IARA GONÇALVES DOS SANTOS

ALFALFA BREEDING FOR CULTIVATION IN THE TROPICS: INSIGHTS ON GENETIC DIVERSITY AND YIELD PERSISTENCE

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

Orientador: Cosme Damião Cruz

VIÇOSA - MINAS GERAIS
2021

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

T
S237a
2021

Santos, Iara Gonçalves dos, 1992-
Alfalfa breeding for cultivation in the tropics : insights on genetic diversity and yield persistence / Iara Gonçalves dos Santos. – Viçosa, MG, 2021.
92 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Inclui apêndice.
Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. *Medicago sativa* - Melhoramento genético. 2. Análise de regressão. 3. Redes neurais (Neurobiologia). I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 633.312

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB6 2523

IARA GONÇALVES DOS SANTOS

ALFALFA BREEDING FOR CULTIVATION IN THE TROPICS: INSIGHTS ON GENETIC DIVERSITY AND YIELD PERSISTENCE

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

APROVADA: 14 de julho de 2021.

Assentimento:

Iara Gonçalves dos Santos
Autora

Cosme Damião Cruz
Orientador

Aos meus familiares e amigos.

AGRADECIMENTOS

A Deus, pelo dom da vida.

Aos meus pais Jonilson Pereira dos Santos (sempre presente) e Marlene Gonçalves dos Santos, e às minhas irmãs Juliane e Gabriela, meu agradecimento por todo o apoio e incentivo. Eu sou porque nós somos!

Ao Arthur Matheus, parceiro de vida, pelo amor dedicado.

Ao professor Dr. Cosme Damião Cruz, pela orientação, ensinamentos e amizade.

À Embrapa Pecuária Sudeste em especial ao pesquisador Dr. Reinaldo de Paula Ferreira pela parceria.

Aos professores Dr. Moysés Nascimento, Dr. Renato Domiciano Silva Rosado e à pesquisadora Dr. Isabela de Castro Sant'Anna, pelas valiosas contribuições e suporte.

À família BIOINFO por caminharem junto comigo.

Ao Programa de Pós-Graduação em Genética e Melhoramento e à Universidade Federal de Viçosa pela formação.

Ao Conselho Nacional de desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão de bolsas de estudos. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

À Iowa State University, em especial aos professores Dr. Lübberstedt e Dr. Frei pela acolhida, suporte e ensinamentos durante o doutorado sanduíche.

E a todos que de alguma forma contribuíram para essa conquista, a minha imensa gratidão!

MUITO OBRIGADA!

“The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.”
(Isaac Asimov)

“The good thing about science is that it's true whether or not you believe in it.”
(Neil deGrasse Tyson)

“What you learn from a life in science is the vastness of our ignorance”
(David Eagleman)

BIOGRAFIA

IARA GONÇALVES DOS SANTOS, filha de Marlene Gonçalves dos Santos e Jonilson Pereira dos Santos, nasceu em 04 de fevereiro de 1992, em Pavão, Minas Gerais.

Estudou na Escola Municipal Professora Davina Santos e concluiu o ensino médio no ano de 2009, na Escola Estadual Caio Nelson de Sena, Pavão, Minas Gerais.

Em março de 2011, ingressou no curso de Agronomia da Universidade Federal de Viçosa, colando grau em janeiro de 2016.

Em março de 2016, iniciou o curso de Mestrado em Genética e Melhoramento pela Universidade Federal de Viçosa, colando grau em julho de 2017.

Em julho de 2017 iniciou o curso de Doutorado em Genética e Melhoramento pela Universidade Federal de Viçosa, submetendo-se à defesa de tese em 14 de julho de 2021.

RESUMO

SANTOS, Iara Gonçalves dos, D.Sc., Universidade Federal de Viçosa, julho de 2021. **Melhoramento de alfafa para cultivo nos trópicos: insights sobre a diversidade genética e persistência da produção.** Orientador: Cosme Damião Cruz.

Alfafa (*Medicago sativa* L.) é considerada a "rainha das forrageiras" e é essencial em rebanhos leiteiros altamente especializados. Alfafa tem potencial de ser cultivada em diferentes regiões edafoclimáticas mas seu cultivo ainda é limitado em regiões de clima tropical. Persistência de produção é um dos gargalos do melhoramento da cultura em regiões tropicais e por isso, todos os esforços para melhorar essa característica ajudarão a expandir o seu cultivo. Esse estudo objetivou investigar se o germoplasma de alfafa mantido pela Embrapa Pecuária Sudeste carrega diversidade genética para caracteres bromatológicos e agronômicos. O estudo também investigou a persistência de produção, qual a melhor forma de acessá-la e como selecionar acessos persistentes baseado em modelos de regressão aleatória e redes neurais artificiais (RNA). Para o estudo da diversidade genética, valores genéticos de 77 acessos de alfafa de background genético temperado avaliados em relação a nove características e em oito cortes foram obtidos para caracterizar a diversidade fenotípica. Marcadores microssatélites foram utilizados para acessar a diversidade molecular. As análises dos dados fenotípicos revelaram a presença de diversidade genética. A variabilidade detectada pelos dados fenotípicos e moleculares indicaram o potencial do germoplasma para geração de populações-base adaptadas a condições de clima tropical. Informações de produção de matéria seca tomadas em 24 cortes foram usadas para acessar a persistência dos acessos que compõe o germoplasma em estudo. Um modelo de regressão aleatória foi ajustado para construir curvas da trajetória dos acessos. As curvas ajustadas mostraram alta amplitude da produção de matéria seca ao longo do tempo, o que sugere alta variabilidade para persistência. O método de três etapas para acessar persistência apresentado nesse estudo envolveu (1) modelo de regressão aleatória para obtenção das trajetórias genéticas, (2) a utilização do método de agrupamento k-médias para definição de grupos de persistência e (3) a utilização de RNA para realizar a mesma classificação definida pelo método k-médias de forma automatizada. A vantagem desse método é que novos acessos de alfafa podem ser submetidos à mesma RNA. Basicamente, quando novos acessos forem avaliados, estes poderão ser classificados de acordo com seus valores genéticos utilizando a mesma RNA treinada anteriormente sem precisar submeter os escores a um novo agrupamento.

O método de persistência salta de três para dois passos e pode ajudar melhoristas de alfafa no processo de tomada de decisão.

Palavras-chave: *Medicago sativa*. Mapas auto-organizáveis. Modelos de regressão aleatória. Redes neurais artificiais.

ABSTRACT

SANTOS, Iara Gonçalves dos, D.Sc., Universidade Federal de Viçosa, July, 2021. **Alfalfa breeding for cultivation in the tropics: insights on genetic diversity and yield persistence.** Adviser: Cosme Damião Cruz.

Alfalfa (*Medicago sativa* L.) is considered "the queen of forages" and plays a key role in highly specialized dairy herds. Alfalfa has the potential to be grown in different edaphoclimatic regions, though its cultivation in tropical regions is still limited. Because yield persistence is one of the bottlenecks of alfalfa breeding in tropical regions, efforts should be done to overcome this problem. This study aimed to investigate whether the alfalfa germplasm held by Embrapa Southeast Livestock has satisfactory genetic diversity regarding bromatological and agronomic traits. The investigation also looked into yield persistence, how to access it, and how to select persistent accessions based on random regression models and artificial neural networks (ANN). Best linear unbiased predictors (BLUPs) of nine traits of seventy-seven alfalfa accessions from a temperate genetic background evaluated in eight harvests were used to estimate the phenotypic diversity. Microsatellite markers assessed the molecular diversity. Phenotypic data analyses revealed the presence of genetic diversity. The genetic variability obtained by both phenotypic and molecular information indicated the potential of the germplasm for developing base populations adapted to tropical conditions. Dry matter yield taken from 24 cuttings was used to assess the persistence. A random regression model was used to build trajectory curves of the accessions. The fitted curves showed a great amplitude regarding dry matter yield over time, which suggested a high variability regarding persistence. The three-step method for accessing persistence presented in this study included (1) a random regression model to obtain persistence trends, (2) a k-means method to define different persistence clusters, as well as (3) an ANN to perform classification of persistent accessions in an automated way. The upside of this method is to evaluate different alfalfa accessions using the same ANN. Basically, when new accessions are evaluated, they will be classified according to their genetic value scores using the same ANN previously fitted, with no need for a new clustering step. The persistence method jumps down from three to two steps and can help alfalfa breeders in the decision-making process.

Keywords: *Medicago sativa*. Self-organizing maps. Random regression models. Artificial neural network.

SUMÁRIO

GENERAL INTRODUCTION

Alfalfa (Medicago sativa L.), also known as the "queen of forages", is the most important and grown forage legume in the world (Bouton, 2012). It is an autotetraploid with 2n = 4x = 32, perennial, and is propagated by seeds. It presents self-incompatibility and self-sterility, in addition to severe inbreeding depression (Kopp, 2011). Due to the high genetic complexity of the crop (Flajoulot et al., 2005), most of the developed varieties are synthetic populations with different levels of genetic diversity.

The factors that guarantee its wide use are its high nutritional quality (~22 – 26% protein content, good palatability and digestibility, high levels of vitamins A, E, K, and minerals such as calcium, potassium, magnesium and phosphorus), high biomass production (the record yield of one alfalfa hectare is 22.4 tons), nitrogen fixation, and low seasonality (Bouton, 2012; Comeron et al., 2015). Compared to species such as corn, sugarcane, and elephantgrass, alfalfa forage has a much higher nutritional quality, which increases its importance for high-yield production systems.

Alfalfa cultivars have been developed, during the crop's long domestication, adapted to a wide range of climatic situations, from oases on the Sahara fringe to temperate zones. However, alfalfa cultivation has been more frequent in temperate climate regions, in part due to breeding efforts. The United States of America cultivated about 7.05 million alfalfa hectares in 2018, with a production of 57.8 million tons of dry matter (USDA, 2019). Other important producers are Canada and Argentina. In Brazil, alfalfa was first introduced in Rio Grande do Sul and currently occupies an area of about 40,000 hectares (Ferreira and Vilela, 2015). Although the crop has great potential for cultivation in regions with different soil and climate conditions, a large part of its production is concentrated in the southern region, especially in the States of Paraná and Rio Grande do Sul (Santos et al., 2018).

The factors that limit the use of alfalfa in Brazilian production systems include the lack of knowledge about management practices, low soil fertility, and the low availability of cultivars adapted to tropical conditions (Ferreira and Vilela, 2015). There are currently 10 alfalfa cultivars registered in Brazil. They are: KF 911 and CUF 101, which are maintained by Fazendas Reunidas do Rio de Contas; Super Leitera, from Central Riograndense de Agroinsumos; Alfafa, from Agristar do Brasil; Monarca SP INTA, from Otimiza Consultoria e Gestão Rural Ltda; WL-325 HQ, WL-525 HQ, Trifecta, and Crioula, from Itapuã Indústria e Comércio de Produtos de Alfafa Ltda., and BRS Tropluz, from Embrapa (MAPA, 2019).

Embrapa Livestock Southeast is currently focused on incorporating alfalfa into high-yield milk production systems in Brazil. Embrapa's crop improvement has been possible thanks to a partnership with Argentina's National Institute of Agricultural Technology (INTA). Since 1987, INTA has been conducting an alfalfa breeding program in partnership with private companies. The partnership agreement determines that INTA is responsible for selecting parents to compose synthetic populations whereas the partner company multiplies the seeds to sell them (Basigalup, 2016). The germplasm of Embrapa Livestock Southeast, provided by INTA, has been studied, aiming at the formation of new promising synthetic populations. This germplasm is composed of 77 accessions of alfalfa from a temperate genetic background that, once characterized and explored, can generate synthetic populations adapted to the tropical climate.

The very first step to carrying out such a characterization is to study the genetic diversity based on traits related to the main objectives of the alfalfa breeding program. Once groups of similar accessions are identified, it is possible to choose different parents and outline the planning of crosses to compose promising synthetic populations. Cruz et al. (2011) stated that the importance of studies on genetic diversity for breeding, comes down to the fact that

crosses involving genetically different parents are more suitable to maintain genetic variability in advanced generations, which is especially important for alfalfa cultivation.

Phenotypic (e.g., nutritional, physiological, morphological, and agronomic) and molecular data can be used to estimate a species' genetic diversity. Despite the fact that phenotypic data is valuable for quantifying accession diversity, it contains measurement inaccuracies that affect the final information (Silva et al. 2018). Molecular markers such as microsatellites have been successfully used to characterize the genetic diversity of alfalfa (Annicchiarico et al. 2016; Herrmann et al., 2018; Julier et al. 2018). Molecular data combined with conventional phenotypic analysis delivers better diversity comprehension and also the prospect of getting promising populations. (Julier et al. 2018).

When compared to biased phenotypic diversity estimates, DNA markers are a tool for characterizing and estimating genetic diversity neutrally. Microsatellites (SSRs) are widely adopted among the different types of markers because they have the advantage of being codominant, highly polymorphic, and easily replicated (Azevedo et al., 2012). Because of the alfalfa polyploidy, SSR markers are considered dominant in the pattern of presence or absence of the band when used in the crop. Although the presence and absence of assessment limits marker information, it can yield valuable information on allelic richness.

After the assessment of phenotypic or molecular dissimilarities, the first step in guiding crosses in alfalfa would be to perform predictive analysis to evaluate genetic diversity, as the establishment of a synthetic population demands a considerable number of parents. Predictive techniques rely on all types of data (morphological, physiological, and molecular) that is usually subjected to a dissimilarity measure. Methods such as principal component analysis, canonical variables, hierarchical clustering, and optimization have been employed as an alternative for simultaneous accessions comparison, improving the accuracy of genetic diversity estimates. (Preisigke et al., 2015). Another very interesting approach is self-organizing

maps (SOM), a computational intelligence technique that allows people to visualize patterns and classify data based on distances established between them (Kohonen, 2014).

Self-organizing maps are an unsupervised learning neural network that uses a competitive mechanism to recognize similarities between input patterns. The technique uses a specific activation function, such as the Euclidean distance, to preserve notions of distance (Nascimento et al., 2018). The learning process starts by assigning synaptic weights to different neurons, then a competitive process begins to determine which neuron is the winner. After this step, a cooperative process takes place in which the winning neuron sets the approach of other neurons. After the neighborhood has been created, the adaptation period starts, in which the synaptic weights are adjusted. After all iterations, the map is organized into a topological structure that reflects the closeness of the accessions under examination (Santos et al., 2019).

It is feasible to integrate information from molecular and phenotypic groups to guide crossings between alfalfa accessions to compose new synthetic populations and enhance the odds of success in forage breeding programs. Among the objectives of alfalfa breeding, the nutritive value of the forage stands out. This parameter is determined by different attributes such as in vitro dry matter digestibility, considered the best individual criterion of forage nutritive value for ruminants (Wilkins, 2003). Tolerance to biotic and abiotic stresses it is also important, as well as dry matter yield, which will always be a key trait as production costs dissipate as yield increases. However, all the other traits have to be present in a persistent material. Persistence is one of the main factors affecting perennial crops over time. Persistence can be defined as the forage's capacity to re-establish itself satisfactorily after one or more harvests (Jones and Tracy, 2017).

The regrowth ability of forage species is reduced over time. Weeds typically invade crop areas shortly after harvest, reducing forage yield and nutritive value (Wilkins, 2003). This process is slower in areas with persistent plants, allowing for better use and a longer useful life

of the cultivated areas. Combining high persistence and high yield in alfalfa is not straightforward, especially if the forage is breed to produce silage or hay. According to Takasaki et al. (1989) forage species that develop a high proportion of tillers tend to have less persistence and retain fewer non-structural carbohydrates compared to species that have a low proportion. For all of the mentioned reasons, it is critical that alfalfa cultivars remain persistent to maintain yield across future harvests.

In order to measure persistence in alfalfa, several harvests must be evaluated at different seasons of the year. From an appropriate statistical model, it is possible to model trajectory curves that characterize each accession. This type of multidimensional data (created by the continual evaluation of a trait throughout time) is called longitudinal (Fitzmaurice et al., 2011; Resende et al., 2014).

Some characteristics of longitudinal data or repeated measures over time are: the presence of positive correlation between different measurements taken in the same individual, which invalidates the important prerequisite of independence required by various statistical techniques, and heterogeneity of variances (Sun et al., 2017). According to Resende et al. (2014), there are three major approaches for repeated measures over time: (i) Analysis of Variance (ANOVA) of repeated measures (if the sphericity condition is met), (ii) covariance patterns in the multivariate matrix, and (iii) fitting curves or polynomial equations in quantitative variables. Among these options, (iii) have been extensively used to model growth patterns in plants and animals (Su et al., 2017; Rocha et al., 2018).

Random regression models are used to predict the trajectories of observations taken over time. In other words, the trajectory or growth pattern is adjusted through covariance functions while also accessing the variation among trajectories. Some trajectories have a continuous pattern, whilst others seem to increase or decrease over time (Schaeffer, 2016).

Fitting random regression models in the context of yield persistence involves both fixed (years, harvests, etc) and random effects (genetic and permanent environment). The covariance function used to fit the trajectory of a genotype can adjust fixed (one function for each level of effect) and random effects (model order determined from testing different models). Then, the genetic trajectory and permanent environmental curves are adjusted for each genotype. The residual matrix is also incorporated in the mixed model equations, which indicates that the observations are weighted by the magnitude of the residual variance. The greater the residual variance, the lesser weight the equation estimates. As longitudinal data has particular characteristics regarding residuals, it is important to test different structures in order to improve the fit of random regression models. If residuals from harvests obtained in the same season of the year, or from harvests within the same year do not differ, then it is possible to adjust to a less parameterized model. The first step in adjusting a good random regression model to obtain persistence curves is to plot the data and understand its general trend, and test different models and structures before using it (Schaeffer, 2016).

REFERENCES

Annicchiarico P, Nazzicari N, Ananta A, Carelli M, Wei Y, Brummer EC (2016) Assessment of Cultivar Distinctness in Alfalfa: A Comparison of Genotyping-by-Sequencing, Simple-Sequence Repeat Marker, and Morphophysiological Observations. Plant Genome 9:1-12. doi: 10.3835/plantgenome2015.10.0105

Annicchiarico P, Barrett B, Brummer EC, Julier B, Marshall AH (2015) Achievements and challenges in improving temperate perennial forage legumes. *Critical Reviews in Plant Sciences*, 34:327-380. doi: g/10.1080/07352689.2014.898462

Araújo SI (2005) Uso de modelos de regressão aleatória na análise de dados longitudinais no melhoramento genético vegetal. Tese (Doutorado em Genética e Melhoramento) - Universidade Federal de Viçosa. Viçosa, 111p.

Azevedo ALS, Costa PP, Machado JC, Machado MA, Pereira AV, Lédo FJS (2012) Cross Species Amplifcation of Pennisetum glaucum Microsatellite Markers in Pennisetum purpureum and Genetic Diversity of Napier Grass Accessions. Crop Sci. 52:1776-1785. doi: 10.2135/cropsci2011.09.0480

Basigalup DH (2016) Producción de alfalfa en Argentina. In: Jornada Nacional de Forrajes Conservados, 7, Buenos Aires. [Resúmenes…] Buenos Aires: Instituto Nacional de Tecnología Agropecuaria. p. 83-85.

Biazzi E, Nazzicari N, Pecetti L, Brummer EC, Palmonari A, Tava A, et al. (2017) Genome-wide association mapping and genomic selection for alfalfa (Medicago sativa) forage quality traits. PLoS One 12:e0169234. doi:10.1371/journal.pone.0169234

Comeron EA, Ferreira RP, Vilela D, Kuwahara FA, Tupy O (2015) Utilização da alfafa em pastejos para alimentação de vacas leiteiras. In: Ferreira RP, Vilela D, Comeron EA, Bernardi ACC, Karam D (Ed.). *Cultivo e utilização da alfafa em pastejo para alimentação de vacas leiteiras*. Brasília: Embrapa, p.13-16.

Creste S, Tulmann Neto A, Figueira A (2001) Detection of single sequence repeat polymorphisms in denaturing polyacrilamide sequencing gels by silver staining. Plant Mol. Biol. Rep. 19:299-306. doi: 10.1007/BF02772828

Cruz CD, Bhering LL, Ferreira RP (2015) Procedimentos biométricos aplicados ao melhoramento genético da alfafa. In: R.P. Ferreira, D.H. Basigalup, and J.O. Gieco, editors, Melhoramento Genético da Alfafa. Embrapa Editor, São Carlos, Brazil. p.225-260.

Cruz CD, Ferreira FM, Pessoni LA (2011) Biometria aplicada ao estudo da diversidade genética. Visconde do rio Branco: Suprema. 620p.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical bulletin 19:11–15.

Ferreira RP, Basigalup DH, Vasconcelos ES, Cruz CD, Pereira AV (2008) Genética quantitativa e métodos de melhoramento em alfafa. In: Ferreira, R.P., Rassini, J.B., Rodrigues, A.A., Freitas, A.R., Camargo, A.C., Mendonça, F.C., editors, Cultivo e utilização da alfafa nos trópicos. Embrapa Editor, São Carlos, Brasil. p.171-205.

Ferreira RP, Vilela D (2015) Potencial de utilização da alfafa. In: Ferreira RP, Vilela D, Comeron EA, Bernardi ACC, Karam D (Ed.). Cultivo e utilização da alfafa em pastejo para alimentação de vacas leiteiras. Brasília: Embrapa, p.13-16.

Fitzmaurice GM, Laird NM, Ware JH (2011) Applied longitudinal analysis 2nd ed. Boston: Wiley, 998p.

Herrmann D, Flajoulot S, Barre P, Huyghe C, Ronfort J, Julier B (2018) Comparison of morphological traits and molecular markers to analyse diversity and structure of alfalfa (*Medicago sativa* L.) cultivars. Genet. Resour. Crop. Evol. 65:527-540. doi: 10.1007/s10722-017-0551-z

Jones GB, Tracy BF (2017) Persistence and productivity of orchardgrass and orchardgrass/alfalfa mixtures as affected by cutting height. Grass and Forage Science, 73:544-552. doi: 10.1111/gfs.12309

Julier B, Barre P, Lambroni P, Delaunay S, Thomasset M, Lafaillette F, et al. (2018) Use of GBS markers to distinguish among lucerne varieties, with comparison to morphological traits. Mol. Breeding 38:133. doi: 10.1007/s11032-018-0891-1

Kirkpatrick M, Lofsvold D, Bulmer M (1990) Analysis of the inheritance, selection and evolution of growth trajectories. Genetics, 24:979-993.

Kohonen T (2014) MATLAB Implementations and Applications of the Self-Organizing Map. Helsinki: Unigrafia Oy

Kopp MM (2011) Origem, evolução e domesticação da alfafa. In: Ferreira RP, Basigalup DH, Gieco JO (Ed), Melhoramento Genético da Alfafa. Embrapa Editor, São Carlos, Brazil. p.225-260.

MAPA (2019) CultivarWeb. Disponível em: http://sistemas.agricultura.gov.br/snpc/cultivarweb/cultivares_registradas.php. Acesso em: 22/03/2019.

Nascimento M, Nascimento ACC, Cruz CD (2018) SOM - Mapas Auto-Organizáveis de Kohonen. In: Nascimento M. and Cruz CD (Eds.). Inteligência computacional aplicada ao melhoramento genético. Viçosa: Editora UFV. 414p.

Preisigke SC, Neves LG, Araújo KL, Barbosa NR, Serafim ME, Krause W (2015) Multivariate analysis for the detection of passiflora species resistant to collar rot. Bioscience Journal, 31: 1700-1707.

Rassini JB (2002) Manejo da água na irrigação da alfafa num Latossolo Vermelho-Amarelo. Pesquisa Agropecuária Brasileira, 37:503-507.

Santos IG, Carneiro VQ, Silva Junior AC, Cruz CD, Soares PC (2019) Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. Acta Sci. Agron. 41:e39803. doi: 10.4025/actasciagron.v41i1.39803

Schaeffer LR (Ed.) (2016). *Random Regression Models*. Available at: http://animalbiosciences.uoguelph.ca/~lrs/BOOKS/rrmbook.pdf.

Silva MJ, Pastina MM, Souza VF, Schaffert RE, Carneiro PCS, Noda RW, et al. (2018) Phenotypic and molecular characterization of sweet sorghum accessions for bioenergy production. PLoS ONE 12: e0183504. doi: 10.1371/journal.pone.0183504

Takasaki Y, Isoda A, Nojima H, Oizumi H (1989) Behaviours of annual and perennial grass species in the same genus. In Proceedings of the XVI International Grassland Congress, Nice, France. vol. 1, 449– 450. Nice: Association Franc¸aise pour la Production Fourrage´re.

