ÍNDICES DE DISSIMILARIDADE E MÉTODOS DE AGRUPAMENTO EM DADOS MOLECULARES DOMINANTES COM PERDAS DE DADOS

Gustavo Martins Sturm[1], João Felipe De Brites Senra[2], Marcia Flores Da Silva Ferreira[4], Moyses Nascimento[3], Adésio Ferreira[4]

1. Pós-Graduando em Produção Vegetal da Universidade Federal do Espírito Santo (gustavosturm@hotmail.com)
2. Graduando em Agronomia da Universidade Federal do Espírito Santo CCA/UFES
3. Professor e Doutorando da Universidade Federal de Viçosa
4. Professores Doutores da Universidade Federal do Espírito Santo

RESUMO

Este trabalho teve como objetivo avaliar a influência da perda de dados sobre os métodos de agrupamento UPGMA e Vizinho mais próximo, quanto aos índices de similaridade de Jaccard e Nei e Li. Foram estabelecidas quatro populações paternais (base), cada uma composta por cinco indivíduos diplóides e homozigotos para todos os locos. Em cada população avaliou-se 50 e 100 locos (marcas), e para cada loco foram simulados dois alelos. A partir das populações base obteve-se as F1s respectivas. A partir de cada população F_1 procederam-se cinco retrocruzamentos (RC_1, RC_2, RC_3, RC_4 e RC_5), com cada população base, cada qual com cinqüenta indivíduos, para posterior obtenção da População de Trabalho. Sobre a população de trabalho foi imposta a perda de dados na proporção de 0, 10, 20, 30, 40 e 50%. A perda de dados alterou drásticamente os agrupamentos e os índices de dissimilaridade, afetando de maneira significativa a interpretação dos resultados.E ainda não houve correlação entre os níveis de perda e número de grupos formados nos diferentes níveis de cortes nos agrupamentos estudados.

PALAVRAS-CHAVE: Diversidade genética, marcadores moleculares, estatística genômica

DISSIMILARITY INDEX AND CLUSTERING METHODS IN DOMINANT MOLECULAR DATA WITH LOSS OF DATA

ABSTRACT

This study aimed to evaluate the influence of data loss on the methods of UPGMA clustering and Nearest Neighbor, about the similarity indices of Jaccard and Nei and Li. Was established four parental populations (base), each one consisting of five diploids and homozygous for all loci subjects. In each population was evaluated 50 and 100 loci (marks), and for each locus two alleles were simulated. From the base population was obtained the respective F1s. From each F1 population held up five backcrosses (BC1, BC2, BC3, BC4 and BC5), with each base population, each one with fifty subjects, for subsequent acquisition of the Working Population. On the population of working was imposed data loss ratio of 0, 10, 20, 30, 40 and 50%. Data

loss drastically changed the groupings and the indices of dissimilarity, significantly affecting the interpretation of results. And there was no correlation between levels of loss and number of groups formed in different levels of cuts in the clusters studied.

KEYWORDS: Genetic diversity, molecular markers, statistical genomics

INTRODUÇÃO

O estudo da dissimilaridade genética tem por objetivo identificar genitores para a obtenção de híbridos com maior efeito heterótico e que proporcionem maior segregação entre os recombinantes, possibilitando o aparecimento de transgressivos (CRUZ; CARNEIRO, 2003). O conhecimento da distância genética entre genótipos de uma população de interesse é importante para um programa de melhoramento, pois permite a organização do germoplasma e uma amostragem mais eficiente de genótipos (NIENHUIS et al., 1993).

A determinação ou avaliação da divergência genética apresenta duas naturezas de avaliação: natureza quantitativa e de natureza preditiva (CRUZ; CARNEIRO, 2003). Nas avaliações de natureza quantitativa, comumente são utilizadas análises dialélicas. Em que de acordo com Filho et al., (2008), possibilitam determinar a capacidade geral e específica de combinação e a heterose manifestada nos híbridos. No entanto, a necessidade de avaliações de p genitores e de todas as suas combinações híbridas $p(p\text{-}1)/2$, aliada ao fato de que, em algumas culturas, a polinização manual é onerosa, difícil de ser executada e com pouca probabilidade de êxito na obtenção de semente híbrida, pode inviabilizar o estudo, principalmente quando o valor de p é elevado (CRUZ; REGAZZI, 1997).

Desta forma, as análises de natureza preditiva apresentam grande importância no estudo da diversidade genética, pois, através de informações sobre diferenças morfológicas (descritores e ou dados agronômicos), fisiológicas (isoenzimas) e moleculares, é possivel determinar quais genótipos são mais divergentes, e assim os mais prováveis a possibilitarem ganhos de seleção. Entre estes, os dados moleculares são os mais indicados para estas análises. Pois, não são influenciados pelo ambiente (condições de campo), informando portanto a divergência genética entre os indivíduos.

