UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
EMBRAPA MANDIOCA E FRUTICULTURA
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS
CURSO DE MESTRADO

SELEÇÃO DE DESCRITORES E ANÁLISE DE AGRUPAMENTO EM ACESSOS DE TABACO

ANTONIO LEANDRO DA SILVA CONCEIÇÃO

CRUZ DAS ALMAS - BAHIA
MARÇO – 2015

SELEÇÃO DE DESCRITORES E ANÁLISE DE AGRUPAMENTO EM ACESSOS DE TABACO

ANTONIO LEANDRO DA SILVA CONCEIÇÃO

Engenheiro Agrônomo
Universidade Federal do Recôncavo da Bahia, 2013

Dissertação submetida ao Colegiado de Curso do Programa de Pós-Graduação em Recursos Genéticos Vegetais da Universidade Federal do Recôncavo da Bahia e Embrapa Mandioca e Fruticultura, como requisito parcial para obtenção do Grau de Mestre em Recursos Genéticos Vegetais.

Orientador: Prof. Dr. Carlos Alberto da Silva Ledo

Coorientador: Prof. Dr. Ricardo Franco Cunha Moreira

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
EMBRAPA MANDIOCA E FRUTICULTURA
MESTRADO EM RECURSOS GENÉTICOS VEGETAIS
CRUZ DAS ALMAS - BAHIA – 2015

FICHA CATALOGRÁFICA

C744s

Conceição, Antonio Leandro da Silva.
Seleção de descritores e análise de agrupamento em acessos de tabaco / Antonio Leandro da Silva Conceição._ Cruz das Almas, BA, 2015.
111f.; il.

Orientador: Carlos Alberto da Silva Ledo.
Coorientador: Ricardo Franco Cunha Moreira.

Dissertação (Mestrado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas.

1.Fumo – Cultivo. 2.Fumo – Recursos genéticos vegetais. 3.Diversidade genética – Descritores. 4.Germoplasma vegetal – Análise. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Lopes, Everaldo Antônio. III.Título.

CDD: 633.71

Ficha elaborada pela Biblioteca Universitária de Cruz das Almas - UFRB.

COMISSÃO ORGANIZADORA

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLOGICAS
EMBRAPA MANDIOCA E FRUTICULTURA
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS

COMISSÃO EXAMINADORA DA DEFESA DE DISSERTAÇÃO DE ANTONIO LEANDRO DA SILVA CONCEIÇÃO

Orientador: Prof. Dr. Carlos Alberto da Silva Ledo

Embrapa Mandioca e Fruticultura

Prof. Dr.ª Eva Maria Rodrigues Costa

Universidade Federal do Recôncavo da Bahia

Prof. Dr.ª Simone Alves Silva

Universidade Federal do Recôncavo da Bahia

Dissertação homologada pelo Colegiado do Curso de Mestrado em Recursos Genéticos Vegetais em....., Conferindo o Grau de Mestre em Recursos Genéticos Vegetais em.....

DEDICATÓRIA

“Dedico esta Dissertação primeiramente a Deus, por sempre está ao meu lado me dando forças”.

Aos meus pais Sinésio e Vera,

Por terem me dado a vida e me ensinarem bons valores.

As minhas avós, Maria e Araci (Nenzinha),

Pelo amor e carinho que sempre semearam na família.

“Cada sonho que você deixa para trás, é um pedaço do seu futuro que deixa de existir”.

(Steve Jobs)

AGRADECIMENTOS

A Deus, Obrigado pela tua grandeza, pelo seu amor incondicional. Obrigado pelo carinho, pelo cuidado com minha família, por nunca desistir de mim, por me amparar em todos os momentos.

A Universidade Federal do Recôncavo da Bahia pela oportunidade de realização deste curso.

A Fundação de Apoio à Pesquisa do Estado da Bahia (FAPESB), pelo suporte financeiro dado a este estudo.

Aos professores do Programa de Pós-Graduação em Recursos Genéticos Vegetais, pelos ensinamentos transmitidos e amigável convivência.

Ao meu orientador Prof. Dr. Carlos Alberto da Silva Ledo pela orientação na minha dissertação, pelos conhecimentos transmitidos e imprescindível contribuição na realização desse trabalho.

Ao meu co-orientador Prof. Dr. Ricardo Franco Cunha Moreira pela atenção, disponibilidade e pelos conhecimentos transmitidos.

Ao Prof. Dr. Cosme Damião Cruz e a Doutoranda Isabela de Castro Sant'Anna da Universidade Federal de Viçosa, pela atenção, disponibilidade e contribuição na realização desse trabalho.

Aos Professores Endrigo Sampaio, Ronaldo Fiúza e Silvia Patrícia pelo incentivo e pelos ensinamentos transmitidos e amigável convivência.

A todos os meus amigos do curso de Pós-graduação em Recursos Genéticos Vegetais e demais amigos dos cursos de Pós-graduação da UFRB que tive o prazer de conhecer, em especial a Sandra Afonso, Von Daniken, Thâmara Lima, Any, Lorena, Juliana, Tamara Leal e Raysa, amigos estes de grande importância no meu crescimento profissional e pessoal.

Aos meus amigos, Clailto e Maurício pelo auxilio e contribuição para a realização desse trabalho.

A empresa Ermor Tabarama Tabacos do Brasil Ltda, pela parceria e infraestrutura e aos seus funcionários pelo auxilio que na realização desse trabalho.

A Embrapa Mandioca e Fruticultura Tropical pelo apoio institucional.

Agradeço a todas as pessoas que passaram pela minha vida e de alguma forma me incentivaram a realizar esse feito.

SUMÁRIO

Página

SELEÇÃO DE DESCRITORES E ANÁLISE DE AGRUPAMENTO EM ACESSOS DE TABACO

Autor: Antonio Leandro da Silva Conceição
Orientador: Carlos Alberto da Silva Ledo
Co-orientador: Ricardo Franco Cunha Moreira

RESUMO: Objetivou-se com este estudo selecionar descritores morfoagronômicos e determinar sua importância relativa na caracterização, bem como propor um número mínimo capaz de quantificar a divergência genética existente entre acessos de tabaco tipo Sumatra e aplicar diferentes metodologias de análise de agrupamento com base na análise simultânea e individual por meio dos descritores quantitativos e qualitativos selecionados. Para a caracterização dos acessos inicialmente foram usados 43 descritores morfoagronômicos, sendo 17 quantitativos e 26 qualitativos. Foi realizada a identificação dos descritores quantitativos redundantes, por dois procedimentos: 1) seleção direta, proposta por Jolliffe e 2) Seleção baseada no coeficiente de Singh. A seleção dos descritores qualitativos foi realizada por meio do nível de entropia dos caracteres (H), proposto por Renyi. Foram selecionados 10 descritores quantitativos e 8 descritores qualitativos. Todos esses descritores selecionados são importantes na caracterização do germoplasma de tabaco em estudo. O descarte de 58% dos descritores não provocou perda de informação considerável, uma vez que os descritores redundantes estão correlacionados aos descritores remanescentes. Para a análise de divergência foram utilizados os métodos de agrupamento UPGMA e WARD, em que o método UPGMA foi o que melhor explicou a divergência genética entre os acessos em estudo. Foram utilizadas para as análises individuais as distâncias de mahalanobis e euclidiana média, para obtenção das matrizes oriundas dos dados quantitativos. Para os dados qualitativos originais e dados quantitativos transformados foi utilizada a distância de Cole-Rodgers. Para as análises simultâneas dos dados quantitativos e qualitativos foram testadas três metodologias distintas: O algoritmo de Gower; Soma algébrica de matrizes individuais e integração de dados por meio da transformação de caracteres quantitativos em multicategóricos por duas

estratégias distintas (Regra de Sturges e Regra da raiz quadrada). Os valores de correlação entre as matrizes de dados quantitativos transformados e quantitativos originais foram significativos a 1 e a 5% de probabilidade, sendo um deles de alta magnitude, oferecendo suporte para extrapolar os resultados de um conjunto de dados para outro. Com isso, a estratégia da raiz quadrada foi a mais indicada, com correlação de 0,75 e 0,82 entre as matrizes de dissimilaridade dos dados codificados e quantitativos originais. Como critério para definição do número ótimo de grupos foi usado o índice Pseudo-t^2, com este, foi possível a formação de 3 grupos pelo método UPGMA para todas as metodologias de análise simultânea utilizadas. O acesso A14 (125 PD) possui comportamento distinto dos demais e as metodologias de análise simultânea, com base nas matrizes de distâncias geradas, captaram essa divergência. Todas as matrizes de análise conjunta foram comparadas, mostrando grande correspondência entre as mesmas, com altas correlações que variaram de 0,824 a 0,998. Os resultados deste trabalho mostraram que as metodologias de análise simultânea foram eficazes em relevar a existência de divergência genética entre acessos de *Nicotiana tabacum* L. tipo Sumatra e mostram a importância da combinação de métodos, uma vez que puderam otimizar, de forma considerável, a interpretação dos resultados para maior conhecimento do germoplasma em estudo.

Palavras-chave: Divergência genética, análises simultâneas, análise multivariada.

SELECTION OF DESCRIPTORS AND CLUSTER ANALYSIS IN ACCESSIONS OF TOBACCO

Author: Antonio Leandro da Silva Conceição
Advisor: Carlos Alberto da Silva Ledo
Co-advisor: Ricardo Franco Moreira Cunha

ABSTRACT: This study aimed to select descriptors morphological and agricultural traits and determine their relative importance in the characterization, as well as propose a minimum number to quantify the genetic divergence between accessions of tobacco type Sumatra and apply different methods of cluster analysis based on the simultaneous analysis and individual by means of quantitative descriptors and qualitative selected. For the characterization of accessions were initially used 43 descriptors morphological and agricultural traits, being 17 quantitative and 26 qualitative. Was the identification of quantitative descriptors redundant, by two procedures: 1) direct selection, proposal by Jolliffe and 2) Selection based on the coefficient of Singh. The selection of qualitative descriptors was performed by means of the entropy level of the characters (H), proposed by, Renyi's series. 10 Were selected quantitative descriptors and 8 qualitative descriptors. The disposal of 58% of the descriptors has not lost considerable information, since the descriptors are correlated with the remaining descriptors. For the analysis of divergence were used methods UPGMA and WARD, in which the UPGMA method was the best explained the genetic divergence among the accessions into study. They were used for the individual analyzes the Mahalanobis distance and Euclidean average for obtaining the arrays from the quantitative data and the qualitative data and original quantitative data processed was used the distance of Cole-Rodgers. For the simultaneous analysis of quantitative and qualitative data were tested three different methodologies: The algorithm of Gower; algebraic Sum of individual arrays and data integration through the transformation of quantitative traits in multicategoric by two distinct strategies (Rule of Sturges and Rule the square root). The correlation values between the arrays of quantitative data processed and original quantitative were significant at 1 and 5% of probability, one of them being of high magnitude,

offering support to extrapolate the results of a set of data to another. With this, the strategy of square root was the most indicated, with a correlation of 0.75 and 0.82 between the arrays of dissimilarity of encoded data and quantitative originals. As a criterion for the definition of the optimal number of groups was used the index Pseudo-t2, this was possible the formation of 3 groups by UPGMA method for all the methodologies of simultaneous analysis used. Access to14 (125 PD) has distinct behavior of others and the methodologies of simultaneous analysis, based on matrices of distances generated, captured this divergence. All arrays of joint analysis were compared, showing great correspondence between them, with high correlations ranged from 0.824 to 0.998. The results of this work showed that the methodologies of simultaneous analysis were effective in identifying the existence of genetic divergence among accessions of Nicotiana tabacum L. type Sumatra and show the importance of the combination of methods, since they have been able to leverage a significant way to the interpretation of the results for greater knowledge of germplasm in study.

Key words: Genetic divergence, simultaneous analysis, multivariate analysis.

INTRODUÇÃO

O tabaco é atualmente a cultura não alimentícia mais importante da agricultura mundial. O Brasil é o segundo maior produtor mundial de tabaco e líder em exportações desde 1993, graças à qualidade e integridade do produto. Em 2013, o tabaco representou 1,35% do total das exportações brasileiras, com US$ 3,27 bilhões embarcados. Da produção de 706 mil toneladas registrada na safra 2012/13, mais de 85% foi destinada ao mercado externo. O principal mercado brasileiro neste período foi a União Europeia com 42% do total dos embarques de 2013, seguida pelo Extremo Oriente (26%), América do Norte (13%), Leste Europeu (7%), África/Oriente Médio (7%) e América Latina (5%) (SINDITABACO, 2014).

A Bahia ocupa a quinta posição do ranking no País, atrás do Rio Grande do Sul, Santa Catarina, Paraná e Alagoas. A produção de fumo na Bahia é localizada em 22 municípios e na região produtora funcionam três empresas exportadoras e oito fabricantes de charutos. A região do Recôncavo baiano agrega ótimas condições para a cultura do fumo: temperado entre 17º e 27º, boa umidade e solos arenosos ou argilosos. O método de produção é de agricultura familiar, com propriedades variando entre 0,7 e 1 hectare. Já a cadeia produtiva do charuto emprega 14 mil pessoas no Recôncavo, a maioria agricultores familiares, sendo que 90% deste são mulheres que aprenderam o ofício com suas mães e avós, repassando as filhas desde 1842 – período em que foi fundada a primeira fábrica, a Juventude, no município de São Félix. A Bahia que já chegou a produzir 240 milhões de charutos por ano, fabrica hoje entre seis milhões e oito milhões de unidades, (SEAGRI, 2014).

A Bahia tem um grande potencial agrícola para a produção de tabaco das variedades Brasil-Bahia, de coloração castanha e Sumatra, originário da

Indonésia e utilizado na confecção de capas claras para charutos. Concentrada na região do Recôncavo, especialmente no entorno do município de Cruz das Almas - BA, a cultura fumageira possui grande importância econômica e social, garantindo o sustento a milhares de pequenos produtores rurais e empregados nas empresas de beneficiamento de fumo (OLIVEIRA, 2006). O tabaco tipo Sumatra é cultivado a décadas na região do Recôncavo da Bahia, com destaque para empresa Ermor Tabarama Tabacos do Brasil Ltda, onde esse tipo de tabaco é o principal utilizado pela empresa. Dentre os principais tipos de tabaco no mundo destinados a produção de capas para charutos, encontra-se o tipo Bahia e o Sumatra (TABACCO JOURNAL INTERNATIONAL, 2000).

Outra finalidade para o tabaco que tem se tornado realidade em alguns países é seu potencial para uso medicinal, na produção de produtos biofarmacêuticos como vacinas, hormônios, anticorpos e insulina (BLINDER et al., 2007).

Na implantação de um programa de melhoramento genético, uma das principais necessidades do melhorista é o conhecimento do germoplasma disponível e a capacidade de identificar plantas que possuam genes de interesse para o programa (WEEDEN et al., 1994).

Todo caráter deve apresentar uma parcela de contribuição na variação do germoplasma analisado. Mas, há uma tendência de que o aumento do número de descritores avaliados ocasione a presença de informações redundantes, posto que essas informações quase sempre estão associadas a outras (DAHER, 1993). Logo, a eliminação de descritores redundantes seria uma decisão vantajosa, pois reduziria o trabalho de tomada de dados sem ocasionar perda na precisão da caracterização, especialmente se esses caracteres forem de difícil mensuração e apresentarem baixa variabilidade e estabilidade de expressão (PEREIRA, 1989).

Na caracterização da diversidade genética das espécies vegetais, animais e de microrganismos, os pesquisadores têm o interesse em agrupar genótipos similares, de maneira que as maiores diferenças ocorram entre os grupos formados. Neste aspecto técnicas multivariadas, como análise discriminante, componentes principais, análise de coordenadas e de agrupamento, podem ser aplicadas neste tipo de estudo. A adoção de uma, entre as técnicas citadas, varia de acordo com o padrão de resultado desejado e com a informação disponível,

seja ela característica morfológica, fisiológica, ecológica ou genético-molecular (CRUZ et. al., 2011). Dentre estas, pode ser destacada a análise de agrupamento que é muito utilizada pelos pesquisadores tanto da área de melhoramento genético vegetal quanto na caracterização morfológica de acessos, ou seja, na caracterização morfológica de coleções de constituições genéticas geralmente mantidas em bancos de germoplasma e ainda pouco conhecidas pelos melhoristas (KOOP et al., 2007).

Segundo Campbell et al. (2010) a preservação das cultivares, raças e parentes silvestres de espécies vegetais importantes, fornecem um fundamento básico para promover e sustentar a agricultura. Sendo assim, os programas de melhoramento vegetal têm investido sobre os recursos genéticos e coleções de germoplasma para desenvolvimento de genótipos melhorados com ganhos significativos na produtividade. Portanto, o estabelecimento de processos voltados para conservação dos recursos genéticos é uma necessidade que tem sido abordada mundialmente. O papel fundamental dos bancos de germoplasma é a manutenção e preservação da variabilidade genética (RAMALHO et al., 2008).

Em determinados casos, como nos estudos de diversidade genética, torna-se necessário à avaliação conjunta de diversas variáveis agronômicas, morfológicas e moleculares. A análise individual para cada tipo de variável pode levar a discrepâncias em relação aos agrupamentos e às inferências em relação à quantificação da variabilidade entre as unidades experimentais ou amostrais a serem agrupadas. Com isso, a analises com métodos que considerem simultaneamente os diversos tipos de variáveis é preferível (LEDO e GONÇALVES, 2012).

O presente trabalho teve por objetivo selecionar descritores morfoagronômicos de acessos de *Nicotiana tabacum* L., e determinar sua importância relativa na caracterização, bem como propor um número mínimo capaz de quantificar a diversidade entre esses acessos, como também, aplicar diferentes estratégias de agrupamento com mistura de variáveis quantitativas e qualitativas e identificar a combinação mais adequada para maior conhecimento do germoplasma em estudo, visando sua conservação e contribuição para o melhoramento genético da espécie.

REVISÃO DE LITERATURA

Origem do tabaco (*Nicotiana tabacum* L.)

O tabaco ou fumo, como é popularmente conhecido, é cultivado há centenas de anos pelo homem. Entretanto, existem duas correntes sobre a difusão da fumicultura pelo mundo, uma delas afirma que o fumo é originário das Américas e a outra afirma que descende de plantas utilizadas como fumo na Ásia desde o século IX, provavelmente em cachimbos. Atualmente, admite-se que a planta é originária dos vales orientais dos Andes bolivianos, difundindo-se pelo território brasileiro através das migrações indígenas, sobretudo da nação Tupi-Guarani (SINDIFUMO, 2007).

O fumo é uma planta anual, autógama, cultivada com fim comercial, com ciclo de vida variando entre 120 a 240 dias. O gênero *Nicotiana* tem cerca de 60 espécies conhecidas, originárias da América do Sul, América do Norte, Austrália e Ilhas do Pacífico Sul. O gênero *Nicotiana* pertence à família Solanaceae e está dividido em três subgêneros *Rustica, Tabacum* e *Petunioides* (GOODSPEED, 1954; GERSTEL,1979; NARAYAN, 1987).

O fumo, devido ao alcalóide nicotina era empregado pelos índios em rituais religiosos e também com fins medicinais. Atualmente, *N. rustica* tem sido utilizada também como fonte de nicotina para produção de inseticida e como fonte de ácido cítrico. Outras espécies como *N. alata, N. sandarae* e *N. glauca* são ornamentais. *N. tabacum* é entre as espécies a mais importante na agricultura atual e no mercado internacional (COLLINS & HAWKS, 1993).

No Brasil são plantados os tipos de fumo Virgínia (81%), Burley (17%), Comum (0,8%) e outros (1,2%), onde se encontram os fumos para capa de charuto, oriental e fumo em corda. Na fabricação do cigarro são usados 40% de fumo Virgínia, 35% de fumo Burley, 15% de fumo Oriental e 10% de talo picado. A mistura destes tipos de fumo na composição do cigarro produz perfeito equilíbrio no sabor e aroma, atendendo a exigências do mercado consumidor (KIST et al., 2004).

Os tipos de tabaco cultivados no Brasil são classificados de acordo com a finalidade de uso e o método de cura. São eles os fumos tipo estufa, galpão,

oriental e outros pequenos grupos. O processo de cura é muito importante para garantir ao produto aroma e boa qualidade química. Neste processo, a clorofila é degradada e os carboidratos são convertidos a açúcares simples. O tabaco do tipo estufa compreendem os grupos varietais Virgínia e Amarelinho. Possuem colheita de folhas individual e cura através de calor artificial em estufas apropriadas. São empregados para misturas na fabricação de cigarros industrializados e possuem alto teor de açúcares. Os do tipo galpão compreendem os grupos varietais Burley, Comum, Dark e Maryland. A colheita é feita pelo corte da planta inteira e a cura é realizada em galpões sem utilização de calor artificial. Estes grupos também são utilizados em misturas na fabricação de cigarros industrializados. Os fumos do tipo oriental compreendem os grupos varietais Izmir, Basma e Gavurkoy. Possuem folhas pequenas e característica marcante pelo forte aroma, razão pela qual são designados fumos tipo *flavor*, importantes na mistura para fabricação de cigarros industrializados devido ao aroma característico e baixos teores de nicotina (MASSOLA et al., 2005).

Descrição e Caracterização da Espécie

O gênero *Nicotiana* L. pertence à família Solanaceae, subfamília Cestroideae, tribo Nicotianeae e subtribo Nicotianinae. Nicotianeae é uma das oito tribos que compõem a subfamília Cestroideae, sendo formada por três subtribos (Nierembergiinae, Nicotianinae e Leptoglossinae) e oito gêneros, segundo Hunziker (2001), ou nove, segundo D'arcy (1991).

Segundo Vieira et al., (2003), o gênero *Nicotiana* L. abriga uma série diversificada de espécies, sendo algumas tóxicas, outras ornamentais e até mesmo espécies possuidoras de substâncias inseticidas (anabasina, nicotina e a nornicotina). Já *N. tabacum* L., é amplamente conhecida por sua importância econômica, como fonte de matéria-prima para a indústria do fumo, por suas propriedades estimulantes e por serem muito utilizadas em investigações científicas nas áreas de farmácia, fisiologia, virologia e plantas transgênicas (GOODSPEED, 1954; HAWKES, 1999; HUNZIKER, 2001).

É uma planta anual, autógama, mas que apresenta um baixo percentual de alogamia (inferior a 3%). Cultivada com fim comercial, com ciclo de vida variando

entre 120 a 240 dias. Trata-se de uma planta herbácea que concentra o alcaloide nicotiana, com folhas grandes, que amadurecem da base para cima, sendo que nas espécies e variedades de maior porte, as folhas basais podem chegar aos 70-75 cm de comprimento. Podem atingir de 90-180 cm de altura. As flores, que aparecem no topo, acima das folhas menores e mais jovens, apresentam cores variáveis (branco, púrpura, rosa, vermelho). São tubulares e possuem tanto os órgãos masculinos e os femininos, podendo tanto se autofecundar ou serem fecundadas pelo pólen de outras plantas do mesmo gênero (HUNZIKER, 2001).

De todas as espécies Nicotiana, a *N. tabacum* é a mais plantada no mundo, é cultivada para a produção de folhas, que são transformadas em formas que possam ser fumadas, mastigadas, e inaladas. O gênero é nomeado em homenagem a Jean Nicot, que em 1561 foi o primeiro a apresentar o tabaco para o francês Royal Court. Nicotiana pertence à família das solanáceas e subdivide-se em três subgenêros: Rustica, Tabaccum e Petunioides (GOODSPEED, 1954; REN & TIMKO, 2001).

Desde o século XIX *Nicotiana* spp. se constituem em importante material para estudo genético devido à facilidade de manipulação das flores e o grande número de sementes produzido (GERSTEL, 1979). A maioria dos fumos cultivados pertence à *N. tabacum* L., um alotetraplóide que apresenta 2n = 4x = 48 cromossomos, distribuídos nos genomas S e T. A espécie *N. tabacum* se originou da hibridização de duas espécies diplóides, *N. sylvestris* (2n=24) com genoma S(maternal), e *N. tomemtosiformis* (2n=24), que apresenta o genoma T (parental) (GERSTEL, 1979; BLAND et al., 1985; OKAMURO & GOLDBERG, 1985; SPERISEN et al., 1991; COLLINS & HAWKS, 1993, SANTOS, 2002). A hibridização natural entre estas espécies de *Nicotiana* ocorreu provavelmente no Nordeste da Argentina ou Sudeste da Bolívia, porque é a região onde as duas espécies convivem na natureza (COLLINS & HAWKS, 1993).

Para a obtenção de folhas bem estruturadas e finas, os fumos para capa não devem receber diretamente muita luz solar, o que se consegue com a utilização de telas com diferentes taxas de sombreamento, permitindo uma redução superior a 40% na luz solar. A luz incidente no interior da área coberta é difusa, atingindo as folhas de fumo mais uniformemente. O sombreamento, ou seja, o cultivo sob telado, proporciona um microclima favorável ao desenvolvimento das plantas de

fumo para a produção de capa (WEHLBURG, 1999). O objetivo desta técnica é proteger a cultura de luminosidade intensa, reduzir o movimento do ar e a evaporação. Esta prática surgiu no Estado de Connecticut (Estados Unidos) como forma de imitar as condições climáticas naturais vigentes em Sumatra (Indonésia); espalhando-se então para a América Central, Caribe e Brasil, especificamente no Estado da Bahia (TOBACCO JOURNAL INTERNATIONAL, 2000).

Importância da Cultura

O tabaco é atualmente a cultura não alimentícia mais importante da agricultura mundial. O Brasil é o segundo maior produtor mundial de tabaco e líder em exportações desde 1993, graças à qualidade e integridade do produto. Em 2014, foram embarcadas 476 mil toneladas do produto, gerando divisas de US$ 2,5 bilhões; 13 países deixaram de embarcar o produto brasileiro e sete novos países passaram a integrar a lista de importadores, totalizando 96 países que levaram o tabaco produzido por mais de 162 mil produtores brasileiros. Apesar da Bélgica e Holanda terem reduzido suas compras em cerca de 30% em comparação com 2013, a União Europeia continua sendo o principal destino do tabaco brasileiro (42%), seguida pelo Extremo Oriente (28%). A China também acompanhou o ano de redução de embarques (-27%) comparado com o ano anterior, mas a principal queda registrada foi para os Estados Unidos (-42%), (SINDITABACO, 2014).

A produção nacional se concentra fortemente no Sul do País, ocupando os Estados do Paraná, Santa Catarina e Rio Grande do Sul. Com destaque para a cidade de Santa Cruz do Sul (RS) onde estão localizadas as sedes das maiores empresas de beneficiamento do tabaco. (ANUARIO, 2011).

A Bahia ocupa a quinta posição do ranking no país, atrás do Rio Grande do Sul, Santa Catarina, Paraná e Alagoas. A produção de fumo na Bahia é localizada em 22 municípios e na região produtora do Recôncavo Baiano funcionam três empresas exportadoras e oito fabricantes de charutos. A região do Recôncavo Baiano agrega ótimas condições para a cultura do fumo. O método de produção é de agricultura familiar, com propriedades variando entre 0,7 e 1 hectare. Já a cadeia produtiva do charuto emprega 14 mil pessoas no Recôncavo, a maioria

agricultores familiares, sendo que 90% deste são mulheres que aprenderam o ofício com suas mães e avós, repassando as filhas desde 1842 – período em que foi fundada a primeira fábrica, a Juventude, no município de São Félix. A Bahia que já chegou a produzir 240 milhões de charutos por ano, fabrica hoje entre seis milhões e oito milhões de unidades (95% do fabrico do país), (SEAGRI, 2014).

Concentrada na região do Recôncavo, especialmente no entorno do município de Cruz das Almas - BA, a cultura fumageira possui grande importância econômica e social, garantindo o sustento a milhares de pequenos produtores rurais e empregados nas empresas de beneficiamento de fumo (OLIVEIRA, 2006). Por seu impacto social, a cultura do tabaco é considerada um fator de promoção humana e de manutenção do homem no campo, gerando mais de 2 milhões de empregos diretos e indiretos no Brasil (AFUBRA, 2010).

Além do tabaco ser destinado a indústria fumageira na produção de cigarros, charutos e etc. O tabaco tem se tornado realidade em alguns países por seu potencial para uso medicinal, na produção de produtos biofarmacêuticos como vacinas, hormônios, anticorpos e insulina (BLINDER et al., 2007).

O tabaco também é importante nas pesquisas envolvendo a tecnologia do DNA recombinante, mutação induzida e outras. Além disso, muitos trabalhos importantes de genética quantitativa foram realizados utilizando plantas do gênero *Nicotiana*, devido à morfologia floral que facilita a obtenção de autofecundações e cruzamentos, e uma grande quantidade de sementes por fruto (FARIAS, 2007). A obtenção de proteínas virais animais em células vegetais é um caminho promissor. É possível, por exemplo, obter antígenos de hepatite B e raiva em plantas do tabaco (KOPROWSKI e YUSIBOV, 2001; SCHATZMAYR, 2002). O Tabaco é usado também para produzir anticorpos para combater vários outros tipos de doenças, como por exemplo a febre amarela. Recentemente com o surto de ebola em vários países da África ocidental, foi criado um medicamento experimental contra esse vírus, desenvolvido por um laboratório americano em colaboração com uma empresa canadense, onde o anticorpo foi produzido partir de folhas de tabaco. Por meio de *Agrobacteriumtumefasciens* as sequências codificantes dos anticorpos desejados podem ser inseridas no genoma da planta, para que esta passe a produzi-los. (NANOCELL NEWS, 2014).

Uma variedade transgênica de tabaco foi desenvolvida pela Universidade de Cornell, nos Estados Unidos, e pelo Centro de Pesquisas Rothamsted, no Reino Unido. O tabaco geneticamente modificado (GM) possui dois genes de cianobactéria, as algas azuis. Plantas e cianobactérias realizam a fotossíntese (transformação de dióxido de carbono, água e luz em oxigênio e energia), porém, as algas azuis conseguem fazer o processo mais rápido do que a maioria dos vegetais. Esta é a primeira vez que uma planta foi desenvolvida pela engenharia genética para fixar todo o seu carbono por uma enzima cianobacteriana. É um primeiro passo importante no desenvolvimento de plantas com uma fotossíntese mais eficiente. De acordo com o recente estudo publicado pela revista científica Nature, culturas que apresentarem a característica da rápida fixação de carbono das cianobactérias poderão produzir mais. O aumento de produtividade de maneira sustentável e a conservação de recursos naturais é uma preocupação global, uma vez que a população mundial já é de 7 bilhões de habitantes e deve chegar a 9 bilhões até 2050 (CORNELL CHRONICLE, 2014).

Melhoramento Genético do Tabaco

No melhoramento de fumo muitos são os métodos utilizados. A seleção massal foi responsável pelo desenvolvimento dos principais tipos de fumo usados pelas indústrias fumageiras (Matzinger & Wernsman, 1979). Quando o objetivo é combinar características desejáveis encontradas em duas ou mais cultivares, o método genealógico é o procedimento mais adequado. Entretanto, quando alguma característica ou resistência à moléstia se encontra em outra espécie do gênero *Nicotiana* ou tipo de fumo o método retrocruzamento é o mais indicado (LEGG & SMEETON, 1999).

Para o melhoramento genético, a utilização de genes oriundos de espécies selvagens nem sempre é acompanhada de ganhos agronômicos, uma vez que a introdução de genes utilizando hibridações e repetidos retrocruzamentos pode ocasionar o arraste gênico, ou seja, em conjunto com a introdução do gene de interesse podem ser incorporados outros genes a ele ligados geneticamente, de efeitos deletérios (BROWN, 2002).

