Ciência Rural, Santa Maria, v.51:5, e20190984, 2021

http://doi.org/10.1590/0103-8478cr20190984

ISSNe 1678-4596

BIOLOGY

Identification of patterns related to linkage groups or disequilibrium by factor analysis

Cristiano Ferreira de Oliveira[1*] Gabriely Teixeira[1] Alex da Silva Temoteo[1] Moysés Nascimento[1] Cosme Damião Cruz[1]

[1]Universidade Federal de Viçosa, Departamento de Estatística, 36570-900, Viçosa, MG, Brasil. E-mail: cristiano.oliveira@ufv.br. [*]Corresponding author.

ABSTRACT: *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 (AF) 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 haplotypes 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 AF was able to identify patterns of association between markers, identifying subgroups of markers that reflect factor binding groups and also linkage disequilibrium groups.*

Key words: *linkage disequilibrium, factor analysis, SNP, haplotype blocks, linkage groups, QTL.*

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

RESUMO: *Padrões empíricos de desequilíbrio de ligação (LD) podem ser utilizados para aumentar o poder estatístico do mapeamento genético. Este trabalho foi realizado com o objetivo de verificar a eficácia da análise de fatores (AF) aplicada a conjuntos de dados de marcadores moleculares do tipo SNP, visando identificar grupos de ligação e blocos de haplótipos. O conjunto de dados SNPs utilizado foi oriundo de um processo de simulação de uma população F2, contendo 2000 marcas com informações de 500 indivíduos. A estimação das cargas fatoriais (loadings) da AF foi feita de duas formas, considerando a matriz de distâncias entre os marcadores (A) e considerando a matriz de correlação (R). O número de fatores (k) a ser utilizado foi estabelecido com base no gráfico scree-plot e com base na proporção da variância total explicada. Os resultados indicam que as matrizes A e R conduzem a resultados similares. Com base no scree-plot considerou-se k igual a 10 e os fatores interpretados como sendo representativos dos grupos de ligação. O segundo critério conduziu a um número de fatores igual a 50, e os fatores interpretados como sendo representativos dos blocos de haplótipos. Isto mostra o potencial da técnica que permite obter resultados aplicáveis a qualquer tipo de população, corroborando a interpretação de estudos genômicos. O trabalho demonstrou que a AF foi capaz de identificar padrões de associação entre marcadores, identificando subgrupos de marcadores que refletem grupos de ligação fatorial e também grupos de desequilíbrio de ligação.*

Palavras chave: *desequilíbrio de ligação, análise de fatores, SNP, blocos de haplótipos, grupos de ligação, 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 (RESENDE et al., 2014; BORÉM & CAIXETA, 2016; NADEEM et al., 2018). The use of molecular information in the process of genetic breeding brings great benefits; phenotypic information, combined with genotypic

Received 12.11.19 Approved 10.08.20 Returned by the author 12.10.20
CR-2020-0984.R1

information, provides greater precision to predict its genetic value (SPINDEL et al., 2015; MEUWISSEN et al., 2016).

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 (GHOLAMI et al., 2015; CAETANO, 2009).

LD refers to the non-independence of alleles in different locations. Considering a locus with alleles *A* and *a* with frequencies p_A and $(1 - p_A)$, respectively, and a second with alleles *B* and *b* with frequencies p_B and $(1 – p_B)$. If the loci are independent, the frequency expected for the AB haplotype will be $p_A\, p_B$. 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 (PRITCHARD & PRZEWORSKI, 2001; CARNEIRO & VIEIRA, 2002). 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, 2002; PATIL, 2001; REICH et al., 2001; WANG, N. et al., 2002).

SHIFMAN et al., 2003 measured the LD between pairs of SNP using the absolute value of Lewontin's D′ (|D′|) and r statistics. They found that measuring LD with r or r^2 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 et al., 2014) 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.

MATERIALS 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 allow 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-brother 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 a number of variables of interest, $X_1, X_2, ..., X_p$, is linearly related to a smaller number of the non-observable factors, $F_1, F_2, ..., F_k$.
Considering $Y_p = [Y_1, Y_2, \ldots, Y_p]$, 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:
$Y - \mu = \Gamma F + \epsilon$
such that pΓk = [γ_ij], which is a matrix of coefficients denominated by factorial loads and has rank k ≤ p. ${}_mF_1$ 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:

$$\begin{array}{ccccccccccc}
Y_1-\mu_1 & = & \gamma_{11}F_1 & + & \gamma_{12}F_2 & + & \ldots & + & \gamma_{1k}F_k & + & \epsilon_1\\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots\\
Y_i-\mu_i & = & \gamma_{i1}F_1 & + & \gamma_{i2}F_2 & + & \cdots & + & \gamma_{ik}F_k & + & \epsilon_i\\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots\\
Y_p-\mu_p & = & \gamma_{p1}F_1 & + & \gamma_{p2}F_2 & + & \cdots & + & \gamma_{pk}F_k & + & \epsilon_p
\end{array}$$

If Cov(F) = I_k or, in other words, if the factors are not correlated, we call the model an orthogonal factorial.