USDA (2019) Crop Production 08/10/2018. Disponível em: https://www.nass.usda.gov/Publications/Todays_Reports/reports/crop0818.pdf. Acesso em: 22/03/2019.

Wilkins PW, Humphreys MO (2003) Progress in breeding perennial forage grasses for temperate agriculture. Journal of Agricultural Science, 140:129-150. doi: 10.1017/S0021859603003058

CHAPTER 1

EXPLORING THE DIVERSITY OF ALFALFA WITHIN BRAZIL FOR TROPICAL PRODUCTION

VIÇOSA – MINAS GERAIS

2021

RESUMO

SANTOS, Iara Gonçalves dos, D.Sc., Universidade Federal de Viçosa, julho de 2021. **Exploração da diversidade genética de alfafa no Brasil para produção tropical.** Orientador: Cosme Damião Cruz.

Alfafa (*Medicago sativa* L.) é uma leguminosa forrageira de grande interesse devido às suas características nutricionais que garantem alto retorno para os sistemas especializados de produção de leite. Apesar da alfafa apresentar potencial para cultivo em diferentes regiões edafoclimáticas, a produção da forrageira em regiões tropicais é limitada. Os objetivos desse estudo foram realizar a caracterização fenotípica e molecular do germoplasma de alfafa da Embrapa Pecuária Sudeste e identificar o potencial do germoplasma para gerar populações-base adaptadas a regiões tropicais. Os valores genéticos de nove características de acessos de alfafa avaliadas em 77 acessos de *background* genético temperado foram obtidos para caracterizar a diversidade fenotípica. Marcadores microssatélite foram utilizados para acessar a diversidade molecular. Foi detectada variância nas informações fenotípicas pela análise de deviance para as características avaliadas. A correlação entre matrizes de dissimilaridade baseadas em valores genéticos e informações moleculares foi baixa. Baseado nas informações moleculares e fenotípicas, uma população-base promissora seria composta por Pro INTA Patricia, Pro INTA Super Monarca, Mecha, Magna 601, WL 525, ACA 900, Bacana, CUF 101, Crioula e Ruano. Esses acessos possuem pelo menos 50 parentais, apresentam alta produção de matéria seca além de alelos favoráveis para habilidade de rebrota, produtividade e persistência. A variabilidade genética observada para a maioria das características indica o alto potencial do germoplasma para o desenvolvimento de populações-base adaptadas ao clima tropical. Uma vez que populações sintéticas forem selecionadas, elas poderão integrar programas de melhoramento de alfafa em outras regiões tropicais no mundo.

Palavras-chave: Mapas auto-organizáveis. SSR. BLUP.

ABSTRACT

SANTOS, Iara Gonçalves dos, D.Sc., Universidade Federal de Viçosa, July, 2021. **Exploring the Diversity of Alfalfa within Brazil for Tropical Production.** Adviser: Cosme Damião Cruz.

Alfalfa (*Medicago sativa* L.) is a forage legume of great interest because of its role in milk production schemes. Although it has the potential to be cultivated in different edaphoclimatic regions, the fodder production in tropical regions is limited. The objectives of this study were to perform phenotypic and molecular characterization of alfalfa germplasm and to identify the potential of the germplasm to generate base populations adapted to tropical conditions. The genetic values of nine traits from seventy-seven alfalfa accessions of a genetic background amenable to a temperate climate were obtained to characterize phenotypic diversity, and microsatellite markers were used to assess molecular diversity. Phenotypic information based on joint deviance analysis revealed the presence of genetic diversity. The correlation between the dissimilarity matrices of genetic values and molecular data was low. Based on phenotypic and molecular data, a great base population would be composed of Pro INTA Patricia, Pro INTA Super Monarca, Mecha, Magna 601, WL 525, ACA 900, Bacana, CUF 101, Crioula, and Ruano. These populations have at least 50 distinct parents, presented high dry matter yield, besides favorable alleles for regrowth ability, biomass yield, and persistence. The genetic variability observed for most traits indicates a high potential for the development of alfalfa base populations adapted to the tropical condition. Once adapted synthetic alfalfa populations are selected, they may be integrated into breeding programs in other tropical regions of the world.

Keywords: Self-organizing maps. SSR. BLUP.

INTRODUCTION

Alfalfa is a forage legume that produces a high yield and protein content. It is palatable and digestible and has a high capacity of nitrogen fixation in soil. In addition, it has low yield seasonality and high vitamin and mineral levels (Annicchiarico et al. 2015). These attributes guarantee the possibility of using this crop in specialized dairy herds, with excellent results for milk production (Comeron et al. 2015).

In intensive milk production systems, using alfalfa as part of animal feed reduces production costs (Ferreira et al. 2008). In Brazil, most of the alfalfa production is concentrated in the South Region, especially in the states of Paraná and Rio Grande do Sul (Santos et al. 2018). However, there is great potential for cultivation in regions with different soil and climatic conditions. Alfalfa production in Brazil is limited by the low availability of cultivars adapted to tropical conditions (Ferreira and Vilela 2015).

Cultivated alfalfa is autotetraploid with $2n = 4x = 32$, it is perennial, propagated by seeds, and exhibits autoincompatibility and autosterility, along with a severe inbreeding depression (Kopp 2011). Due to the large genetic complexity of the crop (Flajoulot et al. 2005), most cultivated varieties are synthetic populations with different levels of genetic diversity.

Currently, Embrapa Southeast Livestock maintains research related to the incorporation of alfalfa in dairy production systems. Seventy-seven accessions have been characterized, aiming for the formation of promising synthetic populations. It can be considered that the only cultivar well adapted to Brazilian climatic conditions is the Crioula (Ferreira and Vilela 2015). It is worth mentioning that there is low genetic progress reported for the crop, especially for fodder quality (Annicchiarico et al. 2015; Biazzi et al. 2017). Therefore, quantifying the genetic diversity of this germplasm is fundamental to guiding and accelerating the actions of the alfalfa breeding program.

The genetic diversity of the species can be based on phenotypic data (e.g., nutritive value in addition to physiological, morphological, and agronomic traits) and molecular data (e.g., single nucleotide polymorphisms and single sequence repeats), among other characteristics. Although the phenotypic data offer valuable contributions for quantifying the diversity between accessions, they carry measurement errors that influence the final information (Silva et al. 2017). This fact does not reduce the importance of having phenotypic information. Some small institutions have only phenotypic information due to the costs of obtaining molecular data. Molecular markers, such as microsatellites, have been successfully used to characterize the genetic diversity of alfalfa (Annicchiarico et al. 2016; Herrmann et al. 2018; Julier et al. 2018).

Molecular information combined with conventional phenotypic analysis increases the understanding of alfalfa diversity and the chances of obtaining promising populations (Julier et al. 2018). Flajoulot et al. (2005) identified high diversity in an alfalfa breeding pool using phenotypic and molecular information. However, considering the traits of interest, this pool would not be able to generate a large variety differentiation. Annicchiarico et al. (2017) highlighted the importance of combining phenotypic and molecular information. In their study, among-population variance was over eightfold smaller than the within-population variance indicating of modest genetic differentiation between populations even for geographically distant alfalfa germplasm. Even though investigating genetic diversity of alfalfa germplasm is common in temperate regions, there is a lack of information on the performance and genetic diversity of alfalfa germplasm in Brazilian climatic conditions.

Self-organizing maps (SOM) have shown to be efficient in capturing genetic diversity in populations under genetic drift, inbreeding, selection, and migration (Oliveira et al. 2020) and can be useful to map genetic diversity of alfalfa accessions. Regardless of the kind of data (phenotypic or molecular) they are useful for visualizing patterns. SOMs are a class of two-dimensional artificial neural networks that organize data, preserving neighborhood notions

through the adoption of a specific activation function, such as Euclidian distance (Kohonen 2014). Briefly, SOM learning is achieved in three states. Firstly, synaptic weights are attributed to different neurons, and then a competition process begins. The set of genetic values of each accession is allocated to the neuron that best represents it (this neuron is called "winner"). Secondly, the cooperation stage begins. The winning neuron determines the approximation of the other neurons in the order of proximity. Finally, the neurons that establish their neighborhood proceed to the adaptation phase, where there is a weight adjustment. After all iterations, the map is organized in a topological structure that reflects the proximity among populations under study (Santos et al. 2019).

The objectives of this study were (i) to perform phenotypic and molecular characterization of an alfalfa germplasm, (ii) to examine the relationship between dissimilarity-based phenotypic and molecular information and (iii) identify the potential of the germplasm to generate base populations adapted to the tropical condition.

MATERIALS AND METHODS

Experimental information. Seventy-seven alfalfa accessions of a temperate genetic background were evaluated in eight different cuttings, and each was analyzed in different months from 2015 to 2017 [11/12/2015, 02/03/2016, 05/09/2016, 08/12/2016, 11/28/2016, 02/14/2017, 05/26/2017, 08/28/2017]. The accessions constitute synthetic populations developed by the National Agricultural Technology Institute (INTA) and evaluated by Embrapa Southeast Livestock (Table S1). Since 1987, INTA has been conducting an alfalfa breeding program in partnership with private companies. During the process INTA is responsible for selection of parents for the formation of synthetic populations while partner company multiplies the seeds and markets them (Basigalup 2016). The parents of the populations that compose this germplasm were selected duo to their persistence besides resistance to the main alfalfa pests

and diseases. The experiment was performed at the experimental field of Embrapa Southeast Livestock, located in the municipality of São Carlos, São Paulo, Brazil [21° 57'42 "S, 47° 50'28" W, 860 m].

The experiment was carried out in randomized complete blocks with three replicates. The plots consisted of four rows of four meters in length spaced 0.20 m apart, and the useful area corresponded to the two central rows, eliminating 0.50 m at the ends of the lines. Sowing was performed on October 2015 using a seeding rate of 20 kg ha^{-1}. Seeds of the accessions were inoculated with strains of *Rhizobium melilotti* - SEMIA 116. After each cutting, the plants received cover fertilization according to the soil analysis. The irrigation management was performed by a central pivot system, according to Rassini (2002). Cultural management procedures for pests and diseases were carried out according to Ferreira et al. (2008).

Phenotypic data. Nine traits related to nutritive value and forage yield were phenotypically characterized. The traits were selected based on the list of alfalfa descriptors for cultivar registration in Brazil (MAPA 2019). Agronomic traits included the following: plant height (Ht, in cm), measured from the ground to the top of the inflorescence in the day of harvest; dry matter yield (DMY, in kg ha-1), obtained by manually cutting the plants at eight to 10 cm above ground when each cultivar reached the flowering stage; and disease score (DS), determined by values from 0 to 3 (0 - high incidence; 1 - medium incidence; 2 - low incidence; 3 - no disease), according to the percentage of leaf area attacked by *Leptosphaerulina briosiana*; *Cercospora medicaginis*; *Phoma medicaginis*; *Leptotrochila medicaginis*, and *Uromyces striatus* in each plot, according to Vasconcelos et al. (2010).

The samples that were used to determine the nutritive value traits were collected before each cut, air-dried at 65°C, and then they were ground through a 1 mm sieve. The micro-Kjedahl method was used to determine the nitrogen content. Crude protein content (CP, in %) was calculated by multiplying the nitrogen content by 6.25 (CP = N x 2.25). Neutral detergent fiber

(NDF, in %), acid detergent fiber (ADF, in %), lignin (L, in %), and in vitro dry matter digestibility (IVDMD, in %) were quantified according to the methods of Van Soest et al. (1963) and Goering and Van Soest (1970). Stem/leaf ratio (SLR, in %) was calculated by determining the ratio between the number of stems and leaves collected in each accession.

Molecular data. New and healthy leaves were collected in each genotype block to compose a sample of each accession. Each sample was collected in the field and immediately immersed in liquid nitrogen. DNA extraction was performed according to the CTAB (cetyltrimethylammonium bromide) protocol modified by Doyle & Doyle (1987). Quantification of the extracted DNA was checked in a NanoDrop® ND-1000 Spectrophotometer, and the concentration of each DNA sample was obtained in μg/μL. The samples were diluted to 10 μg/μL concentration in Milli-Q water and stored in a freezer at -20°C for use in subsequent polymerase chain reaction (PCR) experiments. DNA quality was checked in agarose gel (1%) stained with ethidium bromide.

Twenty-nine microsatellite (SSR) markers previously described in the literature were selected (Table 1) and the PCR protocol consisted of 30 ng of DNA, 10 μL of "5X PCR reaction buffer" and 1.25 U of Taq DNA Polymerase enzyme (both supplied by Promega), 10 mM dNTP, and 20 pmol of each primer, with a final reaction volume of 20 μL. Polymerase chain reactions were performed on a thermocycler with initial denaturation step at 94°C (5 min), 35 cycles of 94°C (1 min), primer pair specific annealing temperatures of 60 or 55°C (1 min), and 72°C (1 min), and a final extension cycle at 72°C (7 min). Amplification products stained with ethidium bromide were separated on a 3% agarose gel with electrophoresis using a current of 100 V for one hour to separate fragments. Gels were analyzed in a Gel Doc ™ XR + (Bio-Rad) photodocumentation system. Successfully amplified PCR products were run on 6% denaturing polyacrylamide gel and stained with silver nitrate solution, according to Creste et al. (2001).

The alleles were visualized on a white light table, and their size was estimated in comparison to the 10 bp ladder (Invitrogen) (Figure 1).

Table 1. SSR markers selected in the literature.

SSR	References	GL[a]	References	QTL[b,c]	References
aa660573	Eujayl et al., 2003; Sledge et al., 2005	4	Robins et al., 2007a	Y, H, R	Robins et al., 2007b
al372288	Eujayl et al., 2003; Sledge et al., 2005	7	Robins et al., 2007a	Y, H, P	Robins et al., 2007b; Robins et al., 2008
al373004	Eujayl et al., 2003; Sledge et al., 2005	7	Robins et al., 2007a	B, P	Robins et al., 2007a; Robins et al., 2008
aw686836	Eujayl et al. 2003; Sledge et al., 2005	5	Robins et al., 2007a	B	Robins et al., 2007a
aw695900	Eujayl et al. 2003; Sledge et al., 2005	5	Robins et al., 2007a	B	Robins et al., 2007a
aw775062	Eujayl et al. 2003; Sledge et al., 2005	5	Robins et al., 2007a	B	Robins et al., 2007a
bf648174	Eujayl et al. 2003; Sledge et al., 2005	7	Robins et al., 2007a	Y, H	Robins et al., 2007b
bg449206	Eujayl et al. 2003; Sledge et al., 2005	5	Robins et al., 2007a	B	Robins et al., 2007a
bg645450	Eujayl et al., 2003; Sledge et al., 2005	7	Robins et al., 2007a	P	Robins et al., 2008
MTR58	Julier et al., 2003; Odorizzi et al., 2015	1	Julier et al., 2003	-	-
B21E13	Julier et al., 2003; Odorizzi et al., 2015	2	Julier et al., 2003	-	-
MTIC103	Julier et al., 2003; Odorizzi et al., 2015	8	Julier et al., 2003	-	-
MTIC210	Julier et al., 2003; Odorizzi et al., 2015	2	Julier et al., 2003	-	-
MTIC247	Julier et al., 2003; Odorizzi et al., 2015	1	Julier et al., 2003	-	-
MTIC451	Julier et al., 2003; Flajoulot et al., 2005; Grandon et al., 2013	2	Julier et al., 2003	-	-
B14B03	Julier et al., 2003; Flajoulot et al., 2005; Grandon et al., 2013	5	Julier et al., 2003	-	-
MTIC432	Julier et al., 2003; Flajoulot et al., 2005; Grandon et al., 2013	7	Julier et al., 2003	-	-
MTLEC2A	Diwan et al., 1997; Grandon et al., 2013	3	Diwan et al., 1997	-	-
AFct32	Diwan et al., 1997; Grandon et al., 2013	3	Diwan et al., 1997	-	-
AFca11	Diwan et al., 1997; Grandon et al., 2013	6	Diwan et al., 1997	-	-
be126	Qiang et al., 2015	4	Li et al. 2011; Narasimhamoorthy et al. 2007; Sledge et al. 2005	-	-
aw267840	Qiang et al., 2015	4	Li et al. 2011; Narasimhamoorthy et al. 2007; Sledge et al. 2005	-	-
aa05	Qiang et al., 2015	5	Li et al. 2011; Narasimhamoorthy et al. 2007; Sledge et al. 2005	-	-
aw01	Qiang et al., 2015	6	Li et al. 2011; Narasimhamoorthy et al. 2007; Sledge et al. 2005	-	-
bg288	Qiang et al., 2015	7	Li et al. 2011; Narasimhamoorthy et al. 2007; Sledge et al. 2005	-	-
mtic188	Qiang et al., 2015	8	Li et al. 2011; Narasimhamoorthy et al. 2007; Sledge et al. 2005	-	-
MTIC299	Julier et al., 2003; Flajoulot et al., 2005	8	Julier et al., 2003	-	-
MTIC189	Julier et al., 2003; Flajoulot et al., 2005	3	Julier et al., 2003	-	-
MTIC93	Julier et al., 2003; Flajoulot et al., 2005	6	Julier et al., 2003	-	-

[a] Linkage group; [b] Quantitative Trait Loci; [c] Y is yield, H is height, R is regrowth ability, B is biomass yield, and P is persistence.

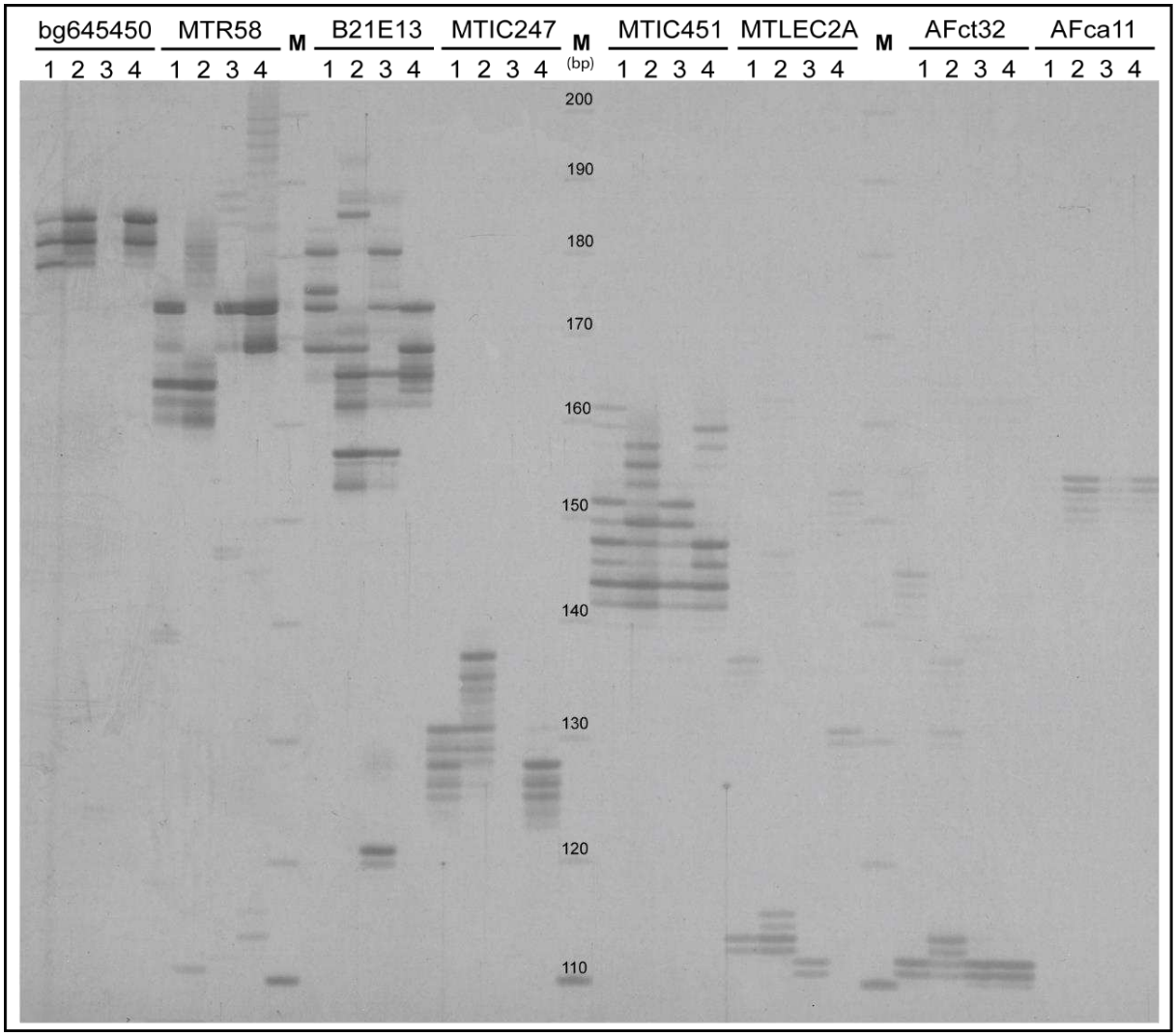


Figure 1. Microsatellite amplification profile of markers *bg645450*, *MTR58*, *B21E13*, *MTIC247*, *MTIC451*, *MTLEC2A*, *AFct32*, *AFca11* in 6% polyacrylamide gels. Individuals 1 to 4 are indicated. M = 10 pb ladder (Invitrogen).

Statistical Model. Phenotypic data were analyzed using the mixed model as follows:

$$y = Xm + Zg + Wp + Ti + \varepsilon$$

where y is the vector of phenotypic data, m is the vector of fixed effects of the cutting-replication combination added to the overall mean, g is the random vector of genetic effects with $g \sim N(0, \sigma_g^2)$, p is the random vector of permanent environment (plots), i is the vector of the genotype x measurement effects with $i \sim N(0, \sigma_i^2)$, and ε is the vector of random residuals with $\varepsilon \sim N(0, \sigma^2)$. X, Z, W and T are the incidence matrices for the described effects.

The random effects were tested by a Likelihood Ratio Test (LRT) at 5% probability. The adjusted means for each trait considering each accession were obtained via Best Linear Unbiased Predictor (BLUP), and the variance components were estimated via Residual Maximum Likelihood (REML). The analyses of mixed models were performed using the software Selegen REML/BLUP (Resende 2016). Genotypic correlation between pairs of traits was estimated based on the Pearson's correlation coefficient for the genetic values, which correspond to the BLUPs, using the R package PerformanceAnalitics (R Core Team 2017).

Genetic and Molecular Diversity Measurements. To quantify genetic diversity, the dissimilarity matrix was calculated by the BLUPs Average Euclidean distance, using the software Selegen-REML/BLUP (Resende 2016). Molecular diversity analysis began with filtering of the SSRs by considering a minimum band presence of 5% and a call rate less than 95%. The dissimilarity matrix was obtained by the software Genes (Cruz 2013), using the arithmetic complement of the Jaccard index, applied to binary patterns of multicategorical variables with variable class numbers, as follows:

$$d_{ii'} = \frac{1}{v}\sum \frac{(b + c)}{(a + b + c)}$$

where dii' is the dissimilarity value obtained for the marker pair i and i', a is the number of coincidences of type 1-1 for each marker pair, b is the number of disagreements of type 1-0 for each marker pair, and c is the number of disagreements of type 0-1 for each marker pair.

To visualize the diversity of accessions, SOMs were constructed, using the genetic values of the accessions as the inputs in this analysis. No outputs were stipulated a priori because creating a SOM is an unsupervised technique. The maps were designed in a 10 by 10 arrangement, which represents a map of 10 lines of 10 neurons each, with hexagonal topology and 5000 iterations. To ensure clustering consistency and repeatability, 10 independent replicates of the maps were constructed using the software MATLAB (MATLAB 2010) integrated with the software Genes (Cruz 2016).

RESULTS

Phenotypic Data. The LRT detected significant differences ($p < 0.01$) among genotypic effects for the accessions in all traits, except for *in vitro* dry matter digestibility and acid detergent fiber (table 2). The accessions x-cuttings interaction demonstrated significant effects ($p < 0.01$) for plant height, dry matter yield, stem/leaf ratio, crude protein, and neutral detergent fiber.

Table 2. Combined deviance analysis for eight alfalfa cuttings regarding plant height - Ht (cm), dry matter yield - DMY ($kg.ha^{-1}$), disease score - DS (%), stem / leaf ratio - SLR (%), lignin - L (%), crude protein - CP (%), *in vitro* dry matter digestibility - IVDMD (%), neutral detergent fiber - NDF (%) and, acid detergent fiber - ADF (%).

