FÁBIO MEDEIROS FERREIRA

# DIVERSIDADE EM POPULAÇÕES SIMULADAS COM BASE EM LOCOS MULTIALÉLICOS

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

VIÇOSA
MINAS GERAIS – BRASIL
2007

**Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV**

T
F383d
2007

Ferreira, Fábio Medeiros, 1977-
Diversidade em populações simuladas com base em locos multialélicos / Fábio Medeiros Ferreira. – Viçosa, MG, 2007.
xiii, 177f. : il. ; 29cm.

Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 154-177.

1. Genética - Simulação por computador. 2. Genética de populações - Simulação por computador. 3. Marcadores genéticos - Simulação por computador.
I. Universidade Federal de Viçosa. II.Título.

CDD 22.ed. 576.5

FÁBIO MEDEIROS FERREIRA

# DIVERSIDADE EM POPULAÇÕES SIMULADAS COM BASE EM LOCOS MULTIALÉLICOS

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

APROVADA: 18 de maio de 2007.

| | |
|---|---|
| _____ | _____ |
| Prof. Luiz Antônio dos Santos Dias | Prof. Pedro Crescêncio Souza Carneiro |
| (Co-Orientador) | (Co-Orientador) |
| _____ | _____ |
| Prof. Marcos Ribeiro Furtado | Prof[a]. Ana Maria Waldschmidt |

_____

Prof. Cosme Damião Cruz

(Orientador)

*A todos os meus familiares,*
*À minha irmã, Luciana,*
*Ao meu querido afilhado e sobrinho, Thiago*
*À minha querida e amada, Rafaela,*
*ofereço.*
*Aos meus pais, Laudelino e Maria de Lourdes,*
*meu alicerce e fonte de amor,*
*dedico.*

## AGRADECIMENTOS

A Deus, por trilhar o meu caminho, com paz e muita proteção.

À Universidade Federal de Viçosa, pela oportunidade e excelência do ensino recebido.

Ao CNPq, pela concessão da bolsa de estudos durante esses quatro anos.

Ao professor Cosme Damião Cruz, pela orientação, transmissão de conhecimentos, incentivo e confiança no meu trabalho. Em especial, à amizade.

Ao professor José Ivo Ribeiro Júnior, pela orientação durante o mestrado e, acima de tudo, pelo estímulo ao meu crescimento profissional.

Aos professores Carlos Henrique Osório Silva, Cláudio Horst Bruckner, Luiz Alexandre Peternelli, Luiz Antônio dos Santos Dias, Márcio Henrique Pereira Barbosa, Marcos Ribeiro Furtado, Pedro Crescêncio Souza Carneiro, pela amizade e trabalho conjunto.

Aos professores Adair José Regazzi, Aluízio Borém, Ana Maria Waldschmidt, Carlos Sigueyuki Sediyama, Derly José Henriques da Silva, Luiz Orlando de Oliveira, José Marcelo Soriano Viana, Paulo Roberto Cecon, Glauco Viera Miranda, Tuneo Sediyama, Múcio Silva Reis, Ney Sussumu Sakiyama, Ricardo Frederico Euclydes, Rita Márcia Andrade Vaz de Mello, pela agradável convivência e contribuição à minha formação profissional.

Aos amigos Carlão, Cláudia Fogaça, Cleiton, Edmar, Fabiano Caliman, Giselda, Gustavo, Gustavo Fialho, Leandro Barbosa, Léo, Leandro Vagno, Leonarda, Lívia, Luiz Pessoni, Marcão, Márcia Costa, Maurinho, Odilon, Rodrigo Rocha, Talles, Thiago Lívio e Willian pela camaradagem e momentos de descontração durante esses quatro anos.

Aos demais amigos, fora do âmbito estudantil; companheiros de república; colegas do curso de Genética e Melhoramento; do laboratório de Bioinformática e da UFV, pela força durante esses anos em Viçosa.

À Rita de Cássia, pela amizade. À Rosemary e demais funcionários, pela ajuda prestada durante o curso.

Ao Grupo de Estudos em Genética e Melhoramento, pela oportunidade e proveitosa experiência.

À cidade de Viçosa, em nome de seus moradores e comerciantes, pela acolhida e prestação de serviço.

## BIOGRAFIA

Fábio Medeiros Ferreira, filho de Laudelino da Silva Ferreira e de Maria de Lourdes Medeiros Ferreira, nasceu no Rio de Janeiro, Estado do Rio de Janeiro, no dia 25 de agosto de 1977.

Nessa cidade estudou no Colégio Pedro II – Unidade São Cristóvão, de 1985 a 1995. Em maio de 2001, graduou-se Engenheiro Agrônomo pelo Centro de Ciências Agrárias da Universidade Federal do Espírito Santo.

Iniciou, em abril de 2001, o curso de Mestrado em Genética e Melhoramento, na Universidade Federal de Viçosa, em Viçosa, Minas Gerais, tendo defendido tese em 27 de fevereiro de 2003.

Em março de 2003, iniciou o curso de Doutorado em Genética e Melhoramento, também na Universidade Federal de Viçosa. Durante o doutorado, desenvolveu atividades de pesquisa nas áreas de Biometria, Simulação e Diversidade Genética. Junto a outros colegas de curso, ajudou no desenvolvimento do Grupo de Estudos em Genética e Melhoramento da UFV. No dia 18 maio de 2007, submeteu-se aos exames finais de defesa de tese.

**SUMÁRIO**

## RESUMO

FERREIRA, Fábio Medeiros, D.Sc., Universidade Federal de Viçosa, maio de 2007. **Diversidade em populações simuladas com base em locos multialélicos**. Orientador: Cosme Damião Cruz. Co-Orientadores: Luiz Antonio dos Santos Dias e Pedro Crescêncio Souza Carneiro.

A disponibilidade de novos referenciais teóricos e práticos continua sendo bem vinda e importante na orientação sobre o uso de *softwares* e métodos biométricos, para que se possa aproveitar melhor o conjunto de dados e interpretar corretamente os resultados. Neste estudo foram simuladas populações-base, submetidas a algumas gerações de acasalamento ao acaso e autofecundação, formação de populações híbridas $F_1$, gerações segregantes $F_n$ e de retrocruzamento, com informações provenientes de 20 locos codominantes e multialélicos. Várias análises biométricas foram aplicadas em nível intra e interpopulacional, utilizando-se de sete programas computacionais disponibilizados gratuitamente na *internet*. Com isso objetivou-se: i) avaliar a diversidade genética entre e dentro das populações; ii) comparar as metodologias empregadas; iii) verificar a potencialidade e funcionalidade dos programas computacionais e iv) auxiliar pesquisadores no uso de vários métodos biométricos para análise de dados genéticos e na execução das análises através dos programas. Destacaram-se as medidas, número efetivo de alelos ($A_e$), conteúdo de informação polimórfica (PIC), índice Shannon-Wiener (H'), heterozigosidade observada ($H_o$), índice de fixação/endogamia (*f*) e heterozigosidade esperada ($H_e$ ou $\overline{D}_j$), como ótimos descritores da diversidade genética intrapopulacional. Para um grande tamanho amostral ($N_i$ = 200), os diferentes estimadores do coeficiente de fixação/endogamia pouco diferiram em relação a suas estimativas, assim como para as medidas viesadas e não viesadas de heterozigosidade esperada. Os valores de tamanho efetivo ($N_e$) sendo inversamente proporcional ao coeficiente de endogamia, revelaram que autofecundações proporcionam menores $N_e$. As populações sob acasalamento ao acaso tiveram valores de $N_e \cong N_i$, pois atendem aos pressupostos de uma população idealizada de cruzamentos aleatórios.

Populações que sofrem autofecundação, como as gerações segregantes $F_n$, foram unânimes quanto ao desequilíbrio dentro de seus locos. O desequilíbrio nos locos de populações híbridas ocorre para aqueles em que não há diferenças de freqüências alélicas entre as populações cruzadas. O processo de hibridação nas populações de retrocruzamento pode levar um loco ao equilíbrio de Hardy-Weinberg (EHW) ou não, o que vai depender das combinações gênicas e genotípicas entre o hibrido $F_1$, ou a geração antecedente de retrocruzamento, com o genitor recorrente. Desvios nas proporções do EHW servem à definição de hipóteses alternativas em relação ao excesso e deficiência de heterozigotos. Os testes de qui-quadrado, razão de verossimilhança e teste exato com permutações foram concordantes na detecção ou não do EHW. Com os avanços das gerações de acasalamento ao acaso tende-se ao equilíbrio gamético e diminuição do coeficiente de desequilíbrio. Os testes de qui-quadrado ($\chi^2$) e razão de verossimilhança ($G^2$) foram bastante concordantes em relação à detecção ou não de desequilíbrio gamético. No geral, as medidas de distância empregadas no estudo da diversidade interpopulacional formaram grupos bem semelhantes, com a aplicação dos métodos de agrupamento de Tocher, UPGMA e projeção bidimensional (2D). No entanto, medidas de distância com filosofias afins, proporcionaram em algumas situações agrupamentos idênticos, como no caso das distâncias Euclidiana média ($D_E$) e de Roger ($D_R$); Latter (1972 e 1973) ($D_{L72}$ e $D_{L73}$) e Reynolds ($D'_{RWC}$); Rogers modificada ($D_{GS}$), mínima ($D_m$) e Reynolds ($D_{RWC}$) e; complemento do cosseno ($D_{COS}$), comprimento da corda ($D_{CC}$) e Nei et al. (1983) ($D_{N83}$). Comparações ao nível genotípico, a exemplo da distância genotípica de Hedrick ($D_H$) proporcionam informação adicional, a qual distâncias gênicas podem não promover. As medidas de distância geométricas ($D_E$, $D_R$, $D_{GS}$, $D_{COS}$ e $D_{CC}$) apresentaram maior eficiência de projeção gráfica. Verificou-se que com o avanço nas gerações de retrocruzamento, há realmente uma recuperação do genoma parental recorrente, refletidos na maior diferenciação genética entre população de retrocruzamento com o genitor doador, em situação de não seleção de genótipos. A análise de variância de freqüências alélicas promoveu uma partição mais eqüitativa da variação de indivíduos dentro de populações e de

genes dentro de indivíduos, ao passo que a análise molecular de variância (AMOVA) atribuiu grande parte da variação a genes dentro de indivíduos. A variação entre populações e a soma da variação dentro de populações e dentro de indivíduos, foi quantificada de maneira idêntica pelos métodos de freqüências alélicas e de Nei. Valores das estatísticas F foram iguais pelos métodos de análise de variância de freqüências alélicas e AMOVA. A análise via estatísticas F permitiu o entendimento de como se distribui a variação entre populações parentais e gerações segregantes. Os programas utilizados se mostraram adequados para o estudo de diversidade genética de populações, com locos multialélicos. Destacam-se os programas GENES e PowerMarker, em função da variedade de métodos de análise direcionados a este estudo. O estudo simulado possibilitou realizar comparações metodológicas e de programas computacionais de maneira satisfatória, sendo este um referencial teórico que pode auxiliar pesquisadores nas inferências sobre a genética de populações.

## ABSTRACT

FERREIRA, Fábio Medeiros, D.Sc., Universidade Federal de Viçosa, May, 2007. **Diversity in populations simulated on the basis of multiallelic loci.** Adviser: Cosme Damião Cruz. Co-Advisers: Luiz Antonio dos Santos Dias and Pedro Crescêncio Souza Carneiro.

The availability of the new theoretical and practical referential systems has been welcome and important as a guide to using softwares and biometric methods, since it makes possible a better use of the data file and the correct interpretation of the results. So, the following simulations were accomplished: base-populations subjected to either randomly mating and selfings; formation of hybrid populations $F_1$; segregant generations $F_n$ and backcrossing generations provided with information from twenty codominant and multiallelic loci. Several biometric analyses were applied at intra- and inter-populational levels**,** by using seven computer programs that were gratuitously available in internet. The objectives were: i) to evaluate the genetic diversity either between and within populations; ii) to compare the applied methodologies; iii) to verify both potentiality and functionality of the computer programs; and iv) to help the researchers either in using several biometric methods for the genetic data analysis and in accomplishing the analyses through the programs. The following measures were distinguished as optimum descriptors of the intra populational: the allele effective numbers ($A_e$); polymorphic information content (PIC); Shannon-Wiener index (H'); observed heterozygosity ($H_o$); fixation/endogamy index (f); and expected heterozygosity ($H_e$ or $\overline{D}_j$), as optimum descriptors of the intrapopulational genetic diversity. For a high sample size ($N_i$ = 200), the different estimators of the endogamy/fixation coefficient slightly differed concerning to their estimates, as well as for the biased and non-biased expected heterozygosity. Because the effective-sized values ($N_e$) were inversely proportional to the endogamy coefficient, they revealed selfings to provide lower $N_e$. The populations under random mating showed values $N_e \cong N_i$, since they fulfill the presuppositions of an idealized population with random matings. Those populations under selfing such as the segregant generations $F_n$ were

unanimous concerning to disequilibrium inside their loci. The disequilibrium in the hybrid population loci occurs for those where there are no different allelic frequencies between crossed populations. The hybridization process in the backcrossing populations can or cannot lead one locus to Hardy-Weinberg equilibrium (HWE), what will depend from both genic and genotypic combinations between the hybrid $F_1$ or the antecedent backcrossing generation with the recurrent genitor. Deviations in the HWE proportions are useful to the definition of alternative hypotheses concerning to either excess or deficiency of heterozygotes. The tests of qui-square, likelihood ratio**,** and the exact test with permutations were in accordance for the HWE detection or not. As the random mating advances**,** there is a tendency to reach the gametic equilibrium and the disequilibrium coefficient to be decreased. The qui-square tests $\chi^2$ and likelihood ratio ($G^2$) were in concordance in relation to the detection or non-detection of the gametic disequilibrium. In general, the distance measures used in the interpopulational diversity study rather composed very similar groups, when applying the clustering methods by Tocher, UPGMA and two-dimensional projection (2D). Though, the distance measures with similar philosophies and under some situations provided identical groupings, such as the distances: the mean Euclidian ($D_E$) and Roger ($D_R$); Latter (1972 and 1973) ($D_{L72}$ and $D_{L73}$) and Reynolds ($D`_{RWC}$); modified Rogers ($D_{GS}$), minimum ($D_m$) and Reynolds ($D_{RWC}$)**;** as well as the cosine complement ($D_{COS}$), cord length ($D_{CC}$) and Nei et al. (1983) ($D_{N83}$). Comparisons at genotypic level, such as the Hedrick genotypic distance ($D_H$) provide additional information at which the gene distances probably cannot promote**.** The geometric distance measures ($D_E$, $D_R$, $D_{GS}$, $D_{COS}$ and $D_{CC}$) showed higher efficiency in graphic projection. It was verified that as the backcrossing generations advanced, the recovery of the recurrent parental genome really occurs, which is reflected on higher genetic differentiation between the backcrossing population and the donor genitor. The variance analysis of the allelic frequencies promoted an equal partition of either individual variations within populations and genes within individuals, whereas the molecular variance analysis (AMOVA) ascribed great part of the variation to the genes within individuals. The variation among populations and the

sum of the variation within populations and within individuals were identically quantified by the allelic frequencies and Nei methods. The values of the statistics F were equal, based on the allelic frequency variance analysis and AMOVA. The analysis by statistics F allowed to understand how the variation is distributed between parental populations and segregant generations. The programs under study showed to be appropriate for the study of the genetic diversity in populations with multiallelic loci. The programs GENES and PowerMarker are distinguished as a function of the several analysis methods directed to this study. The simulated study made possible to accomplishing satisfactory comparisons among either the available methodologies and the computer programs, therefore constituting a theoretical referential that can help the researchers to draw inferences on the populational genetics.

# 1. INTRODUÇÃO

O conhecimento sobre a diversidade genética intra e, ou, interpopulacional é de fundamental importância ao melhoramento genético de plantas, pois permite ao melhorista direcionar sua estratégia de seleção, realizando e, ou, predizendo, mudanças em magnitude e sentido desejados, além de fornecer bases para a compreensão evolutiva da espécie alvo (Dias, 1998; Cruz, 2005). Especificamente este conhecimento tem proporcionado importantes contribuições no gerenciamento de bancos de germoplasma e na conservação de recursos genéticos, onde a riqueza alélica e genotípica das espécies encontra-se em seu estado natural ou resguardada pelos acessos, o que permite a médio e longo prazo ser explorada pelos melhoristas. No melhoramento, estudos sobre a diversidade genética, têm auxiliado na identificação de linhagens mais produtivas e combinações parentais adequadas à obtenção de híbridos altamente heteróticos, que possibilitem maior segregação em recombinações, com o aparecimento de genótipos transgressivos.

Os modelos teóricos desenvolvidos na Genética de Populações respondem aos mais variados questionamentos sobre o comportamento das populações ou espécies. Por exemplo, o modelo reprodutivo de acasalamento ao acaso possui uma função importante em outros modelos da genética de populações porque freqüentemente serve como um ponto de partida para a consideração de situações mais realísticas, embora algumas complicações o acompanhem, como a dependência sobre a característica e a subestruturação da população (Hartl & Clark, 1997). O equilíbrio de Hardy-Weinberg também é outro modelo básico que permite desvendar as associações existentes entre alelos de um mesmo loco ou de locos diferentes, caracterizando assim a estrutura genética de uma população e dando condições de interpretar quais são os fenômenos ecológicos e genéticos atuantes sobre a mesma.

A teoria da genética de populações está preocupada, principalmente, em entender duas variáveis bastante conectadas: a freqüência gênica e a freqüência genotípica. Elas são descritores básicos, no entanto, ênfase pode ser dada a heterozigosidade ou diversidade gênica.

A freqüência de heterozigotos assume importância uma vez que cada indivíduo heterozigoto carrega consigo diferentes alelos e representa a existência de variação. Em populações de autofecundação onde a variação resulta da presença contínua de diferentes homozigotos, a medida de diversidade gênica é a mais indicada (Weir, 1996).

Certamente o rejuvenescimento da Genética de Populações ocorreu com o desenvolvimento da biologia molecular (Nei & Kumar, 2000), em que os marcadores moleculares e bioquímicos tornaram-se novas fontes de investigação acerca da diversidade genética das espécies. Nos tempos atuais, a análise de dados genéticos é rotineiramente usada na pesquisa para genes que regulam doenças em humanos ou características de interesse econômico em plantas cultivadas e animais domesticados. No entanto, os benefícios da genética molecular somente serão atingidos se as análises biométricas forem aplicadas de forma apropriada.

Teorias estatísticas foram desenvolvidas para genes que realmente representem regiões codificadoras de proteínas ou que dependem das propriedades físicas de produtos gênicos, ou ainda para aqueles que não possuam associação com regiões codificadoras. Existem vários métodos de análise genético-estatísticas específicas e direcionadas exclusivamente para dados de marcadores com o intuito de fornecer informações sobre a diversidade das unidades taxonômicas (espécies, populações, cultivares, acessos, etc).

Muitos destes métodos estão descritos de maneira dispersa na literatura através de publicações em periódicos. Para se ter uma idéia, dos sete programas computacionais utilizados neste trabalho, a maioria baseou-se em um único livro texto ("*Genetic Data Analysis"*, de Bruce S. Weir, 1996) como referencial teórico em seus manuais. O que pode acabar ocasionando preferências por alguma(s) metodologia(s). Além disso, há elevado número de *softwares* para a análise genética de populações. Poucos trabalhos de revisão se destinam a comparação entre programas, informando a potencialidade e funcionalidade dos mesmos.

Em algumas ocasiões o geneticista e, ou, melhorista pode ter dúvidas quanto a metodologia a ser empregada. Isto porque muitas delas são complementares ou se sobrepõem. Outras, embora se assemelhem, são

fundamentadas em princípios e pressuposições diferenciadas, a exemplo das medidas de distância e as diferentes maneiras de se estimar os índices de fixação (estatísticas F de Wright). Um conhecimento mais amplo deve ser considerado, mesmo que a análise seja realizada com a ajuda de um *software*. Um clique em um ícone do programa pode não ser uma tarefa tão simples, pois o usuário, por desconhecimento, pode estar escolhendo as opções de análise que não se ajustam ao seu conjunto de dados, e consequentemente, não trarão respostas condizentes.

Assim, 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. Justifica-se, portanto, a disponibilidade de novos referenciais teóricos e práticos, que orientem a utilização de aplicativos computacionais e recursos biométricos para um melhor aproveitamento dos dados e interpretação correta dos resultados obtidos.

Populações constituídas experimentalmente, como aquelas desenvolvidas no melhoramento genético, possuem sistemas controlados de acasalamento e apresentam um comportamento mais previsível da sua variabilidade genética, ao contrário das populações naturais. Neste sentido, as populações experimentais podem ser consideradas como modelos, o que possibilita estudos comparativos, com a aplicação de diversas metodologias. Adicionalmente, o uso da simulação torna-se uma alternativa viável na detecção e comprovação de técnicas estatísticas mais eficientes.

O presente trabalho baseou-se num estudo simulado de populações submetidas a acasalamentos ao acaso, autofecundação e hibridações planta a planta ou por mistura de pólen. Foram constituídas populações base, populações híbridas, gerações de acasalamento ao acaso, autofecundações, gerações segregantes $F_n$ e de retrocruzamento, com informações provenientes de 20 locos codominantes e multialélicos. Diversas análises biométricas foram aplicadas em nível intra e interpopulacional, utilizando-se de sete programas disponibilizados na *internet*. Dentro deste contexto, os objetivos foram os seguintes:

i) avaliar a diversidade genética entre e dentro das populações simuladas;

ii) comparar metodologias genético-estatísticas com propósitos afins no estudo de estrutura e diversidade genética;

iii) verificar o desempenho de programas computacionais no estudo da diversidade genética entre e dentro de populações, e submetê-los a comparações sobre suas funcionalidades;

iv) orientar pesquisadores interessados na área quanto a utilização de métodos biométricos na genética de populações e na execução das análises através de alguns programas computacionais.

# 2. REVISÃO BIBLIOGRÁFICA

## 2.1. Germoplasmas e populações experimentais em plantas

Um programa de melhoramento ou de conservação de germoplasma estará propenso ao sucesso a partir do conhecimento da quantidade de variação existente na espécie sob estudo. A maioria das espécies apresenta grande diversidade genética, sendo possível selecionar e recombinar genótipos mais adaptadas, de melhor qualidade e mais eficientes (Cruz, 2005).

Assim como em outros setores, o desenvolvimento tecnológico e científico permitiu ao melhoramento genético ganhar o *status* de ciência e passar a exercer um papel preponderante no desenvolvimento qualitativo e quantitativo das plantas cultivadas. A teoria do melhoramento é fundamentada nos vários ramos da genética, como a Mendeliana, de Populações, Quantitativa, Citologia e Molecular, além da Evolução. Embora atuem em conjunto para elucidar os mais variados fenômenos genéticos, é válido destacar a importância da Genética de Populações, principalmente com os avanços da biologia molecular. Novas perspectivas na pesquisa em conservação de espécies e na biologia populacional proporcionaram rejuvenescimento desta área (Nei & Kumar, 2000).

Na genética de populações, a palavra população nem sempre se refere à espécie como um todo; conceitua-se como um grupo de indivíduos, pertencentes a uma mesma espécie, vivendo dentro de uma área suficientemente restrita, com sistema de acasalamento supostamente conhecido, de forma que qualquer indivíduo possa se acasalar com qualquer outro do sexo oposto (Hartl & Clark, 1997), possibilitando a formação de descendentes em freqüência proporcional à contribuição gamética de seus genitores (Cruz, 2005). Essa definição assume um componente genético (indivíduos pertencentes a mesma espécie) e um componente espacial (convívio dentro de uma mesma área).

Sabe-se que o conhecimento do padrão reprodutivo e estrutura genética de uma população permitem estabelecer e relacionar proporções

genotípicas e alélicas dentro e entre populações no espaço e no tempo. A manipulação de modelos de cruzamento e seleção tem sido a base do melhoramento genético, tanto animal quanto vegetal. De maneira geral, têm sido utilizados vários tipos de cruzamentos e, consequentemente, populações constituídas de genótipos característicos, visando à obtenção de materiais genéticos com atributos desejáveis. De acordo com Liu (1997), as populações obtidas por cruzamentos controlados entre genitores selecionados podem ser consideradas como autênticas populações de melhoramento ou experimentais. Nelas a estrutura e variabilidade genética são previsíveis, senão conhecidas. Assim o melhorista tem a possibilidade de predizer mudanças em magnitude e sentido desejado e formular hipóteses acerca do seu comportamento genético. Enquadram-se as populações de híbridos $F_1$, gerações avançadas $F_n$ (ou $S_n$), retrocruzamentos entre outras. Estas populações, ao serem obtidas, passam por seleções e avaliações, até que se tornem de uso comercial.

As populações base, embora não constituam populações experimentais propriamente ditas, são fundamentais nos programas de melhoramento, pois a partir delas são extraídos um ou mais genótipos superiores, sendo que um ou alguns destes genótipos, após os trabalhos do melhorista, se tornarão cultivares. As populações base devem ser detentoras de variabilidade genética suficiente para o seu devido fim.

É fato nos programas de melhoramento a busca por “novos” alelos. Eles, ao serem incorporados nos cultivares, que serão lançados, propiciarão atributos desejáveis, sejam de caráter comercial ou puramente agronômico. Evidentemente que novos alelos podem surgir em qualquer população por meio de mutação. Os acessos (genótipos) constituintes dos bancos de germoplasma, como espécies selvagens, raças locais, linhagens, variedades etc, são unidades genéticas tão importantes quanto as anteriormente citadas, pois inserem-se no processo de pré-melhoramento, antecedente a fase inicial de um programa de melhoramento e tem como objetivo, justamente, identificar novos genes, favoráveis e úteis a população base e aos materiais elites.

### 2.1.1. Populações naturais

Nos últimos 20 anos uma preocupação latente tem sido dirigida aos efeitos maléficos do desrespeito aos ecossistemas. A fragmentação de espécies, a perda da variabilidade genética e mesmo a extinção de espécies são processos nítidos e preocupantes em muitas regiões. Dentro deste contexto, muitos estudos biológicos têm focado suas atenções às populações naturais de espécies selvagens (nativas), a exemplo do patauá (*Oenocarpus bataua* Mart.), a macaúba (*Acrocomia aculeata* (Jacq.) Lodd. ex Mart.), o umbu (*Spondias tuberosa* Arr. Câm.), a castanha sapucaia (*Lecythis pisonis*), entre outros. Estas espécies encontram-se em ritmo de domesticação (Paiva & Valois, 2001) e com potencial agrícola, florestal, medicinal e energético. Explorá-las de forma sustentável e preservar sua diversidade genética passa a ser o objetivo central da conservação genética e de novos programas de melhoramento.

As populações naturais são aquelas geradas por acasalamentos ocorridos naturalmente, ou seja, sem controle artificial. Trata-se de unidades sobre as quais incide o manejo para a conservação e utilização dos recursos naturais, além de fonte de germoplasma para os programas de melhoramento genético (Robinson, 1998). Algumas espécies podem ser geradas por acasalamentos preferenciais ou mesmo por completa autofecundação, ao invés do acasalamento ao acaso unicamente. Exemplos de populações naturais são dados por famílias de meios-irmãos, populações mistas, cujas árvores são constituídas por sementes oriundas de polinização cruzada e autopolinização e, populações com sobreposição de gerações (Liu, 1997). As populações naturais também são populações locais. Estas unidades locais, possivelmente estruturadas geograficamente, são de grande interesse, pois é a partir delas que a evolução adaptativa se origina, com mudanças sistemáticas nas freqüências alélicas (Hartl & Clark, 1997). Populações situadas no centro da distribuição geográfica da espécie geralmente são mais densas, contínuas e contém maior variabilidade genética. Populações marginais são mais isoladas quanto ao fluxo gênico e, por isso, geneticamente menos variáveis. Entretanto, populações pequenas com baixa variabilidade podem, alternativamente, representar populações

adaptadas a um ambiente ecológico geograficamente restrito. As informações sobre a variabilidade em populações naturais são fundamentais para o progresso de duas áreas de grande interesse atual: a especiação de florestas tropicais e a conservação de recursos genéticos (Moraes & Derbyshire, 2002).

A grande quantidade de variabilidade genética encontrada dentro das populações naturais, a exemplo da seringueira (Paiva et al., 1994), associada à forma de distribuição dos indivíduos nas populações naturais, é forte indicativo de que o cultivo racional de espécies autóctones permitirá novos enfoques e novas estratégias de melhoramento genético. Novos panoramas podem ser adotados nos programas, como evitar a uniformidade genética. Para espécies perenes, cujas populações base são compostas atualmente por um número reduzido de material clonal, mudaria para um amplo conjunto de plantas sexuadas, de modo a explorar a variabilidade existente nas populações naturais (Paiva & Valois, 2001). Nessa linha de raciocínio o melhoramento genético deixaria de explorar somente o vigor híbrido dos cruzamentos com clones produtivos e resistentes. O foco poderia ser os descendentes de plantas com características favoráveis, passando a explorar o melhoramento de populações. Daí obtém-se sementes melhoradas a médio e a longo prazo, visando ao pequeno produtor, em detrimento de uma maior produtividade das espécies, mas com a garantia de produção sem a necessidade de controle fitossanitário (Paiva & Kageyama, 1993). Trabalhos realizados com os mais variados tipos de espécies arbóreas, florestais e, ou, frutíferas, como o pinho caribenho (Zheng & Ennos, 1999); dendê (Barcelos et al., 2002); canela-amarela (Moraes & Derbyshire, 2002); cagaita (Zucchi et al., 2003); pimenta-de-macaco (Gaia et al., 2004); pequi (Melo Júnior et al., 2004) e pimenta-longa (Wadt & Kageyama, 2004) vêm demonstrando a preocupação com a manutenção do equilíbrio biológico das populações no ecossistema tropical.

Até mesmo em culturas anuais esta visão tem prevalecido. É o caso da cultura do arroz, em que Gao & Hong (2000) inferiram sobre a estrutura genética e sistema reprodutivo de populações silvestres (*Oryza rufipogon* Griff.) em diferentes regiões da China. Os autores constataram que a

conservação desses germoplasmas é potencialmente viável para uma futura exploração da diversidade genética.

### 2.1.2. Acessos

Os bancos de germoplasma são detentores de uma representação gênica bastante ampla. Na verdade, cada unidade de germoplasma deve ser uma cópia única do material genético e representativa do organismo vivo de interesse atual ou potencial. Consequentemente, o germoplasma é o elemento dos recursos genéticos que maneja a variabilidade genética inter e intraespecífica, para conservá-la e utilizá-la na pesquisa em geral, especialmente em programas de melhoramento genético. A conotação do termo germoplasma é eminentemente técnica, não envolvendo sua importância econômica e política, ao contrário do termo recursos genéticos que agrega um valor econômico presente nos genes (Querol, 1993). Existem cinco categorias de germoplasmas, segundo Hoyt (1992): i) parentes silvestres; ii) populações locais (do termo em inglês *landraces*); iii) cultivares que foram substituídas; iv) linhagens experimentais, mutações e outros produtos dos programas de melhoramento e v) cultivares modernas.

Basicamente, existem duas estratégias de conservação: *in situ* e *ex situ*. A conservação genética *in situ* de uma espécie é a manutenção de amostras de populações da mesma e de parentes próximos em seu ecossistema natural, com condições de diversidade genética que permitam o seu desenvolvimento, a manutenção de suas populações e a continuidade da evolução (Kageyama et al., 2001). É mais indicada para espécies silvestres das plantas cultivadas, forrageiras, fruteiras e espécies florestais. A conservação *ex situ* caracteriza-se pelo manejo do germoplasma fora do local de origem, tanto autóctones como exótico - que não apresentam uma utilidade imediata sem uma seleção prévia para adaptação em uma determinada área (Hallauer & Miranda Filho, 1988). A conservação *ex situ* encontra-se nos bancos e coleções de germoplasma, que são as estruturas físicas onde armazena-se os acessos (Valois et al., 2001). Como o mercado é muito competitivo e dinâmico, tempo é um recurso teoricamente escasso,

assim muitos programas de melhoramento evitam trabalhar com genótipos selvagens, raças locais e materiais exóticos disponíveis em coleções. Embora o cenário tenha apresentado mudanças na concepção da interação de recursos genéticos e programas de melhoramento, ainda existe uma lacuna entre essas atividades (Nass, 2001).

Alegando o estreitamento da base genética e a vulnerabilidade do cultivo, Almeida et al. (2005) realizaram estudo em que identificaram dentre 207 acessos do banco de germoplasma de cacau pertencente ao CEPLAC de Rondônia, alguns genótipos de interesse para o programa de melhoramento do cacaueiro na região. Pelos mesmos motivos, Brown-Guidera et al. (2000) avaliaram o parentesco entre 105 genótipos de soja do germoplasma do Norte dos Estados Unidos. Os autores sugeriram que alguns grupos de plantas introduzidas (PIs) no programa de melhoramento são extremamente úteis aos melhoristas dada a sua diversidade genética.

### 2.1.3. Populações base

O processo de formação da população base difere com o sistema reprodutivo da espécie a ser melhorada. Em espécies alógamas a forma de reprodução natural ocorre com o acasalamento entre plantas, assim, as plantas de uma geração qualquer são naturalmente oriundas de gametas femininos e masculinos de diferentes plantas. O melhoramento genético de espécies alógamas é feito aumentando-se as freqüências dos alelos favoráveis das características de importância econômica e, ou, agronômica que serão submetidas à seleção (Souza Jr, 2001). O foco dos programas das espécies alógamas são as populações híbridas, ou seja, o cultivar híbrido ou clone.

As plantas de espécies alógamas de propagação assexuada, como a cana-de-açúcar, mandioca, eucalipto etc, normalmente são muito heterozigotas e, por isso, apresentam elevada depressão por endogamia. As populações a serem submetidas à seleção são formadas pelos mais variados cruzamentos entre clones. A geração $F_1$ oriunda destes cruzamentos é a referencial, isto é, nela que se inicia o processo de seleção.

Assim, a geração $F_1$ proveniente do cruzamento de dois clones apresenta variabilidade genética, e o nível dessa variabilidade depende do nível de heterozigosidade dos clones e de sua divergência genética.

As espécies de propagação sexuada, como o milho e mamão, têm suas populações base constituídas de grupos heteróticos distintos e neles incluem-se populações, linhagens, sintéticos, de tal maneira que dentro de um grupo os cruzamentos não manifestam heterose ou esta é muito baixa, enquanto que entre grupos os níveis de heterose são elevados. Esta característica é fundamental às populações base, pois a heterose só se manifesta quando há divergência genética entre as populações e quando o caráter tiver alto nível de dominância unidirecional (Falconer & Mackay, 1996). Os grupos heteróticos são produzidos utilizando-se geralmente de delineamentos genéticos, a exemplo dos cruzamentos dialélicos. Marcadores moleculares também permitem a alocação e a diferenciação das linhagens quanto a seu grupo heterótico. As populações base são produzidas dentro de cada grupo heterótico. Podem ser formadas populações de base genética ampla - intercruzamento de diversas populações - ou de base estreita, denominadas de sintéticos e que são produzidas pelo intercruzamento de algumas linhagens homozigóticas. Além dessas, inclui-se as populações formadas pelo cruzamento de duas linhagens, escolhidas com base em sua performance *per se* e em cruzamentos com materiais de outros grupos heteróticos (Souza Jr, 2001).

Destacam-se o arroz, aveia, feijão, soja, tomate e trigo, como espécies autógamas anuais que ocupam grandes áreas produtivas e de cultivo mundial. Entre as perenes autógamas, que ocorrem em número reduzido, cita-se o pêssego e a nectarina. As espécies autógamas são aquelas que possuem flores hermafroditas e que reproduzem predominantemente por meio da autopolinização. Aceita-se que no máximo 5% de taxa de polinização cruzada possa existir entre as autógamas (Borém, 1998; Ramalho et al., 2001). Uma cultivar autógama é representada por um genótipo homozigótico, ou seja, uma linhagem, ou uma mistura deles.

A formação de população base em plantas autógamas tem sido realizada quando o programa de melhoramento faz uso do processo de

seleção recorrente. A seleção recorrente é qualquer processo cíclico de melhoramento que envolve a obtenção de famílias, avaliação e intercruzamento das melhores. Espera-se, aumentar gradativamente a freqüência de alelos favoráveis e por conseqüência melhorar a expressão fenotípica da característica sob seleção, sem reduzir a variabilidade genética da população (Borém, 1998). Exemplos, do seu emprego em espécies autógamas, estão relatados nas culturas da soja (Werner e Wicox, 1990) e feijão (Ranalli, 1996). Tendo em vista que um programa de seleção recorrente é demorado, o melhorista deve estar atento à escolha dos genitores na formação da população base. Recomenda-se 10 a 20 genitores como um número satisfatório na composição da população submetida a ciclos de seleção recorrente (Ramalho et al., 2001).

Embora não se constituía como uma população base em essência, em algumas espécies autógamas, o processo de seleção se inicia numa população básica desenvolvida ao longo do tempo a partir do cultivo de agricultores. Em algumas culturas de subsistência como o arroz e feijão, os agricultores não tem o hábito de comprar sementes anualmente, e acabam reutilizando os grãos como sementes por vários anos. Nessa condição, espera-se que ocorra variabilidade dentro desta população. Essa variabilidade é oriunda de misturas mecânicas de linhagens diferentes, seguidas muitas vezes de fecundação cruzada entre elas no campo e, da ocorrência de mutação. Assim, é esperado que o germoplasma usado por vários anos pelos agricultores seja uma mistura de linhas puras. Considerando o número de gerações de cultivo sucessivo e a área semeada anualmente, o número de linhagens genotipicamente diferentes dentro da população é enorme (Ramalho et al., 2001). Além disso, com a ação da seleção natural, espera-se que permaneçam apenas aquelas combinações gênicas e genotípicas que sejam mais adaptadas à região.

### 2.1.4. Populações híbridas

Os tipos de híbridos encontrados no mercado são simples, triplos e duplos. Um hibrido simples é produto do cruzamento de duas linhagens de

grupos heteróticos distintos. Um hibrido triplo é produzido pelo cruzamento de um hibrido simples proveniente de linhagens de um mesmo grupo com uma linhagem de outro grupo heterótico e um hibrido duplo é produzido pelo cruzamento de dois híbridos simples de grupos heteróticos distintos. O produção híbridos visa capitalizar a heterose. Por apresentar elevada depressão por endogamia, as linhagens de espécies alógamas geralmente apresentam baixa produtividade e, consequentemente, a produção de híbridos simples é prejudicada para algumas espécies. No entanto a composição dos híbridos triplos e duplos tende a contornar este problema (Borém, 1998).

Segundo Xavier (2003), a hibridação pode ser encarada como uma forma de explorar a heterose para características que representam maior produtividade, ou como meio de reunir atributos específicos de interesse que estão separados nos genitores. Os métodos preditivos de heterose são aqueles que relacionam a divergência dos genitores com o desempenho dos híbridos (Miranda et al., 1988).

A hibridação entre populações que possuem diferentes combinações adaptativas de genes pode aumentar consideravelmente a dimensão do conjunto gênico quanto a genes dotados de valores adaptativos diferentes, desde que os híbridos possam dar origem a progênies segregantes em gerações posteriores. Provavelmente, a maioria dos genótipos segregantes obtidos de tais populações híbridas terá valor adaptativo menor do que qualquer uma das populações ascendentes, porém uma pequena porção deles pode resultar em melhor adaptação a certos ambientes (Paiva & Valois, 2001).

Nos programas de espécies alógamas para reunir em um único genótipo alelos desejáveis, que se encontram em linhagens distintas, opta-se por cruzamentos biparentais ou multiparentais, de maneira a alterar a contribuição de cada genitor nos locos segregantes. Por exemplo, no cruzamento $[(P_1 \times P_2) \times P_3] \times P_4$, os genitores $P_1$ e $P_2$ contribuem, em média, igualmente com 12,5%, $P_3$ com 25% e $P_4$ com 50% dos alelos na população híbrida.

No melhoramento de milho, em nível mundial, o setor público e privado têm usado uma estratégia válida, porém de alto risco futuro. Trata-se

da obtenção de novas populações e, ou, linhagens a partir de gerações avançadas de híbridos comerciais (Nass, 2001). Seguindo esta tendência, provavelmente em um futuro próximo todas as populações híbridas serão muito semelhantes e estarão compartilhando o mesmo *background* genético, restringindo consideravelmente a base genética dos materiais cultivados.

Alguns programas de melhoramento de espécies autógamas também têm direcionado esforços a produção de populações híbridas, como é o caso da cultura do tomate. A vantagem das cultivares híbridas sobre as de polinização aberta em tomate é justificada pelos aumentos potenciais na produção de 25 a 40%, maturação mais precoce, melhor uniformidade, maior vigor inicial e de desenvolvimento, melhor qualidade, resistência a doenças e capacidade de adaptação mais ampla (Mello et al., 1988).

### 2.1.5. Populações segregantes de gerações avançadas ($F_n$)

Nos programas de melhoramento de plantas autógamas o objetivo final é ter na geração $F_n$, linhagens com alelos favoráveis para a grande maioria de locos. As sementes da geração $F_1$ são heterozigóticas para todos os locos cujos genitores contrastavam. A população $F_2$, derivada do intercruzamento dos indivíduos $F_1$'s ou da autofecundação deles, é a população segregante de maior variabilidade genética, pois nela encontram-se amostrados maior número de arranjos genotípicos. Com o avanço das gerações de endogamia, a freqüência de heterozigotos diminui e de homozigotos aumenta até a geração $F_\infty$, quando todos os locos estarão em homozigose.

Existem vários métodos de condução das populações segregantes. Ramalho et al. (2001) os agrupram em métodos que não separam as fases de endogamia e seleção e aqueles que separam as duas fases. No primeiro grupo destaca-se o método de seleção massal e o método genealógico. No segundo grupo encontra-se o método da população (*bulk*), descendência de uma única semente (SSD, do inglês *single seed descent*) e o método do *bulk* dentro de famílias (maiores detalhes em Borém, 1998; Ramalho et al., 2001).

Cita-se o uso de populações segregantes avançadas na obtenção dos cultivares de cevada, BRS 224 e BRS 225 (Minella et al., 2005a; Minella et al., 2005b, respectivamente). O processo de seleção foi basicamente o mesmo até se chegar as duas cultivares. Para cada caso, uma única planta foi selecionada na população $F_2$. A população $F_2$ foi desenvolvida no campo, onde algumas plantas foram selecionadas e avançadas até a geração $F_5$, em *bulk*, pelo método SSD em casa-de-vegetação. As plantas selecionadas de progênies $F_6$ cresceram no campo e foram colhidas em *bulk* para aumentar a quantidade de sementes. Uma linha endogâmica (CEV 96501 e CEV 96503) foi selecionada e colhida em *bulk* da população $F_7$. Após os ensaios, registro e proteção das cultivares, foram designadas como BRS 224 e BRS 225.

### 2.1.6. Populações de retrocruzamento

Populações de retrocruzamento são confeccionadas com o intuito de melhorar a expressão fenotípica de uma característica de um dado cultivar para a qual ele é deficiente. O termo retrocruzamento se refere aos repetidos cruzamentos dos indivíduos da população segregante com uma das linhagens genitoras. O genitor que contém o alelo que confere fenótipo desejado é denominado de não recorrente ou doador, ou seja, é utilizado apenas uma vez nos cruzamentos. No retrocruzamento a geração $F_1$ é cruzada com um dos genitores. O genitor que é submetido aos sucessivos cruzamentos com os indivíduos da população segregante é denominado de recorrente. O genitor recorrente deve ser cuidadosamente escolhido, pois deve apresentar bons atributos para todas as características, exceto aquelas que serão doadas pelo não recorrente (Borém, 1998). Provavelmente é o método mais adotado no melhoramento de espécies autógamas. Trata-se de uma alternativa eficiente para a transferência de pequenas proporções genômicas - geralmente caracteres de controle qualitativo - de genótipos não adaptados ou de espécies selvagens para genótipos elite (Lorencetti et al., 2006). Também tem sido utilizado na adaptação de germoplasma exótico (Nass, 2001).