Contudo, como qualquer outra informação, é necessário avaliar a robustez dos resultados e de acordo com Hackett e Broadfoot (2003) em estudos de diversidade o ideal é que o conjunto de dados moleculares não tenha valores perdidos e erros de genotipagem. Assim, neste contexto o objetivo deste trabalho foi de avaliar a influência de diferentes níveis de perda de dados moleculares dominantes sobre os métodos de agrupamento UPGMA e Vizinho mais próximo, quanto aos índices de similaridade de Jaccard e Nei e Li.

METODOLOGIA

Foram estabelecidas quatro populações paternais (base), cada uma composta por cinco indivíduos diplóides e homozigotos para todos os locos. Em cada população avaliou-se 50 e 100 locos (marcas), e para cada loco foram

simulados dois alelos. A partir dos cruzamentos aleatórios entre as duas populações, foram obtidas duas populações F_{1s}, de acordo com o esquema a seguir:
- Cruzamento 1: População Base 1 (PB_1) x População Base 2 (PB_2) = F_1 (50 marcas).
- Cruzamento 2: População Base 3 (PB_3) x População Base 4 (PB_4) = F_1 (100 marcas).

A partir de cada população F_1 procederam-se cinco retrocruzamentos (RC_1, RC_2, RC_3, RC_4 e RC_5), com cada população base, cada qual com cinqüenta indivíduos, para posterior obtenção da População de Trabalho. De cada retrocruzamento foram retirados os dez genótipos mais similares, segundo o índice de similaridade de Jaccard e método de agrupamento de Tocher modificado. Foram estabelecidas assim quatro populações de trabalho (PT) cada uma formada por 55 indivíduos, de acordo com os esquemas abaixo:

- População de Trabalho 1 (PT_1) 55 indivíduos de 50 marcas cada:
PB_1 5 indivíduos
F_1 X PB_1 = RC_1 – Amostra de 10 indivíduos
RC_1 X PB_1 = RC_2 – Amostra de 10 indivíduos
RC_2 X PB_1 = RC_3 – Amostra de 10 indivíduos
RC_3 X PB_1 = RC_4 – Amostra de 10 indivíduos
RC_4 X PB_1 = RC_5 – Amostra de 10 indivíduos

- População de Trabalho 2 (PT_2) 55 indivíduos de 50 marcas cada:
PB_2 5 indivíduos
F_1 X PB_2 = RC_1 – Amostra de 10 indivíduos
RC_1 X PB_2 = RC_2 – Amostra de 10 indivíduos
RC_2 X PB_2 = RC_3 – Amostra de 10 indivíduos
RC_3 X PB_2 = RC_4 – Amostra de 10 indivíduos
RC_4 X PB_2 = RC_5 – Amostra de 10 indivíduos

- População de Trabalho 3 (PT_3) 55 indivíduos de 100 marcas cada:
PB_3 5 indivíduos
F_1 X PB_3 = RC_1 – Amostra de 10 indivíduos
RC_1 X PB_3 = RC_2 – Amostra de 10 indivíduos
RC_2 X PB_3 = RC_3 – Amostra de 10 indivíduos
RC_3 X PB_3 = RC_4 – Amostra de 10 indivíduos
RC_4 X PB_3 = RC_5 – Amostra de 10 indivíduos

- População de Trabalho 4 (PT_4) 55 indivíduos de 100 marcas cada:
PB_4 5 indivíduos
F_1 X PB_4 = RC_1 – Amostra de 10 indivíduos
RC_1 X PB_4 = RC_2 – Amostra de 10 indivíduos
RC_2 X PB_4 = RC_3 – Amostra de 10 indivíduos
RC_3 X PB_4 = RC_4 – Amostra de 10 indivíduos
RC_4 X PB_4 = RC_5 – Amostra de 10 indivíduos

Considerando a constituição genotípica dos indivíduos de cada população de trabalho, foi realizado o escoreamento de cada loco (por exemplo: AA e Aa = 1; aa = 0) e obtido o padrão de bandas de cada indivíduo para todos os locos

avaliados. Sobre os padrões de bandas obtidos foi imposta a perda de marcas na proporção de 10, 20, 30, 40 e 50% (cinco níveis de perda), de forma aleatória.

A perda foi aleatória e sobre o total de marcadores e de dados (marcas) em cada marcador. Os dados são referentes ao número de genótipos, ou seja, 55 indivíduos. Desta forma, uma perda de 20% sobre a população de trabalho um, significa que em dez marcadores, escolhidos aleatoriamente, serão perdidos onze dados.