Also, considering the assumptions that E(Y) = μ, E(F) = E(ϵ) = 0, Cov(Y)=Σ, Cov(ϵ)=Ψ, and Cov(F,ϵ)=0, such that,

$$\Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0\\ 0 & \Psi_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \Psi_p \end{bmatrix}$$

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

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

To measure data’s adequacy , the LEDOIT & WOLF (2002) 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 take into account 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.

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, 2010), 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.

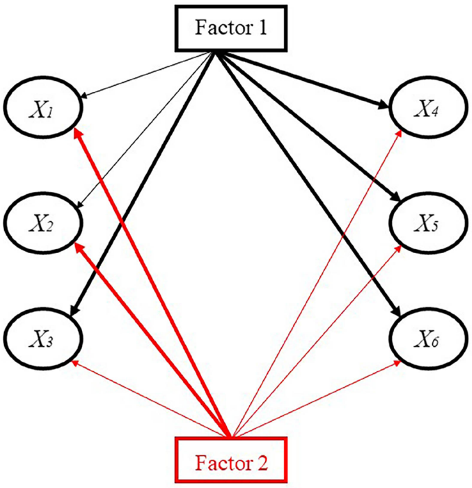

Figure 1 - Grouping scheme using Analysis of Factors, with six variables and two factors. Thicker lines indicate higher factorial load. Group 1 consists of X_3, X_4, X_5, and X_6. Group 2 consists of X_1 and X_2.

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.

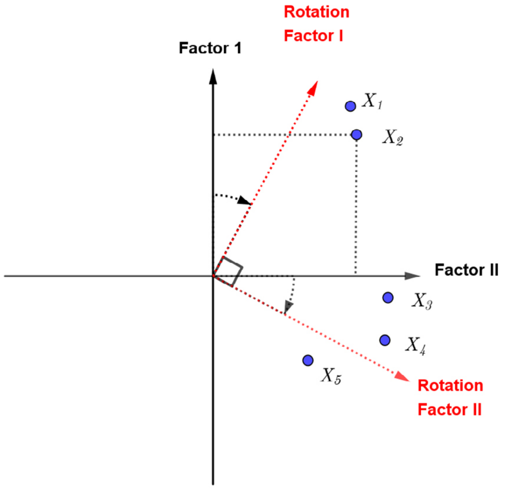

Figure 2 - Illustration representing an orthogonal factorial rotation.

QTL analysis has been used to identify molecular markers responsible for the variation in an observed characteristic (LI, C. et al., 2016; DHARIWAL et al., 2018). 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 4 a) and 4 b), respectively, in view of the values of LOD (log of odds ratio) obtained in the analyses of each interval in each binding group. For the purpose of dimensionality reduction, the researcher may choose samplings prioritizing brands that represents the regions with higher concentration of putative controlling loci of the characteristics included in the analyses.

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

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 et al., 2003). High values of r^2 can be found between markers belonging to the same block as verified by GABRIEL (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