Traits	Effects	
	Accessions[a]	Accessions x-Cuttings[b]
Ht	111.2**	155.09**
DMY	42.58**	33.18**
DS	36.15**	0.28^{ns}
SLR	11.77**	14.58**
L	26.63**	0.15^{ns}
CP	47.02**	26.65**
IVDMD	2.41^{ns}	0.39^{ns}
NDF	18.23**	8.45**
ADF	3.76^{ns}	3^{ns}

[a]Likelihood ratio test using the chi-squared test with one degree of freedom, reduced model without accession effect. [b]Likelihood ratio test using the chi-squared test with one degree of freedom, reduced model without accessions x-cuttings effect. **, *Significant according to a chi-squared test at the 0.01 and 0.05 probability level, respectively. ^{ns}Not significant at the 0.05 probability level by the chi-squared test.

The correlations among nutritive value traits were significant according to a t-test ($p<0.05$), with the exception of some correlations involving lignin, such as the correlation between lignin and crude protein (0.060), lignin and *in vitro* dry matter digestibility (0.016), and lignin and stem/leaf ratio (0.022) (Figure 2). All correlations involving the agronomic traits were significant. In addition, the correlations involving crude protein or *in vitro* dry matter digestibility and any agronomic traits had high and negative values.

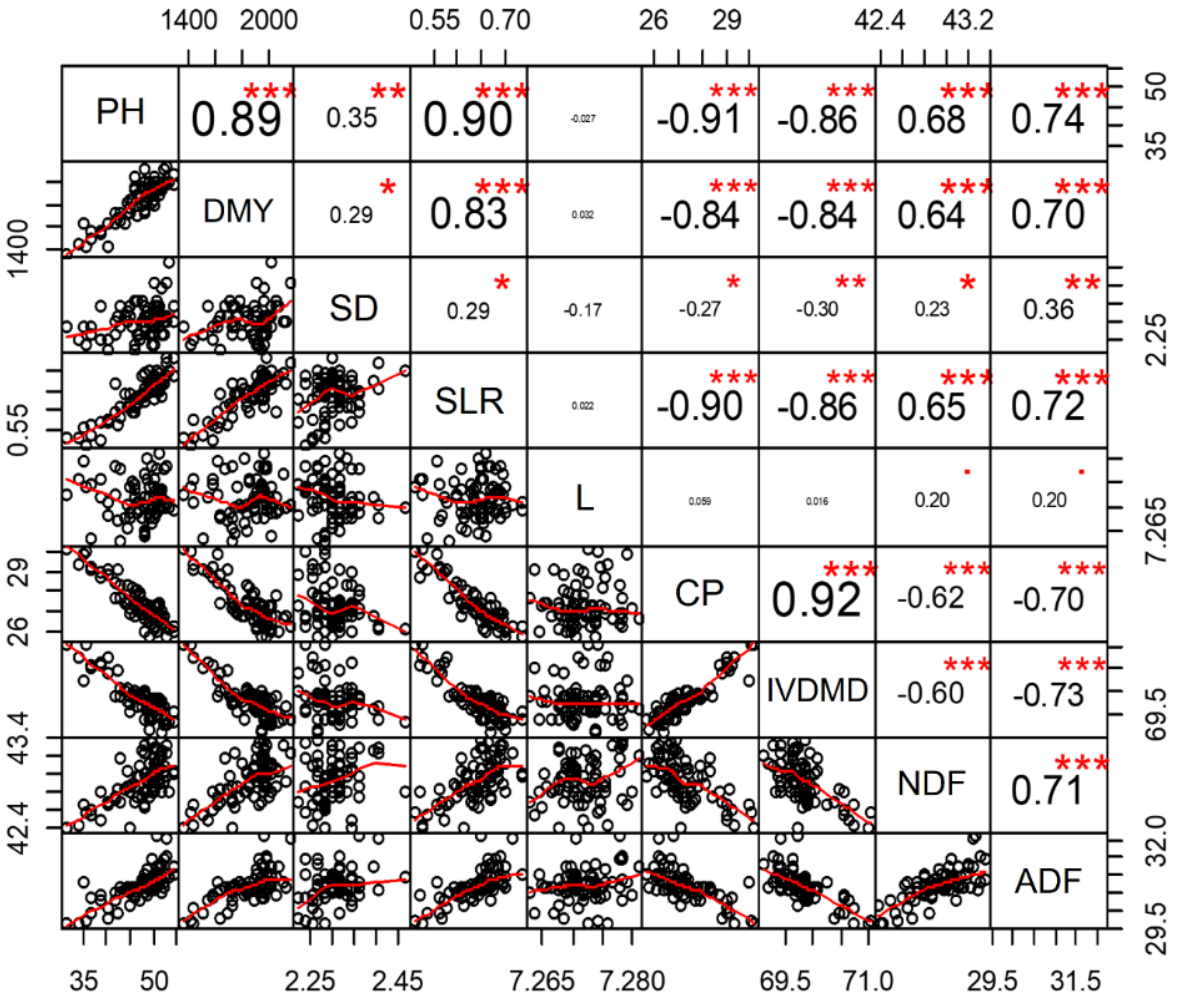


Figure 2. Genotypic correlations between alfalfa traits. Above the main diagonal are the correlation values between pairs of traits; the size of the numbers in the figure is proportional to the value of the correlation. Below the main diagonal is the correlation direction (positive or negative) between the corresponding pairs of traits. Plant height - Ht, dry matter yield - DMY, disease score - DS, stem / leaf ratio - SLR, lignin - L, crude protein - CP, in vitro dry matter digestibility - IVDMD, neutral detergent fiber - NDF and, acid detergent fiber - ADF. *, ** and ***, significant according to a t-test at the 0.001, 0.01 and 0.05 probability level, respectively.

Molecular Data. After conducting quality control, 10 of the 15 markers that successfully amplified their targets were selected (Table 3). Due to alfalfa polyploidy, SSR were evaluated as dominant markers in a pattern of presence and absence; thus, a total of 54 alleles were evaluated. Of the selected markers, two were monomorphic, and the others were polymorphic.

The *mtic 189* and *mtic 451* markers exhibited the highest allele numbers, with a total of 10 for each.

Table 3. SSRs used to determine the genetic diversity of alfalfa accessions

SSR name		Primers sequences (5' - 3')	Linkage group	Annealing temperature (°C)	Number of alleles per locus
al373004	forward	CAACCAACTCAATGCCACTC	7	60	1
	reverse	ACTTTGGAGCCATCATCACA			
aw686836	forward	TTTTATTTGTGGTCATTAGCCTCT	5	60	2
	reverse	GAAACCAAGATCCCCACACA			
aw695900	forward	GCAACCATCTAAACCCAACAA	5	60	1
	reverse	AGGCTAATCGACGGGAAAAT			
MTR58	forward	GAAGTGGAAATGGGAAACC	1	55	9
	reverse	GAGTGAGTGAGTGTAAGAGTGC			
B21E13	forward	GCCGATGGTACTAATGTAGG	2	55	7
	reverse	AAATCTTGCTTGCTTCTCAG			
MTIC451	forward	GGACAAAATTGGAAGAAAAA	2	55	10
	reverse	AATTACGTTTGTTTGGATGC			
MTLEC2A	forward	CGGAAAGATTCTTGAATAGATG	3	55	3
	reverse	TGGTTCGCTGTTCTCATG			
AFca11	forward	CTTGAGGGAACTATTGTTGAGT	6	55	5
	reverse	AACGTTTCCCAAAACATACTT			
aw01	forward	ACCTGTTCTAAGGGAGATTTCG	6	55	6
	reverse	CAGGGGAAGCATACAAAACC			
MTIC189	forward	CAAACCCTTTTCAATTTCAACC	3	55	10
	reverse	ATGTTGGTGGATCCTTCTGC			

The markers *al373004*, *aw686836* and *aw695900*, known to be in regions of Quantitative Trait Loci (QTL) for dry matter yield and persistence (Robins et al. 2007a; Robins et al. 2008), were identified in all accessions. One allele was identified for the marker *al373004*, two for the marker *aw686836*, and one for the marker *aw695900*. The accessions with the highest number of different alleles were Activa, CUF 101, Don Enrique, LPS 8500, Prointa Patrícia 1, Verdor and Crioula, with three alleles each.

Genetic Diversity. The 10 SOMs, which were generated by genetic values, were used to organize the accessions into 11 clusters. As the organization of accessions was the same in all 10 maps, only one is shown here (Figure 3). Clusters G1 and G2 exhibited the highest averages for dry matter yield, while clusters G5, G6, and G7 were characterized by the smallest means for this trait (Table 4). Although the nutritive value traits had few variations, there was an observed tendency of maps to group the accessions according to correlations. Considering the crude protein trait, the clusters of higher means for this trait were those with the lowest dry matter yield. Considering the traits stem/leaf ratio, neutral detergent fiber, and acid detergent fiber, those groups of higher averages for dry matter yield concentrated the highest means for these traits.

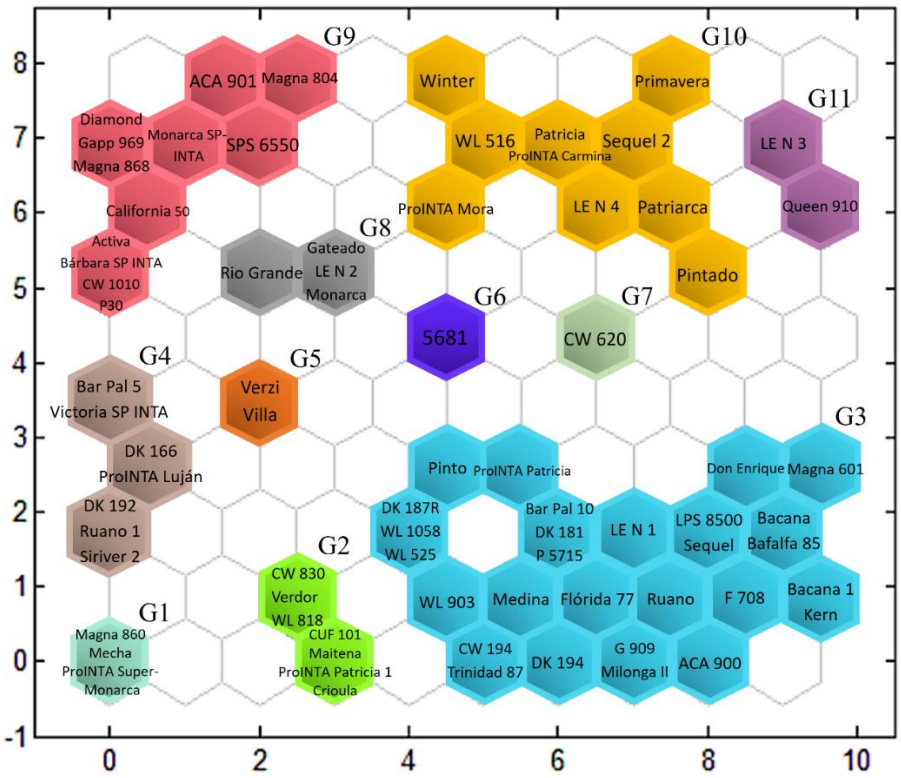


Figure 3. Self-organizing map for the phenotypic data. Accessions belonging to the same group were identified using equal colors on the map. Numbers on x- and y-axis represent the neurons of the 10 × 10 topology. G1, G2, G3, G4, G5, G6, G7, G8, G9, G10, and G11, corresponding to group 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11, respectively.

Table 4. Means of the genotypic values for the plant height - Ht (cm), dry matter yield - DMY (kg ha^{-1}), disease score - DS (%), stem / leaf ratio - SLR (%), lignin - L (%), crude protein - CP (%), *in vitro* dry matter digestibility - IVDMD (%), neutral detergent fiber - NDF (%), and acid detergent fiber - ADF (%), of alfalfa accessions in different clusters.

Traits	Cluster										
	1	2	3	4	5	6	7	8	9	10	11
Ht	50.05	50.30	47.50	43.96	39.27	39.38	35.09	47.90	46.72	45.97	43.00
DMY	2030.54	1988.59	1844.82	1784.58	1649.15	1576.19	1485.44	1864.72	1825.89	1753.39	1711.93
DS	2.35	2.33	2.29	2.29	2.29	2.34	2.26	2.30	2.32	2.32	2.31
SLR	0.65	0.68	0.64	0.62	0.58	0.58	0.54	0.64	0.63	0.62	0.64
L	7.27	7.27	7.27	7.27	7.27	7.27	7.27	7.27	7.27	7.27	7.28
CP	26.85	26.60	27.31	27.83	29.07	28.91	29.23	27.26	27.68	27.46	27.96
IVDMD	69.61	69.54	69.87	70.09	70.05	70.61	70.72	69.82	69.89	69.87	69.98
NDF	42.95	42.93	42.92	42.91	42.85	42.78	42.76	42.93	42.94	42.89	42.93
ADF	30.86	30.88	30.68	30.57	30.63	30.29	29.73	30.66	30.56	30.55	30.46

Similar to the genotypic maps, the organization obtained for all 10 maps was the same in the molecular maps. The accessions were organized in four clusters, according to the 54 SSRs (Figure 4). Sixty-eight of the 77 accessions were in group G3 of the molecular diversity map. Some cultivars such as Monarca SP INTA, Mecca, Sequel, and Condor are parents of several accessions clustered in this group.

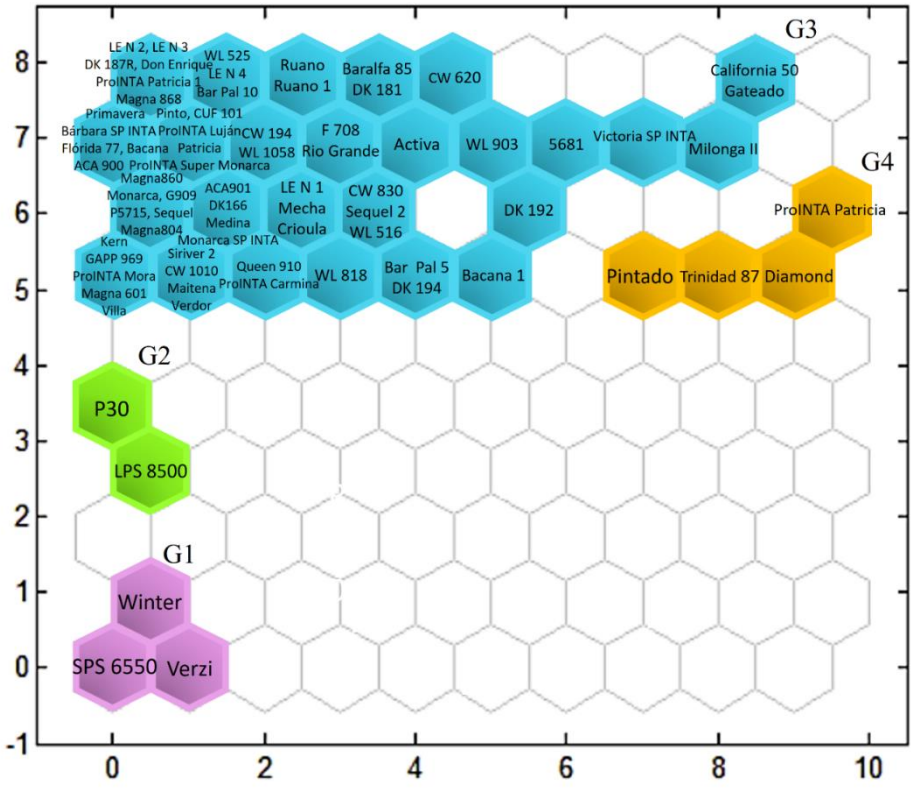


Figure 4. Self-organizing map for the SSR data. Accessions belonging to the same group were identified using equal colors on the map. Numbers on x- and y-axis represent the neurons of the 10 × 10 topology. G1, G2, G3, and G4 corresponding to group 1, 2, 3, and 4, respectively.

DISCUSSION

The results obtained by the LRT indicate the presence of genetic variability among the alfalfa accessions. The genotypes ×-cuttings interactions considering disease score, lignin, *in vitro* dry matter digestibility, and acid detergent fiber, were not significative (Figures S1). This result indicates that selection for these traits can be made at any time along alfalfa cuttings. The

significance of the genotype ×-cuttings interaction to traits such as dry matter yield is commonly observed in forage species (Rocha et al. 2016; Fernandes et al. 2017) and suggests that the selection needs to be performed more carefully, since that the interaction may influence the genotypic response.

The ability of alfalfa to better adapt to tropical conditions can be first evaluated by a higher productivity reached by the crop along the different cuts. The cultivar Crioula, which is currently the most cultivated in Brazil, has an annual production of approximately 14 to 22 t ha^{-1} (Ferreira and Vilela 2015). Considering the importance of productivity and the significance of the accessions by cuttings interaction, it can be affirmed that the most efficient method of alfalfa selection is through the joint evaluation of cuts made over time and not through isolated information alone.

However, in addition to knowledge regarding yield, characterizing the alfalfa germplasm requires the study of the accessions' nutritive values, which is one of the main objectives of forage breeding programs (Jones and Tracy 2017; Wilkins 2003). Rassini et al. (2007) suggest optimal nutritive values in alfalfa that agree with the values found in the evaluated germplasm (Table 4). Limited alfalfa yield in tropical regions is directly related to the lack of well-adapted cultivars. Problems related to salt and heat tolerance, persistence, insect, and nematode resistance are major challenges to be overcome in tropical climate regions. The first step to achieve this adaptation is to use the right breeding strategies. Therefore, the choice of promising germplasm to compose base populations is critical.

In molecular analyses for tetraploid alfalfa with codominant markers, genotypic information is commonly transformed into allele information of absence or presence (Flajoulot et al. 2005). Although SSR information is reduced, information on the species' allelic richness can be revealed (Azevedo et al. 2012). Even though there are few SSR markers, compared to SNPs, they can highly discriminate among alfalfa accessions. This fact was verified by

Annicchiarico et al. (2016), who compared the genetic diversity of alfalfa cultivars by means of SNPs, SSR, and morphophysiological data. The authors concluded that molecular characterization is less challenging than morphophysiological characterization and that this could complement the differentiation of alfalfa cultivars.

The organization of genetic diversity in the map of the BLUPs supported the results describing the correlations between traits. For example, accessions of higher dry matter yields (Figure S2) were found in higher concentration in groups whose crude protein averages were the lowest (Table 4). In addition, the map was able to separate the accessions with the highest yield in its lower portion, the least-yielded ones in the median portion, and those of mean yield in the upper portion. Although poorly explored in plant breeding, SOMs have proven to be quite effective in organizing genetic diversity. Santos et al. (2019) had a great result using SOMs to organize the genetic diversity of rice genotypes.

Due to the presence of a complex genomic structure, severe inbreeding depression, and cross-pollination (Nagl et al. 2011), the correct choice of parents for alfalfa poly-crossing is fundamental for the genetic progress of the crop. The organization of the genetic diversity of the accessions revealed important information in defining breeding strategies for alfalfa tropical production. To compose crosses that will form synthetic populations of high dry matter yield, for example, the choice of accessions of high yield belonging to different groups in the lower portion of the map of BLUPs could be prioritized.

The SOM based on SSR markers was also faithful to the information of the molecular analysis. The accessions that presented the highest concentrations of favorable alleles for persistence and dry matter yield were organized in the same group. The accession Prointa Patrícia, which was in G4 in the molecular map, was originated from CW4496 population in addition to individuals from unknown populations that present high persistence (Table S1). The parent's selection to develop this accession resulted from recurrent selection for resistance to

Phytophothora megasperma (INTA 2019). The Trinidad 87 accession, located in the same molecular group of Prointa Patricia, also belonged to the same group in the map of the BLUPs. This accession is probably also the result of a selection for persistence, a fact that demonstrates the important linkage between phenotypic and molecular information to decision making in plant breeding programs.

The map based on molecular information reduced the number of groups compared to the genetic map, but this does not invalidate the results. Once the genetic constitution of the accessions is known, it is easy to understand the causes that led to this reduced quantity. Many accessions have a close genetic background, as they can share the same parents. The accession Mecca, for example, participated in the crosses that gave rise to ACA 900, Activa, DK 194, Gapp 969, Prointa Carmina, and Prointa Super Monarca. ACA 900 was a part of the crosses that originated DK 194. Condor is a parent of Activa, DK 166, and Prointa Mora, and so on (Table S1).

Clusters obtained using SOM do not have the same interpretation as clusters from conventional techniques, such as Tocher clustering, for example. Certainly, using the data in a Tocher clustering would lead to more groups. However, the relationships among accessions would be underestimated since only linear relationships would be observed. Using SOM enables more than a simple subdivision into groups but allows understanding of the proximity and organization of a set of genetic materials. In the SOM of molecular information, cluster reduction does not mean low genetic diversity, but that many accessions share the same information. It is perfectly acceptable, for example, that accessions located at the borders in a large cluster to be used for the formation of one same alfalfa base population.

In addition, discrepancies between molecular and genetic diversity have been reported by other authors for sorghum (Silva et al. 2017), wheat (Soriano et al. 2016), or white clover (Annicchiarico and Carelli 2014). Although Herrmann et al. (2018) found similarities between

the molecular and morphological diversity of alfalfa cultivars, large discrepancies were also observed. While the morphological traits were able to clearly separate cultivar groups belonging to different geographical regions, the molecular markers did not return a clear division structure. Herrmann et al. (2018) suggests that low correlation values can be explained by the fact that the markers extract a neutral or non-adaptive diversity and that the phenotypic diversity represents a selected natural or artificial diversity.

Studying genetic diversity is critical in breeding programs. By identifying more similar clusters of accessions, it is possible to guide the use of different parents and to plan crossings to generate promising synthetic populations. According to Cruz et al. (2011), the importance of genetic diversity studies for plant breeding comes down to the fact that crosses involving genetically different parents are better able to maintain genetic variability in advanced generations. This variability is especially important for alfalfa that has a high sensitivity to inbreeding (Kopp 2011).

Based on the analysis of phenotypic and molecular data, and the accession's performance over the years (Table S2), a great base population would be composed of ProINTA Patricia, Pro INTA SuperMonarca, Mecha, Magna 601, WL 525, ACA 900, Bacana, CUF 101, Crioula, and Ruano. These populations have at least 50 distinct parents, presented high DMY, besides favorable alleles for regrowth ability, biomass yield, and persistence.

Alfalfa has great potential to occupy new Brazilian areas. However, since the available accessions have origins in a temperate climate, adaptation and cultivar development in a tropical climate requires controlled breeding actions. Annicchiarico et al. (2016) argue that independent information generated from molecular and phenotypic data may enhance the distinguishability of alfalfa cultivars. For this distinction to be truly effective, the pooling of molecular and phenotypic information is of fundamental relevance at the beginning of the breeding program. The genetic variability observed in this study for the great majority of the traits indicates the

high potential for development of alfalfa cultivars of high yield and high nutritive value. Using the approaches introduced here, breeders will be able to select accessions based on molecular and genotypic maps for the formation of promising base populations adapted to tropical conditions. Once adapted alfalfa synthetic populations are selected, these may be integrated into breeding programs in other tropical regions of the world.

REFERENCES

Annicchiarico P, Wei Y, Brummer EC (2017) Genetic structure of putative heterotic populations of alfalfa. Plant Breeding 136:671-678. https://doi.org/10.1111/pbr.12511

Annicchiarico P, Nazzicari N, Ananta A, Carelli M, Wei Y, Brummer EC (2016) Assessment of cultivar distinctness in alfalfa: A comparison of genotyping-by-sequencing, simple-sequence repeat marker, and morphophysiological observations. The Plant Genome 9:1-12. https://doi.org/10.3835/plantgenome2015.10.0105

Annicchiarico P, Barrett B, Brummer EC, Julier B, Marshall AH (2015) Achievements and challenges in improving temperate perennial forage legumes. Critical Reviews in Plant Science 34:327-380. https://doi.org/10.1080/07352689.2014.898462

Annicchiarico P, Carelli M (2014) Origin of ladino white clover as inferred from patterns of molecular and morphophysiological diversity. Crop Sci. 54:2696–2706. https://doi.org/10.2135/cropsci2014.04.0308

Azevedo ALS, Costa PP, Machado JC, Machado MA, Pereira AV, Lédo FJS (2012) Cross species amplifcation of *Pennisetum glaucum* microsatellite markers in Pennisetum purpureum and genetic diversity of napier grass accessions. Crop Sci. 52:1776-1785. https://doi.org/10.2135/cropsci2011.09.0480

Basigalup DH (2016) Producción de alfalfa en Argentina. In: Jornada Nacional de Forrajes Conservados, Buenos Aires, Argentina. 83-85p.