Cada porcentagem de perda representa outro padrão de banda em cada população de trabalho, e o padrão sem perda constituiu a testemunha. Assim o delineamento ficou composto de seis padrões de banda (uma testemunha e cinco níveis de perda), quatro populações de trabalho, dois índices de similaridade (Jacard e Nei e Li) e dois métodos de agrupamento (Vizinho mais próximo e UPGMA), totalizando 96 tratamentos.

Cada tratamento gerou um dendograma, sendo que em cada um destes foram realizados cortes em três níveis de dissimilaridade, 50, 65 e 80%. Desta forma foram avaliados a alteração do número de grupos, comparado com a testemunha, nos três níveis de cortes de dissimilaridade.

Nos dendogramas os indivíduos de um a cinco correspondem aos genótipos da população base da população de trabalho correspondente, de seis a 15 os indivíduos da amostra do RC_1, de 16 a 25 amostra do RC_2, de 26 a 35 amostra do RC_3, de 36 a 45 RC_4 e 46 a 55 RC_5.

O processo de simulação dos dados para obtenção das populações foi através do módulo de “simulação” do aplicativo computacional em Estatística e Genética GENES (CRUZ, 1997). A perda de dados foi aleatória e simulada pelo aplicativo computacional GQMOL (CRUZ, 2010), através do módulo “perda de dados” do aplicativo.

RESULTADOS E DISCUSSÃO

O número de grupos formados, quanto aos índices de dissimilaridade de Jaccard e Nei e Li com os agrupamentos de UPGMA VMP, apresentaram grande variação comparada ao padrão, sem perda de dados (Tabelas de 1 a 12).

TABELA 1: Número de grupos no corte de 50% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VPM), para População de Trabalho 1.

Corte de dissimilaridade a 50%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	44	53	37	50
10	44	53	37	51
20	45	49	37	43
30	40	48	35	45
40	46	46	36	45

50	39	48	31	45

TABELA 2: Número de grupos no corte de 65 % de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), para População de Trabalho 1.

Corte de dissimilaridade a 65%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	29	40	21	35
10	29	43	22	37
20	26	32	21	19
30	23	26	17	17
40	34	34	16	26
50	21	35	16	30

TABELA 3: Número de grupos no corte de 80 % de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), para População de Trabalho 1.

Corte de dissimilaridade a 80%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	10	12	06	11
10	11	20	08	16
20	11	5	08	05
30	12	7	06	05
40	11	11	07	11
50	09	18	06	13

TABELA 4: Número de grupos no corte de 50% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), para População de trabalho 2.

Corte de dissimilaridade a 50%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	46	53	40	50
10	29	24	19	14
20	28	23	19	14
30	28	25	19	13
40	27	22	18	16
50	26	20	15	15

TABELA 5: Número de grupos no corte de 65% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), para população de trabalho 2.

Corte de dissimilaridade a 65%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	37	37	24	34
10	19	13	11	12
20	19	13	11	12
30	19	13	11	11
40	18	16	09	10
50	15	15	10	09

TABELA 6: Número de grupos no corte de 80% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), para População de trabalho 2.

Corte de dissimilaridade a 80%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	17	14	12	12
10	11	11	08	07
20	11	09	08	06
30	11	09	06	06
40	08	08	05	03
50	09	07	06	05

TABELA 7: Número de grupos no corte de 50% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), para População de Trabalho 3.

Corte de dissimilaridade a 50%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	31	26	14	10
10	48	54	43	52
20	45	53	42	51
30	42	54	42	49
40	47	51	43	50
50	41	48	37	38

TABELA 8: Número de grupos no corte de 65% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), para População de Trabalho 3.

Corte de dissimilaridade a 65%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	14	06	04	03
10	41	45	32	39
20	35	45	24	40
30	35	41	29	39
40	38	36	28	27
50	26	21	18	16

TABELA 9: Número de grupos no corte de 80% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), População de Trabalho 3.

Corte de dissimilaridade a 80%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	03	03	03	02
10	19	20	12	14
20	13	20	09	12
30	13	20	09	19
40	18	10	13	05
50	10	08	05	04

TABELA 10: Número de grupos no corte de 50% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), População de Trabalho 4.

Corte de dissimilaridade a 50%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	37	31	22	17
10	43	50	37	33
20	42	47	34	31
30	42	49	36	41
40	43	45	36	28
50	37	47	31	34

TABELA 11: Número de grupos no corte de 65% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), População de Trabalho 4.

Corte de dissimilaridade a 65%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	22	17	12	12
10	27	16	19	11
20	29	17	17	11
30	26	25	18	17
40	29	17	21	11
50	22	18	15	14

TABELA 12: Número de grupos no corte de 80% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para os índices de dissimilaridade de Jacard e Nei e Li, agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), População de Trabalho 4.