Os principais objetivos do melhoramento genético do fumo são: a melhoria da qualidade e produtividade das lavouras, a obtenção de resistência às principais doenças que atacam a cultura, como as viroses (TMV, PVY e TSWV), a murcha bacteriana (*Ralstonia solanacearum*), os nematoides de galhas (*Meloidogyne incognita* e *M. javanica*), a tolerância ao amarelão (complexo de fungos de solo), e a obtenção de cultivares com baixos teores de alcalóides, especialmente as nitrosaminas (TSNA's), que são cancerígenas (GAVILANO et al., 2006).

O mofo azul é a principal doença da cultura do fumo, causada pelo fungo *Peronospora tabacina*, que danifica as folhas inviabilizando-as. A Bahia é declarada pelo Ministério da Agricultura (Mapa) como Área Livre de Mofo Azul, praga que atinge a cultura do tabaco e impede as exportações do produto para outras partes do mundo. A Bahia está autorizada a exportar charuto para o mercado chinês. Reuniões de trabalho foram realizadas na China, e uma delegação de técnicos chineses veio à Bahia, comprovando a qualidade e sanidade do tabaco e charutos produzidos no Recôncavo Baiano. Essa decisão foi devido ao intenso trabalho que vem sendo realizado há mais de três anos pelo Ministério da Agricultura, governo da Bahia através da Secretaria Estadual da Agricultura, Embrapa Mandioca e Fruticultura, prefeituras dos municípios onde a cultura está presente, da Universidade Federal do Recôncavo da Bahia, do Fórum dos Secretários da Agricultura do Recôncavo, da Câmara Setorial do Charuto, e do Sinditabaco, com o objetivo de reabilitar a cultura do fumo no Recôncavo e recuperar milhares de empregos diretos e indiretos perdidos na região com o fechamento de diversas fábricas, entre elas a Suerdieck (JORNAL BAHIA ON LINE, 2012).

Durante a etapa de geração de variabilidade nos programas de melhoramento, podem ser usados métodos convencionais (através de cruzamentos) e biotecnológicos, incluindo cultura de tecidos e células, obtenção de somaclones, hibridação somática (fusão de protoplastos) e produção de haplóides e duplo-haplóides (LEWIS et al., 2007).

Em geral, os fumicultores estão interessados nos atributos que aumentem características tais como: resistência a moléstias, altos rendimentos de folha, melhorias na qualidade, facilidade de colheita e cura. Por outro lado, as indústrias de fumo desejam alta produção de lâmina e diminuição de talo, composição

química e física equilibrada necessárias para a produção de misturas com aroma e sabor apropriado (LEGG & SMEETON, 1999).

As variedades de fumo desenvolvidas pelos programas de melhoramento são predominantemente linhas puras. Em menor escala, o desenvolvimento de híbridos tem sido utilizado especialmente quando o objetivo é resistência à pragas e doenças. Como exemplo específico de cultivares híbridos podem ser citadas as cultivares de fumo resistentes à TMV (Tabacco Mosaic Virus) (LEGG & SMEETON, 1999).

Caracterização da Variabilidade Genética

Na implantação de um programa de melhoramento genético, uma das principais necessidades do melhorista é o conhecimento do germoplasma disponível e a capacidade de identificar plantas que possuam genes de interesse para o programa (WEEDEN et al., 1994).

O grau de relacionamento genético entre genótipos pode ser estimado através de diferentes métodos, podendo ser baseado em dados de genealogia, caracteres morfológicos e marcadores moleculares ao nível de DNA (MELCHINGER et al., 1994).

Na caracterização da diversidade genética das espécies vegetais, animais e de microrganismos, os pesquisadores têm o interesse em agrupar genótipos similares, de maneira que as maiores diferenças ocorram entre os grupos formados. Neste aspecto, técnicas multivariadas, como análise discriminante, componentes principais, análise de coordenadas e de agrupamento, podem ser aplicadas neste tipo de estudo. A adoção de uma, entre as técnicas citadas, varia de acordo com o padrão de resultado desejado e com a informação disponível, seja ela característica morfológica, fisiológica, ecológica ou genético-molecular (CRUZ et. al., 2011). Dentre estas, pode ser destacada a análise de agrupamento que é muito utilizada pelos pesquisadores tanto da área de melhoramento genético vegetal quanto na caracterização morfológica de novos acessos, ou seja, na caracterização morfológica de coleções de constituições genéticas geralmente mantidas em bancos de germoplasma e ainda pouco conhecidas pelos melhoristas (KOOP et al., 2007).

A análise de correlação cofenética (SOKAL e ROHLF, 1962) associada à análise de agrupamento, podem ser empregadas para aumentar a confiabilidade das conclusões frente a interpretação dos dendrogramas. A correlação cofenética é uma análise que estabelece uma correlação entre a matriz de similaridade ou dissimilaridade com o dendrograma gerado através desta, ou seja, compara as reais distâncias obtidas entre os acessos com as distâncias representadas graficamente sujeitas ao acúmulo de erro supracitado (KOOP et al., 2007).

Caracteres fenotípicos têm sido utilizados desde os tempos de Mendel em estudos de genética e melhoramento. As principais limitações do uso deste tipo de marcador para a caracterização de germoplasma são o efeito do ambiente, a ação gênica, a epistasia e a pleitropia, que podem dificultar a avaliação. Apesar disso, diversos autores têm utilizado marcadores fenotípicos para caracterizar germoplasma de diferentes espécies (SANTOS, 2002; SUDRÉ et al., 2006; COIMBRA et al., 2010; COSTA, 2012; CONCEIÇÃO et al., 2014). O estudo da diversidade genética por meio de caracteres fenotípicos, principalmente os de natureza quantitativa, é de interesse no melhoramento aplicado, tendo em vista sua importância econômica e a necessidade de se obter êxito na escolha adequada de combinações híbridas superiores (VILLELA, 2013).

Características fenotípicas, influenciadas por fatores ambientais, têm sido usadas para caracterização e registro de cultivares de fumo em diferentes países, com base na descrição recomendada pela UPOV (Union pour La Protection des Obtentions Variétales). Caracteres como forma e tamanho da folha, número de folhas, altura da planta e comprimento de internódios são importantes porque influenciam o manejo, o rendimento e a composição química das folhas. Características desejáveis incluem a insensibilidade ao florescimento precoce, que reduz o número de folhas por planta; resistência ao acamamento; amadurecimento uniforme das folhas; ausência de excessiva sensibilidade à quebra da folha e inclinação. Ainda, cultivares do tipo Burlei não devem ter estatura muito elevada e possuir um diâmetro moderado do caule para evitar rachaduras por ocasião da colheita. Informações sobre o mecanismo de herança de muitos destes caracteres de interesse agronômico foram obtidas e contribuíram de forma efetiva para o melhoramento genético da espécie. Diversos mutantes de grande efeito sobre o fenótipo foram identificados para caracteres de

folha, altura da planta e comprimento de internódios (HUMPHREY et al., 1964; SMITH, 1950; LEGG & COLLINS, 1982).

Seleção de Descritores Morfoagronômicos

Nas coleções de germoplasma, o termo descritor é utilizado para se referir a um atributo ou caráter que se observa ou se mensura nos acessos (QUEROL, 1993), sendo capaz de discriminar um acesso de outro. Nos bancos de germoplasma, frequentemente há um grande número de acessos que necessita ser avaliado, além de ser regra geral as observações e a mensuração de um grande número de caracteres (PEREIRA, 1989). Em muitos casos, são obtidos sem nenhum critério sobre sua real contribuição para a variabilidade e esse tipo de procedimento, além de produzir a duplicação da mesma informação, tem contribuído para uma análise multivariada, confusa e de difícil interpretação (DIAS, 1994).

No geral, todo caráter deve apresentar uma parcela de contribuição na variação do germoplasma analisado. Mas, há uma tendência de que o aumento do número de descritores avaliados ocasione a presença de informações redundantes, posto que essas informações quase sempre estão associadas a outras (DAHER, 1993). Logo, a eliminação dos redundantes seria uma decisão vantajosa, pois reduziria o trabalho de tomada de dados sem ocasionar perda na precisão da caracterização, especialmente se esses caracteres forem de difícil mensuração e apresentarem baixa variabilidade e estabilidade de expressão (PEREIRA, 1989).

O descarte deve se mostrar efetivo na representação da variação total, além de proporcionar uma redução nos gastos com mão-de-obra e no tempo destinado à tomada de dados. A seleção de descritores tem sido realizada com base em várias análises estatísticas, podendo-se mencionar: a regressão e interdependência de dados, o coeficiente de repetitividade, variáveis canônicas e componentes principais (CRUZ, 1990). Contudo, a análise de componentes principais vem se destacando como a metodologia mais empregada em bancos e/ou coleções de germoplasma, pois além de identificar os caracteres mais importantes na contribuição de variação total disponível entre os indivíduos

analisados, fornece indicação para eliminar os que pouco contribuem (DIAS, 1994; ALVES, 2002).

Jolliffe (1973) impulsionou o emprego da análise de componentes principais (ACP) no descarte de caracteres a partir da publicação de seus trabalhos. Analisando quatro métodos de descarte com base em dados simulados e reais, concluiu que esse procedimento era satisfatório quando o número de caracteres rejeitados fosse igual ao de componentes principais que apresentassem variâncias inferiores a 0,7. Posteriormente, Mardia et al. (1979), complementando essa metodologia, recomendaram o descarte com base na observação dos componentes principais que apresentassem autovalores inferiores a 0,70 e, em cada um desses componentes, fosse descartado o caráter com maior coeficiente de ponderação em valor absoluto (autovetor). Esse procedimento foi denominado, por Cruz (1990), de seleção direta.

Wilches (1983), aplicando a análise de componentes principais em 34 variedades de amendoim, propôs uma seleção prévia antes da utilização da metodologia de Jolliffe (1973) e descartou os caracteres altamente influenciados pelo ambiente, com base na informação da análise de variância univariada realizada para cada caráter. Cruz (1990) menciona outros trabalhos que empregaram a seleção prévia por meio de outras análises estatísticas.

Pereira (1989) iniciou a utilização dessa metodologia no descarte de caracteres redundantes, quando caracterizou 208 acessos de mandioca com base em 28 caracteres e conseguiu descartar 50% dos caracteres analisados, o que proporcionou redução no trabalho e facilidade na interpretação dos dados. Daher (1993), empregando a mesma metodologia na aplicação de 22 caracteres em 60 acessos de capim-elefante, obteve uma redução de 63,6% no conjunto analisado.

Alterações foram propostas por Strapasson (1997) para aumentar a eficiência do descarte com o emprego da análise de componentes principais (ACP). Cury (1993) modificou parcialmente a metodologia de Jolliffe (1973) quando estudou 20 caracteres em 30 acessos de mandioca, propondo uma nova análise com os remanescentes após o descarte de cada caráter, além da observação da matriz de correlação fenotípica para auxiliar no descarte dos caracteres redundantes. O procedimento foi realizado até não ser possível

discriminar o maior autovetor no último componente principal e considerou, a partir dessa situação, o processo inconsistente. Com essa modificação, reduziu 30% dos caracteres, em vez dos 65% propostos na metodologia inicial, sem perda significativa de informações e concluiu que o número de descarte não deve ser pré-fixado, como sugerido na seleção direta.

A análise de variáveis canônicas (AVC) é um procedimento alternativo à ACP, que apresenta como diferença única, o uso adicional da matriz de covariância residual para a obtenção das combinações lineares das variáveis originais. Essa análise permite saber quais características foram mais importantes para classificar a divergência dos acessos. A importância relativa das variáveis canônicas decresce da primeira para a última, sendo as últimas responsáveis pela explicação mínima da variação total existente. Assim, esta avaliação possibilita o descarte de caracteres que pouco contribuem para a discriminação do material avaliado, reduzindo mão-de-obra, tempo e custos, além de contribuir para uma mensuração mais detalhada das variáveis efetivas para a caracterização (CRUZ & REGAZZI, 2001).

Tal como a análise de componentes principais (ACP), a técnica A análise de variáveis canônicas (AVC) tem sido utilizada também na identificação das variáveis mais importantes num conjunto de descritores, os chamados descritores mínimos, com consequente descarte daquelas menos relevantes ou redundantes para a caracterização do germoplasma (FONSECA & SILVA, 1999; RIBEIRO et al., 1999). A análise de variáveis canônicas (AVC) é um procedimento alternativo à ACP, que apresenta como diferença única, o uso adicional da matriz de covariância residual para a obtenção das combinações lineares das variáveis originais (CRUZ & REGAZZI, 2001).

Cruz (1990) denominou esse procedimento de seleção com reanálise. No segundo trabalho, o número de acessos era bem inferior ao número de caracteres avaliados, os quais pertenciam a diferentes grupos, descartando os redundantes, dentro de cada grupo, com base na metodologia inicial de Jolliffe (1973). Em seguida, procedeu a mais uma análise, utilizando todos os descritores previamente selecionados para definir o conjunto final de descritores e concluiu que, dos 40 caracteres avaliados, apenas oito seriam importantes na quantificação da avaliação dos acessos.

Outras metodologias vêm sendo empregadas na avaliação da eficiência do descarte, como o estudo comparativo dos agrupamentos formados pelo dendrograma (BEKELE et al., 1994) e a comparação por meio de medidas de similaridade, estimativas pelo coeficiente de correlação entre os pares obtidos (r_1) e entre dois conjuntos de componentes (Q_1), utilizada por Strapasson (1997). Essa última metodologia é indicada para condições onde o número de caracteres é superior ao número de acessos.

Afonso et al., (2014), em estudo de seleção de descritores em mandioca realizou a identificação dos descritores redundantes, por dois procedimentos: 1) seleção direta, proposta por Jolliffe (1972, 1973), onde foram eliminados os caracteres que apresentaram maior coeficiente de ponderação em valor absoluto (autovetor), no componente principal de menor autovalor, partindo do último componente até aquele cujo autovalor não excedeu 0,70; e 2) Seleção baseada no coeficiente de Singh (1981). Porém, o descarte final foi realizado com base na informação obtida nos dois procedimentos, sendo indicado para descarte o descritor identificado simultaneamente nos dois procedimentos.

Divergência Genética e Variabilidade Genética

A variabilidade genética é considerada como sendo a capacidade de uma espécie, de uma população ou de uma progênie expressar diferentes fenótipos (RAMALHO et al., 2000), enquanto a divergência genética pode ser medida entre indivíduos, progênies, populações, espécies, cultivares ou qualquer outro tipo de unidade amostral e corresponde a diferenças nas frequências alélicas entre as unidades consideradas. Assim, a variabilidade está diretamente relacionada com a divergência genética, uma vez que a amplitude da variabilidade em uma população segregante é função da divergência genética entre os pais envolvidos (FALCONER, 1987).

Estudos de divergência genética são, ainda, extremamente importantes no contexto de trabalhos de conservação genética e manutenção de bancos de germoplasma (PAULA, 2007). A identificação de acessos semelhantes em populações naturais e, ou bancos de germoplasma, por meio da identificação da

divergência, direcionam trabalhos de coleta de sementes e descarte de materiais redundantes (MARTINELLO et al., 2002; PEREIRA et al., 2003).

A maioria das espécies exploradas agronomicamente teve sua diversidade reduzida em consequência da domesticação e dos processos de seleção e melhoramento de plantas. A diversidade genética é considerada mais baixa em cultivares modernas de espécies autógamas devido ao seu sistema de fecundação e também da sua domesticação fora do centro de origem, onde um número limitado de sementes (ou acessos) foi levado pelos exploradores e serviram como a base genética das cultivares modernas de hoje (SAAVEDRA & SPOOR, 2002).

Em programas de melhoramento, a importância do conhecimento da diversidade genética está no fato de que cruzamentos que envolvam genitores geneticamente divergentes são os mais eficientes em produzir híbridos com maior efeito heterótico na progênie e maior variabilidade genética nas gerações segregantes (FALCONER, 1987).

Essa divergência pode ser avaliada a partir de uma série de marcadores que podem ser morfológicos, fisiológicos, citológicos, proteicos, bioquímicos e moleculares (AMARAL JÚNIOR & THIÉBAUT, 1999). As informações múltiplas de cada acesso ou cultivar são expressas em medidas de dissimilaridade, que representam a diversidade em relação ao conjunto de acessos (CRUZ & CARNEIRO, 2003).

Em diversos trabalhos, as análises de divergência genética são realizadas considerando os caracteres de diferentes naturezas individualmente. MOHAMMADI; PRASANNA, 2003; ZEWDIE et al., 2004; KARASAWA et al., 2005; SUDRÉ et al., 2006; BENTO et al., 2007; OLIVEIRA et al., 2007; CONCEIÇÃO et al., 2014).

Apesar da maioria dos trabalhos em relação a divergência genética, há anos terem sido realizados utilizando-se análise em isolado, nos últimos anos as análises levando em consideração mistura de variáveis vem sendo muito utilizadas. Os resultados demonstram que a análise conjunta, de certo modo, permite uma melhor compreensão do fenômeno biológico quando comparada com as análises individuais para cada tipo de variável (LEDO & GONÇALVES, 2012).

Transformação de Variáveis Quantitativas em Multicategóricas

A transformação de variáveis quantitativas em multicategóricas pode ser utilizada para facilitar sua caracterização com informações preliminares de grande utilidade. Existem vários métodos para se fazer essa transformação, porém estes precisam ser melhor entendidos para que a perda de informações ocorrida na transformação não prejudique significativamente os resultados da análise (BARROSO, 2010).

Estabelecimento do Número de Classes

Segundo Martins et al. (2011), os métodos mais utilizados para a distribuição de valores quantitativos em classes são a regra de Sturges (1926) e a regra do quadrado.

A transformação dos dados quantitativos pode ser realizada com o auxílio do programa estatístico Genes (CRUZ 2014), que dá suporte para o uso desses procedimentos, sendo as estratégias citadas abaixo:

A “Regra de Sturges” e raiz quadrada do número de observações.

“Regra de Sturges”, fornece o número de classes em função do total de observações:

$$k = 1 + 3{,}3 * \log^{10}(n)$$

Onde: K é o número de classes; n é o número total de observações ou seja o número total de dados.

O método que utilizada o cálculo da raiz quadrada também é baseada no número de observações e é dada por:

$$K = \sqrt{N}$$

Onde: K é o número de classes; *N* é o número total de observações ou seja o número total de dados.

Medidas de Dissimilaridade

As medidas de dissimilaridade são importantes em estudos de diversidade genética, pois identificam genitores possíveis de serem utilizados em programas de melhoramento. As medidas de dissimilaridade são diferentes para cada grupo de variáveis: quantitativas, binárias e multicategóricas. Porém, serão abordadas apenas as medidas de dissimilaridade obtidas por variáveis quantitativas e multicategóricas (BARROSO, 2010).

Variáveis Multicategóricas

Variáveis multicategóricas – características morfológicas atribuídas à estrutura de planta, assim como atributos que conferem qualidade aos produtos comercializados, como forma, cor e sabor – são comumente determinadas utilizando-se a distância de Cole-Rodgers (1997), na qual as características que normalmente não podem ser ordenadas são classificadas em escalas, podendo então ser analisadas como características quantitativas discretas (CRUZ & CARNEIRO, 2003).

No melhoramento vegetal os caracteres multicategóricos são comumente avaliados, principalmente aqueles relacionados com particularidades morfológicas e estruturais da planta, além de se ter grande interesse em certos atributos que conferem qualidade ao produto comercializado, como a coloração, a forma e o sabor do fruto (CRUZ & CARNEIRO, 2006).

Variáveis Quantitativas

As medidas mais utilizadas para caracteres quantitativos nos estudos genéticos são: a distância euclidiana, a distância euclidiana média, o quadrado da distância euclidiana média, a distância ponderada e a distância generalizada de Mahalanobis (CRUZ & CARNEIRO, 2006).

Distância Euclidiana Média

A distância euclidiana média tem sido utilizada de forma alternativa à distância euclidiana, pois o valor desta sempre aumenta com o acréscimo do número de características consideradas na análise.

A diferença da distância Euclidiana Média para a de Mahalanobis está na matriz de variâncias e covariâncias que é substituída pela matriz diagonal do inverso do número de variáveis (diag (1/p)).

$$d(X_l X_k) = [(X_l - X_k)^T \, diag(1/p)(X_l - X_k)]^{1/2}$$

Como a distância Euclidiana cresce com o aumento do número de variáveis, essa distância consegue eliminar o efeito do número de variáveis ao utilizar a matriz diag (1/*p*).

Distância Generalizada de Mahalanobis

A distância generalizada de Mahalanobis (D^2) leva em consideração as variâncias e as covariâncias residuais que existem entre as características mensuradas, possíveis de serem quantificadas quando as avaliações são realizadas em genótipos avaliados em delineamentos experimentais. Essa é uma vantagem em relação às distâncias euclidianas.

Quando se dispõe de informações de ensaios experimentais com repetições é possível se obter a matriz de dispersão residual (Ψ) e as médias das características. Com base nessas informações, obtêm-se as estimativas das distâncias de Mahalanobis por meio da expressão:

$$D^2_{ii'} = \delta' \psi^{-1} \delta$$

em que:

$D^2_{ii'}$: é a distância de Mahalanobis entre os genótipos i e i';

Ψ: matriz de variâncias e covariâncias residuais;

δ' = [d_1 d_2 ... d_v], sendo $d_j = Y_{ij} - Y_{i'j'}$;

Y_{ij}: é a média do i-ésimo genótipo em relação à j-ésima variável.

Distância de Gower

A distância de Gower é calculada como a soma dos quadrados da diferença entre as matrizes de distâncias cofenéticas e a original.

A análise simultânea proposta por Gower (1971) é expressa por:

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} \cdot S_{ijk}}{\sum_{K=1}^{p} W_{ijk}}$$

Em que *K* é o número de variáveis (k = 1, 2,…, p = número total de características avaliadas); *i* e *j* dois indivíduos quaisquer; W_{ijk} é um peso dado a comparação *ijk*, atribuindo valor 1 para comparações válidas e valor 0 para comparações inválidas (quando o valor da variável está ausente em um ou ambos indivíduos); S_{ijk} é a contribuição da variável *k* na similaridade entre os indivíduos *i* e *j*, ele possui valores entre 0 e 1. Para uma variável nominal, se o valor da variável *k* é a mesma para ambos os indivíduos, *i* e *j*, então S_{ijk} = 1, caso contrário, é igual a 0; para uma variável contínua $S_{ijk} = 1 - | x_{ik} - x_{jk} | / R_k$ onde x_{ik} e x_{jk} são os valores da variável *k* para os indivíduos *i* e *j*, respectivamente, e R_k é a amplitude de variação da variável *k* na amostra. A divisão por R_k elimina as diferenças entre escalas das variáveis, produzindo um valor dentro do intervalo [0, 1] e pesos iguais.

Comparação entre as Matrizes de Distâncias Genéticas

As matrizes das variáveis transformadas, construídas para todas medidas de dissimilaridade, são comparadas com as obtidas pelos dados originais pelo teste de Mantel (MANLY,1997).

O valor Z de Mantel é dado por:

$$Z = \sum_{i,j=1}^{n} X_{ij} Y_{ij} \ ,$$

Onde X_{ij} e Y_{ij} são elementos das matrizes X e Y a serem comparadas. A significância desse valor de Z pode ser obtida comparando-se esse valor observado com valores de uma distribuição sob hipótese nula, recalculando-se os valores de Z diversas vezes, aleatorizando, em cada uma delas, a ordem dos elementos de uma das matrizes. Este Z calculado após permutações aleatórias é chamado de Z randômico (Z_{rnd}). A estatística Z possui uma relação monotônica com o r de Pearson entre as matrizes (correlação matricial), de modo que ela é de fato utilizada para testar a significância do r (MANLY, 1997). A correlação calculada pelo teste de Mantel varia de -1 a +1, e mede a correlação entre duas matrizes com relação ao Z randômico (Z_{rnd}). Para valores negativos de r, quanto menor a freqüência do $Z_{rnd} \leq Z_{obs,}$ maior a correlação entre as duas matrizes. Para valores positivos de r, quanto menor a freqüência de $Z_{rnd} \geq Z_{obs,}$ maior a correlação. Neste trabalho, 1000 permutações aleatórias serão utilizadas para se testar a significância das correlações matriciais.

Análise de Agrupamento

De acordo com Mingoti (2005), a Análise de Agrupamentos também é conhecida como Análise de Conglomerados ou Análise de Classificação ou *Cluster Analysis*. Seu objetivo é agrupar os elementos da amostra ou população em grupos. Os elementos de um mesmo grupo são homogêneos entre si, no que se refere às variáveis (características) que neles foram medidas. Por outro lado estes grupos já formados são heterogêneos entre eles em relação a estas mesmas características.

Segundo Hair et al. (2005), o objetivo principal da Análise de Agrupamentos é situar as observações homogêneas em grupos, a fim de definir uma estrutura para os dados. Para isto, são abordadas algumas questões básicas que devem ser consideradas durante a análise.

Quando o número de genótipos é relativamente alto, torna-se praticamente inviável o reconhecimento dos grupos que apresentam semelhanças a partir da matriz de dissimilaridade. Contudo, com o uso de técnicas de agrupamento, podem-se classificar os genótipos em vários grupos de forma que exista homogenidade dentro dos grupos e heterogenidade entre os grupos, seguindo o critério de similaridade ou de dissimilaridade (CRUZ & CARNEIRO, 2003).

A primeira decisão na análise se refere à medida de similaridade que deve ser estabelecida, ou seja, deve-se estabelecer a associação de dois objetos baseada nas variáveis da 'variável estatística de agrupamento'. Hair et. al (2005) define a 'variável estatística de agrupamento' como "o conjunto das variáveis que representam as características usadas para comparar objetos na análise de agrupamentos".

Para Mingoti (2005), é indispensável decidir *à priori,* a medida de similaridade que será utilizada para se proceder ao agrupamento de elementos. Para isto, existem medidas apropriadas para análise de variáveis qualitativas e quantitativas. As medidas apropriadas para variáveis quantitativas também são ditas 'de dissimilaridade'. Neste caso, quanto menores os seus valores, mais similares serão os elementos que estão sendo comparados. O próximo passo da análise se refere à formação do agrupamento do método hierárquico a ser empregado. A última decisão na análise refere-se à escolha do número de agrupamentos. Deve-se haver um equilíbrio entre definir a estrutura mais básica com o nível de similaridade dentro dos agrupamentos porque quando o número de agrupamento diminui, a homogeneidade dentro dos grupos necessariamente diminui (HAIR, 2005).

Os métodos de agrupamento utilizados para análise da diversidade podem ser classificados como "métodos hierárquicos" e "técnicas de projeção". Métodos hierárquicos, métodos hierárquicos gerais e aglomerativo em particular, são mais comumente empregados na análise da diversidade genética em espécies vegetais (MOHAMMADI e PRASANNA, 2003).

Métodos de Agrupamento Hierárquicos

Nos métodos hierárquicos, os genótipos são agrupados por um processo que se repete em vários níveis, até que seja estabelecido o diagrama de árvore ou o dendrograma. Neste caso, o maior interesse está na "árvore" e nas suas ramificações e não no número ótimo de grupos. As delimitações podem ser estabelecidas por um exame visual do dendrograma, em que se avaliam pontos de alta mudança de nível, tomando-os em geral como delimitadores do número de genótipos para determinado grupo (CRUZ & CARNEIRO, 2006).

Segundo Mingoti (2005), os métodos de agrupamentos hierárquicos mais comuns e disponíveis na grande maioria dos softwares estatísticos são: os métodos de ligação simples (single linkage); de ligação completa (complete linkage); UPGMA - (Unweighted Pair-Group Method Using Arithmetic Averages); do centróide (centroid method) e de Ward.

As técnicas hierárquicas são as mais amplamente difundidas (SIEGMUND et al., 2004) e envolvem basicamente duas etapas. A primeira se refere à estimação de uma medida de similaridade ou dissimilaridade entre os indivíduos e a segunda, à adoção de uma técnica de formação de grupos (SANTANA & MALINOVSKI, 2002).

A seguir encontram-se os métodos de agrupamentos utilizados nesse trabalho. É importante lembrar que em todos os métodos hierárquicos, a ideia é agrupar os indivíduos mais similares e que estes se diferenciam na forma como atualizam a matriz de distâncias.

Método UPGMA - Método de ligação média entre grupos – (Unweighted Pair-Group Method Using Arithmetic Averages).

Nesse método, a matriz de distâncias é atualizada calculando-se a média das distâncias entre os indivíduos de dois grupos. Assim, se C_1 tem n_1 indivíduos e C_2 tem n_2 indivíduos, a distância entre eles será definida por

$$d(C_1, C_2) = \sum_{l \in C_1} \sum_{k \in C_2} \left(\frac{1}{n_1 n_2} \right) d(X_l, X_k)$$

Esse método visa trabalhar com médias ao invés de valores extremos.

O método hierárquico das medidas da dissimilaridade ponderada (UPGMA) é o mais utilizado em diversidade, tendo vantagem sobre os demais métodos por considerar médias aritméticas das medidas de dissimilaridade, o que evita caracterizar a dissimilaridade por valores extremos entre os genótipos (CRUZ e CARNEIRO, 2003).

Método de Ward

O método de Ward (WARD, 1963) ou de variância mínima consiste em formar grupos a partir de pares que proporcionem a menor soma de quadrados.

Cada elemento é considerado um conglomerado e então, calcula-se a soma de quadrados dentro de cada conglomerado. Esta soma é o quadrado da distância Euclidiana de cada elemento pertencente ao conglomerado em relação ao correspondente vetor de médias do conglomerado.

$$SS_i = \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.} \right)^T \left(X_{ij} - \bar{X}_{i.} \right)$$

em que n_i é o número de elementos do conglomerado C_i quando se está no passo k do processo de agrupamento; X_{ij} é o vetor de observações do j-ésimo elemento pertencente ao i-ésimo conglomerado ; $\bar{X}_{i.}$ é o vetor de médias do conglomerado C_i e SS_i é a soma de quadrados referente a tal conglomerado (MINGOTI, 2005).

Posteriormente, calcula-se a soma de quadrados entre dois conglomerados C_l e C_i que é dado por:

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i} \right] \left(\bar{X}_{l.} - \bar{X}_{i.} \right)^T \left(\bar{X}_{l.} - \bar{X}_{i.} \right)$$

em que $\left[\frac{n_l n_i}{n_l + n_i}\right]$ é um fator de ponderação para quando os conglomerados tiverem tamanhos diferentes (MINGOTI, 2005).

A cada passo do algoritmo, os dois conglomerados que minimizam tal distância são combinados.

Análises Simultânea de Dados

Em determinados casos, como nos estudos de diversidade genética, torna-se necessário à avaliação conjunta de diversas variáveis agronômicas, morfológicas e moleculares. A análise individual para cada tipo de variável pode levar a discrepâncias em relação aos agrupamentos e às inferências em relação à quantificação da variabilidade entre as unidades experimentais ou amostrais a serem agrupadas. Com isso, a analises com métodos que considerem simultaneamente os diversos tipos de variáveis é preferível (LEDO e GONÇALVES, 2012). Segundo Cruz et al. (2011) quando a diversidade genética é estudada a partir de características de diferentes naturezas, diferentes estratégias de análises podem ser recomendadas.

De forma geral, pode-se obter uma matriz de dissimilaridade de três formas distintas: usando apenas variáveis numéricas; usando somente variáveis categóricas; e a utilização conjunta dessas variáveis, sendo que esse último procedimento pode ser realizado por alguns procedimentos: utilizando um coeficiente que calcula a similaridade de uma só vez para essa mistura de variáveis (GOWER, 1971). Uma outra forma de se obter uma matriz conjunta é através da Soma de matrizes: Na metodologia descrita por Cruz et al. (2011) é proposta a soma algébrica das distâncias padronizadas das matrizes individuais. Também é possível realizar a análise simultânea com o processo de “Qualitificação” (Qualitizing**)** que consiste em transformar as variáveis quantitativas em multicategóricas, visando a obtenção de uma matriz única. Martins et al., (2011) utilizou as estratégias de transformação de dados, soma de matrizes de dissimilaridade e o algoritmo de Gower na avaliação da diversidade genética de tomateiro.