Um fator complicador no retrocruzamento surge quando o gene a ser transferido está ligado a um outro regulador de fenótipo(s) indesejável(is), tornando o processo mais demorado. Espera-se que a recuperação do genitor recorrente no $RC_x$ ocorra na proporção de $(2^{x+1}-1)/2^{x+1}$, em que x é número de retrocruzamentos com o genitor recorrente (Cruz, 2005). Só que nem sempre a recuperação do genoma recorrente ocorre da maneira prevista. Blocos gênicos ligados ao(s) gene(s) de interesse (*linkage drag*) oriundos do genótipo doador podem ser inseridos ao genótipo recorrente e as proporções ao longo das gerações não encontram-se como o esperado (Frisch et al., 1999). Marcadores moleculares tornam-se ferramentas importantes no monitoramento da quantidade de DNA do genitor doador durante cada geração de retrocruzamento.

O controle gênico da característica a ser transferida deve ser levado em consideração, em virtude da necessidade ou não de testes da descendência para determinar seu genótipo. Se o alelo transferido é dominante, ele é facilmente identificado nas gerações de retrocruzamento, pois os indivíduos segregantes para o loco em questão terão constituição homozigota recessiva ou heterozigota. Neste caso, os homozigotos são descartados. Se o alelo presente no genitor doador é recessivo, faz-se necessário a cada geração de retrocruzamento submeter os indivíduos da população à autofecundação, de modo a identificar os genótipos recessivos.

Um fator que influencia o número de retrocruzamentos é a adaptação do genitor recorrente. Se ele é não adaptado, o número de retrocruzamentos tende a ser maior. Entretanto, com o emprego de marcadores moleculares também pode-se abreviar o número de gerações de retrocruzamento para a introgressão de genes exóticos (Tanksley et al. 1981).

Moraes et al. (2005) avaliaram a proximidade entre genótipos usados como genitores recorrentes e doadores, para alelos responsáveis pelo alto teor de proteína, orientados por marcadores microssatélites. Os autores definiram grupos de potenciais cruzamentos entre os genótipos recorrentes e genótipos doadores, a partir das menores distâncias genéticas entre recorrente e doador. Esta proposição permite otimizar o processo de

retrocruzamento, mantendo assim as características desejáveis do genitor recorrente. Com outro enfoque, Lorencetti et al. (2006), utilizando dados fenotípicos, atestaram a utilização de retrocruzamento como estratégia eficiente no desenvolvimento de populações segregantes promissoras em aveia. Ainda constataram que a superioridade das populações de retrocruzamentos sobre uma população $F_3$ foi dependente do cruzamento.

Alguns melhoristas questionam a obtenção de populações de retrocruzamento, afirmando que é um processo muito conservador, pois há tendência de permanecer com apenas um cultivar. Entretanto, é fácil visualizar que se durante o processo for identificada uma linhagem superior ao genitor recorrente esse pode ser substituído nos futuros retrocruzamentos, tornando-o um processo mais dinâmico (Ramalho et al., 2001).

## 2.2. Marcadores codominantes

Um determinado fenótipo, uma proteína ou um fragmento de uma seqüência de nucleotídeos do DNA codificador ou não de um gene, possuindo unidade(s) repetitiva(s) ou sendo único no genoma, podem representar uma forma alélica qualquer e serem utilizados como marcadores (Souza, 2001).

Nos estudos de diversidade genética, a partir de marcadores é importante e, talvez essencial, conceituar o termo polimorfismo genético. O polimorfismo refere-se à existência de dois ou mais alelos em um determinado loco com consideráveis freqüências relativas em uma população, geralmente, acima de 1% ou 5% (Nei & Kumar, 2000; Cole, 2003).

Inicialmente, os marcadores utilizados em estudos genéticos e de melhoramento eram controlados por genes de herança mendeliana e associados a características morfológicas, com fenótipo de fácil identificação visual – a exemplo do nanismo, cor de pétalas e morfologia foliar (Ferreira & Grattaplagia, 1998; Robinson, 1998). No entanto, os marcadores

morfológicos apresentavam grandes limitações. Dentre elas cita-se a expressão do marcador influenciada pelo estádio de desenvolvimento da planta e a sua interferência sobre o valor adaptativo. Na maioria dos casos tais marcadores genéticos estão restritos a espécies tidas como sistemas modelo, como o milho, tomate e ervilha e são controlados por genes dominantes, não permitindo distinguir quais são as plantas heterozigotas (Ferreira & Grattaplagia, 1998).

No início da década de 60 do século passado, os estudos genéticos sofreram uma revolução com o surgimento dos marcadores isoenzimáticos. De acordo com Carlini-Garcia et al. (2001), o surgimento dos marcadores bioquímicos e algum tempo depois dos marcadores moleculares, proporcionou um salto qualitativo e quantitativo em estudos sobre a estrutura populacional e o sistema reprodutivo das espécies vegetais. Atribuem-se inúmeras vantagens a estes marcadores quando comparados aos marcadores morfológicos. Em geral, eles apresentam neutralidade em relação a efeitos fenotípicos, com mínimo ou nulo efeito epistático ou pleiotrópico e exibem maior nível de polimorfismo.

A expressão gênica comumente apresentada pelos marcadores bioquímicos ou moleculares é a dominância completa ou codominância exibindo multialelismo. Eles podem ser utilizados para caracterizar o genótipo de uma planta a partir de amostras de células ou tecidos. Os marcadores baseados em DNA podem ser extraídos em qualquer estádio de desenvolvimento da planta, desde que extraídos em quantidades suficientes (Ferreira & Grattaplaglia, 1998). Tais atributos potencializam os estudos de diversidade genética, tornando possível predizer a magnitude da divergência entre dois acessos, bem como conhecer a estrutura e a variabilidade genética das populações, em nível enzimático (um produto expresso de um gene) ou genômico (fragmento na seqüência de nucleotídeos do DNA).

Comumente, é estabelecido o critério de classificação dos tipos de marcadores com base em suas expressões gênicas, podendo ser: i) dominantes, a exemplo do RAPD (polimorfismo de DNA amplificado ao acaso), AFLP (polimorfismo de comprimento de fragmentos amplificados) e ISSR (inter-seqüências simples repetidas); e ii) codominantes, dentre os quais podem ser citadas as isoenzimas, RFLP (polimorfismo de

comprimento no comprimento de fragmentos de restrição) e microssatélites ou SSR (seqüências simples repetidas).

As informações provenientes dos marcadores dominantes são corriqueiramente codificadas na forma binária, com 1 (um) para presença e 0 (zero) para ausência da banda (marca ou loco) no gel. Assim o acesso ou indivíduo heterozigoto para um determinado loco não é distinguido do acesso homozigoto dominante. Neste contexto, a classe de marcadores codominantes permite identificar os dois alelos de um mesmo loco, considerando um organismo diplóide. A seguir são descritos os marcadores codominantes citados anteriormente.

### 2.2.1. Isoenzimas

Market & Moller (1959) designaram como isoenzimas as formas moleculares múltiplas de enzimas que ocorrem num organismo e dentro de uma espécie, como resultado da presença de mais de um gene codificando cada uma das enzimas. Na realidade os organismos comumente sintetizam formas moleculares múltiplas de enzimas com a mesma especificidade enzimática (Caixeta et al. 2006).

Resumidamente a detecção de isoenzimas envolve três processos básicos. No primeiro ocorre a extração da enzima do tecido vegetal escolhido. A segunda etapa visa separar as enzimas por eletroforese e, finalmente, tem-se a revelação das bandas por métodos de coloração histoquímica do gel de poliacrilamida, amido ou ágar.

As isoenzimas de um mesmo grupo são diferentes entre si na seqüência de aminoácidos que possuem e, consequentemente, influenciam na estrutura protéica da enzima. Naturalmente que a banda visualizada é o produto da reação enzimática no meio-suporte, obtendo assim o zimograma, que é o nome atribuído ao conjunto de bandas reveladas no gel. A interpretação dos padrões de bandas resultantes dá-se por meio do número de subunidades da enzima. As enzimas monoméricas são constituídas por um único polipeptídeo, enquanto que as diméricas, triméricas e tetraméricas, possuem dois, três e quatro polipeptídeos, respectivamente.

Ocasionalmente, subunidades de uma enzima podem ser codificadas por mais de um loco gênico, de maneira que a migração das bandas de cada loco é visualizada em zonas diferentes no zimograma (veja em Alfenas et al., 1991).

O princípio fundamental das isoenzimas é que diferenças na mobilidade das mesmas em um campo elétrico são resultantes de diferenças em nível de seqüências de DNA que as codificam. A migração de uma enzima no gel depende de seu peso molecular, sua conformação e carga elétrica.

O controle genético mais freqüente de locos isoenzimáticos em plantas é a segregação monogênica. Tanto o controle genético quanto o número de subunidades da maioria das isoenzimas de plantas já são conhecidos (Caixeta et al. 2006).

Embora em número limitado, vários locos isoenzimáticos podem ser analisados, o que pode atingir uma quantidade entre 20 a 50 locos no genoma da espécie (Ferreira & Grattapaglia, 1998). Embora sejam marcadores seletivamente neutros, estas enzimas possuem funções metabólicas e, conseqüentemente, os locos exibem um baixo nível de polimorfismo. Adicionalmente, tem-se como desvantagem a especificidade enzimática em certos tecidos vegetais, a influencia ambiental na atividade enzimática, o estádio de desenvolvimento da planta na expressão de determinadas enzimas, modificações pós-traducionais (formas múltiplas do produto de um único gene), além da complexidade apresentada por alguns zimogramas, dificultando a interpretação dos mesmos.

Os estudos clássicos de genética de populações foram os precussores na difusão da técnica de análise isoenzimática. Por ser de baixo custo de implementação e operacionalidade, mesmo na era dos marcadores moleculares, sua aplicabilidade se destaca em várias áreas da genética e do melhoramento de plantas, como caracterização da diversidade genética de populações naturais (Zheng & Ennos, 1999; Gao & Hong, 2000; Melo Júnior et al., 2004) e populações derivadas de sistema de cruzamentos controlados (Zheng & Ennos, 1999; Gimenes & Lopes, 2000); estudos evolutivos e análises filogenéticas (Jaaska, 2005) entre outras aplicações.

### 2.2.2. RFLP

Este foi o primeiro tipo de marcador molecular capaz de detectar as variações nas seqüências de nucleotídeos do DNA. Entretanto, estes marcadores surgiram após a descoberta das enzimas de restrição (Linn & Arber, 1968; Meselson & Yuan, 1968), capacitadas a clivar uma molécula de DNA (de fita dupla) em locais específicos, conhecidos como sítios de restrição, formados por quatro a oito pares de bases ao longo da molécula de DNA. O surgimento e primeiro estudo com a técnica de RFLP foi realizado por Grodzicker et al. (1974), em um experimento destinado à detecção de mutações em DNA de vírus.

O polimorfismo entre genótipos diferentes, ou seqüências nucleotídicas nas fitas de DNA, é evidenciado pela fragmentação do DNA por meio de enzimas de restrição e visualizado por hibridização destes fragmentos com seqüências homólogas de DNA marcadas com radioatividade ou compostos que desencadeiam uma reação de luminescência (Ferreira & Grattapaglia, 1998). Em outras palavras, o polimorfismo gerado na técnica de RFLP provém da formação ou supressão de sítios de restrição, decorrentes de substituições ou modificações na seqüência de nucleotídeos do DNA (eventos mutacionais), ocasionados por efeitos de deleções, inserções, inversões ou translocações, provocando mudanças na distância entre dois sítios de restrição vizinhos.

Nos marcadores RFLP não há efeitos pleiotrópicos, epistáticos ou interferências ambientais. Apresentam um nível de polimorfismo genético muito maior que as isoenzimas. A planta pode encontrar-se em qualquer estádio de desenvolvimento e não existe restrições quanto ao tecido da planta utilizado para extração do DNA. Além disso, são amplamente distribuídos no genoma da espécie estudada, pois as sondas utilizadas podem ser obtidas de regiões transcritas (cDNA), que são normalmente regiões de copias únicas, codificantes e mais informativas, ou de DNA genômico clonado ao acaso (Caixeta et al. 2006).

Um aspecto limitante ao uso da técnica de RFLP é a ausência de bibliotecas para espécies de menor relevância e menos pesquisadas no âmbito da biologia molecular. O custo desta técnica pode ser reduzido, uma

vez que as membranas de hibridização podem ser reutilizadas e os vários tipos de enzimas de restrição combinados a diversos tipos de sondas podem gerar inúmeros marcadores RFLP. Em contra partida, o desenvolvimento de marcadores RFLP envolve várias etapas de execução, o que impede sua automação e obtenção de grande volume de dados. Além disso, é um procedimento caro e trabalhoso, pois requer equipamentos em boas condições, investimento em profissionais qualificados e em produtos químicos.

A literatura tem relatado a utilização dos RFLP em estudos de: i) avaliação da relação de parentesco entre acessos (Bernardo et al., 2000); ii) mapeamento genético e detecção de QTL (La Rosa et al., 2003; Waldron et al., 1999); iii) associação entre diversidade genética e geográfica (Barcelos et al., 2002); filogenia e evolução de espécies (Ando et al., 2005), entre outros.

### 2.2.3. SSR ou Microssatélites

Os marcadores oriundos de seqüências simples repetidas (SSR – *simple sequence repeats* ou STR – *short tandem repeats*), conhecidos como microssatélites, consistem em pequenas seqüências com um a seis nucleotideos de comprimento, repetidas em tandem. Em genomas eucariotos, estas seqüências são mais comuns, melhor distribuídas ao acaso e formam locos mais polimórficos.

Regiões contendo microssatélites são amplificadas individualmente através de PCR utilizando-se um par de *primers* específicos com 20 a 30 nucleotídeos complementar as seqüências únicas que flanqueiam o microssatélite. Cada segmento amplificado possui tamanho distinto, composto por várias dezenas até algumas centenas de pares de nucleotídeos, representando assim um alelo diferente do mesmo loco.

Vários *primers* já foram desenvolvidos para amplificar segmentos de DNA específicos e direcionados a variados tipos de estudos genéticos. Geralmente, em cada reação de PCR um par de *primers* é utilizado e representa apenas um loco de microssatélites. No entanto, mais do que um

par de *primers* específicos pode ser utilizado simultaneamente na mesma reação de PCR, método este de genotipagem denominado *multiplex*.

Alguns estudos (Cardle et al. 2000; Morgante et al. 2002) mostram que os microssatélites são amplamente distribuídos no genoma das plantas superiores e com freqüência de distribuição variada quando se considera a infinidade de classes de microssatélites. Isto engloba todas as combinações de um a seis nucleotídeos somadas a variação do número de unidades repetitivas que caracterizam um (alelo) microssatélite

Segundo McCouch et al. (1997) quanto maior o número de nucleotídeos na unidade repetitiva do microssatélite, menor é a sua freqüência dentro do genoma. A variação da classe e unidade repetitiva dos microssatélites também depende da região em que se encontram. Em regiões de seqüências expressas (EST), alguns estudos mostraram que há uma predominância de trinucleotídeos (Cardle et al., 2000; Cordeiro et al., 2001; Morgante et al., 2002; Gao et al., 2003), enquanto que em bibliotecas genômicas foi encontrada maior proporção de clones contendo repetições de dinucleotídeos, como em aveia (Pal et al., 2002) e trigo (Bryan et al., 1997). A hipervariabilidade observada em locos microssatélites é gerada por um mecanismo específico freqüente: repetições de *slippage*, que ocorrem em maior freqüência do que eventos mutacionais ou inserções e deleções nucleotídica, geradoras do polimorfismo em RAPD e AFLP (Powell et al., 1996).

Alguns autores sugeriram polimorfismo em locos de microssatélites na seqüência que os flanqueia, tanto em animais (Macaubaus et al., 1997) quanto em espécies vegetais (Caixeta et al., 2005). Significa dizer que o polimorfismo revelado por alguns pares de *primers*, desenvolvidos para uma espécie, mostraram que a diferença do número de nucleotídeos entre alelos não era originada da multiplicidade do número de nucleotídeos da unidade repetitiva. Esse fenômeno poderia ser explicado por *indels* – inserções e deleções – de um único nucleotídeo ou mesmo de um fragmento longo de DNA nas regiões flaqueadoras dos microssatélites.

Os locos de microssatélites são altamente multialélicos, o que lhes garante o mais elevado conteúdo de informação de polimorfismo quando comparado com outros marcadores (Bryan et al., 1997). A conservação de

sítios de microssatélites entre espécies relacionadas, possibilita, em certos casos, a transferência de marcadores entre espécies usando *primers* heterólogos, derivados de espécies do mesmo gênero ou gêneros afins (Cordeiro et al, 2001; Gao et al., 2003).

Todas estas particularidades reunidas fazem com que os marcadores microssatélites sejam aplicáveis no mapeamento genômico (La Rosa et al., 2003; Okogbenin et al., 2006); detecção de QTL (Prasad et al., 2003); na conservação de germoplasma e estudos de diversidade genética e geográfica (Belaj et al., 2003); estrutura e variabilidade genética (Zucchi et al., 2003); sintenia (Marques, et al., 2002); análise de parentesco (Bernardo et al., 2000); estudos evolutivos entre espécies correlatas (Santacruz-Varela et al., 2004), entre outros.

A grande limitação da tecnologia de microssatélites refere-se à grande quantidade de trabalho envolvida para o desenvolvimento de *primers*, exigindo profissionais qualificados e equipamento sofisticado para o sequenciamento automatizado, além do elevado custo.

## 2.3. Aplicativos computacionais e a simulação de dados genéticos

Na genética são enfatizados temas relativos a evolução, à hereditariedade e ao melhoramento das espécies. Estudos genéticos em geral envolvem grande conjunto de dados, compostos por variáveis (características fenotípicas, multicategóricas, marcadores moleculares e/ou geográficas) e unidades taxonômicas (espécies, populações, cultivares, famílias, indivíduos etc). A complexidade de um estudo genético aumenta quando informações adicionais sobre a espécie são requeridas, além de o pesquisador ter que saber quais as metodologias de análise genética estão disponíveis e se mostram mais adequadas. Também é comum a adoção de modelos matemáticos capazes de caracterizar o fenômeno biológico e, assim, prover estimativas de vários parâmetros genéticos. Diante deste cenário, o uso da informática passou a ser ferramenta imprescindível ao processamento e gerenciamento de dados. O computador passou a ser

equipamento presente em todos os laboratórios de pesquisa do mundo. Por meio do processamento apropriado, os parâmetros são estimados e os fenômenos biológicos interpretados. Portanto, na etapa de análise e interpretação de resultados é fundamental a existência de recursos computacionais e aplicativos eficientes à disposição do pesquisador, o que pode ser comprovado pelas publicações nos mais variados periódicos, em que numa imensa maioria há referências sobre o programa computacional utilizado.

Os aplicativos computacionais, tanto os desenvolvidos em nosso país quanto no exterior, permitem a execução de muitas análises que auxiliam de forma considerável o pesquisador por permitir uma condensação de seus dados sem perda de informações (Cruz, 2001). Atualmente uma ampla variedade de *softwares* estatísticos está disponibilizada, porém nem sempre apresentam análises que satisfaçam a modelos genéticos, biométricos e evolutivos, necessários para realização dos estudos.

Criar um programa computacional não é uma das tarefas mais fáceis. Ressalta-se que na área da genética, processos biológicos complexos estão envolvidos e associados a distribuições probabilísticas, de modo que os profissionais que venham a criar um novo aplicativo devem agregar conhecimentos nas áreas específicas e da informática. Outra particularidade nesta área é que, corriqueiramente, recorre-se a processos de aleatorização, reamostragem e permutações para a estimação e realização de testes de hipótese sob os mais variados parâmetros genéticos, bem como definição de intervalos de confiança e distribuições empíricas, o que exige melhor qualidade, velocidade de processamento e operacionalidade do aplicativo a ser desenvolvido.

De acordo com Cruz (2001), programas destinados a Genética devem ser desenvolvidos atendendo a finalidade básica de análise e processamento de dados, com base em modelos adequados. A interface entre usuário e máquina deve ser de fácil manuseio, objetiva e abrangente, atendendo as exigências do usuário. Adicionalmente, sua compatibilidade com sistemas operacionais triviais, e a característica amigável com planilhas eletrônicas e editores de texto mais utilizados, são fundamentais na difusão e adoção do aplicativo na comunidade científica.

Atualmente não se justifica o uso restrito de determinadas metodologias por modismo. A diversificação de técnicas de análise segue paralelamente a existência de programas computacionais. Um exemplo evidente é demonstrado pelas técnicas multivariadas, que permitem a interpretação simultânea de inúmeras variáveis.

Alguns programas eficientes e aplicáveis à área de Genética, em especial à Genética de Populações, encontram-se disponibilizados, muitos deles distribuídos gratuitamente, em geral, disponíveis na *internet*. Para se ter uma idéia da quantidade de programas existentes com este enfoque, estão nos endereços eletrônicos http://www.biology.lsu.edu/general/software.html e http://www.nslij-genetics.org/soft/ mais de 30 aplicativos computacionais dirigidos à análise de dados genéticos.

Numa linha de programas gratuitos, destaca-se o Arlequin versão 3.1 (Excoffier et al., 2005), em http://cmpg.unibe.ch/software/arlequin3/; GDA (Genetic Data Analysis) versão 1.1 (Lewis & Zaykin, 2002), em http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php/; GENEPOP versão interativa disponibilizada na *internet* (Raymond & Rousset, 1995), em http://genepop.curtin.edu.au/; GENES versão 2007.0.0 (CRUZ, 2006a) em http://www.ufv.br/dbg/genes/gdown.htm; POPGENE versão 1.32 (Yeh et al., 1997), em http://www.ualberta.ca/~fyeh/download.htm/; PowerMarker versão 3.25 (Liu & Muse, 2005), em http://statgen.ncsu.edu/powermarker/downloads.htm/; e TFPGA (Tool for Population Genetic Analyses) versão 1.3 (Miller, 1997), em http://www.marksgeneticsoftware.net/tfpga.htm/.

Os programas Arlequin, GDA, GENEPOP, POPGENE e TFPGA, são de fácil acesso, possuem técnicas sofisticadas implementadas, poderosas ferramentas estatísticas e de fácil manuseio, fazendo deles uma alternativa atraente para realização de análises biométricas nos estudos genéticos. Embora programas computacionais apresentem sobreposições de análises, cada programa tem características únicas a serem oferecidas aos usuários (Labate, 2000).

O programa Arlequin versão 3.1 (Excoffier et al., 2005) proporciona aos usuários inferir sobre a genética de populações a partir de várias

metodologias básicas e testes estatísticos, de modo que informações genéticas e características demográficas de uma coleção de populações amostradas possam ser obtidas. É um programa com várias opções de análise, o que significa que o usuário pode gastar um tempo até aprender a manuseá-lo por completo. A análise de dados é dividida em duas principais categorias: métodos intrapopulacionais (medidas descritivas; equilíbrio de Hardy-Weinberg; desequilíbrio de ligação, etc) e interpopulacionais (análise molecular de variância, distâncias genéticas etc).

O programa Arlequin pode avaliar vários tipos de dados, sejam eles no formato haplotípico ou genotípico (diplóides). Os tipos básicos de dados são seqüências de DNA, RFLP (dados binários, presença e ausência da banda), microssatélite (número de repetições de *motif*, ou seja, comprimento alélico), padrões e freqüências alélicas. Dados padrões são aqueles nos quais as bases moleculares do polimorfismo não são particularmente definidas, ou quando diferentes alelos são considerados como mutacionalmente eqüidistantes um dos outros no mesmo loco (codificação não ordinal comumente usada em dados codominantes). O formato genotípico deve ser caracterizado como de fase gamética conhecida ou desconhecida. Se a fase gamética for conhecida, o genótipo pode ser considerado como a união de dois haplótipos bem definidos. Caso contrário, considera-se os alelos presentes em cada loco como codominantes, no qual pode se definir se existe ou não um alelo recessivo. A introdução do arquivo de dados é feita via “arquivos de projeto” (“*project files*”). O usuário pode criar um projeto num editor de texto qualquer desde que defina qual editor irá utilizar, ou usar a opção “*project wizard*”, no qual é aberta uma janela em que são definidos os elementos essenciais do projeto, como o tipo de dado (genotípico ou haplótipo, número de populações etc). O programa é capaz de exportar e importar arquivos de dados no próprio formato Arlequin, GENEPOP (Raymond & Rousset, 1995), BIOSYS (Swofford & Selander, 1981), PHYLIP (Felsenstein, 1993), MEGA versão 1.0 (Kumar et al., 1994) e WinAmova (Excoffier et al., 1992). Não existem opções interativas de exclusão ou inclusão de locos ou populações a serem analisadas. A saída dos resultados se dá no formato HTML. O manual possui 145 páginas que descrevem detalhadamente as metodologias e apresenta os referenciais

teóricos. Exemplos de utilização do programa Arlequin são vistos em testes de equilíbrio de Hardy-Weinberg e desequilíbrio de ligação em populações de milho (Reif et al., 2004) e na execução da análise de variância molecular (AMOVA) em trigo (Dreisigacker et al., 2004).

O aplicativo computacional GDA versão 1.1 (Lewis & Zaykin, 2002) tem suas análises fundamentadas no livro de Bruce S. Weir *"Genetic Data Analysis"* (Weir, 1996). Analisa também dados haplotípicos e diplóides codominantes. O arquivo de dados deve estar no formato Nexus (Maddison et al., 1997). Este formato facilita a análise de subconjuntos de dados (*"blocks"* do arquivo NEXUS) sem alterar o arquivo original. É lida uma seqüência de instruções que definem qual arquivo deve ser lido, qual análise a ser executada e onde armazenar os resultados. Permite a inclusão e exclusão de locos e, ou, populações a serem analisadas. As análises são agrupadas em quatro categorias: estatísticas descritivas, estatísticas F, distâncias e desequilíbrio. Seu manual, com 45 páginas, explica como usar o programa e definir comandos, mas contém poucos detalhes e referências acerca dos métodos. Também é possível exportar e importar dados dos programas BIOSYS (Swofford & Selander, 1981) e GeneStrut (Constantine et al., 1994). Exporta o próprio formato NEXUS, o GeneStat-PC (Lewis & Whitkus, 1989) e SAS (SAS, 1989). O GDA possui uma interface simples, mas grande poder estatístico, satisfazendo grande parte das exigências sobre a análise de marcadores codominantes (Labate, 2000). Exemplos de utilização do GDA são vistos nas análises de diversidade genética e estrutura de população de coco (*Cocos nucifera L.)* com a utilização de distâncias genéticas (Meerow et al., 2003); realização de análise de variância de freqüências alélicas em populações naturais de pequizeiro (Melo Júnior, et al., 2004) e construção de árvores filogenéticas em populações de milho pipoca (Santacruz-Varela et al. 2004).

O GENEPOP versão 3.4 (Raymond & Rousset, 1995) é mais um *software* de análise genética de populações. Os dados são analisados no formato de arquivos texto e o programa é executado no sistema DOS. Existe uma versão *on line* disponibilizada em http://genepop.curtin.edu.au/. Entretanto, a versão DOS é atualizada periodicamente, mais do que a versão da *internet*, e pode conter algumas opções não disponíveis na *web*.

Dados haplotípicos e diplóide codominantes são passíveis de avaliação. Na versão *on line* a saída dos resultados se dá no formato HTML e, ou, via *email*, opções estas definidas pelo usuário. Dentre as seções de análises, destaca-se: teste exato do equilíbrio de Hardy-Weinberg; desequilíbrio de ligação; informações básicas; estatísticas $F_{ST}$ e outras correlacionadas. Os dados podem ser exportados para os formatos do FSTAT (Goudet, 1995), BIOSYS (Swofford & Selander, 1981), LINKDOS (Garnier-Gere & Dillmann, 1992) e Arlequin. O manual possui 36 páginas, com os procedimentos bem detalhados e descrição dos métodos. Testes do modelo de isolamento por distância em populações naturais de *Cryptocarya Aschersoniana* Mez (Moraes & Derbyshire, 2002) e investigação do equilíbrio de Hardy-Weinberg e desequilíbrio genotípico em pinhos (*Pinus caribaea* Morelet) (Zheng & Ennos, 1999) são alguns exemplos da utilização do GENEPOP.

O programa GENES é um pacote genético-estatístico capacitado a realizar inúmeros tipos de análises biométricas, tanto para características fenotípicas, morfológicas e de marcadores moleculares, que têm servido à identificação de genótipos superiores nos programas de melhoramento e caracterização de populações. As análises se dividem nas grandes seções: estatística; análise multivariada, biometria e diversidade genética. Atualmente tem-se direcionado esforços na composição e aperfeiçoamento do *software* em análises biométricas para o estudo da diversidade genética de unidades taxonômicas, o que inclui inferências sobre a estruturação genética das mesmas. Em relação aos dados de marcadores moleculares, avalia-se dados do tipo binário ou genotípico (os alelos de um loco recebem codificação numérica não ordinal). O arquivo de dados a ser analisado pode ser delineado no próprio editor do programa ou em editores de texto e planilha Excel (*Microsoft Office*). As extensões *.prn, *.dat, *.txt e *.xls são lidas pelo aplicativo. As técnicas de diversidade genética disponibilizadas são destinadas ao estudo intrapopulacional (estatísticas descritiva, equilíbrio de Hardy-Weinberg, desequilíbrio gamético) e interpopulacional (medidas de distâncias, técnicas multivariadas, análise de variâncias para dados de marcadores, etc). A saída dos resultados é fornecida num editor próprio, mas de dimensões limitadas. Os resultados podem ser exportados para o editor de texto Word e a planilha Excel. Há um conversor de dados que

permite exportá-los nos formatos do Arlequin, NEXUS, POPGENE e TFPGA. Recentemente, foram publicados três manuais (Cruz, 2006a; 2006b; 2006c), porém não há uma abordagem específica para a seção diversidade genética. Aplicações do GENES são vistas em estudos de diversidade para diversas espécies, em que índices de dissimilaridade e técnicas de agrupamento são utilizadas (Faleiro et al., 2004; Bertini et al., 2005)

O POPGENE versão 1.32 é mais um aplicativo destinado à análise da variação genética entre e dentro de populações naturais usando marcadores codominantes e dominantes, oriundos de dados haplotípicos ou diplóides. O arquivo de dados é constituído em um editor de texto qualquer. As análises estão dispostas em um menu simples, que permite a execução de inúmeras metodologias de uma única vez. Podem ser realizadas análises do tipo estimação de freqüências alélicas, diversidade genética, distâncias genéticas, estatísticas F, estrutura multiloco etc, considerando locos e populações individualmente ou conjuntamente. Antes de se realizar as análises é possível incluir e excluir locos e populações. A saída dos resultados pode ser salva em um arquivo texto ou copiados e colados em outros editores. O manual apresenta 28 páginas, mas pouco detalhado e a ajuda *on line* não está disponível. Exemplos de aplicação são vistos na obtenção de freqüências alélicas em locos isoenzimáticos e microssatélites em populações de milho pipoca (Santacruz-Varela et al. 2004) e no uso de técnicas de agrupamento (Barcelos et al., 2002).

PowerMarker versão 3.25 (Liu & Muse, 2005) apresenta uma boa interface entre métodos estatísticos novos e tradicionais para a análise genética de populações. Antes de sua instalação é requerida a instalação do *Microsoft. NET framework* e a existência da planilha eletrônica do Excel (versão 2000 ou superior). Trabalha com dados haplotípicos e diplóides, cuja fase gamética pode ser definida como conhecida ou desconhecida. Informações de marcadores microssatélites, polimorfismo de nucleotídeos (SNP) e RFLP são exemplos de dados passíveis de análise. A sua interface gráfica se assemelha a do *Windows Explorer*, em que pastas contêm os arquivos e é permitido criar novas pastas. Assim como no Arlequin é necessário criar um projeto para organização dos dados e pastas. O conjunto de dados a ser analisado deve estar digitado em um editor de texto,

com definições das colunas dos descritores (grupos, populações e, ou, indivíduos) e dos locos. Defini-se no programa as colunas que são tidas como descritores e as colunas que são representadas pelos marcadores. É possível importar dados de freqüência, distância em formatos específicos, além de tabelas e textos. Os dados podem ser exportados nos formatos originais, de tabelas, NEXUS e do Arlequin. Possibilita ainda a exclusão, inclusão e partição do conjunto de dados. Realiza análises descritivas (heterozigosidades, nível de polimorfismo, estimação de freqüências alélicas, índice de fixação, testes de desequilíbrio, etc); de estrutura (estatísticas F) e filogenia (medidas de distâncias e construção de árvores). Seu manual possui 33 páginas com descrição detalhada sobre a manipulação dos dados e suporte teórico sobre as metodologias, com os respectivos referenciais teóricos. Aplicações práticas são vistas no cálculo de estatísticas descritivas de populações (Liu et al., 2003; Fukunaga et al., 2005), distâncias genéticas e construção de árvores filogenéticas (Liu et al., 2003); teste de desequilíbrio (Garris et al., 2005) e definição de coleções núcleo (Fukunaga et al., 2005)

O TFPGA versão 1.3 (Miller, 1997) analisa tanto dados haplotípicos quanto diplóides, dominantes ou codominantes. Os dados sob análise podem ser adicionados no formato de arquivo texto nos moldes do TFPGA. Os resultados são liberados no formato texto, podendo ser modificados e, ou, salvos. Seu menu inclui itens de análise como estatísticas descritivas, estatísticas F, distâncias genéticas, construção de dendrograma (UPGMA) e teste de equilíbrio de Hardy-Weinberg. O manual de 30 páginas inclui descrições das metodologias disponíveis, comentários sobre erros comuns do programa, além de sugestão de literatura. Alguns exemplos práticos do uso do TFPGA são vistos no cálculo dos índices de fixação em populações estruturadas (Wadt & Kageyama, 2004) e teste de equilíbrio de Hardy-Weinberg (Moraes & Derbyshire, 2002).

Todos estes programas podem ser executados no sistema operacional Windows XP e 128 MB RAM permite executá-los sem maiores transtornos.

Outra grande contribuição da informática é viabilizar estudos de fenômenos, via simulação de situações mais complexas, em que são

estabelecidas pressuposições e definidos parâmetros, de tal forma que o efeito de certos fatores sejam controláveis e possam ser convenientemente estudados (Cruz, 2001). A simulação tem sido definida como o modo de reproduzir, por meio de recursos computacionais, o comportamento de um sistema real, para estudar seu funcionamento em condições alternativas (Dachs, 1988), envolvendo modelos lógicos, que permitam descrever adequadamente o sistema natural (Naylor et al., 1971). Portanto a modelagem é outro aspecto fundamental na simulação. Um modelo deve ser simples o suficiente para ser funcional e interpretável, mas de desempenho comparável ao modelo real.

A importância da simulação ganhou grande espaço no âmbito cientifico. Impulsionada pela informática, a simulação não se limita apenas a modelos que representam a entidade a ser investigada e sim a uma metodologia para avaliação destes. Técnicas de reamostragem e permutação e alguns algoritmos são em essência processos de simulação que em geral visam a estimação de parâmetros e definições de distribuições. Em estudos de genética de populações com o uso de marcadores, freqüentemente utiliza-se de métodos de reamostragem para a estimação de parâmetros genéticos populacionais e seus respectivos desvios-padrão (Carlini-Garcia et al., 2003). Exemplos práticos e atuais da importância da simulação nos estudos genéticos, sob vários contextos são vistos em:

a) estudos visando encontrar uma solução ou valor ótimo;

A grande variação no número de marcadores moleculares utilizados em estudos de diversidade genética levou alguns pesquisadores a tentar desvendar um número ideal ou ótimo em relação a esta quantidade. Pouco se sabe a respeito do número de marcas e de indivíduos necessários para se predizer com acurácia a divergência genética entre e dentro de populações ou indivíduos. Dias et al. (2004) revisaram 139 estudos sobre diversidade genética e verificaram que, em média, o número de marcas utilizados é de 160, 281 e 25 para RAPD, RFLP, e locos SSR, respectivamente. Os autores consideram este número médio de marcas pequeno para se obter análises acuradas.

Alguns estudos de simulação presentes na literatura com este objetivo, basicamente foram realizados com processo de reamostragem

*boostrap* (Efron & Tibshirani 1993). Visando obter a variância amostral e o número ótimo de marcas RFLP, Tivang et al. (1994) observaram que 284 a 377 bandas foram necessárias para estimar a distância genética entre 37 linhagens endogâmicas de milho associadas a um coeficiente de variação fixo de 10%, independente da enzima de restrição usada. Pejic et al. (1998), estudando a similaridade genética de 33 linhagens endogâmicas de milho, observaram que acima de 150 bandas (seja RFLP, RAPD, AFLP e SSR) houve diminuta resposta no ganho em precisão pela adição de novas bandas. Fanizza et al. (1999) avaliaram 10 acessos de *Vitis vinifera* e verificaram que o agrupamento formado com 400 marcas não apresentou distorção quando comparado ao agrupamento formado com todas as marcas (932 bandas). Picoli et al. (2004) obtiveram resultados similares em um estudo com 84 genótipos de *Eucalyptus* sp. Este valor (≅ 400 bandas) pode ser tomado como ponto de referência a outros estudos de diversidade, embora possa ser limitante para outros conjuntos gênicos ou espécies de plantas (Dias et al., 2004).

Moraes (2003) verificou que 44 pares de *primers* de microssatélites foram requeridos para se obter valores de correlação de 95% e estresse de 6,44%, em relação a projeção gráfica, comparado à amostra padrão de 57 pares de *primers*. Já Tardin et al. (2003), utilizando apenas 55 marcas polimórficas de RAPD, concluíram que 50 seria um número satisfatório para se estudar a diversidade genética entre acessos de alface (L*actuca sativa* L.).

b) eficiência de metodologias

Alguns trabalhos compararam a eficiência dos coeficientes de similaridade em expressar o grau de divergência genética de espécies vegetais a partir de marcadores RAPD e AFLP (Duarte et al. 1999; Emygdio, 2003; Meyer, 2004). Inúmeras simulações foram feitas no conjunto original de dados e obtidas novas estimativas de distância, consequentemente, novos agrupamentos foram gerados. Estes autores constataram que o número de grupos formados se altera conforme o coeficiente de similaridade utilizado. Já Laval et al. (2002) avaliaram o comportamento das principais distâncias genéticas encontradas na literatura sob diferentes modelos de

mutação, direcionadas a espécies de animais, com dados de marcadores microssatélites. Com estudo simulado os autores concluíram que a distância genética de Reynolds et al. (1983) foi a melhor.

Nei et al. (1983) examinaram a acurácia e eficiência de três diferentes métodos de reconstrução de árvores filogenéticas e diferentes medidas de distância genética, a partir de dados de freqüência alélica, por intermédio de simulação. O processo de simulação permitiu aos autores chegar a conclusões satisfatórias e generalizadas para dados de diferenças nucleotídicas. Takezaki & Nei (1996), também através de simulações com dados de locos microssatélites, verificaram que a distância padronizada de Nei (Nei, 1972) e a de Goldstein (Goldstein, et al., 1995) foram as mais apropriadas para estimar o tempo de divergência evolutiva.

c) Testes, obtenção de distribuições e estimação de parâmetros genéticos

Quando populações naturais são estudadas, dados com repetições não são disponibilizados, como em ensaios experimentais, de maneira que populações e indivíduos são amostrados sobre as condições ecogeográficas da espécie. Assim erros nas estimativas dos parâmetros genéticos não podem ser obtidos como ocorre usualmente na experimentação. O desenvolvimento dos computadores em tempos recentes permitiu que métodos de reamostragem como *jackknife* (Efron & Tibshirani 1993) e *bootstrap* sejam freqüentemente aplicados para a estimação de tais erros (Carlini-Garcia et al., 2006). Numa seqüência de trabalhos, Carlini-Garcia e colaboradores (Carlini-Garcia et al., 2001; 2003; 2006) aplicaram o método de reamostragem *bootstrap* para inferir a respeito da adequação da amostragem sobre locos, indivíduos, populações e, indivíduos e populações concomitantemente, em dados de populações naturais reais e simuladas e, consequentemente, seu reflexo na estimação dos parâmetros de índice de fixação total (F ou $F_{IT}$), índice de fixação intrapopulacional (*f* ou $F_{IS}$), a divergência interpopulacional (θ ou $F_{ST}$) e taxa aparente de cruzamento ($t_a$), bem como nas suas respectivas variâncias, erros associados, distribuição empírica e intervalos de confiança. Os autores ainda destacaram a importância do processo de simulação tanto na estimação de parâmetros

quanto na elucidação de sistemas biológicos ocorrentes nas populações naturais. Estratégias de amostragem servem como guia para investigações desta natureza, onde não existe conhecimento prévio sobre a estrutura genética e o sistema de acasalamento das populações.

Em um outro enfoque, Guo & Thompson (1992) desenvolveram uma versão permutada do teste exato de Fisher que tem servido aos testes de hipótese de nulidade para o equilíbrio de Hardy-Weinberg e testes de excesso e deficiência de heterozigotos para locos e populações individualmente e conjuntamente. Esta versão baseia-se no método (algoritmo) de Monte Carlo e da cadeia de Markov, também consideradas como técnicas de simulação, cujo princípio é considerar todas as categorias de problemas ou de sistemas que têm base probabilística ou estocástica.

Obviamente que o estudo de simulação somente se justifica se soluções analíticas não existem ou o grau de dificuldades e o número de variáveis envolvidas não permitem a realização de inferências adequadas sobre o problema. O uso da simulação deve ser encarado pelos geneticistas e melhoristas como uma alternativa para a resolução de seus problemas, detecção e comprovação de técnicas estatísticas mais eficientes, comparações metodológicas entre outras (Ferreira, 2001).

A área de simulação em genética permitiu a criação de novos métodos e, consequentemente, *softwares*, com simuladores específicos e capazes de realizar estudos didáticos e pesquisas, a exemplo do programa GENES (Cruz, 2006a), PowerMarker (Liu & Muse, 2005) e GDA (Lewis & Zaykin, 2002).

### 2.4. Análises biométricas no estudo da diversidade genética

Para que o geneticista ou o melhorista possam inferir sobre a variabilidade genética das populações diversas metodologias de análise de dados genéticos foram desenvolvidas ao longo dos anos. Trata-se de um processo contínuo e que caminha juntamente com os avanços na área da biologia molecular e da informática. Elas podem ser caracterizadas pelo

nível de inferência a que se destinam, ou seja, dentro e ,ou, entre as unidades taxonômicas.

É comum a utilização de estatísticas descritivas nos estudos de populações, pois elas dão uma idéia inicial de como se encontra o polimorfismo genético de uma população. Medidas como: número de alelos por loco ou total; número de alelos efetivos, raros; exclusivos; polimórficos; conteúdo de polimorfismo, heterozigosidade observada e esperada, coeficiente de fixação/endogamia (*f*), riqueza genotípica (índice Shannon-Wiener), entre outras, são alguns dos descritores que ajudam a caracterizar uma população.

Não menos importante é o entendimento da relação entre o tamanho efetivo ($N_e$) e o tamanho real de uma população de plantas. Define-se $N_e$ como o tamanho de uma população idealizada - não sofre com os efeitos da mutação, seleção e migração – que tem a mesma quantidade de deriva genética atuante nas freqüências alélicas ou a mesma taxa de decréscimo de heterozigotos que a população sob estudo (Falconer, 1987; Vencovsky & Crossa, 1999). O tamanho efetivo pode ser tomado como uma medida da representatividade genética de uma amostra de indivíduos e é um importante parâmetro em genética quantitativa e de populações, pois auxilia na mensuração da taxa de deriva genética e endogamia (exemplos em Wang, 1997; Santiago & Caballero, 1998).

