XIV INIC Encontro Latino Americano de Iniciação Científica
X EPG Encontro Latino Americano de Pós Graduação
IV INIC Jr Encontro Latino Americano de Iniciação Científica Júnior

NÍVEIS DE PERDA EM DADOS MOLECULARES CODOMINANTES SOB ÍNDICES DE DISSIMILARIDADE E AGRUPAMENTOS

Catarina Denise Entringer Contreiro[1], João Felipe De Brites Senra1, Wellington Clarindo[1], Sérgio Henriques Saraiva[1], Marcia Flores Da Silva Ferreira[1], Moyses Nascimento[2], Adésio Ferreira[1]

[1]Universidade Federal do Espírito Santo Centro de Ciências Agrárias, Departamento de Produção Vegetal, . Alto Universitário, s/nº, CEP 29500-000 Alegre, ES. E-mail: catarinaentringer@hotmail.com, joaofelipeagronomo@hotmail.com, welbiologo@gmail.com, sergio@cca.ufes.br, mfloressf@gmail.com, adesioferreira@gmail.com; [2]Universidade Federal de Viçosa, Departamento de Estatística, Avenida P.H. Rolfs, s/nº, CEP 36571-000 Viçosa, MG. E-mail: moysesnascim@ufv.br.

Resumo- O presente estudo objetivou avaliar os efeitos da perda de dados moleculares codominantes, em diferentes níveis de perdas de dados quanto aos métodos de agrupamento UPGMA e Vizinho mais próximo referentes aos índices de dissimilaridade Ponderado (Cruz, 2006a) e índice Smouse (d2) (Smouse e Peakall, 1999). Foram estabelecidas quatro populações paternais (base), cada uma composta por cinco indivíduos diplóides e homozigotos para todos os locos. Em cada população avaliou-se 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 F1 procederam-se cinco retrocruzamentos (RC1, RC2, RC3, RC4 e RC5), 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%. O índice Ponderado e o índice de Smouse apresentam a mesma sensibilidade em todos os níveis e perda. Também os métodos de agrupamento UPGMA e Vizinho mais próximo comportaram-se semelhantes independente dos índices e níveis de perda.

Palavras-chave: análise multivariada, diversidade genética, biometria, bioinformática

Área do Conhecimento:

Introdução

A variabilidade genética é o ponto de partida em todo programa de melhoramento. Pois, sem esta os ganhos com a seleção são difíceis ou até mesmo improváveis. Assim, o melhorista ao selecionar genótipos contrastantes para realizar cruzamentos, deve antes de qualquer coisa estudar a diversidade dentro e entre as populações disponíveis, averiguando o nível de variabilidade genética que ele dispõe, ou seja, quais alelos a, ou as populações possuem, e qual sua taxa de freqüência (CRUZ; CARNEIRO, 2006).

Os avanços na biologia molecular abriram novas perspectivas para a pesquisa em conservação de espécies e para os estudos de biologia populacional. E com a utilização de marcadores moleculares é possível a detecção da variabilidade (base para o melhoramento) existente diretamente ao nível do DNA (Milach, 1998).

Um grande número de metodologias para a quantificação e estudo da diversidade entre e dentro de populações são encontrados na literatura, no entanto, a abrangência dos estudos, de informações, de métodos e de material biológico tem levado a certa dificuldade em escolher e aplicar corretamente as metodologias disponíveis e interpretar, convenientemente, o significado dos resultados das análises biométricas (Cruz; Carneiro, 2006).

Nesse contexto em que a perda de dados moleculares consiste em um dos processos que podem interferir nos resultados da diversidade genética dos indivíduos estudados. Assim, o presente trabalho objetivou estudar os efeitos da perda de dados moleculares codominantes, em diferentes níveis de perdas de dados quanto aos métodos de agrupamento UPGMA e Vizinho mais próximo referentes aos índices de dissimilaridade Ponderado (Cruz, 2006a) e índice Smouse (d^2) (Smouse e Peakall, 1999), proporcionando um referencial teórico e prático, que orientará a utilização dos recursos biométricos, permitindo o melhor aproveitamento dos mesmos e conseqüentemente interpretações corretas dos resultados obtidos.

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 100 locos (marcas), e para cada loco foram simulados dois alelos. A partir dos cruzamentos aleatórios entre as populações, foi obtida uma população F_1, de acordo com o esquema a seguir:

Cruzamento 2: População Base 3 (PB_3) x População Base 4 (PB_4) = F_1 (100 marcas).