Outra estratégia de análise conjunta que tem sido utilizada é o *Modified location Model* (MLM) proposto por Franco et al. (1998). Esse método apresenta duas fases, a primeira o método de agrupamento de Ward (WARD Jr., 1963) é aplicado sobre a matriz de dissimilaridade de Gower (GOWER, 1971). Na segunda fase, o vetor de médias das variáveis quantitativas é estimado pelo procedimento MLM para cada subpopulação, independente da variável qualitativa. Estes procedimentos foram utilizados recentemente por Barbé et al. (2010), Cabral et al. (2010), Sudré et al. (2010), dentre outros.

Sarkar et al. (2011) apresenta a comparação dos seguintes métodos de análise conjunta de variáveis quantitativas e qualitativas: PCAMIX - análise de componentes principais de variáveis mistas (DE LEEUW e VAN RIJCKEVORSEL, 1980); INDOMIX - escalonamento da diferença individual com restrições sobre coordenadas ortonormais para mistura de variáveis (CARROL e CHANG, 1970); PRINQUAL - componentes principal dos dados qualitativos, quantitativos ou mistura (WINSBERG e RAMSAY, 1983); EM algoritmo de maximização da esperança (DEMPSTER et al., 1977); e RNA - rede neural artificial (KOHONEN, 1988). Esses métodos são utilizados, principalmente, na área de taxonomia, botânica, psicologia e educação, não sendo encontrados trabalhos realizados no Brasil com dados agronômicos.

A análise de agrupamento a partir de variáveis quantitativas e qualitativas simultaneamente tem sido muito utilizada no Brasil nos últimos 5 anos em estudos relacionados ao melhoramento genético de plantas, principalmente nos estudos de diversidade genética. Os resultados demonstram que a análise conjunta, de certo modo, permite uma melhor compreensão do fenômeno biológico quando comparada com as análises individuais para cada tipo de variável (LEDO e GONÇALVES, 2012).

REFERÊNCIAS

AFONSO, S. D. J.; LEDO, C. A. da S.; MOREIRA, R. F. C.; SILVA, S. de O. e; LEAL, V. D. de J.; CONCEIÇÃO, A. L. da S. Selection of descriptors in a morphological characteristics considered in cassava accessions by means of

multivariate techniques. **Journal of Agriculture and Veterinary Science**, v. 7, *Issue 1 Ver. V*, p. 13-20, Feb. 2014.

AFUBRA - Associação dos Fumicultores do Brasil. Disponível em: <http://www.afubra.com.br/index.php/home>. Acesso em: 01 dez. 2014.

ALVES, R. M. **Caracterização genética de populações de cupuaçuzeiro, Theobroma grandiflorum (Will ex Spreng) Schum., por marcadores microssatélites e descritores botânico-agronômicos**.Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba. p.146, 2002.

AMARAL JÚNIOR, A. T., THIÉBAUT, J. T. L. **Análise Miltivariada na avaliação da diversidade genética em recursos genéticos vegetais**. Apostila: CCTA – UENF. p. 55,1999.

ANUÁRIO Brasileiro do Fumo 2011. Santa Cruz do Sul: Gazeta do Sul. 2011.

BARBÉ, T.C.; AMARAL JÚNIOR, A.T.; GONÇALVES, L.S.A.; RODRIGUES, R.; SCAPIM, C.A. Association between advanced generations and genealogy in inbred lines of snap bean by the Ward-Modified Location Model. **Euphytica**, v. 173, p. 337-343, 2010.

BARROSO, N. C. **Categorização de dados quantitativos para estudos de diversidade genética.** Dissertação de Mestrado em Estatística Aplicada e Biometria, Universidade Federal de Viçosa, Minas Gerais – Brasil. Dezembro, p. 99, 2010.

BEKELE, F.L.; KENNEDY, A. J.; McDAVID, C.; LAUCKNER, F.B.; BEKELE, I. Numerical taxonomic studies on cacao (Theobroma cacao L.) in Trinidad. **Euphytica**, Dordrecht, v.75, n. 3, p.231-240, 1994.

BENTO, C. S.; SUDRÉ, C. P.; RODRIGUES, R.; RIVA, E. M.; PEREIRA, M. G. Descritores qualitativos e multicategóricos na estimativa da variabilidade fenotípica entre acessos de pimentas. **Scientia Agraria**, v. 8, n. 2, p. 149-156, 2007.

BINDLER, G; VAN DER HOEVEN, R.; GUNDUZ, I.; PLIESKE, J.; GANA, M.; ROSSI, L.; GADANI, F.; DONINI, P. A microssatelite marker based linkage map of tobacco. **Theoretical and Applied Genetics**, New York, v. 114, p. 341-349, 2007.

BLAND, M.M.; MATZINGER, D.F.; LEVINGS, C.S. Comparison of the mitochondrial genome of *Nicotiana tabacum* with its progenitor species. **Theoretical and Applied Genetics**, New York, v. 69, p. 535-541, 1985.

BROWN, J.K.M. Yield penalties of disease resistance in crops. **Current Opinion Plant Biology**, Oxford, v. 5, p. 339-344, 2002.

CABRAL, P.D.S.; SOARES, T.C.B.; GONÇALVES, L.S.A.; AMARAL JÚNIOR, A.T.; LIMA, A.B.P.; RODRIGUES, R.; MATTA, F.P. Quantification of the diversity among common bean accessions using Ward-MLM strategy. **Pesquisa Agropecuária Brasileira,** Brasília, v.45, n.10, p.1124-1132, 2010.

CAMPBELL, B. T Gotmare, V.; Dessauw, D.; Gband, M.; Du, X.; Jia, Y.; Constable, G.; Dillon, S.; Abdurakhmonov, I.Y.; Abdukarimov, A.; Rizaeva, S.M.; Abdullaev, A.A.; Barrose, P.A.V.; Padua, J.G.; Hoffman, L.V. & Podolnaya, L. **Status of the Global Cotton Germplasm Resources.** Crop Science, v. 50, jul./aug., 2010.

CARROL, J.D.; CHANG, J.J. Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart–Young decomposition. **Psychometrika**, v. 35, p. 283–319, 1970.

COIMBRA, R.R.; MIRANDA, G.V.; CRUZ, C.D.; MELO, A.V. de.; ECKERT, F.R. Caracterização e divergência genética de populações de milho resgatadas do

Sudeste de Minas Gerais. **Revista Ciência Agronômica**, v. 41, n. 01, p. 159-166, 2010.

COLLIS, W. K.; HAWKS, S.N. **Principles of Flue-Cured Tobacco Production.** Raleigh: N. C. State University, p. 301, 1993.

CONCEIÇÃO, A. L. da S.; SILVA, M. dos S. da.; SANTOS, C. C. dos.; ARAUJO, G. de M.; MOREIRA, R. F. C. Variabilidade genética e importância relativa de caracteres em acessos de tabaco (*Nicotiana tabacum* L.) Tipo broad leaf por meio de marcadores fenotípicos. **Enciclopédia Biosfera**, Centro Científico Conhecer - Goiânia, v.10, n.19; p.1900-1907, Dez. 2014.

CORNELL CHRONICLE. **Daily news from Cornell University**. Disponível em: < http://www.news.cornell.edu/stories/2014/09/plant-engineered-more-efficient photosynthesis>. Acesso Dez. 2014.

COSTA, T. P. P. **Caracterização Morfoagronômica de Genótipos de Tabaco na Região do Recôncavo da Bahia.** Dissertação de Mestrado em Recursos Genéticos Vegetais, Universidade Federal do Recôncavo da Bahia, Cruz das Almas, BA, Brasil. Maio, 2012.

CRUZ, C.D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas**. Tese (Doutorado em genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba. p.188, 1990.

CRUZ, C. D.; REGAZZI, A. J. Divergência Genética. In: **Modelos Biométricos Aplicados ao Melhoramento Genético**. CRUZ, C. D.; REGAZZI, A. J.-2. Ed.rev.- Viçosa: UFV, p. 287-323, 2001.

CRUZ, C.D., CARNEIRO, P.C.S. Modelos biométricos aplicados ao melhoramento genético. Viçosa: UFV, v.2, p. 585, 2003.

CRUZ, C. D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. 3. ed. Viçosa: UFV, 480 p, 2004.

CRUZ, C.D. **Programa Genes: análise multivariada e simulação**. Viçosa: Ed. UFV, p. 175, 2006.

CRUZ, C.D.; FERREIRA, F.M.; PESSONI, L.A.; **Biometria aplicada ao estudo da diversidade genética.** Visconde do Rio Branco-MG, Suprema, 620p, 2011.

CRUZ, C.D. Programa Genes - **Aplicativo computacional em genética e estatística**. Disponível em: <www.ufv.br/dbg/genes/genes.htm>. 2014.

CURY, R. **Dinâmica evolutiva e caracterização de germoplasma de mandioca (*Manihot esculenta*, Crantz) na agricultura autóctone do Sul do Estado de São Paulo**. Dissertação (Mestrado) – Escola Superior de Agricultura "Luiz de Queiroz, Universidade de São Paulo, Piracicaba. p.103, 1993.

DAHER, R. F. **Diversidade morfológica e isoenzimática em capim elefante (**Pennisetum purpureum Schum.). Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Universidade Federal de Viçosa, Viçosa, MG.p. 110, 1993.

D'ARCY, W. G. The Solanaceae since 1976, with a review of its biogeography. In: HAWKES, J. G.; LESTER, R. N.; NEE, M.; ESTRADA, N. (Ed.). **Solanaceae III: taxonomy, chemistry, evolution**. Kew: The Royal Botanic Gardens; London: The Linnean Society of London, p. 75-137, 1991.

DE LEEUW, J.; VAN RIJCKEVORSEL, J.L.A. **HOMALS and PRINCALS, some generalization of principal components analysis**. In: Diday E, Lebart L, Page`s JP and TomassoneR (ed.) Data Analysis and Informatics II. North Holland/ Amsterdam: Elsevier Science Publisher, p. 231–242, 1980.

DEMPSTER, A.P.; LAIRD, N.M.;RUBIN, D.B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society**, v. 39, p. 1–38, 1977.

DIAS, L. A. dos S. **Divergência genética e análise multivariada na predição de híbridos e preservação de germoplasma de cacau** (Theobroma cacao L.). Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba. p.94, 1994.

FALCONER, D. S. **Introdução à genética quantitativa**. Tradução de Martinho de Almeida e Silva e José Carlos da Silva. Viçosa: UFV, p. 279, 1987.

FARIAS, G. J; GERALD, I. O. Melhoramento genético de fumo (*Nicotiana tabacum* L.). LGN 5799 – Seminários em Genética e Melhoramento de Plantas. Programa de pós-graduação em genética e melhoramento de plantas. ESALQ/USP. Disponível em: < http://www.genetica.esalq.usp.br/pub/seminar/GJFarias-200702-Resumo.pdf>. 2007.

FONSECA, J. R.; SILVA, H. T. Identificação de duplicidades de acessos de feijão por meio de técnicas multivariadas. **Pesquisa Agropecuária Brasileira**, Brasília, v.34, n.3, p. 409-414, 1999.

FRANCO, J.; CROSSA, J.; VILLASEÑOR, J.; TABA, S.; EBERHART, S.A. Classifying genetic resources by categorical and continuous variables. **Crop Science**, v. 38, p. 1688-1696, 1998.

GAVILANO LB, COLEMAN NP, BURNLEY LE, BOWMAN ML, KALENGAMALIRO NE, HAYES A, BUSH L, SIMINSZKY B. Genetic engineering of *Nicotiana tabacum* for reduced nornicotine content. **J. Agric Food Chem.** Nov 29;54 (24):9071-8, 2006.

GERSTEL, D. U. Tobacco. *Nicotiana tabacum*. (Solanaceae) In: SIMMONS, N. W. (Ed.) **Evolution of Crop Plants.** New York: Longman, p. 273-277, 1979.

GOODSPEED, T. H.; WHEELER H-M; HUTCHISON, P. C. Taxonomy of *Nicotiana.* In: GOODSPEED, T. H. **The genus Nicotiana**. Waltham: Chronica Botanica. v. 16, pt. 6, p. 321-492, 1954.

GOWER, J.C. A general coefficient of similarity and some of its properties. **Biometrics**, Arlington, v. 27, n. 4, p. 857-874, 1971.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. Ed Bookman, Porto Alegre, p. 593, 2005.

HAWKES, J. G. In: NEE, M.; SYMON, D. E.; LESTER, R. N. **The economic importance of the family Solanaceae**. Kew: The Royal Botanic Gardens; London: The Linnean Society of London. p. 1-8, 1999.

HUMPHREY, A.B.; MATZINGER, D.F.; MANN, T.J. Inheritance of leaf shape in flue-cured tobacco (*Nicotiana tabacum* L.). **Heredity**, Oxford, v. 19, p. 615-628, 1964.

HUNZIKER AT. *The genera of Solanaceae*. Ruggell: A.R.G. Gantner Verlang K.G. ROOS, E.E.; MOORE, F.D. Effect of seed coating performance of lettuce seeds in greenhouse soil tests. **Journal American Society Horticultural Science**, v.100, p.573-576, 2001.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. II: real data. Journal of the **Royal Statistical Society Series C - Applied Statistics**, v. 22, p. 21-31, 1973.

JOLLIFFE, I.T. Discarding variables in a principal component analysis. I. Artificial data. **Applied Statistics**, v.21, p.160-173, 1972.

JORNAL BAHIA ONLINE. Disponivel em: <http://www.jornalbahiaonline.com.br/noticia/19660/bahia_vai_exportar_charuto_para_a_china_para_recuperar_economia_do_reconavo/11>. Acesso em 15 de Fev. de 2015.

KARASAWA, M.; RODRIGUES, R.; SUDRÉ, C. P.; SILVA, M. P.; RIVA, E. M.; AMARAL JUNIOR, A. T. Aplicação de métodos de agrupamento na quantificação da divergência genética entre acessos de tomateiro. **Horticultura Brasileira**, v. 23, n. 4, p.1000-1005, 2005.

KIST, B. B.; SANTOS, C.; REETZ, E.; BELING, R. R.; CORRÊA, S.; RIGON, L.; MÜLLER, I. **Anuário Brasileiro do Fumo.** Santa Cruz do Sul: Gazeta Grupo de Comunicações, p. 160, 2004.

KOHONEN, T. **Self-Organization And Associative Memory**, Springer-Verlag, 2nd Edition, Berlin, Germany, 1988.

KOOP, M. M.; SOUZA, V.Q.; COIMBRA, J.L.M. ; LUZ, V.K. ; MARINI, N. ; OLIVEIRA, A.C. Melhoria da correlação cofenética pela exclusão de unidades experimentais na construção de dendrogramas. **Revista da Faculdade de Zootecnia, Veterinária e Agronomia** (Uruguaiana), v. 14, p. 46-53, 2007.

KOPROWSKI, HILLARY *et al.* 'The green revolution: plants as heterologous expression factors'. ***Vaccine***, vol. 19, pp. 2.735-41, 2001.

LEDO, C. A da S.; GONÇALVES, L.S.A. **Novas abordagens multivariadas em experimentação com fruteiras**. **XXII Congresso Brasileiro de Fruticultura.** Bento Gonçalves, RS, 2012.

LEGG, P.D.; COLLINS, G.B. Inheritance of a short-internode trait in tobacco. **Canadian Journal of Genetics and Cytology**, Ottawa, v. 24, p. 653-659, 1982.

LEGG, P.D.; SMEETON, B.W. Breeding and genetics. In: DAVIS, D.L.; NIELSEN, M. (Ed.). **Tobacco, Production, Chemistry and Technology**. [S.l.:s.n.], p.32-48, 1999.

LEWIS, R.S.; LINGER, L.R.; WOLFF, M.F.; WERNSMAN, E.A. The negative influence of N- mediated TMV resistance on yield in tobacco: linkage drag versus pleiotropy. **Theoretical and Applied Genetics**, New York, v. 115, p. 169-178, 2007.

MAHALANOBIS, P.C. On the generalized distance in statistic. **Proceedings of the National Institute of Sciences of India,** New Delhi, v.2, p.49-55, 1936.

MANLY B.F.J. **Randomization, Bootstrap and Monte Carlo Methods in Biology**. Chapman and Hall, London, p. 399, 1997.

MARDIA, K.L.; KENT, J.T.; BIBBY, J.M. **Multivariate analysis**. London: Academic Press, p.521, 1979.

MARTINELLO, G. E.; LEAL, N. R.; JÚNIOR, A. T. A.; PEREIRA, M. G.; DAHER, R. F. Divergência genética em acessos de quiabeiro com base em marcadores morfológicos. **Horticultura Brasileira,** Brasília, v.20, n.1, p.52-58, 2002.

MARTINS, F.A.; CARNEIRO, P.C.S; SILVA, D.J.H. DA.; CRUZ, C. D.; CARNEIRO, J.E. DE S. **Integração de dados em estudos de diversidade genética de tomateiro.** Pesquisa Agropecuária Brasileira, vol.46, n. 11, p.1496-1502, 2011.

MASSOLA, N.S.; PULCINELLI, C.E.; JESUS, W.C.; GODOY, C.V. Doenças do fumo. In: KIMATI, H.; AMORIM, L.; REZENDE, J.A.M.; BERGAMIN FILHO, A.; CAMARGO, L.E.A. **Manual de fitopatologia**. São Paulo: Agronômica Ceres, v. 2, p. 361-370, 2005.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: Editora UFMG, p. 297, 2005.

MELCHINGER, A.E.; GRANER, A.; SINGH M.; MESSMER, M. M. Relationships among European barley germplasm: I. Genetic diversity among winter and spring cultivars revealed by RFLPs. **Crop Science**, Madison, v. 34, p. 1191-1199, 1994.

MOHAMMADI, S.; PRASANNA, B. Analysis of genetic diversity in crop plants salient statistical tools and considerations. **Crop Science**, v. 43, n. 4, p. 1235-1248, 2003.

NANOCELL NEWS. **O jornal eletrônico do Instituto NANOCELL**. Disponível em: < http://www.institutonanocell.org.br/zmapp-a-esperanca-mundial-contra-o-virus-ebola/ >. Acesso Dez. 2014.

NARAYAN, R.K. Nuclear DNA changes, genoma differentiation and evolution in *Nicotiana* (Solanaceae). **Plant Systematic and Evolution**, Vienna, v.157, p. 161-180, 1987.

OKAMURO, J.; GOLDBERG, B. Tobacco single-copia DNA is highly homologous to sequences present in the genomes of its diploid progenitors. **Molecular General Genetics**, New York, v. 198, p. 290-298, 1985.

OLIVEIRA, J. M. C. de. A cultura do fumo na Bahia: refletindo sobre a convençãoquadro. **Revista Bahia Agrícola**, Salvador, v. 7, n. 2, p. 59-65, 2006.

OLIVEIRA, M. S. P.; AMORIM, E. P.; SANTOS, J. B.; FERREIRA, D. F. Diversidade genética entre acessos de açaizeiro baseada em marcadores RAPD. **Ciencia e Agrotecnologia**, v. 31, n. 6, p. 1645-1653, 2007.

PAULA. R. C. **Repetibilidade e divergência genética entre matrizes de *Pterogyne nitens* Tul. (Fabaceae – Caesalpinioideae) por caracteres**

biométricos de frutos e de sementes e parâmetros da qualidade fisiológica de sementes. 2007. 128 p. Tese (Livre-Docência em Silvicultura) – Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, Jaboticabal, 2007.

PEREIRA, V. A. **Utilização de análise multivariada na caracterização de germoplasma de mandioca (Manihot esculenta Crantz.)**. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, p. 180, 1989.

PEREIRA, F. H. F.; PUIATTI, M.; MIRANDA, G. V.; SILVA, D. J. H.; FINGER, F. L. Divergência genética entre acessos de taro utilizando caracteres qualitativos de inflorescência. **Horticultura Brasileira,** Brasília, v.21, n.3, p.520-524, 2003.

QUEROL, D. **Recursos genéticos, nosso tesouro esquecido**. Tradução Joselita Wasniewski. Rio de Janeiro: ASPTA, p. 206, 1993.

SAAVEDRA G; SPOOR W. Genetic base broadening in autogamous crops: *Lycopersicum esculentum Mill*. As a model. Managing **Plant Genetic Diversity.** 443: p. 291-299, 2002.

SINDITABACO - Sindicato Interestadual da Indústria do Tabaco. Disponível em: < http://sinditabaco.com.br/brasil-exporta-us-25-bilhoes-em-2014/>. Acesso em 05 de Fev. de 2014.

RAMALHO, M. A. P.; SANTOS, J. B.; PINTO, C. A. B. P. **Genética na agropecuária.** Lavras, UFLA, p. 472, 2000.

RAMALHO, M. A. P.; SANTOS, J. B.; PINTO, C. A. B. P. **Genética na agropecuária**. v. 2, Lavras: Editora UFLA, p. 464, 2008.

REN, N.; TIMKO, M.P. AFLP analysis of genetic polymorphism and evolutionary relationships among cultivated and wild *Nicotiana* species. **Genome**, Ottawa, v. 44, n. 4, p. 559-571, 2001.

RIBEIRO, F. E.; SOARES, A. R.; RAMALHO, M. A. P. Divergência genética entre populações de coqueiro-gigante-do-Brasil. **Pesquisa Agropecuária Brasileira**, Brasília, v. 34, n. 9, p. 1615-1622, 1999.

SANTANA, C. M.; MALINOVSKI, J. R. Uso da análise multivariada no estudo de fatores humanos em operadores de motosserra, **Cerne**, v. 8, n. 2, p. 101-107, 2002.

SANTOS, M. **Caracterização fenotípica e molecular de genótipos de fumo no Sul do Brasil.** Dissertação de Mestrado em Fitotecnia, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil. 122p. Abril, 2002.

SARKAR, R. K.; RAO, A. R.; WAHI, S. D.; BHAT, K. V. A comparative performance of clustering procedures for mixture of qualitative and quantitative data – an application to black gram. **Plant Genetic Resources**: Characterization and Utilization 9(4); 523–527, 2011.

SCHATZMAYR, HERMANN. 'Use of plants as vectors for production of biomedical products'. **Virus Reviews & Research**, vol. 7, p. 21-7, 2002.

SEAGRI - Secretaria da Agricultura, Pecuária, Irrigação, Pesca e Aquicultura. Disponivel em: < http://www.seagri.ba.gov.br/noticias/2014/09/22/exporta%C3%A7%C3%A3o-de-fumo-mant%C3%AAm-cultivo-local>. Acesso em. 15 de Dez. 2014.

SIEGMUND, K.D.; LAIRD, P.W.; LAIRD-OFFRINGA, I.A. A comparison of cluster analysis methods using DNA methylation data**. Bioinformatics**, v. 20, n.12, p.1896-1904, 2004.

SINDITABACO - Sindicato Interestadual da Indústria do Tabaco. Disponível em: <http://www.sindifumo.com.br>. Acesso em: 20 Dez, 2014.

SINGH, D. **The relative importance of characters affecting genetic divergence**. The Indian Journal of Genetic and Plant Breeding, New Delhi, v. 41, p. 237-245, 1981.

SMITH, H.H. Differential photoperiod response for interspecific gene transfer. **Journal of Heredity**, Cary, v. 41, p. 198-203, 1950.

SOKAL, R.R.; ROHLF, F.J. The comparison of dendrograms by objective methods. **Taxon**, Berlin, v.11, p. 30-40, 1962.

SPERISEN, C.; RYALS, J.; MEINS, F. Comparison of clones genes provides evidence for intergenomic exchange of DNA in the evolution of a tobacco glucan endo-1.3-ß-glucosidase gene family. **Proceeding of National Academy of Sciences USA,** Washington, v. 88, p. 1820-1824, 1991.

STRAPASSON, E. **Seleção de descritores na caracterização de germoplasma de Paspalum através de componentes principais**. (Mestrado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba. p. 95, 1997.

STURGES, H.A. The choice of a class interval. **Journal of the American Statistical Association**, v. 21, p. 65-66, 1926.

SUDRÉ, C.P.; CRUZ, C.D.; RODRIGUES, R.; RIVA, E. M.; AMARAL JÚNIOR, A. T. do.; SILVA, D.J. H. da.; PEREIRA, T.N.S.P. Variáveis multicategóricas na determinação da divergência genética entre acessos de pimenta e pimentão. **Horticultura Brasileira**, v. 24, n. 01, p. 88-93, 2006.

SUDRÉ, C.P.; GONÇALVES, L.S.A.; RODRIGUES, R.; AMARAL JÚNIOR, A.T.; RIVA-SOUZA, E.M.; BENTO, C.S. Genetic variability in domesticated *Capsicum* spp as assessed by morphological and agronomic data in mixed statistical analysis. **Genetics and Molecular Research**, v. 9, n. 1, p. 283-294, 2010.

TOBACCO JOURNAL INTERNATIONAL. **Tobacco Encyclopedia**. Mainz: Ed. Voges, p. 279, 2000.

UPOV (*Union pour la Protection des Obtentions Variétales*). Disponível em http://www.upov.int/tabaco. Acesso em: 20 dez. 2014.

VIEIRA, P. C.; FERNANDES, J. B.; ANDREI, C. C. Plantas inseticidas. In: SIMÕES, C. M. O.; SCHENKEL, E. P.; GOSMANN, G.; MELLO, J. C. P.; MENTZ, L. A.; PETROVICK, P. R. (Org.). **Farmacognosia: da planta ao medicamento**. 5.ed. Porto Alegre: Editora da Universidade/ UFRGS /Florianópolis: p. 903-918, 2003.

VILLELA, O. T. **Diversidade fenotípica e molecular de cultivares brasileiras de soja portadoras de gene RR.** Dissertação de Mestrado em Agronomia, Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, SP, Brasil. Abril, 2013.

WARD Jr., J.H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, p. 236-244, 1963.

WEEDEN, N.F.; TIMMERMAN, G.M.; LU, J. Identifying mapping genes of economic significance. **Euphytica**, Wageningen, v. 73, p. 191-198, 1994.

WEHLBURG, A.F. Cigars and Cigarellos. In D.L. Davis and M.T. Nielsen, eds. Tobacco: Production, Chemistry and Technology. **Blackwell Science** Publication. p. 440-451, 1999.

WILCHES, M. O. Evaluación de treinta y cuatro variedades de mani mediante técnicas multivariadas. **Revista ICA**, v. 18, n.1, p. 67-76, 1983.

WINSBERG, S.; RAMSAY, J.O. Monotone spline transformations for dimension reduction. **Psychometrika**, v. 48, p. 575–595, 1983.

ZEWDIE, Y.; TONG, N.; BOSLAND, P. Establishing a core collection of Capsicum using a cluster analysis with enlightened selection of accessions. **Genetic Resources and Crop Evolution**, v. 51, n. 2, p. 147-151, 2004.

CAPÍTULO I

SELEÇÃO DE DESCRITORES MORFOAGRONÔMICOS EM ACESSOS DE TABACO POR MEIO DE TÉCNICAS MULTIVARIADAS

---

[1]Artigo a ser submetido ao periódico Genetics and Molecular Research

SELEÇÃO DE DESCRITORES MORFOAGRONÔMICOS EM ACESSOS DE TABACO POR MEIO DE TÉCNICAS MULTIVARIADAS

Autor: Antonio Leandro da Silva Conceição
Orientador: Carlos Alberto da Silva Ledo
Co-orientador: Ricardo Franco Cunha Moreira

RESUMO: O objetivo deste trabalho foi selecionar descritores morfoagronômicos e determinar sua importância relativa na caracterização, bem como propor um número mínimo capaz de quantificar a divergência entre acessos de tabaco no intuito de fornecer maior conhecimento da variabilidade genética para conservação e melhoramento genético da espécie. Para a caracterização de 15 acessos de tabaco tipo Sumatra foram usados 43 descritores morfoagronômicos, sendo 17 quantitativos e 26 qualitativos. Foi realizada a identificação dos descritores quantitativos redundantes, por dois procedimentos: 1) seleção direta, proposta por Jolliffe e 2) Seleção baseada no coeficiente de Singh. Para auxiliar na decisão de descarte, foram estimados os coeficientes de correlação de Spearman entre todos os descritores. A seleção dos descritores qualitativos foi realizada por meio do nível de entropia dos caracteres (H), proposto por Renyi. Foram selecionados 10 descritores quantitativos: altura da planta, número de folhas, diâmetro médio do caule, Comprimento da 3ª folha, comprimento da 10ª folha, largura da base da 10ª folha, ângulo de inserção da 10ª folha, comprimento dos internódios, engrossamento do tubo da flor e comprimento da corola e 8 descritores qualitativos: Coloração do caule, coloração das folhas, coloração da nervura central, face inferior, superfície da lâmina foliar, perfil longitudinal da folha, margem da lâmina foliar: 10ª folha, ponta da lâmina foliar: 10ª folha e cor da corola. Todos esses descritores selecionados são importantes na caracterização do germoplasma de tabaco em estudo. O descarte de 58% dos descritores não provocou perda de informação considerável, uma vez que os descritores redundantes estão correlacionados aos descritores remanescentes. Onde esse descarte também possibilitará a redução de custos e dará mais dinâmica ao manejo e caracterização da cultura.

Palavras-chave: Variabilidade, *Nicotiana tabacum* L., descarte de descritores

SELECTION OF DESCRIPTORS A MORPHOLOGICAL CHARACTERISTICS CONSIDERED IN ACCESS OF TOBACCO BY MEANS OF MULTIVARIATE TECHNIQUES

Author: Antonio Leandro da silva Conceição
Advisor: Carlos Alberto da Silva Ledo
Co-advisor: Ricardo Franco Cunha Moreira

ABSTRACT: The objective of this work was to select descriptors morphological and agricultural traits and determine their relative importance in the characterization, as well as propose a minimum number to quantify the divergence among accessions of tobacco in order to provide greater knowledge of genetic variability for conservation and genetic improvement of the species. For the characterization of accessions of tobacco type Sumatra were used 43 descriptors morphological and agricultural traits, being 17 quantitative and 26 qualitative. Was the identification of quantitative descriptors redundant, by two procedures: 1) direct selection, proposal by Jolliffe and 2) Selection based on the coefficient of Singh. To assist in the decision to discard, were estimated the Spearman's correlation coefficients between all descriptors. The selection of qualitative descriptors was performed by means of the entropy level of the characters (H), proposed by Renyi. 10 Were selected quantitative descriptors: plant height, number of leaves, average diameter of stem, Length of 3RD leaf length of 10TH leaf, the base width of 10TH leaf, insertion angle of 10TH leaf length of internodes, tube thickening and the flower corolla length and 8 qualitative descriptors: Coloring of the stem, staining the sheets, coloring of the midrib, lower face, surface of leaf, longitudinal profile of leaf margin leaf: 10TH sheet, tip of leaf blade: 10TH leaf and color of the corolla. All these selected descriptors are important in the characterization of germplasm of tobacco in study. The disposal of 58% of the descriptors has not lost considerable information, since the descriptors are correlated with the remaining descriptors. Where the disposal also will enable the reduction of costs and give more dynamic management and characterization of culture.