Corte de dissimilaridade a 80%				
Perda	Jaccard		Nei e Li	
(%)	UPGMA	VMP	UPGMA	VMP
0	12	08	06	05
10	13	07	09	07
20	13	07	07	06
30	07	08	07	07
40	15	07	10	05
50	10	10	08	06

A variação do número de grupos não seguiu um padrão de diferenciação relativo ao nível de perda, como era esperado. A grande diferenciação foi verificada em todos os níveis de cortes realizados (50%, 65% e 80%) nos agrupamentos para os dois paternais.

O aumento do número de grupos demonstra uma dissimilaridade inexistente nos agrupamentos, induzindo a conclusão que existe variabilidade. Isto na prática pode acarretar, por exemplo, a escolha de genótipos similares para serem os paternais, acreditando que estes são dissimilares. Já a redução do número de grupos induz ao contrário, concluindo similaridade entre os genótipos erroneamente. Sabe-se que a quantificação da dissimilaridade genética, antes de qualquer cruzamento, possibilita aos melhoristas concentrarem seus esforços nas combinações mais promissoras ao ganho de seleção, evidenciando a necessidade da correta quantificação da variabilidade genética.

De acordo com Parteniani & Lonquist (1963), deve existir um nível ótimo de dissimilaridade entre os pais para a obtenção de heterose. (MOLL et al., 1965) estudando a divergência genética do milho concluiu que deve haver um grau ótimo de divergência para que ocorra a expressão máxima da heterose. Assim, dentro deste contexto a perda de dados torna-se um fato indesejável em programas de melhoramento os quais se baseiam num nível mínimo de divergência genética, como tomada de decisão para escolha dos paternais. Pois, aumenta as chances de erros nas análises da diversidade e escolha incorreta dos genitores.

Para a PT_1 o corte a 65% foi o mais afetado, sendo que nas demais populações de trabalho foi o corte a 50%, porém na PT_2 a diferença no módulo do número de variações foi pequena entre 50 e 65% de corte.

Para as populações de trabalho um, dois e três o agrupamento VMP apresentou maior variação, portanto, foi o mais afetado pela perda de dados. Mas, o fato não acorreu na população de trabalho 4 onde o agrupamento UPGMA foi mais alterado (com exceção em Jaccard corte de 50%).

Quanto aos índices de similaridade verificou-se que ambos foram sensíveis à perda de dados, em que não foi possível determinar qual o mais afetado, pois, analisando dentro e entre as populações de trabalho, verificou-se que a diferença entre elas não manteve um padrão de superioridade ou inferioridade.

CONCLUSÕES

A perda de dados alterou drásticamente os agrupamentos e os índices de dissimilaridade, afetando de maneira significativa a interpretação dos resultados.

Não houve correlação entre os níveis de perda e o número de grupos formados nos diferentes níveis de cortes nos agrupamentos independente dos índices de dissimilaridade.

AGRADECIMENTOS

Ao CNPq pela concessão da bolsa.

REFERÊNCIAS BIBLIOGRÁFICAS

CRUZ, C. D. **Programa GENES:** versão Windows. Aplicativo computacional em genética e estatística. Viçosa, UFV, 1997, 442p.

CRUZ, C.D.; SCHUSTER, I. **GQMOL**: aplicativo computacional para análise de dados moleculares e de suas associações com caracteres quantitativos. Versão 2.1. Viçosa, MG: UFV, 2004. Disponível em: <http://www.ufv.br/dbg/gqmol/gqmol.htm>. Acesso em: 20 out. 2010.

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

CRUZ, C.D;. REGAZZI, A.J. Divergência genética In: CRUZ, CD.; REGAZZI, A.J. **Métodos biométricos aplicados ao melhoramento genético**. Viçosa, UFV: Imprensa Universitária, cap. 6, p.287-324. 1997.

FILHO, A. C.; RIBEIRO, N.D.; REIS, R. C. P.; SOUZA, J. R.; 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, nov, 2008.

HACKETT, C.A.; BROADFOOT, L.B. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. **Heredity**, v.90, p.33-38. 2003.

MOLL, R.H.; LONQUIST, J.H.; VÉLEZ FORTUNO, J.; JOHNSON, E.C. The relationship of heterosis and genetic divergence in maize. **Genetic**, v.52, n.1, p.139-144, 1965.

NIENHUIS, J. et al. Genetic similarity among Brassica oleracea genotypes as measured by restriction fragment length polymorphisms. **Journal of the American Society for Horticultural Science**, Alexandria, v.118, n.2, p.298-303, 1993.

PARTENIANI, E.; LONQUIST, J.H. Heterosis in interracial crosses of corn (Zea mays L.). **Crop Sci**., v.3, n. 1, p. 504-507, 1963.