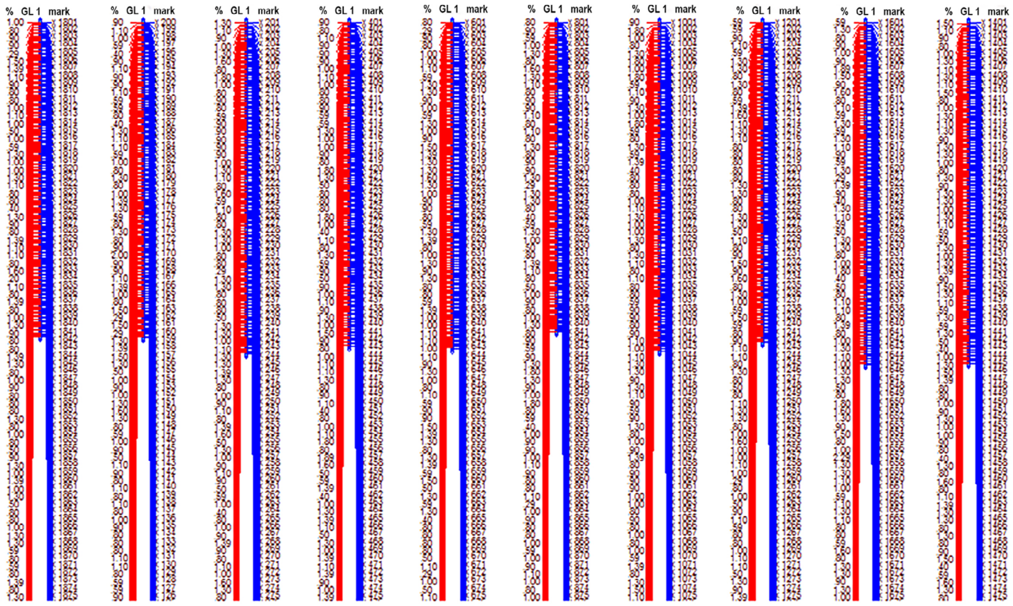

Figure 3 - Genetic map established for a population F2 considering information of 2000 molecular markers of the SNP type.