Key words: Variability, *Nicotiana tabacum* L., descriptors disposal

INTRODUÇÃO

O tabaco é atualmente a cultura não alimentícia mais importante da agricultura mundial. O Brasil é o 2º maior produtor, sendo que cerca de 85% dessa produção é exportada (ADAPAR, 2014). A Bahia ocupa a quinta posição do ranking brasileiro, atrás apenas do Rio Grande do Sul, Santa Catarina, Paraná e Alagoas. A região do Recôncavo Baiano agrega ótimas condições para a cultura do fumo. O método de produção é de agricultura familiar, com propriedades variando entre 0,7 e 1 hectare. Já a cadeia produtiva do charuto emprega 14 mil pessoas no Recôncavo, a maioria agricultores familiares, sendo que 90% destes são mulheres que aprenderam o ofício com suas mães e avós, repassando as filhas desde 1842 – período em que foi fundada a primeira fábrica, a Juventude, no município de São Félix (SEAGRI, 2014).

De todas as espécies do gênero Nicotiana, a *N. tabacum* é a mais cultivadas no mundo. O gênero é nomeado em homenagem a Jean Nicot, que em 1561 foi o primeiro a apresentar o tabaco para o francês Royal Court. Nicotiana pertence à família das solanáceas e subdivide-se em três subgenêros: Rustica, Tabaccum e Petunioides (GOODSPEED, 1954; REN & TIMKO, 2001).

Além do tabaco ser destinado a indústria fumageira na produção de cigarros, charutos e etc. O tabaco tem se tornado realidade em alguns países por seu potencial para uso medicinal, na produção de produtos biofarmacêuticos como vacinas, hormônios, anticorpos e insulina (BLINDER et al., 2007).

O uso de técnicas multivariadas é um dos fatores, que tem impulsionado os estudos sobre diversidade genética entre genótipos (LEDO et al., 2009). As análises multivariadas são ferramentas úteis para a identificação de descritores com maior conteúdo informativo para caracterização de germoplasma e para melhoramento genético, uma vez que fornece informações para eliminar características que contribuem pouco para variação total (CRUZ et al., 2004).

No geral, todo caráter deve apresentar uma parcela de contribuição na variação do germoplasma analisado. Mas há uma tendência de que o aumento do número de descritores avaliados ocasione a presença de caracteres redundantes, posto que esses caracteres quase sempre estão associados a outros (DAHER, 1993). Logo, a eliminação dos redundantes seria uma decisão vantajosa, pois

reduziria o trabalho de tomada de dados sem ocasionar perda na precisão da caracterização, especialmente se esses caracteres forem de difícil mensuração e apresentarem baixa variabilidade e estabilidade de expressão (PEREIRA, 1989).

A redução no número de descritores morfoagronômicos é relatada em vários estudos, os quais utilizaram diversas técnicas de descarte, com o objetivo de otimizar o trabalho de coleta dos dados e identificar os descritores com maior contribuição na divergência genética. Para (OLIVEIRA, 2005), o açaí (*Euterpe oleracea* Mart.), usam-se apenas quatro descritores. Já para a pupunha (*Bactris gasipaes* Kunth), é necessário utilizar 10 descritores morfológicos para discriminar as raças desta espécie (MARTEL et al., 2003). Segundo Afonso et al., 2014, em mandioca (*Manihot esculenta*), dos 35 descritores morfoagronômicos analisados apenas 43% destes foram suficientes para discriminar os acessos estudados. Oliveira et al., (2014), também analisando acessos de mandioca (*Manihot esculenta*) utilizando-se de 51 descritores morfoagronômicos, constatou que 32 deles foram suficientes em razão de sua alta capacidade para discriminar o germoplasma de mandioca e de sua habilidade de manter alguns caracteres agronômicos preliminares, úteis para a caracterização inicial do germoplasma.

Portanto, o objetivo deste trabalho foi selecionar descritores morfoagronômicos e determinar sua importância relativa na caracterização de acessos de tabaco, assim como verificar a associação entre os descritores descartados e os remanescentes.

MATERIAL E MÉTODOS

Foram avaliados 15 genótipos de tabaco da espécie *Nicotiana tabacum* L., tipo Sumatra (Tabela 1). Esses genótipos são provenientes da empresa ERMOR TABARAMA TABACOS DO BRASIL Ltda., localizada no município de Cruz das Almas – BA. Estes materiais são e formam a base de materiais (sementes) fornecidos aos agricultores da região do Recôncavo da Bahia.

O campo de produção da empresa ERMOR onde foi implantado o experimento localiza-se no distrito de São José do Itaporã, que faz divisa com o município de Cruz das Almas-BA. A área experimental possui as seguintes

características: altitude de 208m acima do nível do mar, clima Aw a Am, tropical quente e úmido, segundo a classificação de Köppen. A pluviosidade média anual é de 1220 mm, com maior incidência de chuvas no período compreendido entre março e junho. A umidade relativa do ar é de aproximadamente 80% e a temperatura média anual é de 24,1ºC.

O delineamento experimental utilizado foi o de blocos casualizados com quatro repetições. Cada parcela foi constituída de cinco linhas de 10 plantas e cada linha com 4,5 metros de comprimento, com espaçamento de 1,0 metros entre linhas e 0,42 metros entre plantas.

Os caracteres analisados foram definidos conforme o SINDIFUMO (Subcomissão de Sementes), Ministério da Agricultura, Pecuária e Abastecimento e com base na descrição recomendada pela UPOV e Legislações Americana e Italiana. Foram avaliadas 17 variáveis quantitativas (Tabela 2) e 26 variáveis Qualitativas (Tabela 3). (MAPA, 2015)

Tabela 1. Relação dos acessos de Tabaco provenientes da empresa Ermor Tabarama Tabacos do Brasil Ltda, utilizados no estudo. UFRB, Cruz das Almas, BA. 2014.

Código (Nº)	Acessos	Tipo	Procedência
A1	ER 03-107	Sumatra	ERMOR TABARAMA
A2	ER 04-090	Sumatra	ERMOR TABARAMA
A3	ER 04-095	Sumatra	ERMOR TABARAMA
A4	ER 05-005	Sumatra	ERMOR TABARAMA
A5	ER 05-070	Sumatra	ERMOR TABARAMA
A6	ER 12-040	Sumatra	ERMOR TABARAMA
A7	ER 13-061	Sumatra	ERMOR TABARAMA
A8	ER 13-065	Sumatra	ERMOR TABARAMA
A9	ER 28-027	Sumatra	ERMOR TABARAMA
A10	ER 33-021	Sumatra	ERMOR TABARAMA
A11	ER 33-022	Sumatra	ERMOR TABARAMA
A12	ER 33-023	Sumatra	ERMOR TABARAMA
A13	109 PD	Sumatra	ERMOR TABARAMA
A14	125 PD	Sumatra	ERMOR TABARAMA
A15	221 PD	Sumatra	ERMOR TABARAMA

Para mensuração dos dados foram utilizadas réguas de comprimento de 20 e 60 centímetros, régua de mira de 3 m, paquímetro digital 10 mm e um transferidor para aferir o ângulo de inserção da 10ª folha no caule. Para estimar os dados de produção as folhas das plantas foram colhidas, curadas e fermentadas e mantidas a uma umidade de 28%. Após esse processo foi pesado e obtido o peso seco das folhas de 600 plantas, estendendo para um total de 28.000 planta/hectare dividido pelo quociente de plantas colhidas por repetição.

Análises estatísticas

Seleção dos descritores quantitativos

Os dados obtidos a partir dos descritores quantitativos inicialmente foram submetidos a estatísticas descritivas, onde foram calculados: valores mínimos e máximos, média, desvio padrão e coeficiente de variação. Foi realizado também o teste de Normalidade de Shapiro–Wilk. Essas análises foram realizadas com o auxílio do programa estatístico SAS – Statistical Analysis System (SAS Institute Inc, 2004).

Para auxiliar na decisão de descarte, foram estimados os coeficientes de correlação de Spearman entre todos os descritores, para verificar a associação entre os descritores descartados e os remanescentes. A significância do coeficiente de correlação foi verificada pelo teste de t, com auxílio do SAS – Statistical Analysis System (SAS Institute Inc, 2004).

A identificação dos descritores redundantes foi realizada por dois procedimentos: 1) seleção direta, proposta por Jolliffe (1972, 1973), sendo indicado para descarte todo descritor que apresentou maior coeficiente de ponderação em valor absoluto (autovetor), na variável canônica de autovalor menor, partindo-se da última variável até aquela que seu autovalor não excedeu 0,70 e 2) Seleção baseada no coeficiente de Singh (1981), levando-se em consideração a contribuição relativa de cada característica para a divergência genética. Ambas as análises foram realizadas por meio do programa computacional GENES (CRUZ, 2014).

O descarte definitivo dos caracteres foi efetuado levando-se em consideração as informações coincidentes nos dois métodos, eliminando-se os caracteres sugeridos como redundantes em ambos os procedimentos.

Tabela 2. Relação das variáveis quantitativas de 15 acessos de tabaco estudados. Cruz das Almas, BA. 2014.

Variáveis quantitativas	**Medida expressa em**
Rendimento (REND)	$Kgha^{-1}$
Dias do transplante ao florescimento (DTF)	Dias
Altura total da planta (ALT)	cm
Nº de folhas (NF)	-
Diâmetro médio do caule (DMC)	cm
Índice cilíndrico (IC) =quociente entre diâmetro médio e base da inflorescência	-
Largura da 3º folha (LF3)	cm
Comprimento da 3º folha (CF3)	cm
Largura da 10º folha (LF10)	cm
Comprimento da 10º folha (CF10)	cm
Largura da base da 10º folha (LB10)	cm
Ângulo de inserção 10º folha (ANG10)	-
Comprimento dos internódios (CINT)	cm
Comprimento da flor (CFRL)	cm
Diâmetro do tubo da flor (DTFRL)	mm
Engrossamento do tubo da flor (ETFRL)	mm
Comprimento da corola (CCFLR)	cm

Seleção dos descritores qualitativos

A seleção dos descritores qualitativos foi realizada por meio do nível de entropia dos caracteres (H), proposto por Renyi (1961), de acordo com o seguinte modelo:

$$H = -\sum_{i=1}^{s} p_i \ln p_i$$

Onde a Entropia é uma medida da frequência da distribuição de (n) acessos P = (p_1, p_2 ... p_s), sendo: $p_1 = f_i/n$ e ($p_1 + p_2 + ... + p_s = 1$) desde que ($n = f_1 + f_2 + ... + f_s$), onde f_1, f_2, ... f_n, são as contagens de cada uma das classes (s) no descritor considerado. Esta estimativa da entropia foi realizada com o auxílio do programa SAS – Statistical Analysis System (SAS Institute, 2004). Logo abaixo, na Tabela 3 são relacionados os descritores qualitativos utilizados no estudo.

Tabela 3. Relação das variáveis qualitativas de 15 acessos de tabaco estudados. Cruz das Almas, BA. 2014.

DESCRITORES	SIGLAS	CLASSES
1-Forma da planta	FP	(1) Cônica; (2) Cilíndrica; (3) Elíptica; (4) Cônica invertida
2-Coloração do caule: início do florescimento	CCF	(1) Verde-esbranquiçada; (2) V.clara; (3) V.média; (4) V.escura
3-Presença de brotos, início do florescimento	PB	(1) Ausente ou Muito fraca; (3) Fraca; (5) Média; (7) Forte; (9) Muito forte
4-Tipo de folha	TF	(1) Séssil; (2) Peciolada
5-Forma das folhas medianas centrais (10ª a 15ª Folhas): início do florescimento	FFMC	(1) Lanceolada; (2) Estreito-eliptica; (3) Largo-eliptica; (4) Ovalada; (5) Obovada; (6) Cordiforme; (7) Arredondada
6-Coloração das folhas: 10ª folha, início do florescimento	CF	(1) Verde-amarelada; (2) V.esbranquiçada; (3) V.clara; (4) V.médio; (5) V.escura
7-Coloração da nervura central, face inferior	CNC	(1) Esbranquiçada; (2) Verde-esbranquiçada; (3) Verde
8-Curva da ponta da lâmina foliar: 10ª folha, início do	CPLF	(1) Direto; (2) Curvado para baixo; (3) Curvado para cima

florescimento

9-Superfície da lâmina foliar: 10ª folha, início do florescimento	SLF	(3) fraco; (5) médio; (7) forte
10-Perfil transversal da 10ª folha: início do florescimento	PTF	(1) Côncava; (2) Plana; (3) Convexa
11-Perfil longitudinal da folha	PLF	(1) Reto; (3) Ligeiramente recurvado; (5) Moderadamente recurvado; (7) Fortemente Recurvado
12-Ângulo nervuras laterais em relação à nervura central 10ª folha, início do florescimento	ANGL	(1) Muito agudo; (2) Medianamente agudo; (3) Reto
13-Margem lâmina foliar: 10ª folha, início do florescimento;	MLF	(1) Ausente ou muito fraca; (3) Fraca; (5) Média; (7) Forte
14-Ponta da lâmina foliar: 10ª folha, início do florescimento	PLAMF	(1) Obtusa; (3) Ligeiramente pontiaguda; (5) Medianamente Pontiaguda; (7) Fort. pontiaguda; (9) Extremamente Pontiaguda
15-Formato da aurícula: 10ª folha, início do florescimento	FA	(1) Ausente ou muito fraco; (3) Fraco; (5) Médio; (7) Forte; (9) Muito forte
16-Tipo de flor: presença ou ausência de anteras	TFLOR	(1) Presença; (2) Ausência
17-Cor da corola: início do florescimento	CCOR	(1) Branca; (2) Rosa-clara; (3) Rosa-média; (4) Rosa-forte; (5) Vermelha
18-Formato do limbo da corola: forma da seção vista do alto	FLC	1) Arredondada; (2) Poligonal; (3) Poligonoestelar; (4) Estelar; (5) C/ sépalas muito pronunciadas
19-Desenvolvimento dos estames	DESEN-E	1) Nenhum ou rudimentar; (2) Pleno
20-Desenvolvimento do pistilo em relação aos estames	DESEN-P	(1) Mais curto; (2) Mesmo comprimento; (3) Mais largo
21- Expressão dos ápices da corola	EAC	(1) Ausente ou muito fraca; (3) Fraca; (5) Média; (7) Forte; (9) Muito forte

22-Forma da inflorescência: pleno florescimento, 20 a 30 frutos formados	FI	(1) Esférica; (2) Esférica-aplanada; (3) Cônica-invertida; (4) Cônica-dupla
23-Posição da inflorescência em relação as folhas superiores	PI	(1) Entre as folhas; (2) Acima
24-Densidade da inflorescência; média 20 a 30 frutos maduros	DI	(3) Esparsa; (5) Média; (7) Densa
25-Tipo de deiscência do fruto: na maturação completa	TDF	(1) Não deiscente; (2) Cápsula deiscente
26-Formato do fruto: média de 20 a 30 frutos formados	FF	(1) Arredondada; (2) Alongada; (3) Elíptica

A entropia de um descritor qualquer será tão maior quanto maior for o número de classes fenotípicas desse e quanto mais homogêneo for o balanço entre a frequência dos acessos nas diferentes classes fenotípicas (VIERA et al., 2007). Neste trabalho valores baixos de H associados a > 50% de frequência de acessos em uma determinada classe foram utilizados como critério de descarte do descritor.

RESULTADOS E DISCUSSÃO

Na Tabela 4 estão apresentadas as estatísticas descritivas dos descritores quantitativos. Nesta pode-se observar a amplitude dos valores apresentados para as variáveis estudadas. De acordo com os resultados obtidos, observou-se que os coeficientes de variação oscilaram de 4,34 a 34,40, para número de dias do transplante ao florescimento (DTF) e largura da base da 10ª folha (LB10) respectivamente. Esses resultados vêm a corroborar com os resultados obtidos por Costa (2012) em estudo de diversidade entre acessos de *Nicotiana tabacum* L., onde obteve-se também o menor e o maior coeficiente de variação para as mesmas variáveis. O desvio padrão de 137,41 da variável produção permitiu estimar uma variação em relação à média entre os genótipos 109 PD e ER 13-

065, responsáveis pelo menor e maior rendimento, respectivamente.

Observou-se ainda que, para o teste de normalidade, os resultados indicam que metade das variáveis foram significativas pelo teste de Shapiro-Wilks a 5% de significância, sendo assim, essas variáveis não seguem distribuição normal, portanto calculou-se a correlação de Spearman (Tabela 5) para medir a relação entre as variáveis e para auxiliar na decisão de descarte, verificando a associação entre descritores descartados e os remanescentes.

Tabela 4. Estatística descritiva e teste de normalidade das variáveis quantitativas estudadas. Cruz das Almas, BA. 2014.

Variáveis	Mínimo	Máximo	Média	Desvio Padrão	CV (%)	Teste de Normalidade
REND	420,00	1085,00	794,62	137,41	17,29	0,98^{ns}
DTF	74,00	87,00	81,87	3,55	4,34	0,91*
ALT	189,30	272,60	241,13	22,93	9,51	0,85*
NF	16,60	33,70	25,87	4,08	15,76	0,97^{ns}
DMC	1,63	2,88	2,15	0,30	14,00	0,94*
IC	1,39	3,73	2,24	0,39	17,34	0,97^{ns}
LF3	21,80	28,50	25,09	1,68	6,69	0,93*
CF3	37,75	51,95	43,44	3,66	8,43	0,97^{ns}
LF10	24,25	32,70	29,25	1,77	6,04	0,96^{ns}
CF10	46,90	62,50	52,91	2,65	5,00	0,92*
LB10	4,03	13,90	6,79	2,34	34,40	0,74*
ANG10	32,00	54,00	43,33	5,40	12,45	0,98^{ns}
CINT	5,63	9,85	8,23	1,02	12,39	0,94*
CFOR	4,32	8,26	5,01	0,56	11,14	0,72*
DTFLR	3,42	5,72	4,45	0,40	8,96	0,97^{ns}
ETFLR	7,73	11,14	9,18	0,83	9,01	0,98^{ns}
CCFLR	2,25	3,25	2,75	0,17	6,31	0,98^{ns}

^{ns} não significativo pelo teste de Shapiro-Wilks a 5% de significância. Rendimento - produção (REND), dias do transplante ao florescimento (DTF); altura da planta (ALT); número de folhas (NF); diâmetro médio do caule (DMC); índice cilíndrico (IC); largura da 3ª folha (LF3); Comprimento da 3ª folha (CF3); largura da 10ª folha (LF10); comprimento da 10ª folha (CF10); largura da base da 10ª folha (LB10); ângulo de inserção da 10ª folha (ANG10); comprimento dos internódios (CINT); comprimento da flor (CFLR); diâmetro da flor (DTFLR); engrossamento tubo da flor (ETFLR); comprimento da corola (CCFLR); coeficiente de variação (CV).

As correlações fenotípicas determinadas entre caracteres são atribuídas a fatores genéticos e ambientais (VENCOVSKY & BARRIGA, 1992) e estimadas

com o propósito de mensurar a alteração em um caráter quando se altera outro. Nesse estudo as estimativas dos coeficientes de correlação de Spearman entre os caracteres estudados são apresentadas na tabela 5, em que foi observado correlação positiva e altamente significativa, porém com magnitude média entre os caracteres número de folhas e altura total da planta (r_S= 0,49**). Esse resultado vem a corroborar com o comportamento esperado e também citado por Santos (2002), em estudo de caracterização fenotípica e molecular de genótipos de *Nicotiana tabacum* L., onde os genótipos com maior número de folhas por planta foram os de maior estatura. Essa mesma autora enfatiza que, o tipo moderno de planta não deve apresentar estatura muito elevada, para evitar acamamento.

Constatou-se também correlação positiva e altamente significativa entre a largura e comprimento da 3ª folha (r_S= 0,73**) e entre o diâmetro médio do caule e altura total da planta (r_S= 0,55**). Pode-se notar que os acessos que obtiveram menores médias em relação à altura total da planta, também foram os que apresentaram os menores valores em relação ao número de folhas, sendo esses caracteres altamente correlacionados (Tabela 5).

O comprimento de internódios apresentou correlação negativa altamente significativa com número de folhas (r_S= -0,70**), observou-se que plantas com menores médias em relação ao comprimento dos internódios apresentaram maior número de folhas (Tabela 5).

Observou-se que a variável largura da base da 10ª folha apresentou correlação negativa e altamente significativa com a altura total da planta (r_S= -0,75**), com o número de folhas (r_S= - 0,61**) e com o diâmetro médio do caule (r_S= -0,45**), onde as plantas com menor altura foram as que apresentaram maior comprimento para largura da base da 10ª folha (Tabela 5). O elevado número de plantas mensuradas sugere que as estimativas de r_s apresentam elevada precisão, e com isso, associações de baixa magnitude são significantes. É importante examinar, além da significância estatística, a magnitude do r, que fornece uma significância prática de determinada associação linear (HAIR et al., 2005).

Tabela 5. Estimativas dos coeficientes de correlação de Spearman, com as suas respectivas significâncias, entre as 17 variáveis, Cruz das Almas - BA. 2014.

	REND	DTF	ALT	NF	DMC	IC	LF3	CF3	LF10	CF10	LB10	ANG10	CINT	CFOR	DTFLR	ETFLR
DTF	0,13^{ns}															
ALT	0,28*	0,10^{ns}														
NF	0,33**	0,00^{ns}	0,49**													
DMC	0,36**	0,30*	0,55**	0,26*												
IC	0,27*	0,01^{ns}	0,22^{ns}	0,52**	0,53**											
LF3	0,08^{ns}	0,34**	-0,03^{ns}	-0,31*	0,11^{ns}	-0,04^{ns}										
CF3	-0,09^{ns}	0,29*	-0,24^{ns}	-0,34**	-0,20^{ns}	-0,14^{ns}	0,73**									
LF10	0,39**	0,36**	0,52**	0,05^{ns}	0,63**	-0,01^{ns}	0,24^{ns}	-0,11^{ns}								
CF10	0,28*	0,37**	0,23^{ns}	0,26*	0,17^{ns}	0,01^{ns}	0,02^{ns}	0,18^{ns}	0,27*							
LB10	-0,24^{ns}	0,09^{ns}	-0,75**	-0,61**	-0,45**	-0,32**	0,16^{ns}	0,34**	-0,33**	-0,01^{ns}						
ANG10	-0,04^{ns}	-0,04^{ns}	0,26*	0,25*	0,05^{ns}	0,11^{ns}	-0,28*	-0,36**	-0,10^{ns}	-0,08^{ns}	-0,30*					
CINT	-0,31*	-0,27*	-0,23^{ns}	-0,70**	-0,18^{ns}	-0,28*	0,09^{ns}	0,13^{ns}	-0,10^{ns}	-0,29*	0,35**	-0,05^{ns}				
CFLR	-0,05^{ns}	0,53**	-0,10^{ns}	-0,02^{ns}	-0,13^{ns}	-0,09^{ns}	0,19^{ns}	0,48**	-0,04^{ns}	0,38**	0,22^{ns}	-0,17^{ns}	-0,16^{ns}			
DTFLR	-0,25^{ns}	0,16^{ns}	-0,06^{ns}	-0,14^{ns}	-0,29*	-0,31*	0,00^{ns}	0,00^{ns}	-0,03^{ns}	0,06^{ns}	0,12^{ns}	-0,15^{ns}	0,03^{ns}	0,21^{ns}		
ETFLR	-0,37**	0,30*	-0,16^{ns}	-0,40**	-0,02^{ns}	-0,24^{ns}	0,04^{ns}	0,11^{ns}	-0,03^{ns}	-0,20^{ns}	0,18^{ns}	-0,11^{ns}	0,28*	0,37**	0,50**	
CCFLR	-0,24^{ns}	0,11^{ns}	-0,43**	-0,54**	-0,25*	-0,33**	0,30*	0,30*	-0,18^{ns}	-0,16^{ns}	0,57**	-0,08^{ns}	0,35**	0,16^{ns}	0,05^{ns}	0,30*

^{ns} = não significativo; *significativo ao nível de 5% de probabilidade ($p < 0,05$); e **significativo ao nível de 1% de probabilidade ($p < 0,01$), pelo teste t.
Rendimento - produção (REND), dias do transplante ao florescimento (DTF); altura da planta (ALT); número de folhas (NF); diâmetro médio do caule (DMC); índice cilíndrico (IC); largura da 3ª folha (LF3); Comprimento da 3ª folha (CF3); largura da 10ª folha (LF10); comprimento da 10ª folha (CF10); largura da base da 10ª folha (LB10); ângulo de inserção da 10ª folha (ANG10); comprimento dos internódios (CINT); comprimento da flor (CFLR); diâmetro da flor (DTFLR); engrossamento tubo da flor (ETFLR); comprimento da corola (CCFLR).

Os resultados da seleção baseada no coeficiente de Singh (1981), levando-se em consideração a contribuição relativa de cada característica para a divergência genética podem ser observados na figura 1.

Segundo o método de Singh (1981), as características que proporcionaram maiores contribuições relativas foram a largura da base da 10ª folha (LB 10ª), com 44,73% de contribuição quanto à diversidade genética dos acessos, sendo esta responsável pela maior percentagem de toda variabilidade dos dados, seguida pelo variável comprimento dos internódios (CINT), com 9,91%, comprimento da 10ª folha (CF10) com 7,69% e altura da planta (ALT), com 7,39%, essas quatro características contribuíram com 68,72% da distribuição total (Figura 1). Este resultado é similar ao observado por Costa (2012) e por Conceição et al. (2014) em seus estudos com *Nicotiana tabacum* L., onde também foi encontrado uma maior contribuição relativa para diversidade na largura da base da folha. Estas informações evidenciam que essa característica possui grande importância na diferenciação de acessos de tabaco, sendo de grande importância em estudos de diversidade genética da cultura.

A variável largura da base da 10ª folha (LB 10ª) se mostrou uma variável de grande importância na caracterização dos acessos em estudo, com isso, se faz necessário mais estudos acerca dessa variável, que possivelmente irá contribuir para melhores resultados nos estudos de divergência genética da espécie.

Já as variáveis com menor contribuição foram: largura da 3ª folha (LF3), com 0,24%, diâmetro da flor (DTFLR), 0,33%, comprimento da flor (CFLR), com 0,51% e índice cilíndrico (IC), com 0,53% (Figura 1). No estudo realizado por Costa (2012) e por Conceição et al. (2014), também foi observado uma menor contribuição relativa por parte de algumas dessas características citadas no presente trabalho, com isso, pode-se inferir que essas variáveis de menor contribuição são pouco informativas na caracterização da variabilidade genética existente. Sendo assim, algumas dessas variáveis podem ser descartadas, pois de acordo com Rêgo et al. (2003) caracteres que contribuíram com um percentual muito baixo ou que não contribuíram para a variabilidade detectada podem ser descartadas. Como critério para essa análise, variáveis com contribuição relativa inferior a 3,5% foram descartas. Sendo assim, sete descritores, ou seja, 41,17% dos descritores foram descartados nesse processo (Figura 1 e Tabela 7).

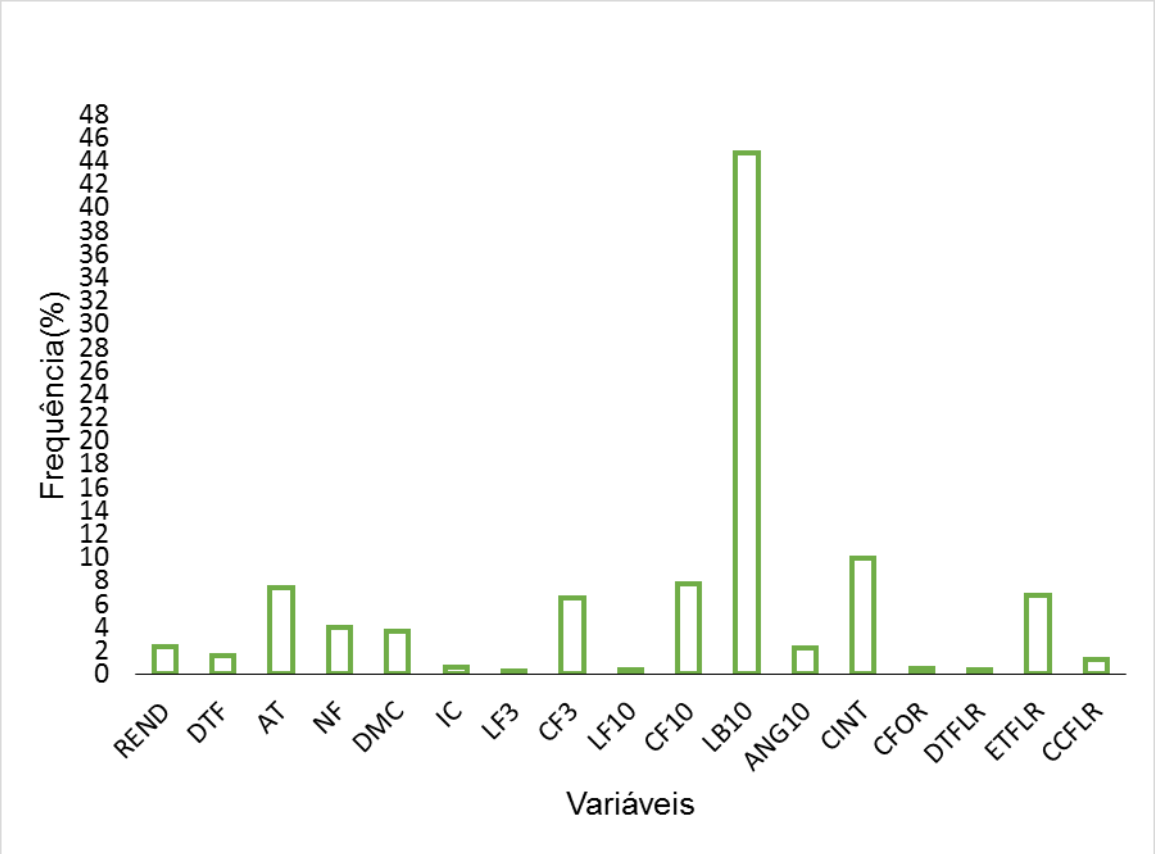

Figura 1. Contribuição relativa dos caracteres para diversidade segundo Singh (1981). Cruz das Almas, 2014.

Rendimento - produção (REND), dias do transplante ao florescimento (DTF); altura da planta (ALT); número de folhas (NF); diâmetro médio do caule (DMC); índice cilíndrico (IC); largura da 3ª folha (LF3); Comprimento da 3ª folha (CF3); largura da 10ª folha (LF10); comprimento da 10ª folha (CF10); largura da base da 10ª folha (LB10); ângulo de inserção da 10ª folha (ANG10); comprimento dos internódios (CINT); comprimento da flor (CFLR); diâmetro da flor (DTFLR); engrossamento tubo da flor (ETFLR); comprimento da corola (CCFLR).

Os resultados da seleção direta, proposta por Jolliffe (1972, 1973), estão apresentados nas tabelas 6 e 7.

Ao analisar as estimativas dos autovalores associados as variáveis canônicas e suas respectivas variâncias relativas e acumuladas obtidas para os 17 caracteres morfológicos quantitativos, percebe-se que as duas primeiras variáveis canônicas conseguiram explicar 90,21% da variação total acumulada, com a primeira variável explicando 65,48% e a segunda 24,73%. Onde a variância acumulada foi concentrada até a 7ª variável canônica, que respondeu por 98,70 % de toda a variação disponível (Tabela 6). Segundo Pereira et al.

(1992), estudando seleção de descritores em mandioca, a distribuição da variância está associada ao número de descritores utilizados na análise, estando concentrada nas primeiras variáveis canônicas, apenas quando se utiliza um número reduzido de descritores de interesse agronômico ou que estejam em um mesmo grupo.

A análise de variáveis canônicas permite realizar o descarte de descritores morfológicos, eliminando, desta forma, aqueles que oferecem pouca importância (contribuição) no estudo de divergência (NEGREIROS et al., 2008).

