the analysis of factors with varimax rotation, considering 50 factors. a - Using the correlation matrix and b - using the matrix A established as a function of distance genetics.

In figure 7, we have the factorial complexes considering ten factors, which can be interpreted as being representative of the linkage groups. In figure 3, these groups are also formed through the construction of a genetic map, which shows the technique's ability to

reproduce this result without the need for genomic studies and, therefore, can be applied with any type of population.

It is also possible to establish patterns of disequilibrium within the same linkage group (HALLDÓRSSON et al., 2004; JASIELCZUK et al., 2020; PATIL et al., 2001). In figure 8, the results are presented in the process of grouping via AF, using 50 factors. The factorial complexes have different interpretations and are now representative of groups of disequilibrium. These groups are partitions of a connection group in which the intensity of the disequilibrium is more intense than that observed in the genetic maps, in which, by transitive property, if A is connected to B and B is connected to C, then A and C are linked even if the distance is equal to or greater than 50 cM, so that A and C can be declared bound, but not necessarily nongame tic disequilibrium.

A technique that allows identifying groups of disequilibrium is more advantageous when the interest is to guide sampling for the purpose of reducing dimensionality. Many conventional techniques to separate the bonding groups perform the tests by considering two marks (loci) at a time (CARNEIRO; VIEIRA, 2002). In view of the data size of SNPs chips, this process can become unfeasible, since it would be necessary to test $\frac{n!}{2!(n-2)!}$ combinations. Using the proposed technique is the result of the haplotypes blocks formed by the markers that are in disequilibrium without the need to test all combinations of two brands.

It is noteworthy that, despite the use of 50 factors, only 42 groups were formed, and the groups 41 and 42 contain only 10 markers when we used the correlation matrix. Using the distance matrix, 40 groups were formed.

Reducing the number of factors has resulted in the reduction of communalities (PREACHER; MACCALLUM, 2002). This can be seen in Figure 9, where when using 10 factors, the highest values were 0.60, while when using 50, values were higher than 0.85. This result showed that dividing the markers into groups of disequilibrium generates a better representation of the relationship structure between the SNPs, compared to the division according to the linkage groups.

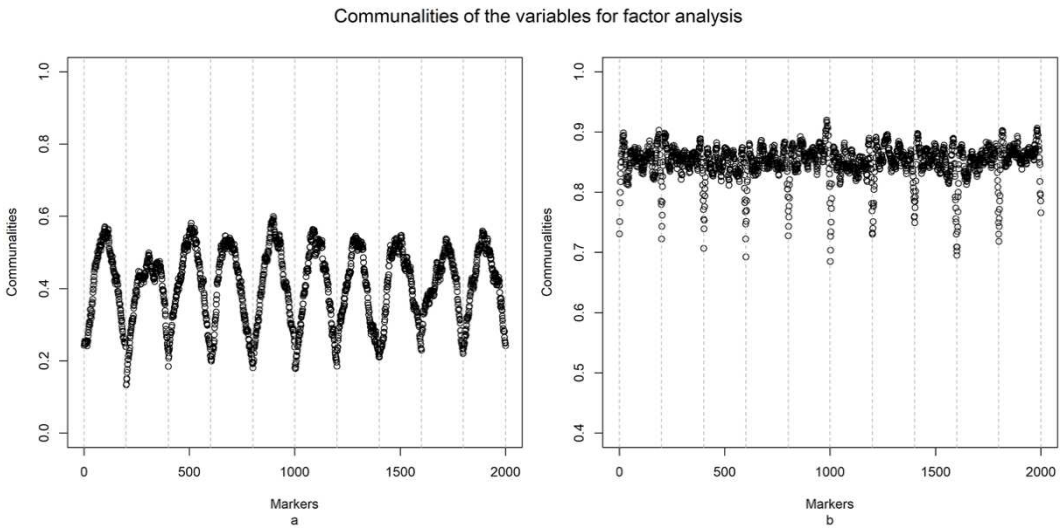


Figure 9: Communalities of the variables for factor analysis, obtained through the correlation matrix. a) Using 10 factors and b) using 50 factors.

CONCLUSION

Recognizing grouping patterns of a set of markers that reflect linkage groups or disequilibrium groups is important for guiding sampling with a view to reduce dimensionality. The study demonstrated that the use of factor analysis is a viable alternative to finality.

Defining the best number of factors (k) is a challenge, since there are several methodologies available in the literature, which all lead us to different results. Therefore, using the average communality to assist in the use of the best number of factors can be efficient.

The factor analysis used for data with high dimensionality, in which the number of variables is higher than the number of individuals, was able to synthesize the degree of association between pairs of markers, identifying subgroups of markers that reflect factor binding groups and linkage disequilibrium groups.

REFERENCES

BORÉM, A.; CAIXETA, E. **Marcadores Moleculares**. Viçosa: UFV, 2016. v. 1

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, n. spe, p. 64–71, jul. 2009.

CARNEIRO, M. S.; VIEIRA, M. L. C. Mapas genéticos em plantas. **Bragantia**, v. 61, n. 2, p. 89–100, ago. 2002.

CATTELL, R. B. The Scree Test For The Number Of Factors. **Multivariate Behavioral Research**, v. 1, n. 2, p. 245–276, abr. 1966.

COLLARD, B. C. Y. et al. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. **Euphytica**, v. 142, n. 1–2, p. 169–196, jan. 2005.

COSTELLO, A. B.; OSBORNE, J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. 2005.

CRUZ, C. Genes software – Extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum Agronomy**, v. 38, p. 547–552, 1 out. 2016.

DALY, M. J. et al. High-resolution haplotype structure in the human genome. **Nature Genetics**, v. 29, n. 2, p. 229–232, out. 2001.

DEON VILELA DE RESENDE, M.; FONSECA E SILVA, F.; FERREIRA AZEVEDO, C. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão, aleatória, seleção genômica, QTL, GWAS, estatística espacial e temporal, competição, sobrevivência.** Viçosa: UFV, 2014. v. 1

DHARIWAL, R. et al. High Density Single Nucleotide Polymorphism (SNP) Mapping and Quantitative Trait Loci (QTL) Analysis in a Biparental Spring Triticale Population Localized Major and Minor Effect Fusarium Head Blight Resistance and Associated Traits QTL. **Genes**, v. 9, n. 1, p. 19, 5 jan. 2018.

GABRIEL, S. B. et al. The Structure of Haplotype Blocks in the Human Genome. **Science**, v. 296, n. 5576, p. 2225–2229, 21 jun. 2002.

GHOLAMI, M. et al. Genome Scan for Selection in Structured Layer Chicken Populations Exploiting Linkage Disequilibrium Information. **PLOS ONE**, v. 10, n. 7, p. e0130497, 7 jul. 2015.

GREENSPAN, G.; GEIGER, D. Model-Based Inference of Haplotype Block Variation. **Journal of Computational Biology**, v. 11, n. 2–3, p. 493–504, mar. 2004.

HAIR, J. F. (ED.). **Multivariate data analysis**. 7th ed ed. Upper Saddle River, NJ: Prentice Hall, 2010.

HALLDÓRSSON, B. V. et al. Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies. **Genome Research**, v. 14, n. 8, p. 1633–1640, ago. 2004.

JASIELCZUK, I. et al. Comparison of linkage disequilibrium, effective population size and haplotype blocks in Polish Landrace and Polish native pig populations. **Livestock Science**, v. 231, p. 103887, jan. 2020.

KUMAR, V. et al. Genome-wide association mapping of salinity tolerance in rice (Oryza sativa). **DNA Research**, v. 22, n. 2, p. 133–145, 1 abr. 2015.

LEDOIT, O.; WOLF, M. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. **The Annals of Statistics**, v. 30, n. 4, 1 ago. 2002.

LI, G. et al. Genome-Wide Association Mapping Reveals Novel QTL for Seedling Leaf Rust Resistance in a Worldwide Collection of Winter Wheat. **The Plant Genome**, v. 9, n. 3, nov. 2016.

LIU, H. et al. **A Neural Network Based on Rough Set (RSNN) for Prediction of Solitary Pulmonary Nodules**. 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing. **Anais**... Em: 2009 INTERNATIONAL JOINT CONFERENCE ON BIOINFORMATICS, SYSTEMS BIOLOGY AND INTELLIGENT COMPUTING. Shanghai, China: IEEE, 2009. Disponível em: <http://ieeexplore.ieee.org/document/5260720/>. Acesso em: 20 ago. 2022

LOEHLIN, J. C. **Latent variable models: An introduction to factor, path, and structural equation analysis**. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2004. p. xi, 317

MCRAE, A. F. et al. Linkage Disequilibrium in Domestic Sheep. **Genetics**, v. 160, n. 3, p. 1113–1122, 1 mar. 2002.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers**, v. 6, n. 1, p. 6–14, 1 jan. 2016.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada - Uma Abordagem Aplicada**. Belo Horizonte: Editora UFMG, 2005.

NADEEM, M. A. et al. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. **Biotechnology & Biotechnological Equipment**, v. 32, n. 2, p. 261–285, 4 mar. 2018.

PALLANT, J. **SPSS survival manual: a step by step guide to data analysis using IBM SPSS**. 6th edition ed. Maidenhead New York: McGraw Hill Education, 2016.