O teorema de Hardy-Weinberg é importante conceitualmente, historicamente, na pesquisa aplicada e em trabalhos com modelos teóricos (Ridley, 2006). Ele se fundamenta no princípio de que para um loco, uma população panmítica que não sofre influências de forças evolutivas (seleção, migração e mutação) atinge o equilíbrio após uma geração de acasalamento ao acaso, de maneira que a relação genotípica torna-se igual ao produto das freqüências alélicas e permanece inalterada com as sucessivas gerações de acasalamento ao acaso. A questão para múltiplos locos conduz a um outro conceito importante, chamado equilíbrio gamético e, ou, de ligação. Se forem comparadas as freqüências genotípicas de uma população real com as relações de Hardy-Weinberg, caso elas se desviem, isso sugere que seleção ou ausência de cruzamentos aleatórios possa estar acontecendo, sendo que explorar a relação entre freqüências para um conjunto de alelos,

dentro ou entre locos é fundamental. A questão geral é se a freqüência de um conjunto de alelos é a mesma que o produto das freqüências alélicas separadas (Weir, 1996). As diferenças entre freqüências conjuntas e o produto de freqüências individuais são chamadas de coeficientes de desequilíbrio.

Várias estratégias para testar as hipóteses formuladas sobre as proporções esperadas no equilíbrio de Hardy-Weinberg ou gamético estão presentes na literatura, como o tradicional teste de qui-quadrado ou testes de razões de verossimilhança e, ainda, o teste exato de Fisher (Weir, 1996). A aplicação destes testes, embora com um mesmo propósito, não deve ser usada indiscriminadamente nos testes de equilíbrio, pois as informações e codificações do conjunto de dados direcionam seu uso. Rousset & Raymond (1995) apresentam testes que baseados em hipóteses alternativas para equilíbrio de Hardy-Weinberg, testam o excesso e a deficiência de heterozigotos na população.

Grande parte dos estudos de diversidade baseiam-se em uma amostra aleatória de locos obtida em populações não estruturadas hierarquicamente (Dias, 1998). Assim, a análise interpopulacional, a partir de dados codominantes se dá por intermédio de medidas de distância, classificadas como geométricas, genéticas e genotípicas. As distâncias pertencentes ao primeiro grupo, não invocam qualquer pressuposição genética ou evolutiva. Ao contrário, nas distâncias genéticas modelos genético-evolutivos estão contextualizados. Sob uma filosofia admite-se que a população ancestral e as populações derivadas, encontram-se em equilíbrio entre mutação e deriva genética, o que implica dizer que a divergência entre as populações é devida ao aparecimento de novos mutantes dentro delas. Deste modo, as distâncias podem ser usadas do ponto de vista filogenético, como estimadores do tempo de divergência (Laval et al., 2002). As distâncias genéticas propostas por Nei e colaboradores (Nei 1972, 1973; Nei et al. 1983) seguem esta linha de raciocínio. Em uma outra linha encontram-se as distâncias genéticas de coancestralidade, a exemplo de Reynolds et al. (1983), cuja flutuação nas freqüências alélicas é atribuída exclusivamente à deriva genética.

Em muitas situações, o pesquisador está interessado na formação de grupos de populações similares. No entanto, se um número elevado de populações (g) for avaliado, muitas combinações, na ordem de $(g^2 - g)/2$, de pares de distâncias serão obtidas. Técnicas multivariadas de agrupamento facilitam a interpretação conjunta das populações, a exemplo dos métodos de otimização de Tocher e projeções gráficas, n-dimensionais ou dendrogramas (Sneath & Sokal, 1973; Cruz & Carneiro, 2003).

O conhecimento do padrão de variação genética entre e dentro de populações introduz o conceito de estrutura genética de populações (Dias, 1998). A ação de forças evolutivas, ou amostragem genética, resultarão em diferenciação intraespecífica, quantificada pelas estatísticas F, também conhecidas como índices de fixação de Wright (1951; 1978). Estas quantidades medem o grau de parentesco de vários pares de alelos. Pelas definições de Wright e, posteriormente, descritas por Cockerham (1969, 1973), em análises de variância de freqüências alélicas, na situação em que indivíduos diplóides são amostrados de uma série de populações, tem-se três medidas básicas:

$F_{IT}$ (ou F) – mede o desvio das freqüências genotípicas da população - conjunto de todas as subpopulações - em relação ao equilíbrio de Hardy-Weinberg. Esses desvios resultam de cruzamentos não ao acaso dentro da população (incluindo a endogamia em todos os níveis). F também é uma medida de correlação entre duas unidades gaméticas que formam um zigoto na população (ou do conjunto de subpopulações amostradas);

$F_{ST}$ (ou θ) – é o coeficiente de ancestria, representando a probabilidade de que dois indivíduos, pertencentes a subpopulações diferentes, possuem um alelo idêntico por descendência. θ é uma correlação de gametas dentro de subpopulações;

$F_{IS}$ (ou *f*) - mede a endogamia em nível de indivíduos, ou seja, mede a probabilidade de que os dois alelos de um loco presentes no mesmo indivíduo sejam idênticos por descendência. *f* é também uma medida de correlação de gametas devido à endogamia dentro das subpopulações.

A partir destes coeficientes Wright também estabeleceu a seguinte relação:

$(1-F_{IT}) = (1 - F_{IS})(1- F_{ST})$.

Estas estatísticas F foram definidas por Cockerham (1973), como: $F_{IS} = \frac{F - \theta}{1 - \theta}$; $F_{ST} = \frac{\theta - f}{1 - f}$; $F_{IT} = \frac{F - f}{1 - f}$, em que $F = F_{IT}$, $f = F_{IS}$ e $\theta = F_{ST}$, e I, S e T representam, indivíduos, subpopulações e população total, respectivamente. Weir & Cockerham (1984) propuseram fórmulas gerais para a estimação de F, θ e *f* para alelos múltiplos e procedimentos *jackknife* para estimação de suas variâncias.

A idéia básica das formulações de Wright e, posteriormente, seguida por Cockerham em seus trabalhos, assume que as populações sob investigação são derivadas de um ancestral comum num mesmo tempo e que todas as populações são igualmente relacionadas, existindo ou não migração entre elas. No entanto, Nei (1986) e Nei & Kumar (2000) argumentam que a aplicação desta estrutura aproxima-se a grupos de populações experimentais. Em populações naturais, este modelo quase nunca se aplica, pois estas possuem relações filogenéticas.

Nei (1977) mostra que as estatísticas F de Wright podem ser definidas como razões entre heterozigosidades, ou estatísticas H (definidas por Nei, 1973), ao invés de correlações entre unidades gaméticas. Estas definições, segundo Nei (1977), são independentes do número de alelos envolvidos, da atuação de forças evolutivas e o sistema reprodutivo da espécie. Neste caso caracterizou-se: $F_{IS} = \frac{H_S - H}{H_S}$; $F_{ST} = \frac{H_T - H_S}{H_T}$; $F_{IT} = \frac{H_T - H}{H_T}$, em que $H_T$, $H_S$ e H, correspondem à diversidade genética (ou heterozigosidade) na população total, entre subpopulações e dentro de subpopulações, respectivamente. Na verdade, $F_{ST}$ foi redefinida como a estatística $G_{ST}$, originalmente denominada de coeficiente de diferenciação gênica (Nei, 1973).

Posteriormente, Excoffier et al. (1992) desenvolveram uma análise hierarquizada diretamente da matriz de quadrados das distâncias Euclidianas, inicialmente aplicáveis a dados haplotípicos. A AMOVA (análise de variância molecular) gera estatísticas $\Phi$, análogas às estatísticas F. Tem sido preferida em muitas ocasiões, por apresentar flexibilidade em relação ao tipo de informações, podendo ser marcadores dominantes ou codominantes e dados de seqüência (Michalakis & Excoffier, 1996).

Exemplos práticos da aplicação da AMOVA em dados de marcadores codominantes são vistas nos trabalhos de Peakall et al. (1995) e Maguire et al. (2002). Alguns exemplos a seguir mostram como as análises biométricas têm sido aplicadas nos estudos de populações.

Análises da diversidade genética, estrutura e relações filogenéticas foram realizadas em 172 acessos do gênero *Zea* sp., em que estão incluídos o milho e o teosinto. Estes acessos eram representantes da distribuição geográfica do teosinto no nordeste do México e sudeste da Nicarágua, a partir de 93 locos microssatélites (Fukunaga et al., 2005). Os autores calcularam algumas medidas descritivas, incluindo número de alelos, heterozigosidade observada e heterozigosidade esperada (diversidade gênica) e número de alelos privados para as espécies, subespécies e raças. Observaram que a espécie *Zea mays* possui substancialmente maiores valores de heterozigosidade e diversidade gênica do que outras espécies diplóides do mesmo gênero. As subespécies *Z. mays* ssp. *mexicana* e *Z. mays* ssp. *parviglumis* apresentaram uma grande quantidade de alelos raros, que futuramente podem ser potenciais fontes de variabilidade genética.

Duzentas e sessenta linhagens de milho, representativas da diversidade genética entre todas as linhagens públicas de importância para o melhoramento da região temperada e para várias linhagens da região tropical e subtropical do mundo, tiveram seu polimorfismo investigado por 94 locos microssatélites (Liu et al., 2003). Os 2039 alelos identificados serviram a avaliação da diversidade e estrutura genética. Calcularam-se as heterozigosidades, número de alelos privados e número de alelos exclusivos por grupo específico. Também realizaram a análise AMOVA, da qual extraíram a estatística $\Phi_{ST}$ ($F_{ST}$ ou distância genética) entre pares de populações. Os autores ainda correlacionaram a matriz de distância $\Phi_{ST}$ com a matriz de distâncias obtida do coeficiente de parentesco de Malécot (Malécot, 1948), testando esta correlação pelo teste de Mantel (Mantel, 1967). Adicionalmente, teste de equilíbrio para um e dois locos foram realizados. Assim, os pesquisadores concluíram que linhagens endogâmicas tropicais e subtropicias possuem um maior número de alelos e maior diversidade genética do que as de clima temperado. As linhagens do grupo

(temperado) Stiff Stalk foram em média mais divergentes em relação aos outros grupos de linhagens endogâmicas.

Comparações da diversidade em amostras equivalentes de linhagens endogâmicas e raças locais (*landraces*) de polinização aberta revelaram que as linhagens capturam menos que 80% dos alelos das raças locais, sugerindo que *landrace*s podem providenciar adicional diversidade genética para o melhoramento de milho.

Análises isoenzimáticas (cinco aloenzimas) de amostras de semente, derivadas de populações naturais e manejadas, de *Pinus caribabea* vars 'bahamensis' e 'caribeae' foram usadas para acessar a estrutura genética de populações em seu espaço nativo e para detectar mudanças ocorridas durante a precoce domesticação da espécie (Zheng & Ennos, 1999). A diversidade genética de cada população foi acessada por intermédio do número de alelos por loco, porcentagem de locos polimórficos (critério a 95%), diversidade genética e heterozigosidade esperada. Dentro de cada população a estrutura genética de cada loco foi testada por meio do teste exato para desvios da proporção do equilíbrio de Hardy-Weinberg. A extensão e direção dos desvios do equilíbrio dentro das populações foram quantificadas pelas estimativas médias ponderadas de *f* ($F_{IS}$) sobre todos os locos. Os autores ainda analisaram o grau de diferenciação entre populações dentro das variedades estimando os valores de $F_{ST}$. O parentesco entre as populações foi estimado pela distância genética padronizada de Nei, não viesada, entre todos os pares de populações. Pelo menos um loco de cada população desviou do equilíbrio de Hardy-Weinberg e as medidas de diversidade para as populações manejadas foram similares às populações genitoras. Todos os valores de $F_{IS}$ foram positivos, indicando deficiência de heterozigotos. Os valores de $F_{ST}$ obtidos foram significativamente maiores que zero e, portanto, indicaram diferenciação genética entre as populações. O método de agrupamento UPGMA separou claramente as populações naturais nos dois grupos de variedades 'bahamensis' e 'caribeae'.

Em arroz, Garris et al. (2005) estudaram a estrutura e diversidade genética de 234 acessos representados nas dimensões geográficas de *Oriza sativa* L., a partir de 169 microssatélites. Os autores calcularam a distância

genética entre pares de acessos pela distância angular do comprimento da corda, que promove boa topologia de árvore filogenética considerando o modelo mutacional dos microssatélites. Medidas como número de alelos por loco, heterozigosidade esperada e conteúdo de informação de polimorfismo (PIC) também foram calculadas, além da realização da AMOVA. Os resultados da AMOVA sobre todos os locos mostraram que 37,5% da variação foi devida a diferenciação entre grupos (populações) e 62,5% dentro de grupos. As estimativas $F_{ST}$ entre pares de populações, oriundos da AMOVA indicaram um alto grau de diferenciação entre os cinco grupos de subpopulações estudadas, valores estes que variaram de 0,20 a 0,42.

Moraes et al. (2005) avaliaram a proximidade entre genótipos usados como genitor recorrente e genitor doador de alelos para alto teor de proteína, orientados por 57 pares de *primers* microssatélites, visando à obtenção de linhagens de soja com altos teores de proteína com o mínimo de gerações de retrocruzamento. Na análise de agrupamento foram utilizados os métodos SAHN do vizinho mais próximo e UPGMA, além do método de otimização de Tocher. Os métodos permitiram identificar grupos de potenciais cruzamentos entre os genótipos recorrentes e genótipos doadores, a partir das menores distâncias genéticas entre recorrentes e doadores.

Zucchi et al. (2003) realizaram um estudo sobre a caracterização da estrutura genética de 10 populações nativas de *Eugenia dysenterica.* As análises biométricas consistiram de testes das proporções de equilíbrio de Hardy-Weinberg, por meio do teste exato de Fisher (Weir, 1996), para 356 pares de *primers* microssatélites. Estimativas F ($F_{IS}$, $F_{IT}$ e $F_{ST}$) foram estimadas segundo a análise de variância de freqüências alélicas propostos por Weir & Cockerham (1984), que se baseia num modelo aleatório, cujas populações amostradas são consideradas representativas da espécie e com uma história evolutiva comum. Os autores utilizaram ainda a distância de Nei e o método de agrupamento UPGMA. O número de alelos por loco e as heterozigosidades também foram calculados. Eles constataram que algumas populações não se encontravam em equilíbrio para a maioria dos locos estudados e vários alelos exclusivos foram encontrados nas populações. A alta diversidade genética entre as populações, detectada por locos

microssatélites, indicou que estes marcadores são altamente sensíveis a detecção de estrutura populacional.

# 3. MATERIAL E MÉTODOS

## 3.1. Material

Inicialmente, foram simuladas três populações base, de tamanho finito e representantes de uma espécie vegetal qualquer de reprodução sexuada. As populações base são independentes e igualmente amostradas por 200 indivíduos em um conjunto de 20 locos codominantes e multialélicos. Os locos foram obtidos aleatoriamente e sem a caracterização de um genoma específico, sendo estes monomórficos ou polimórficos ao nível intrapopulacional. O número de alelos por loco variou de um a cinco, dentro de cada população. Não houve dados perdidos. Na simulação cada loco encontrava-se em equilíbrio de Hardy-Weinberg (EHW).

Adicionalmente, foram simuladas mais 39 populações, cada uma com 200 indivíduos, derivadas de cruzamentos direcionais, autofecundação ou acasalamento ao acaso das três populações base, totalizando 42 populações, conforme apresentado na Tabela 1. Durante o processo de simulação, considerou-se que nos cruzamentos direcionais – hibridações ($F_1$) e retrocruzamentos (RC) – ocorreram acasalamentos dirigidos planta a planta ou por mistura de pólen.

Os indivíduos não foram caracterizados por sexo, apenas considerados potencialmente viáveis e férteis. Para os cruzamentos, sistemas de auto-incompatibilidade não foram simulados, existindo a possibilidade de um indivíduo se autofecundar, ou seja, ser planta macho e fêmea, ao ser tomado aleatoriamente na população.

Tabela 1. Codificação das populações-base, híbridos $F_1$, gerações $F_n$, retrocruzamentos (RC), autofecundações e acasalamentos ao acaso, geradas via simulação

| População | Código[#] | Origem da população |
|---|---|---|
| 1 | $P_1$ | População-base 1 |
| 2 | $P_2$ | População-base 2 |
| 3 | $P_3$ | População-base 3 |
| 4 | $P_1a_1$ | 1ª geração de acasalamento ao acaso de P1 |
| 5 | $P_2a_1$ | 1ª geração de acasalamento ao acaso de P2 |
| 6 | $P_2a_2$ | 2ª geração acasalamento ao acaso de P2 |
| 7 | $P_2a_3$ | 3ª geração de acasalamento ao acaso de P2 |
| 8 | $P_2a_4$ | 4ª geração de acasalamento ao acaso de P2 |
| 9 | $P_3a_1$ | 1ª geração de acasalamento ao acaso deP3 |
| 10 | $P_2s_1$ | 1ª geração de autofecundação de P2 |
| 11 | $P_2s_2$ | 2ª geração de autofecundação de P2 |
| 12 | $P_2s_3$ | 3ª geração de autofecundação de P2 |
| 13 | $P_2s_4$ | 4ª geração de autofecundação de P2 |
| 14 | $P_2s_5$ | 5ª geração de autofecundação de P2 |
| 15 | $H_{12pp}$ | $F_1$ (P1xP2) |
| 16 | $H_{12mp}$ | $F_1$ (P1xP2) |
| 17 | $H_{13pp}$ | $F_1$ (P1xP3) |
| 18 | $H_{13mp}$ | $F_1$ (P1xP3) |
| 19 | $H_{23pp}$ | $F_1$ (P2xP3) |
| 20 | $F_{2(pp)}$[§] | autofecundação de $H_{23pp}$ |
| 21 | $F_{3(pp)}$ | autofecundação de $F_{2(pp)}$ |
| 22 | $F_{4(pp)}$ | autofecundação de $F_{3(pp)}$ |
| 23 | $H_{23(mp)}$ | $F_1$ (P2xP3) |
| 24 | $F_{2(mp)}$ | autofecundação de $H_{23mp}$ |
| 25 | $F_{3(mp)}$ | autofecundação de $F_{2(mp)}$ |
| 26 | $F_{4(mp)}$ | autofecundação de $F_{3(mp)}$ |
| 27 | $RC_{11pp}$[##] | [$F_1$(P1xP3) x P1] (1ª geração) |
| 28 | $RC_{12pp}$ | [($RC_1$) x P1] (2ª geração) |
| 29 | $RC_{13pp}$ | [($RC_2$) x P1] (3ª geração) |
| 30 | $RC_{14pp}$ | [($RC_3$) x P1] (4ª geração) |
| 31 | $RC_{11mp}$ | [$F_1$(P1xP3) x P1] (1ª geração) |
| 32 | $RC_{12mp}$ | [($RC_1$) x P1] (2ª geração) |
| 33 | $RC_{13mp}$ | [($RC_2$) x P1] (3ª geração) |
| 34 | $RC_{14mp}$ | [($RC_3$) x P1] (4ª geração) |
| 35 | $RC_{31pp}$ | [$F_1$(P1xP3) x P3] (1ª geração) |
| 36 | $RC_{32pp}$ | [($RC_1$) x P3] (2ª geração) |
| 37 | $RC_{33pp}$ | [($RC_2$) x P3] (3ª geração) |
| 38 | $RC_{34pp}$ | [($RC_3$) x P3] (4ª geração) |
| 39 | $RC_{31mp}$ | [$F_1$(P1xP3) x P3] (1ª geração) |
| 40 | $RC_{32mp}$ | [($RC_1$) x P3] (2ª geração) |
| 41 | $RC_{33mp}$ | [($RC_2$) x P3] (3ª geração) |
| 42 | $RC_{34mp}$ | [($RC_3$) x P3] (4ª geração) |

[#] a: acasalamento ao acaso; s: autofecundação; $_{pp}$: hibridação planta a planta; $_{mp}$: hibridação por mistura de pólen.

[§] Os termos (pp) ou (mp) indexado nas populações $F_2$, $F_3$ e $F_4$ significam que elas são oriundas da população $F_1$ obtida por cruzamento planta a planta ou mistura de pólen, respectivamente.

[##] Indexação nos retrocruzamentos: Por exemplo, no $RC_{12}$, a população base $P_1$ é a recorrente e trata-se da segunda geração de retrocruzamento.

### 3.2. Informações do conjunto de dados

A caracterização dos indivíduos ocorreu de forma que a descrição genotípica de cada um foi representada por códigos numéricos informativos dos alelos que eles possuem. Por exemplo, na existência de três alelos $A_1$, $A_2$ e $A_3$ em um loco qualquer, os genótipos homozigotos foram descritos por 11, 22 e 33 e os genótipos heterozigotos descritos por 12, 13 e 23. Uma amostra do arquivo de dados gerado está apresentada na Tabela 2. Em cada linha estão definidos os genótipos dos indivíduos, em que na primeira coluna do arquivo está referida a sua respectiva população (Tabela 1) e as demais colunas os respectivos locos. Adotou-se, para a maioria das análises, o arquivo de dados composto por este tipo de codificação. Alguns aplicativos computacionais utilizados neste estudo definem as informações genotípicas de forma diferenciada, mas sempre mantendo a informação básica, os alelos de cada loco.

Nas situações em que esta forma de codificação não foi utilizada, especificou-se o modo pelo qual os dados foram codificados. Todo o processo de simulação e codificação inicial dos dados foi realizado pelo programa GENES versão 2007.0.0 (CRUZ, 2006a), disponível em http://www.ufv.br/dbg/genes/gdown.htm.

Tabela 2. Amostra do arquivo de dados gerado pelo programa GENES vs 2007.0.0

| | Loco (j) | | | | |
|---|---|---|---|---|---|
| População (i) | 1 | 2 | ... | 19 | 20 |
| 1 | 11 | 33 | ... | 34 | 11 |
| 1 | 11 | 23 | ... | 22 | 11 |
| ... | ... | ... | ... | 23 | ... |
| 1 | 11 | 45 | ... | 11 | 11 |
| 2 | 33 | | | 12 | 34 |
| ... | ... | ... | ... | ... | ... |
| 2 | 12 | 23 | ... | 11 | 11 |
| ... | ... | ... | ... | ... | ... |
| 42 | 13 | 35 | ... | 45 | 12 |
| ... | ... | ... | ... | 12 | ... |
| 42 | 11 | 13 | ... | 22 | 23 |

A informação de freqüência alélica e genotípica foi quantificada a partir de 200 indivíduos amostrados dentro de cada população. As freqüências alélicas $[f(A_k)]$ foram estimadas a partir do número de ocorrência das diferentes classes genotípicas. Assim, denominando a ocorrência de homozigotos de $n_{ijkk}$ e de heterozigotos de $n_{ijkk'}$, na população i e loco j com os alelos k e k', tem-se:

$$f(A_k) = \hat{p}_{ijk} = \frac{2n_{ijkk} + \sum_{k<k'}^{a_j} n_{ijkk'}}{2N_i}$$

em que:

$\hat{p}_{ijk}$ expressa a estimativa da freqüência do alelo k (k = 1, 2, ..., $a_j$), do loco j (j = 1, 2, ..., L), na população i (i = 1, 2, ..., g) e $N_i$ representa o tamanho amostral da população i ($N_i = \sum_{k \le k'}^{a_j} n_{ijkk'}$).

A freqüência genotípica [$f(A_kA_k)$ ou $f(A_kA_{k'})$] também foi estimada a partir do número de ocorrência das diferentes classes genotípicas. Assim, no loco j, da população i, tem-se:

$$f(A_kA_k) = \hat{P}_{ijkk} = \frac{n_{ijkk}}{N_i}$$

$$f(A_kA_{k'}) = \hat{P}_{ijkk'} = \frac{n_{ijkk'}}{N_i}$$

Tanto as freqüências alélicas quanto as freqüências genotípicas são estimativas não viesadas e de máxima verossimilhança (Weir, 1996).

### 3.3. Aplicativos computacionais utilizados

As análises biométricas foram realizadas pelos seguintes programas:

a) Arlequin versão 3.1 (Excoffier et al., 2005), em http://cmpg.unibe.ch/software/arlequin3/;

b) GDA (Genetic data Analysis) versão 1.1 (Lewis & Zaykin, 2002), em http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php/;

c) GENEPOP versão interativa disponibilizada na *internet* (Raymond & Rousset, 1995), em http://genepop.curtin.edu.au/;

d) GENES versão 2007.0.0 (CRUZ, 2006a);

e) POPGENE versão 1.32 (Yeh et al., 1997), em http://www.ualberta.ca/~fyeh/download.htm/;

f) PowerMarker versão 3.25 (Liu & Muse, 2005), em http://statgen.ncsu.edu/powermarker/downloads.htm/;

g) TFPGA (Tool for Population Genetic Analyses) versão 1.3 (Miller, 1997), em http://www.marksgeneticsoftware.net/tfpga.htm/.

Os aplicativos foram acessados na *web* em 26 de novembro de 2006.

### 3.4. Análises de diversidade genética

Inferiu-se sobre a estrutura e diversidade genética ao nível intra e interpopulacional por meio de várias técnicas biométricas. Para cada metodologia empregada estão citados os aplicativos computacionais que foram usados.

#### 3.4.1. Em nível intrapopulacional

##### 3.4.1.1. Medidas descritivas

Estatísticas auxiliares foram empregadas para quantificar o conteúdo de informação intrínseco nos locos de cada população e, assim, caracterizá-

las quanto a diversidade genética existente nelas. Seguem as seguintes medidas:

***a - Número médio de alelos por loco ($\overline{A}_j$)***

O número de alelos ($a_j$) corresponde a quantidade de alelos diferentes no loco j, variando num intervalo de $1 < k < a_j$. O número médio de alelos por loco é dado por:

$$\overline{A}_j = \frac{1}{L}\sum_{j=1}^{L} a_j$$

*(Aplicativos usados: Arlequin, GDA, GENEPOP, GENES, POPGENE, PowerMarker e TFPGA)*

***b – Número total de alelos na população ($A_t$)***

Refere-se a soma de todos os alelos dos L locos avaliados dentro da população i, dado por:

$$A_t = \sum_{j=1}^{L} a_j$$

*(Aplicativo usado: GENES)*

***c - Número de alelos raros ($A_r$)***

Expressa o número de alelos com freqüência menor que 0,05 em cada população amostrada.

*(Aplicativo usado: GENES)*

***d - Número efetivo e número efetivo médio de alelos ($A_e$ e $\overline{A}_e$)***

O número efetivo de alelos por loco é dado por:

$$a_e = \frac{1}{\sum_{k=1}^{a_j} \hat{p}_{ijk}^2}$$ (Morgante et al., 1994)

O número efetivo de alelos total ($A_e$), definido por Pejic et al. (1998), é apresentado por:

$$A_e = \sum a_e$$

logo a média é dada por:

$$\bar{A}_e = \frac{1}{L}\sum_{j=1}^{L} a_e$$

*(Aplicativo usado: POPGENE)*

***e – Proporção de alelos na população ($P_a$)***

É dada por:

$$P_a = \frac{\text{(número de alelos da população)}}{\text{(número total de alelos na espécie)}}$$

*(Aplicativo usado: GENES)*

***f - Proporção de locos polimórficos (PLP)***

Três critérios são comumente utilizados para classificar um loco polimórfico (Cole, 2003):

i. loco exibindo polimorfismo em pelo menos um indivíduo da população amostrada, ou seja, não existe alelo fixado;

ii. loco em que o alelo mais comum tem freqüência menor que 99%;

iii. loco em que o alelo mais comum tem freqüência menor que 95%.

A proporção de locos polimórficos é dada por:

$$PLP = \frac{\text{número de locos polimórficos}}{\text{número total de locos}}$$

*(Aplicativos usados: Arlequin, GDA, GENES, POPGENE e TFPGA)*

***g – Número médio de alelos por loco polimórfico ($\bar{A}_p$)***

É dado por:

$$\bar{A}_p = \frac{1}{L_p}\sum_{j=1}^{Lp} a_j$$

em que $L_p$ é o número de locos polimórficos.

*(Aplicativos usados: GDA e GENES)*

***h – Número médio de genótipos por loco ($\bar{g}_j$)***

É dado por:

$$\overline{g}_j = \frac{1}{L}\sum_{j=1}^{L} n_c$$

em que $n_c$ é o número de classes genotípicas do loco j.

*(Aplicativos usados: POPGENE e PowerMarker)*

***i – Conteúdo de informação polimórfica (PIC)***

O PIC de um loco qualquer é dado por:

$$PIC_{ij} = 1 - \sum_{k=1}^{a_j} \hat{p}_{ijk}^2 - \sum_{k<k'}^{a_j} 2\hat{p}_{ijk}^2 \hat{p}_{ijk'}^2 \text{ (Botstein et al., 1980)}$$

O $PIC_i$ para todos os locos é dado pela média dos $PIC_{ij}$ para cada loco.

*(Aplicativos usados: GENES e PowerMarker)*

***j - Índice Shannon-Wiener (Shannon ou Shannon-Weaver - H')***

A função de Shannon-Wiener (Alfenas et al., 1991) proposta para medir a diversidade ou riqueza de espécies em estudos de ecologia, também pode ser empregada para medir a diversidade (riqueza) fenotípica ou genotípica dentro de uma população. Definiu-se a estatística de Shannon-Wiener para a população i no loco j, como:

$$H'_{ij} = -\sum_{k\leq k'}^{a_j} \hat{P}_{ijkk'} \ln(\hat{P}_{ijkk'})$$

Para todos os locos, tem-se:

$$H' = \frac{1}{L}\sum_{j=1}^{L} H'_{ij}$$

*(Aplicativos usados: GENES e POPGENE)*

***k – Heterozigosidade observada ($h_o$)***

Trata-se de uma simples medida de variação genética em uma população, dada por:

$$\hat{h}_o = \sum_{k<k'}^{a_j} \hat{P}_{ijkk'}$$

A heterozigosidade observada média é dada por:

$$\hat{H}_o = \frac{\sum_{j=1}^{L} \hat{h}_o}{L}$$

*(Aplicativos usados: Arlequin, GDA, GENEPOP, GENES, POPGENE, PowerMarker e TFPGA)*

***I – Índice de fixação/endogamia (f)***

O índice de fixação, também conhecido como coeficiente de endogamia, dentro de cada população foi calculado conforme Weir (1996), usando o método dos momentos, cujo estimador é o seguinte:

$$\hat{f}_{ij} = \frac{\sum_{k=1}^{a_j} \left( \hat{p}_{ijkk} - \hat{p}_{ijk}^2 \right) + \frac{1}{2N_i} \left( 1 - \sum_{k=1}^{a_j} \hat{P}_{ijkk} \right)}{\left( 1 - \sum_{k=1}^{a_j} \hat{p}_{ijk}^2 \right) - \frac{1}{2N_i} \left( 1 - \sum_{k=1}^{a_j} \hat{P}_{ijkk} \right)}$$

O índice de fixação médio (considerando L locos) foi computado pela soma do numerador e denominador separadamente e, então, formou-se um quociente a partir destas duas somas (Lewis & Zaykin, 2002).

(*Aplicativos usados: GDA e PowerMarker)*

Outro estimador foi obtido conforme a expressão abaixo:

$$\hat{f}_{ij} = \frac{\hat{h}_e - \hat{h}_o}{\hat{h}_e}$$ (Nei & Kumar, 2000)

em que $\hat{h}_e$ e $\hat{h}_o$ correspondem as estimativas de heterozigosidade esperada e observada, respectivamente.

(*Aplicativos usados: GENES e POPGENE)*

Também é possível obter *f* por intermédio das expressões:

$$\hat{f}_{ijkk'} = \frac{-2N_i T_{ijkk'}}{a_{ijk} a_{ijk'}}$$ (Robertson & Hill, 1984)

em que $\hat{f}_{ijkk'}$ é o estimador do coeficiente de endogamia dentro população i para o loco j, e o genótipo formado pelos alelos k e k', em que k < k'; $T_{ijkk'}$ é o

desvio entre o número de heterozigotos observado e o número de heterozigotos das proporções esperadas no EHW, dado por:

$$T_{ijkk'} = \frac{\left[(2N_i - 1)n_{ijkk'} - n_{ijk}n_{ijk'}\right]}{2(N_i - 1)}$$

Aqui a média dos estimadores $\hat{f}_{ijkk'}$'s corresponde ao coeficiente de endogamia no loco j.

e

$$\hat{f}_{ijk} = 1 - \frac{h_{ok}}{2\hat{p}_{ijk}(1-\hat{p}_{ijk})}$$ (Weir & Cockheram, 1984)

em que $\hat{f}_{ijk}$ é estimado para cada alelo k dentro de cada loco j. A razão entre as somas do numerador e denominador de $\hat{f}_{ijk}$ dão a estimativa do coeficiente para o loco j.

(*Aplicativos usados: GENEPOP)*

***m - Heterozigosidade esperada e, ou, diversidade gênica ($h_e$ ou $D_j$)***

Considerando o loco j, o estimador viesado é dado por:

$$\hat{h}_e = 1 - \sum_{k=1}^{a_j} \hat{p}_{ijk}^2$$ (Nei, 1973)

*(Aplicativos usados: GENEPOP, GENES, POPGENE, PowerMarker e TFPGA)*

Um estimador não viesado é dado por:

$$\hat{h}_e = \hat{D}_j = \frac{\left(1 - \sum_{k=1}^{a_j} \hat{p}_{ijk}^2\right)}{\left(1 - \frac{1+\hat{f}}{2N_i}\right)}$$ (Weir, 1996)

(*Aplicativo usado: PowerMarker)*

Para *f* = 1, tem-se:

$$\hat{h}_e = \frac{N_i}{(N_i - 1)} (1 - \sum_{k=1}^{a_j} \hat{p}_{ijk}^2)$$

(*Aplicativo usado: Arlequin)*

Para *f* = 0, tem-se:

$$\hat{h}_e = \frac{2N_i}{(2N_i - 1)} (1 - \sum_{k=1}^{a_j} \hat{p}_{ijk}^2) \quad \text{(Nei, 1978)}$$

(*Aplicativos usados: GDA, POPGENE e TFPGA)*

A heterozigosidade esperada (diversidade gênica) média é dada por:

$$\hat{H}_e = \frac{\sum_{j=1}^{L} \hat{h}_e}{L}$$

ou

$$\overline{D}_j = \frac{\sum_{j=1}^{L} \hat{D}_j}{L}$$

#### 3.4.1.1.1. Comparações entre as medidas descritivas

Correlacionou-se as medidas descritivas, utilizando as correlações de Pearson e de Spearman. Para correlacioná-las, foram obtidas estimativas de cada uma delas, a partir das informações médias dos 20 locos e das 42 populações (observações). Diferentes estimadores de um mesmo parâmetro, também foram submetidos a comparações.

### 3.4.1.2. Tamanho efetivo populacional

Os tamanhos efetivos de cada população foram estimados, conforme Vencovsky (1992), pela seguinte expressão:

$$\hat{N}_e = \frac{N_i}{1 + \hat{f}}$$

### 3.4.1.3. Equilíbrio de Hardy-Weinberg

Cada loco j (= 1, ..., 20) pertencente a população i (= 1, ..., 42) foi submetido ao teste de equilíbrio de Hardy-Weinberg (EHW). Para os testes que serão apresentados, a hipótese de nulidade ($H_0$) geral é à união aleatória de gametas.

***a - Teste de qui-quadrado***

O teste de qui-quadrado ($\chi^2$) é dado pela seguinte expressão:

$$\chi^2 = \sum_{c} \frac{(O_c - E_c)^2}{E_c} = \sum_{k}^{a_j} \frac{(n_{ijkk} - N_i \hat{p}_{ijk}^2)^2}{N_i \hat{p}_{ijk}^2} + \sum_{k<k'}^{a_j} \frac{(n_{ijkk'} - 2N_i \hat{p}_{ijk} \hat{p}_{ijk'})^2}{2N_i \hat{p}_{ijk} \hat{p}_{ijk'}}$$

em que $O_c$ e $E_c$ representam o número observado e esperado de indivíduos para a c-ésima classe genotípica, respectivamente. Considerando k alelos em um loco j, tem-se as classes genotípicas no EHW dadas pela seguinte proporções:

Relação Genotípica esperada = $\sum_{k \leq}^{a_j} \sum_{k'}^{a_j} \hat{p}_{ijk} \hat{p}_{ijk'}$ .

Esta estatística tem k(k-1)/2 graus de liberdade.

*(Aplicativos usados: GENES, POPGENE, PowerMarker e TFPGA)*

***b - Teste da Razão de verossimilhança (Teste G ou $G^2$)***

A razão de verossimilhança oferece um caminho sistemático para se testar o EHW quando existem mais de dois alelos por loco. O teste é baseado na razão entre a função de máxima verossimilhança do modelo definido pelas freqüências genotípicas esperadas ($L_0$) no EHW e a função de máxima verossimilhança do modelo definido pelas freqüências genotípicas observadas ($L_1$). Como as variáveis (locos ou genes) seguem distribuição multinomial [ X ~ B (N, p)], as funções $L_0$ e $L_1$ foram definidas por:

$$L\,(p_{ijkk'}; n_{ijkk'}) = \lambda \prod_{k \leq k'}^{a_j} \hat{p}_{ijkk'}^{n_{ijkk'}} ,$$

em que:

$$\lambda = \frac{N_i!}{\prod_{k=1}^{a_j} n_{ijkk'}!}$$

Pode-se definir $L_0$ e $L_1$ pelo logarítmo neperiano das respectivas funções $L(p_{ijkk'}; n_{ijkk'})$, de modo que:

$$\ln L_0 = \ln \lambda + \sum_{k=1}^{a_j} n_{ijkk} \ln\left(\frac{n_{ijk}}{2N_i}\right)^2 + \sum_{k}^{a_j}\sum_{k'\neq k}^{a_j} n_{ijkk'} \ln\left(2\frac{n_{ijk}}{2N_i}\frac{n_{ijk'}}{2N_i}\right)^2$$

e

$$\ln L_1 = \ln \lambda + \sum_{k=1}^{a_j} n_{ijkk} \ln\left(\frac{n_{ijkk}}{N_i}\right) + \sum_{k}^{a_j}\sum_{k'\neq k}^{a_j} n_{ijkk'} \ln\left(\frac{n_{ijkk'}}{N_i}\right)$$

O primeiro somatório de cada função corresponde a todos os homozigotos e o segundo a todos os heterozigotos do loco j na população i. O teste para o EHW é dado pela seguinte razão de verossimilhança:

$$G^2 = -2 \ln\left(\frac{L_0}{L_1}\right) = -2(\ln L_0 - \ln L_1)$$

Sob EHW, a quantidade $-2(\ln L_0 - \ln L_1)$ tem distribuição aproximada de qui-quadrado, com k(k-1)/2 graus de liberdade.

*(Aplicativos usados: POPGENE, PowerMarker e TFPGA)*

***c - Teste Exato de Fisher***

O teste exato tem como objetivo verificar se as informações (alélicas e genotípicas) de uma população (amostra) podem ser usadas para rejeitar uma hipótese, na situação em que a probabilidade total condicionada a hipótese desta amostra é pequena, ou, então, menos provável do que ela. Para um melhor entendimento, considere um arranjo qualquer do número de indivíduos das classes genotípicas de um loco j proveniente de $N_i$ indivíduos amostrados na população i. Este arranjo genotípico foi definido em uma tabela de contingência de dimensão k x k, em que k é o número de alelos para o loco considerado.

Esta abordagem permitiu estimar uma probabilidade condicional oriunda da probabilidade de um arranjo de classes genotípicas, assumindo que estas classes estão em EHW e condicionadas as k freqüências alélicas

observadas. Sob este ponto de vista, vários arranjos das classes genotípicas foram obtidos para um particular conjunto de alelos observados. Consequentemente, várias probabilidades condicionais foram estimadas. Haldane (1954) afirmou que não é necessário contabilizar todos os possíveis arranjos genotípicos, apenas os possíveis arranjos dos genótipos heterozigotos.

Generalizando para k alelos, a probabilidade condicional de um arranjo com a quantidade genotípica $n_{ijkk'}$ de heterozigotos (para k < k'), condicionado as quantidades alélicas $n_{ijk}$ observadas, provenientes de k alelos do loco j amostrados na população i, é expressa por:

$$\Pr\left[(n_{ijkk'})\middle|(n_{ijk})\right] = \frac{N_i!\, 2^H \prod_{k=1}^{a_j}(n_{ijk})!}{(2N_i)! \prod_{k \le k'}^{a_j}(n_{ijkk'})!} \quad \text{(Weir, 1996)}$$

em que $H = \sum_{k}^{a_j}\sum_{k' \neq k}^{a_j} n_{ijkk'}$ é o número de indivíduos heterozigotos amostrados do loco j na população i.

A hipótese de nulidade é rejeitada quando a soma das probablidades (condicionais) ordenadas, de forma crescente, for menor do que um nível de significância $\alpha$ estabelecido. Ressalta-se que a probabilidade condiconal proveniente do arranjo genotípico observado na amostra original também deve ser computada no somatório das probabilidades condicionais.

Considerando as condições do presente trabalho, cujo tamanho amostral é grande ($N_i$ = 200) e podendo existir vários alelos por loco, as probabilidades condicionais serão muito pequenas, embora a quantidade relevante seja a soma das probabilidades, ou seja, a probabilidade agregada (valor de *p* ou área de rejeição). Nesta situação o número de arranjos genotípicos possíveis (tabelas de contigência) é elevado. Guo & Thompson (1992) sugeriram uma versão permutada para se obter o valor de *p*. Considere uma população de tamanho $N_i$, em que cada alelo $A_{ijk}$ possue $n_{ijk}$ cópias, num total de $2N_i$ alelos. Procede-se uma desestruturação em todos os $N_i$ indivíduos (genótipos) da população amostrada, de maneira que pares de alelos são tomados aleatoriamente até reconstituí-los novamente. Este

processo é repetido inúmeras vezes. Sob EHW, os alelos estão distribuídos independentemente nos genótipos, assim um arranjo genotípico encontrado pelo processo de permutação (embaralhamento de alelos) corresponde a um dos arrrajos possíveis de serem encontrados sob EHW. A freqüência alélica de cada arranjo genotípico é a mesma que a da população amostrada e suas probabilidades condicionais podem ser calculadas conforme a expressão generalizada dada acima. O valor de *p* (área de rejeição) é dado pela proporção de arranjos genotípicos (ou tabelas de contingência) que tiveram as probabilidades condicionais menores ou iguais a probablidade condicinal do arranjo genotípico original amostrado.

Os aplicativos computacionais tem estabelecido variações para este processo de permutação, com destaque aos métodos MCMC (cadeia de Markov e Monte Carlo, do inglês, *Markov Chain Monte Carlo*). Ao invés de enumerar todos os possíveis arranjos genotípicos, usa-se um processo de "caminhada" aleatória (cadeia de Markov) capaz de explorar eficientemente o espaço de todos os arranjos genotípicos possíveis (tabelas de contigência), mantendo-se o número de alelos da amostra original, ou melhor, as marginais da tabela de contingência original.

Para iniciar o processo MCMC, os programas costumam pedir que seja definido um número de passos de "dememorização", que corresponde ao número de passos que permite a cadeia de Markov "esquecer" o estado inical da tabela de contigência original e que o ponto de partida para realização da análise seja independente do estado inicial, mantendo-se, é claro, as quantidades alélicas originais. Raramente é necessário exceder 1.000 passos de "dememorização". Outro procedimento a ser definido pelos usuários é o processo de *batching (*loteamento), que consiste em subdivisões da quantidade total de permutações a serem executadas. Este processo permite atingir uma convergência a ser checada automaticamente. Basicamente o valor de *p* é calculado para cada *batch* (lote) e o desvio padrão (ou coeficiente de variação) destes valores de *p* é comparado com o critério de convergência ($\alpha$). O processo, como um todo, pára quando o desvio padrão encontra-se menor do que o valor de $\alpha$ especificado pelo usuário, ou seja, quando o critério de convergência for atingido.

Em todos os testes ($\chi^2$ , $G^2$ e exato de Fisher) considerou-se o nível de significância ($\alpha$) igual a 0,05 para a tomada de decisão em rejeitar ou não a hipótese de nulidade. Também foram definidas 1000 dememorizações e 100 *batches*.

*(Aplicativos usados: Arlequin, GDA, GENEPOP, PowerMarker e TFPGA)*

### 3.4.1.4. Desequilíbrio gamético

O estudo de desequilíbrio gamético foi realizado somente entre pares de locos. Esta análise foi feita apenas na população base $P_2$ e gerações de acasalamento ao acaso ($P_2a_1$, $P_2a_2$, $P_2a_3$ e $P_2a_4$). Além disso, os alelos passaram por novo processo de codificação. Codificou-se o alelo mais freqüente de cada loco como 1 e todos os demais como 2. Ressalta-se que para aplicação destes códigos foi levado em consideração que este grupo de populações era representante de uma espécie e, consequentemente, o alelo mais freqüente de cada loco era aquele pertencente a este grupo.