Utilizando-se da metodologia proposta por Jolliffe (1972, 1973), que estabelece a eliminação dos descritores de maior peso nas variáveis canônicas cuja associação entre autovetores e autovalores são inferiores a 0,7, foram descartados dez descritores quantitativos na seguinte ordem: DMC, CF3, IC, LF10, DTF, REND, CFLR, DTFLR, LF3 e NF (Tabela 7). Essas exclusões indicam que 59% dos descritores morfológicos são desnecessários para a caracterização dos acessos de tabaco em estudo.

Ao analisar o descarte preliminar, para os descritores quantitativos, com o uso das estimativas dos coeficientes de ponderação associados as variáveis canônicas de autovetores, verificou-se que o primeiro caráter indicado foi o diâmetro médio do caule (DMC), uma vez que apresentou o maior peso no módulo com a última variável canônica (-0,81), seguido pelos caracteres (CF3), (IC), (LF10) e (DPF) cujos maiores valores próprios do módulo ocorreu em variáveis canônicas elevadas como VC16, VC15 e VC14 e VC13, respectivamente (Tabela 7).

As estimativas da correlação de Spearman, entre o conjunto de descritores redundantes e o dos selecionados, demonstram que o descarte não revelou perda relevante de informação, pois os descritores descartados apresentam correlações significativas com os descritores selecionados (Tabela 5). Para Cury (1993), num processo de descarte envolvendo vários caracteres, é normal que ocorram perdas de informações. De acordo com Pinto et al. (2010), é possível a eliminação de descritores sem perda de informação, pois os mesmos podem estar correlacionados a outros que permaneceram na análise.

Com base na análise simultânea dos dois procedimentos (Singh e Jolliffe), sete caracteres foram coincidentes em relação ao descarte, ou seja, os

descritores eliminados foram os descritores descartados em ambos os métodos, assim fizeram parte do descarte final os seguintes descritores: Rendimento - produção (REND), dias do transplante ao florescimento (DTF); índice cilíndrico (IC); largura da 3ª folha (LF3); largura da 10ª folha (LF10); comprimento da flor (CFLR) e diâmetro da flor (DTFLR), (Tabela 8).

Em relação aos descritores selecionados na análise simultânea (Singh e Jolliffe) temos: altura da planta (ALT); número de folhas (NF); diâmetro médio do caule (DMC); Comprimento da 3ª folha (CF3); comprimento da 10ª folha (CF10); largura da base da 10ª folha (LB10); ângulo de inserção da 10ª folha (ANG10); comprimento dos internódios (CINT); engrossamento tubo da flor (ETFLR) e comprimento da corola (CCFLR), (Tabela 8).

Tabela 6. Estimativas dos autovalores associados as Variáveis Canônicas e suas variâncias total e acumulada, obtidas a partir dos dezessete descritores morfológicos avaliados, nos 15 acessos de tabaco estudados. Cruz das Almas-BA, 2014.

Variáveis Canônicas	Autovalores	Variância (%)	Variância acumulada (%)
1	79,98	65,48	65,48
2	30,21	24,73	90,21
3	4,78	3,91	94,12
4	2,14	1,75	95,88
5	1,48	1,21	97,09
6	1,19	0,98	98,07
7	0,78	0,64	98,70
8	0,62	0,51	99,21
9	0,30	0,25	99,46
10	0,25	0,20	99,66
11	0,19	0,16	99,82
12	0,16	0,13	99,95
13	0,04	0,03	99,98
14	0,02	0,02	100,00
15	0,00	0,00	100,00
16	0,00	0,00	100,00
17	0,00	0,00	100,00

Tabela 7. Estimativas dos coeficientes de ponderação associados as variáveis canônicas de autovetores inferiores 0,70 e identificação dos caracteres com indicação para descarte, em cada componente, pela seleção direta dos 15 acessos de tabaco. Cruz das Almas, 2014.

Descritores	Variáveis Canônicas									
	VC8	VC9	VC10	VC11	VC12	VC13	VC14	VC15	VC16	VC17
REND	.	.	.	.	0.546	0.105	0.024	0.280	0.007	-0.132
DTF	.	.	.	.	.	0.575	0.414	0.200	0.122	-0.283
ALT	-0.166	-0.041	-0.067	-0.043	0.263	0.354	-0.264	-0.511	0.295	-0.106
NF	0.306	0.170	-0.223	-0.067	-0.013	-0.057	0.024	0.478	0.032	0.225
DMC	.	.	.	.	.	.	.	.	.	0.811
IC								-0.650	-0.215	-0.077
LF3	.	1.030	-1.278	0.224	0.029	-0.317	-0.278	0.389	-0.170	0.035
CF3	.	.	.	.	.	.	.	.	0.546	0.231
LF10	.	.	.	.	.	.	0.487	-0.009	-0.046	0.054
CF10	-0.304	0.017	-0.255	-0.143	0.058	-0.153	-0.390	0.142	-0.054	-0.310
LB10	-0.210	-0.071	-0.227	0.054	0.153	0.102	-0.016	-0.077	0.103	0.130
ANG10	-0.332	-0.168	0.000	0.455	0.256	-0.460	0.192	0.057	-0.011	-0.052
CINT	0.266	0.023	-0.186	0.025	0.197	-0.010	0.018	0.449	-0.023	0.097
CFOR	.	.	.	0.457	-0.182	0.390	-0.171	-0.032	-0.397	-0.024
DTFLR	.	.	0.565	0.456	-0.037	0.175	0.043	-0.296	-0.374	0.203
ETFLR	0.231	0.610	0.314	-0.734	0.087	-0.257	-0.197	0.160	-0.002	-0.239
CCFLR	0.042	-0.425	0.392	0.048	-0.108	0.362	-0.288	0.051	0.267	-0.194

Rendimento - produção (REND), dias do transplante ao florescimento (DTF); altura da planta (ALT); número de folhas (NF); diâmetro médio do caule (DMC); índice cilíndrico (IC); largura da 3ª folha (LF3); Comprimento da 3ª folha (CF3); largura da 10ª folha (LF10); comprimento da 10ª folha (CF10); largura da base da 10ª folha (LB10); ângulo de inserção da 10ª folha (ANG10); comprimento dos internódios (CINT); comprimento da flor (CFLR); diâmetro da flor (DTFLR); engrossamento tubo da flor (ETFLR); comprimento da corola (CCFLR).

Tabela 8. Variáveis pré-selecionadas e selecionadas baseadas nos procedimentos de Singh (1981) e Jolliffe (1972). Cruz das Almas, 2014.

Variáveis	Pré-selecionadas		Selecionadas
	Jolliffe (1972)	Singh (1981)	
REND	Desc (6)	Desc (9)	Desc
DTF	Desc (5)	Desc (7)	Desc
ALT	Sel	Sel	**Sel**
NF	Desc (10)	Sel	**Sel**
DMC	Desc (1)	Sel	**Sel**
IC	Desc (3)	Desc (5)	Desc
LF3	Desc (9)	Desc (1)	Desc
CF3	Desc (2)	Sel	**Sel**
LF10	Desc (4)	Desc (3)	Desc
CF10	Sel	Sel	**Sel**
LB10	Sel	Sel	**Sel**
ANG10	Sel	Desc (8)	**Sel**
CINT	Sel	Sel	**Sel**
CFOR	Desc (7)	Desc (4)	Desc
DTFLR	Desc (8)	Desc (2)	Desc
ETFLR	Sel	Sel	**Sel**
CCFLR	Sel	Desc (6)	**Sel**

Rendimento - produção (REND), dias do transplante ao florescimento (DTF); altura da planta (ALT); número de folhas (NF); diâmetro médio do caule (DMC); índice cilíndrico (IC); largura da 3ª folha (LF3); Comprimento da 3ª folha (CF3); largura da 10ª folha (LF10); comprimento da 10ª folha (CF10); largura da base da 10ª folha (LB10); ângulo de inserção da 10ª folha (ANG10); comprimento dos internódios (CINT); comprimento da flor (CFLR); diâmetro da flor (DTFLR); engrossamento tubo da flor (ETFLR); comprimento da corola (CCFLR); Descartados (Desc); Selecionados (Sel).

Na tabela 9, estão apresentados os descritores qualitativos, suas classes fenotípicas, frequência percentual dos acessos em cada uma das classes e o nível de entropia de Renyi. Foram considerados como variáveis descartadas todas aquelas que apresentaram nível de entropia inferior a 0,60.

Os descritores descartados inicialmente por não serem capazes de diferir os acessos são apresentados na análise de entropia com uma frequência de 100%, ou seja, esses acessos ficaram concentrados na mesma classe dos descritores em questão, apresentando nível de entropia igual a zero: presença de brotos, início do florescimento (PB), curva da ponta da lâmina foliar (CPLF), perfil transversal da 10ª folha (PTF), ângulo nervuras laterais em relação à nervura central 10ª folha (ANGL), formato da aurícula: 10ª folha (FA), tipo de flor: presença ou ausência de anteras (TFLOR), formato do limbo da corola: forma da seção vista do alto (FLC), desenvolvimento dos estames (DESEN-E),

desenvolvimento do pistilo em relação aos estames (DESEN-P), expressão dos ápices da corola (EAC), forma da inflorescência (FI), posição da inflorescência em relação as folhas superiores (PI), tipo de deiscência do fruto (TDF) e formato do fruto (FF), (Tabela 9).

Com relação às entropias, verifica-se que as variáveis que apresentaram maiores valores estão relacionadas à cor da corola (CCOR): (H=1,00), coloração das folhas (CF): (H=0,96) e perfil longitudinal da folha (PLF) (H=0,68), em função de apresentarem elevado número de classes e um maior equilíbrio na proporção entre a frequência dos acessos nas diferentes classes fenotípicas. Isso revela variabilidade genética entre os acessos estudados (Tabela 9). Costa, (2012) analisando a variável cor da folha em acessos de *Nicotiana tabacum* L., observou a distribuição dos acessos em quatro classes. Sendo a cor da folha, característica de fundamental importância, por estar relacionado à classificação da cor da capa de fumo.

As variáveis que apresentaram os menores valores em relação ao nível de entropia foram: forma da planta (FP); tipo de folha (TF); forma das folhas medianas centrais (FFMC); densidade da inflorescência (DI), todas essas apresentando um nível de entropia de H=0,41 e com isso serão descartadas por não apresentarem um nível aceitável de entropia que possa ser determinante para discriminação dos acessos em estudo (Tabela 9). De acordo com Ledo et al. (2011), o nível de entropia pode ser utilizado para quantificar a variabilidade presente em descritores qualitativos por meio da observação das frequências relativas das classes para cada descritor avaliado. Desta forma, baixos valores para entropia estão associados a uma menor quantidade de classes fenotípicas para o descritor utilizado e a um maior desequilíbrio na proporção entre a frequência dos acessos nas diferentes classes fenotípicas.

Foram selecionados os seguintes descritores: coloração do caule (CCF), coloração das folhas (CF), coloração da nervura central, face inferior (CNC), superfície da lâmina foliar (SLF), perfil longitudinal da folha (PLF), margem lâmina foliar: 10ª folha (MLF), ponta da lâmina foliar: 10ª folha (PLAMF) e cor da corola (CCOR).

Dos 26 descritores qualitativos definidos conforme o SINDIFUMO (Subcomissão de Sementes), com base na descrição recomendada pela UPOV e

Legislações Americana e Italiana (Tabela 3), foi observado que apenas 8 descritores foram importantes na discriminação da divergência genética entre os acessos em estudo (Tabela 9).

Tabela 9. Descritores qualitativos avaliados, categorias fenotípicas (classes), frequência percentual e nível de entropia dos acessos de tabaco estudados. Cruz das Almas 2014.

Descritores	Classes	Frequência (%)	Nível de Entropia (H)
FP	Cônica	-	0,41
	Cilíndrica	14,29	
	Elíptica	85,71	
	Cônica invertida	-	
CCF	Verde-esbranquiçada	-	0,60
	Verde clara	28,57	
	Verde média	71,43	
	Verde escura	-	
PB	Ausente ou Muito fraca	100%	0
	Fraca	-	
	Média	-	
	Forte	-	
	Muito forte	-	
TF	Séssil	85,71	0,41
	Peciolada	14,29	
FFMC	Lanceolada	-	0,41
	Estreito-eliptica	14,29	
	Largo-eliptica	85,71	
	Ovalada	-	
	Obovada	-	
	Cordiforme	-	
	Arredondada	-	
CF	Verde-amarelada	-	0,96
	Verde esbranquiçada	-	
	Verde clara	28,57	
	Verde médio	57,14	
	Verde escura	14,29	
CNC	Esbranquiçada	28,57	0,60
	Verde esbranquiçada	71,43	
	Verde	-	
CPLF	Direto	-	0
	Curvado para baixo	100	
	Curvado para cima	-	
SLF	Fraco	71,43	0,60

	Médio	28,57	
	Forte	-	
PTF	Côncava	-	0
	Plana	-	
	Convexa	100	
PLF	Reto	-	0,68
	Ligeiramente recurvado	42,86	
	Moderadamente recurvado	57,14	
	Fortemente recurvado	-	
ANGL	Muito agudo	-	0
	Medianamente agudo	100	
	Reto	-	
MLF	Ausente ou muito fraca	71,43	0,60
	Fraca	28,57	
	Média	-	
	Forte	-	
PLAMF	Obtusa	-	0,60
	Ligeiramente pontiaguda	28,57	
	Median. Pontiaguda	71,43	
	Fort. Pontiaguda	-	
	Extre. Pontiaguda	-	
FA	Ausente ou muito fraco	-	0
	Fraco	-	
	Médio	-	
	Forte	100	
	Muito forte	-	
TFLOR	Presença	100	0
	Ausência	-	
CCOR	Branca	-	1,00
	Rosa-clara	42,86	
	Rosa-média	42,86	
	Rosa-forte	14,29	
	Vermelha	-	
FLC	Arredondada	-	0
	Poligonal	-	
	Poligonoestelar	100	
	Estelar	-	
	C/ sépalas muito pronunciadas	-	
DESEN-E	Nenhum ou rudimentar	-	0
	Pleno	100	
DESEN-P	Mais curto	-	0
	Mesmo comprimento	100	
	Mais largo	-	
EAC	Ausente ou muito fraca	-	0
	Fraca	-	

	Média	-	
	Forte	100	
	Muito forte	-	
FI	Esférica	-	0
	Esférica-aplanada	-	
	Cônica-invertida	-	
	Cônica-dupla	100	
PI	Entre as folhas	-	0
	Acima	100	
DI	Esparsa	85,71	0,41
	Média	14,29	
	Densa	-	
TDF	Não deiscente	-	0
	Cápsula deiscente	100	
FF	Arredondada	-	0
	Alongada	100	
	Elíptica	-	

Forma da planta (FP); Coloração do caule: início do florescimento (CCF); Presença de brotos (PB); Tipo de folha (TF); Forma das folhas medianas centrais (FFMC); Coloração das folhas (CF); Coloração da nervura central, face inferior (CNC); Curva da ponta da lâmina foliar (CPLF); Superfície da lâmina foliar (SLF); Perfil longitudinal da folha (PLF); Margem lâmina foliar: 10ª folha (MLF); Ponta da lâmina foliar: 10ª folha (PLAMF); Formato da aurícula: 10ª folha (FA); Tipo de flor: presença ou ausência de anteras (TFLOR); Cor da corola (CCOR); Densidade da inflorescência (DI).

Segundo Santos (2002), estudando acessos de *Nicotiana tabacum* L., caracteres como forma e tamanho da folha, número de folhas, altura da planta e comprimento dos internódios são importantes porque influenciam o manejo, o rendimento e a composição química das folhas. Alguns acessos no presente trabalho foram identificados para características de folha, altura da planta e comprimento de internódios, sendo indivíduos interessantes com potencial para serem utilizados em possíveis cruzamentos.

Os descritores selecionados são importantes na caracterização de germoplasma de tabaco, onde disponibilizam informações primordiais para o melhoramento genético dos acessos em estudo e para espécie.

O descarte realizado possibilitará a redução no tempo, na mão-de-obra e nos custos das atividades de avaliação e caracterização da cultura. O presente trabalho fornece informações para considerar que todos os descritores morfoagronômicos selecionados, cerca de 42%, são essenciais no estudo de caracterização por apresentarem contribuições importantes na discriminação da

divergência genética entre os acessos de tabaco *Nicotiana tabacum* L., tipo Sumatra em estudo.

CONCLUSÕES

Os métodos utilizados para o descarte das variáveis foram eficientes na identificação e eliminação de variáveis redundantes.

Os 10 descritores quantitativos selecionados, na análise simultânea de Singh e Jolliffe, altura da planta, número de folhas, diâmetro médio do caule, comprimento da 3ª folha, comprimento da 10ª folha, largura da base da 10ª folha, ângulo de inserção da 10ª folha, comprimento dos internódios, engrossamento tubo da flor e comprimento da corola e os 8 descritores qualitativos selecionados através do nível de entropia, coloração do caule, coloração das folhas, coloração da nervura central, face inferior, superfície da lâmina foliar, perfil longitudinal da folha, margem lâmina foliar: 10ª folha, ponta da lâmina foliar: 10ª folha e Cor da corola são importantes na caracterização do germoplasma de tabaco em estudo.

O descarte de 58% dos descritores não provocou perda de informação considerável, uma vez que os descritores redundantes estão correlacionados aos descritores remanescentes, com possibilidade de redução de custos e melhor dinâmica no manejo e caracterização da cultura.

AGRADECIMENTOS

À Universidade Federal do Recôncavo da Bahia, a empresa Ermor Tabarama Tabacos do Brasil Ltda, pela parceria e infraestrutura e a Fundação de Amparo à Pesquisa do Estado da Bahia (Fapesb) pela concessão da bolsa de mestrado que permitiu o desenvolvimento deste trabalho.

REFERÊNCIAS

ADAPAR - AGÊNCIA DE DEFESA AGROPECUÁRIA DO PARANÁ. Disponível em: < http://www.adapar.pr.gov.br/modules/noticias/article.php?storyid=178>. Acesso em 16 de Dez. de 2014.

AFONSO, S. D. J.; LEDO, C. A. da S.; MOREIRA, R. F. C.; SILVA, S. de O. e; LEAL, V. D. de J.; CONCEIÇÃO, A. L. da S. Selection of descriptors in a morphological characteristics considered in cassava accessions by means of multivariate techniques. **Journal of Agriculture and Veterinary Science**, v. 7, *Issue 1 Ver. V*, p. 13-20, 2014.

BINDLER, G; VAN DER HOEVEN, R.; GUNDUZ, I.; PLIESKE, J.; GANA, M.; ROSSI, L.; GADANI, F.; DONINI, P. A. Microssatellite marker based linkage map of tobacco. **Theoretical and Applied Genetics**, New York, v. 114, p. 341-349, 2007.

CONCEIÇÃO, A. L. da S.; SILVA, M. dos S. da.; SANTOS, C. C. dos.; ARAUJO, G. de M.; MOREIRA, R. F. C. Variabilidade genética e importância relativa de caracteres em acessos de tabaco (*Nicotiana tabacum* L.) Tipo broad leaf por meio de marcadores fenotípicos. **Enciclopédia Biosfera**, Centro Científico Conhecer - Goiânia, v.10, n.19; p.1900-1907, 2014.

COSTA, T. P. P. **Caracterização Morfoagronômica de Genótipos de Tabaco na Região do Recôncavo da Bahia.** Dissertação de Mestrado em Recursos Genéticos Vegetais, Universidade Federal do Recôncavo da Bahia, Cruz das Almas, BA, Brasil. Maio, 2012.

CRUZ, C. D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. 3. ed. Viçosa: UFV, p. 480, 2004.

CRUZ, C.D. Programa Genes - **Aplicativo computacional em genética e estatística**. Disponível em: <www.ufv.br/dbg/genes/genes.htm>. 2014.

CURY, R. **Dinâmica evolutiva e caracterização de germoplasma de mandioca (*Manihot esculenta*, Crantz) na agricultura autóctone do Sul do Estado de São Paulo**. Dissertação (Mestrado) – Escola Superior de Agricultura "Luiz de Queiroz, Universidade de São Paulo, Piracicaba. p.103, 1993.

DAHER, R. F. **Diversidade morfológica e isoenzimática em capim elefante (**Pennisetum purpureum Schum.). Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Universidade Federal de Viçosa, Viçosa, MG.p. 110, 1993.

GOODSPEED, T. H.; WHEELER H-M; HUTCHISON, P. C. Taxonomy of *Nicotiana.* In: GOODSPEED, T. H. **The genus Nicotiana**. Waltham: Chronica Botanica. v. 16, pt. 6, p. 321-492, 1954.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. Ed Bookman, Porto Alegre, p. 593, 2005.

JOLLIFFE, I.T. Discarding variables in a principal component analysis. I. Artificial data. **Applied Statistics**, v.21, p.160-173, 1972.

JOLLIFFE, I.T. Discarding variables in a principal component analysis. II: real data. **Journal of the Royal Statistical Society**. Series C (Applied Statistics), v.22, p. 21-31, 1973.

LEDO, C. A da S.; TAVARES FILHO, L.F.de Q.; OLIVEIRA, M. M de., SILVEIRA, T.C da, SANTOS, A. S.; ALVES, A. A. C.; GONÇALVES, L.S.A. Análise de agrupamento utilizando variáveis quantitativas e qualitativas para o estudo da diversidade genética em genótipos de mandioca silvestre. **XIII Congresso Brasileiro de Mandioca.** Botucatu, SP, 591-595, 2009.

LEDO, C. A. S da.; ALVES, A. A. C.; SILVEIRA, T. C. da.; OLIVEIRA, M. M. de.; SANTOS, A. S.; TAVARES FILHO, L. F. de Q. Caracterização morfológica da

coleção de espécies silvestres de *Manihot* (Euphorbiaceae – Magnoliophyta) da Embrapa Mandioca e Fruticultura**. Pesquisa Agropecuária Brasileira**, v.53, dezembro, 2011.

OLIVEIRA, E.J. de.; OLIVEIRA FILHO, O.S. de.; SANTOS, V. da. S. Selection of the most informative morphoagronomic descriptors for cassava germplasm. **Pesquisa Agropecuária Brasileira**, Brasilia, DF, v.49, n.11, p.891-900, nov. 2014.

OLIVEIRA, M. S. P. **Caracterização molecular e morfoagronômica de germoplasma de açaizeiro**. 2005. 171f. Tese (Doutorado) - Universidade Federal de Lavras, Lavras, 2005.

MAPA – Ministério da Agricultura, Pecuária e Abastecimento. Instruções para Execução dos Ensaios de Distingüibilidade, Homogeneidade e Estabilidade de Cultivares de Tabaco (*Nicotiana tabacum* L.). Disponível em: <http://www.agricultura.gov.br/arq_editor/file/vegetal/RegistroAutorizacoes/Formularios%20Proe%C3%A7%C3%A3o%20Cultivares/TABACO%20FORMULARIO%2001%2008%202008%20P.doc>. Acesso em: 15 fev, 2015.

MARTEL, J. H. I., FERRAUDO, A. S.; MOROÔ, J. R.; PERECIN, D. Estatística multivariada na discriminação de raças amazônicas de Pupunheira (*Bactris gasipaes* Kunth) em Manaus (Brasil). **Revista Brasileira de Fruticultura,** Jaboticabal, v.25, n.1, p. 115-118, 2003.

NEGREIROS, J. S.; ALEXANDRE, R. S.; ÁLVARES, V. S.; BRUCKNER, C. H.; CRUZ, C. D. Divergência genética entre progênies de maracujazeiro-amarelo com base em características das plântula. **Revista Brasileira de Fruticultura**, Jaboticabal, v.30, n.1, p.197-201, 2008

PEREIRA, V. A. **Utilização de análise multivariada na caracterização de germoplasma de mandioca (Manihot esculenta Crantz.)**. Tese (Doutorado em

Genética e Melhoramento de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, p. 180, 1989.

PINTO, J. F. N.; REIS, E. F. dos; FALEIRO, F.G.; BARBOSA, E. C. C.; NUNES, H. F.; Pinto, JEEDER, F. N. Seleção de descritores vegetativos para caracterização de acessos de guariroba (Syagrus oleracea (Mart.) Becc.). **Revista Brasileira de Fruticultura** [online]. v.32, n.3, p. 832-840, 2010.

REGO, E.R. do; REGO, M.M. do; CRUZ, C.D.; CECON, P.R.; AMARAL, D.S.S.L.; FINGER, F. Genetic diversity analysis of peppers: a comparison of discarding variable methods. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 3, n. 1, p. 19-26, 2003.

REN, N.; TIMKO, M.P. AFLP analysis of genetic polymorphism and evolutionary relationships among cultivated and wild *Nicotiana* species. **Genome**, Ottawa, v. 44, n. 4, p. 559-571, 2001.

RENYI, A. **On measures of entropy and information**. Fourth Berkeley Symposium, Berkley, 1960. p. 547-561, 1961.

SANTOS, M. **Caracterização fenotípica e molecular de genótipos de fumo no Sul do Brasil.** Dissertação de Mestrado em Fitotecnia, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil. 122p. Abril, 2002.

SAS INSTITUTE. **SAS user's guide**: statistic: version 9.1.3. Cary: SAS Institute, p. 846, 2004.

SINGH, D. The relative importance of characters affecting genetic divergence**. The Indian Journal of Genetics and Plant Breeding**, v.41, n.1, p.237 - 245, 1981.

SEAGRI - Secretaria da Agricultura, Pecuária, Irrigação, Pesca e Aquicultura. Disponivel em: <

http://www.seagri.ba.gov.br/noticias/2014/09/22/exporta%C3%A7%C3%A3o-de-fumo-mant%C3%AAm-cultivo-local>. Acesso em. 15 de Dez. 2014.

UPOV (*Union pour la Protection des Obtentions Variétales*). Disponível em http://www.upov.int/tabaco. Acesso em: 20 dez. 2014.

VENCOVSKY, R.; BARRIGA, P. Associação entre caracteres. In: _____. **Genética biométrica no fitomelhoramento**. Ribeirão Preto: Sociedade Brasileira de Genética, p. 335-434, 1992.

VIEIRA, E. A.; FIALHO, J. de F.; SILVA, M. S.; FALEIRO, F. G. **Variabilidade genética do banco ativo de germoplasma de mandioca do cerrado acessada por meio de descritores morfológicos**. Planaltina: Embrapa Cerrados. (Boletim de Pesquisa e Desenvolvimento, 129). p. 15 , 2007.

CAPÍTULO II

COMPARAÇÃO DE MÉTODOS DE AGRUPAMENTO EM ACESSOS DE TABACO

---

[2]Artigo a ser submetido ao periódico Pesquisa Agropecuária Brasileira

COMPARAÇÃO DE MÉTODOS DE AGRUPAMENTO EM ACESSOS DE TABACO

Autor: Antonio Leandro da Silva Conceição
Orientador: Carlos Alberto da Silva Ledo
Co-orientador: Ricardo Franco Cunha Moreira

RESUMO: O estudo da diversidade genética de tabaco é de interesse para a conservação dos recursos genéticos, ampliação da base genética e aplicações práticas, onde a existência de variabilidade genética e seu conhecimento são fatores indispensáveis para o sucesso dos programas de melhoramento da cultura. O objetivo deste trabalho foi aplicar diferentes metodologias de análise de agrupamento com base na análise simultânea e individual de descritores quantitativos e qualitativos selecionados de 15 acessos de tabaco. Os métodos de agrupamento utilizados foram o UPGMA e WARD e a consistência do ajuste desses agrupamentos foram medidos pelo coeficiente de correlação cofenética, em que o método UPGMA foi o que melhor explicou a divergência genética entre os acessos em estudo. Foram utilizadas para as análises individuais as distâncias de Mahalanobis e euclidiana média para obtenção das matrizes oriundas dos dados quantitativos e para os dados qualitativos originais e, para os dados quantitativos transformados foi utilizada a distância de Cole-Rodgers. Para as análises simultâneas dos dados quantitativos e qualitativos foram testadas três metodologias distintas: O algoritmo de Gower; Soma algébrica de matrizes individuais e integração de dados por meio da transformação de caracteres quantitativos em multicategóricos por duas estratégias distintas (Regra de Sturges e Regra da raiz quadrada). A comparação entre as matrizes de dissimilaridade obtidas a partir dos dados de diferentes naturezas foi realizada pelo teste Z de Mantel. Os valores de correlação entre as matrizes de dados quantitativos transformados e quantitativos originais foram significativos a 1 e a 5% de probabilidade, sendo um deles de alta magnitude, oferecendo suporte para extrapolar os resultados de um conjunto de dados para outro. Com isso, a estratégia da raiz quadrada foi a mais indicada, com correlação de 0,75 e 0,82 entre as matrizes de dissimilaridade dos dados codificados e quantitativos

originais. Como critério para definição do número ótimo de grupos foi usado o índice Pseudo-t^2, com este foi possível a formação de 3 grupos pelo método UPGMA para todas as metodologias de análise simultânea utilizadas. O acesso A14 (125 PD) possui comportamento distinto dos demais e as metodologias de análise simultânea, com base nas matrizes de distâncias geradas, captaram essa divergência. Todas as matrizes de análise conjunta foram comparadas, mostrando grande correspondência entre as mesmas, com altas correlações que variaram de 0,824 a 0,998. Os resultados deste trabalho mostraram que as metodologias de análise simultânea foram eficazes em revelar a existência de divergência genética entre acessos de *Nicotiana tabacum* L. tipo Sumatra e mostram a importância da combinação de métodos, uma vez que puderam otimizar de forma considerável a interpretação dos resultados para maior conhecimento do germoplasma em estudo.

Palavras-chave: Divergência genética, *Nicotiana tabacum* L., análises simultâneas

COMPARISON OF CLUSTERING METHODS IN TOBACCO ACCESS

Author: Antonio Leandro da silva Conceição
Advisor: Carlos Alberto da Silva Ledo
Co-advisor: Ricardo Franco Cunha Moreira

ABSTRACT: The study of genetic diversity of tobacco is of interest for the conservation of genetic resources, expanding the genetic basis and practical applications, where the existence of genetic variability and their knowledge are factors, which are essential to the success of the breeding programs of culture. The objective of this study was to apply different methods of cluster analysis based on the simultaneous analysis and individual descriptors quantitative and qualitative selected of 15 accessions of tobacco. The grouping methods used were the UPGMA and WARD adjusting consistency of these groupings were measured by the coefficient of correlation cofenetica, in that the UPGMA method was the best explained the genetic divergence among the accessions into study. They were used for the individual analyzes the Mahalanobis distance and Euclidean average for obtaining the arrays from the quantitative data and the qualitative data and original quantitative data processed was used the distance of Cole-Rodgers. For the simultaneous analysis of quantitative and qualitative data were tested three different methodologies: For the simultaneous analysis of quantitative and qualitative data were tested three different methodologies: The algorithm of Gower; Algebraic Sum of individual arrays and data integration through the transformation of quantitative traits in multicategoric by two distinct strategies (Rule of Sturges and Rule the square root). The comparison between dissimilarity matrices obtained from data of different natures was performed by Mantel Z test. The correlation values between the arrays of data processed and quantitative quantitative originals were significant 1 and 5% probability, one of them being of high magnitude, offering support to extrapolate the results of a set of data to another. With this, the strategy of square root was the most indicated, with a correlation of 0.75 and 0.82 between the arrays of dissimilarity of encoded data and quantitative originals. As a criterion for the definition of the optimal number of groups was used the index Pseudo-t2, this was possible the formation of 3 groups by UPGMA method for all the methodologies of simultaneous analysis used.