Biazzi E, Nazzicari N, Pecetti L, Brummer EC, Palmonari A, Tava A, Annicchiarico P (2017) Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. PLoS One 12:e0169234. https://doi.org/10.1371/journal.pone.0169234

Comeron EA, Ferreira RP, Vilela D, Kuwahara FA, Tupy O (2015) Utilização da alfafa em pastejos para alimentação de vacas leiteiras. In: Ferreira RP, Vilela D, Comeron EA, Bernardi ACC, Karam D (eds) Cultivo e utilização da alfafa em pastejo para alimentação de vacas leiteiras. Embrapa, Brasilia, pp 13-16

Creste S, Tulmann Neto A, Figueira A (2001) Detection of single sequence repeat polymorphisms in denaturing polyacrilamide sequencing gels by silver staining. Plant Molecular Biology Reporter 19:299-306. https://doi.org/10.1007/BF02772828

Cruz CD, Ferreira FM, Pessoni LA (2011) Biometria aplicada ao estudo da diversidade genética. Suprema, Visconde do Rio Branco.

Cruz CD (2013) GENES – a software package for analysis in experimental statistics and quantitative genetics. Acta Sci Agron 5:271–276.

Cruz CD (2016) Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Sci Agron 38:547-552. https://doi.org/10.4025/actasciagron.v38i3.32629

Diwan N, Bhagwat AA, Bauchan GB, Cregan PB (1997) Simple sequence repeat DNA markers in alfalfa and perennial and annual Medicago species. Genome 40:887-895. https://doi.org/10.1139/g97-115

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical bulletin 19:11–15.

Eujayl I, Sledge MK, Wang L, May GD, Chekhovskiy K, Zwonitzer JC, Mian MA (2003) *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for Medicago spp. Theor Appl Genet 108:414–422. https://doi.org/10.1007/s00122-003-1450-6

Fernandes FD, Braga GJ, Ramos AKB, Jank L, Carvalho MA, Maciel GA, Karia CT, Fonseca CEL (2017) Repeatability, number of harvests, and phenotypic stability of dry matter yield and quality traits of *Panicum maximum* jacq. Acta Sci Agron 39:149-155. https://doi.org/10.4025/actascianimsci.v39i2.32915

Ferreira RP, Vilela D (2015) Potencial de utilização da alfafa. In: Ferreira RP, Vilela D, Comeron EA, Bernardi ACC, Karam D (eds) Cultivo e utilização da alfafa em pastejo para alimentação de vacas leiteiras. Embrapa, Brasilia, pp 13-16

Ferreira RP, Basigalup DH, Vasconcelos ES, Cruz CD, Pereira AV (2008) Genética quantitativa e métodos de melhoramento em alfafa. In: Ferreira RP, Rassini JB, Rodrigues AA, Freitas AR, Camargo AC, Mendonça FC (eds) Cultivo e utilização da alfafa nos trópicos. Embrapa, Brasilia, pp 171-205

Flajoulot S, Ronfort J, Baudoin P, Barre P, Huguet T, Huyghe C, Julier B (2005) Genetic diversity among alfalfa (*Medicago sativa*) cultivars coming from a breeding program, using SSR markers. Theor Appl Genet 111:1420–1429. https://doi.org/10.1007/s00122-005-0074-4

Goering HK, Van Soest PJ (1970) Forage fiber analysis (Apparatus, Reagents, Procedures and Some Applications). Agriculture Handbook 379. US Government Printing Office.

Grandon NG, Alarcón Y, Moreno MV, Arolfo V, Orodizzi A, Basigalup DH, Gieco JO, Bruno C (2013) Genetic diversity among alfalfa genotypes (*Medicago sativa* L.) of non-dormant cultivars using SSR markers and agronomic traits. Revista de la Facultad de Ciencias Agrarias 45:181-195.

Herrmann D, Flajoulot S, Barre P, Huyghe C, Ronfort J, Julier B (2018) Comparison of morphological traits and molecular markers to analyse diversity and structure of alfalfa (*Medicago sativa* L.) cultivars. Genet Resour Crop Ev 65: 527-540. https://doi.org/10.1007/s10722-017-0551-z

Instituto Nacional de Tecnología Agropecuaria (2001). Prointa Patricia. https://inta.gob.ar/variedades/prointa-patricia. Accessed 20 December 2019

Jones GB, Tracy BF (2017) Persistence and productivity of orchardgrass and orchardgrass/alfalfa mixtures as affected by cutting height. Grass Forage Sci 73:544-552. https://doi.org/10.1111/gfs.12309

Julier B, Barre PG, Lambroni P, Delaunay S, Thomasset M, Lafaillette F, Gensollen V (2018) Use of GBS markers to distinguish among lucerne varieties, with comparison to morphological traits. Molecular Breeding 38:133. https://doi.org/10.1007/s11032-018-0891-1

Julier B, Flajoulot S, Barre PG, Cardinet G, Santoni S, Huguet T, Huyghe C (2003) Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago*

sativa) using microsatellite and AFLP markers. BMC Plant Biology 3:9. https://doi.org/10.1186/1471-2229-3-9

Kohonen T (2014) MATLAB implementations and applications of the self-organizing map. Unigrafia Ou, Helsinki

Kopp MM (2011) Origem, evolução e domesticação da alfafa. In: Ferreira RP, Basigalup DH, Gieco JO (eds) Melhoramento Genético da Alfafa, Embrapa, São Carlos, pp 225-260

Li X, Wang X, Wei Y, Brummer EC (2011) Prevalence of segregation distortion in diploid alfalfa and its implications for genetics and breeding applications. Theor Appl Genet 123:667–679. https://doi.org/10.1007/s00122-011-1617-5

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Research 27:209–220

Ministério da Agricultura e Pecuária - MAPA (2019) Formulários para registro de cultivares. http://www.agricultura.gov.br/assuntos/insumos-agropecuarios/insumos-agricolas/sementes-e-mudas/registro-nacional-de-cultivares-2013-rnc-1/formularios-para-registro-de-cultivares. Accessed 20 December 2019

Matlab (2010) Matlab Version 7.10.0. Natick, MA: The Math Works Inc.

Nagl N, Taski-Ajdukovic K, Barac G, Baburski A, Seccareccia I, Milic D, Katic S (2011) Estimation of the genetic diversity in tetraploid alfalfa populations based on RAPD markers for breeding purposes. International Journal of Molecular Sciences 12:5449-5460. https://doi.org/10.3390/ijms12085449

Narasimhamoorthy B, Bouton JH, Olsen KM, Sledge MK (2007) Quantitative trait loci and candidate gene mapping of aluminum tolerance in diploid alfalfa. Theor Appl Genet 114:901–913. https://doi.org/10.1007/s00122-006-0488-7

Oliveira MS, Santos IG, Cruz CD (2020) Self-organizing maps: a powerful tool for capturing genetic diversity patterns of populations. Euphytica 216:49. https://doi.org/10.1007/s10681-020-2569-0

Qiang H, Chen Z, Zhang Z, Wang X, Gao H, Wang Z (2015) Molecular diversity and population structure of a worldwide collection of cultivated tetraploid alfalfa (Medicago

sativa subsp. sativa L.) germplasm as revealed by microsatellite markers. Plos One 10:e0124592. https://doi.org/10.1371/journal.pone.0124592

R Core Team (2017) The R project for statistical computing- R version 3.4.3. http://www.r-project.org

Rassini JB (2002) Manejo da água na irrigação da alfafa num latossolo vermelho-amarelo. Pesqui Agropecu Bras 37:503-507

Rassini JB, Ferreira RP, Rodrigues AA, Barioni Junior W, Moreira A, Machado R (2007) Qualidade da forragem de alfafa na região Sudeste do Brasil. Embrapa. https://ainfo.cnptia.embrapa.br/digital/bitstream/CPPSE/17079/1/PROCIJBR2007.00088.pdf. Accessed 20 December 2019

Resende MDV (2016) Software Selegen-REML/BLUP: a useful tool for plant breeding. Crop Breed Appl Biotech 16:330-339. https://doi.org/10.1590/1984-70332016v16n4a49

Robins JG, Luth D, Campbell A, Bauchan GR, He C, Viands DR, Hansen JL, Brummer C (2007a) Genetic mapping of biomass production in tetraploid alfalfa. Crop Sci 47:1–10. https://doi.org/10.2135/cropsci2005.11.0401

Robins JG, Bauchan GR, Brummer EC (2007b) Genetic mapping forage yield, plant height, and regrowth at multiple harvests in tetraploid alfalfa (Medicago sativa L.). Crop Sci 47:11–18. https://doi.org/10.2135/cropsci2006.07.0447

Robins JG, Hansen JL, Viands DR, Brummer EC (2008) Genetic mapping of persistence in tetraploid alfalfa. Crop Sci 48:1780–1786. https://doi.org/10.2135/cropsci2008.02.0101

Rocha JRASC, Machado JC, Carneiro PCS, Carneiro JC, Resende MDV, Lédo FJS, Carneiro JES (2016) Bioenergetic potential and genetic diversity of elephantgrass via morpho-agronomic and biomass quality traits. Industrial Crops and Products 95:485-492. https://doi.org/10.1016/j.indcrop.2016.10.060

Santos IG, Carneiro VQ, Silva Junior AC, Cruz CD, Soares PC (2019) Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. Acta Sci Agron 41:e39803. https://doi.org/10.4025/actasciagron.v41i1.39803

Santos IG, Cruz CD, Nascimento M, Rosado RDS, Ferreira RP (2018) Direct, indirect and simultaneous selection as strategies for alfalfa breeding on forage yield and nutritive value. Pesqui Agropecu Trop 48:178-189. https://doi.org/10.1590/1983-40632018v4851950

Silva MJ, Pastina MM, Souza VF, Schaffert RE, Carneiro PCS, Noda RW, Carneiro JES, Damasceno CMB, Parrella RAC (2017) Phenotypic and molecular characterization of sweet sorghum accessions for bioenergy production. PLoS One 12:e0183504. https://doi.org/10.1371/journal.pone.0183504

Singh D (1981) The relative importance of characters affecting genetic divergence. Indian Journal of Genetics and Plant Breeding 41:237–245

Sledge MK, Ray IM, Jiang G (2005) An expressed sequence tag SSR map of tetraploid alfalfa (*Medicago sativa* L.). Theor Appl Genet 111:980–992. https://doi.org/10.1007/s00122-005-0038-8

Soriano JM, Villegas D, Aranzana MJ, Del Moral LFG, Royo C (2016) Genetic structure of modern durum wheat cultivars and mediterranean landraces matches with their agronomic performance. PLoS One 11:e0160983. https://doi.org/10.1371/journal.pone.0160983

Van Soest PJ (1963) Use of detergents in the analysis of fibrous feed. II. A rapid method of the determination of fiber and lignin. Journal of the Association of Official Agricultural Chemists 46:829-35

Vasconcelos ES, Ferreira RP, Cruz CD, Moreira A, Rassini JB, Freitas AR (2010) Estimativas de ganho genético por diferentes critérios de seleção em genótipos de alfafa. Revista Ceres 57:205-210

Wilkins PW, Humphreys MO (2003) Progress in breeding perennial forage grasses for temperate agriculture. Journal of Agricultural Science 140:129-150. https://doi.org/10.1017/S0021859603003058

SUPPLEMENTARY INFORMATION

Table S1. Description of the genetic materials that originated some evaluated accessions. As more materials participate in the crosses as broader is the accession genetic base. Accessions that have more than 10 parents are considered of broad genetic base. The parents that compose the crosses were chosen based on yield and resistance to pests and diseases. The parents had evaluated in alfalfa breeding programs in Argentina and in the United States for more than 20 years.

Accession	Genetic Origin	Adaptation
1 5681	Ancher (34%) Meteor (16%) Cuf01 (6%) 555 (6%) Mercury (3%), 524. 526. 5331. 5432. Apollo, NCMP10	Central Argentina
2 ACA 900	CW2820 (16%) DK189 (14%) CW2818, CW2817, Mecca, Express UC332, UC176, UC196, UC226, UC276, UC296, UC222, UC263, UC189, UC231, UC195	Northern Argentina, Santa Fé, and Córdoba
3 ACA 901	UC332, SW9212, SW8210	Argentina
4 Activa	VS448 (17%), CW2820 (14%), Mecca (12%), DK189 (10%), Condor (8%), UC332 (6%), CW2815 (5%), CW2817 (5%), VS446 (3%), Armona (3%), Sundor (3%), Express (2%), DK187 (2%), UC176 (2%), UC196 (2%), UC222 (2%), UC236 (2%), UC276 (2%)	Argentina
5 Bacana	Beacon (305), DK189 (30%), Coronado (20%), Zaino (20%)	Argentina
6 Bacana 1		Argentina
7 Bar Pal 5		Argentina
8 Bar Pal 10		Argentina
9 Baralfa 85	CA G, WI457, WI514, Cuf101	Argentina
10 Barbara SP INTA	Monarca SP INTA (25%), WL516 (25%), Armona (25%), C/W331 (25%)	Pampas Argentinos
11 California 50		Southern Califórnia, Southest Arizona
12 CUF 101	UC Cargo (55%), UC Salton (1%), UC76 (22%), 1972 Breeding Mixture (20%), Niagara N71 (2%)	Southern Califórnia, Southest Arizona
13 CW 1010		Argentina
14 CW 194		Argentina
15 CW 620	SPS6550 (11%), DK166 (8%), Express (7%), N650 (6%), Archer (2%), AlfaStar (2%), Bighorn (1%), Cal/West Seeds (63%)	Argentina
16 CW 830	Alfa200 (9%), Monarca SPI (9%), DK189, Weston (3%), WL525HQ (1%), Cal/West Seeds (69%)	Argentina
17 Diamond		Argentina
18 DK 166	Express, Condor, Valley+, VS626, Shenandoah, VS481, Mede	Sacramento, California, Idaho, Novo México, and Buenos Aires
19 DK 181		Argentina
20 DK 187 R		Argentina
21 DK 192		Argentina
22 DK 194	Topacio, DK192. Grasis, ACA900, Super Supreme, Mecca, F969, DK191, CalWest Seeds	Argentina, Southest USA
23 Don Enrique	Aurora (20%), 5683 (18%), Vitoria SP INTA (30%), Primavera (20%)	Argentina
24 F 708		Argentina
25 Florida 77		Argentina, Southest USA

26	G 909	82-296 F (35%), C346 (32%), C442 (2%), Alazan (14%)	Argentina
27	GAPP 969	Mecca (10%), DK189 (10%), UC332 (10%), Cal/West Seeds (70%)	Argentina, Southest USA
28	Gateado	GH97-53/55/56/58/60/65 (52%), FG1150/1151/1159/1188 (24%), 5683 (10%), Yolo (7%), ARG96-1 (4%), Tahoe (3%)	Argentina
29	Kern		Argentina
30	LE N 1		Argentina
31	LE N 2		Argentina
32	LE N 3		Argentina
33	LE N 4		Argentina
34	LPS 8500		Argentina
35	Magna 601	Sutter varieties	Argentina
36	Magna 804		Argentina
37	Magna 860		Argentina
38	Magna 868		Argentina
39	Maitena		Argentina
40	Mecha	FG94-55 (18%), FG94-48 (17%), FG94-49 (15%), FG94-46 (13%), FG94-13 (2%), FG94-12 (1%), FG94-15 (1%), DK193 (20%), DK180ML (13%)	Argentina
41	Medina		Argentina
42	Milonga II	FG 988 (30%), FG1164 (6%), Coronado (31%), DK193 (33%)	Argentina
43	Monarca		Argentina
44	Monarca SP INTA		Argentina
45	P 30		Argentina
46	P 5715		Argentina
47	Patriarca		Argentina
48	Patricia		Argentina
49	Pintado		Argentina
50	Pinto		Argentina
51	Primavera		Argentina
52	ProINTA Carmina	Monarca SP INTA, 5929, Mecca, Sequel	Argentina
53	ProINTA Luján		Argentina
54	ProINTA Mora	Monarca (30%), Rocío ISP (17%), DK189 (13%), AL102 (9%), Sima15 (7%), 5929 (5%), Condor (4%), P30, Sundor, 581, DK181, WL605, AL, Sima1/3/6/16/29	Argentina
55	ProINTA Patricia	CW4496 e materiais de alta persistência	Argentina
56	ProINTA Patricia 1		Argentina
57	ProINTA Super Monarca	Yolo (20%), Maxidor (18%), Monarca SP INTA (46%), Mecca (16%)	Pampas Argentinos
58	Queen 910	Araucana (10%), Super Lenchera (10%), Máxima (10%), Falcon (15%), DK192 (15%), Trinidad87 (20%), Monarca SP INTA (20%)	Argentina
59	Rio Grande		Argentina
60	Ruano		Argentina
61	Ruano 1		Argentina
62	Sequel		Argentina
63	Sequel 2		Argentina
64	Siriver 2		Argentina
65	SPS 6550		Argentina
66	Trinidad 87		Argentina
67	Verdor		Argentina
68	Verzi		Argentina

69	Victoria SP INTA	WL 313 (20%), 77-T-25 (40%), 77-78 CaB (40%)	Argentina
70	Villa	Máxima (25%), Monarca SP INTA (25%), Super Lechera (25%), Armona (25%)	Argentina
71	Winter		Argentina
72	WL 1058	Rosillo (53%), FG10A215 (23%), FG8M809 (16%), FGA016TF (4%), FG99x49/50 (3%), DK193 (1%)	Pampas Argentinos
73	WL 516		Santa Fé, Córdoba, Buenos Aires
74	WL 525	WL516, 86-222-CA, 898, Maxidor	Santa Fé, Córdoba, Buenos Aires
75	WL 818	Pintado (54%) WL 611 (46%)	Argentina
76	WL 903		Argentina
77	Crioula		Argentina, Brazil

Table S2. Top 10 alfalfa accessions per year. The ranking considered one cut in 2015, four in 2016, and three in 2017.

Years		
2015	2016	2017
CUF101	Ruano	Siriver 2
Verdor	Siriver 2	F 708
Ruano	DK 192	Monarca
Siriver 2	Magna 860	Ruano
Bal Par 10	ProINTA Super Monarca	Florida 77
DK 192	Mecha	Verdor
LEN N 3	CW 830	LE N 4
ProINTA Patricia	CW 194	Mecha
F 708	ProINTA Patricia	DK 194
Mecha	Maitena	Villa

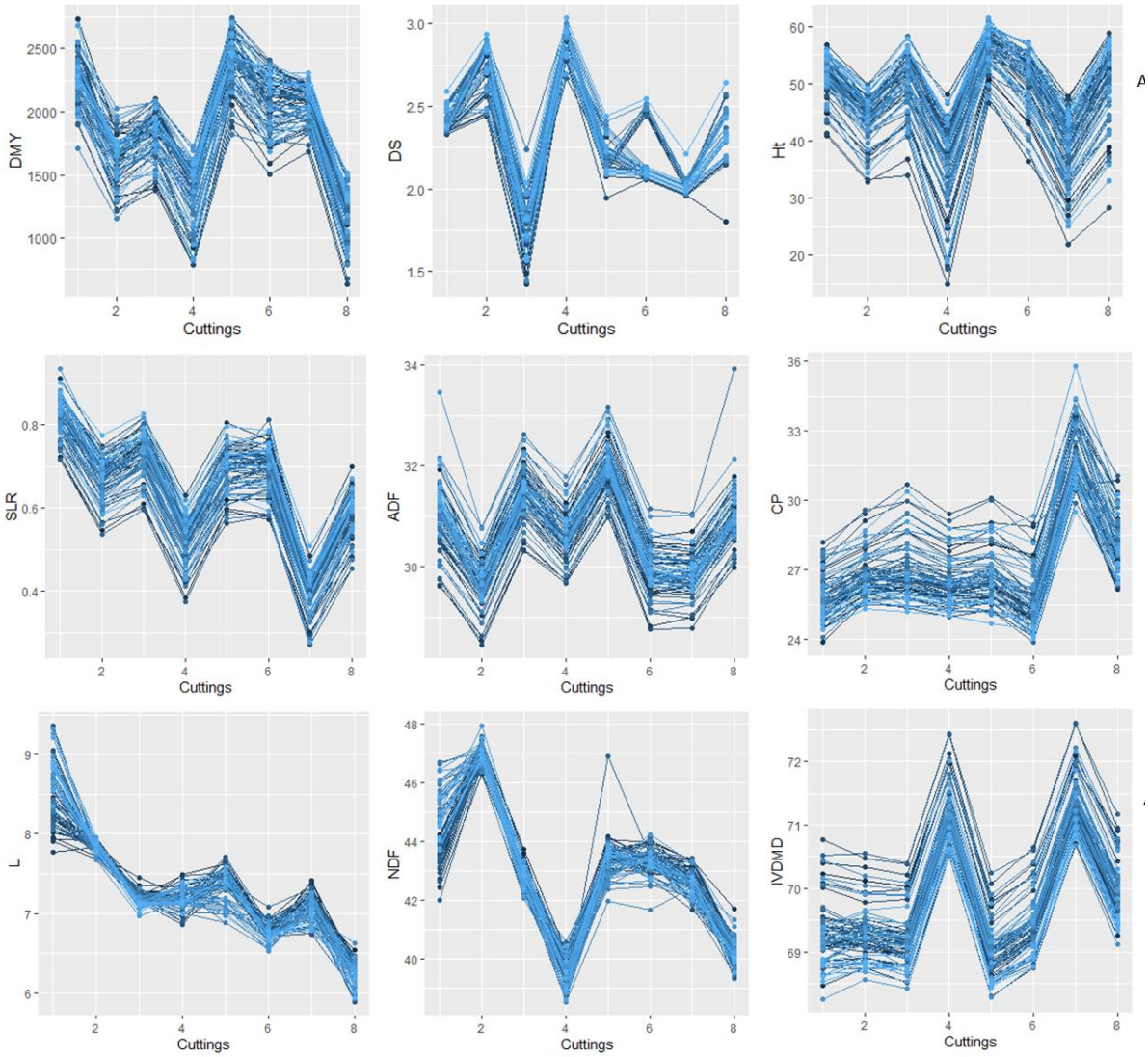


Figure S1. Genotypic values of 77 alfalfa accession through eight different cuttings.

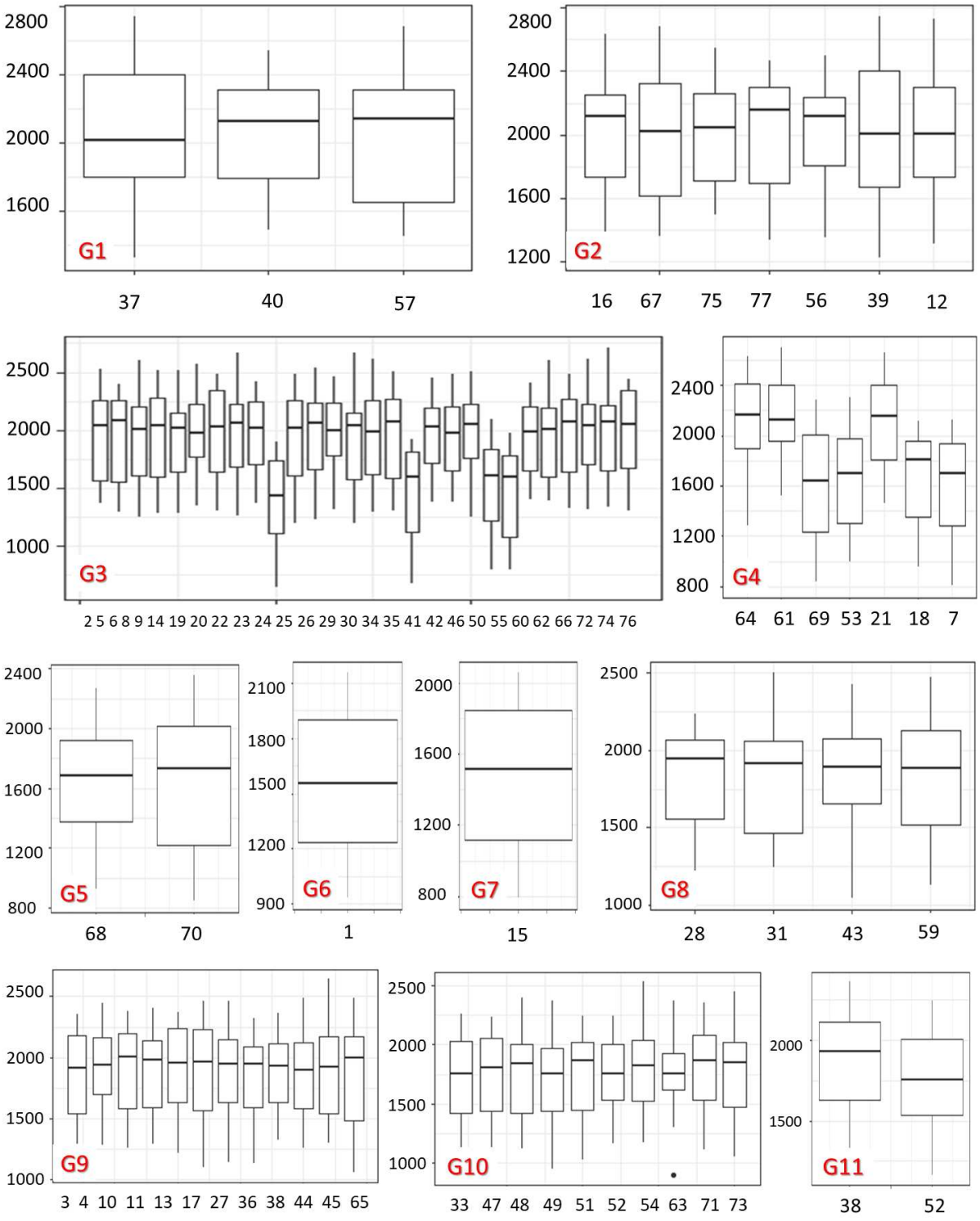


Figure S2. Clustering of 77 alfalfa accessions for dry matter yield - DMY (kg ha^{-1}) according to the groups obtained in the self-organizing map. Numbers on x- and y-axis represent the genotypes described in Table S1 and DMY, respectively. G1, G2, …, G11 represent different clusters.