PATIL, N. et al. Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. **Science**, v. 294, n. 5547, p. 1719–1723, 23 nov. 2001.

PREACHER, K. J.; MACCALLUM, R. C. [No title found]. **Behavior Genetics**, v. 32, n. 2, p. 153–161, 2002.

PRITCHARD, J. K.; PRZEWORSKI, M. Linkage Disequilibrium in Humans: Models and Data. **The American Journal of Human Genetics**, v. 69, n. 1, p. 1–14, jul. 2001.

R CORE TEAM. **R: A Language and environment for Statistical Computing**. Available from: <https://www.r-project.org/>.

REICH, D. E. et al. Linkage disequilibrium in the human genome. **Nature**, v. 411, n. 6834, p. 199–204, maio 2001.

SHIFMAN, S. Linkage disequilibrium patterns of the human genome across populations. **Human Molecular Genetics**, v. 12, n. 7, p. 771–776, 1 abr. 2003.

SPINDEL, J. et al. Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. **PLOS Genetics**, v. 11, n. 2, p. e1004982, 17 fev. 2015.

TERRA, T. G. R. et al. QTLs identification for characteristics of the root system in upland rice through DNA microarray. **Acta Scientiarum. Agronomy**, v. 38, n. 4, p. 457, 2 set. 2016.

WANG, N. et al. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. **The American Journal of Human Genetics**, v. 71, n. 5, p. 1227–1234, nov. 2002.

CAPÍTULO 2

Análise de fatores para redução de dimensionalidade em estudos de predição genômica

RESUMO

OLIVEIRA, Cristiano Ferreira, D.Sc., Universidade Federal de Viçosa, dezembro de 2022. **Análise de fatores para redução de dimensionalidade em estudos de predição genômica**. Orientador: Cosme Damião Cruz. Coorientadores: Moysés Nascimento.

O desequilíbrio de ligação (LD) é um conceito chave para a seleção genômica. A forma com que um fenótipo é expresso está ligado a variações genéticas, que por sua vez levam a muitas associações estatísticas entre marcadores em LD com locos de características quantitativas (QTLs). Em geral não é conhecido todos os locos de QTLs para determinada característica, além de existir muitas propostas diferentes que visam construir grupos de LD. Consequentemente, construir modelos preditivos com bom desempenho se torna um grande desafio devido ao grande número de marcas e as associações não conhecidas entre marcadores e fenótipo. Muitas abordagens diferentes têm sido propostas a fim de selecionar marcadores com o intuito de aumentar o desempenho de modelos preditivos, de forma geral são elas selecionar marcadores relatados em estudos de associação genômica entre SNPs e o fenótipo de interesse, estimar a significância dos efeitos das marcas para um modelo preditivo ou selecionar de forma espaçada os marcadores ao longo do genoma. As duas primeiras abordagens citadas têm como principal desvantagem estarem relacionadas ao fenótipo a ser avaliado, ou seja, para cada fenótipo deverá ser feito um novo estudo a fim de fazer uma seleção de marcadores adequada aquele fenótipo. Já a última abordagem, possui como grande vantagem, produzir painéis de baixa densidade de marcadores que poderão ser utilizados para predição de valor genético de qualquer fenótipo de interesse. Por outro lado, praticar a seleção espaçada pode remover blocos inteiros de marcadores em LD, o que poderá impactar negativamente o desempenho da seleção se o bloco contiver marcadores em desequilíbrio com QTLs. Assim este estudo foi proposto com o objetivo de conduzir uma seleção espaçada orientada por blocos de desequilíbrio. A Análise de Fatores (AF) foi usada para construir os blocos de haplótipos e sintetizar a relação linear entre marcadores e criar fatores comuns interpretados como blocos de LD. A fim de garantir a permanência de todos os grupos de LD após a seleção utilizou-se a seleção espaçada dentro dos blocos formados a partir da AF. Utilizamos um conjunto de dados de soja contendo 41985 marcadores do tipo SNPs com informação de 20087 acessos de soja. Três painéis de SNPs foram considerados, contendo 1% das marcas por bloco, 5% e 100% (painel completo). Para avaliar o êxito da proposta,

mensuramos o impacto da redução do painel de SNPs na acurácia seletiva e na acurácia preditiva de dois modelos de *machine learning*, *Lasso* e *Random Forest* em uma tarefa de predição dos valores dos fenótipos peso de 100 sementes, altura da planta e rendimento de sementes. Os painéis de baixa densidade foram capazes de alcançar o mesmo desempenho de acurácia seletiva que painel completo. Não foi encontrada diferença significativa de acurácia seletiva ao utilizar os painéis reduzidos ou o painel completo e para uma característica não foi encontrado diferença significativa de acurácia preditiva.

Palavras-chave: SNP. GWS. Seleção de Mrcadores. Análise Fatorial. Soja. Aprendizado de Máquina. Blocos de Haplótipos.

ABSTRACT

OLIVEIRA, Cristiano Ferreira, D.Sc., Universidade Federal de Viçosa, march, 2022. **Factor Analysis for dimensionality reduction in genomic prediction studies**. Adviser: Cosme Damião Cruz. Co-adviser: Moysés Nascimento.

Linkage disequilibrium (LD) is a key concept for genomic selection. The way a phenotype is expressed is linked to genetic variations, which in turn lead to many statistical associations between markers in LD with quantitative trait loci (QTLs). In general, not all QTL loci are known for a given trait, and there are many different proposals that aim to build LD groups. Consequently, building predictive models with good performance becomes a major challenge due to many markers and the unknown associations between markers and phenotype. Many different approaches have been proposed to select markers to increase the performance of predictive models, in general, they are to select markers reported in genomic association studies between SNPs and the phenotype of interest, to estimate the significance of the effects of the markers for a predictive model or to select the markers in a spaced way along the genome. The main disadvantage of the first two approaches mentioned is that they are related to the phenotype to be evaluated, that is, for each phenotype a new study must be carried out to select the appropriate markers for that phenotype. The last approach, on the other hand, has the great advantage of producing panels of low-density of markers that can be used to predict the genetic value of any phenotype of interest. But, practicing spaced selection can remove entire blocks of markers in LD, which can negatively impact selection performance if the block contains markers in linkage disequilibrium with QTLs. Thus, this study was proposed with the objective of conducting a spaced selection guided by blocks of disequilibrium. Factor Analysis (FA) was used to build haplotype blocks and synthesize the linear relationship between markers and create common factors interpreted as LD blocks. To guarantee the permanence of all LD groups after selection, a spaced selection was made in each LD block. We used a soybean dataset containing 41985 SNPs markers with information from 20087 soybean plants. Three panels of SNPs were considered, containing 1% of the marks per block, 5%, and 100% (full panel). To evaluate the success of the proposal, we measured the impact of the reduction of the SNP panel on the selective accuracy and on the predictive accuracy of two machine learning models, Lasso and Random Forest, in a task to predict the values of the phenotypes weight of 100 seeds, height of plant and seed yield. Low-density panels were able to achieve the same selective accuracy

performance as a full panel. No significant difference in selective accuracy was found when using the reduced panels or the full panel.

Keywords: Dissertation. SNP. GW. Marker Selection. Factor Analysis. Soybean. Machine Learning. Haplotype Blocks.

INTRODUÇÃO

A soja (*Glycine max*) é originada de clima temperado, com ampla adaptação aos climas subtropicais e tropicais, é considerada uma das mais importantes leguminosas, com produção mundial de safra de 2021/22 estimada em aproximadamente 380 milhões de toneladas (USDA-FAS, https://apps.fas.usda.gov/). A soja contém cerca de 40% de proteínas e 20% de óleo nas sementes. Rica em proteínas e óleo, é uma cultura economicamente importante para a alimentação animal e a produção de óleo e produtos alimentares. Isso tem incentivado os melhoristas a desenvolverem cultivares com maior qualidade, resistência a pragas e doenças e com alta produtividade.

Nas últimas décadas verifica-se que houve aumento significativo na pesquisa de soja enfatizando o uso de marcadores genéticos para fins de predição. Mapas genético molecular de apenas algumas centenas de marcadores RFLP cresceu para milhares de locos englobando marcadores RFLP, RAPD, SSR e SNP.

O uso dos marcadores moleculares tornou possível o desenvolvimento de técnicas de seleção de indivíduos utilizando marcadores, como a seleção genômica ampla (*Genome Selection – GS*) proposta por Meuwissen et al. (2001). Uma importante pressuposição para estudos de seleção genômica é o desequilíbrio de ligação (LD) entre marcadores moleculares e locos controladores de características quantitativas (QTL), e entre marcadores.

A análise de fatores (AF) é geralmente utilizada em dados cujo número de observações é maior que o número de variáveis (COSTELLO et al., 2005; HAIR, 2010). A técnica vem sendo utilizada no melhoramento genético em dados fenotípicos com o objetivo de criar variáveis latentes (fatores comuns) que possam representar um conjunto de fenótipos sem perda de significância biológica (PAIXÃO et al., 2022; MAZZA et al., 2016; TEIXEIRA et al., 2016). OLIVEIRA et al., 2021 utilizaram dados simulados e verificaram que a técnica pode ser utilizada para sintetizar a estrutura de blocos de LD utilizando a correlação entre marcadores como medida auxiliar.