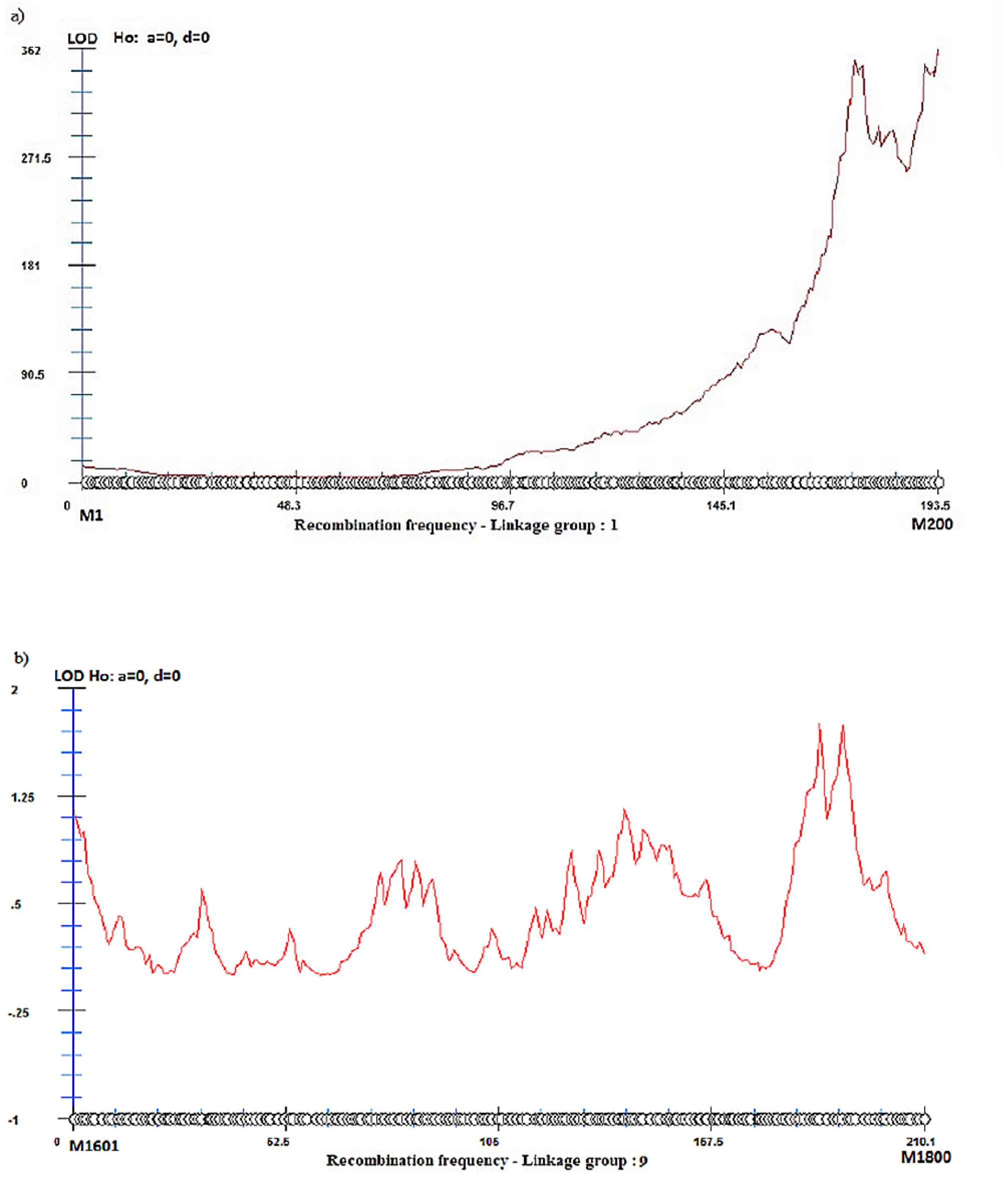

Figure 4 - Genetic map and QTL detection map for a quantitative characteristic measured in a F2 population, evidencing the situations: a) Detection of QTLs (peak above LOD referential equal to 3; b) Absence of QTL for the characteristic of interest.

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.

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

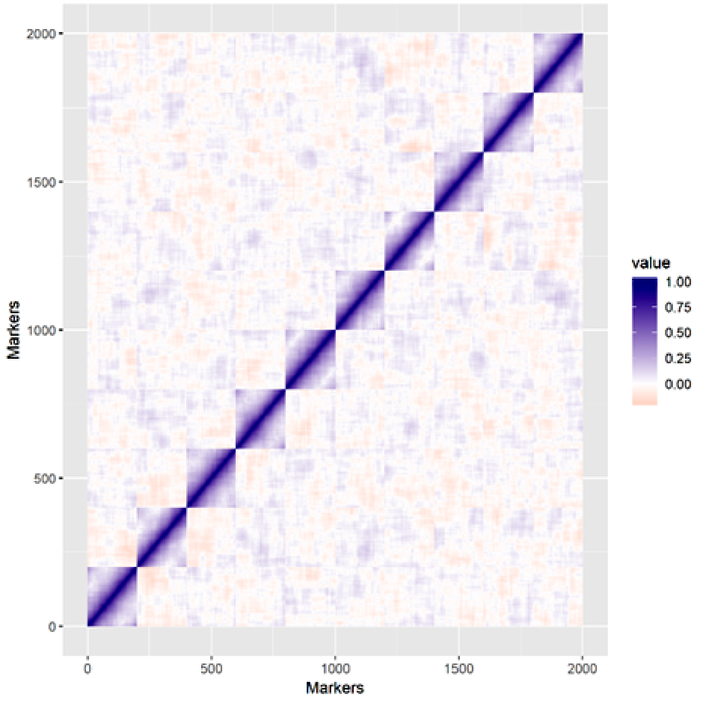

Figure 5 - Correlation map between pairs of markers highlighting the intensity of the associations within the disequilibrium blocks.

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 & WOLF (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.

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

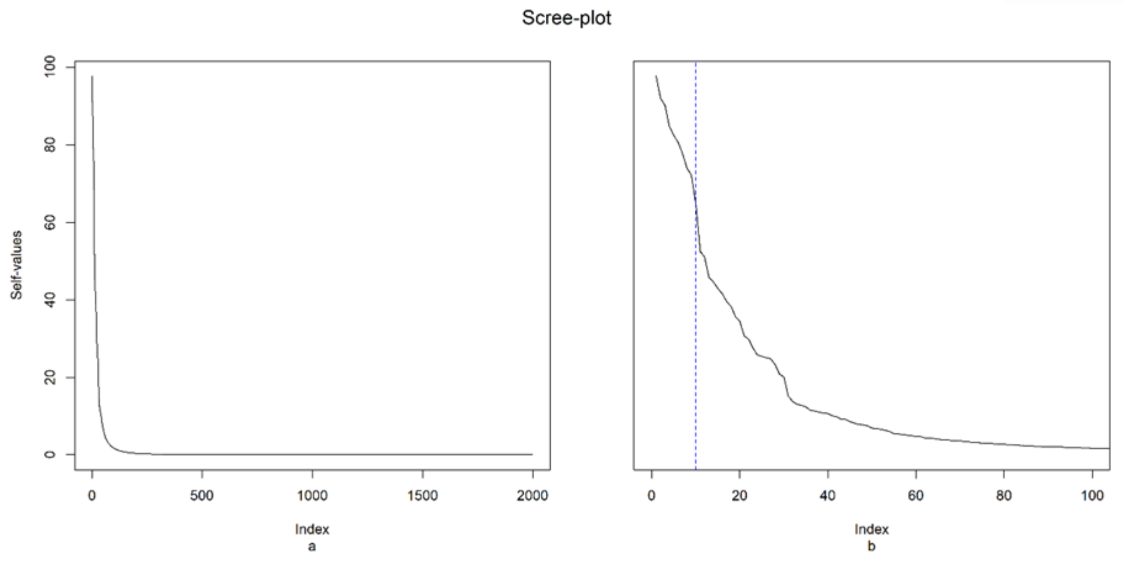

Figure 6 - a) Estimates of self-values in descending order, obtained from the correlation matrix between pairs of markers and b) Highlight of the 100 first estimates of the self-values in descending order.