Access to14 (125 PD) has distinct behavior of others and the methodologies of simultaneous analysis, based on matrices of distances generated, captured this divergence. All arrays of joint analysis were compared, showing great correspondence between them, with high correlations ranged from 0.824 to 0.998. The results of this study showed that the methodologies of simultaneous analysis were effective in identifying the existence of genetic divergence among accessions of *Nicotiana tabacum* L. Type Sumatra and show the importance of the combination of methods, since they have been able to leverage a significant way to the interpretation of the results for greater knowledge of germplasm in study.

Key words: Genetic divergence, *Nicotiana tabacum* L., simultaneous analysis

INTRODUÇÃO

Brasil é líder mundial em exportação de tabaco. Atendendo aos mais exigentes padrões internacionais, o Brasil é o segundo maior produtor mundial de tabaco e líder em exportações desde 1993, graças à qualidade e integridade do produto. Em 2014, o tabaco representou 1,11% do total das exportações brasileiras, com US$ 2,5 bilhões embarcados. Da produção de 735 mil toneladas registrada na safra 2013/14, mais de 85% foi destinada ao mercado externo. O principal mercado brasileiro neste período foi a União Europeia com 42% do total dos embarques de 2014, seguida pelo Extremo Oriente (28%), América do Norte (10%), Leste Europeu (8%), África/Oriente Médio (6%) e América Latina (6%), (SINDITABACO, 2015).

O estudo da diversidade genética em tabaco é de interesse para a conservação dos recursos genéticos, ampliação da base genética e aplicações práticas em programas de melhoramento. Estudos com análises de diversidade genética em *Nicotiana tabacum* L. são escassos no Brasil, principalmente levando em consideração a análise simultânea de dados. Na literatura encontram-se alguns estudos levando em consideração análises em isolado e/ou com mistura de variáveis. Zhang et al. (2006) estudaram diversidade genética entre acessos de flue-cured tobacco (*Nicotiana tabacum* L.) a partir de marcadores moleculares. Davalieva et al. (2010), analisaram variabilidade genética de variedades de tabaco na República da Macedónia determinadas por análise de marcadores microssatélites. Costa (2012) estudou acessos de *Nicotiana tabacum* L., com base em caracteres qualitativos e quantitativos isoladamente e em simultâneo por meio do algoritmo de Gower. Darvishzadeh et al. (2013) estudaram variação genética em tabaco oriental (*Nicotiana tabacum* L.) por marcadores agromorfológicos e traços simples de repetição de sequência. Mwadzingeni et al. (2013) estudaram diversidade genética de variedades exóticas de tabaco do tipo Flue-Cured no Zimbabwe, também usando como base características fenotípicas e repetições de sequências simples. Conceição et al. (2014) analisaram variabilidade genética e importância relativa de caracteres em acessos de tabaco (*Nicotiana tabacum* L.) tipo broad leaf por meio de caracteres quantitativos.

Segundo Moura et al. (2010), embora a análise conjunta das variáveis quantitativas e qualitativas seja potencialmente um indicador mais completo da variabilidade existente nos bancos de germoplasma, poucos trabalhos têm utilizado esta estratégia. Provavelmente, isso ocorre devido à falta de conhecimento das técnicas estatísticas que permitem essa abordagem, desses dados conjuntamente, bem como pela tendência dos pesquisadores em dar mais importância àquelas variáveis diretamente relacionadas com caracteres trabalhados em programas de melhoramento (GONÇALVES et al., 2008).

A análise individual para cada tipo de variável pode levar a discrepâncias em relação aos agrupamentos e às inferências em relação à quantificação da variabilidade entre as unidades experimentais ou amostrais a serem agrupadas. Com isso, as análises com métodos que considerem simultaneamente os diversos tipos de variáveis são preferíveis (LEDO e GONÇALVES, 2012). O aumento no uso de técnicas multivariadas para quantificação da divergência genética tem sido verificado já que essas análises permitem considerar simultaneamente inúmeras características (SUDRÉ et al., 2007).

Algumas metodologias de integração de dados de diferentes naturezas foram propostas para a análise de diversidade genética considerando simultaneamente diferentes grupos de variáveis. Destacam-se as estratégias de transformação de dados utilizadas por Martins et al. (2011), soma algébrica de matrizes individuais por Cruz et al. (2011) e através da técnica proposta por Gower (1971), por meio de um algoritmo que estima a similaridade entre dois indivíduos utilizando dados com distribuições contínuas e discretas.

Segundo Cargnelutti Filho (2008), do ponto de vista do melhorista de plantas, o processamento dos dados por diversos métodos de agrupamento e com base em diversas medidas de dissimilaridade e a consideração das particularidades de cada um, é adequada para uma melhor tomada de decisão em relação à escolha de cultivares para cruzamentos. Portanto, é interessante despertar o interesse por pesquisas que expõem a combinação de métodos, uma vez que, podem resultar num aperfeiçoamento dos resultados de uma análise. Ainda que cada técnica possua suas particularidades e objetivos específicos de pesquisa, o ajuste de duas ou mais técnicas podem proceder a invenção de uma nova técnica (ALVES, 2007).

Diante do exposto, o objetivo deste trabalho foi avaliar a divergência genética, por meio de diferentes metodologias de análise de agrupamento com base na análise simultânea e individual de descritores quantitativos e qualitativos selecionados em15 acessos de tabaco (*Nicotiana tabacum* L.) tipo Sumatra.

MATERIAL E MÉTODOS

Foram avaliados 15 acessos de tabaco da espécie *Nicotiana tabacum* L., conforme apresentados abaixo. Esses acessos são provenientes da empresa ERMOR TABARAMA TABACOS DO BRASIL Ltda., localizada no município de Cruz das Almas – BA. Estes materiais são e formam a base de materiais (sementes) fornecidos aos agricultores da região do Recôncavo da Bahia.

Os acessos de tabaco estudados foram: ER 03-107 (A1); ER 04-090 (A2); ER 04-095 (A3); ER 05-005 (A4); ER 05-070 (A5); ER 12-040 (A6); ER 13-061 (A7); ER 13-065 (A8); ER 28-027 (A9); ER 33-021 (A10); ER 33-022 (A11); ER 33-023 (A12); 109 PD (A13); 125 PD (A14); 221 PD (A15).

O delineamento experimental utilizado foi o de blocos casualizados com quatro repetições. Cada parcela foi constituída de cinco linhas de 10 plantas e cada linha teve 4,5 metros de comprimento com espaçamento de 1,0 metros entre linhas e 0,42 metros entre plantas.

Foram avaliadas 18 variáveis para a para a análise de agrupamento, sendo dez variáveis quantitativas: altura da planta (cm), número de folhas, diâmetro médio do caule (cm), Comprimento da 3ª folha (cm), comprimento da 10ª folha (cm), largura da base da 10ª folha (cm), ângulo de inserção da 10ª folha, comprimento dos internódios (cm), engrossamento tubo da flor (mm) e comprimento da corola (cm). E oito variáveis qualitativas: Coloração do caule: início do florescimento (CCF); Coloração das folhas (CF); Coloração da nervura central, face inferior (CNC); Superfície da lâmina foliar (SLF); Perfil longitudinal da folha (PLF); Margem lâmina foliar: 10ª folha (MLF); Ponta da lâmina foliar: 10ª folha (PLAMF); Cor da corola (CCOR).

Os caracteres analisados foram definidos conforme o SINDIFUMO (Subcomissão de Sementes), Ministério da Agricultura, Pecuária e Abastecimento

e com base na descrição recomendada pela UPOV e Legislações Americana e Italiana (MAPA, 2015).

Análises estatísticas

Análise de diversidade a partir da caracteres quantitativos e qualitativos em isolado.

As matrizes de dissimilaridade obtidas individualmente para os dados quantitativos, basearam-se na distância de Mahalanobis (D^2) e euclidiana média padronizada, já para os dados quantitativos transformados e qualitativos originais foi utilizada a distância de Cole Rodgers.

Análise de diversidade a partir da caracteres quantitativos e qualitativos simultaneamente.

Para a análise simultânea dos dados foram utilizadas as seguintes estratégias:

i) **A análise simultânea proposta por Gower (1971)**.

expresso por:

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} \cdot S_{ijk}}{\sum_{K=1}^{p} W_{ijk}}$$

Em que *K* é o número de variáveis (k = 1, 2,..., p = número total de características avaliadas); *i* e *j* dois indivíduos quaisquer; W_{ijk} é um peso dado a comparação *ijk*, atribuindo valor 1 para comparações válidas e valor 0 para comparações inválidas (quando o valor da variável está ausente em um ou

ambos indivíduos); S_{ijk} é a contribuição da variável *k* na similaridade entre os indivíduos *i* e *j*, ele possui valores entre 0 e 1. Para uma variável nominal, se o valor da variável *k* é a mesma para ambos os indivíduos, *i* e *j*, então $S_{ijk} = 1$, caso contrário, é igual a 0; para uma variável contínua $S_{ijk} = 1 - | x_{ik} – x_{jk} | / R_k$ onde x_{ik} e x_{jk} são os valores da variável *k* para os indivíduos *i* e *j*, respectivamente, e R_k é a amplitude de variação da variável *k* na amostra. A divisão por R_k elimina as diferenças entre escalas das variáveis, produzindo um valor dentro do intervalo [0, 1] e pesos iguais.

ii) **Soma Algébrica de matrizes individuais**

Na metodologia descrita por Cruz et al. (2011) é proposta a soma algébrica das distâncias padronizadas das matrizes individuais. Após a obtenção das matrizes de dissimilaridade, individualmente para cada conjunto de caracteres, as mesmas são somadas algebricamente, obtendo-se outra matriz de dissimilaridade (SOMA) visando a obtenção de uma matriz única.

Antes da soma das matrizes uma medida de distância foi utilizada para cada tipo de variável, sendo que para as variáveis quantitativas foi calculada a distância generalizada de Mahalanobis (D^2). Ainda para os dados quantitativos, foi calculada também a distância euclidiana média padronizada. Para as variáveis multicategóricas foi utilizada a distância de Cole-Rodgers (COLE-RODGERS et al.,1997).

Após a obtenção das matrizes de dissimilaridade individualmente para cada conjunto de caracteres, as mesmas foram somadas algebricamente, tomando-se o cuidado de realizar a padronização dos dados, obtendo-se outra matriz de dissimilaridade (SOMA), a qual foi comparada com as matrizes originais por meio do teste Z de Mantel.

iii) **Transformação de dados Quantitativos em Multicategóricos**

O processo de “Qualitificação” (Qualitizing**)** consiste em transformar as variáveis quantitativas em multicategóricas, visando a obtenção de uma matriz única.

Para a conversão dos dados quantitativos em multicategóricos foi utilizada duas estratégias distintas para decidir o número de classes durante a transformação dos dados. As estratégias utilizadas são citadas abaixo:

A “Regra de Sturges” e Raiz quadrada do número de observações.

“Regra de Sturges”, fornece o número de classes em função do total de observações:

$$k = 1 + 3{,}3 * \log^{10}(n)$$

Onde: K é o número de classes; n é o número total de observações ou seja o número total de dados.

O método que utilizada o cálculo da raiz quadrada também é baseada no número de observações e é dado por:

$$K = \sqrt{N}$$

Onde: K é o número de classes; *N* é o número total de observações ou seja o número total de dados.

Foram testados os dois métodos de transformação, visando verificar qual deles apresenta menor viés dos dados originais, considerando as diferentes particularidades.

Após as conversões, novas matrizes de dissimilaridade foram obtidas para cada estratégia, tendo-se utilizado como medida de dissimilaridade o coeficiente de dissimilaridade, proposto por Cole-Rodgers (COLE-RODGERS et al., 1997)

Os dados transformados foram integrados aos dados qualitativos de origem para obtenção de uma só matriz e também foram testados na metodologia de soma algébrica de matrizes.

Foram estimados os coeficientes de correlação entre essas matrizes e a matriz de dissimilaridade obtida dos dados quantitativos originais, a 5% de probabilidade, pelo teste Z de Mantel e pelo teste t, para a determinação da forma mais adequada de conversão dos dados quantitativos.

Matrizes de dissimilaridade obtidas

A seguir são apresentadas todas as matrizes que foram obtidas no presente estudo utilizando os acessos de tabaco (*Nicotiana tabacum* L.).

1) Matriz dados quantitativos: Distância de Mahalanobis (D^2).
2) Matriz dados quantitativos: Distância euclidiana média padronizada.
3) Matriz dados qualitativos: Distância de Cole-Rodgers.
4) Matriz dados quantitativos transformados pela Regra de Sturges: Distância utilizada: cole Rodgers.
5) Matriz dados quantitativos transformadas pelo cálculo da Raiz Quadrada: Distância utilizada: cole Rodgers.
6) Matriz conjunta 1 - Matriz C1: metodologia utilizada (Algoritmo de Gower).
7) Matriz conjunta 2 - Matriz C2 (soma de matriz Mahalanobis + matriz Cole Rodgers): metodologia utilizada (Soma algébrica das matrizes).
8) Matriz conjunta 3 - Matriz C3 (soma de matriz Euclidiana Média + matriz Cole Rodgers): metodologia utilizada (Soma algébrica das matrizes).
9) Matriz C4 (soma de matriz dados quantitativos transformados em qualitativos pela regra de Sturges + dados qualitativos originais): metodologia utilizada (transformação de dados e Soma algébrica das matrizes).
10) Matriz C5 (dados quantitativos transformados pela regra de Sturges integrados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, assumindo que todos os dados analisados são “qualitativos”): metodologia utilizada (transformação de dados)
11) Matriz C6 (soma de matriz dados quantitativos transformados em qualitativos pelo cálculo da raiz quadrada + dados qualitativos originais): metodologia utilizada (transformação de dados e Soma algébrica das matrizes).

12) Matriz C7 (dados quantitativos transformados pelo cálculo da raiz quadrada integrados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers): metodologia utilizada (transformação de dados)

Técnicas de Agrupamento Utilizadas

A partir de cada uma das matrizes de dissimilaridade, procedeu-se à análise de divergência genética pelos métodos de agrupamento UPGMA e WARD.

UPGMA (Unweighted Pair-Group Method Using Arithmetic Averages) - Método de ligação média entre grupos.

Nesse método, a matriz de distâncias é atualizada calculando-se a média das distâncias entre os indivíduos de dois grupos. Assim, se C_1 tem n_1 indivíduos e C_2 tem n_2 indivíduos, a distância entre eles será definida por

$$d(C_1, C_2) = \sum_{l \in C_1} \sum_{k \in C_2} \left(\frac{1}{n_1 n_2} \right) d(X_l, X_k)$$

Método de Ward

O método de Ward (WARD, 1963) ou de variância mínima consiste em formar grupos a partir de pares que proporcionem a menor soma de quadrados.

Cada elemento é considerado um conglomerado e então, calcula-se a soma de quadrados dentro de cada conglomerado. Esta soma é o quadrado da distância Euclidiana de cada elemento pertencente ao conglomerado em relação ao correspondente vetor de médias do conglomerado

$$SS_i = \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.} \right)^T \left(X_{ij} - \bar{X}_{i.} \right)$$

em que n_l é o número de elementos do conglomerado C_i quando se está no passo k do processo de agrupamento; X_{ij} é o vetor de observações do *j*-ésimo elemento pertencente ao *i*-ésimo conglomerado ; $\overline{X}_{i.}$ é o vetor de médias do conglomerado C_i e SS_i é a soma de quadrados referente a tal conglomerado (MINGOTI, 2005).

Posteriormente, calcula-se a soma de quadrados entre dois conglomerados C_l e C_i que é dado por:

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i}\right] \overline{X}_{l.} - \overline{X}_{i.}^{\ T} \ \overline{X}_{l.} - \overline{X}_{i.}$$

em que $\left[\frac{n_l n_i}{n_l + n_i}\right]$ é um fator de ponderação para quando os conglomerados tiverem tamanhos diferentes (MINGOTI, 2005). A cada passo do algoritmo, os dois conglomerados que minimizam tal distância são combinados.

A validação dos agrupamentos foi determinada pelo coeficiente de correlação cofenética (CCC).

Um importante fator a ser considerado na definição de qual método de agrupamento hierárquico utilizar é sua consistência e adequação aos dados. Alguns métodos estatísticos de análises, como, por exemplo, a análise de correlação cofenética associada à análise de agrupamento, podem ser empregados para aumentar a confiabilidade das conclusões frente à interpretação dos dendrogramas (SOKAL & ROHLF, 1962).

O coeficiente de correlação cofenética (CCC) mede o grau de ajuste da matriz de dissimilaridade (matriz fenética) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz cofenética).

O CCC foi obtido por (BUSSAB et al., 1990):

$$r_{Cof} = r_{FC} = \frac{C\hat{o}v(F,C)}{\sqrt{\hat{V}(F).\hat{V}(C)}}$$

em que:

Côv (F, C): covariância entre os elementos da matriz fenética e cofenética;

$\hat{v}$ (F): variância dos elementos da matriz fenética;

$\hat{v}$ (C): variância dos elementos da matriz cofenética

Definição do número de grupos

Uma dificuldade encontrada na etapa final dos estudos que se utilizam de algoritmos hierárquicos de agrupamento é a falta de critérios objetivos para identificar o número ideal de grupos formados, uma vez que na prática este número é dado simplesmente por uma inspeção gráfica visual ou estabelecido em pontos de alta mudança de nível dos dendrogramas (MILLIGAN, 1985).

Alguns dos critérios para a determinação do número ideal de agrupamentos citados por Mingoti (2007) são: análise de comportamento do nível de fusão (distância), análise de comportamento do nível de similaridade, análise da soma dos quadrados entre grupos, correlação semiparcial e estatística pseudo t^2.

Nesse estudo, para definição do número de grupos foi utilizado o índice pseudo t^2, proposto por Duda e Hart (1973) através do pacote “NbClust” (CHARRAD et al., 2013), pertencente ao programa estatístico R (R CORE TEAM, 2014). Segundo Mingoti (2007), o fundamento do critério pseudo-t^2 está relacionado com o teste de hipótese, ou seja, é como se em cada passo do processo de agrupamento estivesse sendo feito um teste para comparação dos vetores de médias dos dois grupos que se uniram para formar um novo grupo. Por conseguinte, interessam os maiores valores de pseudo-t^2, vez que estariam relacionados com a menor probabilidade de significância do teste, e dessa forma, estaria rejeitando a igualdade de vetores de médias com maior significância.

Correlação entre as matrizes

A existência de correlação entre todas as matrizes de dissimilaridade foi verificada pelo teste Z de Mantel e pelo teste t, a 5% de probabilidade. O valor Z de Mantel (MANLY,1997) é dado por:

$$Z = \sum_{i,j=1}^{n} X_{ij} Y_{ij} \ ,$$

Onde X_{ij} e Y_{ij} são elementos das matrizes X e Y a serem comparadas. A significância desse valor de Z pode ser obtida comparando-se esse valor observado com valores de uma distribuição sob hipótese nula, recalculando-se os valores de Z diversas vezes, aleatorizando, em cada uma delas, a ordem dos elementos de uma das matrizes. Este Z calculado após permutações aleatórias é chamado de Z randômico (Z_{rnd}). A estatística Z possui uma relação monotônica com o r de Pearson entre as matrizes (correlação matricial), de modo que ela é de fato utilizada para testar a significância do r (MANLY, 1997). A correlação calculada pelo teste de Mantel varia de -1 a +1, e mede a correlação entre duas matrizes com relação ao Z randômico (Z_{rnd}). Para valores negativos de r, quanto menor a freqüência do $Z_{rnd} \leq Z_{obs,}$ maior a correlação entre as duas matrizes. Para valores positivos de r, quanto menor a freqüência de $Z_{rnd} \geq Z_{obs,}$ maior a correlação. Neste trabalho, 1000 permutações aleatórias foram utilizadas para se testar a significância das correlações matriciais.

Softwares utilizados nas análises estatísticas

As matrizes baseadas na distância de Mahalanobis, euclidiana média para os dados quantitativos e as matrizes obtidas pela distância de Cole-Rodgers para os dados qualitativos originais e dados quantitativos transformados foram calculadas utilizando o programa Genes (CRUZ, 2014). Para a obtenção da matriz de distância genética da análise conjunta utilizando o algoritmo de Gower foi utilizado o programa estatístico R (R CORE TEAM, 2014). Já para a análise simultânea por meio da SOMA algébrica de matrizes individuais padronizadas e integração das variáveis qualitativas e quantitativas por meio da transformação

dos dados pelos critérios da regra de Sturges e cálculo da raiz quadrada foi usado o programa Genes (CRUZ, 2014). As matrizes construídas para todas medidas de dissimilaridade utilizadas no estudo, foram correlacionadas sua significância foi calculada pelo teste t e o teste Z de Mantel com 10.000 permutações (MANTEL, 1967), no programa Genes (CRUZ, 2014). Nesse estudo o numero de grupos foi estabelecido pelo índice pseudo-t^2 do Pacote NbClust do programa estatístico R (R CORE TEAM, 2014).

RESULTADOS E DISCUSSÃO

Através da análise de agrupamento, foi observado que o método UPGMA foi o que melhor explicou a diversidade genética entre os acessos de tabaco, pois apresentou maior coeficiente de correlação cofenética quando comparado com o método Ward tanto para as matrizes obtidas com a análise simultânea dos dados qualitativos e quantitativos, quanto para as matrizes obtidas com a análise em isolado desses dados. O maior coeficiente de correlação apresentado para o método UPGMA foi em relação a matriz obtida pela distância euclidiana média e matriz conjunta 3 (Matriz C3), (0,96**) e o menor coeficiente de correlação foi obtido pela matriz de dissimilaridade dos dados qualitativos de origem obtida pela distância de Cole Rodgers com (0,87**). Essas estimativas indicam que há alta confiabilidade na representação dos dados de dissimilaridade para a realização dos agrupamentos (Tabela 1). Conforme sugerem Bussab et al. (1990), análises de agrupamento são aceitáveis se produzirem um coeficiente de correlação cofenético a partir de 0,80. Entretanto, outros autores como Rohlf & Fisher (1968), consideram como bons resultados valores superiores a 0,91.

Alguns estudos recentes realizados com culturas distintas, utilizando diferentes métodos de agrupamento, mostram que o método UPGMA foi o que obteve melhores resultados, como encontrados por Aramendiz-Tatis et al. (2011), estudando divergência genética entre genótipos de berinjela observaram que o método de agrupamento hierárquico UPGMA mostrou ser mais fidedigno do que os métodos Ward e Vizinho Mais Próximo (VMP), uma vez que comparando os valores de correlação cofenética entre os diferentes métodos, o UPGMA teve valor de 0,9, enquanto Ward e VMP tiveram valores de 0,7 e 0,8,

respectivamente. Essa maior fidedignidade do UPGMA em relação ao método de Ward e VMP pode ser explicado pelo fato de que este método é baseado em médias aritméticas, enquanto o VMP considera o menor valor entre dois genótipos e o Ward considera a menor soma de quadrados em cada etapa do processo de formação dos grupos.

Moura et al. (2010) analisando acessos de *Capsicum* spp., também observaram que o agrupamento hierárquico UPGMA foi mais confiável do que os agrupamentos Ward e VMP, uma vez que, comparativamente as inferências da correlação cofenética foram superiores, obtendo-se valores de 0,82, 0,59 e 0,62 para UPGMA, Ward e VMP, respectivamente. Rocha et al. (2009) estudando divergência genética entre acessos de tomateiro do grupo cereja observaram que o método de agrupamento UPGMA mostrou-se mais confiável que os métodos Ward e Vizinho Mais Próximo, pois, ao serem comparados os valores de correlação cofenética (CCC) entre os diferentes métodos, o UPGMA obteve um valor de 0,90, enquanto que Ward e VMP tiveram 0,36 e 0,89, respectivamente. Gonçalves et al. (2008), também trabalhando com acessos de tomates, obtiveram CCC maiores no método de agrupamento UPGMA em relação aos métodos de agrupamento vizinho mais próximo (SL) e WARD. Segundo Mohammad & Prasanna (2003), quanto maior CCC, menor será a distorção provocada ao agrupar os acessos.

Costa (2012), estudando diversidade genética em acessos de *Nicotiana tabacum* L., por meio da distância generalizada de Mahalanobis utilizando o método de agrupamento UPGMA, obteve o CCC de 0,95** e com a distância de Cole-Rodgers, 0,85** com o mesmo método, resultados estes, semelhantes aos obtidos no presente trabalho.

Segundo Rocha et al. (2010), a melhor adequação dos dados ao utilizar o método UPGMA, pode ser explicado pelo fato de que este método se baseia nas médias aritméticas das medidas de dissimilaridade, enquanto por exemplo o método Ward, considera a menor soma de quadrados em cada etapa do processo de formação dos grupos.

Tabela 1. Coeficientes de correlação cofenético (CCC) das matrizes de dissimilaridade, a partir de dados de caracteres quantitativos, quantitativos transformados e multicategóricos analisados individualmente e em simultâneo com utilização dos métodos UPGMA e WARD. Cruz das Almas-BA, 2014.

Matrizes	**CCC UPGMA**	**CCC WARD**
Quantitativos (Mahalanobis)	0,94**	0,93**
Quantitativos (Euclidiana Média)	0,96**	0,86**
Multicategóricos (Qualitativos Originais)	0,87**	0,85**
Quantitativos trans. (Regra de Sturges)	0,88**	0,73**
Quantitativos trans. (Raiz Quadrada)	0,88**	0,79**
Matriz C1 (Matriz Conjunta 1)	0,91**	0,84**
Matriz C2 (Matriz Conjunta 2)	0,92**	0,75**
Matriz C3 (Matriz Conjunta 3)	0,96**	0,80**
Matriz C4 (Matriz Conjunta 4)	0,91**	0,82**
Matriz C5 (Matriz Conjunta 5)	0,91**	0,83**
Matriz C6 (Matriz Conjunta 6)	0,89**	0,82**
Matriz C7 (Matriz Conjunta 7)	0,89**	0,82**

Matriz C1 (Algoritmo de Gower); Matriz C2 (soma de matriz Mahalanobis + matriz Cole Rodgers); Matriz C3 (soma de matriz Euclidiana Média + matriz Cole Rodgers); Matriz C4 (soma de matriz dados quantitativos transformados em qualitativos pela regra de Sturges + dados qualitativos originais); Matriz C5 (dados quantitativos transformados pela regra de Sturges anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, assumindo que todos os dados analisados são "qualitativos"); Matriz C6 (soma de matriz dados quantitativos transformados em qualitativos pelo cálculo da raiz quadrada + dados qualitativos originais); Matriz C7 (dados quantitativos transformados pelo cálculo da raiz quadrada integrados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers).

Segundo Cruz et al. (2011), quando o estudo de diversidade genética é feito a partir de vários tipos de variáveis, podem ser recomendadas diferentes estratégias de análise, como a conversão de todas as variáveis em um único padrão, fazendo-se uso por exemplo, da transformação das variáveis quantitativas em multicategóricas. Segundo Martins et al. (2011) os métodos mais utilizados para a distribuição de valores quantitativos em classes são a regra de Sturges (1926) e a da raiz quadrada.

Para esse trabalho foram utilizados dois critérios de transformação dos dados quantitativos em qualitativos: a Regra de Sturges e o cálculo da raiz quadrada, ambos baseados no número de observações para definir o número de classes. Observou-se que o número ideal de classes para a categorização dos dados quantitativos quando as classes são estabelecidas a partir da estratégia da raiz quadrada, para esse conjunto de dados o número de classes ideal foi quatro. Já para os dados transformados pela regra de Sturges, o número de classes ideal

foi cinco. Antes de apresentar o resultado das análises de agrupamento feitas com as variáveis transformadas, foi realizada uma comparação entre as variáveis quantitativas e as transformadas para observar qual transformação distorce menos os dados originais, considerando diferentes particularidades (Tabela 2).

Foi observado (Tabela 1), que as matrizes obtidas pelos dois critérios de transformação obtiveram ótima consistência em relação a matriz original alcançando maior magnitude pelo método UPGMA, ambas com o CCC de 0,88*. Conforme sugerem Bussab et al. (1990), análises de agrupamento são aceitáveis se produzirem um coeficiente de correlação cofenético a partir de 0,80. Na Tabela 2, ao comparar as matrizes transformadas com as demais matrizes individuais a partir dos dados originais, pode-se observar que o resultado mais eficiente foi obtido pelos dados transformados pelo cálculo da raiz quadrada, com uma correlação de 0,75** e 0,82** com a matriz calculada pelas distâncias de Mahalanobis e Euclidiana média, respectivamente. Portanto, esse critério é o que menos distorce os dados originais, evidenciando que não houve perda relevante de informações e que essa matriz a partir de dados transformados também poderá ser usada na análise de agrupamento.

Vale ressaltar que, quanto maior foi o número de classes utilizadas na codificação, menor foi a correlação com as matrizes a partir dos dados quantitativos originais, ou seja, a estratégia da raiz com o número de classes estimado em 4, obteve correlações de maior magnitude do que as obtidas pelo critério que utilizou a regra de Sturges, que estimou o número de classes em 5.

Mesmo com a melhor correlação obtida pela matriz de dissimilaridade gerada a partir do critério do cálculo da raiz quadrada, ambos os critérios serão utilizados a seguir na obtenção das matrizes para agrupar os acessos, para avaliar como estarão distribuídos esses indivíduos dentro dos grupos e entre os grupos com análise individual e simultânea dos dados, obviamente dando mais ênfase e respaldo a estratégia que obteve melhor resultado nessas correlações.

Para correlação entre as matrizes obtidas pelas estimativas de distância euclidiana média e generalizada de Mahalanobis foi encontrado uma correlação de alta magnitude 0,89**, o que evidencia boa concordância entre as medidas. Assim, há possibilidade de ambas formarem agrupamentos semelhantes (Tabela 2). Cargnelutti Filho et al. (2008), também observaram correlação linear

significativa e de alta magnitude (r= 0,92) entre as estimativas das distâncias euclidiana média padronizada (D) e generalizada de Mahalanobis (D^2) no estudo da divergência genética em cultivares de feijão. Coeficiente de correlação linear alto, de 0,97 entre as estimativas de distância euclidiana média e generalizada de Mahalanobis também foi encontrado em estudos sobre divergência genética em milho (CRUZ, 1990). De acordo com Cruz, (1990), concordância ou discordância entre as matrizes obtidas pelas estimativas de distância são dependentes da magnitude das correlações residuais que possam existir entre os caracteres considerados.

Nesse estudo também foi observado uma baixa correlação entre as matrizes obtidas a partir de caracteres quantitativos e multicategóricos de origem. Martins (2011) estudando acessos de tomate também obteve uma baixa correlação 0,47**, entre as matrizes obtidas da análise de caracteres quantitativos e multicategóricos. Ao avaliar a correlação entre medidas de dissimilaridade utilizadas na determinação da divergência genética de mandioca, para caracteres multicategóricos e quantitativos, Afonso (2013) observou coeficiente de correlação igual a 0,18. Gomes (2007) também estudando divergência genética de mandioca observou coeficiente de correlação igual a 0,09. Este autor atribuiu a inexistência de correlação entre as duas medidas de dissimilaridade à diferença do controle genético nos diferentes tipos de caracteres analisados.

Segundo Martins et al. (2011) a diversidade observada a partir de um conjunto de dados não pode ser extrapolada aos demais, em virtude dos baixos valores de correlação entre as matrizes de dissimilaridade, para cada conjunto de dados. A baixa correlação entre as matrizes oriundas de marcadores fenotípicos de natureza quantitativa e qualitativa para estimativas de dissimilaridade genética entre pares de acessos, indica que a combinação dos diferentes marcadores em uma única análise é recomendada para melhor distinguibilidade dos acessos estudados. Com isso, pode-se inferir que os dados analisados individualmente nesse trabalho para os caracteres quantitativos e qualitativos provavelmente irão produzir resultados divergentes nas análises de agrupamento, evidenciando dessa forma, a utilização da análise simultânea dos dados para melhor compreensão da variabilidade existente entre os acessos em estudo, pois a

análise simultânea contempla os diferentes tipos de caracteres em uma única análise.