CHAPTER 2

ACCESSING THE DRY MATTER YIELD PERSISTENCE IN ALFALFA GERMPLASM WITHIN BRAZIL THROUGH RANDOM REGRESSION MODELS

VIÇOSA – MINAS GERAIS

2021

RESUMO

SANTOS, Iara Gonçalves dos, D.Sc., Universidade Federal de Viçosa, julho de 2021. **Acessando a persistência da produção de matéria seca de alfafa cultivada no Brasil por meio de modelos de regressão aleatória.** Orientador: Cosme Damião Cruz.

Persistência é fundamental para o cultivo de alfafa em regiões de clima tropical, entretanto o melhoramento para essa característica constitui um desafio para os programas de melhoramento. Os objetivos desse estudo foram acessar a persistência da produção de matéria seca de alfafa avaliada em condições tropicais, e propor um método baseado em regressão aleatória utilizando redes neurais artificiais (RNA) para seleção de acessos de alfafa persistentes. A produção de matéria seca (PMS) de 77 acessos de alfafa avaliados em 24 cortes foi mensurada para avaliar a persistência utilizando diferentes modelos de regressão aleatória. Foi proposto um método para selecionar materiais genéticos persistentes baseado na trajetória genética dos acessos ao longo dos cortes (curvas de persistência). As curvas ajustadas revelaram grande amplitude em relação à PMS ao longo do tempo, o que sugere alta variabilidade para persistência. O método proposto para acessar persistência é baseado em três etapas, incluindo a (i) obtenção da trajetória genética dos acessos, (ii) o agrupamento dos acessos para definição de grupos de persistência utilizando o método k-médias e (iii) o ajuste de RNA para mimetizar a classificação obtida pelo método k-médias de forma automatizada. Uma vez que novos acessos forem avaliados em programas de melhoramento de alfafa, eles poderão ser classificados de acordo com seus valores genéticos utilizando a mesma rede ajustada nesse trabalho. A RNA será a responsável pela tomada de decisão quanto à persistência da produção de matéria seca de alfafa.

Palavras-chave: *Medicago sativa* subs. *Sativa*. Trajetória genética. Rede neural artificial.

ABSTRACT

SANTOS, Iara Gonçalves dos, D.Sc., Universidade Federal de Viçosa, July, 2021. **Accessing the dry matter yield persistence in alfalfa germplasm within Brazil through random regression models.** Adviser: Cosme Damião Cruz.

Persistence plays a key role in alfalfa cultivation in tropical areas, but it is still a bottleneck for breeding programs. The objectives of the study were (i) to access the dry matter yield persistence of alfalfa accessions evaluated under tropical conditions, and (ii) to propose a method for selecting persistent accessions based on Random Regression (RR) models using an artificial neural network (ANN). The dry matter yield of 77 alfalfa accessions from 24 cuttings was measured to evaluate persistence using different RR models. A persistency method was proposed based on the trajectory curves of the accessions. The fitted curves showed a great amplitude regarding DMY overtime, which suggests high variability in persistence. The three-step method for accessing persistence presented in this study included a RR model to obtain trends of persistence, a k-means method to define different persistency clusters, and then an ANN to perform persistence classification in an automated way. When new accessions are evaluated by the alfalfa breeding program, they will be classified according to their genetic value scores using the same ANN previously fitted. The ANN will be responsible for the decision-making process.

Keywords: *Medicago sativa*. Genetic trajectory. Artificial neural network.

INTRODUCTION

Cultivated alfalfa (*Medicago sativa* ssp. *sativa*) is the most important perennial legume forage in the world, with about 32 million hectares cultivated mainly in temperate regions. The forage is well-known for its yield potential and ability to produce more protein per hectare than grains or oilseed crops (Bouton, 2012). The multi-purpose of alfalfa extends to hay production, grazing, silage, green manure, culinary, and medicinal industry. Using alfalfa in specialized dairy herds reduces production costs (Comeron et al., 2015), not to mention the recently developed concept of pasture-based livestock industries, which is changing the way livestock is produced all over the world. The Chinese government, for example, has applied subsidy policies for grassland ecological protection and so the revitalization of alfalfa for dairying has become more profitable (Shi & Smith, 2017).

Most of the alfalfa production in Brazil is located in the South Region, especially in the states of Paraná and Rio Grande do Sul (Santos et al., 2018). Though alfalfa cultivars are still limited in Brazil, Brazilian farmers have a great interest in inserting the crop into their dairy systems in regions with different soil and climatic conditions. There is an urgent need for cultivars well-adapted to tropical environments and for consistent information on how to grow the crop under those conditions. In order to develop cultivars with an expressive yield in tropical environments, it is important to understand the major factors affecting alfalfa performance in tropical areas.

The dynamics related to regrowth seems to play a role in alfalfa yield improvement (Lamb et al., 2008), especially in the tropics where environmental conditions are highly unstable. More than high yield, elite cultivars need to maintain their yield over different hervests. Persistence is a complex and critical component in perennial forage crops and is defined as the survival of a stand over time. Under a genetic stand point, persistence is the maintenance of the genetic values obtained by each accession over time, which means that the

genetic trajectory of an accession determines its persistence. Rather than observing individual yield increments (in each harvest), the trajectory determines the persistence and gives the breeder the right insights about the trait. Factors affecting persistence include biotic and abiotic resistance and the way the forage is managed. Animal grazing, for instance, imparts a unique form of shredding-type defoliation that has a lot more impact on alfalfa's persistence than harvesting. Perennial crops usually have their yields decreased over time, leading to a loss of their nutritive value. In persistent materials, this process is slower, which holds the forage quality for longer. Accessing the long-term persistence of alfalfa requires selection for persistence per se with an emphasis on long-term nurseries and family selection approaches (Robins et al., 2008).

Statistically, persistence data is classified as longitudinal. Longitudinal data analysis has to consider the continuous scale of the data over time because repeated measures of the same individual are inherently dependent. The adoption of a function to describe the structure of variances and covariances between cuts has become a popular approach for the genetic study of longitudinal data (Azevedo, 2012; Fitzmaurice, 2011). Random regression (RR) models, which can capture changes in traits over time without making assumptions about variances and errors, have been widely used in daily milk and egg production (Kranis et al., 2006). RR models can fit genetic values based on the decomposition of the variance into a genetic and a permanent environmental effect that is not assumed to be constant during the whole harvesting period (Sun et al., 2017).

Artificial neural networks (ANN) can be used as an auxiliary tool to automate the process of identifying persistent accession in the alfalfa breeding program. ANN are learning techniques based on the biological model of the brain. Since ANN can capture patterns other than linear, they can achieve high efficiency dealing with complex traits (Santos et al. 2018). ANNs capture relationships not perceived by stochastic models and have the advantage of

generalize information. Once a persistence criterion is set, ANNs can learn with the criterion and to perform the classification of new accessions based on their genotypic values (that come from an adequate statistic model).

In this study, we evaluated 77 alfalfa accessions with a temperate genetic background from which 24 harvests were collected to evaluate persistence. The parents of the populations that compose this germplasm were selected due to their persistence in temperate environments as well as their resistance to the main alfalfa pests and diseases. The objectives of the present study were (i) to access the dry matter yield persistence of alfalfa accessions evaluated under tropical conditions, and (ii) to propose a method for selecting persistent accessions based on RR models using artificial intelligence.

MATERIALS AND METHODS

Plant material. Seventy-seven alfalfa accessions that have a temperate background were evaluated at 24 different harvesting times from 2015 to 2017 [11/12/2015, 12/08/2015, 01/04/2016, 02/03/2016, 03/08/2016, 04/04/2016, 05/09/2016, 06/07/2016, 07/12/2016, 08/12/2016, 09/13/2016, 10/31/2016, 11/28/2016, 12/21/2016, 01/17/2017, 02/14/2017, 03/13/2017, 04/18/2017, 05/26/2017, 06/21/2017, 08/28/2017, 09/26/2017, 10/26/2017, 11/28/2017]. The experiment was carried out at Embrapa Southeast Livestock's experimental field in the municipality of São Carlos, São Paulo, Brazil [21° 57'42 "S, 47° 50'28" W, 860 m]. The experiment was performed in randomized complete blocks with three replicates. The plots consisted of four rows of four meters in length spaced 0.20 m apart, and the useful area corresponded to the two central rows, eliminating 0.50 m at the ends of the lines. Additional details about the germplasm origin, experimental design, and management practices can be found at (https://doi.org/10.1007/s10681-020-02606-w). Meteorological information for the period of evaluation is shown in Figure S1. In order to access persistence, dry matter yield

(DMY, in kg ha^{-1}) was periodically determined when each accession reached 10% of flowering in the 24 harvests. The alfalfa samples were dried at 65°C for 72 h until reaching a constant weight.

Statistical analysis. For fitting RR models using Legendre polynomials, the harvests were scaled to range -1 to +1 according to Schaeffer (2016), as follows

$$t_x = -1 + 2\left[\frac{(h_x - h_{min})}{(h_{max} - h_{min})}\right],$$

where h_x refers to the harvest *x*; h_{min} is the time of the first harvest; and, h_{max} is the time of the last harvest. DMY data across the 24 alfalfa harvests was used to fit different RR models through Legendre polynomials, according to the following model:

$$y_{ijk} = F_k + \beta_m \phi_m(a_{ij}^*) + \sum_{m=2}^{k_g-1} g_{im}\phi_m(a_{ij}^*) + \sum_{m=2}^{k_p-1} p_{ikm}\phi_m(a_{ij}^*) + e_{ijk}$$

where y_{ijk} is the measure on the i^{th} accession (i = 1, 2, …, 77), on the j^{th} harvest (j = 1, 2, …, 24) on the k^{th} replication. F_k is the fixed effect of replication (k = 1, 2, 3). βm is the fixed regression coefficient to model the overall trajectory of DMY overtime. g_{im} and p_{ikm} are the random regression coefficients for the genetic and permanent environmental effect on the ik^{th} plot (ik = 1, 2, …, 231). k_β ($k_\beta = 4$), k_g($k_g = 3$ or 4) and k_p($k_p = 3$ or 4) represent the order of the covariance functions to describe the fixed, genetic, and permanent environment effects, respectively. a_{ij}^* is the j^{th} harvest on the i^{th} accession [standardized from -1 to 1]. $\phi_m(a_{ij}^*)$ are the Legendre polynomials for a_{ij}^* regarding the fixed regression and the random effects of genetic and permanent environment, considering k_β, k_g, and k_p. e_{ijk} is the residual random effect. The matricial model is shown as follows

$$y = Xb + Zg + Wp + e$$

where y is the vector of phenotypic data, b is the vector of the fixed effects (replication) added to the overall mean, g is the random vector of genetic effects, p is the random vector of the

permanent environment, and e is the random vector of residuals. X, Z, and W are incidence matrices for fixed, genetic, and permanent environment effects, respectively, considering each harvest in a Legendre scale. It was assumed $g \sim N(0, K_g \otimes I)$, $p \sim N(0, K_p \otimes I)$, and $e \sim N(0, R \otimes I)$ for genetic values, permanent environment, and residuals, respectively. I represents the identity matrix. $\otimes$ denotes the direct product, Kg and Kp are matrices of covariance between RR coefficients for genetic and permanent environment effects, respectively. R denotes a matrix of residual variances.

Twelve RR models were tested considering different orders of the Legendre polynomial for the random effects (genetic and permanent environment) and four different residual variance structures (homogeneous, homogeneous within the year, homogeneous within the season of the year, or diagonal) (Table S1). The models were compared on goodness-of-fit by the Schwarz's Bayesian information criterion (BIC) scores as follows

$$\mathrm{BIC} = -2\log \mathrm{L} + \mathrm{k}\log(\mathrm{n} - \mathrm{r})$$

where *logL* denotes the logarithm of the likelihood function, *k* is the number of parameters estimated in the model, *n* is the total number of observations, and *r* is the rank of the incidence matrix of fixed effect (Wolfinger, 1993).

Estimated genetic values (EGV_i) for each accession in each harvest were determined on the original scale based on the estimates of the chosen model according to the following expression

$$\mathrm{EGV_i} = \sum_{\mathrm{m}=0}^{\mathrm{M}} \mathrm{g_{im}}\phi_{\mathrm{m}}(\mathrm{a_{ij}^{*}})$$

where $\hat{\mathrm{g}}_{\mathrm{im}}$ is the matrix of RR coefficient of order *m* for the genetic effects of the accessions. Variance components for genetic and permanent environment effects were obtained on the original scale, according to Kirkpatrick et al. (1990):

$$\hat{\sigma}_{\mathrm{x}} = \phi_{\mathrm{m}}(\mathrm{a_{ij}^{*}})\hat{\mathrm{K}}_{\mathrm{x}}\phi_{\mathrm{m}}(\mathrm{a_{ij}^{*}})'$$

where the index *x* denotes both the genetic or permanent environment effects. Broad sense heritability (h_j^2) on cutting *j* was obtained as follows

$$h_j^2 = \frac{\hat{\sigma}_{g_j}^2}{\hat{\sigma}_{g_j}^2 + \hat{\sigma}_{p_j}^2 + \hat{\sigma}_{e_j}^2}$$

where $\hat{\sigma}_{g_j}^2$ is the genetic variance on harvest *j*, $\hat{\sigma}_{p_j}^2$ is the permanent environment variance on harvest *j*, and $\hat{\sigma}_{e_j}^2$ is the residual variance on harvest *j*. Accuracy ($\hat{r}_{ij}$) estimates for each harvest was obtained as follows

$$\hat{r}_{ij} = \sqrt{1 - \frac{PEV_{ij}}{\hat{\sigma}_g^2}}$$

where PEV_{ij} is the prediction error variance of the accession i in harvest j extracted from the diagonal of the inverse of the coefficient matrix of mixed model equations.

In order to capture the genetic trend of each accession, we built trajectory curves using the EGVs. The scores of the curves were clustered using the k-mean algorithm. K-means is a clustering technique that uses a function to classify individuals into the centroid of their respective group (Nascimento et al., 2018) based on previous information about the number of clusters. For this alfalfa germplasm, we define four persistence groups to avoid the straight division between persistent versus non persistent. So, for the k-means method we settled the following groups: C1: High persistence, C2: Persistent, C3: Modest persistence, C4: Non-persistent. In order to confirm the number of clusters, the Tocher clustering approach was used.

Once the persistence clusters were defined, an Artificial Neural Network (ANN) approach was used to capture the previous classification but in an automated way. Besides the ANN for the EGV data, an ANN for phenotypic information was also tested to check how the noise present in the phenotypic data could affect the ANN's efficiency. Since the aim was to compare the different dataset, the ANN topology was the same for both phenotypic and EGV data. Before using the ANNs, the matrix of 77 (accessions) by 25 (the first column

corresponding to the respective k-means cluster + EGVs or phenotypic value of each harvest) was partitioned so 75% of the information within each cluster was used in the training and the remaining 25% in the validation step. Multilayer Perceptron (MLP) networks were used for both data with a logistic activation function (logsig). The ANN topology included three neurons in one hidden layer and the backpropagation training (trainbr) algorithm with 5,000 epochs. The input layer corresponded to the 24 information of each accession belonging to each k-mean group, whereas the output layer was made up of a single neuron represented by a vector of known elements (labeled C1, C2, C3, and C4). The apparent error rate (APER) was used as a criterion of efficiency of the ANNs. The statistical analyses were performed on ASReml 4.1 (Gilmour et al., 2015), software R (R Development Core Team, 2020), and software Genes (Cruz, 2016).

RESULTS

Random Regression model selection. The best RR model according to the BIC criteria and ASReml warning code for parameters included a random third-order Legendre polynomial for both genetic and permanent environment effects, and a diagonal residual variance structure (Table 1).

Table 1. Random regression models fitted through Legendre polynomials to describe the dry matter yield trajectory over 24 alfalfa cuttings. The general description for the models is $G_xP_yR_z$, where G_x, P_y stands for the degrees of Legendre polynomials for genetic and permanent environment effects, respectively. R_z denotes the residual variance structure that may assume homogeneous, homogeneous within year, homogeneous within season of year, or diagonal structures. The goodness-of-fit for each model was accessed by the Schwarz Bayesian information criteria (BIC).

Model	Effect's degree		P[a]	LogL	Constraints[b]	BIC
	Genetic	Perm Env.				
$G_3P_3R_1$	3	3	21	converged	P	11702.85

$G_3P_3R_2$	3	3	22	converged	P	11691.88
$G_3P_3R_8$	3	3	28	converged	P	11657.15
$G_3P_3R_{24}$	3	3	44	converged	P	11368.24
$G_3P_2R_1$	3	2	17	converged	P	11610.97
$G_3P_2R_2$	3	2	18	converged	P	11601.56
$G_3P_2R_8$	3	2	24	converged	P	11567.80
$G_3P_2R_{24}$	3	2	40	converged	P	11290.15
$G_2P_3R_1$	2	3	17	converged	P	11553.35
$G_2P_3R_2$	2	3	18	converged	P	11542.54
$G_2P_3R_8$	2	3	24	converged	P	11511.76
$G_2P_3R_{24}$	2	3	40	converged	P	11209.57
$G_2P_2R_1$	2	2	13	converged	P	11489.03
$G_2P_2R_2$	2	2	114	converged	P	11479.95
$G_2P_2R_8$	2	2	20	converged	P	11451.09
$G_2P_2R_{24}$	2	2	36	converged	P	11153.17

[a]Number of parameters
[b]ASReml warning code for the fitted parameters. P indicates a positive definite matrix.

According to the BIC criteria, the more parametrized the residual structure was, the better was the model. Regardless of the order of genetic or permanent environment effects, models with diagonal residual variance (one variance per harvest) fitted better than any other model. Models with one variance adjusted in each season of the year performed better than all models but the ones with a diagonal residual variance structure, and so on.

Genetic parameters of DMY overtime. The genetic variance trajectory fitted by $G_2P_2R_{24}$ had an uptrend pattern over the whole period (Figure 1). The comparison between the genetic variance of the first (24,560) and last (149,568) cuttings showed a six-fold higher variance in the last compared to the first. The higher increase was observed from the 21th harvest on. The permanent environment variance had a downtrend pattern until the 13th harvest and then got an uptrend pattern until the last cutting. As observed in the genetic trajectory, the higher increase in the permanent environment trajectory was observed from the 21th harvest on. The permanent

environment variance was higher than the genetic variance only in the first harvest (27,381.18 versus 24,560.1).

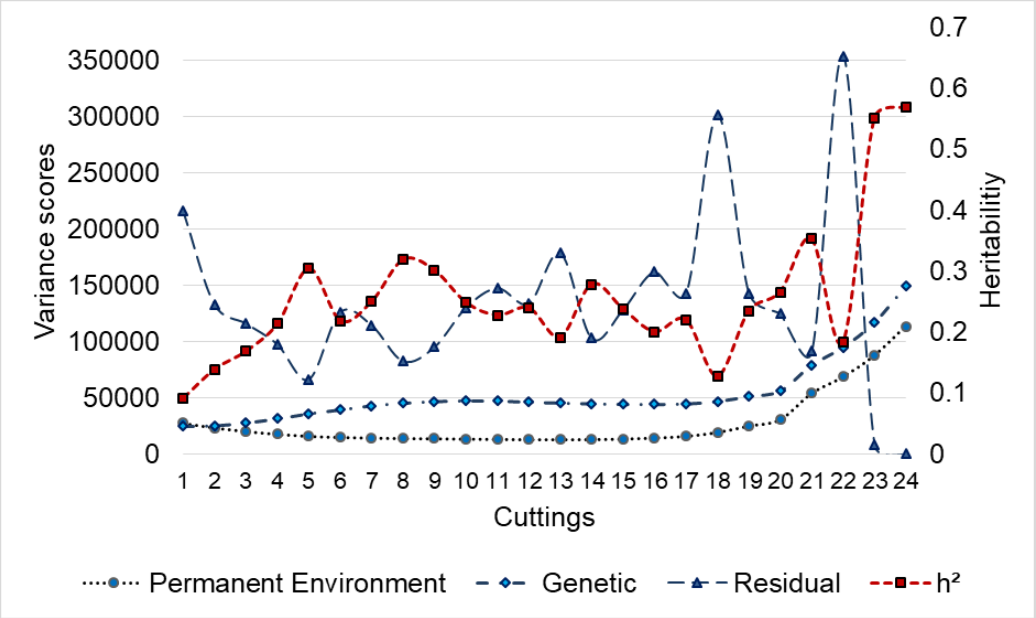
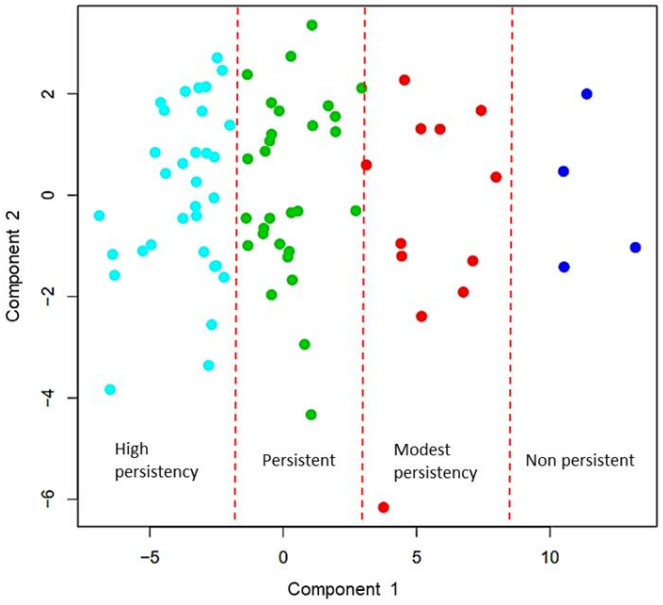


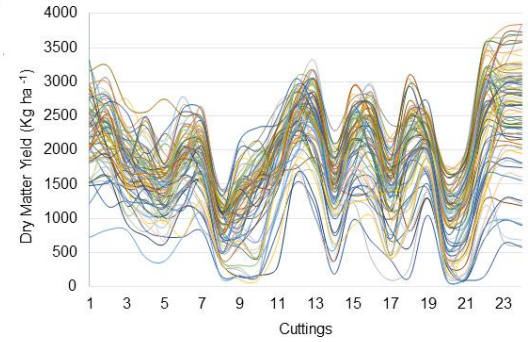
Figure 1**.** Trends for genetic, permanent environment, and residual variances and heritability (h^2) obtained by a Random Regression model through Legendre polynomials.