#### *a - Teste da razão de verossimilhança (LOD e S)*

A fase gamética dos locos é desconhecida, não sendo possível a aplicação do teste exato baseado na cadeia de Markov (Excoffier et al., 2006). Deste modo, o desequilíbrio gamético é testado por meio da razão de verossimilhança, contabilizando as proporções genotípicas observadas e esperadas. Para isso, considere dois locos quaisquer, o loco 1 (A) e o loco 2 (B), como o par de locos a ser testado. O loco 1 possui os alelos 1 ($A_1$) e 2 ($A_2$), e o loco 2 alelos 1 ($B_1$) e 2 ($B_2$). Assim, as freqüências gaméticas ($p_{ijk,j'k}$) esperadas, em função das freqüências alélicas ($p_{ijk}$), são apresentadas da seguinte maneira:

$p_{i11,21} = p_{i11} \cdot p_{i21}$

$p_{i11,22} = p_{i11} \cdot p_{i22} = p_{i11} \cdot (1- p_{i21})$

$p_{i12,21} = p_{i12} \cdot p_{i21} = (1 - p_{i11}) \cdot p_{i21}$

$p_{i12,22} = p_{i12} \cdot p_{i22} = (1 - p_{i11}) \cdot (1- p_{i21})$

Por analogia, considera-se que as freqüências gaméticas observadas são dadas por:

$p_{i11,21} = (p_{i11} \,.\, p_{i21}) + D_{11,21}$

$p_{i11,22} = [p_{i11} \,.\, (1- p_{i21})] - D_{11,21}$

$p_{i12,21} = [(1 - p_{i11}) \,.\, p_{i21}] - D_{11,21}$

$p_{i12,22} = [(1 - p_{i11}) \,.\, (1- p_{i21})] + D_{11,21}$

Assim, as freqüências genotípicas esperadas são produtos das freqüências gaméticas. Por exemplo, O genótipo $A_1A_1B_1B_1$ tem freqüência esperada ($p_{i111,211}$) igual a:

$p_{i111,211} = p_{i11,21} \,.\, p_{i11,21} = [(p_{i11} \,.\, p_{i21}) + D_{11,21}] \,.\, [(p_{i11} \,.\, p_{i21}) + D_{11,21}] = (p_{i11}^2 p_{i21}^2 + 2D_{11,21} p_{i11} p_{i21} + D_{11,21}^2)$

e o genótipo $A_2A_2B_2B_2$ tem freqüência esperada ($p_{i122,222}$) dada por:

$p_{i122,222} = p_{i12,22} \,.\, p_{i12,22} = [(p_{i12} \,.\, p_{i22}) + D_{11,21}] \,.\, [(p_{i12} \,.\, p_{i22}) + D_{11,21}] = (p_{i12}^2 p_{i22}^2 + 2D_{11,21} p_{i12} p_{i22} + D_{11,21}^2)$

A função logarítmica (suporte) de verossimilhança para o modelo com dois locos é a seguinte:

$$\ln L_1 (D_{11,21}, p_{ijk}; n_{ijkk',j'kk'}) = n_{i111,211} \ln (\hat{p}_{i11}^2 \hat{p}_{i21}^2 + 2D_{11,21} \hat{p}_{i11} \hat{p}_{i21} + D_{11,21}^2) + \ldots + n_{i122,222} (\hat{p}_{i12}^2 \hat{p}_{i22}^2 + 2D_{11,21} \hat{p}_{i12} \hat{p}_{i22} + D_{11,21}^2)$$

em que $n_{ijkk',j'kk'}$ é o número de indivíduos com o genótipo $A_kA_{k'}B_kB_{k'}$ (para todo k = k' ou k $\neq$ k'). O estimador de verossimilhança para $D_{11,21}$ é:

$$\frac{\partial \ln L_1}{\partial D_{11,21}}$$

A hipótese de nulidade foi testada por:

$$LOD = Log_{10} \left( \frac{L_1}{L_0} \right)$$ (Schuster & Cruz, 2004)

em que $L_0 = L_1 (D_{11,21} = 0, p_{ijk}; n_{ijkk',j'kk'})$ .

Se $L_1 > L_0$, o LOD escore é positivo. Conclui-se que os locos estão em desequilíbrio quando o LOD é maior que 3, ou seja, probabilidade de 1000 para 1.

*(Aplicativo usado: GENES)*.

Outra maneira apresentada pela literatura para a construção da razão de verossimilhança é estimar a composição gamética das populações pelo algoritmo EM (*Expectation-Maximization*). A função de verossimilhança é a seguinte:

$$L \propto \prod_{jkk',j'kk'} (c_{ijkk',j'kk'}\, p_{ijk,j'k}\, p_{ijk',j'k'})^{n_{ijkk',jkk'}}$$

O produtório é realizado sobre todos os genótipos. $c_{ijkk',j'kk'}$ assume valor igual a 1 no caso de genótipo duplo homozigoto, valor 4 se for duplo heterozigoto e valor 2 em outras situações. Aqui não há a preocupação em se estimar o coeficiente de desequilíbrio, mas apenas testar a associação (não) aleatória entre os alelos dos dois locos (Slatkin & Excoffier, 1996).

A função de verossimilhança dos dados assumindo equilíbrio gamético ($L_0$) também é construída considerando que as freqüências gaméticas ($p_{ijk,j'k}$) esperadas são obtidas pelo produto das freqüências alélicas ($p_{ijk}$). A função de verossimilhança dos dados, sem assumir equilíbrio gamético ($L_1$) é obtida aplicando-se o algoritmo EM (*Expectation-Maximization*) para estimar as freqüências gaméticas (maiores detalhes em Excoffier & Slatkin, 1995; Slatkin & Excoffier, 1996). Assim a hipótese de nulidade é dada por:

$$S = -2 \ln\left(\frac{L_0}{L_1}\right) = -2\,(\ln L_0 - \ln L_1)$$

Para grandes amostras, esta estatística segue distribuição de qui-quadrado, o que pode não ocorrer com pequenas amostras com grande número de alelos por loco. Sugere-se o seguinte procedimento de permutação (Excoffier et al., 2006).

i) Permuta-se os alelos entre indivíduos para um loco apenas;

ii) Estima-se novamente a verossimilhança destes dados ($L_1$') pelo algoritmo EM. A função $L_0$ não é afetada pelo processo de permutação.

iii) Repete-se os passos i) e ii) 1000 vezes para se obter a distribuição nula de $L_1$.

Definiu-se o nível de significância ($\alpha$) igual a 0,05.

*(Aplicativo usado: Arlequin)*

O coeficiente de desequilíbrio clássico (Lewontin & Kojima, 1960) é então estimado por:

$$\hat{D}_{11,21} = \hat{p}_{i11,21} - \hat{p}_{i11} \cdot \hat{p}_{i21}$$

em que $\hat{p}_{i11,21}$ é a freqüência gamética estimada pelo algoritmo EM (Excoffier & Slatkin, 1995).

*(Aplicativo usado: PowerMarker)*

Outro coeficiente de desequilibro utilizado foi $D'_{11,21}$ (Lewontin, 1964), que refere-se ao coeficiente de desequilíbrio clássico ($D_{11,21}$) padronizado pelo valor máximo ($D_{11,21máx}$), dado a seguir:

$$D'_{11,21} = \frac{D_{11,21}}{D_{11,21máx}}$$

em que $D_{11,21máx}$ pode assumir um dos seguintes valores:

$\min [p_{i11}.(1- p_{i21}); (1 - p_{i11}).p_{i21}]$ se $D_{11,21} > 0$;

$- \text{máx} [(p_{i11}.p_{i21}); (1 - p_{i11}).(1- p_{i21})]$ se $D_{11,21} < 0$

*(Aplicativo usado: PowerMarker)*

O quadrado do coeficiente de correlação ($r^2$) também utilizado como medida do desequilíbrio, é expresso como função de $D_{11,21}$, dado por:

$$r^2 = \frac{D^2_{11,,21}}{p_{i11}(1-p_{i11}).\, p_{ij21}(1-p_{21,})}$$

*(Aplicativo usado: PowerMarker)*

Estas medidas ($D_{11,21}$, $D'_{11,21}$ e $r^2$) foram computadas entre todos os pares de alelos para os diferentes locos, assumindo EHW. Como existem apenas dois alelos por locos defini-se apenas uma única estimativa para cada medida estudada, a cada par de loco.

***c - Teste de qui quadrado***

A razão de verossimilhança segue distribuição de qui-quadrado ($\chi^2$) para grandes amostras com número pequeno de alelos por loco. Assim, o teste de $\chi^2$ para $H_0$: $D_{11,21} = 0$ foi calculado por:

$$\chi^2 = \frac{(2N_i)\,\hat{D}^2_{11,21}}{\hat{p}_{i11}\,.(1-\hat{p}_{i11})\,.\,\hat{p}_{i21}\,.(1-\hat{p}_{i21})}$$

com (2 - 1)(2 - 1) = 1 grau de liberdade para os locos 1 e 2 (j e j').

*(Aplicativo usado: PowerMarker)*

### 3.4.1.5. Teste de deficiência e excesso de heterozigotos (Teste U)

O teste de deficiência e excesso de heterozigotos concentra-se na mesma hipótese de nulidade do teste de EHW, que se baseia na união aleatória de gametas e se assemelha ao teste exato de Fisher (Haldane, 1954; Guo e Thompson, 1992; Weir, 1996). A diferença entre eles é a construção da área de rejeição. Para testar o EHW, a probabilidade condicional da população (amostra) observada é usada para definir a área de rejeição. O valor de *p* do teste corresponde ao somatório das probabilidades de todas as tabelas de contingência, mantendo a quantidade alélica original, com probabilidade igual ou menor que a probabilidade condicional da população observada. Quando as hipótese alternativas ($H_a$'s) são referentes à deficiência ou excesso de heterozigotos, sugere-se a utilização de um teste de maior poder, denominado teste U (*score test*).

Conforme Rousset & Raymond (1995), o teste U é construído a partir do escore:

$$U = \left.\frac{\delta \log L_2}{\delta f}\right|_{f=0,\, p_{ijk}} = \sum_{k=1}^{a_j} \frac{n_{ijkk}}{\hat{p}_{ijk}} - N_i$$

em que $L_2$ é uma função de máxima verossimilhança, definida por:

$$L_2(f, p_{ijk}; N_i) = \frac{N_i!}{n_{ij11}! \ldots n_{ijkk}!}[\hat{p}_{ij1}^2 + \hat{f}\ \hat{p}_{ij1}^2(1-\hat{p}_{ij1}^2)]^{n_{ij11}} .\ [2\hat{p}_{ij1}\hat{p}_{ij2}(1-\hat{f})]^{n_{ij12}} \ldots$$

$$\ldots [\hat{p}_{ijk}^2 + f\ \hat{p}_{ijk}^2(1-\hat{p}_{ijk}^2)]^{n_{ijkk}}$$

e $\hat{f}$ é o coeficiente de fixação/endogamia dentro da população, considerando a situação de endogamia regular, em que as demais estimativas de $f_{ijkk'}$ são iguais a $\hat{f}$.

À semelhança do processo de permutação descrito no teste de EHW (seção 3.4.1.3.) é definido o valor de *p*. Pelo teste U foram testados cada loco j de cada população i.

*Teste U múltiplo (global)*

Na verdade os resultados explorados foram dos testes U globais, ou seja, considerando todos os 20 locos para cada população i. Conforme as definições dos autores referidos, o teste U global foi apresentado da seguinte forma:

$$U_g = \sum_{j=1}^{L} U_j = \sum_{j=1}^{L}\sum_{k=1}^{a_{ij}} \frac{n_{ijkk}}{\hat{p}_{ijkk}} - N_i$$

*(Aplicativos usados: GENEPOP)*

### 3.4.2. Em nível interpopulacional

#### 3.4.2.1. Divergência genética

A avaliação da diversidade genética entre as 42 populações foi investigada por intermédio de distâncias geométricas, genéticas e genotípica. Considerando as informações dos 20 locos na estimação das distâncias, foram utilizadas as seguintes medidas:

***a - Distância Euclidiana média***

É dada por:

$$D_{E,ii'} = \sqrt{\sum_{j=1}^{L}\sum_{k=1}^{a_j}(\hat{p}_{ijk} - \hat{p}_{i'jk})^2}$$

*(Aplicativos usados: GENES e PowerMarker)*

***b - Distância de Rogers (1972)***

É dada por:

$$D_{R,ii'} = \frac{1}{L}\sum_{j=1}^{L}\sqrt{\frac{1}{2}\sum_{k=1}^{n_j}\left(\hat{p}_{ijk} - \hat{p}_{i'jk}\right)^2}$$

*(Aplicativos usados: GENES e PowerMarker)*

***c - Distância de Rogers modificada (Goodman & Stuber, 1983)***

É dada por:

$$D_{GS,ii'} = \frac{1}{\sqrt{2L}}\sqrt{\sum_{j=1}^{L}\sum_{k=1}^{a_j}\left(\hat{p}_{ijk} - \hat{p}_{i'jk}\right)^2}$$

*(Aplicativos usados: GENES e TFPGA)*

***d - Distância angular (Cavalli-Sforza & Edwards, 1967)***

Utilizou-se as seguintes distâncias angulares:

- *Complemento aritmético do cosseno*

$$D_{COS,ii'} = \sqrt{\frac{1}{L}\sum_{j=1}^{L}(1 - \sum_{k=1}^{a_j}\sqrt{\hat{p}_{ijk}\hat{p}_{i'jk}})}$$

*(Aplicativo usado: GENES)*

- Comprimento da corda

$$D_{CC,ii'} = \frac{2}{\pi L}\sum_{j=1}^{L}\sqrt{2\,(1 - \sum_{k=1}^{a_j}\sqrt{\hat{p}_{ijk}\hat{p}_{i'jk}})}\,,$$

*(Aplicativos usados: GENES e PowerMarker)*

***e – Distância de Nei et al. (1983)***

É dada por:

$$D_{N83,ii'} = \frac{1}{L}\sum_{j=1}^{L}(1-\sum_{k=1}^{a_j}\sqrt{\hat{p}_{ijk}\hat{p}_{i'jk}})$$

*(Aplicativos usados: PowerMarker)*

Nenhuma das medidas de distância geométrica descritas acima envolve qualquer modelo evolutivo.

***f - Distância genética padronizada de Nei (1972)***

É dada pelo logaritmo neperiano da identidade gênica ($I_{N72,ii'}$), definda por:

$$D_{N72,ii'} = -\ln(I_{N72,ii'}) = -\ln\left(\frac{J_{ii'}}{\sqrt{J_i.J_{i'}}}\right) = -\ln\frac{\frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{a_j}\hat{p}_{ijk}\hat{p}_{i'jk}}{\sqrt{\frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{a_j}\hat{p}_{ijk}^2 \quad \frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{a_j}\hat{p}_{i'jk}^2}}$$

em que,

$$J_{ii'} = \frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{a_j}\hat{p}_{ijk}\hat{p}_{i'jk}\ ,\ J_i = \frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{a_j}\hat{p}_{ijk}^2 \text{ e } J_{i'} = \frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{a_j}\hat{p}_{i'jk}^2$$

*(Aplicativos usados: GDA, GENES, POPGENE, PowerMarker e TFPGA)*

***g – Distância mínima (Nei 1973)***

É dada por:

$$D_{m,ii'} = \frac{(J_i + J_{i'})}{2} - J_{ii'}$$

*(Aplicativos usados: PowerMarker e TFPGA)*

***h - Distância de Reynolds et al. (1983)***

É dada por:

$$D_{RWC,ii'} = = \frac{\sum_{j=1}^{L}\left\{\sum_{k=1}^{a_j}(\hat{p}_{ijk}-\hat{p}_{i'jk})^2 - \frac{1}{2(2N-1)}\left[2-\sum_{k=1}^{a_j}(\hat{p}_{ijk}^2+\hat{p}_{i'jk}^2)\right]\right\}}{2\sum_{j=1}^{L}(1-\sum_{i=1}^{a_j}\hat{p}_{ijk}\hat{p}_{i'jk})}$$

em que $N = N_i = N_{i'}$, ou seja, considera-se as populações com o mesmo tamanho amostral.

*(Aplicativos usados: GDA e TFPGA)*

Nesta mesma expressão, ignorando-se os termos envolvidos com o tamanho amostral, tem-se:

$$D'_{RWC,ii'} = \frac{\sum_{j=1}^{L}\sum_{k=1}^{a_j}(\hat{p}_{ijk}-\hat{p}_{i'jk})^2}{2\sum_{j=1}^{L}(1-\sum_{i=1}^{a_j}\hat{p}_{ijk}\hat{p}_{i'jk})}$$

*(Aplicativos usados: PowerMarker)*

***i – Distâncias de Latter (1972 e 1973)***

As distâncias de Latter são definidas, respectivamente por:

$$D_{L72,ii'} = \frac{(J_i + J_{i'}) - J_{ii'}}{1 - J_{ii'}}$$

e

$$D_{L73,ii'} = - \ln (1 - D_{L72})$$

*(Aplicativo usado: PowerMarker)*

***h - Distância Genotípica de Hedrick (1971)***

É dada pelo complemento aritmético da identidade genotípica ($I_{H\,ii'}$), sendo definda por:

$$D_{H,ii'} = \frac{1}{L}\sum_{j=1}^{L}(1-I_{H,ii'}) = \frac{1}{L}\sum_{j=1}^{L}\left[1-\frac{\sum_{k}^{a_j}\leq\sum_{k'}^{a_j}\hat{P}_{ijkk'}\hat{P}_{i'jkk'}}{\frac{1}{2}\left(\sum_{k}^{a_j}\leq\sum_{k'}^{a_j}\hat{P}_{ijkk'}^2+\sum_{k}^{a_j}\leq\sum_{k'}^{a_j}\hat{P}_{i'jkk'}^2\right)}\right]$$

*(Aplicativo usado: GENES)*

#### 3.4.2.1.1. Agrupamento

A partir das informações provenientes das matrizes de distância procedeu-se o agrupamento das populações através dos métodos de otimização de Tocher, UPGMA e projeção gráfica bidimensional, descritos a seguir:

***a - Método de Otimização de Tocher***

Na matriz de distância foi identificado o par de populações mais similares. Essas populações formaram o grupo inicial. A partir daí foi avaliada a possibilidade de inclusão de novas populações, adotando-se o critério de que a distância média intragrupo deve ser menor que a distância média intergrupo. A entrada de uma população em um grupo sempre aumenta o valor médio da distância dentro do grupo. Assim definiu-se o nível máximo de distância intragrupo, a partir do maior valor ($\theta$) de distância ($D_{ii'}$) encontrado no conjunto das menores distâncias para cada população (Cruz & Carneiro, 2003).

Neste caso, a distância entre a população i e o grupo formado pelas populações i' e i'' é dada por:

$$d_{(i'i'')i} = d_{i'i} + d_{i''i}$$

Durante o processo de agrupamento há necessidade de avaliar o acréscimo ($d_{(grupo)i}$) no total da diversidade do grupo, no estágio t, pela inclusão de uma população i de maior similaridade ao grupo. A inclusão, ou não, da população i no grupo foi, então, realizada considerando que o acréscimo médio promovido pela sua inclusão no grupo previamente estabelecido foi menor que $\theta$:

Se $\frac{d_{(grupo)i}}{g_o} \leq \theta$, inclui-se a população i no grupo;

Se $\frac{d_{(grupo)i}}{g_o} > \theta$, a população i não é incluído no grupo.

sendo $g_o$ o número de populações que constitui o grupo inicialmente formado.

*(Aplicativo usado: Genes)*

***b – Obtenção do dendrograma***

Aplicou-se o método da ligação média entre grupos (UPGMA), em que são utilizadas as médias aritméticas (não ponderadas) das medidas de distância (Sneath & Sokal, 1973). Como regra geral, a construção do dendrograma é estabelecida pelas populações de maior similaridade. Entretanto, a distância entre uma população i e um grupo, formado pelas populações i' e i'', é dada por:

$$d_{(i'i'')i} = \text{média } (d_{i'i}; d_{i''i}) = \frac{d_{i'i} + d_{i''i}}{2}$$

ou seja, $d_{(i'i'')i}$ é dada pela média do conjunto das distâncias dos pares de populações (i' e i) e (i'' e i).

A distância entre dois grupos é dada por:

$$d_{(ii')(i'' i''')} = \text{ média } (d_{ii''} ; d_{ii'''}; d_{i'i''} d_{i'i'''}) = \frac{d_{ii''} + d_{iii'''} + d_{i'i''} + d_{i'i'''}}{4}$$

ou seja, a distância entre dois grupos formados, respectivamente, pelas populações (i e i') e (i'' e i''') é determinada pela média entre os elementos do conjunto, cujos elementos são distâncias entre pares de populações de grupos (i e i''), (i e i'''), (i' e i''), (i' e i''').

*- Definição do número de grupos*

A determinação do número de grupos foi definida pelo critério estatístico proposto por Mojena (1977), que se baseia no tamanho relativo dos níveis de fusões (distâncias) no dendrograma. A proposta é selecionar o número de grupos no estágio t que, primeiramente, satisfizer a seguinte inequação:

$$\alpha_t > \theta_c$$

em que $\alpha_t$ é o valor de distâncias dos níveis de fusão correspondentes ao estágio t (t = 1,2...g); $\theta_c$ é o valor referencial de corte, dado por:

$$\theta_c = \overline{\alpha} + k\hat{\sigma}_\alpha$$

em que:

$\overline{\alpha}$ e $\hat{\sigma}_{\alpha}$ são a média e o desvio-padrão não-viesado dos valores de $\alpha$, respectivamente, e *k* uma constante. Adotou-se o valor de *k* = 1,25 como regra de parada na definição do número de grupos.

*(Aplicativo usado: Genes)*

***c - Projeção gráfica bidimensional (Cruz & Viana, 1994)***

Neste tipo de projeção as medidas de dissimilaridade são convertidas em escores relativos a duas variáveis X e Y, que, quando representadas em gráficos de dispersão, irão refletir, no espaço bidimensional (2D), as distâncias originalmente obtidas a partir do espaço $\beta$-dimensional ($\beta$ = número de locos, alelos ou genótipos utilizados para obtenção das distâncias). O procedimento consiste em calcular as coordenadas das medidas mais divergentes e, a seguir, daquelas que demonstram, em ordem decrescente, maior diversidade com as populações já consideradas (Cruz, 2006a).

Sendo i e i´ as populações mais divergentes, a próxima população i'' a ser considerada será aquela de maior valor d(ii')i'', dado por:

$$d_{(ii')i''} = d_{ii''} + d_{i'i''}$$

O mesmo critério é usado para a próxima população $\ell$, ou seja, escolhe-se $\ell$, tal que o valor $d_{(ii'i'')\ell}$ seja o maior entre todos. Assim, tem-se:

$$d_{(ii'i'')\ell} = d_{i\ell} + d_{i'\ell} + d_{i''\ell}$$

Assim, a coordenada da população $\ell$ é estimada considerando que:

a) A população i apresenta coordenada (0, 0), estabelecida arbitrariamente;

b) A população i' apresenta coordenada ($d_{ii'}$, 0), também estabelecida arbitrariamente;

c) A população i'' apresenta coordenada ($X_{i''}$, $Y_{i''}$) sendo $X_{i''}$ e $Y_{i''}$ estimadas matematicamente;

d) A população $\ell$ apresenta coordenada ($X_{\ell}, Y_{\ell}$), estimada pelo sistema de equações, que pode ser colocado sob notação matricial Y = $X\beta + \varepsilon$, obtendo-se:

$$Y = \begin{bmatrix} d^2_{\ell i'} - d^2_{\ell i} - d^2_{ii'} \\ \\ d^2_{\ell i''} - d^2_{\ell i} - d^2_{ii''} \end{bmatrix};\ X = -2\begin{bmatrix} X_{i'} & Y_{i'} \\ X_{i''} & Y_{I''} \end{bmatrix};\ \beta = \begin{bmatrix} X_\ell \\ \\ Y_\ell \end{bmatrix}\ \text{e}\quad \varepsilon = \begin{bmatrix} e_1 \\ \\ e_2 \end{bmatrix}$$

Visa-se minimizar a distorção entre a distância original e a distância gráfica.

Para as demais coordenadas são acrescentadas linhas no vetor Y e na matriz X, as quais passam a ter as dimensões $(g_p - 2) \times 1$ e $(g_p - 2) \times 2$, respectivamente, sendo $g_p$ o número de populações até então projetadas.

A solução do sistema é obtida por $X'X\hat{\beta} = X'Y$, de forma que a coordenada estimada para a população $\ell$ apresenta a menor distorção de distância com as demais, cujas coordenadas já foram estabelecidas.

Após a obtenção das coordenadas de cada população, calculou-se a eficiência da projeção gráfica, por meio do coeficiente de correlação simples entre as estimativas das distâncias originais e das distâncias gráficas. Outras estatísticas utilizadas para medir a eficiência da projeção foram:

*Grau de distorção (1-$\alpha$)*

Sendo $\alpha = \dfrac{\sum\limits_{i<}\sum\limits_{i'} d^2_{gii'}}{\sum\limits_{i<}\sum\limits_{i'} d^2_{oii'}}$

em que $d^2_{gii'}$ e $d^2_{oii'}$ são as distâncias gráficas (espaço bidimensional) e originais (espaço $\beta$-dimensional), respectivamente, para todos os pares de populações.

*Estresse (s)*

É dado por:

$$s = 100\sqrt{\frac{\sum_{i<}\sum_{i'}\left(d_{gii'} - d_{oii'}\right)^2}{\sum_{i<}\sum_{i'} d_{oii'}^2}}$$

Segundo Kruskal (1964), valores de estresse entre 10 e 20% representam uma projeção gráfica ruim.

*(Aplicativo usado: Genes)*

#### 3.4.2.1.2. *Comparações entre as medidas de distância*

Além da inspeção visual dos agrupamentos obtidos pelas diferentes metodologias, também estimou-se o coeficiente de correlação simples (r) entre as medidas de distância. O teste de Mantel (Mantel, 1967) foi utilizado para avaliar a significância do coeficiente de correlação entre duas matrizes de distância. Inicialmente estimou-se a correlação entre pares de matrizes. Depois foram obtidas várias outras estimativas de correlações a partir de um conjunto desestruturado de dados no qual se espera obter estimativas próprias da hipótese de nulidade.

O processo de desestruturação do conjunto de dados é realizado a partir das informações de uma das matrizes a serem correlacionadas. Promove-se previamente o embaralhamento por critério de permutação aleatória, ou seja, linhas e colunas da matriz são permutadas *s* vezes e, consequentemente, *s* estimativas de correlações são obtidas. O teste de hipótese é, então, feito com base nestas *s* estimativas obtidas. O valor crítico de significância foi obtido no vetor de ordem crescente ($\rho_{(t)}$, t = 1, 2, ..,.*s*) de estimativas de correlação. Foram realizadas 1000 permutações e adotou-se um nível de 5% de probabilidade para o teste, de modo que as 50 estimativas dos extremos (25 maiores e 25 menores estimativas de correlação) foram definidas como os limites da área de rejeição. O valor de r será significativo quando inferior ao valor correspondente à ordem 26ª ou superior ao valor correspondente à ordem 974ª deste conjunto de estimativas.

### 3.4.2.2. Diversidade genética entre e dentro de populações de retrocruzamento

O estudo da variabilidade entre e dentro de populações de retrocruzamento com seu respectivo genitor recorrente, ao longo das gerações foi realizado por intermédio das estatísticas F (ou índices de fixação) definidas por Wright (1951, 1978). Basicamente, dois tipos de análises foram realizadas para estimar os parâmetros genéticos $F_{IS}$, $F_{IT}$ e $F_{ST}$.

a) Estatísticas H de Nei (1973; 1977);

b) análises de variância: de freqüências alélicas (Cockerham, 1969; 1973; Weir & Cockerham, 1984) e molecular (Excoffier et al., 1992; Peakall et al., 1995)

As análises consideraram as fontes de variação populações e indivíduos dentro de populações. No entanto, as análises de variância ainda contabilizaram a fonte de variação entre alelos dentro de indivíduos, conforme apresentado a seguir. Foram realizadas as seguintes comparações (contrastes):

1) $P_1$ x $RC_{31(mp)}$;
2) $P_1$ x $RC_{34(mp)}$;
3) $P_3$ x $RC_{31(mp)}$;
4) $P_3$ x $RC_{34(mp)}$;

#### 3.4.2.2.1. Estatísticas H de Nei

Nei (1973; 1977) mostrou que as estatísticas F de Wright podem ser definidas como razões entre estatísticas H (heterozigosidades), ao invés de correlações entre unidades gaméticas. Considerando apenas um loco, os estimadores dos índices $F_{IS}$, $F_{IT}$ e $F_{ST}$ são:

$\hat{F}_{IS} = \frac{\hat{h}_S - \hat{h}_o}{\hat{h}_S}$ (índice de fixação dentro de populações);

$\hat{F}_{IT} = \frac{\hat{h}_T - \hat{h}_o}{\hat{h}_T}$ (índice de fixação total);

$\hat{F}_{ST} = \hat{G}_{ST} = \frac{\hat{h}_T - \hat{h}_S}{\hat{h}_T} = \frac{\hat{D}_{ST}}{\hat{h}_T}$ (divergência entre as populações).

em que os indexadores I, S e T representam indivíduos, populações e total da população, respectivamente;

$\hat{h}_T = 1 - \sum_{k=1}^{a_j} \hat{\bar{p}}_k^2$, que corresponde ao estimador (viesado) da heterozigosidade esperada total das populações, sendo $\hat{\bar{p}}_k = \sum_{i=1}^{g} w_i \hat{p}_{ijK}$;

$\hat{h}_S = 1 - \sum_{i=1}^{g} w_i \sum_{k=1}^{a_j} \hat{p}_{ijk}^2$, que corresponde ao estimador (viesado) da heterozigosidade esperada dentro das populações;

$\hat{h}_o = \sum_{i=1}^{g} w_i \sum_{k \le k'}^{a_j} \hat{P}_{ijkk'}$, que corresponde ao estimador da heterozigosidade observada dentro das populações, em que $w_i = 1/g$ é o peso da i-ésima população.

Estimadores não viesados de $h_S$ e $h_T$, podem ser obtidos com a correção para erros de amostragem (Nei & Kumar, 2000). Se a média harmônica ($\tilde{n}$) do tamanho amostral de cada população for > 30, seu efeito nas estimativas de $\hat{h}_S$ e $\hat{h}_T$ será negligenciável, logo a correção para tamanho da amostra é necessária apenas quando $\tilde{n}$ for < 30.

Para L locos sob estudo as heterozigosidades são calculadas através da média aritmética dos estimadores de $h_T$, $h_S$ e $h_O$ obtidos em cada loco.

A magnitude relativa da diferenciação gênica entre populações pode ser medida por:

$$\overline{G}_{ST} = \frac{\overline{D}_{ST}}{\overline{H}_{ST}}$$

Ela expressa a proporção da diversidade total explicada por diferenças entre populações e matematicamente equivale à estatística $F_{ST}$ de Wright, variando de 0 a 1.

*(Aplicativos usados: Genes e POPGENE)*

### 3.4.2.2.2. Análises de variância de freqüências alélicas

A análise de variância proposta por Cockerham (1969; 1973) e, generalizada por Weir & Cockheram (1984), para organismos diplóides, permite estimar o grau de parentesco F ($F_{IT}$) entre alelos dentro de indivíduos de todas as populações, assim como a coancestralidade $\theta$ ($F_{ST}$) entre alelos de diferentes indivíduos na mesma população (Weir, 1996). Inicialmente, considere apenas o alelo k do loco j. A análise de variância é resumida na Tabela 2.

Pelo método dos momentos, as estimativas não viesadas dos componentes de variância entre populações ($\sigma_P^2$), entre indivíduos dentro das populações ($\sigma_I^2$) e de alelos dentro de indivíduos ($\sigma_G^2$), são obtidas como funções dos quadrados médios (QM) observados para populações (QMP), indivíduos (QMI) e alelos (QMG), respectivamente, da seguinte maneira:

$$\sigma_G^2 = \text{QMG}$$

$$\sigma_I^2 = \frac{1}{2}(\text{QMI} - \text{QMG})$$

$$\sigma_P^2 = \frac{1}{2n_a}(\text{QMI} - \text{QMG})$$

Tabela 3. Resumo da análise de variância para dados genotípicos em populações aleatórias

| Fonte de variação | Graus de liberdade | SQ | QM | E (QM)* |
|---|---|---|---|---|
| Entre populações | g -1 | $SQP = 2\sum_{i=1}^{g} n_{ijk}(\hat{p}_{ijk} - \bar{\hat{p}}_{ijk})^2$ | $QMP = \frac{SQP}{s-1}$ | $\sigma_G^2 + 2\sigma_I^2 + 2n_a\sigma_P^2$ |
| Indivíduos dentro de populações | $\sum_{i=1}^{g}(n_{ijk} - 1)$ | $SQI = \sum_{i=1}^{g} n_{ijk}(\hat{p}_{ijk} + \hat{P}_{ijkk} - 2\hat{p}_{ijk}^2)$ | $QMI = \frac{SQI}{\sum_{i=1}^{g}(n_{ijk} - 1)}$ | $\sigma_G^2 + 2\sigma_I^2$ |
| Alelos dentro de indivíduos | $\sum_{i=1}^{g} n_{ijk}$ | $SQG = \sum_{i=1}^{g} n_{ijk}(\hat{p}_{ijk} - \hat{P}_{ijkk})$ | $QMG = \frac{SQG}{\sum_{i=1}^{g} n_{ijk}}$ | $\sigma_G^2$ |
| Total | $2\sum_{i=1}^{g} n_{ijk} - 1$ | SQT = SQP + SQI + SQG | | |

* $n_a = \frac{1}{g-1}\left(\sum_{i=1}^{g} n_{ijk} - \frac{\sum_{i=1}^{g} n_{ijk}^2}{\sum_{i=1}^{g} n_{ijk}}\right)$

em que $n_{ijk}$ é o número do alelo k do loco j da população i.

Assim, as medidas de diferenciação podem ser estimadas como:

$$\hat{F} = \frac{\hat{\sigma}_P^2 + \hat{\sigma}_I^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2} = 1 - \frac{2n_c QMG}{QMP + (n_c - 1)QMI + n_c QMG} = 1 - \frac{S_D}{S_B}$$

$$\hat{\theta} = \frac{\hat{\sigma}_P^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2} = 1 - \frac{QMP - QMI}{QMP + (n_c - 1)QMI + n_c QMG} = 1 - \frac{S_A}{S_B}$$

Sendo $a_j$ alelos avaliados em cada loco, cada um apresentará uma estimativa dos parâmetros F de Wright. Para combinar as estimativas sobre todos os alelos de um loco j, o numerador e o denominador são combinados separadamente, da seguinte maneira:

$$\hat{F}_j = 1 - \frac{\sum_{k=1}^{a_j} S_{Djk}}{\sum_{k=1}^{a_j} S_{Bjk}}$$

$$\hat{\theta}_j = 1 - \frac{\sum_{k=1}^{a_j} S_{Ajk}}{\sum_{k=1}^{a_j} S_{Bjk}}$$

As estimativas combinadas dos parâmetros F e θ para todos os alelos considerando todos os locos (estimativas globais) são dadas por:

$$\hat{F} = 1 - \frac{\sum_{j=1}^{l}\sum_{k=1}^{a_j} S_{Djk}}{\sum_{j=1}^{l}\sum_{k=1}^{a_j} S_{Bjk}}$$

$$\hat{\theta} = 1 - \frac{\sum_{j=1}^{l}\sum_{k=1}^{a_j} S_{Ajk}}{\sum_{j=1}^{l}\sum_{k=1}^{a_j} S_{Bjk}}$$

O grau de endogamia dentro das populações (*f* = $F_{IS}$) pode ser obtido, a partir das estimativas globais dos parâmetros F e $\theta$, dado por:

$$\hat{f} = \frac{\hat{F} - \hat{\theta}}{1 - \hat{\theta}}$$

*(Aplicativos computacionais: GDA, PowerMarker, TFPGA)*

Os intervalos de confiança a 95% de probabilidade foram obtidos para se verificar a significância das estimativas dos parâmetros F. Para isso foram realizadas 1000 simulações *bootstrap* sobre os locos.

*(Aplicativos computacionais: GDA, PowerMarker, TFPGA)*

### 3.4.2.2.3. Análise de variância molecular (AMOVA)

Como referido por Michalakis & Excoffier (1996), a análise de variância molecular (AMOVA) já foi estendida a dados codominantes (Peakall et al., 1995; Maguire et al., 2002). É possível executá-la para cada loco. Para a sua realização é necessário, primeiramente, computar todas as distâncias (Euclidianas) entre pares de indivíduos, independente da população a qual eles pertençam. Usualmente utiliza-se o quadrado da distância Euclidiana. Para dados genotípicos, Peakall et al. (1995) e Smouse & Peakall (1999) propuseram a seguinte medida:

$$d^2_{tt'(j)} = \frac{1}{2}\sum_{k=1}^{a_j}(y_{tjk} - y_{t'jk})^2$$

em que $y_{tjk}$ refere-se ao número de alelos k que o indivíduo de genótipo t e t' possui no loco j. Aqui não há pesos atribuídos a alelos.

Para obter a distância considerando múltiplos locos, simplesmente soma-se as distâncias de todos os locos. Para os indivíduos t e t', tem-se:

$$d^2_{tt'(j)} = \sum_{j=1}^{L} d^2_{tt'(j)}$$

*AMOVA com dois níveis hierárquicos (Excoffier et al., 1992)*

Após definida a matriz de distâncias entre todos os indivíduos ($D^2$), particionou-se a mesma, agrupando as informações de pares de genótipos dentro de cada população. Soma de quadrados dos desvios total ($SQD_T$), entre populações ($SQD_{EP}$), entre indivíduos dentro populações ($SQD_{EI/DP}$) e dentro de indivíduos ($SQD_{DI}$) foram obtidas. Os componentes de variância, obtidos a partir das esperanças de quadrado médio (EQM), permitiram a estimação das estatísticas F. A análise de variância está resumida na Tabela 4.

Tabela 4. Esquema da AMOVA com dados agrupados em dois níveis hierárquicos

| Fonte de Variação | GL | SQ | E(QM) |
|---|---|---|---|
| Entre Populações | g – 1 | $SQD_{EP}$ | $n\sigma_a^2 + 2\sigma_b^2 + \sigma_c^2$ |
| Entre indivíduos/Populações | N-g | $SQD_{EI/DP}$ | $2\sigma_b^2 + \sigma_c^2$ |
| Dentro de indivíduos | N | $SQD_{DI}$ | $\sigma_c^2$ |
| Total | 2N – 1 | $SQD_T$ | $\sigma_T^2$ |

Em que n é dado por:

$$n = \frac{2N - \sum_{i=1}^{g} \frac{2N_i^2}{N}}{g-1}$$

em que N é o número total de indivíduos amostrados e $N_i$ é tamanho amostral da população i.

As estatísticas F foram obtidas da seguinte forma:

$$F_{ST} = \Phi_{ST} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_T^2}\ ;\ F_{IT} = \Phi_{IT} = \frac{\hat{\sigma}_a^2 + \hat{\sigma}_b^2}{\hat{\sigma}_T^2} \text{ e } F_{IS} = \Phi_{IS} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_c^2}$$

A significância de $\sigma_a^2$ e $F_{ST}$ foi testada permutando-se os genótipos dos indivíduos entre populações.

A significância de $\sigma_b^2$ e $F_{IS}$ foi testada permutando-se os alelos entre indivíduos dentro de populações.

A significância de $\sigma_c^2$ e $F_{IT}$ foi testada permutando-se os alelos entre indivíduos entre populações.

Esta análise foi realizada sobre todos os locos.

*(Aplicativo usado: Arlequin)*

### 3.5. Comparações entre os aplicativos computacionais

Avaliou-se as particularidades e funcionalidades dos aplicativos computacionais (seção 3.3) em relação as análises biométricas do presente trabalho.

# 4. RESULTADOS E DISCUSSÃO

## 4.1. Diversidade genética em nível intrapopulacional

### 4.1.1. Medidas descritivas

O progresso genético obtido com os ciclos de seleção está diretamente relacionado à variabilidade presente no *pool* gênico e a qualidade da contribuição gênica dos genitores (Dreisigacker et al., 2004). O conhecimento dos níveis de distribuição da diversidade genética é também um pré-requisito para o estabelecimento eficiente do manejo da conservação (Gao & Hong, 2000).

Medidas descritivas têm sido utilizadas rotineiramente em investigações sobre a diversidade genética dentro de populações naturais (Zucchi et al., 2003; Melo Júnior et al., 2004). Aplicam-se também em estudos comparativos entre as várias técnicas de marcadores moleculares quanto à capacidade informativa e discriminatória na identificação genotípica e na análise da diversidade de germoplasmas (Pejic et al., 1998; Belaj et al., 2003).

No presente trabalho foram totalizados 79 alelos pertencentes aos 20 locos codominantes considerando as 42 populações simuladas. O processo de simulação não permitiu a obtenção de alelos exclusivos, ou seja, aqueles presentes unicamente em uma determinada população. O processo de simulação gerou locos monomórficos nas populações $P_1$ e $P_1a_1$ (locos 1, 6, 12, 15 e 20); $P_2$, $P_2a_1$ a $P_2a_4$ e $P_2s_1$ a $P_2s_5$ (locos 15, 17 e 20); $P_3$ e $P_3a_1$ (locos 5 e 11) e $H_{12pp}$ e $H_{12mp}$ (locos 15 e 20).

Na Tabela 5 estão apresentadas as estimativas do número médio de alelos por loco ($\overline{A}_j$), número total de alelos na população ($A_t$), número de alelos raros ($A_r$), número efetivo de alelos total ($A_e$), média do número efetivo de alelos por loco ($\overline{A}_e$) e proporção de alelos na população ($P_a$) em relação aos 20 locos simulados nas 42 populações.