Vale ressaltar que, encontrar um método de transformação que não acarrete em grande perda de informações é de grande valia para o pesquisador, pois segundo Barroso (2010), economizaria tempo e mão de obra na coleta de determinadas variáveis. Porém a adequação do uso de certos tipos de variáveis ainda é questionada. A possibilidade de transformação de um tipo de variável em outra é uma alternativa possível de ser adotada, mas suas consequências precisam ser investigadas.

A baixa correlação entre marcadores qualitativos e quantitativos para estimativas de dissimilaridade genética entre pares de genótipos indica que a combinação dos diferentes métodos é recomendada para maior acurácia na distinguibilidade dos acessos estudados.

Tabela 2. Coeficientes de correlação (r) entre matrizes de dissimilaridade, a partir de dados de caracteres quantitativos, quantitativos transformados (obtidos por diferentes estratégias de recodificação dos dados) e caracteres multicategóricos.

Matrizes comparadas	r
Quantitativos (Mahalanobis) vs. multicategóricos	$0,15^{ns}$
Quantitativos (Mahalanobis) vs. Euclidiana Média	0,89**
Quantitativos (Mahalanobis) vs. Quanti trans Sturges	0,63**
Quantitativos (Mahalanobis) vs. Quanti trans Raiz	0,75**
Quantitativos (Euclidiana Média) vs, multicategóricos	0,37**
Quantitativos (Euclidiana Média) vs. Quanti trans Sturges	0,75**
Quantitativos (Euclidiana Média) vs. Quanti trans Raiz	0,82**
Multicategóricos vs. Quanti trans Sturges	0,37**
Multicategóricos vs. Quanti trans Raiz	0,34**

**Significativo a 1% de probabilidade, pelo teste t e pelo teste de Mantel, baseado em 10.000 simulações, nas comparações de matrizes. Quantitativos (Mahalanobis); Quantitativos (Euclidiana Média padronizada); Quanti trans Sturges (dados quantitativos transformados pela regra de Sturges); Quanti trans Raiz (dados quantitativos, transformados pelo cálculo da Raiz quadrada).

Para cada matriz de dissimilaridade obtida individualmente com base nos grupos de caracteres quantitativos, multicategóricos e quantitativos transformados foi realizada a correlação com as matrizes obtidas pelas diferentes metodologias de análise simultânea dos caracteres em estudo, sendo essa correlação testada pelo teste t e pelo teste pelo teste Z de Mantel (Tabela 3).

Apesar da baixa concordância entre algumas matrizes individuais (Tabela 2), ao proceder a soma destas matrizes de dissimilaridade, de integrar os dados por meio das técnicas de transformações e uso do Algoritmo de Gower, ou seja, correlacionando essas matrizes obtidas através de metodologias distintas de análise conjuntas com as matrizes individuais, encontrou-se na maioria das correlações valores superiores a 0,70, sendo significativos estatisticamente. Este resultado evidencia que a análise simultânea de dados por diferentes metodologias para análise da diversidade, pode ser uma importante ferramenta para reunir conjuntamente as informações independentemente de sua natureza em uma só análise, uma vez que elas mantêm elevada correlação com grande parte das matrizes individuais (Tabela 3).

Tabela 3. Coeficientes de correlação (r) entre matrizes de dissimilaridade, a partir de dados de caracteres quantitativos, quantitativos transformados e multicategóricos em relação as matrizes conjuntas.

Matrizes comparadas	**r**
Quantitativos (Mahalanobis) vs. Matriz C1	0,59**
Quantitativos (Mahalanobis) vs. Matriz C2	0,82**
Quantitativos (Mahalanobis) vs. Matriz C3	0,68**
Quantitativos (Mahalanobis) vs. Matriz C4	0,50**
Quantitativos (Mahalanobis) vs. Matriz C5	0,46**
Quantitativos (Mahalanobis) vs. Matriz C6	0,59**
Quantitativos (Mahalanobis) vs. Matriz C7	0,56**
Multicategóricos vs. Matriz C1	0,86**
Multicategóricos vs. Matriz C2	0,69**
Multicategóricos vs. Matriz C3	0,78**
Multicategóricos vs. Matriz C4	0,78**
Multicategóricos vs. Matriz C5	0,84**
Multicategóricos vs. Matriz C6	0,77**
Multicategóricos vs. Matriz C7	0,81**
Quantitativos (Euclidiana Média) vs. Matriz C1	0,78**
Quantitativos (Euclidiana Média) vs. Matriz C2	0,86**
Quantitativos (Euclidiana Média) vs. Matriz C3	0,87**
Quantitativos (Euclidiana Média) vs. Matriz C4	0,70**
Quantitativos (Euclidiana Média) vs. Matriz C5	0,67**
Quantitativos (Euclidiana Média) vs. Matriz C6	0,75**
Quantitativos (Euclidiana Média) vs. Matriz C7	0,73**
Quanti trans Sturges vs. Matriz C1	0,67**
Quanti trans Sturges vs. Matriz C2	0,68**
Quanti trans Sturges vs. Matriz C3	0,70**

Quanti trans Sturges vs. Matriz C4	0,87**
Quanti trans Sturges vs. Matriz C5	0,82**
Quanti trans Sturges vs. Matriz C6	0,74**
Quanti trans Sturges vs. Matriz C7	0,72**
Quanti trans Raiz vs. Matriz C1	0,68**
Quanti trans Raiz vs. Matriz C2	0,75**
Quanti trans Raiz vs. Matriz C3	0,73**
Quanti trans Raiz vs. Matriz C4	0,71**
Quanti trans Raiz vs. Matriz C5	0,67**
Quanti trans Raiz vs. Matriz C6	0,86**
Quanti trans Raiz vs. Matriz C7	0,83**

** * : Significativo a 1 e 5% de probabilidade pelo teste t. ++ + : Significativo a 1 e 5% de probabilidade pelo teste de Mantel baseado em 10000 simulações. Matriz C1 (Algoritmo de Gower); Matriz C2 (soma de matriz Mahalanobis + matriz Cole Rodgers); Matriz C3 (soma de matriz Euclidiana Média + matriz Cole Rodgers); Matriz C4 (soma de matriz dados quantitativos transformados em qualitativos pela regra de Sturges + dados qualitativos originais); Matriz C5 (dados quantitativos transformados pela regra de Sturges anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, assumindo que todos os dados analisados são "qualitativos"); Matriz C6 (soma de matriz dados quantitativos transformados em qualitativos pelo cálculo da raiz quadrada + dados qualitativos originais); Matriz C7 (dados quantitativos transformados pelo cálculo da raiz quadrada anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers).

Para a determinação da divergência genética, a decisão sobre qual medida de dissimilaridade e conjunto de variáveis que serão utilizadas, depende do objetivo da pesquisa. Contudo, na caracterização e avaliação dos acessos, variáveis quantitativas e qualitativas analisadas conjuntamente proporcionam maior acurácia na identificação de indivíduos contrastantes, pois em uma única análise é contemplado caracteres de naturezas distintas.

Um dos principais problemas na análise de agrupamentos é a definição do melhor algoritmo e do número total de grupos a considerar. Encontra-se na literatura uma série de critérios que podem auxiliar na decisão final (MINGOTI, 2005; HAIR Jr. et al., 2005). Nesse estudo, para definição do número de grupos foi usado o índice Pseudo-t^2 para os agrupamentos obtidos pelo método UPGMA e Ward. Através das distâncias de Mahalanobis, Euclidiana Média, pela distância de Cole-Rodgers a partir dos dados qualitativos originais e quantitativos transformados e pelas distâncias oriundas das metodologias de análise conjunta (Algoritmo de Gower, Soma Algébrica de matrizes individuais e integração de dados por meio de estratégias de transformação (Tabela 4).

Para as distâncias de Mahalanobis, Euclidiana média e dados quantitativos pela estratégia da raiz quadrada houve a formação de 2 grupos distintos e com mesma distribuição dos indivíduos dentro dos grupos, levando em consideração

os agrupamentos obtidos pelo método UPGMA. Apesar de ser observado também a formação de 2 grupos para os acessos que tiveram suas variáveis transformadas pela Regra de Sturges, houve discordância na alocação de um dos indivíduos dentro do grupo em relação as outras distâncias, inclusive da obtida pela estratégia da raiz quadrada. Levando em consideração a Regra de Sturges o indivíduo A14 (125 PD) foi alocado para o grupo I e para as demais distâncias, esse indivíduo esteve presente no grupo II (Tabela 4 e 5). Com isso, se pode inferir que a transformação realizada a partir do cálculo da raiz quadrada fez a distribuição dos indivíduos de maneira similar as matrizes pela distância de Mahalanobis e Euclidiana Média, onde a maior similaridade obtida por esse critério em relação a essas matrizes de distâncias citadas já era esperada devido as maiores correlações obtida entre as mesmas (Tabela 2). Esperava-se que os métodos de transformação de dados que mantivessem um padrão semelhante ao original tivessem bons desempenhos. E foi exatamente o que ocorreu quando agrupados pelo método UPGMA.

Para o agrupamento a partir dos caracteres multicategóricos originais observou-se a formação de 3 grupos, também levando em consideração o método UPGMA (Tabela 4 e 5). Como era esperado, houve discordância no número de grupos formados para com as outras medidas de análise individual analisadas, pois estas apresentaram baixa correlação quando comparadas (Tabela 2). A baixa correlação entre essas medidas de dissimilaridade segundo Gomes (2007), estudando divergência genética em acessos de mandioca, possivelmente está na diferença do controle genético nos diferentes tipos de caracteres analisados. No presente estudo, nota-se que essa diferença, acabou refletindo nos agrupamentos formados, havendo discordância no número de grupos e na distribuição dos indivíduos dentro dos grupos e entre os grupos.

Com base em cada conjunto de caracteres em isolado (quantitativo ou multicategórico), mesmo quando o número de grupos formados foi o mesmo, como foi observado acima, houve discordância entre os procedimentos de agrupamento realizados. Desta forma percebe-se que mesmo quando os dados são analisados isoladamente, há uma dificuldade na análise e interpretação dos resultados de caracterização e avaliação dos genótipos, pois na maioria das vezes resulta na incompleta distinção entre os mesmos.

Como foi observado, a avaliação da diversidade entre os acessos de tabaco inicialmente foi realizada para cada conjunto de caracteres individualmente, e indicou que a divergência baseada em qualquer um dos conjuntos de dados em isolado não reflete a variabilidade total existente. Esse resultado, realmente é um forte indicativo da necessidade de estudar esses dados de maneira simultânea, obtendo assim respostas que realmente contemplem os diferentes tipos de caracteres em uma única análise, para obtenção de respostas mais precisas e abrangentes em relação ao estudo da divergência genética desses acessos.

No Geral, com a utilização do índice pseudo-t^2 para definição do número ótimo de grupos, observou-se uma variação de 2 a 4 grupos formados para as análises em isolado e de 3 a 4 grupos para as estratégias de análises simultâneas utilizadas, levando em consideração os dois métodos de agrupamento, UPGMA e WARD. (Tabela 4). Contudo a análise simultânea dos dados para todas as metodologias testadas pelo método UPGMA apresentam resultados mais consistentes pelo fato do coeficiente de correlação cofenético ter sido de maior magnitude para este método, (Tabela 1), explicando melhor a diversidade existente entre os acessos de tabaco estudados. Com isso foi observado a formação de 3 grupos para todas as metodologias de análise conjunta combinadas a esse método de agrupamento, vale ressaltar também, que houve uma distribuição similar dos indivíduos dentro dos grupos formados, (Tabela 4 e 5), onde ficaram alocados no primeiro grupo 8 acessos: A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A8 (ER 13-065) e A9 (ER 28-027) o que representa, aproximadamente, 53,3% do total dos acessos caracterizados. Esse grupo é caracterizado por possuir valores médios intermediários para todas as variáveis quantitativas analisadas em relação aos demais grupos, e por apresentar plantas predominantemente com coloração do caule verde claro a verde médio, coloração das folhas verde clara a verde médio, coloração da nervura central (face inferior) verde esbranquiçada a esbranquiçada, superfície da lamina foliar fraco e médio, perfil longitudinal da folha moderadamente recurvado, margem da lâmina foliar ausente ou muito fraca, ponta da lâmina foliar medianamente pontiaguda e cor da corola que vai do rosa médio ao rosa forte.

O segundo grupo foi representado por 33,3% dos acessos: A7 (ER 13-061); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023) e A14 (125 PD), é caracterizado principalmente por ser constituído pela a maioria dos acessos com maior número de folhas e menores valores médios para o comprimento dos internódios. Esse grupo também é caracterizado por apresentar acessos com coloração do caule verde claro a verde médio, coloração das folhas verde clara a verde escura, coloração da nervura central (face inferior) verde esbranquiçado a esbranquiçada, superfície da lamina foliar fraco e médio, perfil longitudinal da folha ligeiramente recurvado, margem da lâmina foliar ausente ou muito fraca e fraca, ponta da lâmina foliar ligeiramente pontiaguda, medianamente pontiaguda a fortemente pontiaguda e cor da corola rosa clara.

E por fim o terceiro grupo, constituído pelos acessos A13 (109 PD) e A15 (221 PD) com percentual de 13,3% do total de acessos estudados. Esse grupo é caracterizado por possuir acessos com os menores valores médios para altura de planta, número de folhas, diâmetro médio do caule, e também por ter os maiores valores médios para largura da base da 10ª folha e para o comprimento dos internódios. Esse grupo também é definido por apresentar plantas predominantemente com coloração do caule verde claro, coloração das folhas verde médio, coloração da nervura central (face inferior) do verde esbranquiçada, superfície da lamina foliar fraco, perfil longitudinal da folha ligeiramente recurvado, margem da lâmina foliar ausente ou muito fraca, ponta da lâmina foliar ligeiramente pontiaguda e cor da corola rosa clara (Tabela 4 e 5).

Ainda em relação as análises simultâneas, os agrupamentos em que transformação de dados esteve envolvida, foram observados ótimos resultados, onde foram obtidos 3 grupos e com distribuição dos indivíduos em cada grupo similar aos obtidos pelas outras metodologias de análises conjuntas estudadas. Assim, esse resultado é um indicativo de que a transformação das variáveis pelos critérios da regra de Sturges e principalmente pelo cálculo da raiz quadrada não afetou o estudo da diversidade genética para a análise conjunta dos dados que é feito no contexto multivariado, sendo uma alternativa promissora para análise simultânea de dados em estudo com tabaco (*Nicotiana tabacum* L.).

Entretanto, deverão ser realizados mais estudos com variações no número de acessos e variáveis para afirmar que essas transformações sempre

apresentarão eficácia nos agrupamentos, pois a adequação do uso de certos tipos de variáveis ainda é questionável. Segundo Barroso (2010), a possibilidade de transformação de um tipo de variável em outro é uma alternativa possível de ser adotada pelo pesquisador, mas suas consequências precisam ser investigadas.

Embora a estrutura geral dos agrupamentos em que as estratégias de análise simultânea estiveram envolvidas tenham sido bastante similares nos dois métodos de agrupamento (UPGMA e WARD), com auxílio do índice pseudo-t^2, ainda foram observadas pequenas alterações no modo como os indivíduos foram distribuídos entre os grupos. Isso mostra a importância de escolher o método com melhor consistência, pois este explicará melhor a divergência existente, onde, a aplicação de métodos de agrupamentos indiscriminadamente, sem utilizar critérios que possam aferir sua consistência poderão levar a padrões de agrupamentos equivocados.

Segundo Cargnelutti Filho (2008), do ponto de vista do melhorista de plantas, o processamento dos dados por diversos métodos de agrupamento e com base em diversas medidas de dissimilaridade, e a consideração das particularidades de cada um, é adequada para uma melhor tomada de decisão em relação à escolha de cultivares para cruzamentos. Portanto, é interessante despertar o interesse por pesquisas que expõem a combinação de métodos, uma vez que podem resultar num aperfeiçoamento dos resultados de uma análise (ALVES, 2007).

Uma análise simultânea, quando bem estruturada, pode revelar melhores respostas para uma análise. Contudo, a especificação de uma determinada estratégia não implica em abrir mão da combinação de duas ou mais estratégias para atingir os objetivos, pois essas combinações podem potencializar de forma considerável a interpretação dos resultados.

Tabela 4. Número de Grupos definido a partir do Índice Pseudo-t^2 do pacote NbClust do Programa R, utilizando-se de dois métodos de agrupamentos distintos. Cruz das Almas-BA, 2014.

Matrizes	Número de grupos: NbClust – Índice Pseudot^2			
	Grupos - Método (UPGMA)	Valor do índice	Grupos - Método (WARD)	Valor do índice
Quantitativos (Mahalanobis)	2	1,4288	3	1,4288
Quantitativos (Euclidiana Média)	2	1,4312	3	1,4312
Multicategóricos (Qualitativos Originais)	3	0,1914	4	0,1914
Quantitativos trans. (Regra de Sturges)	2	4,0677	2	4,0889
Quantitativos trans. (Raiz Quadrada)	2	1,03	2	3,9171
Matriz C1 (Matriz Conjunta 1)	3	0,1083	3	2,4147
Matriz C2 (Matriz Conjunta 2)	3	-0,2869	3	3,2262
Matriz C3 (Matriz Conjunta 3)	3	0,6591	3	0,6591
Matriz C4 (Matriz Conjunta 4)	3	0,0743	4	0,0743
Matriz C5 (Matriz Conjunta 5)	3	-0,0853	3	4,6899
Matriz C6 (Matriz Conjunta 6)	3	0,2762	3	0,2762
Matriz C7 (Matriz Conjunta 7)	3	0,5886	3	0,5886

Matriz C1 (Algoritmo de Gower); Matriz C2 (soma de matriz Mahalanobis + matriz Cole Rodgers); Matriz C3 (soma de matriz Euclidiana Média + matriz Cole Rodgers); Matriz C4 (soma de matriz dados quantitativos transformados em qualitativos pela regra de Sturges + dados qualitativos originais); Matriz C5 (dados quantitativos transformados pela regra de Sturges anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, assumindo que todos os dados analisados são "qualitativos"); Matriz C6 (soma de matriz dados quantitativos transformados em qualitativos pelo cálculo da raiz quadrada + dados qualitativos originais); Matriz C7 (dados quantitativos transformados pelo cálculo da raiz quadrada anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers).

Tabela 5. Agrupamento dos acessos por meio dos métodos UPGMA e WARD utilizando o índice pseudo-t^2 do pacote NbClust do Software R. Cruz das Almas-BA, 2014.

Matrizes	Agrupamento dos Acessos														
	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
Quantitativos (Mahalanobis) - UPGMA	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
Quantitativos (Mahalanobis) - WARD	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3
Quantitativos (Euclidiana Média) - UPGMA	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
Quantitativos (Euclidiana Média) - WARD	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3
Multicategóricos (Qualitativos Originais) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	3	3	2	3
Multicategóricos (Qualitativos Originais) - WARD	1	2	2	2	1	2	3	1	2	3	3	4	4	3	4
Quantitativos trans. (Regra de Sturges) - UPGMA	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2
Quantitativos trans. (Regra de Sturges) - WARD	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2
Quantitativos trans. (Raiz Quadrada) - UPGMA	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
Quantitativos trans. (Raiz Quadrada) - WARD	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
Matriz C1 (Matriz Conjunta 1) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C1 (Matriz Conjunta 1) - WARD	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C2 (Matriz Conjunta 2) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C2 (Matriz Conjunta 2) - WARD	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C3 (Matriz conjunta 3) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C3 (Matriz conjunta 3) - WARD	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C4 (Matriz Conjunta 4) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C4 (Matriz Conjunta 4) - WARD	1	2	1	1	2	1	3	1	2	3	3	3	4	3	4
Matriz C5 (Matriz Conjunta 5) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C5 (Matriz Conjunta 5) - WARD	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C6 (Matriz Conjunta 6) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C6 (Matriz Conjunta 6) - WARD	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C7 (Matriz Conjunta 7) - UPGMA	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3
Matriz C7 (Matriz Conjunta 7) - WARD	1	1	1	1	1	1	2	1	1	2	2	2	3	2	3

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Na Tabela 6 são apresentados os acessos mais divergentes e os mais similares a partir das matrizes de dissimilaridades utilizadas nesse estudo. O acesso A14 (125 PD) pode ser considerado o mais divergente em relação aos demais acessos, levando em consideração os resultados de todas as matrizes conjuntas, exceto para Matriz C2 que usou a metodologia da soma algébrica de matrizes (soma de matriz Mahalanobis + matriz Cole Rodgers) que teve maior divergência observada entre os acessos A15 (221 PD) $_x$ A11 (ER 33-022), (Tabela 6). Os indivíduos mais similares entre si para todas as metodologias aplicadas para analise simultânea foram os acessos A15 (221 PD) $_x$ A13 (109 PD), exceto para matriz C2 (matriz conjunta 2). Os acessos A15 (221 PD) $_x$ A13 (109 PD) não houve variação para nenhuma das classes dos caracteres qualitativos analisados entre esses dois acessos, possuem em comum também os menores valores para altura da planta; menor número de folhas; maiores valores para largura da lamina da 10ª folha; maiores valores para comprimento dos internódios.

Como foi observado também em relação ao acesso mais divergente, a matriz Conjunta C2 em questão apresentou maior similaridade entre acessos diferentes aos observados nas demais metodologias de análise simultânea, sendo que para essa matriz, os mais similares foram os acessos A8 (ER 13-065) e o A5 (ER 05-070), (Tabela 6). De acordo com Cruz & Regazzi (2004), deve-se evitar a escolha de indivíduos de mesmo padrão de dissimilaridade nos cruzamentos, de modo a não restringir a variabilidade genética e, assim, evitar reflexos negativos nos ganhos a serem obtidos pela seleção. Para a seleção dos genitores para o cruzamento, deve-se aliar o bom desempenho destes em relação a características de interesse com a divergência genética entre eles.

É interessante salientar que entre as maiores distâncias encontradas, o acesso A14 (125 PD) esteve presente em quase todas as combinações em relação as matrizes de análise conjunta, mostrando-se um indivíduo bastante divergente dos demais acessos. Sendo o acesso A14 o único que possui a ponta da lâmina foliar fortemente pontiaguda, apresenta também as maiores médias para o comprimento das folhas e as menores médias para o comprimento dos internódios e para o engrossamento do tubo da flor e os menores valores para ângulo de inserção da 10ª folha.

Em relação as análises individuais dos dados, para os caracteres quantitativos observou-se que os acessos mais divergentes foram o A15 (221 PD) x A11 (ER 33-022) para matriz de Mahalanobis e entre o A15 (221 PD) x A12 (ER 33-023) para matriz Euclidiana média. Os mais similares para ambas as matrizes quantitativas foram os acessos A2 (ER 04-090) x A9 (ER 28-027).

Para os dados qualitativos (Multicategóricos) não houve uma separação de destaque para um par de acessos mais divergentes e mais similares, sendo observado várias combinações de acessos mais distintos, sendo eles: A10 x A9; A11 x A9 e entre os acessos A14 x (A1, A5, A8), em relação aos acessos mais similares A4 x A2; A8 x A5; A11 x A10 e entre os acessos A15 x A13, muitas combinações de mesma distância podem dificultar a observação do acesso mais promissor para possíveis cruzamentos, levando em consideração características complementares entre os mesmos, lembrando que nem sempre o acesso mais divergente é o mais indicado em cruzamentos, pois este, deve ser além de divergente, complementar. A mesma situação ocorreu para os dados quantitativos transformados em multicategóricos pela regra de Sturges e regra da raiz quadrada. A única semelhança entre essas matrizes obtidas de dados transformados e qualitativos originais pode ser observada em apenas uma das várias combinações apresentadas para os acessos mais similares, ou seja, entre os acessos A15 (221 PD) x A13 (109 PD), que também foram os mais similares para quase todas as matrizes obtidas das análises simultâneas (Tabela 6). Todas as matrizes obtidas nesse estudo, tanto as matrizes que levaram em consideração as análises em isolado quanto as com mistura de variáveis encontram-se em Anexo.

Os dados de diferentes origens, analisados individualmente divergiram quanto aos acessos mais próximos e mais distantes, o que já era esperado, pois as análises em separado de dados de naturezas diferentes tendem a gerar combinações desconcordantes. A partir dessas observações, nota-se mais uma vez a importância da utilização de metodologias que contemplem todas as características em uma única análise, independentemente de sua natureza, pois resultará em respostas mais precisas quanto a variabilidade existente em um conjunto de indivíduos.

A variabilidade observada entre os acessos de tabaco para algumas características estudadas mostram a possibilidade de seleção de indivíduos para o melhoramento do tabaco, considerando características como o número de folhas, comprimento de internódios e largura da base da lâmina foliar.

Tabela 6. Distâncias máximas e mínimas entre os acessos de tabaco estudados, a partir das matrizes de dissimilaridade obtidas por meio de diferentes metodologias.

Matrizes de dissimilaridade		Acessos mais divergentes	Acessos mais similares
Conjunta	Matriz C1	A14 x A1	A15 x A13
	Matriz C2	A15 x A11	A8 x A5
	Matriz C3	A14 x A1	A15 x A13
	Matriz C4	A14 x (A5 x A8)	A15 x A13
	Matriz C5	A14 x (A5 e A8)	A15 x A13
	Matriz C6	A14 x A9	A15 x A13
	Matriz C7	A14 x (A1, A5, A8 e A9)	A15 x A13
Quantitativos	Matriz Mahalanobis	A15 x A11	A9 x A2
	Matriz Euclidiana Média	A15 x A12	A9 x A2
Quantitativos transformados	Matriz quanti trans Sturges (Cole Rodgers)	A12 x A6 A13 x (A4, A6, A11 e A12) A14 x A7 A15 x (A4, A6, A11 e A12	A5 x A2 A9 x A2 A15 x A13
	Matriz quanti trans Raiz (Cole Rodgers)	A13 x (A11 e A12) A14 x A7 A15 x (A11 e A12)	A4 x A1 A15 x A13
Qualitativos	Matriz Cole Rodgers	A10 x A9 A11 x A9 A14 x (A1, A5, A8)	A4 x A2 A8 x A5 A11 x A10 A15 x A13

Matriz C1 (Algoritmo de Gower); Matriz C2 (soma de matriz Mahalanobis + matriz Cole Rodgers); Matriz C3 (soma de matriz Euclidiana Média + matriz Cole Rodgers); Matriz C4 (soma de matriz dados quantitativos transformados em qualitativos pela regra de Sturges + dados qualitativos originais); Matriz C5 (dados quantitativos transformados pela regra de Sturges anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, assumindo que todos os dados analisados são "qualitativos"); Matriz C6 (soma de matriz dados quantitativos transformados em qualitativos pelo cálculo da raiz quadrada + dados qualitativos originais); Matriz C7 (dados quantitativos transformados pelo cálculo da raiz quadrada anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers); A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Como a análise conjunta propiciou melhores interpretações acerca da variabilidade existente entre os acessos de tabaco em estudo, foram comparadas as matrizes de análise simultânea de dados (Tabela 7). Foi observado que houve estimativas de correlação elevada e significativa entre todas as matrizes conjuntas. Essa aparente correspondência foi confirmada pelo alto valor da correlação, que variou de r= 0,824** a r= 0,998, significativo 1 e 5% de probabilidade pelo teste de t e pelo teste Z de Mantel, entre as matrizes obtidas pelas estratégias de integração. Vale ressaltar que as correlações obtidas foram acima de (0,80**), apresentando bastante consistência nos resultados das comparações entre as matrizes. As maiores correlações foram obtidas entre as matrizes conjuntas: Matriz C6 x Matriz C7; Matriz C4 x Matriz C5 e Matriz C1 x Matriz C3. Permitindo inferir que as análises de diversidade genética em tabaco considerando simultaneamente características quantitativas e qualitativas podem ser realizadas utilizando como medida de dissimilaridade o algoritmo de Gower, a soma algébrica de matrizes individuais e por meio da integração de dados pela transformados de caracteres quantitativos (Tabela 7 e Matrizes em anexo).

Martins (2011) trabalhando com caracterização de germoplasma de tomate constatou, em estudos de integração de dados, que a estratégia de conversão de dados foi considerada melhor quando comparada à soma algébrica de matrizes individuais, uma vez que os grupos formados apresentaram maior correspondência com sua origem, e ainda permitiu uma maior discriminação dos acessos.

As combinações entre as estratégias de análises simultâneas de dados e o método de agrupamento UPGMA com auxílio do critério do índice pseudo-t^2, mostraram-se eficientes para discriminar os acessos em estudo. Apresentando resultados semelhantes, como foi observado anteriormente e apresentando alta correlação. Pode-se inferir então, que o uso de qualquer uma das estratégias de análise simultânea ou o uso combinado das mesmas, proporciona maior acurácia na discriminação do conjunto de indivíduos em estudo, em relação as análises em isolados.

Tabela 7. Correlação entre matrizes de dados quantitativos e qualitativos analisados simultaneamente por diferentes metodologias.

Matrizes	Correlação (r)	Nº de dados	Valor de t	Probabilidade (%)	Teste de Mantel		
					Níveis críticos 1%	Níveis críticos 5%	Significância
Matriz **C1** x Matriz **C2**	0,935	105	26,70	.0**	-0,051	0,025	++
Matriz **C1** x Matriz **C3**	0,988	105	64,94	.0**	-0,060	0,020	++
Matriz **C1** x Matriz **C4**	0,906	105	21,77	.0**	-0,008	0,018	++
Matriz **C1** x Matriz **C5**	0,926	105	24,82	.0**	-0,036	0,043	++
Matriz **C1** x Matriz **C6**	0,929	105	25,49	.0**	-0,096	-0,029	++
Matriz **C1** x Matriz **C7**	0,942	105	28,48	.0**	-0,059	0,004	++
Matriz **C2** x Matriz **C3**	0,949	105	30,39	.0**	-0,049	0,011	++
Matriz **C2** x Matriz **C4**	0,824	105	14,77	.0**	-0,013	0,038	++
Matriz **C2** x Matriz **C5**	0,828	105	14,97	.0**	-0,054	0,008	++
Matriz **C2** x Matriz **C6**	0,942	105	28,48	.0**	-0,059	0,004	++
Matriz **C2** x Matriz **C7**	0,880	105	18,82	.0**	-0,037	0,009	++
Matriz **C3** x Matriz **C4**	0,888	105	19,59	.0**	-0,067	-0,008	++
Matriz **C3** x Matriz **C5**	0,897	105	20,60	.0**	-0,097	-0,019	++
Matriz **C3** x Matriz **C6**	0,919	105	23,70	.0**	-0,076	-0,012	++
Matriz **C3** x Matriz **C7**	0,925	105	24,66	.0**	-0,101	-0,014	++
Matriz **C4** x Matriz **C5**	0,995	105	105,28	.0**	-0,045	-0,045	++
Matriz **C4** x Matriz **C6**	0,906	105	21,75	.0**	-0,069	0,007	++
Matriz **C4** x Matriz **C7**	0,913	105	22,71	.0**	-0,086	-0,023	++
Matriz **C5** x Matriz **C6**	0,911	105	22,47	.0**	-0,060	0,012	++
Matriz **C5** x Matriz **C7**	0,923	105	24,42	.0**	-0,088	-0.011	++
Matriz **C6** x Matriz **C7**	0,998	105	156.20	.0**	-0,073	-0.006	++

** * : Significativo a 1 e 5% de probabilidade pelo teste t. ++ + : Significativo a 1 e 5% de probabilidade pelo teste de Mantel baseado em 10000 simulações. Matriz C1 (Algoritmo de Gower); Matriz C2 (soma de matriz Mahalanobis + matriz Cole Rodgers); Matriz C3 (soma de matriz Euclidiana Média + matriz Cole Rodgers); Matriz C4 (soma de matriz dados quantitativos transformados em qualitativos pela regra de Sturges + dados qualitativos originais); Matriz C5 (dados quantitativos transformados pela regra de Sturges anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, assumindo que todos os dados analisados são "qualitativos"); Matriz C6 (soma de matriz dados quantitativos transformados em qualitativos pelo cálculo da raiz quadrada + dados qualitativos originais); Matriz C7 (dados quantitativos transformados pelo cálculo da raiz quadrada anexados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers).