A partir da população F_1 procederam-se cinco retrocruzamentos (RC_1, RC_2, RC_3, RC_4 e RC_5), 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 duas 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 100 marcas cada:
PB_1 5 indivíduos
F_1 X PB_3 = 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 4 (PT_2) 55 indivíduos de 100 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

Assim, totalizando 12 sub-populações da seguinte forma: - as sub-populações - P_1, P_2, Rc_{11}, Rc_{21}, Rc_{31}, Rc_{41}, Rc_{51}, Rc_{12}, Rc_{22}, Rc_{32}, Rc_{42}, Rc_{52}. Em que, Rc_{11} é o retrocruzamento 1 (um) do P1 e Rc_{21} é o retrocruzamento dois do Pai1. Desta forma, obteremos populações com estrutura genética bem definida, para a natureza do marcador, devido que a cada retrocruzamento o perfil genético das populações resultantes se aproximar do perfil de seu genitor (fato corroborado a partir de conhecimentos básicos de genética quantitativa).

Posteriormente foram simuladas perdas de dados nas duas populações no GQMOL (Cruz, 2008), item Simulação → Perdas de dados. Os níveis de perdas de marcadores moleculares variaram de zero (nenhuma perda); 10%, 20%, 30%, 40% e 50%.

Foram utilizados para a análise na pesquisa o índice Ponderado (Cruz, 2006a) e índice Smouse (d^2) (Smouse e Peakall, 1999), por serem de acordo com Cruz (2006a), os coeficientes mais utilizados em trabalhos de diversidade genética com marcadores de natureza codominante.

Para facilitar a interpretação dos resultados foram utilizados métodos de agrupamento que separam um grupo original em vários subgrupos, sempre mantendo homogeneidade dentro dos subgrupos e heterogeneidade entre os subgrupos. Embora existam inúmeros métodos de agrupamento, que se distinguem pelo tipo de resultado a ser fornecido e pelas diferentes formas de definir a proximidade entre um indivíduo e um grupo já formado ou entre dois grupos quaisquer, percebe-se, na prática, que existe uma tendência entre os pesquisadores em se utilizar os Métodos do vizinho mais próximo e vizinho mais distante e Ligação média entre grupos (UPGMA) e Método de Ward. Dessa forma, no presente estudo foram utilizados os métodos do vizinho mais próximo e de UPGMA, na qual são métodos hierárquicos e, portanto, agrupam os genótipos em processos que se seguem em diferentes níveis, não havendo a preocupação com número ótimo de grupos, pois no final é estabelecido um dendrograma ou o diagrama de árvore. Segundo Cruz e Regazzi (2001) o método de agrupamento hierárquico de Vizinho mais próximo adota como critério para formação de grupos a menor distância existente entre um par de genótipos, enquanto o método de agrupamento UPGMA utiliza como base a média das distâncias entre todos os pares de genótipos para formação de cada grupo.

Para a interpretação do dendograma foi realizado comparação de cada nível de perda com o padrão (sem perda), permitindo verificar as diferenças nas formações de grupos quanto aos índices e quanto aos agrupamentos.

Resultados

O número de grupos formados, quanto ao índice Ponderado e índice Smouse para os dados codominantes, com os agrupamentos de: ligação média entre grupos (UPGMA) e Vizinho mais próximo (VMP), apresentaram grande variação comparada ao padrão, sem perda de dados (Tabelas 1, 2, 3, 4, 5 e 6). 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.

Tabela 1 - Número de grupos em cortes de 50% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para o índice ponderado (IP) e índice Smouse (d^2), agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), com paternal um e seus respectivos retrocruzamentos com cem locos.

Corte de dissimilaridade 50%				
Perda (%)	IP		d^2	
	UPGMA	VMP	UPGMA	VMP
0	47	51	40	48
10	34	33	37	38
20	25	31	35	34
30	20	20	28	25
40	16	20	28	39
50	9	9	14	13

Tabela 2 - Número de grupos em cortes de 65% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para o índice ponderado (IP) e índice Smouse (d^2), agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), com paternal um e seus respectivos retrocruzamentos com cem locos.