Em estudos de seleção genômica, uma população de treinamento é genotipada e as características de interesse registradas para treinar um modelo de predição. O efeito de cada polimorfismo de nucleotídeo único (SNP) é estimado. Em etapa posterior, outros candidatos à seleção são genotipados e, utilizando os efeitos dos SNP estimados, o valor genômico estimado é obtido para os candidatos que são comparados. Isto permite fazer uso apenas daqueles

indivíduos que apresentam maior potencial genético. Realizadas as tarefas de ajuste de modelo (treinamento) e uso em validação, os pesquisadores chegaram à conclusão de que esta abordagem de GS é útil no melhoramento genético, pois proporciona alta eficiência seletiva, rapidez na obtenção de ganhos genéticos com a seleção e baixo custo, em comparação com a seleção baseada em dados fenotípicos, principalmente com características que não podem ser medidas diretamente nos candidatos à seleção (JANNINK *et al.*, 2010).

Com o objetivo de aumentar a acurácia dos modelos de predição, alguns métodos de seleção (ou descarte) de marcadores vêm sendo propostos (VALLEJO et al., 2018). Muitos trabalhos foram feitos com o objetivo de identificar subconjuntos de SNPs que possam fornecer alta precisão de previsão genômica de valores genéticos (AKBARZADEH et al., 2021; HABIER; FERNANDO; DEKKERS, 2009; LI et al., 2018; OGAWA et al., 2014). São estratégias muito utilizadas para este fim, considerar os SNPs mais significativos, identificados pela construção do GWAS no conjunto de dados para cada característica (AKARZADEH, DEHKORDI, *et al.*, 2021), ou a seleção subconjuntos dos marcadores uniformemente espaçados ao longo do genoma para previsão genômica (HABIER; FERNANDO; DEKKERS, 2009; LI et al., 2018; OGAWA et al., 2014).

Considerar SNPs identificados pela construção do GWAS conduz ao uso de subconjuntos de SNPs diferentes para diferentes características. Ao utilizar a seleção espaçada como HABIER, FERNANDO, *et al.*, 2009 este problema é superado uma vez que o mesmo painel de marcadores resultantes deste processo de seleção pode ser utilizado para qualquer característica. Porém ao utilizar a seleção espaçada é possível que sejam eliminados blocos de haplótipos em desequilíbrio de ligação nos quais encontram-se SNPs importantes para predizer determinada característica.

Assim, este estudo foi feito com o objetivo de propor o uso da análise de fatores para formar blocos de LD e utilizá-los em um processo de seleção de marcadores uniformemente espaçados dentro de grupos de marcadores altamente correlacionados, entendidos como grupos de marcadores em desequilíbrio de ligação.

MATERIAL

Foi utilizado o banco de dados SoySNP50K iSelect BeadChip (SONG et al., 2013) disponível em https://www.soybase.org. Ele possui 41985 marcadores com informação de 20087 acessos de soja cultivada e selvagem.

Várias características relacionadas aos acessos genotipados também estão disponíveis no repositório. Dentre elas foram utilizadas altura da planta (PH), peso de 100 sementes (SW) e rendimento de sementes (SY) contendo informações de 16275, 16283 e 15976 acessos, respectivamente.

MÉTODOS

Pré-processamento

O pré-processamento dos dados foi feito seguindo as seguintes etapas

a) *Call rate* e *maf*

Foram eliminados marcadores com proporção de genótipos por marcador com dados não ausentes (*Call rate*) maior que 95% e menor frequência alélica (*maf*) inferior a 5%. Esta etapa descartou mais de 5 mil marcas, restando 36371 SNPs.

b) Formação de grupos de desequilíbrio de ligação

O desequilíbrio de ligação (LD) refere-se à não-independência de alelos em diferentes locos. Os padrões de LD podem ser usados para estudar a estrutura e padrões da variação do SNP no DNA (DALY et al., 2001; GABRIEL et al., 2002). Para medir o LD entre pares de SNP é possível utilizar medidas auxiliares tais como os valores das estatísticas r ou r^2 que quantificam o desequilíbrio dentro de uma escala que varia de 0 a 1 (DALY et al., 2001; PATIL et al., 2001; REICH et al., 2001; SHIFMAN, 2003; WANG et al., 2002). Desta forma, os blocos de haplótipos (grupos de desequilíbrio de ligação) são vistos como grupos de marcadores altamente correlacionados.

A fim de inferir sobre quantos e quais seriam os grupos de LD utilizou-se a análise de fatores (AF). A AF pode ser utilizada para examinar padrões reconhecendo grupos em que as correlações entre as variáveis sejam retidas dentro de cada grupo (OLIVEIRA et al., 2021).

Os marcadores moleculares representam as variáveis no modelo fatorial. As variáveis são representadas como uma função linear de variáveis ou fatores comuns, não observáveis, e pelo erro aleatório específico a cada marcador. Com a análise fatorial é possível investigar se

variáveis de interesse X_1, X_2, ..., X_p, estão linearmente relacionados a fatores não observáveis F_1, F2, ..., Fk, com k≤p.

Para estimar as cargas fatoriais (loadings) da análise de fatores utilizou-se a matriz de correlação entre marcas. O número de fatores (k) utilizado foi estabelecido com base no critério da proporção da variância total. Este critério é uma abordagem baseada na obtenção de uma porcentagem cumulativa especificada da variância total extraída por fatores sucessivos. O objetivo é garantir que os fatores derivados expliquem pelo menos uma quantidade especificada de variância (HAIR, 2014).

Desta forma, dada a matriz de correlação R, e $\lambda_1, \lambda_2, \dots, \lambda_p$ os autovalores de R em ordem decrescente, considere $p_1, p_2, \dots, p_n$ tal que $p_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$ e considere também a soma acumulada $s_k = \sum_{i=i}^{k} p_i$. Temos então que k é tal que s_k seja igual a proporção da variância total explicada desejada.

O número k de fatores foi definido de forma que s_k fosse aproximadamente 70%. A formação dos grupos foi realizada e estabelecida por meio de um processo iterativo, tendo como critério que variáveis cuja maior carga fatorial se dava para o i-ésimo fator seriam alocadas no grupo i.

A rotação ortogonal Varimax dos fatores originais foi utilizada com o objetivo de tornar mais fácil a interpretação dos fatores.

c) Imputação dos dados

Os grupos formados a partir da AF foram utilizados no processo de imputação. Assim para cada acesso o valor ausente do marcador foi imputado considerando a moda dos valores das marcas pertencentes ao mesmo grupo de LD.

Redução de dimensionalidade

Utilizar painéis com SNPs uniformemente espaçados traz vantagens em relação a outros métodos de seleção de marcadores. O desempenho deste método é independente do número de QTLs que afetam a característica e dos métodos usados para estimar efeitos dos marcadores nos dados de treinamento (HABIER; FERNANDO; DEKKERS, 2009; OGAWA et al., 2014).

O método proposto utiliza a seleção espaçada de marcadores dentro dos grupos de desequilíbrio. Assim foram selecionadas $p = 1\%$ e $p = 5\%$ dos SNPs dentro de cada grupo de LD formando dois painéis reduzidos de marcas. A seleção foi feita considerando a seguinte regra,

- a_k: número de marcadores no grupo k.
- n: número de marcadores a serem selecionados no grupo k dado pelo inteiro mais próximo e menor que o produto de a_k *e* p e $n \geq 1$.
- a_1: posição do primeiro marcador do grupo k, ou seja 0.
- r: razão da progressão dada pelo inteiro mais próximo e menor que $\frac{a_k - a_1}{n-1}$.

a sequência de SNPs selecionadas no grupo *k* é dada por:

$\{s_k\} = a_1, a_1 + r, a_1 + 2r, \ldots, a_1 + (n-1)\, r$ se $n >= 2$

Se 0<n<2, ou seja, apenas um marcador deve ser selecionado no grupo, considere o marcador na posição central.

Predição de valores genéticos

Para fazer a predição dos valores genéticos utilizou-se dois modelos, o *Lasso* e *Random Forest*.

Dado N observações de uma variável de resultado y_i e p variáveis preditoras associadas $x_i = (x_{i1}, \ldots, x_{ip})^T$, um dos principais objetivos do aprendizado de máquina é fazer a predição da variável resposta y usando as preditoras.

Um modelo de regressão linear assume que:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$$

onde β_0 e $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ são parâmetros desconhecidos e e_i é um termo de erro. O método dos mínimos quadrados estima os parâmetros através da minimização da função objetivo dos mínimos quadrados

$$\min_{\beta_0, \beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_{ij} \right)^2$$

Em geral, os valores estimados para os parâmetros são não nulos, assim em cenários com muitas variáveis preditoras, ou seja, p grande, torna-se difícil fazer a interpretação do modelo e a multicolinearidade nas variáveis de entrada acarreta altos valores para variância dos parâmetros do modelo, comprometendo sua eficiência.

Assim, há a necessidade de restringir ou regularizar o processo de estimativa dos parâmetros do modelo. Na regressão lasso (HASTIE; TIBSHIRANI; WAINWRIGHT, 2020; TIBSHIRANI, 1996) ou regularização l_1, estimamos os parâmetros através da solução do problema

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_{ij} \right)^2 \; sujeito\ a\ \|\beta\|_1 \leq \alpha$$

onde $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ é a norma l1 de β, e $\alpha > 0$ é um parâmetro especificado pelo usuário. Assim, o Lasso minimiza a soma de quadrados do resíduo, desde que a soma do valor absoluto dos coeficientes seja menor que uma constante. Devido à natureza dessa restrição, ela tende a produzir alguns coeficientes que são exatamente 0.