Tabela 5. Estimativas do número médio de alelos por loco ($\overline{A}_j$), número total de alelos na população ($A_t$), número de alelos raros ($A_r$), número efetivo de alelos total ($A_e$), média do número efetivo de alelos por loco ($\overline{A}_e$) e proporção de alelos na população ($P_a$), obtidas em 20 locos codominantes multialélicos para as 42 populações de melhoramento simuladas

| População | $\overline{A}_j$ [1,2,3,4,5,6,7] | $A_t$ [4] | $A_r$ [§4] | $A_e$ [5] | $\overline{A}_e$ [5] | $P_a$ [4] |
|---|---|---|---|---|---|---|
| $P_1$ | 2,45 | 49,00 | 10,00 | 32,39 | 1,62 | 0,62 |
| $P_2$ | 3,00 | 60,00 | 13,00 | 32,85 | 1,64 | 0,76 |
| $P_3$ | 3,00 | 60,00 | 12,00 | 39,14 | 1,96 | 0,76 |
| $P_1a_1$ | 2,45 | 49,00 | 10,00 | 31,82 | 1,59 | 0,62 |
| $P_2a_1$ | 2,95 | 59,00 | 12,00 | 31,91 | 1,60 | 0,75 |
| $P_2a_2$ | 2,95 | 59,00 | 13,00 | 31,70 | 1,59 | 0,75 |
| $P_2a_3$ | 2,95 | 59,00 | 12,00 | 32,06 | 1,60 | 0,75 |
| $P_2a_4$ | 2,90 | 58,00 | 13,00 | 32,28 | 1,61 | 0,73 |
| $P_3a_1$ | 3,00 | 60,00 | 11,00 | 39,01 | 1,95 | 0,76 |
| $P_2s_1$ | 3,00 | 60,00 | 13,00 | 33,38 | 1,67 | 0,76 |
| $P_2s_2$ | 2,95 | 59,00 | 12,00 | 33,25 | 1,66 | 0,75 |
| $P_2s_3$ | 2,95 | 59,00 | 12,00 | 33,43 | 1,67 | 0,75 |
| $P_2s_4$ | 2,95 | 59,00 | 12,00 | 33,43 | 1,67 | 0,75 |
| $P_2s_5$ | 2,95 | 59,00 | 12,00 | 33,47 | 1,67 | 0,75 |
| $H_{12pp}$ | 3,35 | 67,00 | 21,00 | 37,80 | 1,89 | 0,85 |
| $H_{12mp}$ | 3,35 | 67,00 | 20,00 | 38,00 | 1,90 | 0,85 |
| $H_{13pp}$ | 3,35 | 67,00 | 17,00 | 40,56 | 2,03 | 0,85 |
| $H_{13mp}$ | 3,30 | 66,00 | 16,00 | 40,72 | 2,04 | 0,84 |
| $H_{23pp}$ | 3,80 | 76,00 | 24,00 | 40,13 | 2,01 | 0,96 |
| $F_{2(pp)}$ | 3,75 | 75,00 | 23,00 | 40,14 | 2,01 | 0,95 |
| $F_{3(pp)}$ | 3,75 | 75,00 | 23,00 | 40,01 | 2,00 | 0,95 |
| $F_{4(pp)}$ | 3,75 | 75,00 | 23,00 | 39,85 | 1,99 | 0,95 |
| $H_{23(mp)}$ | 3,80 | 76,00 | 23,00 | 40,87 | 2,04 | 0,96 |
| $F_{2(mp)}$ | 3,75 | 75,00 | 22,00 | 40,69 | 2,03 | 0,95 |
| $F_{3(mp)}$ | 3,75 | 75,00 | 20,00 | 40,83 | 2,04 | 0,95 |
| $F_{4(mp)}$ | 3,75 | 75,00 | 20,00 | 40,83 | 2,04 | 0,95 |
| $RC_{11pp}$ | 3,30 | 66,00 | 16,00 | 37,57 | 1,88 | 0,84 |
| $RC_{12pp}$ | 3,15 | 63,00 | 18,00 | 34,88 | 1,74 | 0,80 |
| $RC_{13pp}$ | 3,15 | 63,00 | 20,00 | 33,56 | 1,68 | 0,80 |
| $RC_{14pp}$ | 3,00 | 60,00 | 19,00 | 33,06 | 1,65 | 0,76 |
| $RC_{11mp}$ | 3,35 | 67,00 | 19,00 | 37,32 | 1,87 | 0,85 |
| $RC_{12mp}$ | 3,20 | 64,00 | 18,00 | 35,02 | 1,75 | 0,81 |
| $RC_{13mp}$ | 3,05 | 61,00 | 19,00 | 33,13 | 1,66 | 0,77 |
| $RC_{14mp}$ | 3,05 | 61,00 | 20,00 | 32,79 | 1,64 | 0,77 |
| $RC_{31pp}$ | 3,35 | 67,00 | 17,00 | 41,10 | 2,05 | 0,85 |
| $RC_{32pp}$ | 3,30 | 66,00 | 17,00 | 40,31 | 2,02 | 0,84 |
| $RC_{33pp}$ | 3,30 | 66,00 | 18,00 | 39,74 | 1,99 | 0,84 |
| $RC_{34pp}$ | 3,30 | 66,00 | 18,00 | 39,40 | 1,97 | 0,84 |
| $RC_{31mp}$ | 3,30 | 66,00 | 15,00 | 40,71 | 2,04 | 0,84 |
| $RC_{32mp}$ | 3,30 | 66,00 | 16,00 | 39,71 | 1,99 | 0,84 |
| $RC_{33mp}$ | 3,30 | 66,00 | 18,00 | 39,35 | 1,97 | 0,84 |
| $RC_{34mp}$ | 3,25 | 65,00 | 15,00 | 39,33 | 1,97 | 0,82 |

Programas utilizados: [1]Arlequin; [2]GDA; [3]GENEPOP; [4]GENES; [5]POPGENE; [6]PowerMarker; [7]TFPGA. [§]Alelos com freqüência menor que 0,05.

Para $\overline{A}_j$ foram considerados locos polimórficos e monomórficos, mas alguns autores têm considerado apenas o número de alelos nos locos polimórficos ($\overline{A}_p$) o que não deixa de ser uma alternativa plausível (Gimenes & Lopes, 2000). Os valores de $\overline{A}_j$, $A_t$, $A_r$, $P_a$, $A_e$ e $\overline{A}_e$ variaram de 2,45 a 3,80; 49 a 76; 10 a 24; 0,62 a 0,86; 31,70 a 41,10 e 1,59 a 2,05, respectivamente.

Dentre as populações base, $P_2$ e $P_3$ foram as que exibiram os maiores valores para $\overline{A}_j$, $A_t$, $A_r$, $A_e$, $\overline{A}_e$ e $P_a$. As primeiras gerações de acasalamento ao acaso ($P_1a_1$, $P_2a_1$ e $P_3a_1$) seguiram a padrões semelhantes da quantidade alélica de suas respectivas populações base.

As gerações avançadas de acasalamento ao acaso ($P_2a_1$ a $P_2a_4$) e de autofecundação ($P_2s_1$ a $P_2s_5$) praticamente não diferiram uma das outras e nem de $P_2$, em relação às estimativas dos índices da Tabela 5. O mesmo aspecto é referido em relação às populações híbridas $H_{23pp}$ e $H_{23mp}$ e as respectivas gerações segregante $F_n$. As ligeiras alterações no número de alelos raros de uma geração para outra, nas populações em geral, são atribuídas ao processo de amostragem. Um princípio importante na autofecundação é que o processo de autofecundação ou endogâmico, por si só, não altera freqüências gênicas. Para freqüências alélicas constantes sobre endogamia, pressupõe-se que todos os genótipos devam ter probabilidades iguais de sobrevivência e reprodução, a não ser que a seleção esteja atuando. Então, as freqüências alélicas podem ser alteradas com a endogamia, ou mesmo em qualquer outro sistema de acasalamento (Hartl & Clark, 1997).

Entre as populações híbridas destacaram-se aquelas provenientes do cruzamento $P_2$ x $P_3$ ($H_{23pp}$ e $H_{23mp}$), para as estimativas de $\overline{A}_j$, $A_t$, $A_r$, e $P_a$. Este fato refletiu a importância da contribuição parental na diversidade gênica de híbridos, uma vez que $P_2$ e $P_3$ exibiram maiores valores das medidas descritivas da Tabela 5 em relação a $P_1$.

As populações de retrocruzamento proporcionaram maiores valores para todos os índices da Tabela 5 quando comparados aos das respectivas populações base $P_1$ e $P_3$, justamente por se tratar de um processo de

hibridação. A diminuição dos valores de $\overline{A}_j$, $A_t$, $P_a$, $A_e$ e $\overline{A}_e$ com o avanço das gerações é novamente atribuída a amostragem. No entanto, nas populações de retrocruzamento cujo genitor recorrente foi $P_1$, percebeu-se uma queda mais acentuada nos valores de $\overline{A}_j$, $A_t$, $P_a$, $A_e$ e $\overline{A}_e$ a cada geração, em relação aquelas cujo genitor recorrente foi $P_3$. Assim, constata-se que a variabilidade genética do genitor recorrente é aspecto importante na manutenção da quantidade alélica ao longo das gerações de retrocruzamento. Ressalta-se ainda que no processo de simulação, as gerações de retrocruzamento não passaram por seleção, ou seja, descarte de genótipos indesejáveis.

O número efetivo de alelos por loco ($a_e$) é um índice de diversidade genética intrapopulacional que merece destaque, quando interpretado juntamente com outros índices. Ele representa o número de alelos com freqüências iguais em um loco e que, consequentemente, contribuem de forma efetiva para a diversidade gênica (ou heterozigosidade esperada, $\hat{h}_e$).

O significado de $\overline{A}_e$ ou $A_e$ pôde ser ilustrado pelos locos 5 e 10 da população $F_{4(mp)}$ com 5 alelos em cada loco (Tabela 6). No loco 5 as freqüências alélicas foram 0,8825, 0,0825, 0,0100, 0,0050 e 0,0200. Assim o valor de $\hat{h}_e$ (estimador viesado) e $a_e$ foram iguais a 0,2144 e 1,2720, respectivamente. No loco 10 as freqüências alélicas foram 0,4175, 0,2800, 0,2450, 0,0150 e 0,0425 e os valores de $\hat{h}_e$ e $a_e$ foram iguais a 0,2687 e 3,1770, respectivamente. É sabido que a diversidade gênica num loco é máxima quando as freqüências alélicas são iguais (Liu, 1997). Portanto, nestes dois locos a estimativa de $h_e$ máxima seria igual a 0,80, com freqüências alélicas iguais a 0,20. Nenhum dos locos atingiu este valor de heterozigosidade esperada máxima, porém o loco 10 apresentou maiores valores de $\hat{h}_e$ e $a_e$. Veja que as freqüências alélicas do loco 10 estavam melhor distribuídas entre seus alelos, ou melhor, mais próximas do valor 0,20. Este é um indicativo de que o número efetivo de alelos pode ser usado como um corolário para a heterozigosidade esperada (ou diversidade gênica). Assim, o índice $A_e$ permite comparar populações cujo número e distribuição de alelos difere drasticamente.

Tabela 6. Estimativas de freqüências alélicas, número de alelos raros ($A_r$), número efetivo de alelos ($a_e$) e diversidade gênica ($\hat{h}_e$ #) para os locos 5 e 10 na população $F_{4(mp)}$

| | Loco | |
|---|---|---|
| Freqüência | 5 | 10 |
| alelo 1 | 0,8825 | 0,4175 |
| alelo 2 | 0,0825 | 0,2800 |
| alelo 3 | 0,0100 | 0,2450 |
| alelo 4 | 0,0050 | 0,0150 |
| alelo 5 | 0,0200 | 0,0425 |
| $A_r$ | 3,0000 | 2,0000 |
| $a_e$ | 1,2720 | 3,1770 |
| $\hat{h}_e$ | 0,2144 | 0,2687 |

#Estimador viesado

Outro aspecto que chama atenção é o número de alelos raros. No loco 5 são três alelos raros (com freqüências de 0,0100, 0,005 e 0,0200), enquanto que o loco 10 possui apenas dois (com freqüências de 0,0150 e 0,0425). Se for calculada a diferença entre o número de alelos do loco ($A_j$) e o número efetivo de alelos ($a_e$), verifica-se que esta diferença é maior para o loco 5. Portanto, se as estimativas de $\overline{A}_e$ forem muito menores que $\overline{A}_j$, sugere-se a presença de vários alelos de baixa freqüência (Belaj et al., 2003). Ter-se-á $A_e = A_j$ quando o loco atingir o valor máximo de $\hat{h}_e$ e, consequentemente, $a_e$ será igual ou próxima a unidade (1), a exemplo do que ocorreu no loco 5, em que existia um único alelo de freqüência predominante.

Na verdade, $A_e$ é uma função não linear de $H_e$ (Figura 1), que recaí sobre a inequação de Jensen, ou seja, a esperança de uma função é diferente da função das esperanças para curvas não lineares (Hardy et al., 1988). Portanto, $\overline{A}_e$ deve ser calculada sobre a média das estimativas para cada loco ($a_e$), ao invés de ser calculada a média de $\hat{h}_e$ sobre todos locos e, então, estimar-se $\overline{A}_e$. Por exemplo, na população $P_1$ a estimativa de $\overline{A}_e$ foi de 1,62 sobre a média de $a_e$, enquanto que sobre a média de $\hat{h}_e$, o valor foi de 1,44, conforme esperado para curvas côncavas segundo a inequação de Jensen.

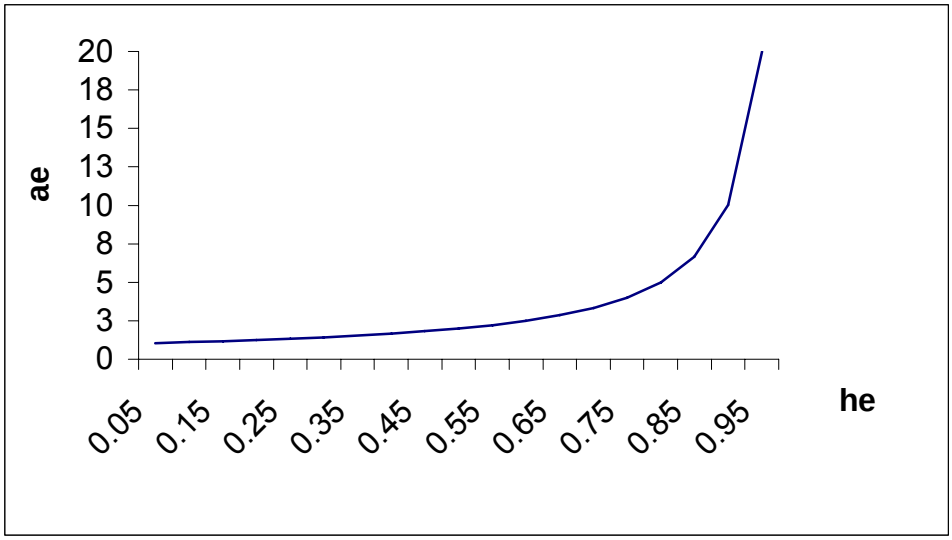


Figura 1. Número efetivo de alelos por loco ($a_e$) como função da heterozigosidade esperada máxima ($h_e$).

O principal objetivo da genética de populações é estudar a geração e a manutenção do polimorfismo genético e compreender os mecanismos de evolução ao nível populacional (Nei & Kumar, 2000). O polimorfismo genético é definido como a existência de dois ou mais alelos com freqüências relativas substanciais (comumente maiores que 1 ou 5%) na população. A origem do polimorfismo genético em um loco, assim como toda variabilidade genética, está associada a processos mutacionais, tais como substituição nucleotídica, inserções e deleções, conversão gênica e recombinação inter-alélica (Nei & Kumar, 2000). Locos de microssatélites freqüentemente apresentam proporção de polimorfismo igual a 100% e elevado número de alelos por loco (Caixeta et al., 2006).

As proporções de locos polimórficos (PLP), bem como o número médio de alelos por loco polimórfico ($\overline{A}_p$) estão dispostos na Tabela 7, conforme os três critérios definidos por Cole (2003). As populações $P_2$ e $P_3$ apresentaram maior polimorfismo genético em relação a $P_1$. Estas medidas de polimorfismo genético corroboram a menor variação genética na população base $P_1$, indicada pelos índices da Tabela 5. A informação dada por $\overline{A}_p$ é relevante, pois um loco pode apresentar polimorfismo, porém com poucos alelos, o que é comum em marcadores isoenzimáticos. Interessante

Tabela 7. Estimativas do percentual de locos polimórficos (PLP) e do número médio de alelos por loco polimórfico ($\overline{A}_p$) obtidas a partir de 20 locos codominantes e multialélicos para 42 populações de melhoramento geradas via simulação

| Populações | PLP[# 4] | $\overline{A}_p$ [# 4] | PLP[## 1,2,4,5,7] | $\overline{A}_p$ [## 2,4] | PLP[### 2,4,7] | $\overline{A}_p$ [### 4] |
|---|---|---|---|---|---|---|
| $P_1$ | 75,00 | 2,45 | 75,00 | 2,93 | 65,00 | 3,08 |
| $P_2$ | 85,00 | 3,00 | 85,00 | 3,35 | 80,00 | 3,31 |
| $P_3$ | 90,00 | 3,00 | 90,00 | 3,22 | 90,00 | 3,22 |
| $P_1a_1$ | 75,00 | 2,45 | 75,00 | 2,93 | 65,00 | 3,08 |
| $P_2a_1$ | 85,00 | 2,95 | 85,00 | 3,29 | 80,00 | 3,25 |
| $P_2a_2$ | 85,00 | 2,95 | 85,00 | 3,29 | 80,00 | 3,25 |
| $P_2a_3$ | 85,00 | 2,95 | 85,00 | 3,29 | 80,00 | 3,25 |
| $P_2a_4$ | 85,00 | 2,90 | 85,00 | 3,24 | 80,00 | 3,19 |
| $P_3a_1$ | 90,00 | 3,00 | 90,00 | 3,22 | 90,00 | 3,22 |
| $P_2s_1$ | 85,00 | 3,00 | 85,00 | 3,35 | 80,00 | 3,31 |
| $P_2s_2$ | 85,00 | 2,95 | 85,00 | 3,29 | 80,00 | 3,31 |
| $P_2s_3$ | 85,00 | 2,95 | 85,00 | 3,29 | 80,00 | 3,31 |
| $P_2s_4$ | 85,00 | 2,95 | 85,00 | 3,29 | 80,00 | 3,31 |
| $P_2s_5$ | 85,00 | 2,95 | 85,00 | 3,29 | 80,00 | 3,31 |
| $H_{12pp}$ | 90,00 | 3,35 | 90,00 | 3,61 | 85,00 | 3,59 |
| $H_{12mp}$ | 90,00 | 3,35 | 90,00 | 3,61 | 85,00 | 3,59 |
| $H_{13pp}$ | 100,00 | 3,35 | 100,00 | 3,35 | 100,00 | 3,35 |
| $H_{13mp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 100,00 | 3,30 |
| $H_{23pp}$ | 100,00 | 3,80 | 100,00 | 3,80 | 100,00 | 3,80 |
| $F_{2(pp)}$ | 100,00 | 3,75 | 100,00 | 3,75 | 100,00 | 3,75 |
| $F_{3(pp)}$ | 100,00 | 3,75 | 100,00 | 3,75 | 100,00 | 3,75 |
| $F_{4(pp)}$ | 100,00 | 3,75 | 100,00 | 3,75 | 100,00 | 3,75 |
| $H_{23(mp)}$ | 100,00 | 3,80 | 100,00 | 3,80 | 100,00 | 3,80 |
| $F_{2(mp)}$ | 100,00 | 3,75 | 100,00 | 3,75 | 100,00 | 3,75 |
| $F_{3(mp)}$ | 100,00 | 3,75 | 100,00 | 3,75 | 100,00 | 3,75 |
| $F_{4(mp)}$ | 100,00 | 3,75 | 100,00 | 3,75 | 100,00 | 3,75 |
| $RC_{11pp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 95,00 | 3,32 |
| $RC_{12pp}$ | 100,00 | 3,15 | 100,00 | 3,15 | 80,00 | 3,31 |
| $RC_{13pp}$ | 100,00 | 3,15 | 95,00 | 3,16 | 75,00 | 3,40 |
| $RC_{14pp}$ | 100,00 | 3,00 | 95,00 | 3,00 | 70,00 | 3,29 |
| $RC_{11mp}$ | 100,00 | 3,35 | 100,00 | 3,35 | 95,00 | 3,32 |
| $RC_{12mp}$ | 100,00 | 3,20 | 100,00 | 3,20 | 95,00 | 3,16 |
| $RC_{13mp}$ | 100,00 | 3,05 | 95,00 | 3,05 | 75,00 | 3,27 |
| $RC_{14mp}$ | 100,00 | 3,05 | 90,00 | 3,11 | 70,00 | 3,36 |
| $RC_{31pp}$ | 100,00 | 3,35 | 100,00 | 3,35 | 100,00 | 3,35 |
| $RC_{32pp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 95,00 | 3,32 |
| $RC_{33pp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 90,00 | 3,33 |
| $RC_{34pp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 90,00 | 3,33 |
| $RC_{31mp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 100,00 | 3,30 |
| $RC_{32mp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 95,00 | 3,32 |
| $RC_{33mp}$ | 100,00 | 3,30 | 100,00 | 3,30 | 95,00 | 3,32 |
| $RC_{34mp}$ | 100,00 | 3,25 | 95,00 | 3,26 | 90,00 | 3,28 |

[#]Loco exibindo polimorfismo em pelo menos um indivíduo da amostra; [##]Loco em que o alelo mais comum tem freqüência menor que 99%; [###]Loco em que o alelo mais comum tem freqüência menor que 95%.

Programas utilizados: [1]Arlequin; [2]GDA; [3]GENEPOP; [4]GENES; [5]POPGENE; [6]PowerMarker; [7]TFPGA.

é que tanto gerações de acasalamento ao acaso quanto de autofecundação da $P_2$ mantiveram este polimorfismo genético. Segundo Hartl & Clark (1997), a proporção do polimorfismo gênico de espécies autógamas é comparável a proporção encontrada em espécies alógamas. Isto porque a autofecundação não elimina a variação genética, ela simplesmente a reorganiza em genótipos homozigotos. Os autores ainda exaltam que espécies autógamas contem menos alelos (recessivos) deletérios do que as alógamas, provavelmente porque o aumento da homozigosidade permite que genes indesejáveis sejam eliminados da população pela seleção natural.

Novamente a superioridade dos híbridos $H_{23pp}$ e $H_{23mp}$ foi destacada em relação aos demais, com 100% de locos polimórficos, ambos com 3,80 de alelos por loco polimórfico. O propósito da hibridação, ou de cruzamentos controlados, é reunir alelos favoráveis presentes em ambos os genitores. Se eles forem materiais superiores e exibirem diversidade genética, manifesta-se a heterose e, populações segregantes avançadas manifestarão ampla variabilidade a ser explorada por técnicas seletivas (Cruz, 2005). Este fato foi verificado nas respectivas gerações segregantes $F_n$ ($F_{2(pp)}$ a $F_{4(pp)}$ e $F_{2(mp)}$ a $F_{4(mp)}$), cujo alto polimorfismo genético, comparado às populações base, foi mantido (Tabela 7).

Populações de retrocruzamento apresentaram valores de PLP e $\bar{A}_p$ superiores às populações base $P_1$ e $P_3$. Houve uma tendência de queda nos valores de PLP e $\bar{A}_p$ com os avanços das gerações, em virtude da proximidade destas populações com o respectivo genitor recorrente. O aumento do polimorfismo genético nestas populações sugere a exploração de variabilidade genética, fazendo do método de retrocruzamento uma estratégia eficiente no desenvolvimento de populações segregantes promissoras, base para seleção de genótipos superiores (Lorencetti et al., 2006).

A Tabela 8 apresenta as estimativas médias do número de genótipos por loco ($\bar{g}_j$), conteúdo de informação polimórfica (PIC), índice Shannon-Wiener (H') e heterozigosidade observada ($H_o$). Mais uma vez a

maior diversidade genética das populações base $P_2$ e $P_3$ em relação a $P_1$ foi constatada por estes índices. Populações que tiveram como ancestral $P_3$

Tabela 8. Estimativas médias do número de genótipos por loco ($\overline{g}_j$), conteúdo de informação polimórfica (PIC), índice Shannon-Wiener (H') e heterozigosidade observada ($H_o$) obtidas a partir de 20 locos codominantes e multialélicos para 42 populações de melhoramento geradas via simulação

| Populações | $\overline{g}_j$ [4,5,6] | PIC [4,6] | H' [4] | H' [5] | $H_o$ [1,2,3,4,5,6,7] |
|---|---|---|---|---|---|
| $P_1$ | 4,00 | 0,26 | 0,81 | 0,51 | 0,29 |
| $P_2$ | 5,25 | 0,29 | 0,93 | 0,58 | 0,33 |
| $P_3$ | 5,40 | 0,38 | 1,15 | 0,74 | 0,46 |
| $P_1a_1$ | 3,90 | 0,25 | 0,76 | 0,49 | 0,29 |
| $P_2a_1$ | 5,25 | 0,27 | 0,89 | 0,56 | 0,30 |
| $P_2a_2$ | 5,05 | 0,27 | 0,88 | 0,55 | 0,31 |
| $P_2a_3$ | 5,00 | 0,27 | 0,88 | 0,56 | 0,32 |
| $P_2a_4$ | 4,95 | 0,28 | 0,90 | 0,56 | 0,31 |
| $P_3a_1$ | 5,50 | 0,38 | 1,16 | 0,74 | 0,44 |
| $P_2s_1$ | 5,60 | 0,30 | 0,93 | 0,60 | 0,18 |
| $P_2s_2$ | 5,10 | 0,29 | 0,81 | 0,59 | 0,08 |
| $P_2s_3$ | 4,95 | 0,30 | 0,74 | 0,60 | 0,04 |
| $P_2s_4$ | 4,60 | 0,30 | 0,68 | 0,59 | 0,02 |
| $P_2s_5$ | 4,20 | 0,30 | 0,65 | 0,60 | 0,01 |
| $H_{12pp}$ | 5,15 | 0,34 | 0,96 | 0,68 | 0,47 |
| $H_{12mp}$ | 5,10 | 0,34 | 0,98 | 0,68 | 0,47 |
| $H_{13pp}$ | 5,35 | 0,40 | 1,11 | 0,78 | 0,55 |
| $H_{13mp}$ | 5,25 | 0,40 | 1,11 | 0,79 | 0,56 |
| $H_{23pp}$ | 6,05 | 0,40 | 1,16 | 0,80 | 0,54 |
| $F_{2(pp)}$ | 7,20 | 0,40 | 1,25 | 0,80 | 0,26 |
| $F_{3(pp)}$ | 6,60 | 0,40 | 1,11 | 0,80 | 0,12 |
| $F_{4(pp)}$ | 6,05 | 0,40 | 0,99 | 0,80 | 0,06 |
| $H_{23(mp)}$ | 6,15 | 0,41 | 1,20 | 0,82 | 0,55 |
| $F_{2(mp)}$ | 7,25 | 0,41 | 1,28 | 0,82 | 0,27 |
| $F_{3(mp)}$ | 6,85 | 0,41 | 1,16 | 0,83 | 0,14 |
| $F_{4(mp)}$ | 6,35 | 0,41 | 1,03 | 0,82 | 0,07 |
| $RC_{11pp}$ | 5,40 | 0,36 | 1,09 | 0,71 | 0,43 |
| $RC_{12pp}$ | 5,30 | 0,31 | 0,98 | 0,62 | 0,36 |
| $RC_{13pp}$ | 5,10 | 0,29 | 0,90 | 0,57 | 0,32 |
| $RC_{14pp}$ | 4,60 | 0,27 | 0,86 | 0,54 | 0,31 |
| $RC_{11mp}$ | 5,60 | 0,35 | 1,09 | 0,70 | 0,43 |
| $RC_{12mp}$ | 5,15 | 0,31 | 0,99 | 0,63 | 0,37 |
| $RC_{13mp}$ | 5,05 | 0,28 | 0,88 | 0,56 | 0,33 |
| $RC_{14mp}$ | 4,90 | 0,27 | 0,84 | 0,53 | 0,32 |
| $RC_{31pp}$ | 5,85 | 0,41 | 1,26 | 0,81 | 0,48 |
| $RC_{32pp}$ | 6,05 | 0,40 | 1,23 | 0,78 | 0,47 |
| $RC_{33pp}$ | 6,00 | 0,39 | 1,20 | 0,76 | 0,45 |
| $RC_{34pp}$ | 5,90 | 0,38 | 1,19 | 0,75 | 0,43 |
| $RC_{31mp}$ | 6,15 | 0,41 | 1,24 | 0,80 | 0,50 |
| $RC_{32mp}$ | 5,80 | 0,40 | 1,22 | 0,78 | 0,47 |
| $RC_{33mp}$ | 5,70 | 0,39 | 1,19 | 0,76 | 0,47 |
| $RC_{34mp}$ | 5,85 | 0,39 | 1,20 | 0,76 | 0,45 |

Programas utilizados: [1]Arlequin; [2]GDA; [3]GENEPOP; [4]GENES; [5]POPGENE; [6]PowerMarker; [7]TFPGA.

exibiram as maiores estimativas médias do índice Shannon-Wiener (Figura 2). Maiores valores de PIC (0,41 a 0,40) foram constatados nas populações obtidas por hibridação com $P_3$ e nos descendentes desta hibridação, ou seja, as gerações $F_n$ e de retrocruzamentos. Já as populações de acasalamento ao acaso praticamente não exibiram variação nos índices da Tabela 8, em virtude, obviamente, das condições de EHW.

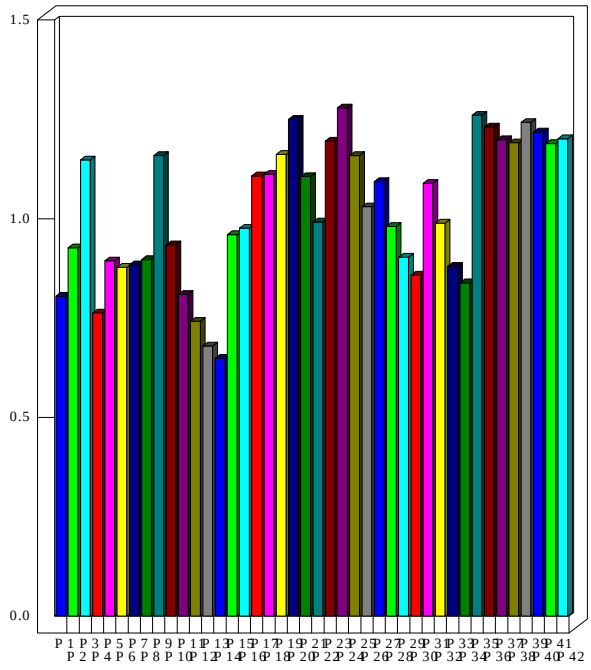


Figura 2. Estimativas médias do índice Shannon-Wiener (H') obtidas a partir de 20 locos codominantes e multialélicos para 42 populações de melhoramento geradas via simulação. Visualização gráfica disponibilizada pelo programa GENES. Representação das populações no gráfico: P4 – $P_1a_1$; P5 - $P_1a_1$; P6 - $P_1a_2$; P7 - $P_1a_3$; P8 - $P_1a_4$; P9 – $P_3a_1$; P10 – $P_2s_1$; P11 – $P_2s_2$; P12 – $P_2s_3$; P13 – $P_2s_4$; P14 – $P_2s_5$; P15 – $H_{12pp}$; P16 – $H_{12mp}$; P17 – $H_{13pp}$; P18 – $H_{13mp}$; P19 – $H_{23pp}$; P20 – $F_{2(pp)}$; P21 – $F_{3(pp)}$; P22 – $F_{4(pp)}$; P23 – $H_{23mp}$; P24 – $F_{2(mp)}$; P25 – $F_{3(mp)}$; P26 – $F_{4(mp)}$; P27 – $RC_{11pp}$; P28 – $RC_{12pp}$; P29 – $RC_{13pp}$; P30 – $RC_{14pp}$; P31 – $RC_{11mp}$; P32 – $RC_{12mp}$; P33 – $RC_{13mp}$; P34 – $RC_{14mp}$; P35 – $RC_{31pp}$; P36 – $RC_{32pp}$; P37 – $RC_{33pp}$; P38 – $RC_{34pp}$; P39 – $RC_{31mp}$; P40 – $RC_{32mp}$; P41 – $RC_{33mp}$; P42 – $RC_{34mp}$.

O avanço de gerações de autofecundação ($P_2s_1$ a $P_2s_4$) proporcionou queda nas estimativas de $\bar{g}_j$, H' (obtido pelo programa

GENES) e $H_o$. Fenômeno semelhante foi observado com as populações $F_n$. Pela teoria de genética de populações é sabido que a cada geração de autofecundação reduz-se à metade a quantidade de genótipos heterozigotos e, conseqüentemente, a riqueza genotípica (H') dessas populações diminui (Figura 2).

As populações $F_2$ ($F_{2(pp)}$ e $F_{2(mp)}$), comparadas aos respectivos híbridos $F_1$ e as gerações $F_n$ subseqüentes, exibiram maior $\bar{g}_j$, H' (Figura 2) e $H_o$. Populações $F_2$ são detentoras de maior variabilidade genética, ou seja, amostra de genótipos. Se esta população $F_2$ é oriunda da autofecundação de uma população $F_1$, cujos genitores são homozigotos contrastantes, para um loco codominante a segregação será de 1:2:1 (Schuster & Cruz, 2004) e, assim, todas as formas genotípicas estarão presentes.

Nos retrocruzamentos, houve tendência dos índices $\bar{g}_j$, PIC, H' e $H_o$ decrescerem ao longo das gerações. As populações de retrocruzamento oriundas do genitor recorrente $P_3$ ($RC_{31pp}$ a $RC_{34pp}$ e $RC_{31mp}$ a $RC_{34mp}$), apresentaram superioridade em todos os índices em relação àquelas cujo genitor recorrente foi $P_1$ ($RC_{11pp}$ a $RC_{14pp}$ e $RC_{11mp}$ a $RC_{14mp}$) (Tabela 8). Espera-se que cada geração t de retrocruzamento seja similar ao genitor recorrente na ordem de $1 - (1/2)^{t+1}$, ou seja, esta é a proporção média de recuperação do genoma recorrente. Se for considerado que as populações genitoras contrastam em $\ell_c$ locos, após t gerações, na ausência de seleção, ter-se-á $\left(\frac{2^{\ell_c}-1}{2^{\ell_c}}\right)^t$ dos indivíduos homozigotos idênticos ao genitor recorrente nos $\ell_c$ locos contrastantes (Borém, 1998).

O índice Shannon-Wiener (H'), também conhecido como índice Shannon-Weaver é um índice bastante utilizado em estudos ecológicos, em que se deseja avaliar a biodiversidade dentro de uma área, ou seja, a riqueza de espécies. No entanto, tem sido empregado especificamente em estudos genéticos (Holcomb et al., 1977; Picoli et al., 2004) como uma medida de diversidade dentro da população e se assemelha a um índice de riqueza genotípica. O valor máximo de H' ($H'_{máx}$) ocorre quando se tem freqüências genotípicas iguais no loco e pode ser definido como $H'_{máx}$ =

$\log(n_c)$, em que $n_c$ é o número de classes genotípicas observadas. Assim, os valores de H' se distribuem no intervalo de 0 a $\log(n_c)$. Quando se deseja que $H'_{máx}$ seja igual a unidade (1), normaliza-se seu valor dividindo-o por $\log(n_c)$. Sugere-se que o logaritmo seja com base 2, base *e* (logaritmo neperiano) ou base 10 (Alfenas et al., 1991), o que pode explicar as diferenças para os valores médios de H', realizados no programa GENES e POPGENE (Tabela 8). Vale ressaltar a importância de índices genotípicos, principalmente em estudos de melhoramento de plantas. A seleção artificial atua diretamente no organismo ao nível genotípico (Hedrick, 1971), portanto, medidas de diversidade gênica nem sempre são capazes de contabilizar e discernir a variabilidade genotípica entre ou dentro de populações. Para um determinado loco, duas populações podem ter as mesmas freqüências alélicas, mas possuírem diferentes arranjos genotípicos.

Além da heterozigosidade, é comum usar o conteúdo de informação de polimorfismo (PIC) para quantificar o polimorfismo genético do loco na população. Botstein et al. (1980) originalmente definiram valores de PIC como a probabilidade de um loco (marcador) ser informativo em condições de acasalamento ao acaso. Este conhecimento é bastante relevante ao mapeamento genético, pois acredita-se que quanto maior o valor de PIC, maior será o conteúdo de informação de ligação (Liu, 1997).

Para um loco cujas freqüências alélicas são iguais, a expressão de $h_e$ fica reduzida a $1 - 1/a_j$ e de $PIC = 1 - (1/a_j - 1/a_j^2 + 1/a_j^3)$, em que $a_j$ é o número de alelos no loco j. Nesta situação, o valor de PIC é sempre inferior a heterozigosidade esperada máxima. No entanto, quanto maior for o número de alelos,mais o valor de PIC se aproximará de $\hat{h}_e$. As informações da Tabela 9 ajudam a compreender este fato. Por exemplo, na população $H_{23(mp)}$, o loco 3 possui apenas dois alelos, com freqüências 0,4775 e 0,5225. Já o loco 9 possui cinco alelos, com freqüências 0,4025, 0,2725, 0,2800, 0,0125 e 0,0325 (dados não apresentados). No loco 3, $\hat{h}_e$ e PIC assumiram os valores de 0,4989 e 0,3745, respectivamente. O loco 9 apresentou $\hat{h}_e$ = 0,6841 e PIC = 0,6341. Percebe-se que a menor diferença entre $\hat{h}_e$ e PIC foi para o loco com maior número de alelos.

Tabela 9. Estimativas de freqüências alélicas, heterozigosidade esperada ($\hat{h}_e$ #) e conteúdo de informação polimórfica (PIC) para dois locos da população $H_{23(mp)}$

| | Loco | |
|---|---|---|
| Freqüência | 3 | 9 |
| alelo 1 | 0,4775 | 0,4025 |
| alelo 2 | 0,5225 | 0,2725 |
| alelo 3 | - | 0,2800 |
| alelo 4 | - | 0,0125 |
| alelo 5 | - | 0,0325 |
| PIC | 0,3745 | 0,6341 |
| $\hat{h}_e$ | 0,4989 | 0,6841 |

#Estimador viesado

Na tabela 10 são apresentadas as estimativas do índice de fixação/endogamia (*f*) e heterozigosidade esperada (ou diversidade gênica, $H_e$ e, ou, $\overline{D}_j$) para os 20 locos simulados. O índice *f* foi estimado por quatro maneiras diferentes: método dos momentos ($f^{\$}$); baseado nas heterozigosidades ($f^{€}$); segundo Weir & Cockerham (1984) ($f^{£}$) e Robertson & Hill (1984) ($f^{¥}$).

Quando os cruzamentos, em uma população, ocorrem entre indivíduos aparentados, chama-se a este padrão de acasalamento de endogamia (Falconer, 1987; Hartl & Clark, 1997). Embora os diferentes estimadores de *f* tenham assumido diferentes estimativas para as populações, quando $f^{\$}$, $f^{€}$ e $f^{£}$ foram ordenadas de forma crescente (ou decrescente) seguiram a uma mesma ordem para as gerações de autofecundação ($P_2s_1$ a $P_2s_4$), $F_n$ ($F_{2(pp)}$ a $F_{4(pp)}$ e $F_{2(mp)}$ a $F_{4(mp)}$) e híbridos $F_1$ ($H_{12pp}$, $H_{12mp}$, $H_{13pp}$, $H_{13mp}$, $H_{23pp}$ e $H_{23mp}$). Para $f^{¥}$ as populações anteriormente referidas alteraram poucas posições no ordenamento. Em relação as demais populações houve maior alternância no ordenamento das populações.

Tabela 10. Estimativas do índice de fixação (*f*) e heterozigosidade esperada ($H_e$ e $\overline{D}_j$) obtidas a partir de 20 locos codominantes e multialélicos para 42 populações de melhoramento geradas via simulação

| Populações | $f$ [ς 2, 6] | $f$ [€4,5] | $f$ [£3] | $f$ [¥3] | $H_e$ [#4,5,6,7] | $H_e$ [## §1,2,5,7] | $\overline{D}_j$ [## ¢6] |
|---|---|---|---|---|---|---|---|
| $P_1$ | 0,0519 | 0,0367 | 0,0392 | 0,0255 | 0,3035 | 0,3042 | 0,3027 |
| $P_2$ | -0,0064 | -0,0115 | -0,0090 | -0,0056 | 0,3276 | 0,3284 | 0,3268 |
| $P_3$ | -0,0399 | -0,0361 | -0,0337 | -0,0153 | 0,4458 | 0,4470 | 0,4448 |
| $P_1a_1$ | 0,0058 | 0,0087 | 0,0112 | 0,0085 | 0,2940 | 0,2947 | 0,2932 |
| $P_2a_1$ | 0,0320 | 0,0253 | 0,0278 | 0,0169 | 0,3109 | 0,3117 | 0,3101 |
| $P_2a_2$ | 0,0061 | 0,0037 | 0,0062 | -0,0006 | 0,3068 | 0,3076 | 0,3061 |
| $P_2a_3$ | -0,0114 | -0,0144 | -0,0119 | -0,0144 | 0,3117 | 0,3125 | 0,3109 |
| $P_2a_4$ | 0,0220 | 0,0099 | 0,0124 | 0,0097 | 0,3144 | 0,3152 | 0,3136 |
| $P_3a_1$ | 0,0015 | 0,0021 | 0,0046 | 0,0010 | 0,4436 | 0,4447 | 0,4425 |
| $P_2s_1$ | 0,4795 | 0,4777 | 0,4796 | 0,4759 | 0,3969 | 0,3382 | 0,3361 |
| $P_2s_2$ | 0,7592 | 0,7583 | 0,7593 | 0,7597 | 0,3349 | 0,3357 | 0,3334 |
| $P_2s_3$ | 0,8702 | 0,8716 | 0,8722 | 0,8838 | 0,3375 | 0,3383 | 0,3359 |
| $P_2s_4$ | 0,9342 | 0,9305 | 0,9308 | 0,9363 | 0,3366 | 0,3374 | 0,3349 |
| $P_2s_5$ | 0,9624 | 0,9597 | 0,9599 | 0,9670 | 0,3376 | 0,3384 | 0,3359 |
| $H_{12pp}$ | -0,1993 | -0,1667 | -0,1643 | -0,1142 | 0,4306 | 0,3885 | 0,3868 |
| $H_{12mp}$ | -0,1924 | -0,1627 | -0,1603 | -0,1179 | 0,3894 | 0,3904 | 0,3886 |
| $H_{13pp}$ | -0,1937 | -0,1899 | -0,1876 | -0,1579 | 0,4607 | 0,4618 | 0,4597 |
| $H_{13mp}$ | -0,2018 | -0,1971 | -0,1948 | -0,1649 | 0,4665 | 0,4677 | 0,4656 |
| $H_{23pp}$ | -0,1604 | -0,1471 | -0,1447 | -0,0799 | 0,4618 | 0,4629 | 0,4608 |
| $F_{2(pp)}$ | 0,4435 | 0,4557 | 0,4577 | 0,4749 | 0,4647 | 0,4658 | 0,4630 |
| $F_{3(pp)}$ | 0,7367 | 0,7455 | 0,7466 | 0,7809 | 0,4633 | 0,4644 | 0,4612 |
| $F_{4(pp)}$ | 0,8625 | 0,8688 | 0,8694 | 0,8946 | 0,4614 | 0,4625 | 0,4592 |
| $H_{23(mp)}$ | -0,1524 | -0,1431 | -0,1407 | -0,0741 | 0,4723 | 0,4735 | 0,4713 |
| $F_{2(mp)}$ | 0,4358 | 0,4361 | 0,4381 | 0,4414 | 0,4697 | 0,4709 | 0,4680 |
| $F_{3(mp)}$ | 0,7143 | 0,7180 | 0,7192 | 0,7535 | 0,4723 | 0,4734 | 0,4702 |
| $F_{4(mp)}$ | 0,8603 | 0,8597 | 0,8604 | 0,8913 | 0,4704 | 0,4716 | 0,4682 |
| $RC_{11pp}$ | -0,0469 | -0,0591 | -0,0566 | -0,0484 | 0,4100 | 0,4110 | 0,4090 |
| $RC_{12pp}$ | 0,0089 | 0,0093 | 0,0118 | 0,0085 | 0,3590 | 0,3599 | 0,3581 |
| $RC_{13pp}$ | 0,0232 | 0,0174 | 0,0199 | 0,0099 | 0,3311 | 0,3319 | 0,3303 |
| $RC_{14pp}$ | 0,0156 | 0,0051 | 0,0076 | 0,0086 | 0,3189 | 0,3197 | 0,3181 |
| $RC_{11mp}$ | -0,0526 | -0,0630 | -0,0606 | -0,0550 | 0,4057 | 0,4067 | 0,4047 |
| $RC_{12mp}$ | -0,0119 | -0,0286 | -0,0261 | -0,0281 | 0,3615 | 0,3624 | 0,3606 |
| $RC_{13mp}$ | -0,0047 | -0,0046 | -0,0021 | 0,0018 | 0,3232 | 0,3240 | 0,3224 |
| $RC_{14mp}$ | -0,0039 | -0,0134 | -0,0109 | 0,0006 | 0,3137 | 0,3145 | 0,3129 |
| $RC_{31pp}$ | -0,0082 | -0,0124 | -0,0099 | 0,0021 | 0,4779 | 0,4791 | 0,4767 |
| $RC_{32pp}$ | -0,0003 | -0,0051 | -0,0026 | -0,0052 | 0,4655 | 0,4666 | 0,4643 |
| $RC_{33pp}$ | 0,0056 | 0,0053 | 0,0079 | 0,0029 | 0,4544 | 0,4555 | 0,4533 |
| $RC_{34pp}$ | 0,0435 | 0,0328 | 0,0353 | 0,0240 | 0,4502 | 0,4513 | 0,4490 |
| $RC_{31mp}$ | -0,0487 | -0,0499 | -0,0474 | -0,0404 | 0,4754 | 0,4766 | 0,4743 |
| $RC_{32mp}$ | -0,0210 | -0,0233 | -0,0208 | -0,0207 | 0,4626 | 0,4638 | 0,4615 |
| $RC_{33mp}$ | -0,0240 | -0,0261 | -0,0236 | -0,0268 | 0,4545 | 0,4556 | 0,4533 |
| $RC_{34mp}$ | 0,0175 | 0,0085 | 0,0110 | 0,0025 | 0,4533 | 0,4544 | 0,4521 |

Programas utilizados: [1]Arlequin; [2]GDA; [3]GENEPOP; [4]GENES; [5]POPGENE; [6]PowerMarker; [7]TFPGA.
[#] Estimador viesado; [##] Estimadores não viesados; [§ 2,5,7] Fator de correção $2N_i/(2N_i - 1)$ e [1] Fator de correção $N_i/(N_i - 1)$; [¢] Fator de correção $1 - [(1 + f)/2n]$.
[ς] Estimador obtido pelo método dos momentos; [€] Estimador baseado nas heterozigosidades observada e esperada; [£] Estimador obtido conforme Weir & Cockerham (1984); [¥] Estimador obtido conforme Robertson & Hill (1984).