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

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

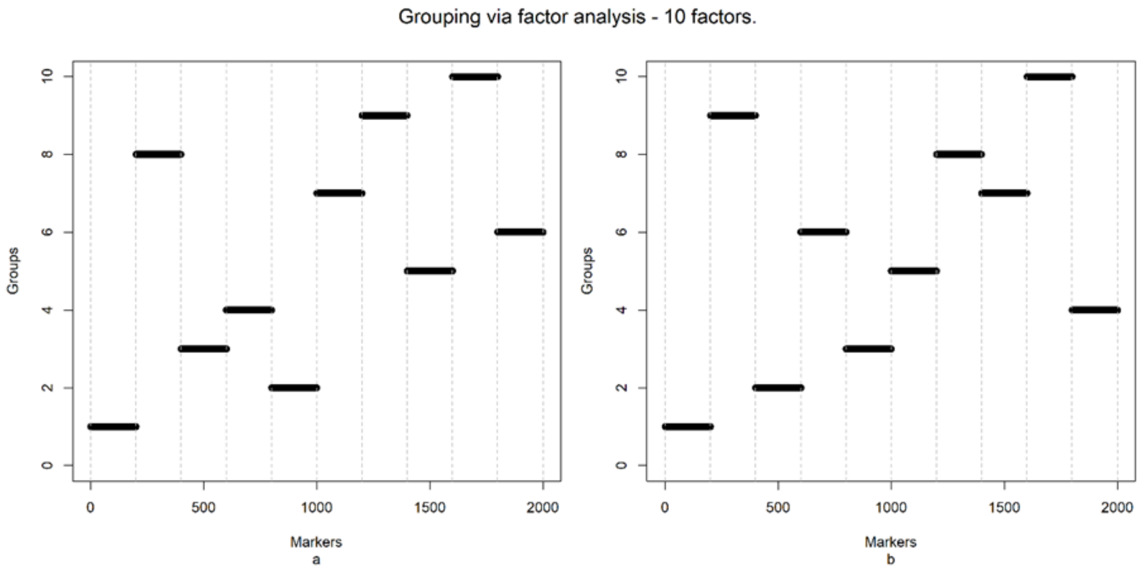

Figure 7 - Groups of factorial complexes obtained through the analysis of factors with varimax rotation, considering 10 factors. a - Using the correlation matrix and b - Using the matrix A established as a function of distances genetics.