Broad-sense heritability had the lower estimate in the first harvest (0.09). The heritability increased slightly from the first to the fifth harvest until the first peak. From harvest six on, small peaks were established before a sharp decrease on harvest 22. The higher estimates were observed in the last two harvests (0.54 and 0.56). Genetic values from the sixth harvest to the 23rd were predicted with more than 90% average accuracy (Figure S2).

Alfalfa's persistence curves and persistence method. The k-means method accomplished 32 accessions in the 'High persistency' cluster (Figure 2, Table S1). The 'Non-persistent' cluster included the four accessions with the lowest EGV scores overtime: Don Enrique, CW 620, Magna 601, and ProINTA Patricia. The 'Persistent' cluster included 29 out of 77 accessions, while the 'Modest persistency' cluster grouped together 12 accessions. The accession

trajectories for all genetic materials were not linear. They had clear cycles of either high or low genetic values overtime.





Total of accessions	C1	C2	C3	C4	APER (%)
	10	9	4	1	
Phenotypic	8	6	2	0	33.3
EGV	10	9	4	1	0

Figure 2. Clustering of alfalfa accessions based on estimated genetic values over 24 cuttings (left). Each cluster (High persistence, Persistent, Modest persistence, and Non persistent) indicates similar dry matter yield trajectories overtime. The different colors of the accessions represent different groups according to the Tocher algorithm. Dry matter yield trajectories for each accession (upper right) show cycles of high or low yield through the cuttings. The Apparent Error Rates (APERs) for validation of the phenotypic and EGV artificial neural networks are shown in the lower right table. Table S2 contains a detailed description of the accessions.

The high persistence accessions had genetic values ranging from -143.6 to 485.8 in the first cutting and from -343.6 to 592.4 in the last one (Figure 3). All of these accessions yielded more than 47 t in 24 cuttings, with an average of 48 t for the whole period, or 24 t/year. The non-persistent accessions had genetic values ranging from -654.82 to -346.5 in the first harvest and from -898.81 to -586.3 in the last one. The average yield was ~30 t in 24 cuttings, or 15 t per year.

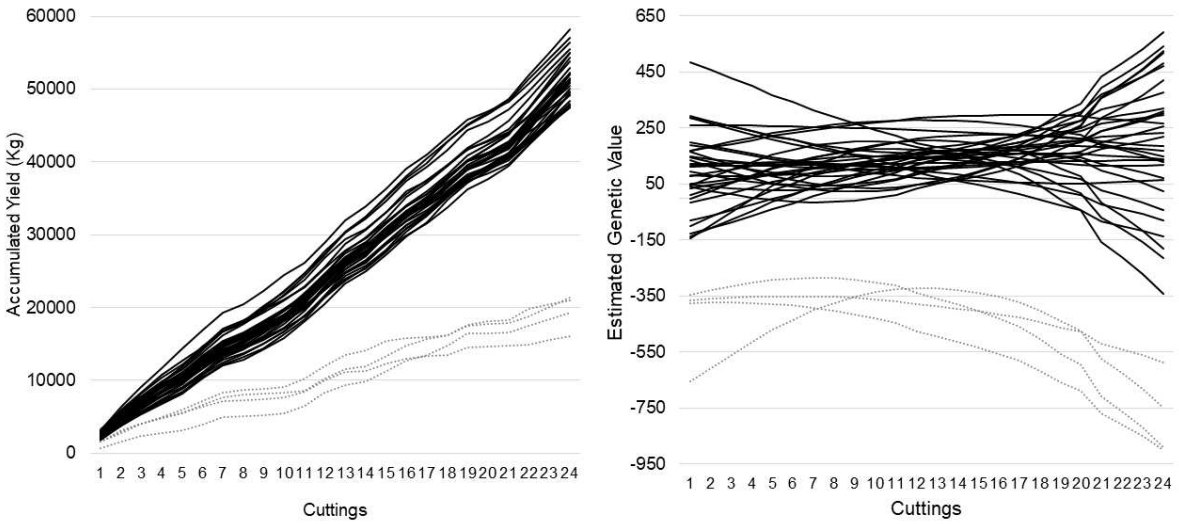


Figure 3. Accumulated alfalfa's dry matter yield (left) and estimated genetic values (EGVs) (right) over 24 cuttings. The top-32 yielding accessions (left) and top-32 EGV trajectories (right) are shown in black lines, while the four less yielding and the four accessions with the lower EGV trajectories are shown in grey lines.

Fifty-three accessions were used to train both the phenotypic and EGV ANNs, while 24 composed the validation step. The apparent error rate (APER) of the persistence ANNs ranged from 33.3% using the phenotypic data to zero using the EBVs (Figure 2) in the validation step. All the accessions were correctly classified in the training for both phenotypic and EGV data.

DISCUSSION

Random Regression model selection. RR models have been extensively used to handle longitudinal data (Campbell et al., 2019) due to their ability to provide a strong framework for modelling trajectories for genetic and non-genetic effects overtime (Sun et al., 2017). Covariances between measures are fitted into the RR model using splines or polynomial functions (Meyer, 2000), utilizing fewer parameters. The 16 different RR models fitted in this study considered Legendre polynomials of third or quartic order (second or third degree) for

both genetic and permanent environment effects. We based the model's choice on BIC as well as on the ASReml warning code to find out the best fit of the Legendre polynomial for both genetic and permanent environment effects (third order). High-order Legendre polynomials rapidly change at the extremes, causing poor estimates of genetic parameters and variance components (Boligon et al., 2012).

Some studies have indicated that better goodness of fit can be accomplished if different residual variances are considered according to periods or time intervals (Oliveira et al., 2019). Alves et al. (2020) modeled reaction norms through RR models using a diagonal residual structure for analysis of multi-environment trials in tree breeding. Rocha (2019) used a diagonal structure after testing three different structures for elephantgrass persistence analysis. In this study, we investigated whether residual variances would be modeled assuming homogeneity for specific pieces of time within the evaluated period. Among the four different variance structures for the residuals, the models with a diagonal structure had a better fit than assuming homogeneity in the whole period, homogeneity within the year, and homogeneity within the season of the year. Our findings confirm that the conditions at each measurement time determine the need to assume a different residual variance per harvest. Genetic parameters of alfalfa accessions cannot consider the same residual variances in the same season of the year, even though they do not affect nutritive value or forage yield traits (Santos et al., 2018). The same can be extended to homogeneous residual variance within the year.

Genetic parameters of DMY overtime. In terms of importance in understanding the genetics of persistence, genetic and permanent environment trends have the upper hand. Permanent environment effect is present in longitudinal data and is estimated by the variance among measures made on a given individual in its plot. Genetic factors such as dominance, epistatic, or maternal effects influence the permanent environment and have a direct impact on plant development, especially in early harvests (Rocha, 2019). The variance of permanent

environment effect value was 27381.18 in the first harvest, while the genetic variance was 24560.1. From the second harvest on, the permanent environment variance held smaller estimates than the genetic variance. Since the word permanent suggests stability (Schaeffer, 2011), it is worth to guaranteeing equal environmental conditions, mainly before the first harvest, to minimize the lack of casualization in the different growing seasons. Meteorological conditions for our experiment on the first cuttings were favorable for good alfalfa establishment (Figure S1).

The genetic trends revealed an overall increasing pattern that did not change significantly during the first year. This is consistent with Wilkins and Humphreys' (2003) assertion that persistence has a major influence on DMY, particularly from the second year of cultivation on. Because persistence refers to the ability of an accession to keep its genetic trajectory of EGV over time, our findings suggest that persistence and DMY genes were more expressed in the second year, resulting in an extreme phenotypic expression and, as a result, higher genetic variance.

Alfalfa's persistence criteria and performance overtime. The fitted curves showed a great amplitude regarding DMY overtime (Figure 2), which suggests a high variability of persistence. In fact, the accessions have significant genetic variance for DMY as stated by Santos et al. (2020). The authors analyzed phenotypic and molecular data and indicated a set of accessions that would compose a good base population exploring traits of nutritive value and forage yield. Different from our approach, residuals were not modeled and accessions previously indicated as good were found to be in the non-persistent group in this study (e.g. ProINTA Patricia and Magna 601).

Overall, the non-persistent accessions decreased their trajectories for the whole period. For the high persistence cluster, some accessions held negative EGVs in the first harvest but compensated their performance in later harvests (Figure 2). One example was the Sequel

accession that had negative EGVs until the sixth harvest and then held an increase until the last harvest. Accumulated yield and persistence are highly correlated, but they have to be carefully examined because an accession with almost no yield in a short period within the 24 cuttings will not have a great influence on the correlation, but it is still not interesting for alfalfa producers. Persistence is a critical trait in forage breeding and depends on survival ability as well as the possession of stolons that allow revegetation (Bouton, 2012). While in some crops, modeling differences between cultivars is sufficient, for alfalfa and other perennial pastures, it is important to model the response over time (Faveri et al., 2015). Predictions of persistence may be required for times other than individual measurements.

Based on APER, ANN can be a useful approach for classifying accessions regarding persistence if the inputs are less biased than the phenotypic data. Our results showed a 33.3% APER for the ANN using phenotypic data compared to 0% when using EBV in the validation step. Even though ANNs work well with noise in the data, they can improve their generalization capacity when more accurate data is available. Because ANNs do not require a linear or any other data structure, they achieve high efficiency by working with complex traits. They can capture relationships that are not captured by stochastic models as well as generalize information to new cases (Santos et al., 2019).

The summary of the persistence method for the first-time-users is: First, it is necessary to fit a RR to capture the genetic trajectories of accessions (ASREML code available as supplementary material). Second, clustering accessions of similar persistence using k-means analysis (available in the software GENES). Third, to establish an ANN to capture the same persistence pattern shown by the k-means clustering (available if requested).

The method for accessing persistence presented in this study involved an RR model to obtain trends of persistence, a k-means method to define different persistence clusters, and then an ANN to perform persistence classification in an automated way. Once the ANN is

established, new EGVs obtained from a RR model can be used directly in the ANN, with no need to use any clustering technique. Researchers will be able to simplify the persistence method by eliminating the clustering step because the ANN can learn how the clustering technique classifies the accessions. When new accessions are evaluated by the alfalfa breeding program, they will be classified according to their EGV scores using the same ANN previously fitted, using two steps rather than three. The ANN will be responsible for the decision-making process.

Three problems can affect alfalfa persistence in tropical areas, according to Bouton (2012): grazing, (Al)-toxic soils, and drought. The challenge for alfalfa cultivation in Brazil is due to the lack of knowledge of how to grow the crop and the lack of registered cultivars. Breeding potential of the available germplasm has shown satisfactory though its temperate background (Santos et al., 2020). Selection of materials showing persistence in tropical areas is not straightforward. The first step is to identify such materials and then to build up crossing strategies to ensure the transfer of the trait from the base to breeding populations and keep it in the cultivars. ANN can make the classification easier as long as the correct information is included in the model. Longitudinal alfalfa data can be simply classified based on its genetic values rather than using individual phenotypic data or using visual criteria from graphical dispersions. Taken together, our findings can help with the understanding of alfalfa persistence behavior in tropical areas. The persistence method can support the selection of genotypes to compose base populations that will generate populations showing consistent DMY overtime.

REFERENCES

Alves, R. S., Resende, M .D. V., Azevedo, C. F., Silva, F. F., Rocha, J. R. A. S. C., Nunes, A. C. P., Carneiro, A. P. S., & Santos, G. A. (2020). Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to

environmental gradients. *Tree Genetics & Genomes, 16,* 1-8. https://doi.org/10.1007/s11295-020-01431-5

Azevedo, A. L. S., Costa, P. P., Machado, J. C., Machado, M. A., Pereira, A. V., & Lédo, F. J. S. (2012). Cross Species Amplifcation of *Pennisetum glaucum* Microsatellite Markers in Pennisetum purpureum and Genetic Diversity of Napier Grass Accessions. *Crop Science, 52,* 1776-1785. https://doi.org/10.2135/cropsci2011.09.0480

Boligon, A. A., Mercadante, M. E. Z., Lôbo, R. B., Baldi, F., Albuquerque, L. G. (2012). Random regression analyses using B-spline functions to model growth of Nellore cattle *Animal, 6,* 212-220.

Bouton, J. H. (2012). Breeding Lucerne for persistence. *Crop and Pasture Science, 63,* 95-106. https://doi.org/10.1071/CP12009

Campbell, M., Momen, M., Walia, H., & Morota, G. (2019). Leveraging Breeding Values Obtained from Random Regression Models for Genetic Inference of Longitudinal Traits. *The Plant Genome, 12,* e180075. https://doi.org/10.3835/plantgenome2018.10.0075

Comeron, E.A., Ferreira, R.P., Vilela, D., Kuwahara, F. A., & Tupy, O. (2015). Utilização da alfafa em pastejos para alimentação de vacas leiteiras. In: R.P. Ferreira, D. Vilela, E.A. Comeron, A.C.C. Bernardi, & D. Karam (Eds.), *Cultivo e utilização da alfafa em pastejo para alimentação de vacas leiteiras* (pp.13-16). Embrapa Editor, Brasília, Brazil.

Cruz, C. D. (2016). Genes Software – extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum, 38,* 547-552.

Faveri, J., Verbyla, A. P., Pitchford, W. S., Venkatanagappa, S., & Cullis, B. R. (2015). Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials. *Crop and Pasture Science, 66,* 947-962. https://doi.org/10.1071/CP14312

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis 2nd ed.* Boston, Wiley.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J., & Thompson, R. (2015). ASReml user guide release 4.1 structural specification. Hemel hempstead VSN Int ltd.

Kranis, A., Su, G., Sorensen, D., & Woolliams, J. A. (2006). The Application of Random Regression Models in the Genetic Analysis of Monthly Egg Production in Turkeys and a Comparison with Alternative Longitudinal Models. *Poultry Science, 86,* 470-475.

Lamb, J. F. S., Sheaffer, C. C., Rhodes, L. H., Sulc, R. M., Undersander, D. J., & Brummer, E. C. (2006). Five Decades of Alfalfa Cultivar Improvement: Impact on Forage Yield, Persistence, and Nutritive Value. *Crop Science, 46,* 902-909. https://doi.org/10.2135/cropsci2005.08-0236

Li, X., Wei, Y., Acharya, A., Hansen, J. L., Crawford, J. L., Viands, D. R., Michaud, R., Claessens, A., & Brummer, E.C. (2015). Genomic Prediction of Biomass Yield in Two Selection Cycles of a Tetraploid Alfalfa Breeding Population. The Plant Genome, 8, 1-10. https://doi.org/10.3835/plantgenome2014.12.0090

Meyer, K. (2000). Random regressions to model phenotypic variation in monthly weights of Australian beef cows. *Livestock Production Science, 65,* 19-38. https://doi.org/10.1016/S0301-6226(99)00183-9

Nascimento, M., Campana, A. C., & Cruz, C. D. (2018) RBF – Redes Funções de Base Radial. In: C.D. Cruz & M. Nascimento (Eds.), *Inteligência Computacional Aplicada ao Melhoramento Genético* (pp.292-309). UFV Editor, Viçosa, Brazil.

Oliveira, H. R., Brito, L. F., Lourenco, D. A. L., Silva, F. F., Jamrozik, J., Schaeffer, L. R., & Schenkel, F. S. (2019). Advances and applications of random regression models: From quantitative genetics to genomics. *Journal of Dairy Science, 102,* 7664-7683. https://doi.org/10.3168/jds.2019-16265

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robins, J. G., Hansen, J. L., Viands, D. R., & Brummer, E. C. (2008). Genetic Mapping of Persistence in Tetraploid Alfalfa. *Crop Science, 48,* 1780-1786. https://doi.org/10.2135/cropsci2008.02.0101

Rocha, J. R. A. S. C. (2019). *Elephantgrass breeding focused on persistence and discover candidate genes* [Doctoral dissertation, Federal University of Viçosa]. Federal University of Viçosa Digital Repository. https://locus.custom.ufv.br//handle/123456789/25208

Santos, I. G., Cruz, C. D., Nascimento, M., Rosado, R. D. S., & Ferreira, R. P. (2018). Direct, indirect and simultaneous selection as strategies for alfalfa breeding on forage yield and nutritive value. *Pesquisa Agropecuária Tropical, 48,* 178-189. http://dx.doi.org/10.1590/1983-40632018v4851950

Santos, I. G., Cruz, C. D., Nascimento, M., & Ferreira, R. P. (2019). Selection index as a priori information for using artificial neural networks to classify alfalfa genotypes. *Genetics and Molecular Research, 18,* 18221. http://dx.doi.org/10.4238/gmr18221

Santos, I. G., Rocha, Rocha, J. R. A. S. C., Vigna, B. B., Cruz, C. D., Ferreira, R. P., Basigalup, D. H., & Marchini, R. M. S. (2020). Exploring the diversity of alfalfa within Brazil for tropical production. *Euphytica, 216, 72*. https://doi.org/10.1007/s10681-020-02606-w

Schaeffer, L. R. (2011). Cumulative permanent environmental effects for repeated records animal models. *Journal of Animal Breeding Genetics, 128,* 95-99. https://doi.org/10.1111/j.1439-0388.2010.00894.x

Schaeffer, L. R. (Ed.). (2016). *Random Regression Models*. Available at: http://animalbiosciences.uoguelph.ca/~lrs/BOOKS/rrmbook.pdf.

Shi, S., & Smith, K. F. (2017). The Current Status, Problems, and Prospects of Alfalfa (*Medicago sativa* L.) Breeding in China. *Agronomy, 7,* 1-11. https://doi.org/10.3390/agronomy7010001

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J. L., & Sorrells, M. E. (2017). Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *Plant genome, 2,* 1-12. https://doi.org/10.3835/plantgenome2016.11.0111

Wolfinger, R.D. (1993). Covariance structure in general mixed models. *Communications in Statistics – Simulation and Computation. 22B,* 1079-1106. https://doi.org/10.1080/0361091930881314

SUPPLEMENTARY INFORMATION

Table S1. Description of the 12 random regression models with different degrees of random effects and different residual structures

Model	Polynomial degree for each effect			
	Fixed	Genetic	Permanent environment	Residuals
$G_3P_3R_1$	23	3	3	Homogeneous
$G_3P_3R_2$	23	3	3	Homogeneous within year
$G_3P_3R_8$	23	3	3	Homogeneous within season of the year
$G_3P_3R_{24}$	23	3	3	Diagonal
$G_3P_2R_1$	23	3	2	Homogeneous
$G_3P_2R_2$	23	3	2	Homogeneous within year
$G_3P_2R_8$	23	3	2	Homogeneous within season of the year
$G_3P_2R_{24}$	23	3	2	Diagonal
$G_2P_3R_1$	23	2	3	Homogeneous
$G_2P_3R_2$	23	2	3	Homogeneous within year
$G_2P_3R_8$	23	2	3	Homogeneous within season of the year
$G_2P_3R_{24}$	23	2	3	Diagonal
$G_2P_2R_1$	23	2	2	Homogeneous
$G_2P_2R_2$	23	2	2	Homogeneous within year
$G_2P_2R_8$	23	2	2	Homogeneous within season of the year
$G_2P_2R_{24}$	23	2	2	Diagonal

Table S2. Description of the genetic materials that originated some of the evaluated accessions and persistence classification according to the k-means method. As more materials participate in the crosses as broader the accession's genetic base. Accessions that have more than 10 parents are considered to have a broad genetic base. The parents that compose the crosses were chosen based on yield and resistance to pests and diseases.

Accession	Cluster	Genetic origin	Adaptation
5681	3	Ancher (34%) Meteor (16%) Cuf01 (6%) 555 (6%) Mercury (3%), 524. 526. 5331. 5432. Apollo, NCMP10	Central Argentina
ACA 900	1	CW2820 (16%) DK189 (14%) CW2818, CW2817, Mecca, Express UC332, UC176, UC196, UC226, UC276, UC296, UC222, UC263, UC189, UC231, UC195	Northern Argentina, Santa Fé, and Córdoba
ACA 901	2	UC332, SW9212, SW8210	Argentina
Activa	2	VS448 (17%), CW2820 (14%), Mecca (12%), DK189 (10%), Condor (8%), UC332 (6%), CW2815 (5%), CW2817 (5%), VS446 (3%), Armona (3%), Sundor (3%), Express (2%), DK187 (2%), UC176 (2%), UC196 (2%), UC222 (2%), UC236 (2%), UC276 (2%)	Argentina
Bacana	2	Beacon (305), DK189 (30%), Coronado (20%), Zaino (20%)	Argentina
Bacana 1	1		Argentina
Bar Pal 5	3		Argentina
Bar Pal 10	1		Argentina
Baralfa 85	2	CA G, WI457, WI514, Cuf101	Argentina
Barbara SP INTA	1	Monarca SP INTA (25%), WL516 (25%), Armona (25%), C/W331 (25%)	Pampas Argentinos
California 50	2		Southern Califórnia, Southest Arizona
CUF 101	2	UC Cargo (55%), UC Salton (1%), UC76 (22%), 1972 Breeding Mixture (20%), Niagara N71 (2%)	Southern Califórnia, Southest Arizona
CW 1010	1		Argentina
CW 194	1		Argentina
CW 620	4	SPS6550 (11%), DK166 (8%), Express (7%), N650 (6%), Archer (2%), AlfaStar (2%), Bighorn (1%), Cal/West Seeds (63%)	Argentina
CW 830	1	Alfa200 (9%), Monarca SPI (9%), DK189, Weston (3%), WL525HQ (1%), Cal/West Seeds (69%)	Argentina
Diamond	2		Argentina
DK 166	3	Express, Condor, Valley+, VS626, Shenandoah, VS481, Mede	Sacramento, California, Idaho, New Mexico, and Buenos Aires
DK 181	2		Argentina
DK 187 R	1		Argentina
DK 192	1		Argentina
DK 194	1	Topacio, DK192. Grasis, ACA900, Super Supreme, Mecca, F969, DK191, CalWest Seeds	Argentina, Southest USA

Don Enrique	4	Aurora (20%), 5683 (18%), Vitoria SP INTA (30%), Primavera (20%)	Argentina
F 708	2		Argentina
Florida 77	1		Argentina, Southest USA
G 909	1	82-296 F (35%), C346 (32%), C442 (2%), Alazan (14%)	Argentina
GAPP 969	2	Mecca (10%), DK189 (10%), UC332 (10%), Cal/West Seeds (70%)	Argentina, Southest USA
Gateado	2	GH97-53/55/56/58/60/65 (52%), FG1150/1151/1159/1188 (24%), 5683 (10%), Yolo (7%), ARG96-1 (4%), Tahoe (3%)	Argentina
Kern	1		Argentina
LE N 1	2		Argentina
LE N 2	2		Argentina
LE N 3	3		Argentina
LE N 4	2		Argentina
LPS 8500	1		Argentina
Magna 601	4	Sutter varieties	Argentina
Magna 804	2		Argentina
Magna 860	1		Argentina
Magna 868	1		Argentina
Maitena	1		Argentina
Mecha	1	FG94-55 (18%), FG94-48 (17%), FG94-49 (15%), FG94-46 (13%), FG94-13 (2%), FG94-12 (1%), FG94-15 (1%), DK193 (20%), DK180ML (13%)	Argentina
Medina	2		Argentina
Milonga II	1	FG 988 (30%), FG1164 (6%), Coronado (31%), DK193 (33%)	Argentina
Monarca	2		Argentina
Monarca SP INTA	2		Argentina
P 30	1		Argentina
P 5715	2		Argentina
Patriarca	3		Argentina
Patricia	2		Argentina
Pintado	3		Argentina
Pinto	3		Argentina
Primavera	2		Argentina
ProINTA Carmina	2	Monarca SP INTA, 5929, Mecca, Sequel	Argentina
ProINTA Luján	3		Argentina
ProINTA Mora	2	Monarca (30%), Rocío ISP (17%), DK189 (13%), AL102 (9%), Sima15 (7%), 5929 (5%), Condor (4%), P30, Sundor, 581, DK181, WL605, AL, Sima1/3/6/16/29	Argentina
ProINTA Patricia	4	CW4496, and high persistency materials	Argentina
ProINTA Patricia 1	1		Argentina
ProINTA Super Monarca	2	Yolo (20%), Maxidor (18%), Monarca SP INTA (46%), Mecca (16%)	Pampas Argentinos
Queen 910	3	Araucana (10%), Super Lenchera (10%), Máxima (10%), Falcon (15%), DK192 (15%), Trinidad87 (20%), Monarca SP INTA (20%)	Argentina
Rio Grande	2		Argentina
Ruano	1		Argentina
Ruano 1	1		Argentina