Além de métodos de regressão penalizada, abordagens baseadas em árvores de decisão (BREIMAN, 1984) como *Random Forest* (BREIMAN, 2001) podem ser utilizadas para lidar com problemas de alta dimensão (CHEN; ISHWARAN, 2012).

Random Forest, ou floresta aleatória, é um algoritmo de aprendizado de máquina supervisionado, baseado em aprendizado de conjunto (*ensemble learning*), que pode ser utilizado em problemas de regressão e classificação. Ele constrói várias árvores de decisão e as agrega a fim de obter um resultado preciso.

Ao construir uma árvore de decisão temos forte dependência dos dados de treino. Se os mesmos dados forem fornecidos, a mesma árvore será produzida todas as vezes. Árvores de decisão têm uma tendência ao *overfit*, pois buscam a melhor árvore possível com os dados fornecidos, mas podem falhar na generalização quando dados não observados são fornecidos.

As árvores de decisão que são combinadas em uma *Random Forest* são diferentes, pois cada árvore é construída a partir de um subconjunto aleatório diferente dos dados. Considerando os dados *D* e *a* sendo uma única árvore da floresta, a previsão de uma *Random Forest* é feita da seguinte forma.

1. É criada uma amostra *bootstrap* com substituição em D, obtendo X_a e Y_a.
2. A árvore f_a a é treinada utilizando X_a e Y_a.
3. A previsão final $\hat{p}$ em problemas de regressão é obtida pela média das previsões de cada árvore, ou seja:

$$\hat{p} = \frac{1}{N} \sum_{a=1}^{A} f_a(x)$$

onde N é o número total de árvores na floresta aleatória, *a=1* representa a primeira árvore e *A* a última árvore e $f_a(x)$ representa a previsão da a-ésima árvore.

Treinamento dos modelos e otimização de hiperparâmetros

Com o conjunto de dados genotípicos definido, foram selecionados os indivíduos (acessos) com fenótipo registrado e então combinado as informações genotípicas e fenotípicas. Foram avaliadas as características peso de 100 sementes, altura da planta e rendimento de sementes. O conjunto de dados foi dividido em duas partes, conjunto de treino (compreendendo 80% dos acessos) e conjunto de testes (contendo 20% dos acessos).

Um grande desafio ao trabalhar com algoritmos de aprendizado de máquina está na definição de seus hiperparâmetros. Os hiperparâmetros possuem grande importância para estes algoritmos, pois controlam seus comportamentos no processo de treinamento e afeta o desempenho dos modelos. A otimização Bayesiana é uma estratégia poderosa que pode ser utilizada para obter os melhores hiperparâmetros (SNOEK; LAROCHELLE; ADAMS, 2012; WU et al., 2019)

Então, dado o modelo φ (Lasso ou *Random Forest*), foi feita a busca pelos melhores hiperparâmetros utilizando otimização Bayesiana. A seguintes etapas foram realizadas para este fim:

I. Definição do espaço Θ de hiperparâmetros do modelo: dado o modelo a ser utilizado os hiperparâmetros são considerados e então o intervalo no qual deseja-se avaliar cada um deles é definido.
II. É definida a função objetivo $\phi(\boldsymbol{\theta}, \varphi(\boldsymbol{\theta}|\boldsymbol{X}, y))$, a qual retorna a média da raiz quadrada do erro (*mean squared error)* em 3 folds de um processo de validação cruzada.
III. Encontrar a combinação de hiperparâmetros ótima utilizando os dados de treino: Estamos interessados em resolver $\boldsymbol{\theta}^* = arg\min_{\theta} \phi(\boldsymbol{\theta})$. $\boldsymbol{\theta}^*$é obtido por meio de um processo de otimização Bayesiana usando Processos Gaussianos (GP) (BROCHU; CORA; DE FREITAS, 2010). Para este processo de busca de parâmetros foi utilizado a função gp_minimize da biblioteca scikit-optimize (HEAD et al., 2018).

Após obter $\boldsymbol{\theta}^*$, o modelo $\varphi(\boldsymbol{X}, y|\boldsymbol{\theta}^*)$ é ajustado com os dados de treinamento.

Esta etapa de ajuste e otimização dos modelos foi feita utilizando a linguagem python, sendo o pacote scikit-learn (PEDREGOSA et al., 2011) utilizado para obter os modelos e o pacote scikit-optimize (HEAD et al., 2018) para otimizar os hiperparâmetros. O Lasso possui o hiperparâmetro α para ser otimizado, que como vimos anteriormente está ligado a regularização praticada pelo modelo.

R*andom forest* possui diversos hiperparâmetros que podem ser otimizados, porém foram escolhidos *min_sample_leaf* que se refere ao número mínimo de amostras necessárias para estar em um nó folha, *n_estimators* que corresponde ao número de árvores na floresta e *min_samples_split*. Um ponto de divisão em qualquer profundidade só será considerado se deixar pelo menos *min_samples_split* amostras de treinamento em cada um dos ramos esquerdo e direito.

Foi considerado no processo de otimização α real e contido no intervalo *(0, 1), min_samples_leaf* inteiro contido no intervalo *(1,10)* e *n_estimators* inteiro pertencente ao intervalo *(100, 500)* e *min_samples_split* inteiro no intervalo *(2, 100).*

Divisão dos dados em treinamento e teste

O conjunto de dados foi divido em duas partes. Uma para treinamento e ajuste dos modelos contendo 80% dos acessos e outra contendo 20% dos acessos (população de teste). Portanto, dado um fenótipo e considerando todos os acessos os quais possuem informações referentes ao fenótipo, estes acessos foram divididos em dois grupos, treinamento e teste, de forma aleatória.

Os dados treinamento foram utilizados em 2 etapas, primeiramente otimização dos modelos utilizando otimização bayesiana para encontrar a melhor combinação de hiperparâmetros e em seguida treinamento dos modelos com todos os dados de treinamento utilizando os hiperparâmetros encontrados na primeira etapa.

Para comparar o desempenho dos modelos os dados de teste foram divididos em 20 amostras. Os modelos (Lasso e Random Forest) foram avaliados em 20 repetições utilizando os painéis de SNPs contendo 1%, 5% e 100% de marcas por bloco de haplótipo. O processo de dividir os dados de teste em amostras foi feito a fim de obter um intervalo de confiança para as medidas de acurácia. Na Tabela 1 é apresentado para cada fenótipo o número de informações disponíveis e o número de acessos destinado para treinamento, teste e tamanho das amostras.

Tabela 1: número de observações disponíveis para cada fenótipo.

	Height	Yield	Seed Weight
Total	16275	16283	15976
Treino	13020	13026	12781
Teste	3255	3257	3195
Tamanho de amostra	163	163	160

Acurácia dos modelos

Foram utilizadas duas métricas a fim de avaliar os modelos o erro quadrático médio, $mse = \frac{\sum_{i=1}^{n}(\hat{y}-y_i)^2}{n}$ o qual está ligado a capacidade preditiva do modelo e coeficiente de correlação $r^2_{y,\hat{y}} = \frac{cov(y,\hat{y})}{\sqrt{var(y)var(\hat{y})}}$ que está ligado a capacidade seletiva. Menores valores de mse indica maior acurácia preditiva, ou seja, o modelo é capaz de prever com maior precisão os valores do fenótipo. Já para a acurácia seletiva, maiores valores de correlação indicam maior acurácia seletiva.

O processo de otimização bayesiana considerou a média do mse oriundo de um processo de validação cruzada de 3 folds como métrica para avaliação dos modelos da função objetivo definida para o processo de otimização. Após encontrar a melhor combinação de hiperparâmetros os modelos foram treinados utilizando todos os indivíduos contidos no conjunto de treino concluindo o processo de treinamento dos modelos.

Para avaliar o desempenho dos modelos com os dados de teste foram consideradas as medidas de acurácia mse e r^2. As medidas foram calculadas para cada amostra, característica e segundo o percentual de SNPs por bloco utilizado como variável de entrada do modelo (Figura 1G).