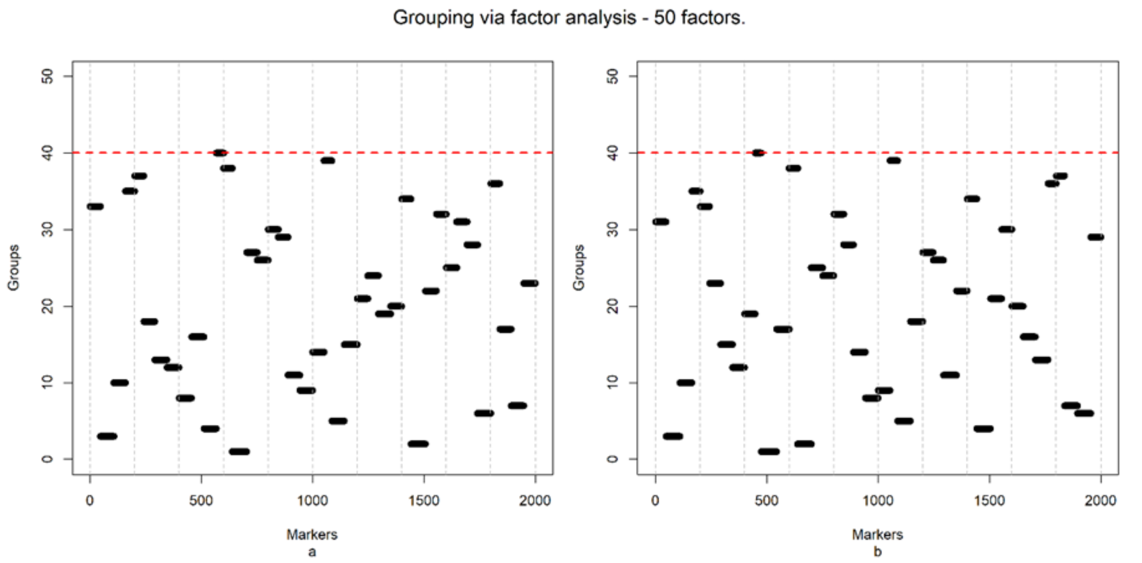

Figure 8 - Groups of factorial complexes obtained through the analysis of factors with varimax rotation, considering 50 factors. a - Using the correlation matrix and b - using the matrix A established as a function of distance genetics.

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 (PATIL, 2001;HALLDORSSON, 2004; JASIELCZUK et al., 2020). 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 equilibrium 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 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.

Figure 9 show the estimates of the communalities that represent how much each marker can be explained by the set of factors established in structural simplification. The use of the correlation matrix or the matrix of distances for AF, the results of commonality are similar, since both quantify the linear relationship between the marks.

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 disequilibrium groups generates a better representation of the relationship structure between the SNPs, compared to the division according to the linkage groups.

CONCLUSION

Recognizing grouping patterns of a set of markers that reflect linkage groups or disequilibrium

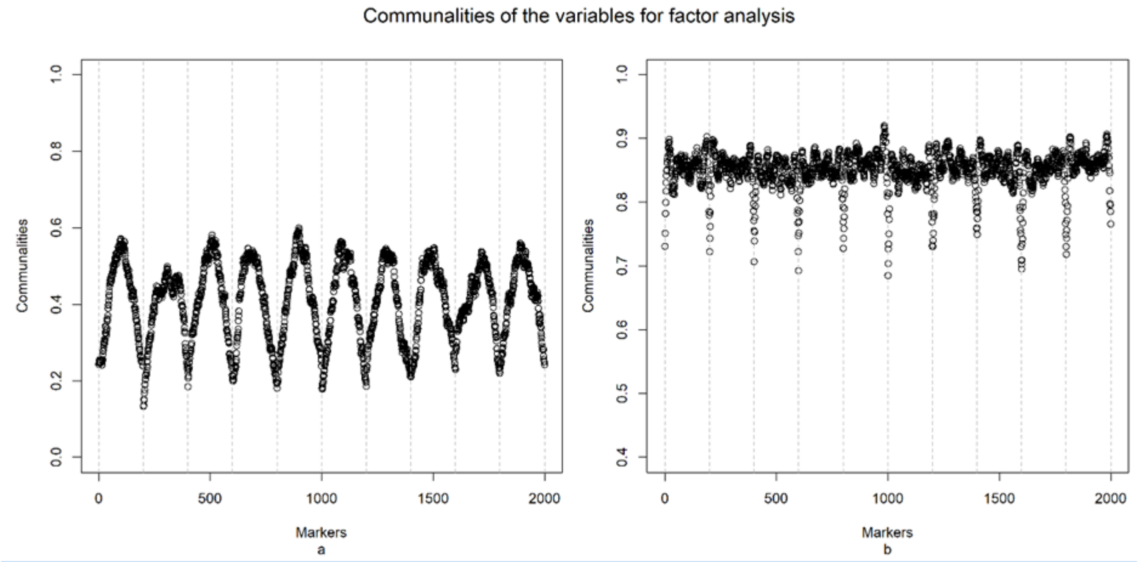

Figure 9 - Communalities of the variables for factor analysis, obtained through the correlation matrix. a) Using 10 factors and b) using 50 factors.

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 also linkage disequilibrium groups.

ACKNOWLEDGEMENTS

We would to thank the “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES), Brazil.

DECLARATION OF CONFLICT OF INTERESTS

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

AUTHORS’ CONTRIBUTIONS

All authors contributed equally for the conception and writing of the manuscript. All authors critically revised the manuscript and approved of the final version.

REFERENCES

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

CAETANO, A. R. SNP markers: basic concepts, applications in animal breeding and management and perspectives for the future.**Revista Brasileira de Zootecnia**, 2009. v.38, n.SUPPL. 1, p.64–71. Available from: <http://dx.doi.org/10.1590/S1516-35982009001300008>. Accessed: Nov. 19, 2019. doi: 10.1590/S1516-35982009001300008.

CARNEIRO, M. S.; VIEIRA, M. L. C. Mapas genéticos em plantas. **Bragantia**, 2002. v.61, p.89–100. Available from: <http://dx.doi.org/10.1590/S0006-87052002000200002>. Accessed: Nov. 19, 2019. doi: 10.1590/S0006-87052002000200002.