Em estudos de diversidade genética os dois conjuntos de caracteres, quantitativos e multicategóricos, são importantes para a caracterização da diversidade de determinado conjunto de indivíduos. Com isso, se faz necessário a identificação e utilização de técnicas que contemplem todas as características, em uma única análise, independentemente de sua natureza.

A integração de dados por meio da transformação de dados quantitativos em qualitativos e a soma algébrica de matrizes individuais são alternativas viáveis em análises de agrupamento com mistura de variáveis em tabaco (*Nicotiana tabacum* L.).

CONCLUSÕES

O método de agrupamento hierárquico UPGMA foi o que melhor explicou a divergência genética dos acessos nesse trabalho.

Na codificação de dados quantitativos em multicategóricos, destaca-se a estratégia da raiz quadrada.

A integração de dados de diferentes naturezas para estudo de diversidade genética pode ser realizada com sucesso pela conversão dos dados quantitativos em multicategóricos, com uso da distância de Cole-Rodgers como medida de dissimilaridade.

A análise de divergência genética em acessos de tabaco a partir da integração dos dados via transformação, soma algébrica de matrizes individuais e pelo algoritmo de Gower resultaram em grupos correspondentes, evidenciando que a combinação de métodos resultou no aprimoramento dos resultados obtidos.

As estratégias aplicadas para análise conjunta dos dados mostraram-se eficazes na distinção dos acessos, sem contudo perder as informações obtidas por cada avaliação em separado e apresentaram alta correlação quando comparadas.

O acesso 125 PD (A14) possui comportamento distinto dos demais acessos e as metodologias de análise simultânea, com base nas matrizes de distâncias geradas, captaram essa divergência.

AGRADECIMENTOS

À Universidade Federal do Recôncavo da Bahia, a empresa Ermor Tabarama Tabacos do Brasil Ltda, pela parceria e infraestrutura e a Fundação de Amparo à Pesquisa do Estado da Bahia (Fapesb) pela concessão da bolsa de mestrado que permitiu o desenvolvimento deste trabalho.

REFERÊNCIAS

ADAPAR - Agência de Defesa Agropecuária do Paraná. Disponível em:< http://www.adapar.pr.gov.br/modules/noticias/article.php?storyid=178>. Acesso em: 17 fev. 2015.

AFONSO, S. D. J. **Seleção de descritores morfológicos e divergência genética em acessos de mandioca.** Dissertação de Mestrado em Recursos Genéticos Vegetais, Universidade Federal do Recôncavo da Bahia, Cruz das Almas, BA, Brasil. Dezembro, 2013.

ALVES, L. B. Tratamento multivariado de dados por análise de correspondência e análise de agrupamentos. **Anais do 13º Encontro de Iniciação Científica e Pós-Graduação do ITA – XIII ENCITA / Instituto Tecnológico de Aeronáutica**. São José dos Campos, SP, Brasil, Disponivel em: <http://www.bibl.ita.br/xiiiencita/MEC17.pdf>. 2007.

ARAMENDIZ-TATIS, H.; SUDRE, C. P.; GONCALVES, L. S.A.; RODRIGUES, R. Potencial agronômico e divergência genética entre genótipos de berinjela nas condições do Caribe Colombiano. **Hortic. Bras**. [online]. vol.29, n.2, pp. 174-180. 2011.

BARROSO, N. C. **Categorização de dados quantitativos para estudos de diversidade genética.** Dissertação de Mestrado em Estatística Aplicada e Biometria, Universidade Federal de Viçosa, Minas Gerais – Brasil. Dezembro, p. 99, 2010.

BUSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D.F. **Introdução à análise de agrupamento**. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, São Paulo. **Anais ...** São Paulo: ABE, p. 105, 1990.

CARGNELUTTI FILHO, A.; RIBEIRO, N. D.; REIS, R. C. P. dos; SOUZA, J. R. de; JOST, E. Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão. **Ciência Rural**, Santa Maria, v.38, n.8, p. 2138-2145, 2008.

COLE-RODGERS, P.; SMITH, D. W.; BOSLAND, P. W. A novel statistical approach to analyze genetic resource evaluations using *Capsicum* as an example. **Crop Science**, v. 37, n. 3, p. 1000-1002, 1997.

CONCEIÇÃO, A. L. da S.; SILVA, M. dos S. da.; SANTOS, C. C. dos.; Araujo, G. de M.; Moreira, R. F. C. Variabilidade genética e importância relativa de caracteres em acessos de tabaco (*Nicotiana tabacum* L.) Tipo broad leaf por meio de marcadores fenotípicos. **Enciclopédia Biosfera**, Centro Científico Conhecer - Goiânia, v.10, n.19; p.1900-1907, Dez. 2014.

COSTA, T. P. P. **Caracterização Morfoagronômica de Genótipos de Tabaco na Região do Recôncavo da Bahia.** Dissertação de Mestrado em Recursos Genéticos Vegetais, Universidade Federal do Recôncavo da Bahia, Cruz das Almas, BA, Brasil. Maio, 2012.

CRUZ, C.D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas**. Tese (Doutorado em Agronomia) - Programa de Pós-graduação em Genética e Melhoramento de Plantas, Escola Superior de Agricultura Luiz de Queiroz, 1990.

CRUZ, C.D., CARNEIRO, P.C.S. Modelos biométricos aplicados ao melhoramento genético. Viçosa: UFV, v.2, p. 585, 2003.

CRUZ, C. D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. 3. ed. Viçosa: UFV, 480 p, 2004.

CRUZ, C.D.; FERREIRA, F.M.; PESSONI, L.A.; **Biometria aplicada ao estudo da diversidade genética.** Visconde do Rio Branco-MG, Suprema, 620p, 2011.

CRUZ, C.D. Programa Genes - **Aplicativo computacional em genética e estatística**. Disponível em: <www.ufv.br/dbg/genes/genes.htm>. 2014.

DARVISHZADEH, R.; MIRZAEI, L.; MALEKI, H. H.; LAURENTIN, H.; ALAVI, S. R. *Genetic variation in oriental tobacco (Nicotiana tabacum* L*.) by agro-morphological traits and simple sequence repeat markers.* **Revista Ciência Agronômica**, vol.44, n. 2, ISSN 1806-6690, 2013.

DAVALIEVA, K.; MALEVA, I.; FILIPOSKI, K.; SPIROSKI, O.; EFREMOV, G. D. Genetic variability of Macedonian tobacco varieties determined by microsatellite marker analysis. **Diversity**, v. 02, n. 04, p. 439-449, 2010.

DUDA, R. O.; HART, P. E. **Pattern classification and scene analysis**. John Wiley & Sons: New York, p.189–225, 1973.

GOMES, C.N. **Caracterização morfo-agronômica e diversidade genética em mandioca *Manihot esculenta* Crantz**. Dissertação (Mestrado) - Universidade Federal de Lavras, Lavras, p. 72, 2007.

GONÇALVES, L.S.A.; RODRIGUES, R.; AMARAL JÚNIOR, A. T.; KARASAWA, M.; SUDRÉ, C.P. Comparison of multivariate statistical algorithms to cluster tomate heirloom accessions. **Genetics and Molecular Research**, v.7, n.4, p.1289-1297, 2008.

GOWER, J.C. A general coefficient of similarity and some of its properties. **Biometrics**, Arlington, v. 27, n. 4, p. 857-874, 1971.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. Ed Bookman, Porto Alegre, p. 593, 2005.

LEDO, C. A da S.; GONÇALVES, L.S.A. **Novas abordagens multivariadas em experimentação com fruteiras**. **Anais do XXII Congresso Brasileiro de Fruticultura.** Bento Gonçalves, RS, 2012.

MAHALANOBIS, P.C. On the generalized distance in statistic. **Proceedings of the National Institute of Sciences of India,** New Delhi, v.2, p.49-55, 1936.

MANLY B.F.J. **Randomization, Bootstrap and Monte Carlo Methods in Biology**. Chapman and Hall, London, p. 399, 1997.

MANTEL, N. The detection of disease clustering and a generalized regression approach. **Cancer Research,** v. 27, n. 02, p. 209-220, 1967.

MAPA – Ministério da Agricultura, Pecuária e Abastecimento. Instruções para Execução dos Ensaios de Distingüibilidade, Homogeneidade e Estabilidade de Cultivares de Tabaco (*Nicotiana tabacum* L.). Disponível em:<http://www.agricultura.gov.br/arq_editor/file/vegetal/RegistroAutorizacoes/Formularios%20Proe%C3%A7%C3%A3o%20Cultivares/TABACO%20FORMULRIO%2001%2008%202008%20P.doc>. Acesso em: 15 fev, 2015.

MARTINS, F.A.; CARNEIRO, P.C.S; SILVA, D.J.H. DA.; CRUZ, C. D.; CARNEIRO, J.E. DE S. **Integração de dados em estudos de diversidade genética de tomateiro.** Pesquisa Agropecuária Brasileira, vol.46, n. 11, p.1496-1502, 2011.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a Data Set. **Psychometrika**, 50, p.159 – 179, 1985.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: UFMG, p. 297, 2005.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.** Belo Horizonte: UFMG, p. 295, 2007.

MOHAMMADI, S.; PRASANNA, B. Analysis of genetic diversity in crop plants salient statistical tools and considerations. **Crop Science**, v. 43, n. 4, p. 1235-1248, 2003.

MOURA, M.C.C.L., Gonçalves, L.S.A., Sudré, C.P., Rodrigues, R., Amaral Júnior, A.T., Pereira, T.N.S. Algoritmo de Gower na estimativa da divergência genética em germoplasma de pimenta. **Horticultura Brasileira**, v. 28(2), p. 155-161, 2010.

MWADZINGENI, L.; KASHANGURA C.; GASURA, E.; GARWE, D.; LEWIS, R. Genetic diversity of Zimbabwean and exotic flue-cured tobacco varieties based on phenotypic traits and simple sequence repeats. **African Journal of Agricultural Research,** Vol. 8(46), p. 5845-5853, 2013.

ROCHA, M.C.; GONÇALVES, L.S.A.; CORRÊA, F.M; RODRIGUES, R.; SILVA SL; ABBOUD A.C.S.; CARMO M.G.F. Descritores quantitativos na determinação da divergência genética entre acessos de tomateiro do grupo cereja. **Ciência Rural,** v. 39, p. 664-670, 2009.

ROCHA, M. G.; GONÇALVES, L. S. A.; RODRIGUES, R.; SILVA, P. R. A.; CARMO, M. G. F.; ABBOUD, A. C. S. Uso do algoritmo de Gower na determinação da divergência genética entre acessos de tomateiro do grupo cereja. **Acta Scientiarum Agronomy**, Maringá, v. 32, n. 3, p. 402-406, 2010.

R CORE TEAM. R**: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: http://www.R-project.org/, 2014.

ROHLF, F.J. & FISHER, D.L. Test for hierarchical structure in random data sets. **Systematic Zoologic 17**: p. 407-412, 1968.

SINDITABACO - Sindicato Interestadual da Indústria do Tabaco. Disponível em: < http://sinditabaco.com.br/sobre-o-setor/exportacoes/>. Acesso em 25 de Fev. de 2015.

SOKAL, R.R.; ROHLF, F.J. The comparison of dendrograms by objective methods. **Taxon**, Berlin, v.11, p.30-40, 1962.

STURGES, H.A. The choice of a class interval. **Journal of the American Statistical Association**, v.21, p.65-66, 1926.

SUDRÉ, C. P.; LEONARDECZ, E.; RODRIGUES, R.; AMARAL JÚNIOR, A. T.; MOURA, M. C. L.; GONÇALVES, L. S. A. Genetic resources of vegetable crops: a survey in the Brazilian germplasm collections pictured through papers published in the journals of the Brazilian Society for Horticultural Science. **Horticultura Brasileira**, v. 25, n. 4, p. 496-503, 2007.

ZHANG, H. Y.; LIU, X. Z.; Li, T. S.; YANG, Y. M. Genetic diversity among flue-cured tobacco (*Nicotiana tabacum* L.) revealed by amplified fragment length polymorphism. **Botanical Studies**, V. 47, p. 223-229, 2006.

CONSIDERAÇÕES FINAIS

A fraca associação entre as matrizes de dissimilaridade em isolado indicou que a melhor estratégia para orientar ações de caracterização, conservação e uso do germoplasma em estudo é por meio de estudos de divergência genética com o emprego de caracteres qualitativos e quantitativos de forma conjunta e complementar.

Propor a utilização de duas ou mais estratégias de análise conjunta de dados levando em consideração as particularidades de cada uma, ao invés de se basear em apenas uma em estudos de diversidade genética, poderá ser mais adequada para uma melhor tomada de decisão em relação à escolha de cultivares mais promissores em programas de melhoramento e para maior conhecimento do germoplasma em estudo visando sua conservação.

Os descritores utilizados na formação dos agrupamentos foram eficientes em quantificar a variabilidade existente. Contudo, novos estudos complementares envolvendo dados moleculares analisados em conjunto aos dados morfoagronômicos possivelmente resultará em respostas ainda mais precisas acerca da variabilidade genética presente entre os acessos em estudo.

Estudos com variação no número de acessos e variáveis são oportunos para inferir sobre a consistência dos agrupamentos obtidos pelas diferentes metodologias de análise conjunta testadas, pois o número de variáveis e/ou de acessos podem estar afetando o agrupamento final. A inserção de dados moleculares em trabalhos futuros com a cultura também pode ser oportuna, pois os marcadores moleculares são estáveis. Onde o uso combinado de marcadores moleculares e descritores agromorfológicos poderá promover resultados mais completos.

ANEXO

Tabela A1. Matriz de dissimilaridade genética (Matriz conjunta 1: Matriz C1) entre os 15 acessos de tabaco, baseado na distância de Gower em relação a 10 variáveis quantitativas e 8 variaveis qualitativas. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,23													
A3	0,23	0,19												
A4	0,22	0,05	0,18											
A5	0,11	0,14	0,18	0,16										
A6	0,24	0,16	0,16	0,16	0,16									
A7	0,35	0,34	0,41	0,36	0,35	0,46								
A8	0,11	0,17	0,15	0,17	0,06	0,15	0,36							
A9	0,18	0,08	0,15	0,11	0,09	0,10	0,41	0,12						
A10	0,50	0,53	0,51	0,50	0,53	0,51	0,30	0,52	0,60					
A11	0,54	0,51	0,56	0,51	0,52	0,54	0,34	0,55	0,57	0,13				
A12	0,41	0,41	0,33	0,39	0,41	0,31	0,46	0,43	0,35	0,34	0,30			
A13	0,47	0,56	0,53	0,58	0,45	0,44	0,47	0,44	0,51	0,47	0,51	0,53		
A14	0,70	0,59	0,58	0,58	0,69	0,58	0,49	0,69	0,65	0,40	0,38	0,43	0,60	
A15	0,51	0,59	0,56	0,61	0,48	0,47	0,51	0,47	0,54	0,52	0,54	0,56	0,05	0,60

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A2. Matriz de dissimilaridade genética (Matriz conjunta 2: Matriz C2) entre os 15 acessos de tabaco, baseado na soma das matrizes de Mahalanobis (10 variáveis quantitativas) + Cole Rodgers (8 variaveis qualitativas). UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,65													
A3	0,65	0,48												
A4	0,63	0,03	0,45											
A5	0,23	0,41	0,46	0,42										
A6	0,65	0,44	0,42	0,43	0,43									
A7	1,04	1,05	1,28	1,04	1,04	1,47								
A8	0,23	0,43	0,44	0,44	0,02	0,42	1,07							
A9	0,45	0,21	0,27	0,23	0,21	0,23	1,26	0,23						
A10	1,55	1,58	1,64	1,53	1,59	1,60	0,76	1,59	1,79					
A11	1,63	1,57	1,75	1,61	1,62	1,67	0,86	1,63	1,80	0,10				
A12	1,16	1,14	0,86	1,13	1,16	0,79	1,20	1,19	0,95	0,88	0,85			
A13	1,86	2,27	1,93	2,15	1,79	1,60	2,04	1,73	2,02	2,08	2,38	2,16		
A14	2,35	1,81	1,89	1,85	2,25	1,83	1,59	2,20	2,05	1,06	0,94	1,19	2,37	
A15	2,23	2,66	2,23	2,50	2,16	1,92	2,41	2,09	2,40	2,49	2,84	2,58	0,04	2,69

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A3. Matriz de dissimilaridade genética (Matriz conjunta 3: Matriz C3) entre os 15 acessos de tabaco, baseado na soma das matrizes de Euclidiana Média (10 variáveis quantitativas) + Cole Rodgers (8 variaveis qualitativas). UFRB, Cruz das Almas - BA. 2014.

Acessos	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
A2	1,04													
A3	0,99	0,88												
A4	0,96	0,32	0,87											
A5	0,61	0,61	0,83	0,73										
A6	1,06	0,80	0,75	0,73	0,72									
A7	1,56	1,52	1,69	1,52	1,50	1,87								
A8	0,53	0,78	0,69	0,79	0,35	0,65	1,47							
A9	0,89	0,37	0,75	0,53	0,44	0,57	1,74	0,60						
A10	2,30	2,43	2,40	2,17	2,45	2,31	1,60	2,35	2,64					
A11	2,30	2,21	2,44	2,18	2,30	2,37	1,71	2,36	2,45	0,86				
A12	1,89	1,84	1,70	1,74	1,92	1,60	2,15	2,01	1,66	1,71	1,32			
A13	2,33	2,62	2,45	2,67	2,21	2,10	2,32	2,14	2,45	2,48	2,63	2,64		
A14	3,30	2,76	2,80	2,82	3,18	2,78	2,46	3,15	3,00	2,26	2,00	2,25	2,96	
A15	2,51	2,80	2,59	2,85	2,37	2,26	2,47	2,31	2,62	2,65	2,78	2,82	0,31	2,97

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A4. Matriz de dissimilaridade genética (Matriz conjunta 4: Matriz C4) entre os 15 acessos de tabaco, baseado da soma da matriz dos dados quantitativos transformados em qualitativos pela regra de Sturges + dados qualitativos originais), Cole Rodgers (10 variáveis quantitativas transformadas) + Cole Rodgers (8 variaveis qualitativas de origem). UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	2,22													
A3	1,69	2,56												
A4	2,22	1,62	1,75											
A5	1,55	0,67	2,29	2,29										
A6	2,22	1,48	1,48	1,48	1,75									
A7	2,63	2,63	2,83	3,17	2,36	3,30								
A8	1,82	2,02	1,48	2,29	1,89	1,48	3,17							
A9	2,29	0,47	2,63	1,55	0,74	1,55	2,56	2,09						
A10	2,77	3,30	2,50	3,30	3,30	3,03	2,22	3,30	3,78					
A11	3,84	3,30	3,84	3,57	3,30	3,57	2,49	3,84	3,24	1,62				
A12	2,90	3,17	3,03	3,17	3,17	3,30	3,17	3,17	2,70	2,70	2,16			
A13	2,97	3,64	3,37	3,91	2,97	3,50	2,90	3,24	3,44	3,03	3,30	3,30		
A14	3,78	3,64	3,37	3,10	4,05	3,37	3,50	4,05	3,84	2,56	2,83	3,03	3,44	
A15	3,24	3,64	3,64	3,91	3,24	3,50	3,17	3,24	3,44	3,03	3,30	3,30	0,27	3,17

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A5. Matriz de dissimilaridade genética (Matriz conjunta 5: Matriz C5) entre os 15 acessos de tabaco, baseado nos dados quantitativos transformados pela regra de Sturges integrados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, (10 variáveis quantitativas transformadas) + (8 variáveis qualitativas de origem) = 18 variáveis qualitativas. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,50													
A3	0,39	0,56												
A4	0,50	0,33	0,39											
A5	0,33	0,17	0,50	0,50										
A6	0,50	0,33	0,33	0,33	0,39									
A7	0,61	0,61	0,67	0,72	0,56	0,78								
A8	0,39	0,44	0,33	0,50	0,39	0,33	0,72							
A9	0,50	0,11	0,56	0,33	0,17	0,33	0,61	0,44						
A10	0,67	0,78	0,61	0,78	0,78	0,72	0,50	0,78	0,89					
A11	0,89	0,78	0,89	0,83	0,78	0,83	0,56	0,89	0,78	0,33				
A12	0,67	0,72	0,67	0,72	0,72	0,72	0,72	0,72	0,61	0,61	0,50			
A13	0,67	0,83	0,78	0,89	0,67	0,78	0,67	0,72	0,78	0,67	0,72	0,72		
A14	0,89	0,83	0,78	0,72	0,94	0,78	0,78	0,94	0,89	0,56	0,61	0,67	0,78	
A15	0,72	0,83	0,83	0,89	0,72	0,78	0,72	0,72	0,78	0,67	0,72	0,72	0,06	0,72

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A6. Matriz de dissimilaridade genética (Matriz conjunta 6: Matriz C6) entre os 15 acessos de tabaco, baseado da soma da matriz dos dados quantitativos transformados em qualitativos pela cálculo da raiz quadrada + dados qualitativos originais), Cole Rodgers (10 variáveis quantitativas transformadas) + Cole Rodgers (8 variaveis qualitativas de origem). UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	2,07													
A3	1,58	1,87												
A4	0,85	1,71	1,62											
A5	0,93	1,14	1,62	1,38										
A6	2,07	1,87	1,38	1,62	1,14									
A7	2,23	1,99	2,92	2,48	1,99	3,13								
A8	0,69	1,87	0,89	1,14	0,73	1,38	2,23							
A9	1,87	0,69	2,15	1,42	0,93	1,18	2,19	1,67						
A10	2,39	2,88	2,64	2,39	2,64	3,13	2,32	2,64	3,33					
A11	3,13	2,88	3,37	2,88	2,88	3,13	2,56	3,37	3,08	0,98				
A12	2,23	2,48	2,56	1,99	1,99	1,83	2,72	2,48	1,79	1,79	1,30			
A13	2,76	3,17	3,17	3,41	2,76	3,01	2,48	2,76	3,21	2,80	3,05	3,05		
A14	3,57	3,41	2,92	3,17	3,57	2,92	3,25	3,57	3,61	2,11	2,60	2,56	3,21	
A15	3,01	3,41	3,41	3,41	3,01	3,01	2,72	3,01	3,21	2,80	3,05	3,05	0,24	2,96

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A7. Matriz de dissimilaridade genética (Matriz conjunta 7: Matriz C7) entre os 15 acessos de tabaco, baseado nos dados quantitativos transformados pelo cálculo da raiz quadrada integrados aos dados qualitativos originais para obtenção de uma única matriz pela distância de cole Rodgers, (10 variáveis quantitativas transformadas) + (8 variáveis qualitativas de origem) = 18 variáveis qualitativas. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,50													
A3	0,39	0,44												
A4	0,22	0,39	0,39											
A5	0,22	0,28	0,39	0,33										
A6	0,50	0,44	0,33	0,39	0,28									
A7	0,56	0,50	0,72	0,61	0,50	0,78								
A8	0,17	0,44	0,22	0,28	0,17	0,33	0,56							
A9	0,44	0,17	0,50	0,33	0,22	0,28	0,56	0,39						
A10	0,61	0,72	0,67	0,61	0,67	0,78	0,56	0,67	0,83					
A11	0,78	0,72	0,83	0,72	0,72	0,78	0,61	0,83	0,78	0,22				
A12	0,56	0,61	0,61	0,50	0,50	0,44	0,67	0,61	0,44	0,44	0,33			
A13	0,67	0,78	0,78	0,83	0,67	0,72	0,61	0,67	0,78	0,67	0,72	0,72		
A14	0,89	0,83	0,72	0,78	0,89	0,72	0,78	0,89	0,89	0,50	0,61	0,61	0,78	
A15	0,72	0,83	0,83	0,83	0,72	0,72	0,67	0,72	0,78	0,67	0,72	0,72	0,06	0,72

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A8. Matriz de dissimilaridade genética entre os 15 acessos de tabaco, baseado na distância de Mahalanobis (D^2), em relação a 10 variáveis quantitativas. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	15,86													
A3	16,05	29,56												
A4	9,25	12,54	16,84											
A5	10,90	2,84	20,94	7,73										
A6	14,74	15,14	6,39	8,06	9,85									
A7	9,87	15,88	25,60	12,26	10,78	21,20								
A8	10,23	10,29	11,89	14,34	9,58	6,13	20,63							
A9	15,73	1,99	25,93	10,39	3,07	10,09	16,65	9,42						
A10	51,18	63,48	87,63	42,82	67,64	71,99	60,59	67,28	67,75					
A11	84,42	60,47	129,40	76,07	78,87	99,05	99,29	82,42	68,73	40,27				
A12	58,80	50,53	97,35	44,09	58,67	72,62	73,89	70,25	53,42	26,43	17,04			
A13	409,59	412,01	277,15	366,01	381,15	307,90	400,51	358,77	394,36	574,27	691,42	606,19		
A14	283,41	232,01	262,48	246,82	245,09	237,81	304,95	224,52	246,08	255,82	210,41	225,36	529,69	
A15	553,75	563,27	394,51	502,01	524,48	433,76	545,45	497,67	540,64	735,46	872,57	770,19	14,55	653,99

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A9. Matriz de dissimilaridade genética entre os 15 acessos de tabaco, baseado na distância Euclidiana média (D), em relação a 10 variáveis quantitativas. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,15													
A3	0,13	0,16												
A4	0,12	0,11	0,15											
A5	0,14	0,07	0,14	0,11										
A6	0,15	0,13	0,11	0,11	0,10									
A7	0,18	0,17	0,16	0,17	0,16	0,15								
A8	0,11	0,13	0,09	0,13	0,12	0,08	0,15							
A9	0,16	0,05	0,18	0,11	0,08	0,12	0,17	0,13						
A10	0,29	0,34	0,33	0,25	0,34	0,30	0,33	0,31	0,34					
A11	0,29	0,26	0,34	0,26	0,29	0,32	0,37	0,31	0,28	0,29				
A12	0,29	0,28	0,36	0,24	0,30	0,33	0,38	0,33	0,28	0,30	0,17			
A13	0,51	0,47	0,41	0,48	0,47	0,43	0,44	0,44	0,48	0,62	0,67	0,68		
A14	0,56	0,52	0,53	0,53	0,52	0,52	0,55	0,51	0,53	0,62	0,53	0,55	0,65	
A15	0,57	0,53	0,46	0,54	0,52	0,48	0,49	0,50	0,54	0,68	0,72	0,74	0,10	0,65

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A10. Matriz de dissimilaridade genética entre os 15 acessos de tabaco, baseado na distância de cole Rodgers, em relação a 10 variáveis quantitativas transformadas pela Regra de Sturges. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,60													
A3	0,40	0,80												
A4	0,60	0,60	0,50											
A5	0,50	0,10	0,70	0,70										
A6	0,60	0,40	0,40	0,40	0,50									
A7	0,60	0,60	0,60	0,80	0,50	0,70								
A8	0,60	0,60	0,40	0,70	0,70	0,40	0,80							
A9	0,70	0,10	0,90	0,50	0,20	0,50	0,50	0,70						
A10	0,50	0,70	0,40	0,70	0,70	0,60	0,60	0,70	0,80					
A11	0,90	0,70	0,90	0,80	0,70	0,80	0,70	0,90	0,60	0,60				
A12	0,70	0,80	0,90	0,80	0,80	1,00	0,80	0,80	0,70	0,70	0,50			
A13	0,80	0,90	0,80	1,00	0,80	1,00	0,70	0,90	0,90	0,90	1,00	1,00		
A14	0,80	0,90	0,80	0,70	0,90	0,80	1,00	0,90	0,90	0,80	0,90	0,90	0,90	
A15	0,90	0,90	0,90	1,00	0,90	1,00	0,80	0,90	0,90	0,90	1,00	1,00	0,10	0,80

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A11. Matriz de dissimilaridade genética entre os 15 acessos de tabaco, baseado na distância de cole Rodgers, em relação a 10 variáveis quantitativas transformadas pelo cálculo da Raiz Quadrada. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,60													
A3	0,40	0,60												
A4	0,10	0,70	0,50											
A5	0,30	0,30	0,50	0,40										
A6	0,60	0,60	0,40	0,50	0,30									
A7	0,50	0,40	0,70	0,60	0,40	0,70								
A8	0,20	0,60	0,20	0,30	0,30	0,40	0,50							
A9	0,60	0,20	0,80	0,50	0,30	0,40	0,40	0,60						
A10	0,40	0,60	0,50	0,40	0,50	0,70	0,70	0,50	0,70					
A11	0,70	0,60	0,80	0,60	0,60	0,70	0,80	0,80	0,60	0,40				
A12	0,50	0,60	0,80	0,40	0,40	0,50	0,70	0,60	0,40	0,40	0,20			
A13	0,80	0,80	0,80	0,90	0,80	0,90	0,60	0,80	0,90	0,90	1,00	1,00		
A14	0,80	0,90	0,70	0,80	0,80	0,70	1,00	0,80	0,90	0,70	0,90	0,80	0,90	
A15	0,90	0,90	0,90	0,90	0,90	0,90	0,70	0,90	0,90	0,90	1,00	1,00	0,10	0,80

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).

Tabela A12. Matriz de dissimilaridade genética entre os 15 acessos de tabaco, baseado na distância de cole Rodgers, em relação a 8 variáveis qualitativas de origem. UFRB, Cruz das Almas - BA. 2014.

Acessos	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A2	0,38													
A3	0,38	0,25												
A4	0,38	0,00	0,25											
A5	0,13	0,25	0,25	0,25										
A6	0,38	0,25	0,25	0,25	0,25									
A7	0,63	0,63	0,75	0,63	0,63	0,88								
A8	0,13	0,25	0,25	0,25	0,00	0,25	0,63							
A9	0,25	0,13	0,13	0,13	0,13	0,13	0,75	0,13						
A10	0,88	0,88	0,88	0,88	0,88	0,88	0,38	0,88	1,00					
A11	0,88	0,88	0,88	0,88	0,88	0,88	0,38	0,88	1,00	0,00				
A12	0,63	0,63	0,38	0,63	0,63	0,38	0,63	0,63	0,50	0,50	0,50			
A13	0,50	0,75	0,75	0,75	0,50	0,50	0,63	0,50	0,63	0,38	0,38	0,38		
A14	1,00	0,75	0,75	0,75	1,00	0,75	0,50	1,00	0,88	0,25	0,25	0,38	0,63	
A15	0,50	0,75	0,75	0,75	0,50	0,50	0,63	0,50	0,63	0,38	0,38	0,38	0,00	0,63

A1 (ER 03-107); A2 (ER 04-090); A3 (ER 04-095); A4 (ER 05-005); A5 (ER 05-070); A6 (ER 12-040); A7 (ER 13-061); A8 (ER 13-065); A9 (ER 28-027); A10 (ER 33-021); A11 (ER 33-022); A12 (ER 33-023); A13 (109 PD); A14 (125 PD); A15 (221 PD).