Corte de dissimilaridade 65%				
Perda (%)	IP		d^2	
	UPGMA	VMP	UPGMA	VMP
0	30	38	23	26
10	12	10	18	13
20	11	11	15	14
30	15	16	17	16
40	11	13	15	18
50	9	9	10	9

Tabela 3 - Número de grupos em cortes de 80% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para o índice ponderado (IP) e índice Smouse (d^2), agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), com paternal um e seus respectivos retrocruzamentos com cem locos.

Corte de dissimilaridade 80%				
Perda (%)	IP		d^2	
	UPGMA	VMP	UPGMA	VMP
0	15	13	7	8
10	6	4	9	3
20	11	10	8	5
30	15	14	12	13
40	8	7	7	10
50	7	8	7	8

Tabela 4 - Número de grupos em cortes de 50% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para o índice ponderado (IP) e índice Smouse (d^2), agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), com paternal dois e seus respectivos retrocruzamentos com cem locos.

Corte de dissimilaridade 50%				
Perda (%)	IP		d^2	
	UPGMA	VMP	UPGMA	VMP
0	50	55	43	55
10	40	49	44	54
20	29	34	35	44
30	23	19	35	40
40	16	13	26	43
50	10	9	19	26

Tabela 5 - Número de grupos em cortes de 65% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para o índice ponderado (IP) e índice Smouse (d^2), agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), com paternal dois e seus respectivos retrocruzamentos com cem locos.

Corte de dissimilaridade 65%				
Perda (%)	IP		d^2	
	UPGMA	VMP	UPGMA	VMP
0	37	50	21	42
10	15	26	23	43
20	13	11	17	19
30	14	15	20	20
40	10	12	15	14
50	9	9	12	9

Tabela 6 - Número de grupos em cortes de 80% de dissimilaridade nos níveis de perdas de dados de 0, 10, 20, 30, 40 e 50%, para o índice ponderado (IP) e índice Smouse (d^2), agrupados pelos métodos de ligação média entre grupos (UPGMA), e do Vizinho mais próximo (VMP), com paternal dois e seus respectivos retrocruzamentos com cem locos.

Corte de dissimilaridade 80%

Perda (%)	IP		d^2	
	UPGMA	VMP	UPGMA	VMP
0	11	36	5	18
10	7	6	7	19
20	11	11	9	10
30	12	13	11	15
40	7	9	7	10
50	7	9	6	9

Discussão

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. E 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.

A variabilidade genética é uma das características mais importantes num programa de melhoramento, e desta forma a avaliação correta deste parâmetro é uma das premissas para o sucesso de qualquer programa de melhoramento. Na escolha de um valor de dissimilaridade mínimo como ponto de partida para a escolha de genótipos como paternais, a perda de dados pode ocultar uma similaridade presente ou ausente, podendo não ocorrer o surgimento de genótipos com o efeito heterótico ou aparecimento dos transgressivos esperados, responsáveis pelos significativos ganhos de seleção, após cruzamentos incorretos.

Os índices de dissimilaridade apresentaram a mesma sensibilidade à perda de dados. A semelhança de sensibilidade pode ter sido em grande parte devido à proximidade que os métodos apresentam de descrever similaridade entre os genótipos.

Conclusão

Em todos os níveis de perda os agrupamentos demonstram incorreta dissimilaridade e similaridade entre e dentro das populações.

O índice Ponderado e o índice de Smouse apresentam a mesma sensibilidade semelhante aos níveis e perda.

A perda de dados apresentou drásticas incoerências nas análises de diversidade genética para dados moleculares codominantes, independente do nível de perda não havendo correlação entre estes fatores.

Referências

- CRUZ, C.D.; CARNEIRO, P.S.C. Modelos biometricos aplicados ao melhoramento genetico. 2. ed. Vicosa: UFV, 2006a. v. 2.

- CRUZ, C.D. **Programa GENES – Análise Multivariada e Simulação**, Viçosa: Imprensa Universitária, 2006b. 285p.

- CRUZ, C.D. **Programa para análises de dados moleculares e quantitativos – GQMOL**. Viçosa: UFV, 2008.

- CRUZ, C. D.; REGAZZI, A. J. **Modelos biométricos aplicados ao melhoramento genético.** UFV, 2001. 390p.

- SMOUSE, P.E. & PEAKALL, R. 1999. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82:561-573.

- Milach, S.C.K. Marcadores de DNA. Aplicações no melhoramento de plantas. Ed.Sandra Cristina Kothe Milach. Porto Alegre, 1998. 141p.