CATTELL, R. B. The Scree Test For The Number Of Factors. **Multivariate Behavioral Research**, 1966. v.1, n.2, p.245–276. Available from:<https://doi.org/10.1207/s15327906mbr0102_10>. Accessed: Nov. 19, 2019. doi: 10.1207/s15327906mbr0102_10.

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**, 2005. v.142, n.1–2, p.169–196. Available from: <https://doi.org/10.1007/s10681-005-1681-5>. Accessed: Nov. 19, 2019. doi: 10.1007/s10681-005-1681-5.

COSTELLO, A. B.; OSBORNE, J. W. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. **Practical Assessment Research & Evaluation**, 2005. v.10, p.9. Avaliable from: <https://doi.org/10.7275/jyj1-4868>. Accessed: Nov. 19, 2019. doi: 10.7275/jyj1-4868.

CRUZ, C. D. Genes Software – extended and integrated with the R , Matlab and Selegen. **Acta Scientiarum Agronomy**, 2016. v.38, n.4, p.547–552. Available from: <http://www.scielo.br/pdf/asagr/v38n4/1807-8621-asagr-38-04-00547.pdf>. Accessed: Nov. 19, 2019. doi: 10.4025/actasciagron.v38i4.32629.

DALY, M. J. et al. High-resolution haplotype structure in the human genome. **Nature Genetics**, 2001. v.29, n.2, p.229–232. Available from: <https://doi.org/10.1038/ng1001-229>. Accessed: Nov. 19, 2019. doi: 10.1038/ng1001-229.

DHARIWAL, R. et al. High density ingle 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**, 5 jan. 2018. v.9, n.1, p.19. Available from: <https://doi.org/10.3390/genes9010019>. Accessed: Nov. 19, 2019. doi: 10.3390/genes9010019.

GABRIEL, S. B. The Structure of haplotype Blocks in the Human Genome. **Science**, 21 jun. 2002. v.296, n.5576, p.2225–2229. Available from: <https://www.sciencemag.org/lookup/doi/10.1126/science.1069424>. Accessed: Nov. 19, 2019. doi: 10.1126/science.1069424.

GHOLAMI, M. et al. Genome scan for Selection in Structured Layer Chicken Populations Exploiting Linkage Disequilibrium Information. **PLOS ONE**, 2015. v.10, n.7, p.1–15. Available from: <https://doi.org/10.1371/journal.pone.0130497>. Accessed: Nov. 19, 2019. doi: 10.1371/journal.pone.0130497.

GREENSPAN, G.; GEIGER, D. Model-Based Inference of Haplotype Block Variation. **Journal of Computational Biology**, 2004. v.11, n.2–3, p.493–504. Available from: <http://doi.org/10.1089/1066527041410300>. Accessed: Nov. 19, 2019. doi: 10.1089/1066527041410300.

HAIR, J. F. J. et al. **Overview of multivariate methods**. 7. ed. USA: Pearson Education Limited, 2014.

HALLDORSSON, B. V. Optimal haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies. **Genome Research**, 1 ago. 2004. v.14, n.8, p.1633–1640. Available from: <http://www.genome.org/cgi/doi/10.1101/gr.2570004>. Accessed: Nov. 19, 2019. doi: 10.1101/gr.2570004.

JASIELCZUK, I. et al. Comparison of linkage disequilibrium, effective population size and haplotype blocks in Polish Landrace and Polish native pig populations. **Livestock Science**, jan. 2020. v.231, p.103887. Available from: <https://doi.org/10.1016/j.livsci.2019.103887>. Accessed: Mar. 01, 2020. doi: 10.1016/j.livsci.2019.103887.

KUMAR, V. et al. Genome-wide association mapping of salinity tolerance in rice (Oryza sativa). **DNA Research**, 2015. v.22, n.2, p.133–145. Available from: <https://doi.org/10.1093/dnares/dsu046>. Accessed: Nov. 19, 2019. doi: 10.1093/dnares/dsu046.

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**, ago. 2002. v.30, n.4, p.1081–1102. Available from: <http://projecteuclid.org/euclid.aos/1031689018>. Accessed: Nov. 19, 2019. doi: 10.1214/aos/1031689018.