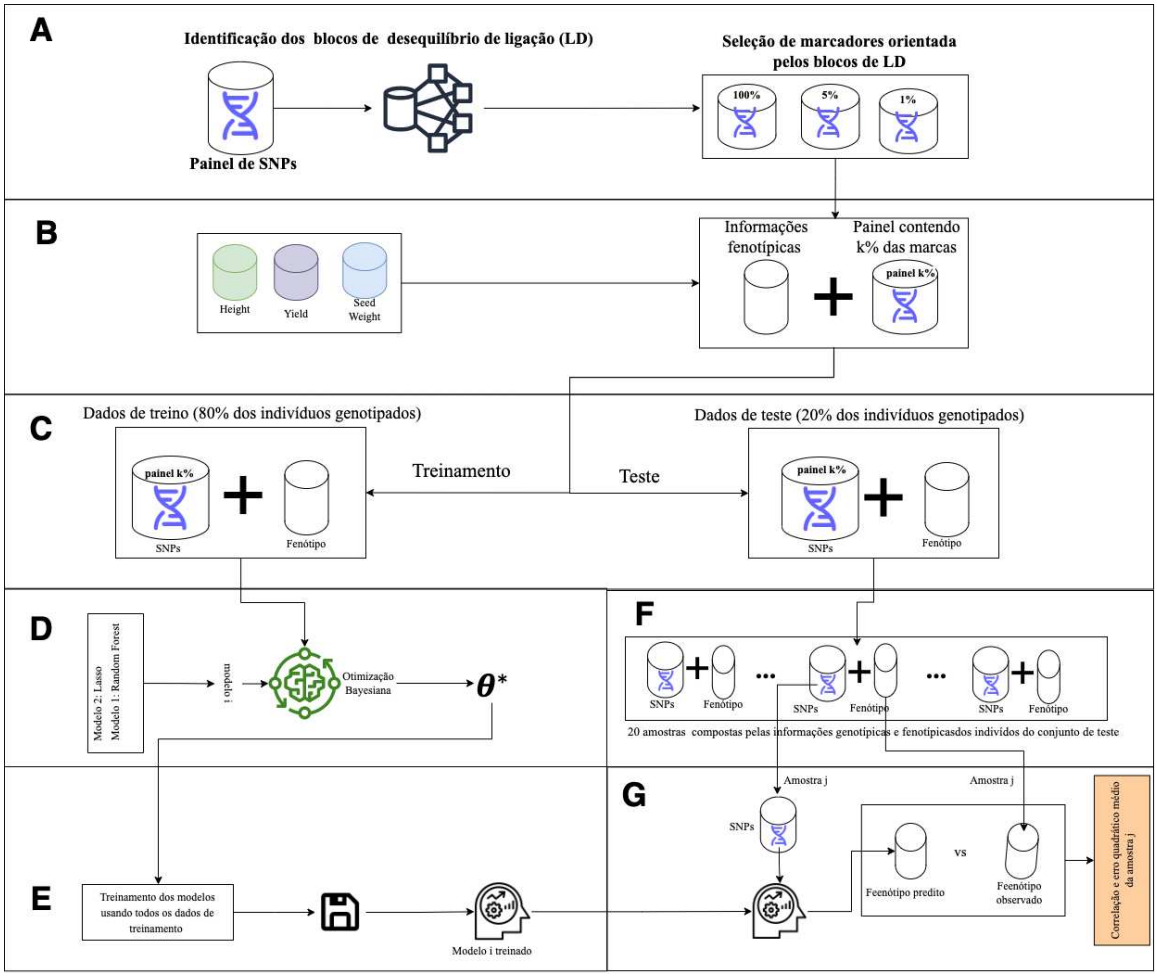


Figura 1: Esquema ilustrativo do processo de seleção de marcas e avaliação dos modelos. **A**: processo de seleção dos marcadores, **B**: adição das informações fenotípicas, **C**: divisão dos dados em treinamento e teste, **D**: busca dos melhores hiperparâmetros $\boldsymbol{\theta}^*$, **E**: treinamento dos modelos utilizando $\boldsymbol{\theta}^*$ e todos os dados de treino, **F**: divisão dos dados de teste em 20 amostras e **G**: Avaliação dos modelos salvos na etapa **E** utilizando as amostras dos dados de teste como ilustrado em F.

Recursos computacionais

Amazon Elastic Compute Cloud (EC2) faz parte da plataforma de *cloud computing* da Amazon, Amazon Web Services (AWS). O EC2 permite que usuários aluguem computadores virtuais nos quais rodam suas próprias aplicações. O EC2 permite a implantação de aplicações escaláveis ao prover um Web service através do qual um usuário pode iniciar uma Amazon Machine Image para criar uma máquina virtual, que a Amazon chama uma "instância",

contendo qualquer software desejado. Um usuário pode criar, lançar e terminar instâncias do servidor, conforme necessário, pagando por hora pelos servidores ativos.

Desta forma foram criadas duas instâncias na AWS, uma m5.4xlarge e uma m5.2xlarge. A m5.4xlarge foi utilizada para realizar o pré-processamento dos dados enquanto a m5.2xlarge foi utilizada para ajustar os modelos.

Os scrips utilizados para realizar todo o processo estão disponíveis no repositório https://github.com/Cristiano2132/SNPs_Selection.

RESULTADOS E DISCUSSÃO

Na Figura 2 pode ser visualizado um padrão de blocos ao considerar os grupos de LD formados utilizando AF e sua posição no cromossomo, resultado similar ao obtido por OLIVEIRA, TEIXEIRA, *et al.*, 2021 em seu trabalho utilizando dados simulados. A formação de grupos e a identificação dos marcadores componentes destes grupos é indicativo da possibilidade de uma seleção ou exclusão de marcadores para estudos futuros de predição.

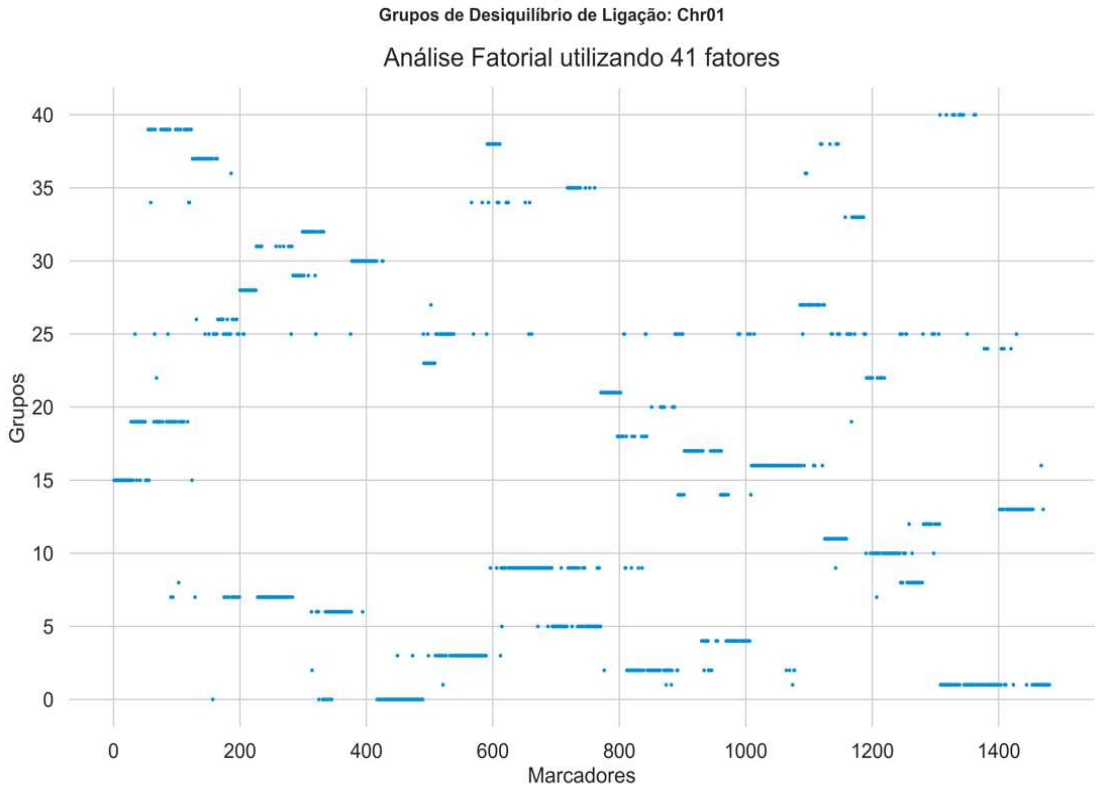


Figura 2: Grupos em desequilíbrio de ligação formados a partir da análise de fatores considerando o cromossomo 1.

Como o eixo X da Figura 2 representa também um ordenamento dos marcadores estudados, tendo por base o mapa genético da soja, constata-se que o desequilíbrio de ligação tende a ser mais forte entre marcadores localizados mais próximos no genoma. Assim padrões similares a blocos de haplótipos podem ser identificados (DALY et al., 2001; OLIVEIRA et al., 2021; PATIL et al., 2001; SHIFMAN, 2003).

O desequilíbrio de ligação é um conceito fundamental para a seleção genômica. É pressuposto que existe desequilíbrio de ligação entre QTL e marcadores, e entre marcadores. Desta forma ao realizar a espaçada dentro de grupos de ligação, garantimos a permanência de marcadores em todos os grupos de LD após a seleção.

Constatada a existência de grupos de LD é possível formular a hipótese de que a predição de valores genéticos poderia ser feita com menos variáveis preditoras (marcadores), com menor tempo computacional e sem reduzir a acurácia em relação ao que seria obtido utilizando-se as informações de todos os marcadores originalmente disponíveis.

Um grande desafio ao trabalhar com modelos de *machine learning* está no custo computacional gasto no processo de otimização de hiperparâmetros (CLAESEN; DE MOOR, 2015). Os hiperparâmetros controlam o comportamento do algoritmo de *machine learning* no treinamento, logo afetam o desempenho do modelo de aprendizado (BROCHU; CORA; DE FREITAS, 2010; SNOEK; LAROCHELLE; ADAMS, 2012; WU et al., 2019). É exibido na Figura 3 o tempo gasto no processo pelos modelos nos diferentes cenários e na Tabela 2 é apresentado os hiperparâmetros encontrados para cada modelo.