As estimativas dos índices de fixação/endogamia ($f^{\$}$, $f^{€}$, $f^{£}$ e $f^{¥}$), para uma mesma população, apresentaram menores diferenças entre si quando as populações foram geradas por autofecundação, ou seja, para aquelas endogâmicas. Notou-se que as estimativas de Robertson & Hill ($f^{¥}$) diferiram das de Weir & Cockerham ($f^{£}$) quando o loco possuia mais de dois alelos. Weir & Cockerham (1984) desenvolveram a técnica que estima o coeficiente de fixação/endogamia dentro da população para cada alelo. O estimador pelo método dos momentos é ponderado para a média dos locos, em que soma-se o numerador e o denominador, para se obter o quociente médio dos locos. O índice *f* de Robertson & Hill é baseado na distribuição dos desvios das proporções do EHW.

O índice de fixação/endogamia é um dos parâmetros mais importantes em genética de populações, por medir o balanço entre homozigotos e heterozigotos nas populações. A explicação para populações ou espécies que contém maior número de locos em homozigose e menor em heterozigose, deve estar associada ao sistema reprodutivo e, ou, à deriva genética (Kageyama et al., 2003). O principal efeito da endogamia é diminuir a heterozigosidade da população, quando comparada a heterozigosidade esperada. Em termos biológicos, é possível quantificá-la pela comparação das proporções de heterozigotos observados e heterozigotos esperados ($f^{€}$). Dentro deste conceito, a estimativa do coeficiente *f* possibilita mensurar a deficiência ou excesso de heterozigotos nas populações. Valores negativos são indicativos de que existe um excesso de heterozigotos na população. Embora nas populações base $P_2$ e $P_3$ tenha sido estimado valores negativos (-0,0064 e -0,0399, respectivamente) estes não são valores tão expressivos a ponto de qualificar as populações como detentoras de excesso de heterozigotos. No entanto, é provável que elas apresentem mais heterozigotos do que $P_1$. Por analogia pode-se ter uma mesma visão das estimativas de *f* nas populações de retrocruzamento. Mesmo existindo valores positivos e negativos, a exemplo das populações $RC_{14mp}$ (-0,0039), $RC_{32pp}$ (-0,0003) e $RC_{12pp}$ (0,0089), $RC_{33pp}$ (0,0056), estes podem ser considerados insignificantes e, consequentemente, o efeito da endogamia sobre elas ser desprezível.

Nos híbridos $F_1$ as estimativas de *f* foram negativas e de magnitude elevada (-,01524 a -0,2018). Nestas populações observa-se maior heterozigosidade observada ($H_o$) do que heterozigosidade esperada ($H_e$). Por outro lado, as gerações de autofecundação ($P_2s_1$ e $P_2s_4$) e segregantes $F_n$ apresentaram aumento progressivo do coeficiente *f* e diminuição de $H_o$ (Tabela 8, 10 e Figura 3). Nas populações base e de acasalamento ao acaso a igualdade nas estimativas de $H_e$ e $H_o$ estão de acordo com o esperado (Figura 3). Segundo Fukunaga et al. (2005) as discrepâncias entre heterozigosidade esperada e observada podem ser devido ao *drop-out* do alelo (falha de um alelo) durante a amplificação na reação de PCR; estruturação da população e endogamia. Em espécies em que há auto-incompatibilidade e a reprodução é exclusivamente assexuada, verifica-se altos níveis de heterozigosidade observada, a exemplo do pomelo (Barkley et al., 2006).

De acordo com Weir (1996) a freqüência de heterozigotos é uma maneira de representar a existência de variação genética, desde que carreguem diferentes alelos. No entanto, há situações em que a variação é resultante da presença contínua de diferentes homozigotos, a exemplo das populações de autofecundação. Nessa condição, o termo diversidade gênica torna-se mais apropriado.

As medidas viesadas e não viesadas de heterozigosidade esperada, ou diversidade gênica ($H_e$ e $\overline{D}_j$) não apresentaram discrepâncias em suas magnitudes quando comparadas em uma mesma população, mantendo ordenamento das populações bem similar. Na verdade o valor esperado de $\hat{D}_j$ para um determinado loco, numa população qualquer, possui um viés de $1 - [(1 + f)\, 2N_i]$ (Weir, 1996). Para populações não endogâmicas ($f = 0$) esse viés é igual a $(2N_i - 1)/2N_i$ e para populações endogâmicas ($f = 1$) esse viés é $(N_i - 1)/N_i$, o que justifica os fatores de correção utilizados nos estimadores não viesados de diversidade gênica. Os estimadores de $H_e$ corrigidos pelos fatores $N_i/(N_i - 1)$ e $2N_i/(2N_i - 1)$ apresentaram os mesmos valores, considerando as quatro casas decimais, em função do tamanho amostral das populações (Tabela 10).

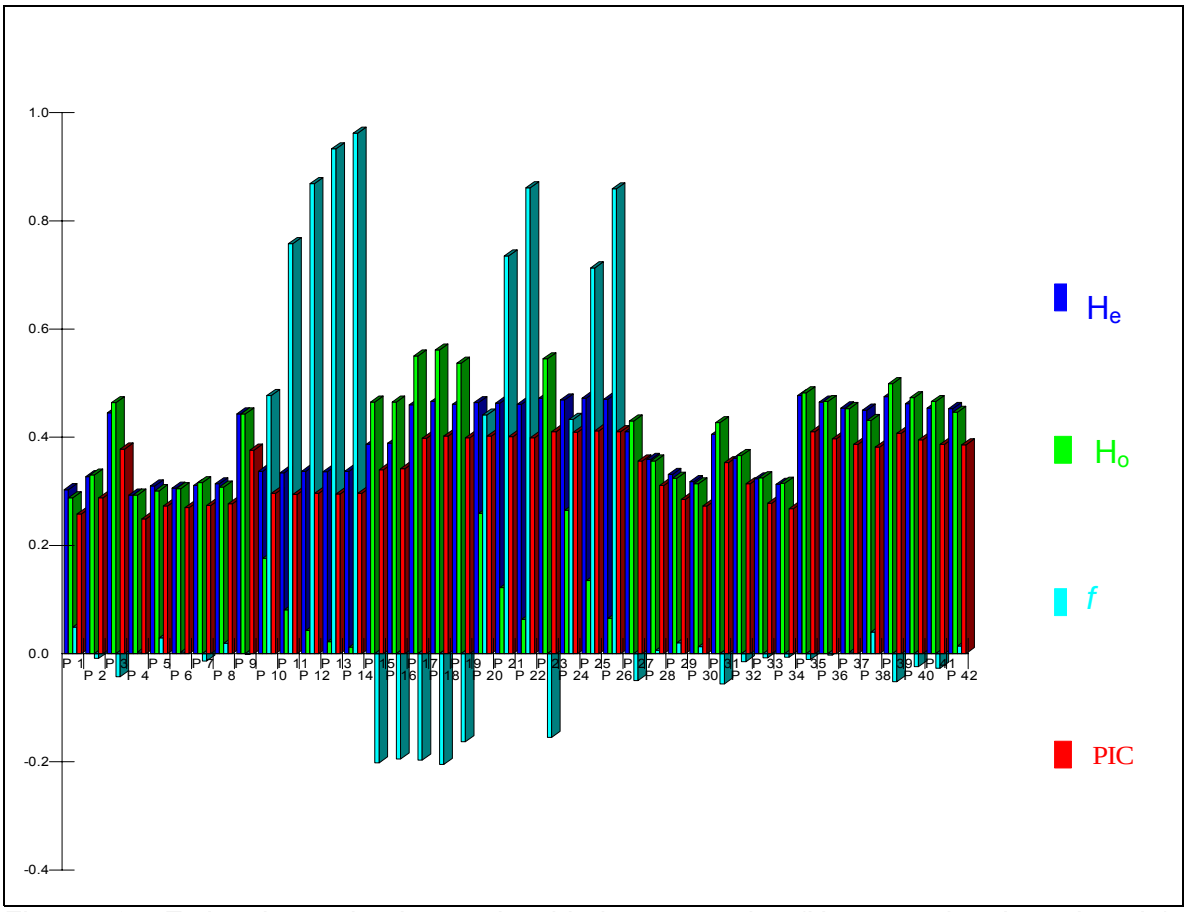


Figura 3. Estimativas da heterozigosidade esperada ($H_e$ – estimador viesado), heterozigosidade observada ($H_o$), índice de fixação/endogamia (*f* – estimador baseado nas heterozigosidades) e conteúdo médio de informação polimórfica (PIC) obtidas a partir de 20 locos codominantes e multialélicos para 42 populações de melhoramento geradas via simulação. Visualização gráfica disponibilizada pelo programa GENES. Representação das populações no gráfico: P4 – $P_1a_1$; P5 - $P_1a_1$; P6 - $P_1a_2$; P7 - $P_1a_3$; P8 - $P_1a_4$; P9 – $P_3a_1$; P10 – $P_2s_1$; P11 – $P_2s_2$; P12 – $P_2s_3$; P13 – $P_2s_4$; P14 – $P_2s_5$; P15 – $H_{12pp}$; P16 – $H_{12mp}$; P17 – $H_{13pp}$; P18 – $H_{13mp}$; P19 – $H_{23pp}$; P20 – $F_{2(pp)}$; P21 – $F_{3(pp)}$; P22 – $F_{4(pp)}$; P23 – $H_{23mp}$; P24 – $F_{2(mp)}$; P25 – $F_{3(mp)}$; P26 – $F_{4(mp)}$; P27 – $RC_{11pp}$; P28 – $RC_{12pp}$; P29 – $RC_{13pp}$; P30 – $RC_{14pp}$; P31 – $RC_{11mp}$; P32 – $RC_{12mp}$; P33 – $RC_{13mp}$; P34 – $RC_{14mp}$; P35 – $RC_{31pp}$; P36 – $RC_{32pp}$; P37 – $RC_{33pp}$; P38 – $RC_{34pp}$; P39 – $RC_{31mp}$; P40 – $RC_{32mp}$; P41 – $RC_{33mp}$; P42 – $RC_{34mp}$.

Como esperado a população $P_3$ foi a que apresentou os maiores valores de diversidade gênica, dentre as populações base. As populações de autofecundação apresentaram estimativas de $H_e$ e $\overline{D}_j$ ligeiramente maiores que as populações de acasalamento ao acaso e $P_2$. Este fato pode ser explicado pela composição homozigota diversificada dessas populações (Weir, 1996). A condição de acasalamento ao acaso manteve os valores de $H_e$ e $\overline{D}_j$ inalterados ao longo das gerações.

Populações híbridas oriundas do cruzamento $P_1$ x $P_3$ e $P_2$ x $P_3$ apresentaram maiores estimativas de heterozigosidade esperada. As estimativas de $H_e$ e $\overline{D}_j$ tiveram um ligeiro decréscimo com os avanços das gerações de retrocruzamento. A diversidade gênica nas populações $F_1$ e $F_n$ foram semelhantes, mas a variação genética das populações híbridas e gerações $F_n$ tem origens diferentes.

Segundo Ott (1992) um loco é considerado polimórfico se $\hat{h}_e \geq 0,1$ e altamente polimórfico se $\hat{h}_e \geq 0,7$. Isto implica dizer que um marcador é considerado polimórfico quando o alelo mais freqüente tem freqüência inferior a 0,95, pois $\hat{h}_e \cong 1 - 0,95^2 \cong 0,1$.

O nível de heterozigosidade esperada depende e muito do tipo de marcador utilizado e, obviamente, das populações estudadas. Gimenes & Lopes (2000), utilizando nove marcadores isoenzimáticos, em 15 populações (raças) de milho, obtiveram na média das populações $\hat{H}_e$ = 0,195. Zheng & Ennos (1999) estudando amostras de sementes de populações naturais de duas variedades de *Pinus caribaea* Morelet, a partir de 8 locos isoenzimáticos, obtiveram $\hat{H}_e$ = 0,26, considerada em nível alto. Em espécies arbóreas tropicais de diferentes estados sucessionais foram obtidos valores de heterozigosidades observada e esperadas de 0,132 a 0,907 (kageyama et al., 2003). Melo Júnior et al. (2004) estudando populações naturais de pequizeiro (*Caryocar brasiliense* Camb.) com isoenzimas, obtiveram elevadas estimativas de $H_e$ (0,45 a 0,53) e $H_o$ (0,583 a 0,817). Takezaki & Nei (1996) definiram níveis de heterozigosidade esperada entre 0,16 e 0,5 para o modelo mutacional de alelos infinitos (IAM) e 0,5 a 0,8 para o modelo mutacional *stepwise* (SMM) em um estudo de simulação. Segundo os autores, para IAM, o valor de $\hat{H}_e$ = 0,5 representa um alto nível de heterozigosidade e $\hat{H}_e$ = 0,16 um baixo nível de heterozigosidade, considerando marcadores genéticos clássicos. No caso de locos microssatélites, a heterozigosidade esperada média encontra-se geralmente entre 0,5 e 0,8. Dentro dos níveis apresentados na literatura, pode-se dizer

que todas as populações simuladas apresentaram heterozigosidade esperada mediana a alta.

As pequenas variações de $H_e$ entre as populações base e as respectivas gerações de acasalamento ao acaso podem ser devido a flutuação de freqüências alélicas. durante o processo de simulação, os indivíduos das populações, inclusive as de acasalamento ao acaso, não estavam livres de autopolinização, embora pouco provável, dado o tamanho amostral ($N_i$ = 200) simulado. Este foi um aspecto importante no processo de simulação, pois o tornou mais próxima de situações realísticas.

Nas hibridações planta a planta e por mistura de pólen não foram observadas diferenças expressivas nas magnitudes dos índices estudados (Tabelas 5, 7, 8 e 10).

Na Tabela 11 estão presentes as estimativas de correlação de Pearson e de Spearman entre as medidas descritivas. As estimativas das duas correlações para $\overline{A}_j$, $A_t$, $A_r$, $P_a$, PLP, $\overline{A}_p$, $\overline{g}_j$, PIC, H' e $H_e$ (ou $\overline{D}_j$) foram todas positivas e de magnitudes medianas a altas entre si, sendo as correlações de Pearson significativas ($P < 0,05$). As exceções foram as correlações com o coeficiente de fixação/endogamia (*f*) e heterozigosidade observada ($H_o$), sendo estas baixas e não significativas. Apenas $A_e$, PIC, H' e $H_e$ apresentaram correlações medianas e significativas com a heterozigosidade observada. Ainda, $A_r$ apresentou baixas correlações com a diversidade gênica ($\overline{D}_j$).

Entre $H_o$ e *f* as estimativas de correlação foram negativas, altas (-0,91 e -0,93) e significativas ($P < 0,05$). O índice de fixação não se correlacionou com as demais estatísticas descritivas. Desvios nas proporções do EHW são originados do coeficiente de endogamia (*f*), que diferente de zero, causa o desequilíbrio dentro do loco (Falconer, 1987; Weir 1996; Hartl & Clark, 1997). Portanto, a ausência ou fraca associação de $H_o$ e *f* com as demais medidas descritivas, que se mostraram altamente correlacionadas com a heterozigosidade esperada, se justifica.

Nybom (2004) compilou 307 trabalhos, publicados de 1993 a 2003, usando marcadores moleculares para avaliação da diversidade genética entre e dentro de populações em espécies silvestres de angiospermas e

gimnospermas. Ele explica que um aumento da população analisada leva a um aumento no número de alelos computados, estando assim correlacionado com a diversidade dentro de população. O autor ainda verificou que o número de locos microssatélites estudados teve um efeito negativo em $H_e$, possivelmente, porque os pesquisadores tendem a investigar mais locos nos táxon, onde cada loco exibe pouco polimorfismo. O número de alelos tem um efeito positivo tanto em $H_e$ quanto em $H_o$, possivelmente porque locos altamente heterozigotos permitem a detecção de um número maior de alelos por loco.

Estes índices de diversidade não podem ser comparáveis entre estimativas obtidas via marcadores isoenzimáticos e microssatélites, pois o poder da técnica de microssatélites em detectar diferentes alelos nos locos, é muito superior às isoenzimas (kageyama et al., 2003).

Tabela 11. Estimativas do coeficiente de correlação de Pearson (abaixo da diagonal) e correlação de Spearman (acima da diagonal) para 19 medidas descritivas da variabilidade genética intrapopulacional a partir das observações médias de 20 locos em 42 populações de geradas via simulação

| | $\overline{A}_j$ | $A_t$ | $A_r$ | $A_e$ | $P_a$ | PLP[#] | $\overline{A}_p$ [#] | PLP[##] | $\overline{A}_p$ [##] | $\overline{g}_j$ | PIC | H'[3] | H'[4] | $H_o$ | $f$ [ς] | $f$ [¥] | $H_e$ [£] | $H_e$ [§] | $\overline{D}_j$ [¢] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\overline{A}_j$ | 1,00 | 0,96 | 0,82 | 0,79 | 0,96 | 0,82 | 0,86 | 0,80 | 0,83 | 0,85 | 0,81 | 0,83 | 0,68 | 0,32 | -0,24 | -0,22 | 0,79 | 0,77 | 0,58 |
| $A_t$ | 0,98* | 1,00 | 0,85 | 0,83 | 1,00 | 0,88 | 0,82 | 0,83 | 0,83 | 0,79 | 0,82 | 0,85 | 0,72 | 0,39 | -0,31 | -0,29 | 0,81 | 0,81 | 0,60 |
| $A_r$ | 0,85* | 0,86* | 1,00 | 0,54 | 0,85 | 0,75 | 0,59 | 0,52 | 0,80 | 0,56 | 0,50 | 0,55 | 0,43 | 0,18 | -0,18 | -0,15 | 0,50 | 0,50 | 0,20 |
| $A_e$ | 0,77* | 0,80* | 0,57* | 1,00 | 0,83 | 0,80 | 0,66 | 0,84 | 0,63 | 0,78 | 0,97 | 0,97 | 0,83 | 0,40 | -0,20 | -0,17 | 0,98 | 0,99 | 0,86 |
| $P_a$ | 0,98* | 1,00* | 0,85* | 0,80* | 1,00 | 0,88 | 0,82 | 0,83 | 0,83 | 0,79 | 0,82 | 0,85 | 0,72 | 0,39 | -0,31 | -0,29 | 0,81 | 0,81 | 0,60 |
| PLP[#] | 0,78* | 0,82* | 0,75* | 0,79* | 0,83* | 1,00 | 0,62 | 0,85 | 0,63 | 0,69 | 0,78 | 0,80 | 0,71 | 0,38 | -0,29 | -0,28 | 0,76 | 0,78 | 0,53 |
| $\overline{A}_p$ [#] | 0,89* | 0,88* | 0,67* | 0,62* | 0,88* | 0,47* | 1,00 | 0,79 | 0,80 | 0,77 | 0,74 | 0,77 | 0,55 | 0,16 | -0,12 | -0,12 | 0,69 | 0,68 | 0,58 |
| PLP[##] | 0,82* | 0,85* | 0,53* | 0,90* | 0,85* | 0,81* | 0,72* | 1,00 | 0,57 | 0,76 | 0,89 | 0,90 | 0,77 | 0,38 | -0,30 | -0,30 | 0,85 | 0,85 | 0,77 |
| $\overline{A}_p$ [##] | 0,89* | 0,89* | 0,81* | 0,58* | 0,89* | 0,51* | 0,94* | 0,59* | 1,00 | 0,65 | 0,63 | 0,67 | 0,44 | 0,08 | -0,01 | 0,02 | 0,61 | 0,61 | 0,39 |
| $\overline{g}_j$ | 0,87* | 0,84* | 0,62* | 0,74* | 0,84* | 0,69* | 0,76* | 0,78* | 0,71* | 1,00 | 0,86 | 0,87 | 0,82 | 0,20 | -0,03 | -0,01 | 0,84 | 0,82 | 0,72 |
| PIC | 0,81* | 0,81* | 0,53* | 0,98* | 0,81* | 0,76* | 0,67* | 0,91* | 0,59* | 0,80* | 1,00 | 0,99 | 0,84 | 0,36 | -0,18 | -0,16 | 0,98 | 0,98 | 0,88 |
| H'[3] | 0,82* | 0,85* | 0,56* | 0,98* | 0,85* | 0,80* | 0,70* | 0,95* | 0,63* | 0,80* | 0,98* | 1,00 | 0,84 | 0,31 | -0,13 | -0,11 | 0,97 | 0,98 | 0,86 |
| H'[4] | 0,65* | 0,68* | 0,48* | 0,88* | 0,68* | 0,77* | 0,45* | 0,81* | 0,39* | 0,78* | 0,86* | 0,87* | 1,00 | 0,56 | -0,35 | -0,33 | 0,85 | 0,85 | 0,79 |
| $H_o$ | 0,12 | 0,12 | 0,17 | 0,40* | 0,12 | 0,36* | -0,09 | 0,28 | -0,13 | 0,10 | 0,35* | 0,31* | 0,60* | 1,00 | -0,91 | -0,90 | 0,37 | 0,38 | 0,36 |
| $f$ [ς] | 0,14 | 0,14 | -0,03 | -0,07 | 0,14 | -0,12 | 0,31* | 0,04 | 0,32* | 0,15 | 0,01 | 0,03 | -0,34* | -0,93* | 1,00 | 0,99 | -0,18 | -0,18 | -0,18 |
| $f$ [¥] | 0,19 | 0,19 | 0,01 | -0,03 | 0,19 | -0,10 | 0,37* | 0,07 | 0,37* | 0,18 | 0,04 | 0,07 | -0,32* | -0,91* | 1,00* | 1,00 | -0,15 | -0,16 | -0,14 |
| $H_e$ [£] | 0,79* | 0,79* | 0,53* | 0,98* | 0,80* | 0,77* | 0,64* | 0,90* | 0,57* | 0,79* | 1,00* | 0,98* | 0,88* | 0,37* | -0,02 | 0,01 | 1,00 | 1,00 | 0,87 |
| $H_e$ [§] | 0,76* | 0,79* | 0,53* | 0,99* | 0,80* | 0,80* | 0,61* | 0,92* | 0,56* | 0,76* | 0,98* | 0,99* | 0,89* | 0,37* | -0,03 | 0,01 | 0,99* | 1,00 | 0,87 |
| $\overline{D}_j$ [¢] | 0,56* | 0,57* | 0,22 | 0,89* | 0,58* | 0,46* | 0,57* | 0,79* | 0,44* | 0,62* | 0,89* | 0,89* | 0,75* | 0,28 | 0,05 | 0,08 | 0,90* | 0,90* | 1,00 |

$\overline{A}_j$: número médio de alelos por loco; $A_t$: número total de alelo na população; $A_r$: número de alelos raros, $A_e$: número efetivo total de alelos; $P_a$: proporção de alelos na população; PLP: proporção de locos polimórficos; $\overline{A}_p$: número médio de alelos por loco polimórfico; $\overline{g}_j$: número médio de genótipos por loco; PIC: conteúdo médio de informação polimórfica, H': índice Shannon-Wiener; $H_o$: heterozigosidade observada média; *f*: índice de fixação/endogamia; $H_e$ e $\overline{D}_j$: heterozigosidade esperada.

[#]Considerado polimórfico o loco em que o alelo mais comum tem freqüência menor que 99%; [##] Loco em que o alelo mais comum tem freqüência menor que 95%.
[3] Obtido pelo programa GENEPOP; [4] Obtido pelo programa GENES; [ς]Estimador obtido pelo método dos momentos; [¥] Estimador obtido conforme Robertson e Hill (1984).
*(P < 0,05); [£]Estimador viesado; [§]Estimador não viesado com fator de correção $2N_i/(2N_i - 1)$; [¢] Estimador não viesado com fator de correção $1 - [(1 + f)/2n]$, com *f* [ς].

### 4.1.2. Tamanho efetivo populacional

A maneira mais conveniente de tratar um desvio particular da estrutura de reprodução idealizada é expressar a situação, em termos do número efetivo de indivíduos que se acasalam, ou melhor, o tamanho efetivo da população (Falconer, 1987). A compreensão do tamanho efetivo é oportuna no sentido do planejamento dos esquemas de acasalamento, visando reduzir a endogamia em populações finitas.

Tamanho efetivo de população representa o número de indivíduos que contribuem efetivamente para a variância de amostragem, ou taxa de endogamia, desde que acasalados de acordo com as premissas da população idealizada (ausência de seleção, migração e mutação) e panmítica.

A Tabela 12 apresenta as estimativas do tamanho efetivo ($N_e$) de cada população. Interpreta-se que a população constituída a partir de 200 indivíduos proporciona valor de variância das freqüências alélicas ou da taxa de endogamia equivalente à fornecida pelo intercruzamento de $N_e$ indivíduos da população idealizada.

As populações base $P_1$, $P_2$, $P_3$, as de acasalamento ao acaso, $P_2a_1$ a $P_3a_1$, e de retrocruzamento, $RC_{11pp}$ a $RC_{14mp}$, tiveram seus tamanhos efetivos mais próximos do tamanho amostral ($N_i$ = 200) quando comparadas às demais populações. Esta proximidade ocorrente nas populações base e de acasalamento ao acaso é porque elas atendem aos pressupostos de uma população idealizada de cruzamentos aleatórios. Concomitantemente, quando o valor do coeficiente de fixação/endogamia (*f*) é negligenciável, o valor de $N_e$ praticamente se iguala ao número de indivíduos amostrados (Moraes et al., 1999; 2002).

As populações de retrocruzamento com os avanços das gerações se tornam mais similares geneticamente em relação às respectivas populações base recorrentes. Após um número suficiente de retrocruzamentos, as progênies serão heterozigotas para os alelos em transferência, mas homozigotas para todos aqueles cujo genitor recorrente é homozigoto (Borém, 1998). Com o avanço das gerações de retrocruzamento, sugere-se um aumento de cruzamentos endogâmicos. Como $N_e$ é inversamente

Tabela 12. Estimativas dos tamanhos efetivos ($N_e$[#]) das 42 populações simuladas, obtidas a partir dos diferentes estimadores de coeficiente de fixação/endogamia intrapopulacionais (*f*)

| População | $N_e$ [ς] | $N_e$ [€] | $N_e$ [£] | $N_e$ [¥] |
|---|---|---|---|---|
| $P_1$ | 190,13 | 192,92 | 192,46 | 195,03 |
| $P_2$ | 201,29 | 202,33 | 201,82 | 201,13 |
| $P_3$ | 208,31 | 207,49 | 206,98 | 203,11 |
| $P_1a_1$ | 198,85 | 198,28 | 197,78 | 198,31 |
| $P_2a_1$ | 193,80 | 195,06 | 194,59 | 196,68 |
| $P_2a_2$ | 198,79 | 199,26 | 198,77 | 200,12 |
| $P_2a_3$ | 202,31 | 202,92 | 202,41 | 202,92 |
| $P_2a_4$ | 195,69 | 198,04 | 197,55 | 198,08 |
| $P_3a_1$ | 199,70 | 199,58 | 199,08 | 199,80 |
| $P_2s_1$ | 135,18 | 135,35 | 135,17 | 135,51 |
| $P_2s_2$ | 113,69 | 113,75 | 113,68 | 113,66 |
| $P_2s_3$ | 106,94 | 106,86 | 106,83 | 106,17 |
| $P_2s_4$ | 103,40 | 103,60 | 103,58 | 103,29 |
| $P_2s_5$ | 101,92 | 102,06 | 102,05 | 101,68 |
| $H_{12pp}$ | 249,78 | 240,01 | 239,32 | 225,78 |
| $H_{12mp}$ | 247,65 | 238,86 | 238,18 | 226,73 |
| $H_{13pp}$ | 248,05 | 246,88 | 246,18 | 237,50 |
| $H_{13mp}$ | 250,56 | 249,10 | 248,39 | 239,49 |
| $H_{23pp}$ | 238,21 | 234,49 | 233,84 | 217,37 |
| $F_{2(pp)}$ | 138,55 | 137,39 | 137,20 | 135,60 |
| $F_{3(pp)}$ | 115,16 | 114,58 | 114,51 | 112,30 |
| $F_{4(pp)}$ | 107,38 | 107,02 | 106,99 | 105,56 |
| $H_{23(mp)}$ | 235,96 | 233,40 | 232,75 | 216,01 |
| $F_{2(mp)}$ | 139,30 | 139,27 | 139,07 | 138,75 |
| $F_{3(mp)}$ | 116,67 | 116,41 | 116,33 | 114,06 |
| $F_{4(mp)}$ | 107,51 | 107,54 | 107,50 | 105,75 |
| $RC_{11pp}$ | 209,84 | 212,56 | 212,00 | 210,17 |
| $RC_{12pp}$ | 198,24 | 198,16 | 197,67 | 198,31 |
| $RC_{13pp}$ | 195,47 | 196,58 | 196,10 | 198,04 |
| $RC_{14pp}$ | 196,93 | 198,99 | 198,49 | 198,29 |
| $RC_{11mp}$ | 211,10 | 213,45 | 212,90 | 211,64 |
| $RC_{12mp}$ | 202,41 | 205,89 | 205,36 | 205,78 |
| $RC_{13mp}$ | 200,94 | 200,92 | 200,42 | 199,64 |
| $RC_{14mp}$ | 200,78 | 202,72 | 202,20 | 199,88 |
| $RC_{31pp}$ | 201,65 | 202,51 | 202,00 | 199,58 |
| $RC_{32pp}$ | 200,06 | 201,03 | 200,52 | 201,05 |
| $RC_{33pp}$ | 198,89 | 198,95 | 198,43 | 199,42 |
| $RC_{34pp}$ | 191,66 | 193,65 | 193,18 | 195,31 |
| $RC_{31mp}$ | 210,24 | 210,50 | 209,95 | 208,42 |
| $RC_{32mp}$ | 204,29 | 204,77 | 204,25 | 204,23 |
| $RC_{33mp}$ | 204,92 | 205,36 | 204,83 | 205,51 |
| $RC_{34mp}$ | 196,56 | 198,31 | 197,82 | 199,50 |

[#] $\hat{N}_e = N_i/(1 + f)$, metodologia apresentada por Vencovsky (1997), em $N_i$ é o tamanho amostral da população.

[ς] Coeficiente de fixação/endogamia (*f*) obtido pelo método dos momentos; [€] *f* estimado pelas heterozigosidades observada e esperada; [£] Estimador obtido conforme Weir e Cockerham (1984); [¥] Estimador obtido conforme Robertson e Hill (1984).

proporcional ao coeficiente de endogamia, justifica-se a queda no tamanho efeitvo ao longo das gerações de retrocruzamento, embora os coeficientes *f* ao longo das gerações não tenham sido expressivamente maiores que zero (Tabela 10), uma vez que as populações recorrentes $P_1$ e $P_3$ foram simuladas nas condições de EHW.

As populações que sofreram autofecundações, como $P_2s_1$ a $P_2s_4$ e $F_{2(pp)}$ a $F_{4(pp)}$ e $F_{2(mp)}$ a $F_{4(mp)}$ tiveram os menores tamanhos efetivos, em consonância aos estudos realizados por Moraes et al. (1999) e Moraes et al. (2002) em populações naturais de *Cryptocarya Aschersoniana* (Lauraceae) e *Cryptocarya moschata* Nees (Lauraceae), respectivamente. Ao contrário, populações híbridas ($H_{12pp}$, $H_{12mp}$, $H_{13pp}$, $H_{13mp}$, $H_{23pp}$ e $H_{23mp}$) apresentaram os maiores valores de $N_e$. Cruzamentos mais contrastantes, ou seja, $H_{12pp}$, $H_{12mp}$, $H_{13pp}$ e $H_{13mp}$ proporcionaram valores de $N_e$ ainda mais elevados. Melo Júnior et al. (2004) estudando 5 populações naturais de pequizeiro (*Caryocar brasiliense* Camb.) obtiveram a maioria das relações $N_e/N_i$ superiores a unidade, sugerindo que os valores de $N_e$ representam, geneticamente, populações com grande potencial panmítico. Por conseqüência, afirmaram a inexistência de endogamia nessas populações, uma vez que o tamanho efetivo calculado para cada uma delas foi superior ao número de indivíduos amostrados. Além disso, as taxas de heterozigosidade foram altas, à semelhança do presente estudo (Tabela 8 e 10). Os heterozigotos por carregarem dois alelos diferentes por loco, representam um maior número de indivíduos na população (Póvoa, 2002).

As populações $H_{12pp}$ e $H_{23pp}$ tiveram seus tamanhos efetivos ligeiramente superiores aquelas obtidas por mistura de pólen ($H_{12mp}$ e $H_{23mp}$). Populações de retrocruzamentos de hibridação por mistura de pólen apresentaram regularmente $N_e$'s superiores as respectivas populações de retrocruzamento de planta a planta, à semelhança do ocorrido com os híbridos $H_{13pp}$ e $H_{13mp}$.

Em virtude do tamanho amostral utilizado, diferenças entre o tamanho efetivo de uma população obtida por hibridação planta a planta e por mistura de pólen não foram efetivamente constatadas. No entanto, populações pequenas, que geram progênies de irmãos-completos (cruzamentos planta a planta) terão menor *pool* genético do que as de

meios-irmãos (cruzamentos por *bulk*, ou mistura de pólen). Sebbenn et al. (2005) observaram que em populações não exploradas de caixeta (*Tabebuia cassinoides* (Lamarck) A. P. de Candolle), espécie arbórea tropical, o tamanho efetivo reduziu-se devido à maior taxa de autofecundação e aos cruzamentos planta a planta (irmãos-completos).

Para Vencovsky (1987) a questão de coletar e preservar germoplasmas não pode ser vista apenas sob a ótica de alelos individuais. Blocos gênicos também têm grande importância para o melhoramento, por isso, manter tamanhos efetivos maiores, quando possível, não deixa de ser uma estratégia apropriada.

### 4.1.3. Equilíbrio de Hardy-Weinberg

Uma população, suficientemente grande e sob acasalamento ao acaso, mantém constantes suas freqüências gênicas e genotípicas ao longo de gerações, na ausência de migração, mutação e seleção. Assim freqüências genotípica são determinadas pelas freqüências gênicas (Falconer, 1987). Estas premissas invocam o equilíbrio de Hardy-Weinberg (EHW).

Um dos modelos mais importantes na genética de populações é o acasalamento ao acaso, o que significa que cada indivíduo tem igual possibilidade de se acasalar com qualquer outro indivíduo da população. Em outras palavras, diz-se que os casais têm as mesmas freqüências de se acasalarem, como se fossem gerados pela “colisão” aleatória entre indivíduos. Isto significa dizer que a chance de um organismo, de genótipo definido, se acasalar com outro, é igual a freqüência deste genótipo na população (Hartl & Clark, 1997). No entanto, o ponto importante é que não haverá tendências para que o acasalamento ocorra entre indivíduos com genótipos semelhantes ou entre aqueles relacionados por ascendência (Falconer, 1987).

O EHW foi testado em todas as populações, por três diferentes métodos: o teste de qui-quadrado ($\chi^2$), o teste de razão de verossimilhança,

conhecido como teste $G^2$ e o teste exato de Fisher. No geral, os testes foram concordantes quanto à detecção de (des)equilíbrio nos locos simulados. A exceção de poucos locos (Tabela 13), as populações base mostraram estar realmente em EHW.

A grande maioria dos locos das populações de acasalamento ao acaso também seguiram as proporções do EHW, conforme o esperado. Mas em alguns poucos locos a hipótese de nulidade ($H_0$: EHW, ou encontro ao acaso de gametas) foi rejeitada ($P > 0,05$). O loco 16 (Tabela 13), não encontrava-se em desequilíbrio na $P_2$, mas nas duas gerações subseqüentes de acasalamento ao acaso ($P_2a_1$ e $P_2a_2$) encontrou-se em estado de desequilíbrio e nas duas últimas gerações ($P_2a_3$ e $P_2a_4$), retornou ao estado de EHW. A esta circunstância, novamente, atribuí-se ao processo de simulação, onde ligeiras variações das freqüências alélicas contribuíram com este acontecimento.

Todas as populações que passaram pelo processo de autofecundação, foram unânimes quanto ao desequilíbrio de seus locos (dados não apresentados). Isto porque, o excesso de genótipos homozigotos, ou a deficiência de heterozigotos (Tabela 8 e Figura 3), promovido pela endogamia, gera desvios nas proporções do EHW (Rousset & Raymond, 1995).

As populações híbridas, a exemplo de $H_{12pp}$, apresentaram a maioria dos seus locos em desequilíbrio. Um fato interessante foi observado no loco 5, em desequilíbrio nas populações $P_1$ e $P_2$, passou a estar em EHW na população híbrida $H_{12pp}$ (Tabela 13). Isto porque as freqüências gênicas no loco 5 foram alteradas durante a hibridação, a ponto de caracterizar o loco em estado de equilíbrio. O EHW ocorrerá para os genes em que não há diferenças de freqüências alélicas entre as populações cruzadas.

As populações de retrocruzamento tiveram a grande maioria dos seus locos em equilíbrio (dados não apresentados). O processo de hibridação no retrocruzamento pode levar um loco ao EHW ou não, o que vai depender das combinações gênicas e genotípicas entre o hibrido $F_1$, ou a geração antecedente de retrocruzamento, e o genitor recorrente, considerando ainda a não execução de seleção na população de retrocruzamento. Este fato pode ser constatado através das populações

$RC_{34pp}$ e $RC_{34mp}$, que apresentaram diferentes locos em (des)equilíbrio (Tabela 13).

Tabela 13. Resumo da análise de equilíbrio de Hardy-Weinberg[#] (EHW) pelos testes de qui-quadrado (a), razão de verossimilhança (b) e teste exato (c) para as populações base ($P_1$, $P_2$ e $P_3$), de acasalamento ao acaso ($P_1a_1$, $P_2a_1$, e $P_3a_1$), híbrida ($H_{12pp}$) e de retrocruzamento ($RC_{34pp}$ e $RC_{34mp}$) em 20 locos codominantes e multialélicos gerados via simulação

| População | Loco | Graus de liberdade | Teste | | |
|---|---|---|---|---|---|
| | | | a | b | c[##] |
| $P_1$ | 5 | 3 | ns | ns | *(1,2,7) |
| $P_2$ | 5 | 10 | *(4,6,7) | ns | *[1] |
| $P_3$ | 10 | 6 | *(4,5,6,7) | *(5,7) | *(1,3,6,7) |
| $P_3$ | 15 | 3 | ns | *(7) | ns |
| $P_1a_1$ | 7 | 1 | *(4,5,6,7) | ns | *(2,7) |
| $P_1a_1$ | 19 | 3 | ns | *(7) | ns |
| $P_2a_1$ | 3 | 1 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $P_2a_1$ | 7 | 6 | *(4,5,6,7) | ns | ns |
| $P_2a_1$ | 16 | 3 | *(4,5,6,7) | *(5,6) | *(1,6) |
| $P_2a_2$ | 16 | 3 | *(4,5,6,7) | *(5,6) | *(1,3,6,7) |
| $P_3s_1$ | 18 | 6 | ns | *(7) | *(1,2,6,7) |
| $H_{12pp}$ | 1 | 3 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 2 | 10 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 3 | 1 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 4 | 6 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 8 | 3 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 9 | 6 | ns | *(7) | *(1,2,7) |
| $H_{12pp}$ | 11 | 3 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 14 | 6 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 16 | 3 | *(4,6,7) | *(6,7) | *(1,2,3,6,7) |
| $H_{12pp}$ | 18 | 6 | *(4,5,6,7) | *(5,6,7) | *(1,3,6,7) |
| $H_{12pp}$ | 19 | 3 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $RC_{34pp}$ | 13 | 3 | *(4,5,6,7) | *(5,6,7) | *(1,2,3,6,7) |
| $RC_{34pp}$ | 16 | 10 | ns | *(7) | ns |
| $RC_{34pp}$ | 17 | 6 | *(4,5,6,7) | ns | *(2,6) |
| $RC_{34pp}$ | 18 | 6 | ns | *(7) | *(1,2) |
| $RC_{34pp}$ | 19 | 3 | *(4,5,6,7) | *(5) | *(1,2) |
| $RC_{34pp}$ | 20 | 6 | *(4,5,6,7) | *(5) | *(1,2,3,6,7) |
| $RC_{34mp}$ | 4 | 3 | *(4,5,6,7) | *(5,6,7) | * (1,2,3,6,7) |
| $RC_{34mp}$ | 10 | 6 | ns | *(7) | ns |
| $RC_{34mp}$ | 15 | 3 | ns | *(7) | ns |

Programas utilizados: [1]Arlequin; [2]GDA; [3]GENEPOP; [4]GENES; [5]POPGENE; [6]PowerMarker; [7]TFPGA.

[#] $H_0$: EHW. [##] Valor de *p* obitdo a partir de 1000 dememorizações e 100 *batches*.

*(P < 0,05); ns (P > 0,05)

Quando um loco em uma população segue as proporções do EHW, sugere-se que o(s) coeficiente(s) de desequilíbrio (D's) dentro do loco seja(m) igual(is) a zero. Portanto, o teste de hipótese para o EHW pode ser construído alegando-se que $H_0$: D = 0. Para grandes amostras, a estimativa de verossimilhança de D apresenta aproximadamente distribuição normal (Weir, 1996). Como o quadrado de uma variável normal padronizada possui distribuição qui-quadrado com um grau de liberdade (Ferreira, 2005), os desvios do EHW, podem ser testados pelo teste de qui-quadrado ($\chi^2$). Embora o índice de fixação/endogamia (*f*) reflita desvios do EHW, Weir (1996) atribui desvantagem a este parâmetro em descrever tais desvios. Segundo o autor, como o coeficiente *f* é estimado como razões de freqüências genotípicas, é difícil determinar as propriedades estatísticas destas razões. Testes estatísticos também podem ser definidos como razões de verossimilhança ($\lambda$), em que uma aproximação de $\chi^2$ torna-se apropriada a distribuição de $\lambda$ (Weir, 1996).

O problema que circunda estes dois testes é que eles são sensíveis a pequenos valores esperados nas classes genotípicas. O que se tem feito no teste de qui-quadrado é usar a correção de Yates (1934). Porém, em situações de múltiplos alelos, em que alguns possuem freqüências bem pequenas, a proposta de correção ou até mesmo uso do $\chi^2$ clássico pode levar a resultados inadequados (Guo & Thompson, 1992). Outra solução seria agrupar classes genotípicas de modo a aumentar o número esperado de uma classe. Mas esta é uma solução pobre do ponto de vista estatístico, pois, obviamente há perda de informação de classes genotípicas. Alternativamente, a literatura tem adotado o teste exato de Fisher. Este teste de probabilidade define um ordenamento crescente das probabilidades condicionais [Pr(S)] de todas as possíveis amostras (arranjos genotípicos ou tabelas de contingência 2 x 2), mantendo-se o número observado de alelos do loco (quantidade marginal da tabela de contingência). O valor de *p* está associado à soma de todas as Pr(S)'s menores que a Pr(S) do arranjo genotípico observado.