Sequel	1		Argentina
Sequel 2	2		Argentina
Siriver 2	1		Argentina
SPS 6550	2		Argentina
Trinidad 87	1		Argentina
Verdor	1		Argentina
Verzi	2		Argentina
Victoria SP INTA	3	WL 313 (20%), 77-T-25 (40%), 77-78 CaB (40%)	Argentina
Villa	3	Máxima (25%), Monarca SP INTA (25%), Super Lechera (25%), Armona (25%)	Argentina
Winter	2		Argentina
WL 1058	1	Rosillo (53%), FG10A215 (23%), FG8M809 (16%), FGA016TF (4%), FG99x49/50 (3%), DK193 (1%)	Pampas Argentinos
WL 516	3		Santa Fé, Córdoba, Buenos Aires
WL 525	1	WL516, 86-222-CA, 898, Maxidor	Santa Fé, Córdoba, Buenos Aires
WL 818	1	Pintado (54%) WL 611 (46%)	Argentina
WL 903	1		Argentina
Crioula	1		Argentina, Brazil

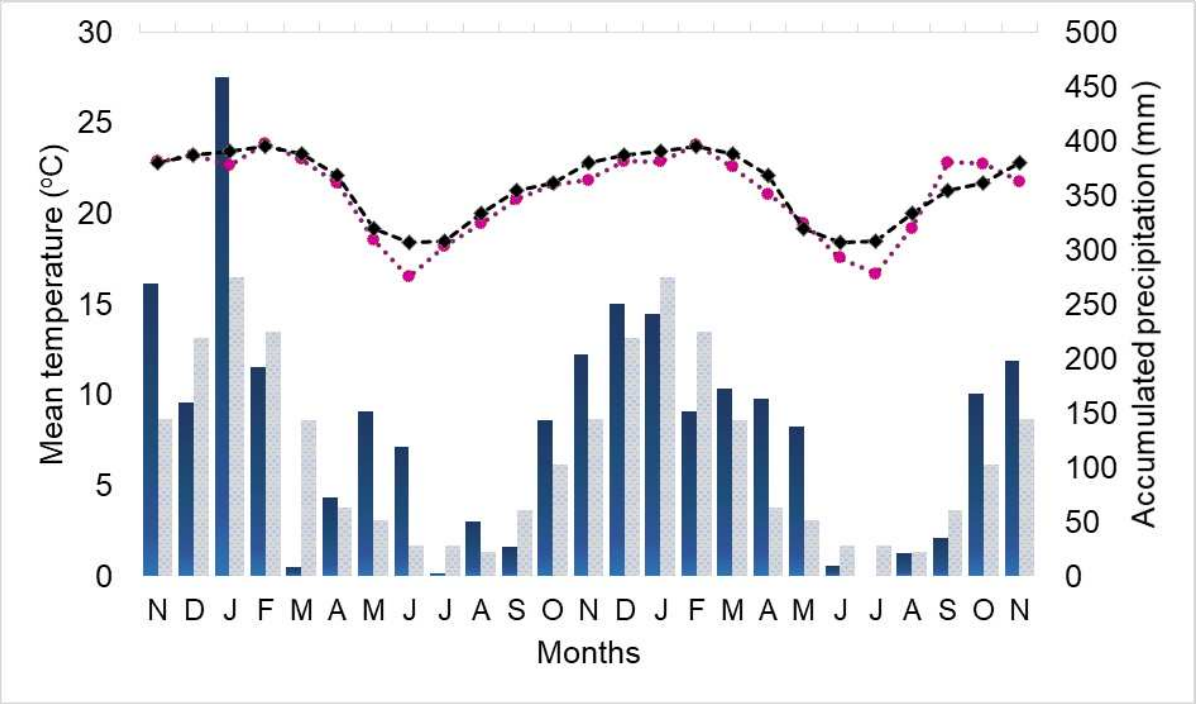


Figure S1. Meteorological data for São Carlos, São Paulo, Brazil. Mean temperature (dark blue bars) and accumulated precipitation (purple line) for the period from November 2015 to November 2017. Monthly averages from 1992 to 2010 were represented by light blue bars (mean temperature) and a black line (accumulated precipitation average).

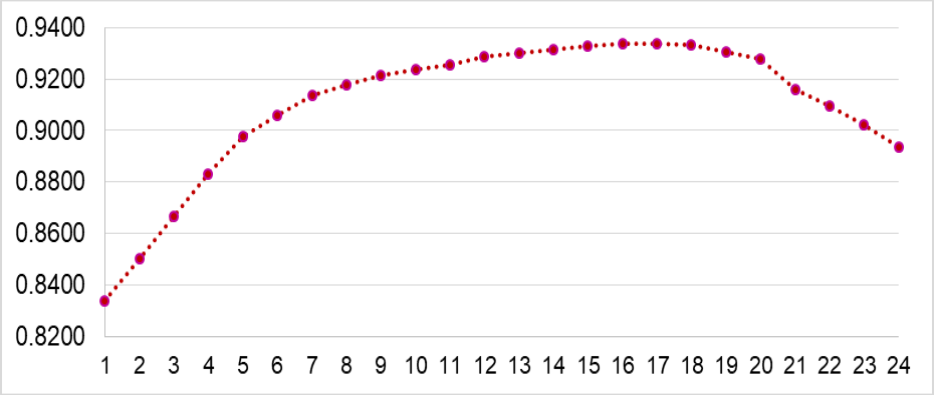


Figure S2. Accuracy estimates of 24 alfalfa cuttings obtained by a Random Regression model through Legendre polynomials.

GENERAL CONCLUSIONS

Based on the investigation with regard to the genetic diversity of the alfalfa germplasm hold by Embrapa Southeast Livestock, we can affirm that the accessions have the potential to generated synthetic populations adapted to tropical conditions. However, the information generated by the genetic diversity analyses should not be evaluated by itself, since the persistence criteria was not used as input for the SOM analysis. Our findings on persistence revealed a large variation of this trait in the germplasm, which means that it is possible to achieve real genetic gains if we cross the right accessions. The information generated by the genetic diversity analyses with the persistence classification are complementary approaches to guide crosses and consequently, to obtain superior alfalfa cultivars. Considering all the information generated by this study we can infer that polycrosses including Pro INTA SuperMonarca, Mecha, WL 525, ACA 900, Bacana, CUF 101, Crioula, and Ruano are promissing. These accessions have at least 50 distinct parents, present favorable alleles for regrowth ability, biomass yield, and were classified as persistent.

APPENDIX A

This document describes how to perform the analyses regarding persistence presented in this study. First, the ASREML code for adjusting random regression models; second, how to get the genetic parameters (estimated genetic values, variance components, heritabilities, and accuracy) and rescale them from the Legendre to the real scale. Third, the k-means method for classifying accessions. Finally, the Multilayer Perceptron network that aids in the automation of alfalfa persistence classification.

ASREML code for fitting the Random Regression model considering a diagonal variance structure for the residuals and a second-degree Legendre polynomial for both genetic and permanent environment effects.

```
 !NOGRAPHS !WORKSPACE !RENAME !ARGS 1 // !DOPART $1

Title: RRM
 Med 24 !I
 Gen 77 !SORT
 MedRep *
 Perm *
 Grad *
 MS

data.txt !skip 1 !AISINGULARITIES !MAXIT 8000

MS ~ mu MedRep leg(Grad,23) !r leg(Grad,2).Gen leg(Grad,2).Perm !f mv

24 1 2
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
```

231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID
231 0 ID

leg(Grad,2).Gen 2
leg(Grad,2) 0 US !GP
((2^2+3*2+2)*0.5)*0
Gen

leg(Grad,2).Perm 2
leg(Grad,2) 0 US !GP
((2^2+3*2+2)*0.5)*0
Perm

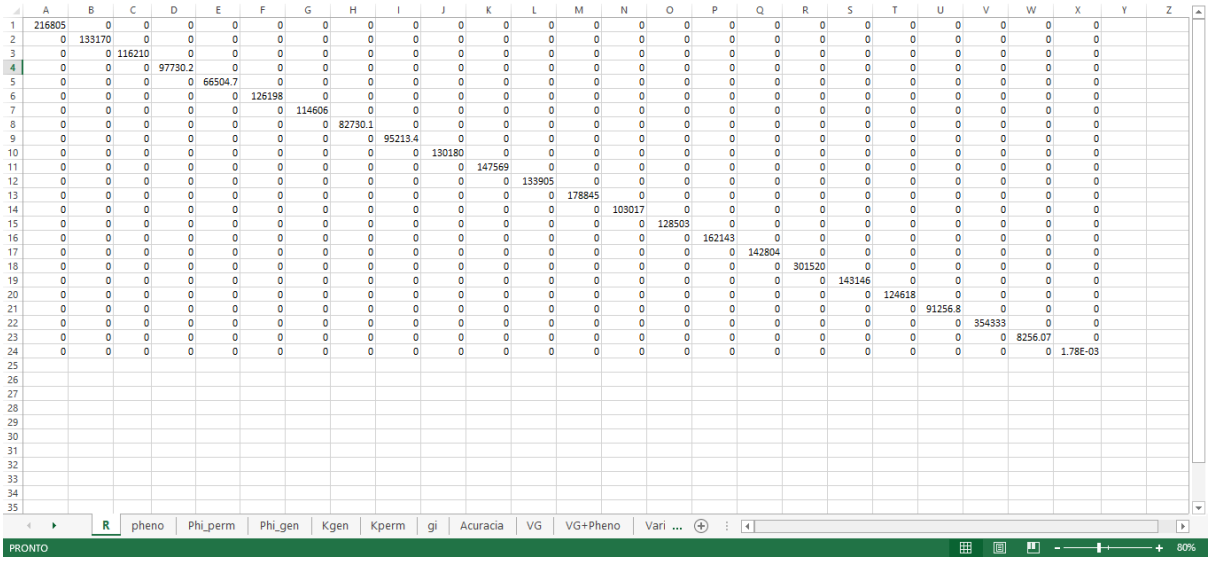
ASREML releases its outputs on a Legendre scale. Set the matrices up on Excel or any other software prior to the transformation into a real scale. Use the file .asr to check the LogL value (if you need to compare models), variance and covariance between Legendre coefficients for the genetic and permanent environment as well as the residuals.

```
Notice: LogL values are reported relative to a base of -30000.000
Notice:    196 singularities detected in design matrix.
  1 LogL=-6934.61     S2=  1.0000       5472 df     :  11 components restrained
  2 LogL=-6920.20     S2=  1.0000       5472 df     :  11 components restrained
  3 LogL=-6791.81     S2=  1.0000       5472 df     :   6 components restrained
  4 LogL=-6513.43     S2=  1.0000       5472 df     :   6 components restrained
  5 LogL=-6027.23     S2=  1.0000       5472 df
  6 LogL=-5580.54     S2=  1.0000       5472 df
  7 LogL=-5356.65     S2=  1.0000       5472 df     :   2 components restrained
  8 LogL=-5346.04     S2=  1.0000       5472 df     :   1 components restrained
  9 LogL=-5227.03     S2=  1.0000       5472 df     :   1 components restrained
 10 LogL=-5183.65     S2=  1.0000       5472 df     :   1 components restrained
 11 LogL=-5170.50     S2=  1.0000       5472 df     :   1 components restrained
 12 LogL=-5167.63     S2=  1.0000       5472 df     :   1 components restrained
 13 LogL=-5167.22     S2=  1.0000       5472 df
 14 LogL=-5167.17     S2=  1.0000       5472 df
 15 LogL=-5167.17     S2=  1.0000       5472 df
 16 LogL=-5167.17     S2=  1.0000       5472 df
 17 LogL=-5167.17     S2=  1.0000       5472 df
```

```
         - - - Results from analysis of MS - - -

Source                Model  terms     Gamma     Component    Comp/SE   % C
Residual               5544   5472
Variance[  1]           231      0   216067.      216067.       9.77   0 P
Variance[  2]           231      0   132354.      132354.       9.54   0 P
Variance[  3]           231      0   113963.      113963.       9.61   0 P
Variance[  4]           231      0   102008.      102008.       9.72   0 P
Variance[  5]           231      0   71037.5      71037.5       9.55   0 P
Variance[  6]           231      0   124500.      124500.      10.13   0 P
Variance[  7]           231      0   113096.      113096.      10.13   0 P
Variance[  8]           231      0   81867.0      81867.0       9.93   0 P
Variance[  9]           231      0   89999.0      89999.0       9.95   0 P
Variance[ 10]           231      0   123398.      123398.      10.11   0 P
Variance[ 11]           231      0   148507.      148507.      10.22   0 P
Variance[ 12]           231      0   131718.      131718.      10.12   0 P
Variance[ 13]           231      0   175593.      175593.      10.26   0 P
Variance[ 14]           231      0   103552.      103552.       9.96   0 P
Variance[ 15]           231      0   133562.      133562.      10.14   0 P
Variance[ 16]           231      0   165576.      165576.      10.24   0 P
Variance[ 17]           231      0   147097.      147097.      10.21   0 P
Variance[ 18]           231      0   308521.      308521.      10.49   0 P
Variance[ 19]           231      0   149720.      149720.      10.35   0 P
Variance[ 20]           231      0   131562.      131562.      10.37   0 P
Variance[ 21]           231      0   96780.2      96780.2      10.49   0 P
Variance[ 22]           231      0   359451.      359451.      10.65   0 P
Variance[ 23]           231      0   7460.04      7460.04      10.30   0 P
Variance[ 24]           231      0  0.116784E-02 0.116784E-02   0.00   0 B
```

```
Covariance/Variance/Correlation Matrix UnStructured leg(Grad,2).Gen
 0.8885E+05  0.5837      0.1652
 0.2093E+05  0.1448E+05  0.4059
  3589.       3559.       5311.
Covariance/Variance/Correlation Matrix UnStructured leg(Grad,2).Perm
 0.2616E+05  0.3635      0.3061
  8252.      0.1970E+05  0.6183
  4243.       7439.       7347.
```

Prepare a diagonal matrix with the residual variances (M1). In our example:

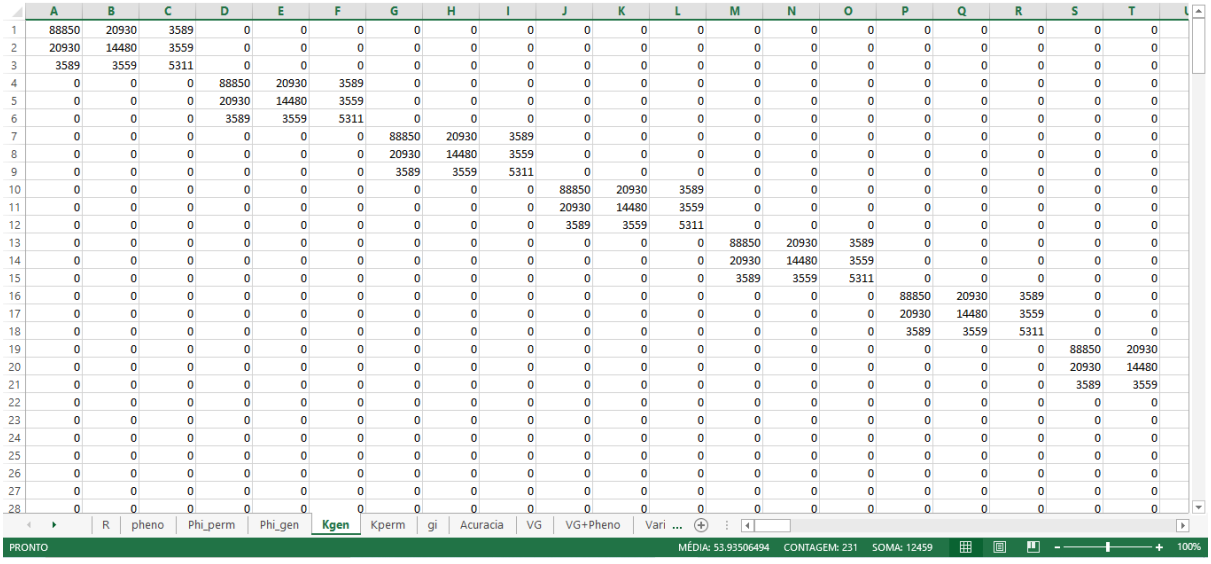


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	216805	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
2	0	133170	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
3	0	0	116210	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
4	0	0	0	97730.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
5	0	0	0	0	66504.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
6	0	0	0	0	0	126198	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
7	0	0	0	0	0	0	114606	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
8	0	0	0	0	0	0	0	82730.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
9	0	0	0	0	0	0	0	0	95213.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
10	0	0	0	0	0	0	0	0	0	130180	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
11	0	0	0	0	0	0	0	0	0	0	147569	0	0	0	0	0	0	0	0	0	0	0	0	0		
12	0	0	0	0	0	0	0	0	0	0	0	133905	0	0	0	0	0	0	0	0	0	0	0	0		
13	0	0	0	0	0	0	0	0	0	0	0	0	178845	0	0	0	0	0	0	0	0	0	0	0		
14	0	0	0	0	0	0	0	0	0	0	0	0	0	103017	0	0	0	0	0	0	0	0	0	0		
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128503	0	0	0	0	0	0	0	0	0		
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	162143	0	0	0	0	0	0	0	0		
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	142804	0	0	0	0	0	0	0		
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	301520	0	0	0	0	0	0		
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	143146	0	0	0	0	0		
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	124618	0	0	0	0		
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91256.8	0	0	0		
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	354333	0	0		
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8256.07	0		
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.78E-03		

R | pheno | Phi_perm | Phi_gen | Kgen | Kperm | gi | Acuracia | VG | VG+Pheno | Vari ...

PRONTO 80%

Prepare Kg (M2) and Kp (M3) block diagonal matrices that contain variances and covariances between Legendre coefficients. In our example, the Kg matrix would be:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	88850	20930	3589	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	20930	14480	3559	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	3589	3559	5311	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	88850	20930	3589	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	20930	14480	3559	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	3589	3559	5311	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	88850	20930	3589	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	20930	14480	3559	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	3589	3559	5311	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	88850	20930	3589	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	20930	14480	3559	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	3589	3559	5311	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	88850	20930	3589	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	20930	14480	3559	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	3589	3559	5311	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88850	20930	3589	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20930	14480	3559	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3589	3559	5311	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88850	20930
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20930	14480
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3589	3559
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

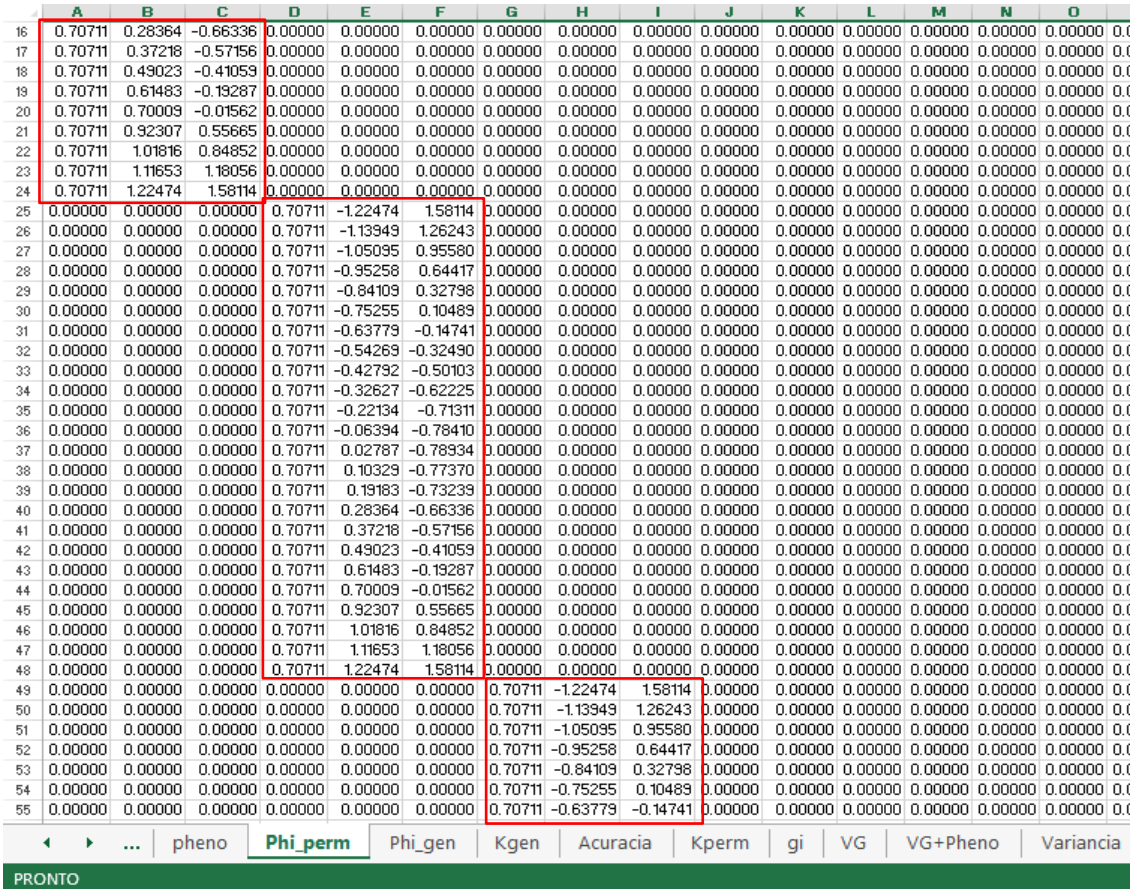
R | pheno | Phi_perm | Phi_gen | Kgen | Kperm | gi | Acuracia | VG | VG+Pheno | Vari ...

PRONTO MÉDIA: 53.93506494 CONTAGEM: 231 SOMA: 12459 100%

The .res file contains the $\phi_m(a_{ij}^*)$ scores for one individual in the 24 cuttings (shown in days).

```
leg(Grad,2)                    has          3  levels
        0.00000  0.70711  -1.22474   1.58114
       26.00000  0.70711  -1.13949   1.26243
       53.00000  0.70711  -1.05095   0.95580
       83.00000  0.70711  -0.95258   0.64417
      117.00000  0.70711  -0.84109   0.32798
      144.00000  0.70711  -0.75255   0.10489
      179.00000  0.70711  -0.63779  -0.14741
      208.00000  0.70711  -0.54269  -0.32490
      243.00000  0.70711  -0.42792  -0.50103
      274.00000  0.70711  -0.32627  -0.62225
      306.00000  0.70711  -0.22134  -0.71311
      343.35000  0.70711  -0.09886  -0.77511
      354.00000  0.70711  -0.06394  -0.78410
      382.00000  0.70711   0.02787  -0.78934
      405.00000  0.70711   0.10329  -0.77370
      432.00000  0.70711   0.19183  -0.73239
      460.00000  0.70711   0.28364  -0.66336
      487.00000  0.70711   0.37218  -0.57156
      523.00000  0.70711   0.49023  -0.41059
      561.00000  0.70711   0.61483  -0.19287
      587.00000  0.70711   0.70009  -0.01562
      624.35000  0.70711   0.82256   0.27924
      655.00000  0.70711   0.92307   0.55665
      684.00000  0.70711   1.01816   0.84852
      714.00000  0.70711   1.11653   1.18056
      747.00000  0.70711   1.22474   1.58114
```

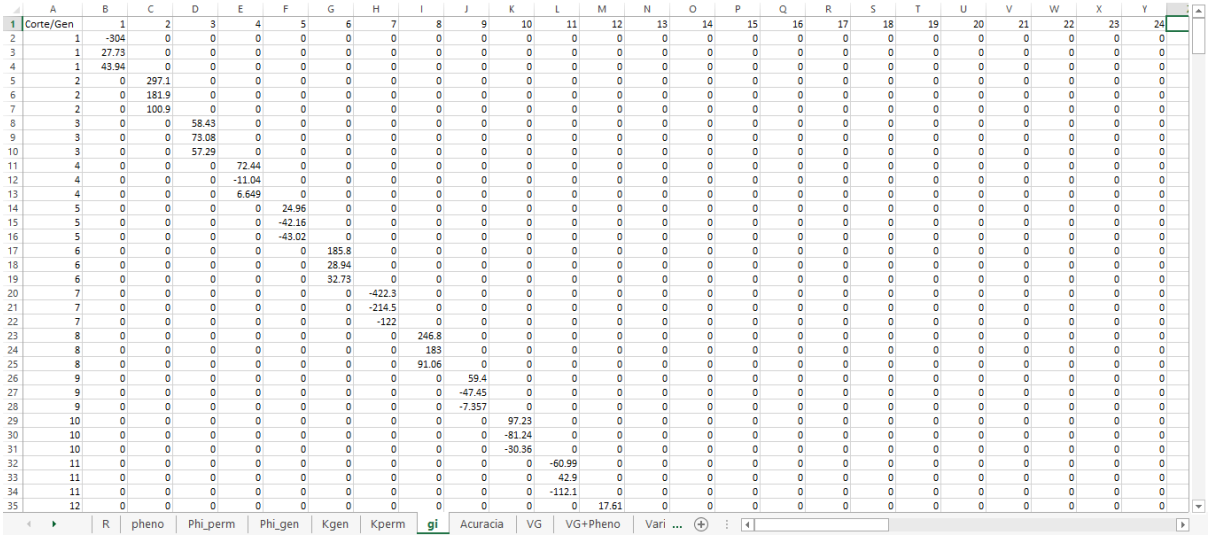
Prepare a diagonal block matrix with this information for all accessions (M5). In our example, the new matrix would have 1,848 lines (77 accessions x 24 harvests) and 231 columns (3 levels x 77 accessions).