Dentre os dois modelos biométricos empregados, o *Lasso* demandou menor custo computacional. NEVES, CARVALHEIRO, et al., 2012 estudaram o desempenho preditivo de dez diferentes métodos estatísticos empregados na seleção genômica e verificaram maior custo computacional do modelo *Random Forest* em relação ao *Lasso*.

Além disso, pode-se observar a redução do tempo gasto no processo de busca de hiperparâmetros ao utilizar painéis de baixa densidade, e esta redução do tempo é percebida para ambos os modelos (Figura 3 A e B).

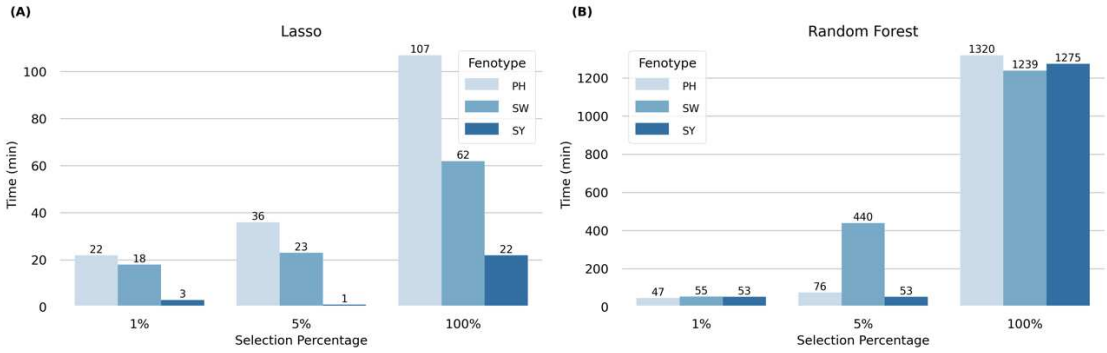


Figura 3: Tempo gasto no processo de otimização e ajuste do modelo de Regressão Lasso (A) e *Random Forest* (B) utilizando cada um dos subconjuntos de marcadores e considerando as variáveis peso de 100 sementes (SW), altura da planta (PH) e rendimento de sementes (SY).

A eficiência da predição com redução de dimensionalidade, orientada pelos resultados da análise de fatores, e sem a redução de dimensionalidade é abordada a seguir. Os dois modelos de predição, com abordagens biométricas distintas, *Lasso* e *Random Forest*, apresentaram bom desempenho ao utilizar os painéis de baixa densidade para predizer todas as características propostas, sendo capaz de obter uma acurácia igual ou próxima a acurácia obtida utilizando o painel completo.

VALLEJO, SILVA, *et al.*, 2018 estudaram precisão das previsões genômicas para resistência a BCWD em trutas arco-íris com painéis SNP de densidade reduzida, e obtiveram bom desempenho dos modelos avaliados ao utilizar aproximadamente 25% dos marcadores, porém em nenhum cenário ultrapassou ou obteve a mesma acurácia dos modelos ao utilizar o painel completo de SNPs.

LI, ZHANG, *et al.*, 2018 avaliaram a eficiência de três métodos baseados em árvores (*Random Forest, Gradient Boosting Machine* (GBM) e *XgBoost*) na identificação de um subconjunto de SNPs e a utilização de painéis de SNP de densidade reduzida desenvolvidos por sub-amostragem de SNPs uniformemente espaçados do painel completo. Foram utilizados 40.184 marcadores SNP de 2.093 bovinos Brahman tropicai. Neste estudo os marcadores uniformemente espaçados não apresentaram bom desempenho. O modelo preditivo apresentou maior acurácia ao utilizar os painéis de baixa densidade obtidos por meio do uso de modelos de *machine learning*. O melhor desempenho observado ocorreu ao utilizar o painel de

aproximadamente 5% dos dados obtido a partir do GBM. Ao utilizar os demais painéis reduzidos o modelo apresentou acurácia inferior em relação à utilização do painel completo.

Nas figuras 4 e 5 é apresentado o desempenho dos modelos de predição com o conjunto de dados de teste. As médias e IC foram obtidos utilizando os valores de correlação e mse oriundos das 20 amostras construídas aleatoriamente a partir dos dados de teste. Os resultados indicam que o painel de baixa densidade é capaz de alcançar o mesmo desempenho em termos de acurácia seletiva (Figura 4).

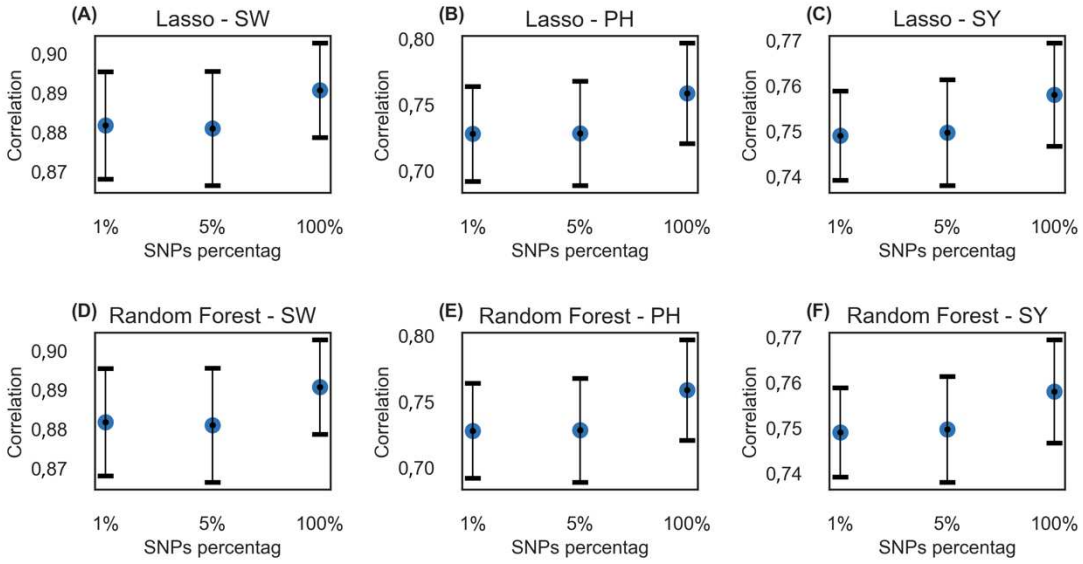


Figura 4: Correlação média e intervalos de confiança dos valores de média da correlação entre valores observados e preditos para as características peso de 100 sementes (SW), altura da planta (PH) e rendimento de sementes (SY). Cada IC e a média são oriundos de 20 repetições.

Para rendimento não houve aumento significativo do erro quadrático médio (Figura 5 C e F) já para altura de planta e peso de semente utilizar os painéis de baixa densidade resultou em aumento significativo do erro quadrático médio (Figura 5 A, B, D e F).