O teste exato apresenta algumas particularidades interessantes: (a) não faz uso de distribuição assintótica; (b) é uma distribuição de

probabilidade independente de parâmetros (desconhecidos) sobre a hipótese de nulidade, um importante requerimento que junto com (a) leva ao uso de distribuições condicionais particulares, como a apresentada por Haldane (1954), que independe das freqüências (alélicas e genotípicas) paramétricas; e (c) usa as probabilidades de uma particular configuração (arranjos genotípicos) como um teste estatístico (Rousset & Raymond, 1995).

Levene (1949) mostrou que probabilidades condicionais podem ser obtidas por uma razão de verossimilhança entre a função de verossimilhança de uma particular amostra S e a soma das funções de verossimilhança de todas as possíveis amostras. Assim a distribuição condicional de qualquer estatística sobre a hipótese de nulidade pode ser computada. Diferentes testes estatísticos definem diferentes ordenamentos de possíveis amostras, porém o valor de *p* é definido igualmente como a soma de probabilidades exatas de amostras de ordem mais extrema, de modo que todos os testes são exatos (Rousset & Raymond, 1995)

Outro problema surge quando o número de indivíduos é elevado e os locos investigados possuem vários alelos, de maneira que obter todas as tabelas de contingência torna-se computacionalmente “impossível”. Técnicas de permutação, como métodos MCMC (método Monte Carlo e cadeia de Markov) tem sido a alternativa mais viável e utilizada pelos programas computacionais da área de genética populacional (Guo & Thompson, 1992).

### 4.1.4. Desequilíbrio gamético

O termo desequilíbrio gamético também tem sido referido como desequilíbrio de ligação, desequilíbrio de fase gamética e associação alélica (Flint-Garcia et al., 2003). Trata-se da associação não aleatória de alelos de diferentes locos nos gametas. Em uma população panmítica, com locos segregando independentemente, na ausência de forças evolutiva, locos polimórficos estarão em equilíbrio gamético (Falconer & Mackay, 1996). O conhecimento do desequilíbrio permite elucidar fenômenos genéticos e evolutivos ocorridos ao longo de gerações em populações ou espécies.

Na Tabela 14 estão apresentadas as estimativas de três medidas de desequilíbrio e os resultados de três testes para o equilíbrio gamético, em alguns pares de locos da população base $P_2$ e gerações subseqüentes de acasalamento ao acaso ($P_2a_1$, $P_2a_2$, $P_2a_3$ e $P_2a_4$). Dos 40 resultados computados, em apenas 15 houve concordância do teste de qui-quadrado ($\chi^2$) com os dois testes de razões de verossimilhança (LOD e $G^2$) quanto à significância ($P < 0,05$) ou não significância ($P > 0,05$). No entanto, 34 resultados foram concordantes entre o teste $\chi^2$ e $G^2$. O par de locos 3 e 14 foi o único em que houve concordância na detecção de desequilíbrio com os três testes utilizados.

Não foi constatado desequilíbrio gamético para uma grande quantidade de pares de locos nestas populações, pois $P_2$ foi simulada nas condições de EHW. Houve tendência de queda nos coeficientes de desequilíbrio ao longo das gerações de acasalamento ao acaso. A diferença entre $D_{11,21}$, $D'_{11,21}$ e $r^2$ recai sobre suas escalas. O coeficiente de desequilíbrio $r^2$ refere-se ao quadrado do coeficiente de correlação entre dois locos. Contudo, ao menos que estes dois locos tenham idênticas freqüências alélicas, um valor igual a unidade (1) não é possível. Já $D'_{11,21}$ tem sua escala baseada na freqüência alélica observada, de modo que ele varia de 0 a 1 mesmo se as freqüências alélicas diferirem entre os locos. $D'_{11,21}$ será menor que 1 se todos os quatro possíveis haplótipos (gametas) forem observados. No entanto, presume-se que um evento de recombinação tenha ocorrido entre estes dois locos (Flint-Garcia et al., 2003). Embora nem $r^2$ nem $D'_{11,21}$ mostrem-se adequados para pequenos tamanhos amostrais e pequenas freqüências alélicas, cada um tem vantagens distintas. Segundo Flint-Garcia et al. (2003), $r^2$ consegue resumir um histórico mutacional e de recombinação, enquanto $D'_{11,21}$ mede apenas a história de recombinação, sendo, portanto, uma estatística mais acurada para estimar diferenças de recombinação.

Tabela 14. Resumo dos testes de qui-quadrado e razão de verossimilhança (LOD e $G^2$) do equilíbrio gamético para alguns pares de locos na população base $P_2$ e gerações subseqüentes de acasalamento ao acaso, $P_2a_1$, $P_2a_2$, $P_2a_3$ e $P_2a_4$

| Par de locos | População | Freqüências alélicas | | | | Medidas de desequilíbrio | | | Qui-quadrado[6] | Verossimilhança | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $p_{i11}$[#] | $p_{i12}$ | $p_{i21}$ | $p_{i22}$ | $D_{11,21}$[4,6] | $D'_{11,21}$ | $r^2$ | | LOD[4] | $G^{2\ (1)}$ |
| | $P_2$ | 0,5400 | 0,4600 | 0,8850 | 0,1150 | 0,0621 | 1,0000 | 0,1525 | 61,0169* | 10,1181 * | * |
| | $P_2a_1$ | 0,5775 | 0,4225 | 0,9125 | 0,0875 | 0,0399 | 0,7902 | 0,0818 | 32,7373* | 4,6152 * | * |
| 3 e 14 | $P_2a_2$ | 0,5850 | 0,4150 | 0,9100 | 0,0900 | 0,0394 | 0,7485 | 0,0781 | 31,2438* | 4,7083 * | * |
| | $P_2a_3$ | 0,5625 | 0,4375 | 0,9025 | 0,0975 | 0,0438 | 0,7978 | 0,0884 | 35,3663* | 4,7635 * | * |
| | $P_2a_4$ | 0,5550 | 0,4450 | 0,9025 | 0,0975 | 0,0425 | 0,7857 | 0,0832 | 33,2741* | 4,2407 * | * |
| | $P_2$ | 0,7100 | 0,2900 | 0,9125 | 0,0875 | 0,0254 | 1,0000 | 0,0392 | 15,6666* | -0,1114 ns | * |
| | $P_2a_1$ | 0,7200 | 0,2800 | 0,9275 | 0,0725 | 0,0116 | 0,5714 | 0,0099 | 3,9694* | -0,0704 ns | ns |
| 4 e 8 | $P_2a_2$ | 0,7350 | 0,2650 | 0,9375 | 0,0625 | 0,0005 | 0,0308 | 0,0000 | 0,0091ns | -0,0093 ns | ns |
| | $P_2a_3$ | 0,7250 | 0,2750 | 0,9300 | 0,0700 | 0,0114 | 0,5897 | 0,0099 | 3,9716* | -0,0763 ns | * |
| | $P_2a_4$ | 0,7450 | 0,2550 | 0,9200 | 0,0800 | 0,0034 | 0,1676 | 0,0008 | 0,3343 ns | -0,0233 ns | ns |
| | $P_2$ | 0,7650 | 0,2350 | 0,8225 | 0,1775 | 0,0256 | 0,6126 | 0,0249 | 9,9522* | -0,0955 ns | * |
| | $P_2a_1$ | 0,7650 | 0,2350 | 0,8375 | 0,1625 | 0,0202 | 0,5283 | 0,0166 | 6,6548* | -0,0774 ns | ns |
| 5 e 10 | $P_2a_2$ | 0,7675 | 0,2325 | 0,8300 | 0,1700 | 0,0328 | 0,8302 | 0,0428 | 17,1059* | -0,1101 ns | * |
| | $P_2a_3$ | 0,7675 | 0,2325 | 0,8225 | 0,1775 | 0,0282 | 0,6834 | 0,0305 | 12,2132* | -0,0931 ns | * |
| | $P_2a_4$ | 0,7575 | 0,2425 | 0,8275 | 0,1725 | 0,0232 | 0,5553 | 0,0206 | 8,2315* | -0,0965 ns | * |
| | $P_2$ | 0,8300 | 0,1700 | 0,6025 | 0,3975 | 0,0226 | 0,2208 | 0,0151 | 6,0526* | 0,6511 ns | * |
| | $P_2a_1$ | 0,8450 | 0,1550 | 0,6300 | 0,3700 | 0,0203 | 0,2080 | 0,0135 | 5,4051* | 0,6530 ns | ns |
| 6 e 16 | $P_2a_2$ | 0,8475 | 0,1525 | 0,6150 | 0,3850 | 0,0052 | 0,0890 | 0,0009 | 0,3570ns | -0,0230 ns | ns |
| | $P_2a_3$ | 0,8450 | 0,1550 | 0,5850 | 0,4150 | 0,0064 | 0,0707 | 0,0013 | 0,5164 ns | 0,0552 ns | ns |
| | $P_2a_4$ | 0,8325 | 0,1675 | 0,5675 | 0,4325 | 0,0129 | 0,1356 | 0,0049 | 1,9411 ns | 0,1661 ns | ns |

Tabela 14. Continuação...

| Par de locos | População | Freqüências alélicas | | | | Medidas de desequilíbrio | | | Qui-quadrado[6] | Verossimilhança | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $p_{i11}$[#] | $p_{i12}$ | $p_{i21}$ | $p_{i22}$ | $D_{11,21}$[4,6] | $D'_{11,21}$ | $r^2$ | | LOD[4] | $G^2$ (1) |
| 8 e 12 | $P_2$ | 0,9125 | 0,0875 | 0,8350 | 0,1650 | 0,0301 | 0,4124 | 0,0825 | 33,0184* | 3,1305* | * |
| | $P_2a_1$ | 0,9275 | 0,0725 | 0,8400 | 0,1600 | 0,0135 | 0,2225 | 0,0203 | 8,1232* | 0,9576 ns | * |
| | $P_2a_2$ | 0,9375 | 0,0625 | 0,8550 | 0,1450 | 0,0166 | 0,3109 | 0,0380 | 15,1970* | 1,4659 ns | * |
| | $P_2a_3$ | 0,9300 | 0,0700 | 0,8650 | 0,1350 | 0,0036 | 0,0595 | 0,0017 | 0,6826ns | 0,0702 ns | ns |
| | $P_2a_4$ | 0,9200 | 0,0800 | 0,8500 | 0,1500 | 0,0010 | 0,0152 | 0,0001 | 0,0455 ns | 0,0047 ns | ns |
| 9 e 18 | $P_2$ | 0,8150 | 0,1850 | 0,3900 | 0,6100 | 0,0315 | 0,2789 | 0,0276 | 11,0463* | -0,0756 ns | * |
| | $P_2a_1$ | 0,8600 | 0,1400 | 0,4200 | 0,5800 | 0,0447 | 0,5505 | 0,0681 | 27,2495* | -0,1345 ns | * |
| | $P_2a_2$ | 0,8650 | 0,1350 | 0,4300 | 0,5700 | 0,0384 | 0,4993 | 0,0516 | 20,6331* | -0,1119 ns | * |
| | $P_2a_3$ | 0,8650 | 0,1350 | 0,4225 | 0,5775 | 0,0193 | 0,2480 | 0,0131 | 5,2494* | -0,0655 ns | ns |
| | $P_2a_4$ | 0,8675 | 0,1325 | 0,4200 | 0,5800 | 0,0354 | 0,4612 | 0,0449 | 17,9457* | -0,1297 ns | * |
| 10 e 12 | $P_2$ | 0,8225 | 0,1775 | 0,8350 | 0,1650 | 0,0293 | 1,0000 | 0,0426 | 17,0577* | -0,1464 ns | * |
| | $P_2a_1$ | 0,8375 | 0,1625 | 0,8400 | 0,1600 | 0,0260 | 1,0000 | 0,0370 | 14,7832* | -0,1654 ns | * |
| | $P_2a_2$ | 0,8300 | 0,1700 | 0,8550 | 0,1450 | 0,0247 | 1,0000 | 0,0347 | 13,8942* | -0,0974ns | * |
| | $P_2a_3$ | 0,8225 | 0,1775 | 0,8650 | 0,1350 | 0,0240 | 1,0000 | 0,0337 | 13,4722* | -0,1439 ns | * |
| | $P_2a_4$ | 0,8275 | 0,1725 | 0,8500 | 0,1500 | 0,0259 | 1,0000 | 0,0368 | 14,7148* | -0,1292 ns | * |
| 13 e 16 | $P_2$ | 0,5350 | 0,4650 | 0,6025 | 0,3975 | 0,0257 | 0,1392 | 0,0111 | 4,4416* | -0,0367 ns | * |
| | $P_2a_1$ | 0,5100 | 0,4900 | 0,6300 | 0,3700 | 0,0213 | 0,1174 | 0,0078 | 3,1084 ns | -0,0345 ns | ns |
| | $P_2a_2$ | 0,5150 | 0,4850 | 0,6150 | 0,3850 | 0,0179 | 0,0957 | 0,0054 | 2,1619 ns | -0,0279 ns | ns |
| | $P_2a_3$ | 0,5200 | 0,4800 | 0,5850 | 0,4150 | 0,0259 | 0,1302 | 0,0111 | 4,4381* | -0,0378 ns | ns |
| | $P_2a_4$ | 0,5450 | 0,4550 | 0,5675 | 0,4325 | 0,0259 | 0,1315 | 0,0110 | 4,4037* | -0,0416 ns | ns |

[#] $p_{ijk}$, freqüência alélica do alelo k(= 1, 2), pertencente ao loco j (1, 2) na população i(=$P_2$, $P_2a_1$, $P_2a_2$, $P_2a_3$ e $P_2a_4$).

* (P < 0,05) e ns (P > 0,05).

Programas utilizados: [1]Arlequin, [4]GENES e [6]PowerMarker.

Mecanismos que promovem desequilíbrio gamético são a seleção, a ligação fatorial e deriva genética, aleatória e não aleatória (Ridley, 2006). Se a seleção favorece indivíduos com combinações particulares de alelos, então ela produz desequilíbrio gamético. Para locos ligados, um número maior de gerações é necessário para que a recombinação realize a sua função de tornar as associações genéticas aleatórias. Locos fracamente ligados não irão apresentar desequilíbrio de ligação por muito tempo. Na ausência de seleção e em uma população infinita e de cruzamentos aleatórios, a quantidade de desequilíbrio de ligação sofre uma queda exponencial a uma taxa igual à de recombinação entre dois locos.

Para os pares de locos 5-10, 6-16 e 9-18, duas geração de acasalamento ao acaso foram suficientes para que entrassem em equilíbrio, enquanto o par 8-12 necessitou de três gerações, considerando as informações provenientes dos testes de desequilíbrio conjuntamente. Para locos fortemente ligados, o desequilíbrio gamético pode persistir indefinidamente, a exemplo do que aconteceu nos pares 3-14 e 10-12, em que quatro gerações não foram suficientes para retirá-los da condição de desequilíbrio gamético.

Processos aleatórios possuem a propriedade interessante de serem capazes de causar desequilíbrio de ligação persistente, não apenas transitório. Se a deriva aleatória produz, ao acaso, um excesso de um haplótipo em uma geração, o desequilíbrio gamético terá aparecido, o que não aconteceu ao longo das gerações nos pares de locos apresentados na Tabela 14. Acrescenta-se que isto pode ser verdadeiro para todos os quatro haplótipos: a amostragem aleatória que produz um excesso de qualquer um deles irá perturbar o estado de equilíbrio. Qualquer gameta poderá ser “favorecido” ao acaso, de forma que é igualmente provável que o desequilíbrio seja $D > 0$ ou $D < 0$ (Ridley, 1996).

Estas associações persistem por mais tempo em locos fortemente ligados, de modo que quanto mais elevada for a taxa de recombinação mais rápida será a destruição da associação. Entretanto, como a taxa de recombinação entre dois locos diminui, o tempo que os alelos podem estar associados entre si de forma não-aleatória aumenta (Ridley, 2006).

Outro fator gerador de desequilíbrio gamético diz respeito aos cruzamentos não aleatórios. Geralmente o desequilíbrio reduz-se mais rapidamente em espécies alógamas, quando comparado com espécies de autofecundação (Nordborg, 2000). Cruzamentos não aleatórios proporcionam aumentos (ou diminuição) de certos haplótipos, fazendo com que eles tenham uma freqüência em excesso (deficiência) sobre a de cruzamentos aleatórios. A alta homozigosidade dos genes, em espécies autógamas, implica que recombinação raramente resultará novos haplótipos que ainda não estão presentes nos parentais. A predominância de autofecundação tende a retardar a proximidade ao equilíbrio gamético, porque para atingi-lo são necessárias recombinações entre duplos heterozigotos, que são raros nas populações autógamas (Hartl & Clark, 1997).

Outro fator, como fluxo gênico entre indivíduos de populações geneticamente distintas e seguidos por intercruzamentos, resultam na introdução de diferentes cromossomos ancestrais e diferentes freqüências alélicas. Frequentemente, o resultado do desequilíbrio se estende a sítios não ligados, mesmo em diferentes cromossomos, mas que são quebrados rapidamente com o processo de acasalamento ao acaso (Pritchard & Rosenberg, 1999).

### 4.1.5. Excesso e deficiência de heterozigotos

Os desvios das proporções do EHW indicam que efeitos da seleção, mistura de populações e acasalamentos não aleatórios ocorrem na população e é este o primeiro passo nas investigações sobre estruturação genética de populações (Rousset & Raymond, 1995).

A Tabela 15 apresenta os valores de probabilidade associados ao teste U global de excesso e deficiência de heterozigotos nas 42 populações. Não houve populações base com excesso de heterozigotos ($P > 0,05$), embora $P_3$ tenha tido um valor de $P = 0,0768$. O coeficiente de fixação/endogamia para $P_3$ foi negativo (Tabela 10), apesar de negligenciável, não sendo suficiente para caracterizá-la como uma

Tabela 15. Valores de probabilidade (P) associados ao teste U[3] para excesso e deficiência de heterozigotos (global) a partir de 20 locos codominantes e multialélicos em 42 populações de melhoramento simuladas

| População | Teste Global* | |
|---|---|---|
| | Excesso | Deficiência |
| $P_1$ | 0,9787 | 0,0187* |
| $P_2$ | 0,4476 | 0,5675 |
| $P_3$ | 0,0768 | 0,9392 |
| $P_1a_1$ | 0,7398 | 0,2748 |
| $P_2a_1$ | 0,8465 | 0,1525 |
| $P_2a_2$ | 0,4761 | 0,5208 |
| $P_2a_3$ | 0,0632 | 0,9337 |
| $P_2a_4$ | 0,8213 | 0,1704 |
| $P_3a_1$ | 0,3524 | 0,6089 |
| $P_2s_1$ | 1,0000 | 0,0000* |
| $P_2s_2$ | 1,0000 | 0,0000* |
| $P_2s_3$ | 1,0000 | 0,0000* |
| $P_2s_4$ | 1,0000 | 0,0000* |
| $P_2s_5$ | 1,0000 | 0,0000* |
| $H_{12pp}$ | 0,0000* | 1,0000 |
| $H_{12mp}$ | 0,0000* | 1,0000 |
| $H_{13pp}$ | 0,0000* | 1,0000 |
| $H_{13mp}$ | 0,0000* | 1,0000 |
| $H_{23pp}$ | 0,0000* | 1,0000 |
| $F_{2(pp)}$ | 1,0000 | 0,0000* |
| $F_{3(pp)}$ | 1,0000 | 0,0000* |
| $F_{4(pp)}$ | 1,0000 | 0,0000* |
| $H_{23(mp)}$ | 0,0000* | 1,0000 |
| $F_{2(mp)}$ | 1,0000 | 0,0000* |
| $F_{3(mp)}$ | 1,0000 | 0,0000* |
| $F_{4(mp)}$ | 1,0000 | 0,0000* |
| $RC_{11pp}$ | 0,0000* | 1,0000 |
| $RC_{12pp}$ | 0,8310 | 0,2064 |
| $RC_{13pp}$ | 0,8307 | 0,1557 |
| $RC_{14pp}$ | 0,4995 | 0,4995 |
| $RC_{11mp}$ | 0,0000* | 1,0000 |
| $RC_{12mp}$ | 0,0122* | 0,9871 |
| $RC_{13mp}$ | 0,6840 | 0,3108 |
| $RC_{14mp}$ | 0,6494 | 0,3617 |
| $RC_{31pp}$ | 0,7565 | 0,2548 |
| $RC_{32pp}$ | 0,3831 | 0,6313 |
| $RC_{33pp}$ | 0,3338 | 0,6703 |
| $RC_{34pp}$ | 0,9979 | 0,0033* |
| $RC_{31mp}$ | 0,0000* | 1,0000 |
| $RC_{32mp}$ | 0,0385* | 0,9637 |
| $RC_{33mp}$ | 0,0201* | 0,9809 |
| $RC_{34mp}$ | 0,6390 | 0,4060 |

[3] Testes realizados pelo programa GENEPOP.

* Significativo quando P < 0,05.

população com excesso de heterozigotos. Para $P_1$ foi detectado deficiência de heterozigotos ($P < 0,05$), o que vai de encontro com as análises anteriores. As populações de acasalamento ao acaso, como esperado, não apresentaram nem excesso nem deficiência de heterozigotos. Desvios no EHW causam variação no(s) coeficiente(s) de desequilíbrio dentro do loco (D's), de modo que estes desvios são atribuídos a um excesso ou deficiência de heterozigotos, levando a valores negativos ou positivos de desequilíbrio, respectivamente, o que permite a formulação de hipótese alternativas (Weir, 1996).

As populações de autofecundação e gerações $F_n$ apresentaram-se com deficiência de heterozigotos. Já as populações de híbridos $F_1$ e primeiras gerações de retrocruzamento, a exceção de $RC_{31pp}$, foram consideradas pelo teste U como detentoras de excesso de heterozigotos. Espera-se que o excesso de heterozigotos seja proporcinal às diferenças de freqüências alélicas das populações cruzadas. Em populações naturais arbóreas, o excesso de heterozigotos corresponde a um impedimento sistemático de cruzamentos endogâmicas (Kageyama et al., 2003).

Pelos valores de probabilidade, à medida que as gerações de retrocruzamento foram avançando houve tendência de se ter um balanço entre heterozigotos e homozigotos ($P > 0,05$), conforme o esperado, pois as freqüências genotípicas tenderão para aquelas de acasalamento ao acaso do genitor recorrente. As populações $RC_{11pp}$, $RC_{11mp}$, $RC_{12mp}$, $RC_{31mp}$, $RC_{32mp}$ e $RC_{33mp}$, exibiram excesso de heterozigotos, de acordo com o teste U. Resultado que não causa estranheza, pois elas apresentaram altas heterozigosidades observadas e esperadas (Tabelas 8 e 10).

## 4.2. Diversidade genética em nível interpopulacional

### 4.2.1. Divergência genética

Grande parte dos estudos de diversidade genética baseia-se em uma amostra aleatória de locos obtida em populações não estruturadas hierarquicamente. Inúmeras medidas têm sido propostas para expressar o

grau de (dis)similaridade entre amostras de populações (Dias, 1998). No campo do melhoramento de plantas quantificar o grau de dissimilaridade entre gêneros, espécies, subespécies, populações e materiais elites melhorados também tem sido primordial. Neste sentido, os marcadores moleculares têm contribuído, com destaque, na detecção do parentesco genético entre diferentes germoplasmas em bancos de sementes e programas de melhoramento; na predição da heterose; busca por grupos heteróticos promissores para constituição de híbridos; identificação de duplicatas nos bancos de germoplasma; na avaliação do fluxo gênico ao longo do tempo e na identificação de variedades essencialmente derivadas de planta protegida.

As Tabelas 16 a 22 mostram o agrupamento realizado pelo método de Tocher das 42 populações. No geral, a conformação dos grupos foi semelhante para as treze medidas de distância utilizadas. Agrupamentos idênticos foram obtidos entre as distâncias: Euclidiana média ($D_E$) e Roger ($D_R$) (Tabela 16); Latter (1972 e 1973) ($D_{L72}$ e $D_{L73}$) e Reynolds - desconsiderando os termos envolvidos no tamanho amostral - ($D'_{RWC}$) (Tabela 18); de Rogers modificada ($D_{GS}$), mínima ($D_m$) e Reynolds ($D_{RWC}$) (Tabela 19); complemento do cosseno ($D_{COS}$) e Nei et al. (1983) ($D_{N83}$) (Tabela 20). Agrupamentos diferenciados foram obtidos pela distância do comprimento da corda ($D_{CC}$), distância genotípica de Hedrick ($D_H$) e distância genética padronizada de Nei ($D_{N72}$), nas Tabelas 17, 21 e 22, respectivamente. Com as distâncias $D_E$, $D_R$, $D_{CC}$ e $D_{N72}$, o método de Tocher agrupou a população base $P_2$ junto as suas gerações de autofecundação e acasalamento ao acaso. Quando o método de otimização fez uso das distâncias $D_{L72}$, $D_{L73}$, $D'_{RWC}$, $D_{GS}$, $D_m$, $D_{RWC}$, $D_{COS}$ e $D_{N83}$, $P_2$ foi agrupada apenas com as populações de autofecundação. Já com a distância $D_H$, a população $P_2$ formou grupo com as populações de acasalamento ao acaso.

Tabela 16. Agrupamento definido pelo método de Tocher gerado a partir da distância Euclidiana média ($D_E$) e de Roger ($D_R$), para as 42 populações de melhoramento simuladas

| Grupo | Populações |
|---|---|
| 1 | 13 14 12 11 10 2 8 7 5 6 |
| 2 | 21 22 20 19 23 24 25 26 |
| 3 | 3 9 38 37 |
| 4 | 30 34 1 4 33 29 |
| 5 | 15 16 |
| 6 | 17 18 |
| 7 | 35 39 |
| 8 | 27 31 |
| 9 | 28 32 |
| 10 | 40 41 |
| 11 | 36 |
| 12 | 42 |

Tabela 17. Agrupamento definido pelo método de Tocher gerado a partir da distância genotípica de Hedrick ($D_H$), para as 42 populações de melhoramento simuladas

| Grupo | Populações |
|---|---|
| 1 | 13 14 12 11 |
| 2 | 5 6 7 8 2 |
| 3 | 22 26 21 25 |
| 4 | 33 34 30 29 1 4 28 32 |
| 5 | 19 23 |
| 6 | 15 16 |
| 7 | 20 24 |
| 8 | 35 39 |
| 9 | 17 18 |
| 10 | 27 31 |
| 11 | 3 9 38 37 41 42 |
| 12 | 36 40 |
| 13 | 10 |

Tabela 18. Agrupamento definido pelo método de Tocher gerado a partir da distância de Latter (1972 e 1973) ($D_{L72}$ e $D_{L73}$) e de Reynolds ($D'_{RWC}$), para as 42 populações de melhoramento simuladas

| Grupo | Populações |
|---|---|
| 1 | 13 14 12 11 10 2 |
| 2 | 21 22 20 19 23 24 25 26 |
| 3 | 5 6 7 8 |
| 4 | 3 9 38 37 42 |
| 5 | 30 34 1 29 33 |
| 6 | 15 16 |
| 7 | 17 18 |
| 8 | 35 39 |
| 9 | 27 31 |
| 10 | 40 41 |
| 11 | 28 32 |
| 12 | 4 |
| 13 | 36 |

Tabela 19. Agrupamento definido pelo método de Tocher gerado a partir da distância de Rogers modificada ($D_{Gs}$), mínima ($D_m$) e Reynolds ($D_{RWC}$) para as 42 populações de melhoramento simuladas

| Grupo | Populações |
|---|---|
| 1 | 13 14 12 11 10 2 |
| 2 | 21 22 20 19 23 24 25 26 |
| 3 | 5 6 7 8 |
| 4 | 3 9 38 37 42 |
| 5 | 30 34 1 29 33 4 |
| 6 | 15 16 |
| 7 | 17 18 |
| 8 | 35 39 |
| 9 | 27 31 |
| 10 | 28 32 |
| 11 | 40 41 |
| 12 | 36 |

Tabela 20. Agrupamento definido pelo método de Tocher gerado a partir da distância de complemento do cosseno ($D_{COS}$) e Nei et al. (1983) ($D_{N83}$), para as 42 populações de melhoramento simuladas

| Grupo | Populações |
|---|---|
| 1 | 13 14 12 11 10 2 |
| 2 | 21 22 20 19 23 |
| 3 | 25 26 24 |
| 4 | 3 9 |
| 5 | 6 7 5 8 |
| 6 | 1 4 |
| 7 | 17 18 |
| 8 | 35 39 |
| 9 | 37 38 |
| 10 | 15 16 |
| 11 | 27 31 |
| 12 | 29 30 33 34 |
| 13 | 36 40 |
| 14 | 41 42 |
| 15 | 28 32 |

Tabela 21. Agrupamento definido pelo método de Tocher gerado a partir da distância de complemento da corda ($D_{CC}$), para as 42 populações de melhoramento simuladas

| Grupo | Populações |
|---|---|
| 1 | 13 14 12 11 10 2 5 7 6 8 |
| 2 | 21 22 20 19 23 |
| 3 | 25 26 24 |
| 4 | 3 9 |
| 5 | 1 4 |
| 6 | 17 18 |
| 7 | 35 39 |
| 8 | 30 34 33 29 |
| 9 | 37 38 41 42 |
| 10 | 15 16 |
| 11 | 36 40 |
| 12 | 27 31 |
| 13 | 28 32 |

Tabela 22. Agrupamento definido pelo método de Tocher gerado a partir da distância genética padronizada de Nei ($D_{N72}$), para as 42 populações de melhoramento simuladas

| Grupo | Populações |
|---|---|
| 1 | 13 14 12 11 10 2 5 7 6 8 |
| 2 | 21 22 20 19 23 |
| 3 | 25 26 24 |
| 4 | 3 9 |
| 5 | 1 4 |
| 6 | 17 18 |
| 7 | 35 39 |
| 8 | 30 34 33 29 |
| 9 | 37 38 41 42 |
| 10 | 15 16 |
| 11 | 36 40 |

As populações $H_{23pp}$ e $H_{23mp}$ agruparam-se com suas respectivas gerações $F_n$, quando utilizadas as distância $D_E$, $D_R$, $D_{L72}$, $D_{L73}$, $D'_{RWC}$, $D_{GS}$, $D_m$ e $D_{RWC}$. Grupos com os pares de populações híbridas $H_{12pp}$ e $H_{12mp}$, $H_{13pp}$ e $H_{13mp}$, e de retrocruzamento, $RC_{11pp}$ e $RC_{11mp}$, $RC_{31pp}$ e $RC_{31mp}$, foram frequentemente formados.

As populações $P_3$ e $P_3a_1$ foram agrupadas com as últimas gerações de retrocruzamento ($RC_{33pp}$, $RC_{34pp}$, $RC_{33mp}$ e $RC_{34mp}$) a partir das distâncias $D_E$, $D_R$, $D_H$, $D_{L72}$, $D_{L73}$, $D'_{RWC}$, $D_{GS}$, $D_m$ e $D_{RWC}$, mostrando assim a maior proximidade genética do genoma das gerações avançadas de retrocruzamento com seu genitor recorrente. O mesmo foi verificado para a população $P_1$ e $P_1a_1$, em relação às populações $RC_{13pp}$, $RC_{14pp}$, $RC_{13mp}$ e $RC_{14mp}$, para as mesmas distâncias.

Com $D_{COS}$ e $D_{N83}$ foram formados 15 grupos (Tabela 20), enquanto que $D_{N72}$ permitiu formar apenas 11 grupos. Foi consenso entre as medidas de distância, não discriminar em grupos separados as populações de hibridação planta a planta e sua respectiva população oriunda por mistura de pólen, atestando assim a semelhança entre ambas, em termos de similaridade genética ou genotípica neste estudo.

O panorama geral dos grupos formados pelos dendrogramas, obtidos com as diferentes matrizes de distância, foi semelhante. Alguns dendrogramas foram idênticos quanto à definição de seus grupos, levando-se em conta o critério estatístico (Mojena, 1977). Na Figura 4, estão

representadas as distâncias $D_E$ e $D_R$; Figura 5, $D_{L72}$, $D_{L73}$ e $D'_{RWC}$; Figura 6, $D_{COS}$ e $D_{CC}$; e Figura 7, $D_{GS}$ e $D_{RWC}$. A semelhança dos agrupamentos entre as distâncias de Latter (1972 e 1973) e de Reynolds et al. (1983) justifica-se, uma vez que elas foram propostas com base no modelo de diferenciação genética entre populações exclusivamente por deriva genética (Nei & Kumar, 2000; Laval et al., 2003). Assim, espera-se que os valores destas distâncias não sejam afetados pelos tipos de marcadores usados, isto é, com diferentes tipos de taxas evolutivas, a exemplo das isoezimas e microssatélites (Mohammadi & Prasanna, 2003). Por outro lado, para medidas de distância de diferenças genéticas absolutas, como a distância genética mínima (Nei, 1973), são esperados diferentes resultados dependendo da taxa evolutiva do marcador empregado.

Assim como no método de otimização de Tocher, uma mesma população submetida aos dois tipos de hibridação esteve sempre agrupada. O que proporcionou ligeiras discrepâncias entre os dendrogramas foram os grupos formados, segundo o critério estatístico. Em todos os dendrogramas a população $P_2$ foi agrupada com suas respectivas populações de autofecundação e acasalamento ao acaso. As populações de retrocruzamento $RC_{11pp}$ a $RC_{14mp}$ e $RC_{31pp}$ a $RC_{34mp}$ estiveram juntas com seus respectivos genitores recorrentes $P_1$ ($P_1a_1$) e $P_3$ ($P_3a_1$). A exceção foi o agrupamento promovido pelas distâncias $D_E$ e $D_R$ (Figura 4), cujas populações $RC_{11pp}$ e $RC_{11mp}$ formaram conjunto com os híbridos $H_{13pp}$ e $H_{13mp}$.

Pela conformação dos dendrogramas, as populações $H_{12pp}$ e $H_{12mp}$ se destacaram das demais populações, formando, geralmente, um conjunto separado (Figuras 4 a 11). As populações segregantes $F_n$ ora agruparam-se com os híbridos $H_{23pp}$ e $H_{23mp}$ (Figura 4, 6 e 7), ora com a população $P_2$ e gerações de acasalamento ao acaso e autofecundação (Figura 8 e 10), ora com gerações de retrocruzamentos cujo parental recorrente foi $P_3$ (Figura 5, 9, e 11).

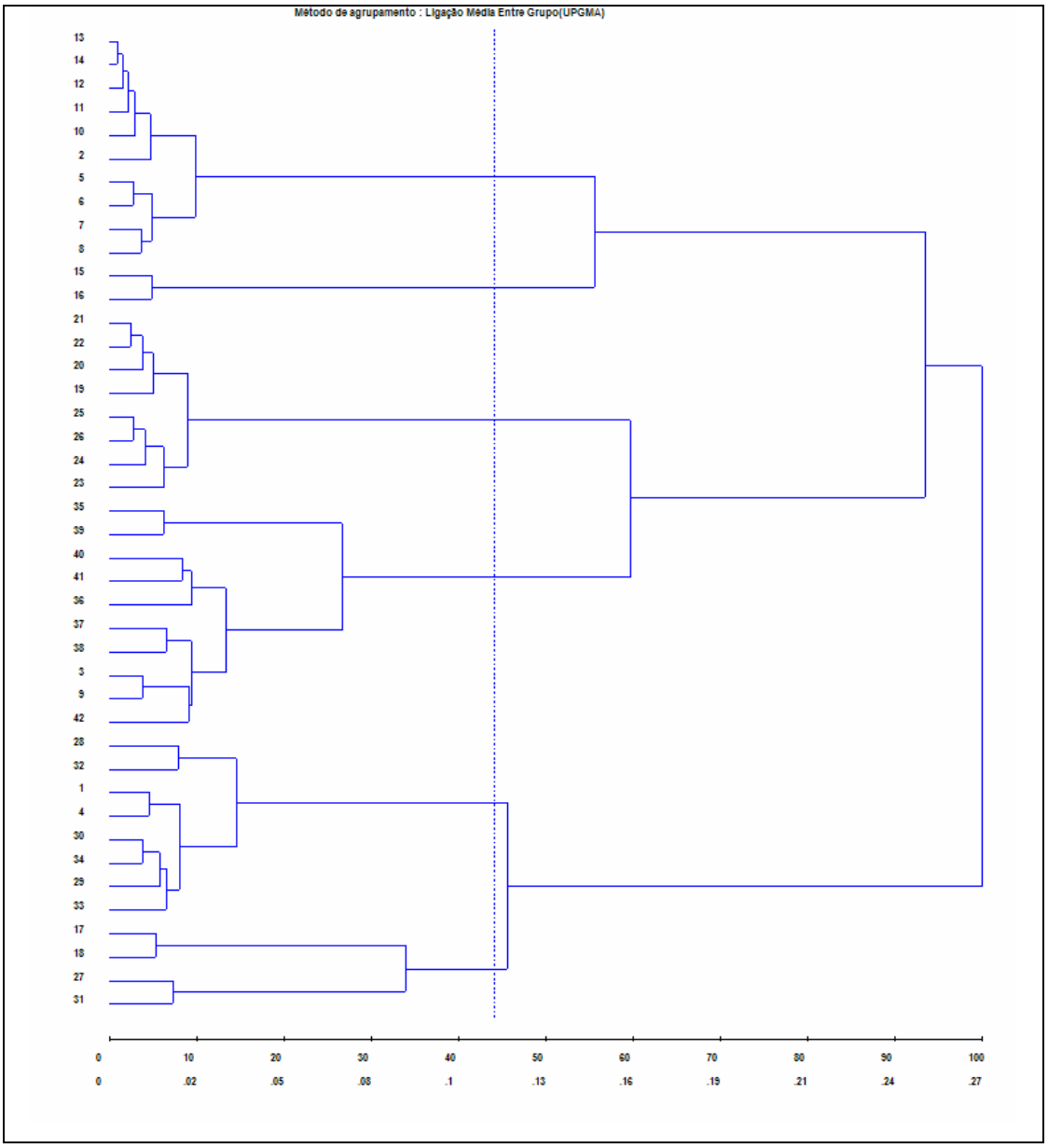


Figura 4. Dendrograma obtido pelo método UPGMA a partir da matriz de distância de Rogers para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,1212, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

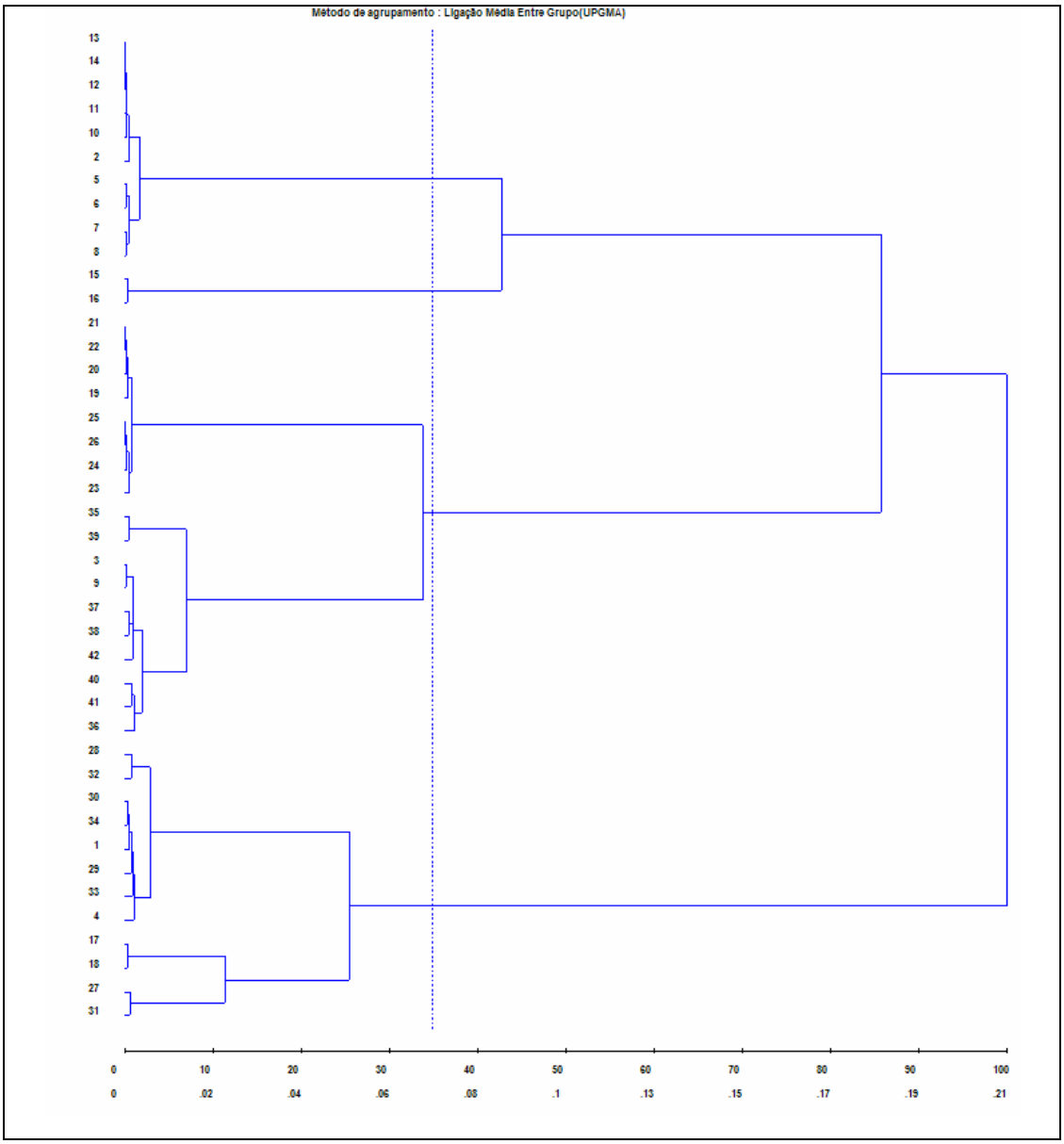


Figura 5. Dendrograma obtido pelo método UPGMA a partir da matriz de distância de Reynolds et al. (1983) (ignorando o termo associado a tamanho amostral) para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,0757, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