The genetic values in the Legendre scale are in the file .sln

mu	1	2241.	60.39
leg(Grad,2).Gen	1.001	-341.9	106.7
leg(Grad,2).Gen	1.002	335.9	106.7
leg(Grad,2).Gen	1.003	61.91	106.7
leg(Grad,2).Gen	1.004	80.61	106.7
leg(Grad,2).Gen	1.005	29.27	106.7
leg(Grad,2).Gen	1.006	208.7	106.7
leg(Grad,2).Gen	1.007	-479.4	106.7
leg(Grad,2).Gen	1.008	282.2	106.7
leg(Grad,2).Gen	1.009	67.89	106.7
leg(Grad,2).Gen	1.010	105.6	106.7
leg(Grad,2).Gen	1.011	-67.46	106.7
leg(Grad,2).Gen	1.012	20.58	106.7
leg(Grad,2).Gen	1.013	135.7	106.7
leg(Grad,2).Gen	1.014	246.9	106.7
leg(Grad,2).Gen	1.015	-708.5	106.7
leg(Grad,2).Gen	1.016	319.7	106.7
leg(Grad,2).Gen	1.017	-40.01	106.7
leg(Grad,2).Gen	1.018	-241.9	106.7
leg(Grad,2).Gen	1.019	-20.06	106.7
leg(Grad,2).Gen	1.020	119.4	106.7
leg(Grad,2).Gen	1.021	423.4	106.7
leg(Grad,2).Gen	1.022	213.1	106.7
leg(Grad,2).Gen	1.023	-861.0	106.7
leg(Grad,2).Gen	1.024	55.82	106.7
leg(Grad,2).Gen	1.025	101.9	106.7
leg(Grad,2).Gen	1.026	236.6	106.7
leg(Grad,2).Gen	1.027	-126.3	106.7
leg(Grad,2).Gen	1.028	-20.22	106.7
leg(Grad,2).Gen	1.029	164.7	106.7
leg(Grad,2).Gen	1.030	30.95	106.7
leg(Grad,2).Gen	1.031	11.12	106.7
leg(Grad,2).Gen	1.032	-380.9	106.7

Prepare a block diagonal matrix for all accessions (M6). In our example, the new matrix would have 231 (3 levels x 77 accessions) and 77 columns.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Corte/Gen	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	1	-304	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	27.73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	43.94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	2	0	297.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	2	0	181.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	2	0	100.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	3	0	0	58.43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	3	0	0	73.08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3	0	0	57.29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	4	0	0	0	72.44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	4	0	0	0	-11.04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	4	0	0	0	6.649	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	5	0	0	0	0	24.96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	5	0	0	0	0	-42.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	5	0	0	0	0	-43.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	6	0	0	0	0	0	185.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	6	0	0	0	0	0	28.94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	6	0	0	0	0	0	32.73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	7	0	0	0	0	0	0	-422.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	7	0	0	0	0	0	0	-214.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	7	0	0	0	0	0	0	-122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	8	0	0	0	0	0	0	0	246.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	8	0	0	0	0	0	0	0	183	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	8	0	0	0	0	0	0	0	91.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	9	0	0	0	0	0	0	0	0	59.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	9	0	0	0	0	0	0	0	0	-47.45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	9	0	0	0	0	0	0	0	0	-7.357	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	10	0	0	0	0	0	0	0	0	0	97.23	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	10	0	0	0	0	0	0	0	0	0	-81.24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	10	0	0	0	0	0	0	0	0	0	-30.36	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	11	0	0	0	0	0	0	0	0	0	0	-60.99	0	0	0	0	0	0	0	0	0	0	0	0	0
33	11	0	0	0	0	0	0	0	0	0	0	42.9	0	0	0	0	0	0	0	0	0	0	0	0	0
34	11	0	0	0	0	0	0	0	0	0	0	-112.1	0	0	0	0	0	0	0	0	0	0	0	0	0
35	12	0	0	0	0	0	0	0	0	0	0	0	17.61	0	0	0	0	0	0	0	0	0	0	0	0

R | pheno | Phi_perm | Phi_gen | Kgen | Kperm | gi | Acuracia | VG | VG+Pheno | Vari ...

Now we are finally ready to transform the scores into real scale. Use the R codes bellow to complete this step

```
#########################################################################
setwd("C:\\Users\\Iara\\OneDrive\\Doutorado")
# Estimated genetic values for each genotype in each harvest
phi = read.table("Phi_g.txt", h=F) #M5
gen = read.table("g.txt", h=F) #M6
phi_g = as.matrix(phi)
g_i = as.matrix(gen)
dim(phi)
dim(gen)
class(g_i)
class(phi_g)
VG = phi_g %*% g_i
head(VG)
tail(VG)
VG_vector = VG[VG!=0] #creates a vector putting aside null elements
#cutting = c(1:24)
#final_data = cbind(cutting,VG_vector)
matrix_data = matrix(VG_vector, nrow=24,ncol=77)
matrix_data
ts.plot(dados_matriz, ylab="Persistence scores", xlab="Harvests",col=rainbow(4))
#########################################################################
# Genetic Variances for each harvest
phi = read.table("Phi_g.txt", h=F) #M5
kgen = read.table("Kgen.txt", h=F) #M2
phi_g = as.matrix(phi)
k_gen = as.matrix(kgen)
sigma2= phi_g %*% k_gen%*%t(phi_g)
sigma2_resumida = sigma2[1:24,1:24]
sigma2_vector = diag(sigma2_resumida) ##creates a vector putting aside null elements
sigma2_vector
ts.plot(sigma2_vector, ylab="Variance scores", xlab="Harvests",col=rainbow(4))
```

```
#####################################################################
# Permanent Variances for each harvest

phi = read.table("Phi_p.txt", h=F) #M5
kperm = read.table("Kperm.txt", h=F) #M3
phi_p = as.matrix(phi)
k_perm = as.matrix(kperm)
sigma2= phi_p %*% k_perm%*%t(phi_p)
sigma2_resumida = sigma2[1:24,1:24]
sigma2_vector = diag(sigma2_resumida) ##creates a vector putting aside null elements
sigma2_vector
ts.plot(sigma2_vector, ylab="Variance scores", xlab="Harvests",col=rainbow(4))
#####################################################################
#Accuracy

rm(list=ls())
r <- read.table("C:\\Users\Iara\OneDrive\Doutorado\Asreml\\dados.txt", header = TRUE)
head(r)


names(r)
r <- na.omit(r)
##> Convert numeric vectors into factor
r$med <- factor(r$Med)
r$gen <- factor(r$Gen)
r$medrep <- factor(r$MedRep) # Number of environments
r$grad <- factor(r$Grad)
r$perm <- factor(r$Perm)

install.packages("orthopolynom")
require(orthopolynom)
require(Matrix)

NB <- 24  # Number of harvests
# Polynomial order
OPgen <- 2
OPperm <- 2
# Polynomial degree
```

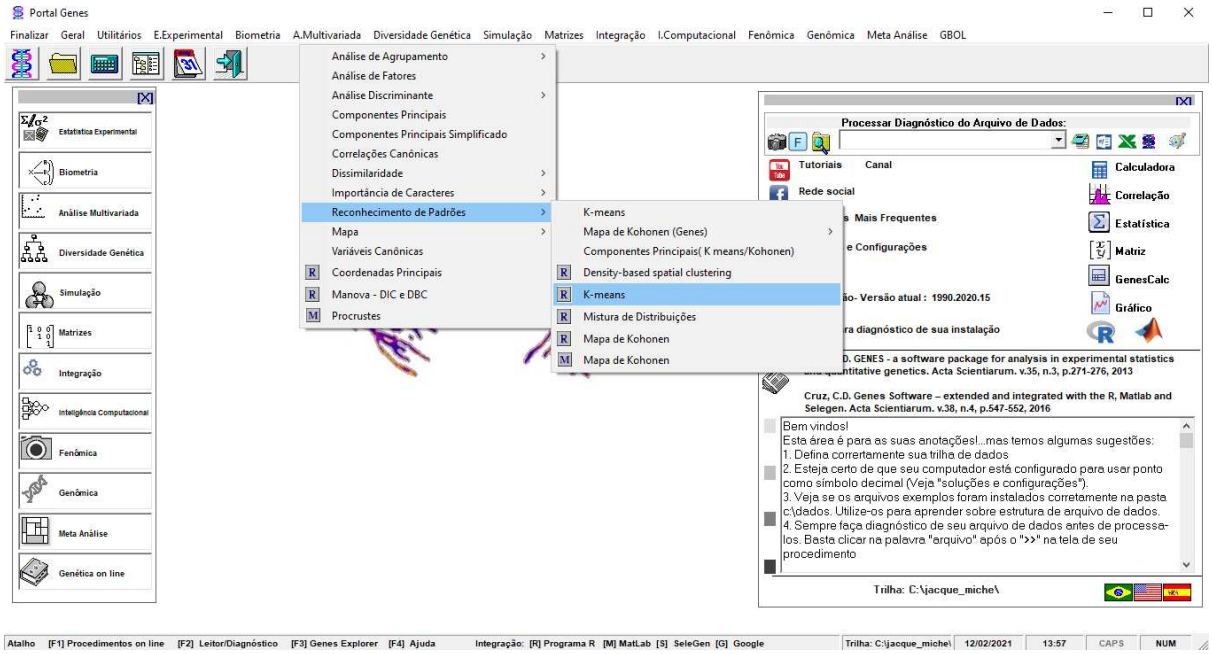
```
GPgen <- 3
GPperm <- 3
EF <- 72 # Harvests x rep (fixed effects)
# Random effects (accessions * 2)
EAgen <- nlevels(r$gen)*GPgen
EAperm <- nlevels(r$perm)*GPperm
# Number of accessions, permanent environment
NG <- nlevels(r$gen)
NPerm <- nlevels(r$perm)

# Kg and Kperm matrices
Kg=matrix(c(88850,20930,3589,20930,14480,3559,3589,3559,5311),GPgen,GPgen)
Kg

Kperm=matrix(c(26160,8252,4243,8252,19700,7439,4243,7439,7347),GPperm,GPperm)
Kperm

#Environment gradient (days of each harvest, considering harvest 1 in the day 0)
timei=c(0,26,53,83,117,144,179,208,243,274,306,354,382,405,432,460,487,523,561,587,655,
        684,714,747)

#Residual variance components (the ones on M1)
ve = matrix(diag(c(rep(216805,table(r$grad)[1]),rep(133170,table(r$grad)[2]),
rep(116210,table(r$grad)[3]),rep(97730.2,table(r$grad)[4]),rep(66504.7,table(r$grad)[5]),rep(
        126198,table(r$grad)[6]),rep(114606,table(r$grad)[7]),rep(82730.1,table(r$grad)[8]),r
        ep(95213.4,table(r$grad)[9]),rep(130180,table(r$grad)[10]),rep(14756,table(r$grad)[1
        1]),rep(133905,table(r$grad)[12]),rep(178845,table(r$grad)[13]),rep(103017,table(r$g
        rad)[14]),rep(128503,table(r$grad)[15]),rep(162143,table(r$grad)[16]),rep(142804,tab
        le(r$grad)[17]),rep(301520,table(r$grad)[18]),rep(143146,table(r$grad)[19]),rep(1246
        18,table(r$grad)[20]),rep(91256.8,table(r$grad)[21]),rep(354333,table(r$grad)[22]),re
        p(8256.07,table(r$grad)[23]),rep(1.78E-
        03,table(r$grad)[24]))),sum(table(r$grad)),sum(table(r$grad)))
ve

#Analyses
##> Make Phi matrix (Orthogonal Polynomials)
time <- r[,"Grad"]
```

```
min.time <- min(time, na.rm = TRUE)
max.time <- max(time, na.rm = TRUE)
qi <- (2*(time - min.time)/(max.time - min.time))-1

#Polinomyals
leg1 <- legendre.polynomials(OPgen, norm = TRUE)
phi1 <- sapply(leg1, predict, qi)

leg2 <- legendre.polynomials(OPperm, norm = TRUE)
phi2 <- sapply(leg2, predict, qi)

# Make the incidence matrix
y=cbind(r,local_rep=paste(r$local,r$medrep, sep = " . "))
X <- sparse.model.matrix(~ -1+as.factor(local_rep),y)
Z1 <- sparse.model.matrix(as.formula(paste("~ 0 + phi1:","gen",sep="")),r)
Z2 <- sparse.model.matrix(as.formula(paste("~ 0 + phi2:","perm",sep="")),r)
y <- as.matrix(r$DMY) ##### Mudar o nome da varável aqui #########

W <- cbind(X,Z1,Z2)
W <- as(W,"dgCMatrix")

Ig <- diag(nlevels(r$gen))
Iperm <- diag(nlevels(r$perm))

# Values
v=diag(ve)
R=1/v

K1 <- Ig%x%solve(Kg)
K1 <- as(K1,"dgCMatrix")

K2 <- Iperm%x%solve(Kperm)
K2 <- as(K2,"dgCMatrix")

VcovX <- diag(ncol(X))*0
```

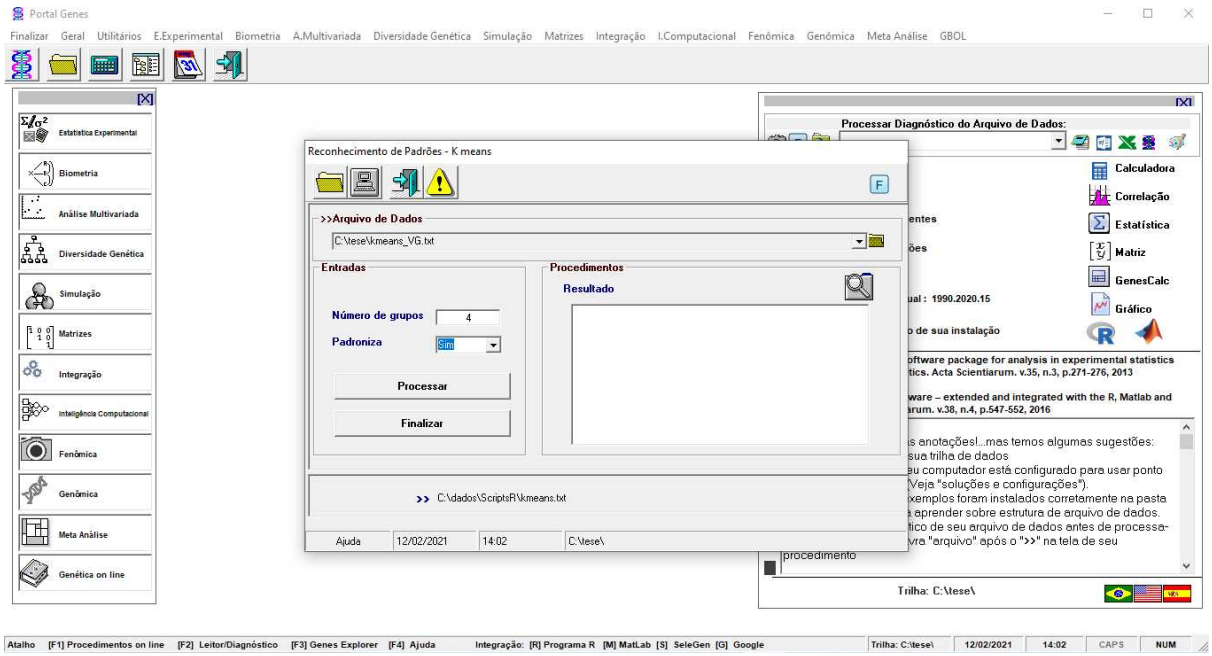
```
K <- bdiag(VcovX,K1,K2)

Rinv=matrix(diag(R),sum(table(r$grad)),sum(table(r$grad)))
Rinv <- as(Rinv,"dgCMatrix")

##> Solution Mixed Model Equations
LHS <- t(W)%*%Rinv%*%W + K
RHS <- t(W)%*%Rinv%*%y
Cinv <- solve(LHS)
sol <- Cinv%*%RHS
sol

#matrix(sol[beggining of the random effects:final of the vaues,], number of accessions,
        polynomial degree)

coef=matrix(sol[(EF+1):(EF+EAgen),], NG, GPgen, byrow=T)
coef

# Polynomials
min.timei <- min(timei, na.rm = TRUE)
max.timei <- max(timei, na.rm = TRUE)

qii <- (2*(timei - min.timei)/(max.timei - min.timei))-1

#legendre.polynomials(pol degree)
leg <- legendre.polynomials(GPgen-1, norm = TRUE)
phi <- sapply(leg, predict, qii)

PHg=phi

# Li e Ls
Cii = list()
PEV = list()

#seq(number of fixed effects + 1, number of accessions x polynomial degree + number of
        fixed effects, pol degree)
```

```
li = seq(EF+1, (NG*GPgen)+EF, GPgen)

#seq(number of fixed effects + ordem do polinomio, number of accessions x polynomial
        degree + number of fixed effects, pol degree)
ls = seq(EF+GPgen, (NG*GPgen)+EF, GPgen)

for(i in 1:NG){
  Cii[[i]] = PHg%*%Cinv[li[i]:ls[i],li[i]:ls[i]]%*%t(PHg)
  PEV[[i]] = diag(Cii[[i]])
}

#matrix(0, number of accessions x number of cuttings, replications)
C22 = matrix(0,NG*NB,NG*NB)
diag(C22) = unlist(PEV)

#C22
Comp.var.gen=PHg%*%Kg%*%t(PHg)
Comp.var.gen

acc=sqrt(1-diag(C22)/rep(diag(as.matrix(Comp.var.gen)),NG))
dados_acc=matrix(acc,NG,NB,byrow=T)
dados_acc
write.table(accuracy "CC:\\Users\Iara\OneDrive\Doutorado\Asreml\\dados_acc.txt", sep="\t")
```

A file containing the genetic values for all accessions in all harvests should be settled up for the K-means clustering technique in the software GENES. In our example, the .txt file has 77 lines x 24 columns.

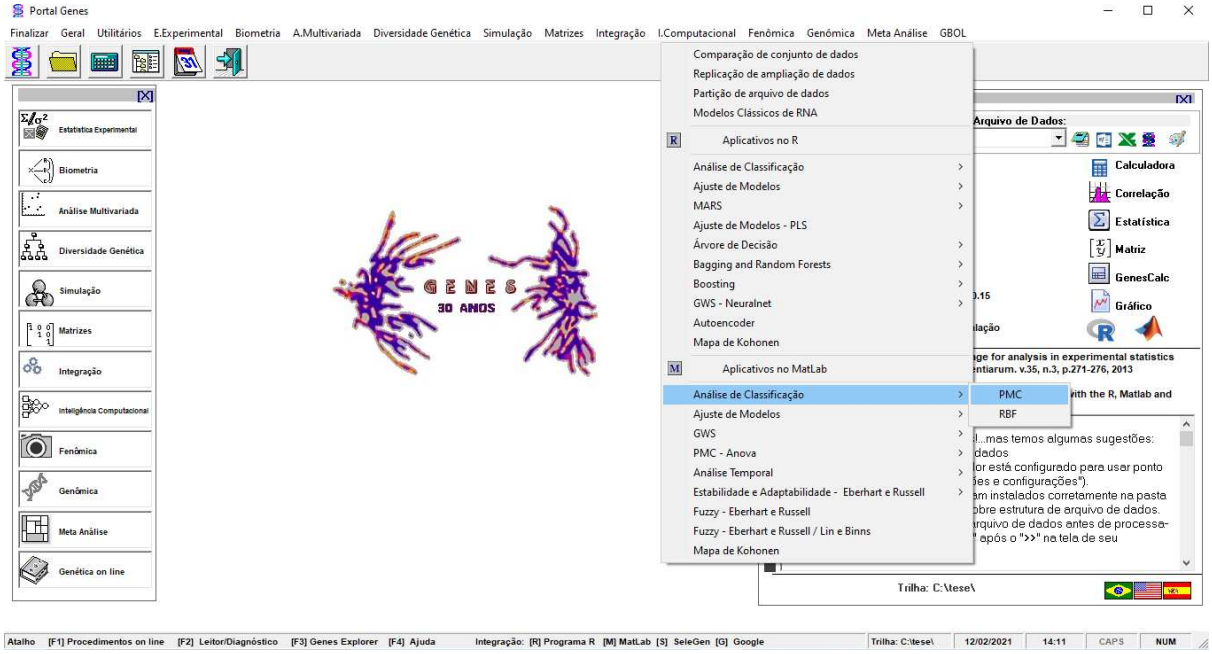


Choose the appropriate number of clusters (the number you consider reasonable for each situation). Standardize the data and click in `Processar`.

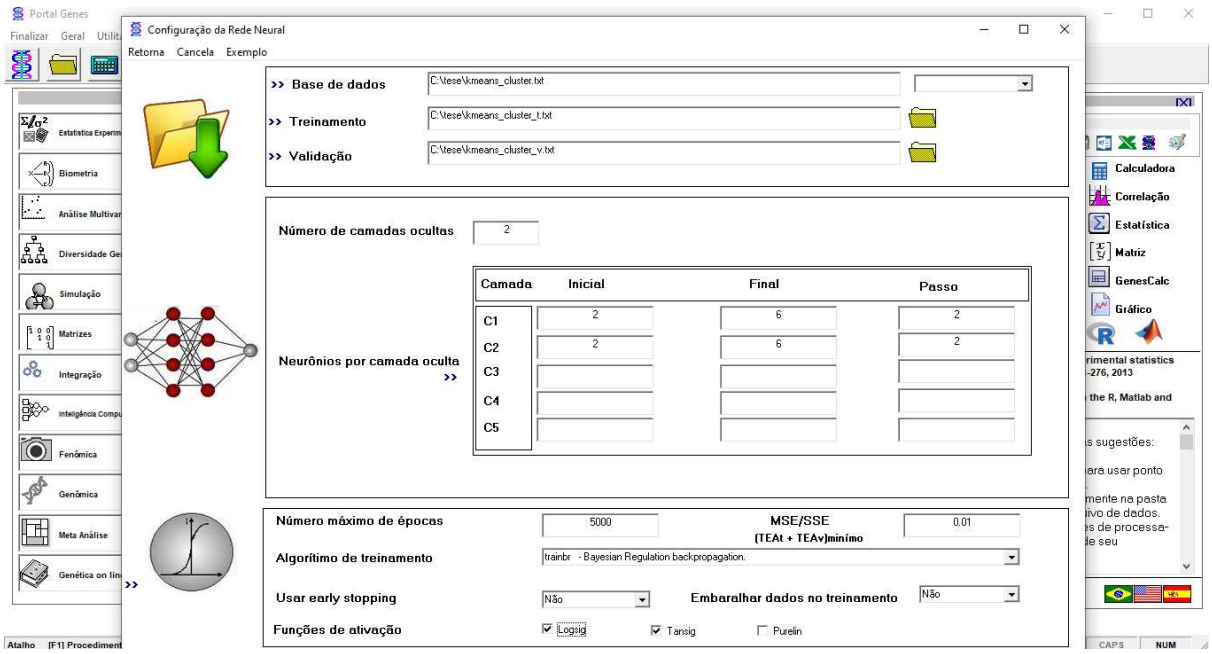


The software Genes will be used to establish the Artificial Neural Network, as follows.

The output of the K-means method contains the accessions' classification in their respective persistence clusters. Prepare a file with that classification along with the genetic values obtained for each accession in each cutting. Go to the software Genes.



Beware to choose the correct parameters and proceed with the analysis.



Click on 'Retorna' and then in 'Gerar rede'. The next time you use new data on this same network, you will click on 'Usar rede para teste'. The network will classify your new data with the same efficiency shown when you generate it.