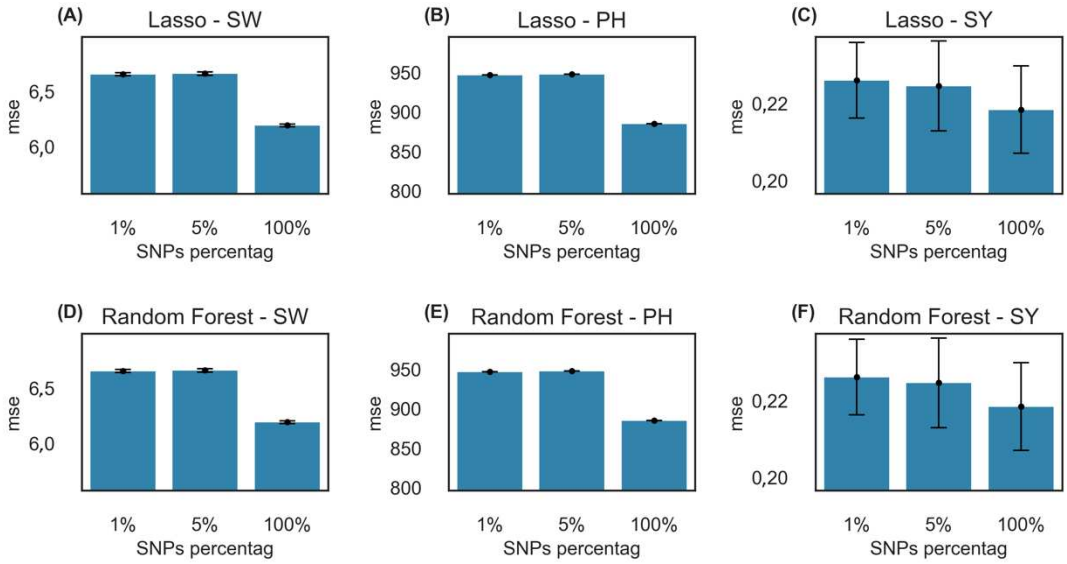


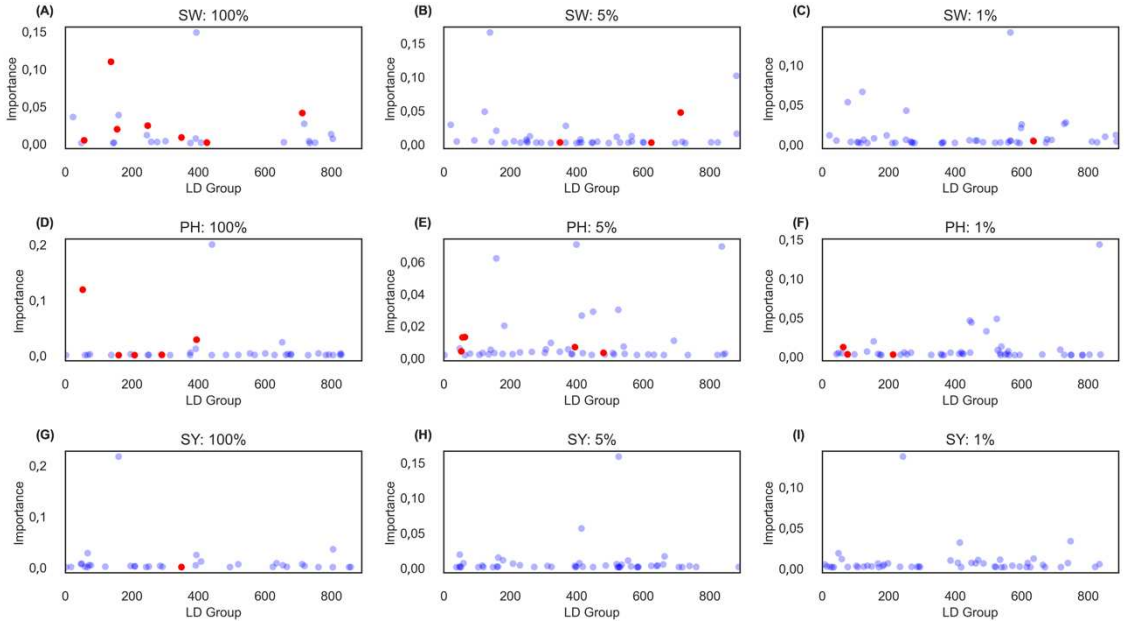
Figura 5: Valor médio do mse (erro quadrático médio) e intervalo de confiança obtido a partir de 20 repetições. (peso de 100 sementes (SW), altura da planta (PH) e rendimento de sementes (SY)).

Na figura 6 pode-se observar os 100 grupos de desequilíbrio com maior importância para as características. Os 100 grupos mais importantes compreendem uma soma acumulada da importância variando de 0.50 a 0.73 (Figura 7).

As figuras 6A, 6B e 6C exibem a importância dos grupos para o peso de semente segundo o modelo RF. Em vermelho destacou-se os blocos de haplótipos que contém loci de QTLs relatados nos estudos de CONTRERAS-SOTO et al., 2017; COPLEY; DUCEPPE; O'DONOUGHUE, 2018; HAO et al., 2012; WANG et al., 2016; YAN et al., 2017; ZHANG et al., 2015, 2016. O modelo construído a partir do painel completo destacou entre os mais importantes maiores números de blocos contendo loci de QTLs relatados nos estudos de GWAS.

Já para altura de planta, os modelos construídos com os painéis de baixa densidade destacaram maior número de loci de QTLs (figura 6C, 6D e 6E) relatados nos trabalhos de CONTRERAS-SOTO et al., 2017b; FANG et al., 2017; ZHANG et al., 2015b, 2015c.

Figura 6: Importância de grupos de LD para a predição das características peso de



100 sementes (SW), altura da planta (PH) e rendimento de sementes (SY) (100 grupos mais importantes). A importância de cada grupo corresponde a soma da importância dos marcadores do grupo segundo o modelo *Random Forest*. Em vermelho estão destacados grupos de LD nos quais existem loci associados a característica avaliada observados em estudos de GWAS.

Considerando agora a variável rendimento de sementes, apenas 1 bloco (Figura 6G) contendo loci de QTLs presentes nos estudos de CONTRERAS-SOTO et al., 2017c; COPLEY; DUCEPPE; O'DONOUGHUE, 2018b; HAO et al., 2012b; ZHANG et al., 2015c se encontra entre os 100 mais importantes segundo o modelo construído com o painel completo. Nos painéis reduzidos não foram assinalados blocos contendo loci de QTLs relatados em estudos de GWAS entre os 100 mais importantes.

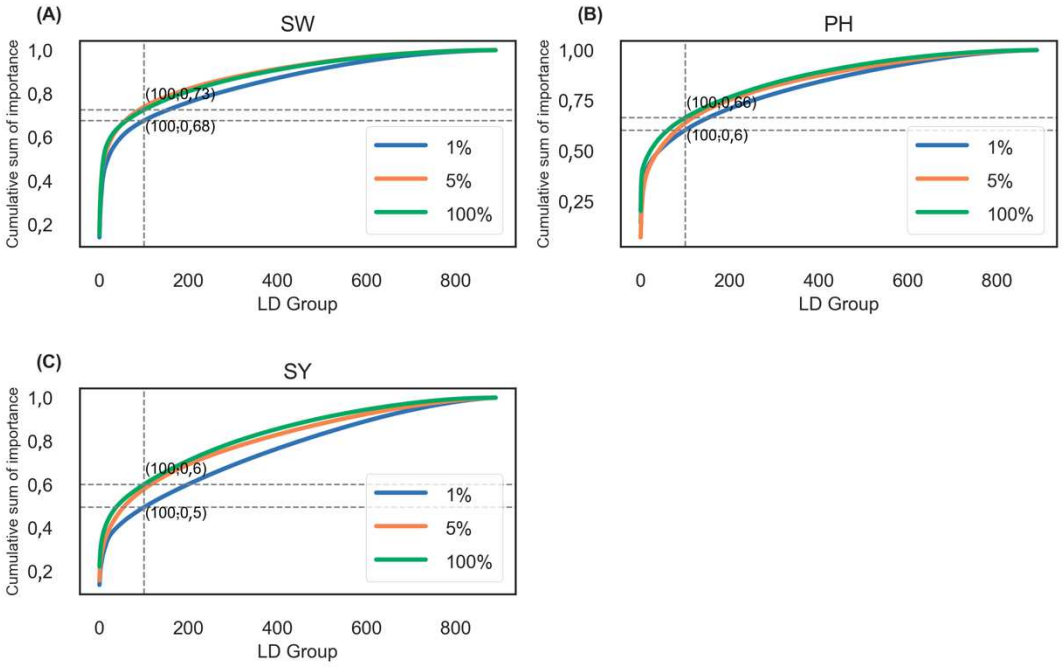


Figura 7: Soma acumulada dos valores de importância de cada grupo de desequilíbrio. A importância de cada grupo foi obtida fazendo a soma da importância dos marcadores do grupo, estimada pelo modelo *Random Forest*. (peso de 100 sementes (SW), altura da planta (PH) e rendimento de sementes (SY)).

CONCLUSÕES

O método proposto se mostrou eficiente no processo de seleção de marcadores. A redução do custo computacional ao considerar painéis de menor densidade é percebida para todos os modelos e características avaliadas. Além disso os resultados mostram que os painéis de SNPs de baixa densidade formados a partir da seleção espaçada dentro dos blocos, formados a partir da AF, são capazes de conservar a acurácia seletiva dos modelos construídos e obter bom desempenho de acurácia preditiva.

O processo de otimização de hiperparâmetros de modelos de *machine learning* é um processo caro computacionalmente que geralmente demanda máquinas de alta performance e requer longos períodos. Utilizar painéis de baixa densidade reduz este custo e traz a possibilidade de explorar modelos mais complexos e considerar espaços de hiperparâmetros maiores no processo de otimização o que poderá resultar em modelos mais acurados.

REFERÊNCIAS

AKBARZADEH, M. et al. GWAS findings improved genomic prediction accuracy of lipid profile traits: Tehran Cardiometabolic Genetic Study. **Scientific Reports**, v. 11, n. 1, mar. 2021.

BREIMAN, L. Classification and regression trees. New York: **Routledge**, 1984. 368 p. v. 1. ISBN 9781315139470.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Outubro 2001.

BROCHU, E.; CORA, V. M.; DE FREITAS, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. **CoRR**, 12 dez. 2010.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, v. 99, n. 6, p. 323–329, jun. 2012.

CLAESEN, M.; DE MOOR, B. Hyperparameter Search in Machine Learning. **CoRR**. 2015.

CONTRERAS-SOTO, R. I. et al. A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-Based Haplotype Analysis. **PLOS ONE**, v. 12, n. 2, p. e0171105, 2 fev. 2017a.

CONTRERAS-SOTO, R. I. et al. A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-Based Haplotype Analysis. **PLOS ONE**, v. 12, n. 2, p. e0171105, 2 fev. 2017b.

CONTRERAS-SOTO, R. I. et al. A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-Based Haplotype Analysis. **PLOS ONE**, v. 12, n. 2, p. e0171105, 2 fev. 2017c.

COPLEY, T. R.; DUCEPPE, M.-O.; O'DONOUGHUE, L. S. Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. **BMC Genomics**, v. 19, n. 1, p. 167, dez. 2018a.

COPLEY, T. R.; DUCEPPE, M.-O.; O'DONOUGHUE, L. S. Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. **BMC Genomics**, v. 19, n. 1, p. 167, dez. 2018b.