LI, C. et al. Single nucleotide polymorphisms linked to quantitative trait loci for grain quality traits in wheat. **The Crop Journal**, fev. 2016. v.4, n.1, p.1–11. Available from: <https://doi.org/10.1016/j.cj.2015.10.002>. Accessed: Nov. 19, 2019. doi: 10.1016/j.cj.2015.10.002.

LI, G. et al. Genome-wide association Mapping Reveals Novel QTL for Seedling Leaf Rust Resistance in a Worldwide Collection of Winter Wheat. Madison, WI: **The Plant Genome**, 2016. v.9. Available from: <https://doi.org/10.3835/plantgenome2016.06.0051>. Accessed: Nov. 19, 2019. doi: 10.3835/plantgenome2016.06.0051.

LIU, H. et al. A Neural network Based on Rough Set (RSNN) for Prediction of Solitary Pulmonary Nodules. **International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing**, 2009. p.135–138. Available from: <http://ieeexplore.ieee.org/document/5260720/>. Accessed: Nov. 19, 2019. doi: 10.1109/IJCBS.2009.105.

LOEHLIN, J. C. **Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis**. 4. ed. New Jersey: LAWRENCE ERLBAUM ASSOCIATES, 2004.

MCRAE, A. F. et al. Linkage disequilibrium in Domestic Sheep. **Genetics**, 1 mar. 2002. v.160, n.3, p.1113 LP – 1122. Available from: <http://www.genetics.org/content/160/3/1113.abstract>. Accessed: Nov. 19, 2019. issn: 0016-6731.

MEUWISSEN, T.; et al.,. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers**, 2016. v.6, n.1, p.6–14. Available from: <https://doi.org/10.2527/af.2016-0002>. Accessed: Nov. 19, 2019. doi: 10.2527/af.2016-0002.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada**: Uma Abordagem Aplicada. Brlo Horizonte: 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**, 4 mar. 2018. v.32, n.2, p.261–285. Available from: <https://doi.org/10.1023/A:1015210025234>. Accessed: Nov. 19, 2019. doi: 10.1023/A:1015210025234.

PALLANT, J. **SPSS survival manual : a step by step guide to data analysis using SPSS**. 4. ed. Maidenhead: Open University Press/McGraw-Hill, 2010.

PATIL, N. Blocks of limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. **Science**, 23 nov. 2001. v.294, n.5547, p.1719–1723. Available from: <https://www.sciencemag.org/lookup/doi/10.1126/science.1065573>. Accessed: Nov. 19, 2019. doi: 10.1126/science.1065573.

PREACHER, K. J.; MACCALLUM, R. C. Exploratory factor Analysis in Behavior Genetics Research: Factor Recovery with Small Sample Sizes. **Behavior Genetics**, 2002. v.32, n.2, p.153–161. Available from: <https://doi.org/10.1023/A:1015210025234>. Accessed: Nov. 19, 2019. doi: 10.1023/A:1015210025234.

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

<https://doi.org/10.1086/321275>. Accessed: Nov. 19, 2019. doi: 10.1086/321275.

R CORE TEAM. **R**: A Language and environment for Statistical Computing. Available from: <https://www.r-project.org/>. Accessed: Nov. 19, 2019.

REICH, D. E. et al. Linkage disequilibrium in the human genome. **Nature**, 10 maio. 2001. v.411, p.199. Available from: <https://doi.org/10.1038/35075590>. Accessed: Nov. 19, 2019. doi: 10.1038/35075590.

RESENDE, M. D. V.; et al. **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: Suprema, 2014.

SHIFMAN, S. et al. Linkage disequilibrium patterns of the human genome across populations. **Human Molecular Genetics**, 2003. v.12, n.7, p.771–776. Available from: <https://doi.org/10.1093/hmg/ddg088>. Accessed: Nov. 19, 2019. doi: 10.1093/hmg/ddg088.

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**, 17 fev. 2015. v.11, n.2, p.e1004982. Available from: <https://doi.org/10.1371/journal.pgen.1004982>. Accessed: Nov. 19, 2019. doi: 10.1371/journal.pgen.1004982.

TERRA, T. G. R. et al. QTLs identification for characteristics of the root system in upland rice through DNA microarray. **Acta Scientiarum. Agronomy**, 2 set. 2016. v.38, n.4, p.457. Available from: <https://doi.org/10.4025/actasciagron.v38i4.30534>. Accessed: Nov. 19, 2019. doi: 10.4025/actasciagron.v38i4.30534.

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**, nov. 2002. v.71, n.5, p.1227–1234. Available from: <https://doi.org/10.1086/344398>. Accessed: Mar. 19, 2019. doi: 10.1086 /344398.