DALY, M. J. et al. High-resolution haplotype structure in the human genome. **Nature Genetics**, v. 29, n. 2, p. 229–232, out. 2001.

FANG, C. et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. **Genome Biology**, v. 18, n. 1, p. 161, dez. 2017.

GABRIEL, S. B. et al. The Structure of Haplotype Blocks in the Human Genome. **Science**, v. 296, n. 5576, p. 2225–2229, 21 jun. 2002.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. Genomic Selection Using Low-Density Marker Panels. **Genetics**, v. 182, n. 1, p. 343–353, 1 maio 2009.

HAIR, J. F. (ED.). **Multivariate data analysis**. 7. ed., Pearson new internat. ed ed. Harlow: Pearson, 2014.

HAO, D. et al. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (Glycine max) landraces across multiple environments. **Theoretical and Applied Genetics**, v. 124, n. 3, p. 447–458, fev. 2012a.

HAO, D. et al. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (Glycine max) landraces across multiple environments. **Theoretical and Applied Genetics**, v. 124, n. 3, p. 447–458, fev. 2012b.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. J. **Statistical learning with sparsity: the lasso and generalizations**. First issued in paperback ed. Boca Raton London New York: CRC Press, Taylor & Francis Group, 2020.

HEAD, T. et al. **Scikit-Optimize/Scikit-Optimize: V0.5.2**. Zenodo, , 25 mar. 2018. Disponível em: <https://zenodo.org/record/1207017>. Acesso em: 10 jan. 2022

JANNINK, J.-L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. **Briefings in Functional Genomics**, v. 9, n. 2, p. 166–177, fev. 2010.

LI, B. et al. Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontiers in Genetics**, v. 9, p. 237, 2018.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, v. 157, n. 4, p. 1819–1829, abr. 2001.

NEVES, H. H.; CARVALHEIRO, R.; QUEIROZ, S. A. A comparison of statistical methods for genomic selection in a mice population. **BMC Genetics**, v. 13, n. 1, p. 100, 2012.

OGAWA, S. et al. Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. **BMC Genetics**, v. 15, n. 1, p. 15, 2014.

OLIVEIRA, C. F. DE et al. Identification of patterns related to linkage groups or disequilibrium by factor analysis. **Ciência Rural**, v. 51, n. 5, p. e20190984, 2021.

PATIL, N. et al. Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. **Science**, v. 294, n. 5547, p. 1719–1723, 23 nov. 2001.

Paixão, P.T.M. et al. Factor analysis applied in genomic selection studies in the breeding of Coffea canephora. **Euphytica**, v. 218, p. 42, 2022.

Mazza, S et al. Factor analysis for genetic evaluation of linear type traits in dual-purpose autochthonous breeds. **Animal**, v. 10, p. 372-380, 2016.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

REICH, D. E. et al. Linkage disequilibrium in the human genome. **Nature**, v. 411, n. 6834, p. 199–204, maio 2001.

SHIFMAN, S. Linkage disequilibrium patterns of the human genome across populations. **Human Molecular Genetics**, v. 12, n. 7, p. 771–776, 1 abr. 2003.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. **arXiv:1206.2944 [cs, stat]**, 29 ago. 2012.

SONG, Q. et al. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. **PLoS ONE**, v. 8, n. 1, p. e54985, 25 jan. 2013.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

VALLEJO, R. L. et al. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: Evidence that long-range LD is a major contributing factor. **Journal of Animal Breeding and Genetics**, v. 135, n. 4, p. 263–274, 2018.

WANG, J. et al. Development and application of a novel genome-wide SNP array reveals domestication history in soybean. **Scientific Reports**, v. 6, n. 1, p. 20728, fev. 2016.

WANG, N. et al. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. **The American Journal of Human Genetics**, v. 71, n. 5, p. 1227–1234, nov. 2002.

WU, J. et al. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. **Journal of Electronic Science and Technology**, v. 17, n. 1, p. 26–40, 1 mar. 2019.

YAN, L. et al. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. **BMC Genomics**, v. 18, n. 1, p. 529, dez. 2017.

ZHANG, H. et al. Genetic dissection of the relationship between plant architecture and yield component traits in soybean (*Glycine max*) by association analysis across multiple environments. **Plant Breeding**, v. 134, n. 5, p. 564–572, out. 2015a.

ZHANG, H. et al. Genetic dissection of the relationship between plant architecture and yield component traits in soybean (*Glycine max*) by association analysis across multiple environments. **Plant Breeding**, v. 134, n. 5, p. 564–572, out. 2015b.

ZHANG, H. et al. Genetic dissection of the relationship between plant architecture and yield component traits in soybean (*Glycine max*) by association analysis across multiple environments. **Plant Breeding**, v. 134, n. 5, p. 564–572, out. 2015c.

ZHANG, J. et al. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. **BMC Genomics**, v. 16, n. 1, p. 217, dez. 2015d.

ZHANG, J. et al. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). **Theoretical and Applied Genetics**, v. 129, n. 1, p. 117–130, jan. 2016.

MATERIAL COMPLEMENTAR

Tabela 2: Valores de parâmetros usados em cada modelo, percentual de marcadores selecionados por bloco de LD (p) e variáveis peso de 100 sementes (SW), altura da planta (PH) e rendimento de sementes (SY).

model	p	variable	parâmetro	valor do parâmetro
Lasso	0,01	PH	alpha	0,0
Lasso	0,01	SW	alpha	0,0002078643980524
Lasso	0,01	SY	alpha	0,0
Lasso	0,05	PH	alpha	0,0
Lasso	0,05	SW	alpha	0,0002078643980524
Lasso	0,05	SY	alpha	0,0011342595463488
Lasso	1,0	PH	alpha	0,0054531258916022
Lasso	1,0	SW	alpha	0,0011342595463488
Lasso	1,0	SY	alpha	0,0011342595463488
Random Forest	0,01	PH	min_samples_leaf	7,0
Random Forest	0,01	PH	min_samples_split	12,0
Random Forest	0,01	PH	n_estimators	107,0
Random Forest	0,01	SW	min_samples_leaf	3,0
Random Forest	0,01	SW	min_samples_split	2,0
Random Forest	0,01	SW	n_estimators	500,0
Random Forest	0,01	SY	min_samples_leaf	3,0
Random Forest	0,01	SY	min_samples_split	2,0
Random Forest	0,01	SY	n_estimators	500,0
Random Forest	0,05	PH	min_samples_leaf	6,0
Random Forest	0,05	PH	min_samples_split	7,0
Random Forest	0,05	PH	n_estimators	500,0
Random Forest	0,05	SW	min_samples_leaf	3,0
Random Forest	0,05	SW	min_samples_split	5,0
Random Forest	0,05	SW	n_estimators	381,0

model	p	variable	parâmetro	valor do parâmetro
Random Forest	0,05	SY	min_samples_leaf	2,0
Random Forest	0,05	SY	min_samples_split	7,0
Random Forest	0,05	SY	n_estimators	500,0
Random Forest	1,0	PH	min_samples_leaf	5,0
Random Forest	1,0	PH	min_samples_split	12,0
Random Forest	1,0	PH	n_estimators	403,0
Random Forest	1,0	SW	min_samples_leaf	3,0
Random Forest	1,0	SW	min_samples_split	8,0
Random Forest	1,0	SW	n_estimators	219,0
Random Forest	1,0	SY	min_samples_leaf	3,0
Random Forest	1,0	SY	min_samples_split	8,0
Random Forest	1,0	SY	n_estimators	219,0

CONCLUSÕES GERAIS

O desequilíbrio de ligação é um conceito fundamental para estudos de seleção genômica. Muitos estudos foram desenvolvidos visando caracterizar e mensurar as relações entre marcadores e marcadores e QTLs. Vimos que a análise de fatores é capaz de descrever estas relações criando fatores comuns a grupos de marcadores e que podemos interpretar estes fatores como grupos de ligação ou de desequilíbrio de ligação.

Entender as relações entre marcadores, em especial devido ao grande número destas variáveis, é um conhecimento valioso que pode ser utilizado para orientar um processo automatizado de seleção de marcadores. É possível encontrar muitas abordagens que visam praticar seleção de SNPs em geral com o objetivo de aumentar a acurácia de modelos preditivos. A seleção espaçada de marcadores ao longo do genoma possui como maior vantagem com relação as demais o fato de que um painel de baixa densidade de marcadores criado a partir desta abordagem, pode ser utilizado no processo de predição de qualquer característica.

Mostramos que esta abordagem pode ser combinada com as informações dos grupos obtidos utilizando análise de fatores e realizar a seleção espaçada dentro dos grupos formados. Os resultados mostraram que está seleção é capaz de praticar uma seleção reduzindo em 95% ou mais o painel de marcas sem perca em termos de acurácia seletiva, e em alguns casos preservando também a acurácia preditiva.

Dentre os grandes desafios presentes na proposta de criar modelos capazes de predizer valores genéticos de indivíduos se destaca a dimensão dos dados, que exigem grande poder computacional e longos períodos no processo de treinamento de modelos e principalmente no processo de otimização dos modelos de *machine learning*. Portanto este trabalho traz oportunidades de explorar modelos e estratégias de otimização inviáveis de serem postos em práticas com volumes tão grande de informação utilizando um painel reduzido de marcadores.