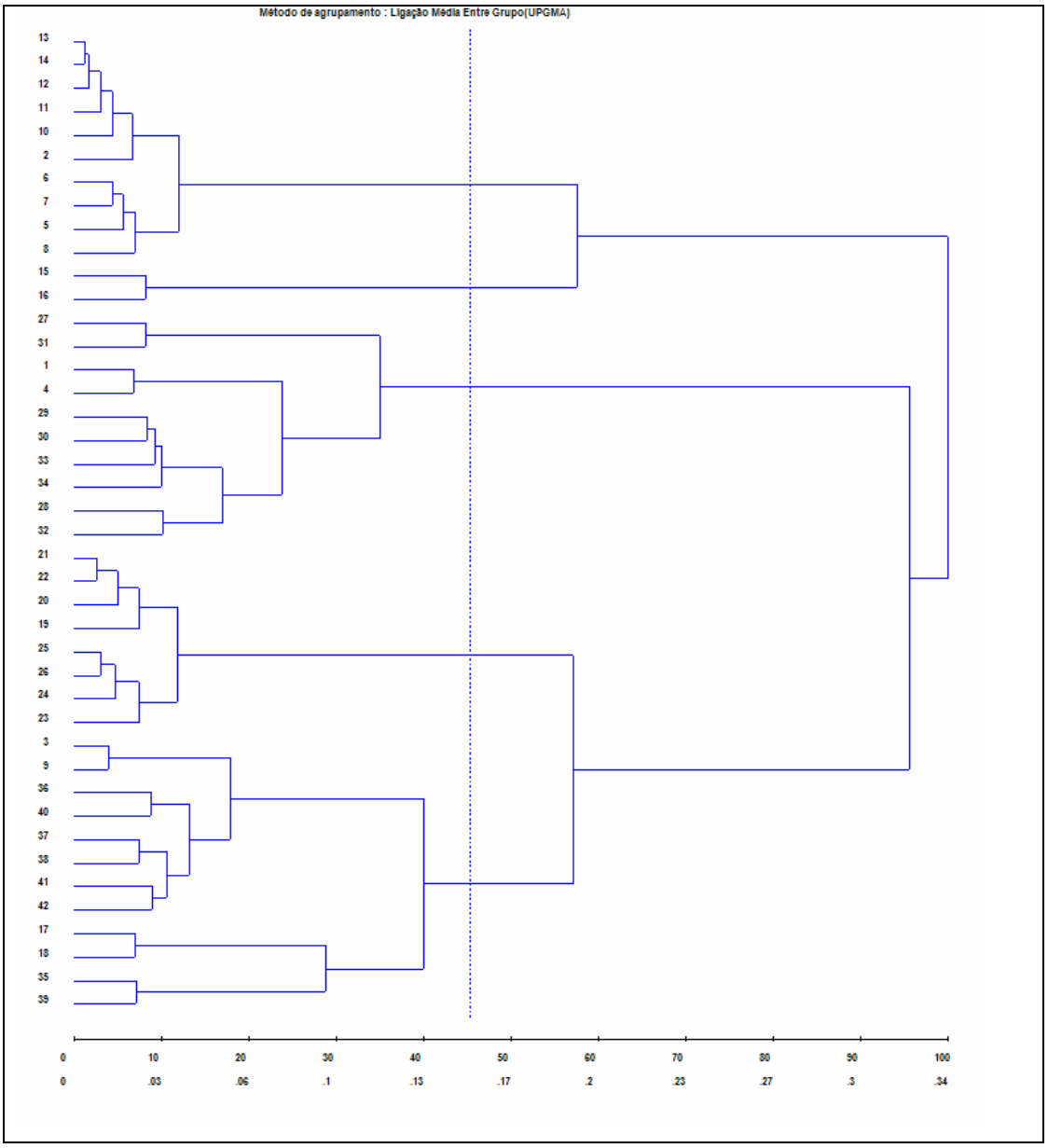


Figura 6. Dendrograma obtido pelo método UPGMA a partir da matriz de distância angular do complemento do cosseno para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,1545, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

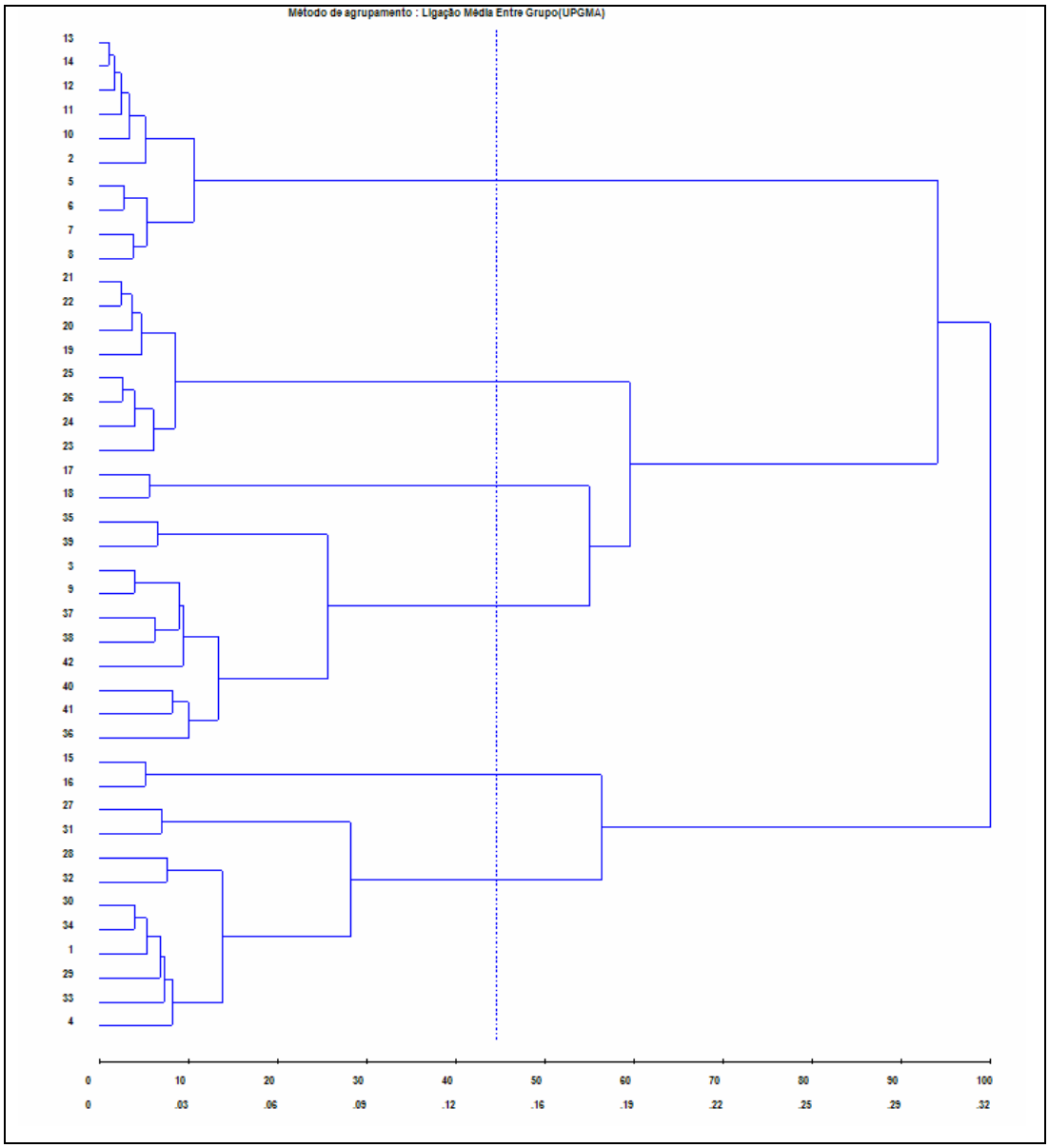


Figura 7. Dendrograma obtido pelo método UPGMA a partir da matriz de distância de Roger modificada para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,1438, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

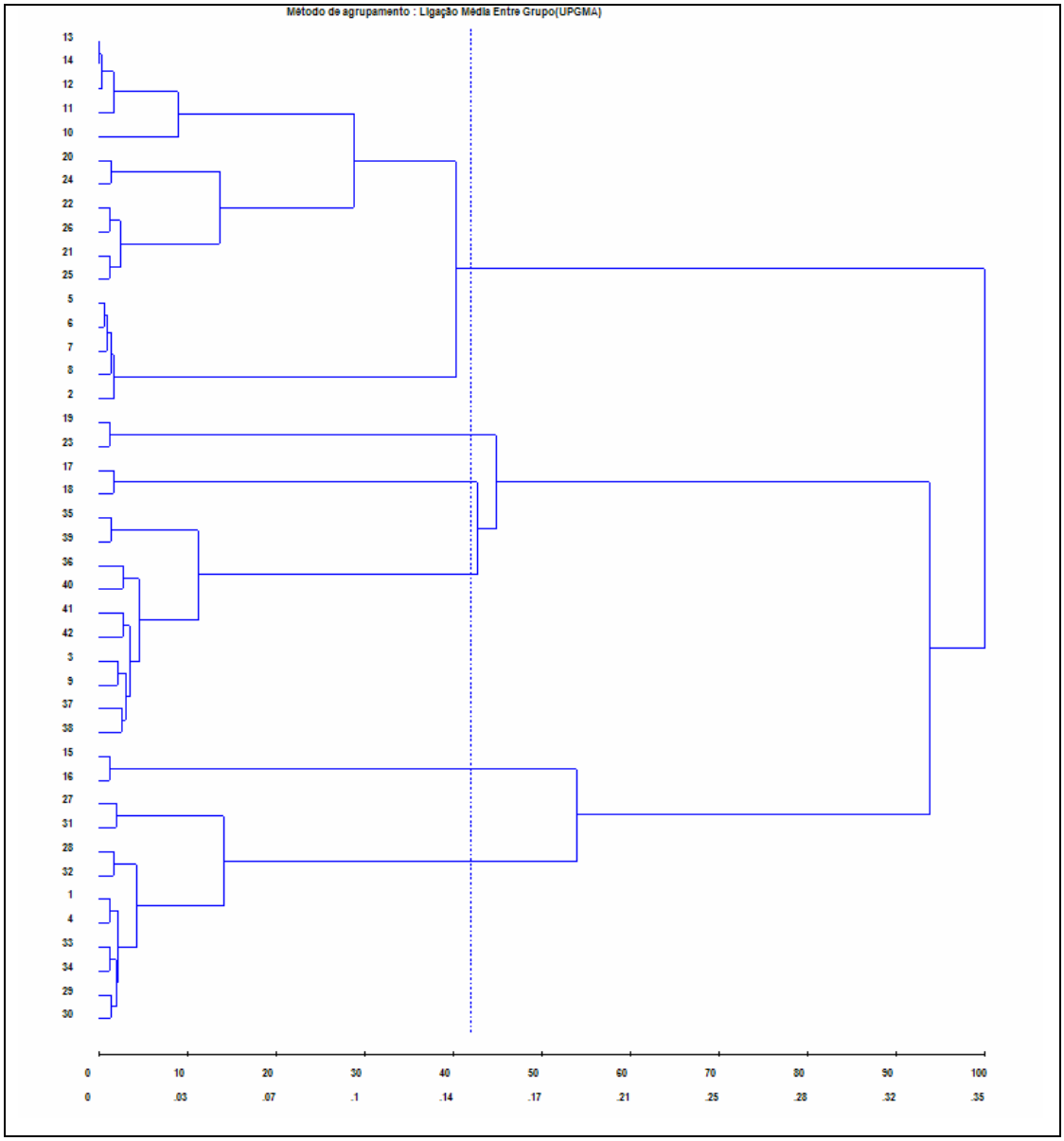


Figura 8. Dendrograma obtido pelo método UPGMA a partir da matriz de distância genotípica de Hedrick para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,1509, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

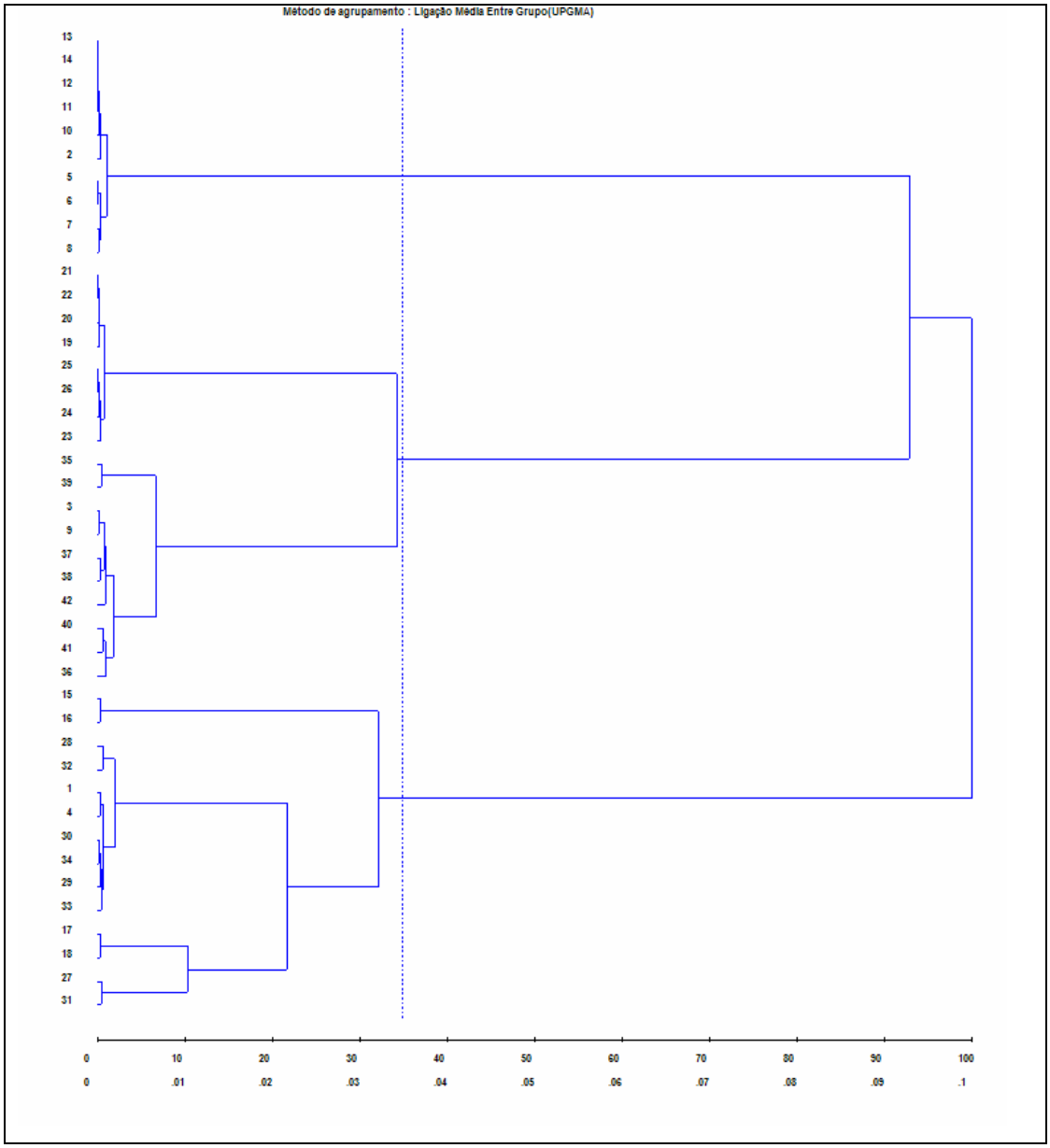


Figura 9. Dendrograma obtido pelo método UPGMA a partir da matriz de distância mínima (Nei 1973) para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,0372, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

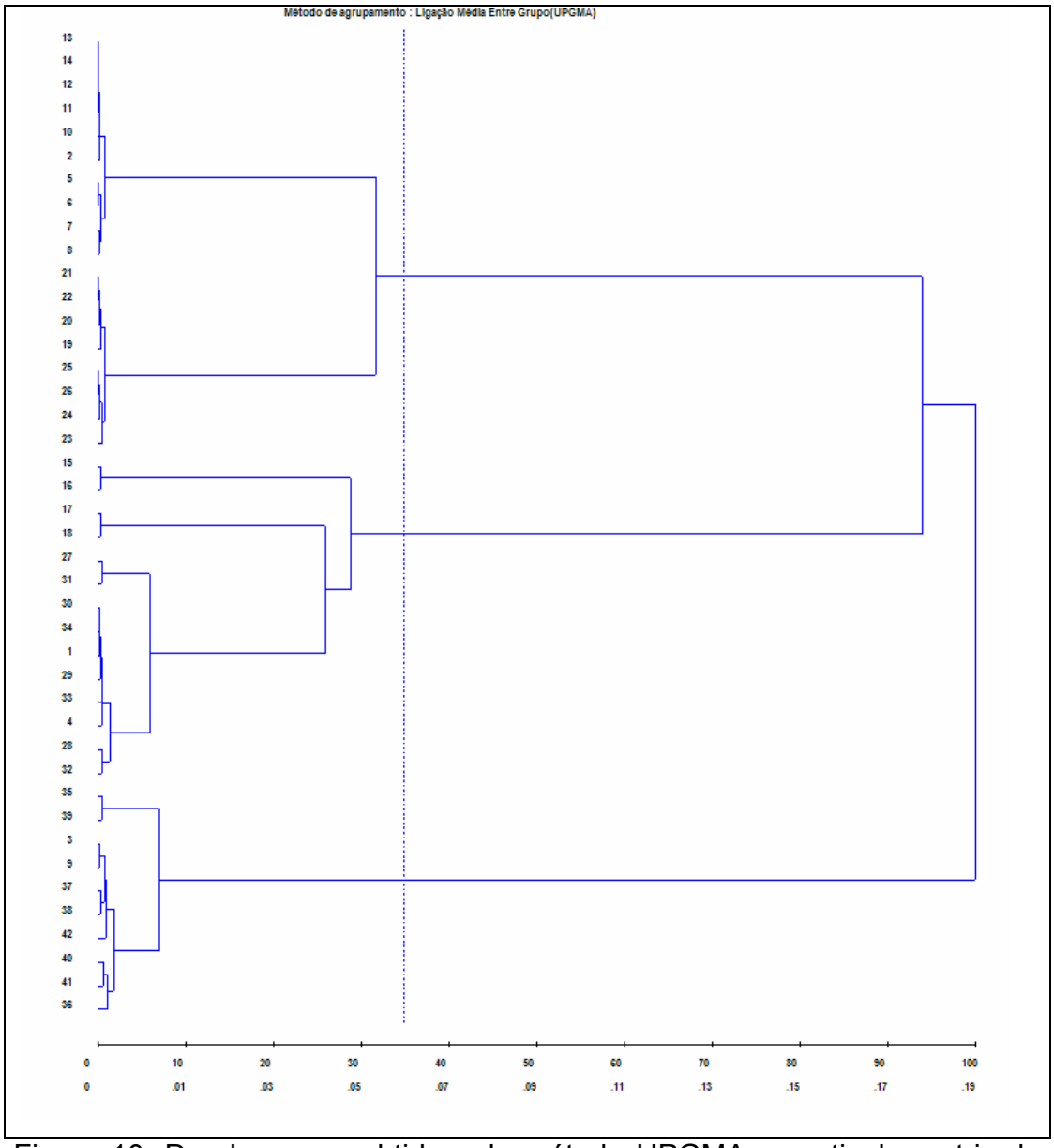


Figura 10. Dendrograma obtido pelo método UPGMA a partir da matriz de distância genética padronizada de Nei para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,0666, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

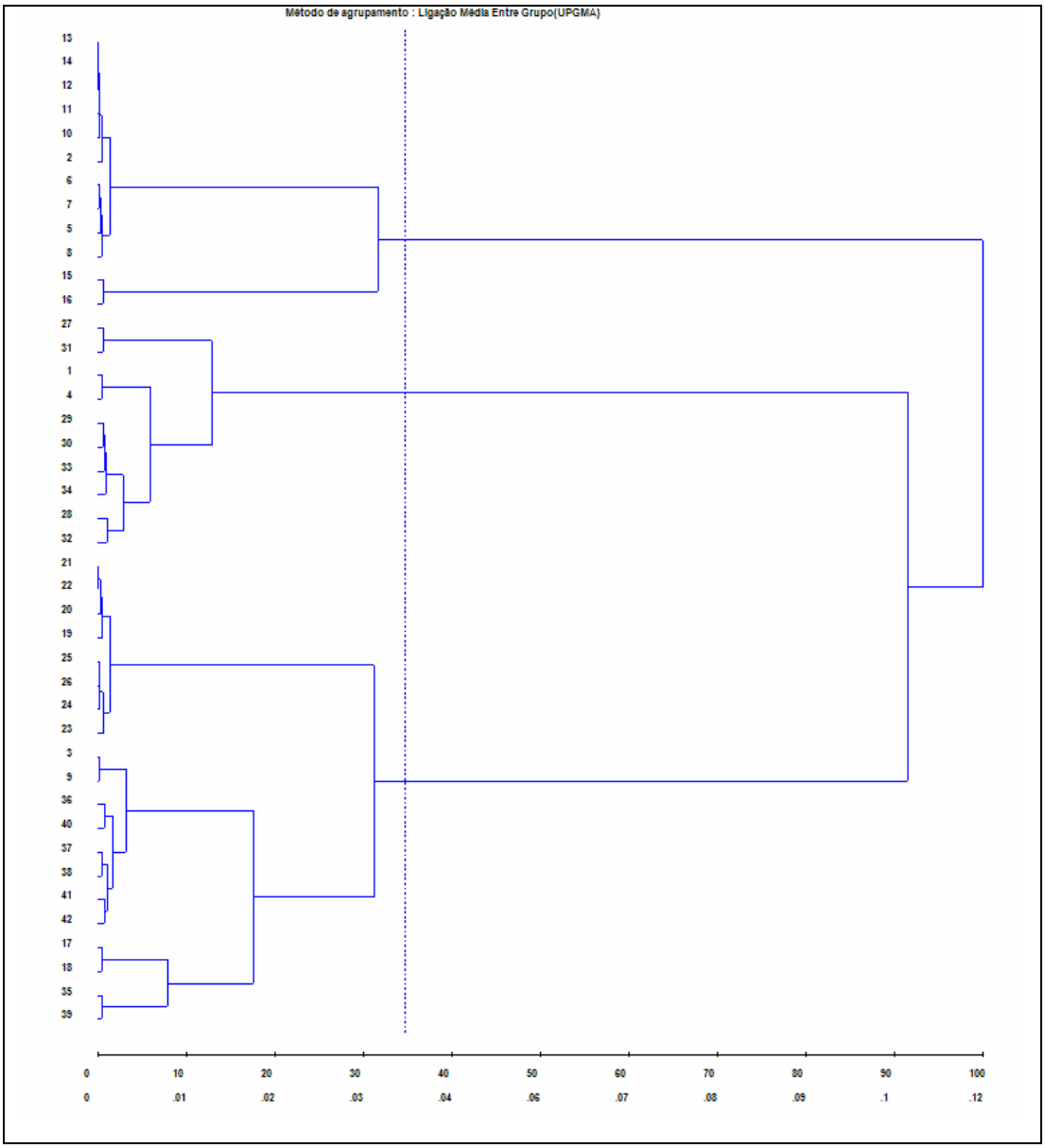


Figura 11. Dendrograma obtido pelo método UPGMA a partir da matriz de distância de Nei et al. (1983) para as 42 populações de melhoramento simuladas. O ponto de corte para a definição dos grupos foi de 0,0423, segundo critério estatístico (Mojena, 1977). Representação das populações no dendrograma: 1 - $P_1$; 2 – $P_2$; 3 – $P_3$; 4 – $P_1a_1$; 5 - $P_1a_1$; 6 - $P_1a_2$; 7 - $P_1a_3$; 8 - $P_1a_4$; 9 – $P_3a_1$; 10 – $P_2s_1$; 11 – $P_2s_2$; 12 – $P_2s_3$; 13 – $P_2s_4$; 14 – $P_2s_5$; 15 – $H_{12pp}$; 16 – $H_{12mp}$; 17 – $H_{13pp}$; 18 – $H_{13mp}$; 19 – $H_{23pp}$; 20 – $F_{2(pp)}$; 21 – $F_{3(pp)}$; 22 – $F_{4(pp)}$; 23 – $H_{23mp}$; 24 – $F_{2(mp)}$; 25 – $F_{3(mp)}$; 26 – $F_{4(mp)}$; 27 – $RC_{11pp}$; 28 – $RC_{12pp}$; 29 – $RC_{13pp}$; 30 – $RC_{14pp}$; 31 – $RC_{11mp}$; 32 – $RC_{12mp}$; 33 – $RC_{13mp}$; 34 – $RC_{14mp}$; 35 – $RC_{31pp}$; 36 – $RC_{32pp}$; 37 – $RC_{33pp}$; 38 – $RC_{34pp}$; 39 – $RC_{31mp}$; 40 – $RC_{32mp}$; 41 – $RC_{33mp}$; 42 – $RC_{34mp}$.

Ao contrário dos resultados obtidos no presente trabalho, Meerow et al., (2003), avaliando 18 populações (acessos) de coco (*Cocos nucifera* L.), num total de 66 indivíduos e 15 locos microssatélites obteve dendrogramas pelo método UPGMA com topologias similares, por meio das medidas de distância de Roger modificada e distância genética padronizada de Nei.

Nas Figuras 12, 13 e 14 estão as projeções bidimensionais (2D) das medidas de distâncias. Na figura 12 a dispersão das populações no plano 2D foi semelhante para a distâncias $D_{N83}$, $D_E$, $D_R$, $D_{GS}$, $D_m$, $D_{RWC}$, $D_{COS}$, $D_{CC}$ e $D_{N72}$, apenas com alterações nas coordenadas dos eixos X e Y. Destaca-se que a medida $D_{N83}$ não foi desenvolvida sob um modelo genético específico e nem é caracterizada como distância métrica e euclidiana (Reif et al., 2005). O mesmo ocorreu para as distâncias $D_{L72}$, $D_{L73}$ e $D'_{RWC}$ (Figura 13).

Destaca-se a projeção 2D com a distância $D_H$, que permitiu melhor discernimento na visualização das populações de autofecundação ($P_2s_2$ a $P_2s_4$), $P_2$ e as gerações de acasalamento ao acaso ($P_2a_1$ a $P_2a_4$), estando $P_2s_1$ entre estes dois grupos. Sabe-se que nas espécies ou populações de plantas há variação de suas taxas de alogamia em diferentes áreas (Hartl & Clark, 1997). Se populações com a mesma freqüência gênica possuem diferentes coeficientes de endogamia, então comparações ao nível genotípico podem proporcionar uma informação adicional, a qual a similaridade gênica não proverá.

A projeção 2D apresentou padrões semelhantes de agrupamento àqueles definidos pelos métodos de Tocher e UPGMA. Takezaki & Nei (1996) comentaram que a distância de Reynolds et al. (1983) é, essencialmente, igual a de Latter (1972) quando o tamanho amostral das populações é grande. Os autores ainda acrescentam dizendo que a distância de Latter (1973) foi igualmente eficiente ou pouco menos que a distância $D_{L72}$ na reconstrução de árvores filogenéticas. As distâncias tidas pela expressão –ln (1 – D), permitem estabelecer uma relação linear aproximada entre o tempo de divergência das populações, a exemplo das distâncias $D_{N72}$ e $D_{L73}$.

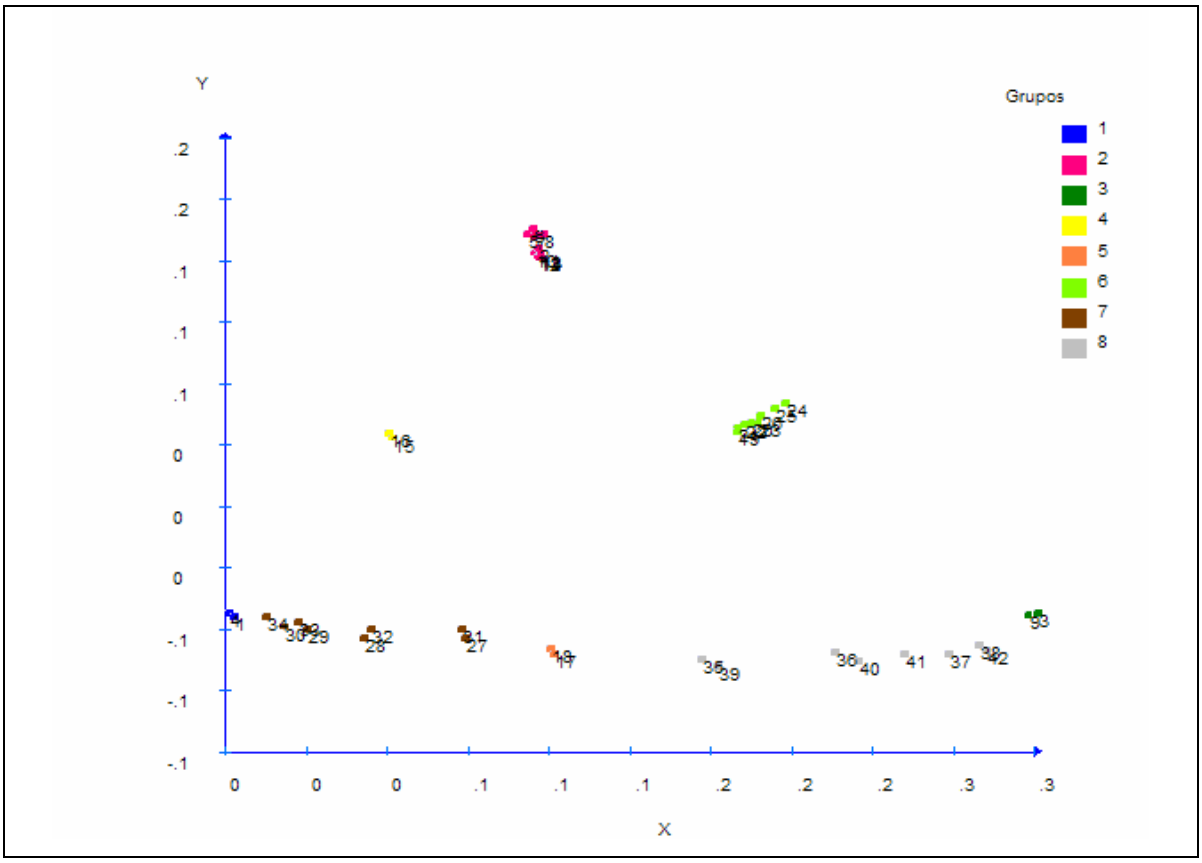


Figura 12. Projeção gráfica bidimensional das 42 populações, obtida com a distância de Nei et al. (1983). Legenda: Grupo 1 – 1 ($P_1$) e 4 ($P_1a_1$); Grupo 2 – 2 ($P_2$), 5 ($P_1a_1$), 6 ($P_1a_2$), 7 ($P_1a_3$), 8 ($P_1a_4$), 10 ($P_2s_1$), 11 ($P_2s_2$), 12 ($P_2s_3$), 13 ($P_2s_4$) e 14 ($P_2s_5$); Grupo 3 – 3 ($P_3$) e 9 ($P_3a_1$); Grupo 4 – 15 ($H_{12pp}$) e 16 ($H_{12mp}$); Grupo 5 – 17 ($H_{13pp}$) e 18 ($H_{13mp}$); Grupo 6 – 19 ($H_{23pp}$), 23 ($H_{23mp}$), 20 ($F_{2(pp)}$), 21 ($F_{3(pp)}$), 22 ($F_{4(pp)}$), 24 ($F_{2(mp)}$), 25 ($F_{3(mp)}$), 26 ($F_{4(mp)}$); Grupo 7 – 27 ($RC_{11pp}$), 28 ($RC_{12pp}$), 29 ($RC_{13pp}$), 30 ($RC_{14pp}$), 31 ($RC_{11mp}$), 32 ($RC_{12mp}$), 33 ($RC_{13mp}$), 34 ($RC_{14mp}$); Grupo 8 – 35 ($RC_{31pp}$), 36 ($RC_{32pp}$), 37 ($RC_{33pp}$), 38 ($RC_{34pp}$), 39 ($RC_{31mp}$), 40 ($RC_{32mp}$), 41 ($RC_{33mp}$), 42 ($RC_{34mp}$).

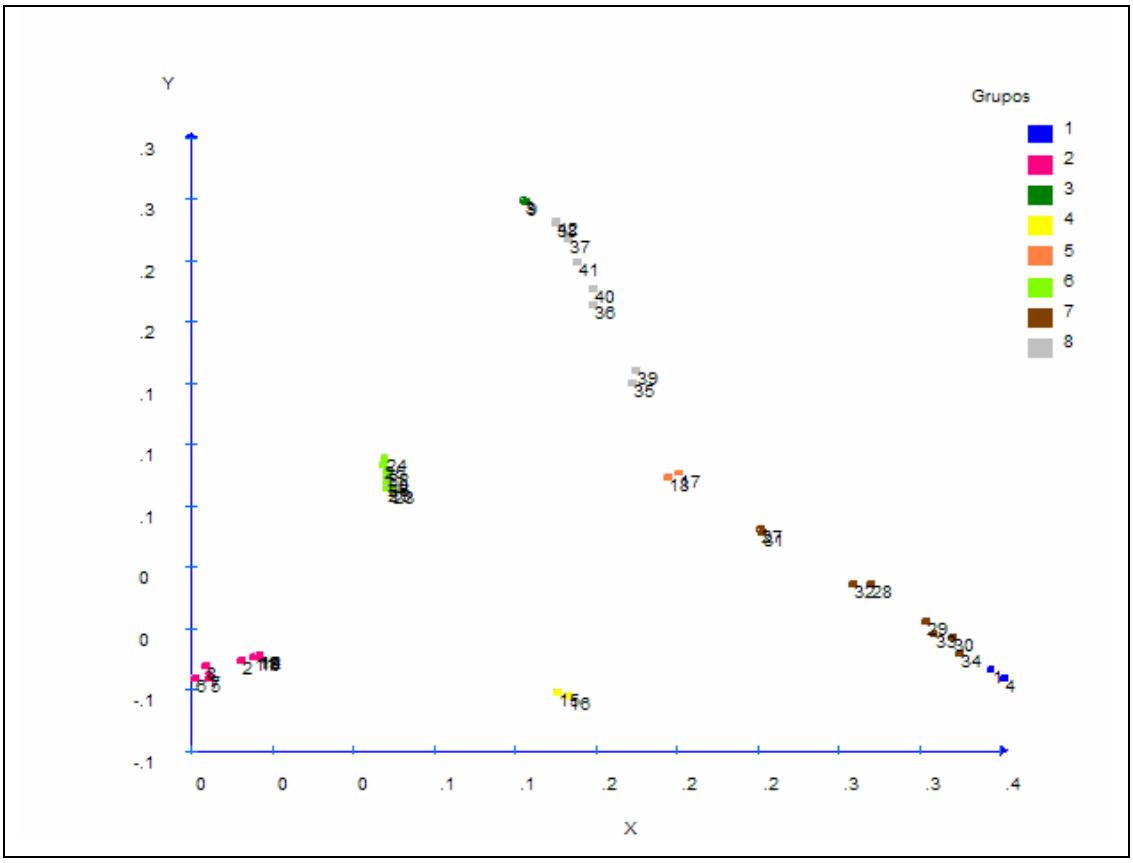


Figura 13. Projeção gráfica bidimensional das 42 populações, obtida com a distância de Latter (1973). Legenda: Grupo 1 – 1 ($P_1$) e 4 ($P_1a_1$); Grupo 2 – 2 ($P_2$), 5 ($P_1a_1$), 6 ($P_1a_2$), 7 ($P_1a_3$), 8 ($P_1a_4$), 10 ($P_2s_1$), 11 ($P_2s_2$), 12 ($P_2s_3$), 13 ($P_2s_4$) e 14 ($P_2s_5$); Grupo 3 – 3 ($P_3$) e 9 ($P_3a_1$); Grupo 4 – 15 ($H_{12pp}$) e 16 ($H_{12mp}$); Grupo 5 – 17 ($H_{13pp}$) e 18 ($H_{13mp}$); Grupo 6 – 19 ($H_{23pp}$), 23 ($H_{23mp}$), 20 ($F_{2(pp)}$), 21 ($F_{3(pp)}$), 22 ($F_{4(pp)}$), 24 ($F_{2(mp)}$), 25 ($F_{3(mp)}$), 26 ($F_{4(mp)}$); Grupo 7 – 27 ($RC_{11pp}$), 28 ($RC_{12pp}$), 29 ($RC_{13pp}$), 30 ($RC_{14pp}$), 31 ($RC_{11mp}$), 32 ($RC_{12mp}$), 33 ($RC_{13mp}$), 34 ($RC_{14mp}$); Grupo 8 – 35 ($RC_{31pp}$), 36 ($RC_{32pp}$), 37 ($RC_{33pp}$), 38 ($RC_{34pp}$), 39 ($RC_{31mp}$), 40 ($RC_{32mp}$), 41 ($RC_{33mp}$), 42 ($RC_{34mp}$).

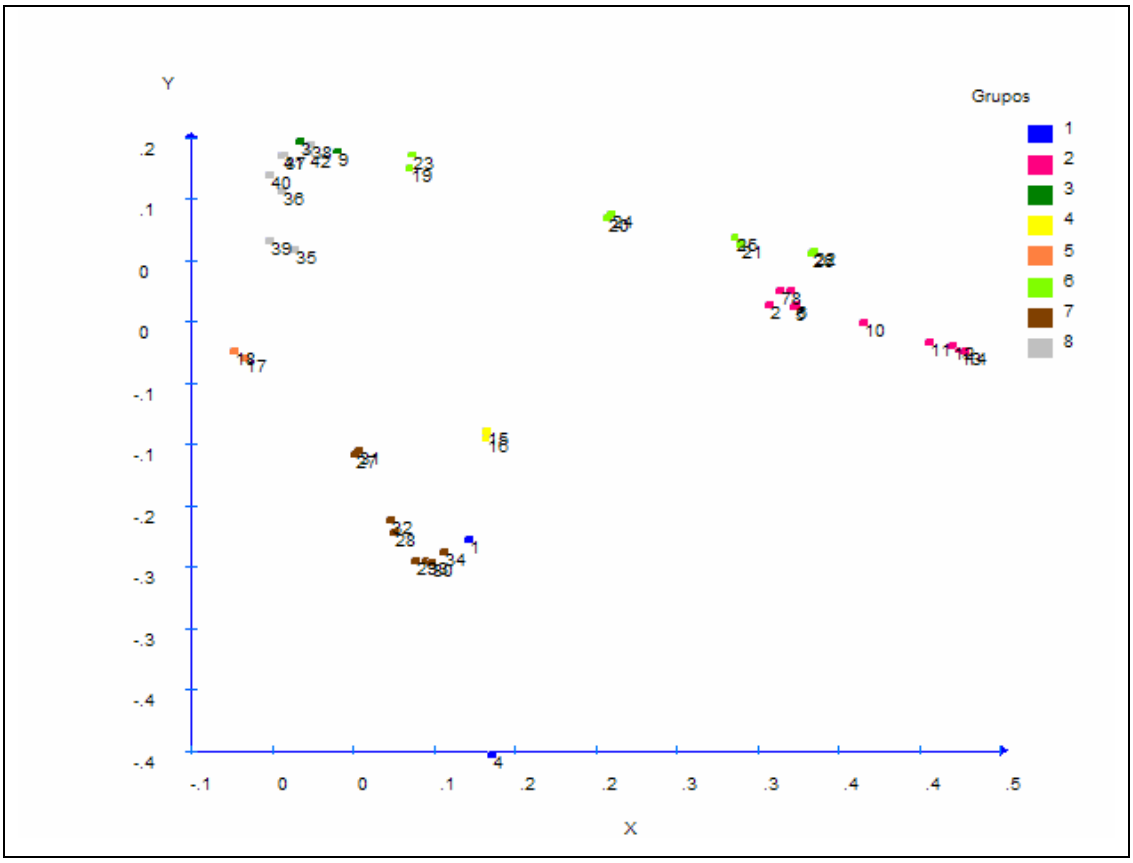


Figura 14. Projeção gráfica bidimensional das 42 populações, obtida com a distância de Hedrick (1971). Legenda: Grupo 1 – 1 ($P_1$) e 4 ($P_1a_1$); Grupo 2 – 2 ($P_2$), 5 ($P_1a_1$), 6 ($P_1a_2$), 7 ($P_1a_3$), 8 ($P_1a_4$), 10 ($P_2s_1$), 11 ($P_2s_2$), 12 ($P_2s_3$), 13 ($P_2s_4$) e 14 ($P_2s_5$); Grupo 3 – 3 ($P_3$) e 9 ($P_3a_1$); Grupo 4 – 15 ($H_{12pp}$) e 16 ($H_{12mp}$); Grupo 5 – 17 ($H_{13pp}$) e 18 ($H_{13mp}$); Grupo 6 – 19 ($H_{23pp}$), 23 ($H_{23mp}$), 20 ($F_{2(pp)}$), 21 ($F_{3(pp)}$), 22 ($F_{4(pp)}$), 24 ($F_{2(mp)}$), 25 ($F_{3(mp)}$), 26 ($F_{4(mp)}$); Grupo 7 – 27 ($RC_{11pp}$), 28 ($RC_{12pp}$), 29 ($RC_{13pp}$), 30 ($RC_{14pp}$), 31 ($RC_{11mp}$), 32 ($RC_{12mp}$), 33 ($RC_{13mp}$), 34 ($RC_{14mp}$); Grupo 8 – 35 ($RC_{31pp}$), 36 ($RC_{32pp}$), 37 ($RC_{33pp}$), 38 ($RC_{34pp}$), 39 ($RC_{31mp}$), 40 ($RC_{32mp}$), 41 ($RC_{33mp}$), 42 ($RC_{34mp}$).

Na Tabela 23, estão dispostos alguns índices capazes de medir a eficiência da projeção das distâncias genéticas. As medidas de distâncias que apresentaram os menores valores de distorção (0,2077 a 3,5004) e estresse (2,2579 a 4,2332) e maiores valores de correlação entre distância original e gráfica (0,99) foram todas aquelas que possuem fundamentos geométricos, como $D_E$, $D_R$, $D_{GS}$, $D_{COS}$ e $D_{CC}$, a exceção de $D_{RWC}$. Segundo Kruskal (1964) valores de estresse acima de 20%, são considerados altos e representam um ajustamento ruim da projeção. A distância de Reynolds et al. (1983) embora não seja fundamentada em propriedades geométricas esteve relacionada a $D_{GS}$, nos agrupamentos de Tocher e UPGMA (Tabela 19 e Figura 7).

Tabela 23. Avaliação da eficiência da projeção de medidas de distâncias[#] no plano bidimensional

| Distância | Distorção (%) | Correlação[##] | Estresse (%) |
|---|---|---|---|
| $D_E$ | 0,2077 | 0,9980 | 3,1631 |
| $D_R$ | 0,2077 | 0,9980 | 3,1631 |
| $D_{GS}$ | 1,2768 | 0,9995 | 2,2579 |
| $D_{COS}$ | 3,5040 | 0,9992 | 4,2332 |
| $D_{CC}$ | 0,9938 | 0,9973 | 3,9864 |
| $D_m$ | 19,7759 | 0,9632 | 26,1362 |
| $D_{N72}$ | 21,0986 | 0,9550 | 28,4451 |
| $D_{L72}$ | 16,2185 | 0,9750 | 20,9833 |
| $D_{L73}$ | 18,8612 | 0,9662 | 24,8061 |
| $D_{N83}$ | 20,1569 | 0,9603 | 26,5420 |
| $D_{RWC}$ | 1,2816 | 0,9995 | 2,2691 |
| $D'_{RWC}$ | 16,2185 | 0,9750 | 20,9833 |
| $D_H$ | 4,2734 | 0,9291 | 18,5434 |

[#]Distâncias: $D_E$: Euclidiana média; $D_R$: Rogers; $D_{GS}$: Rogers modificada; $D_{COS}$: angular complemento aritmético do cosseno; $D_{CC}$: angular comprimento da corda; $D_{N83}$: Nei et al. (1983); $D_{N72}$: genética padronizada de Nei; $D_m$: mínima; $D_{RWC}$: Reynolds; $D'_{RWC}$ Reynolds (ignorando o termo associado ao tamanho amostral); $D_{L72}$: Latter (1972); $D_{L73}$: Latter (1973); $D_H$: genotípica de Hedrick.

[##]Correlação de Pearson entre as distâncias originais e as distâncias gráficas

A caracterização de uma distância no espaço euclidiano é importante porque esta é uma pressuposição explícita ou implícita em várias técnicas de análise multivariada. A propriedade euclidiana é um aspecto desejável, mas o principal critério para a escolha de uma medida de distância são suas propriedades genéticas (Dias, 1998).

Segundo Weir (1996) distâncias geométricas devem satisfazer critérios diferentes em relação às distâncias genéticas, o que não impede que os modelos evolutivos sejam apropriados para as distâncias geométricas. Goodman (1972) relata que uma distância geométrica igual não indica igual divergência biológica e que uma falta de correspondência entre distância e grau de divergência pode ocorrer se distâncias baseadas em mutação são aplicadas a situações de deriva genética.

As demais distâncias, $D_m$, $D_{N72}$, $D_{L72}$, $D_{L73}$, $D_{N83}$, $D'_{RWC}$ e $D_H$, propostas com base em modelos evolutivos e conhecidas como distâncias genéticas, proporcionaram projeção menos eficiente. A distância genética

padronizada de Nei ($D_{N72}$) foi a que apresentou os mais altos valores de distorção (21,0986) e estresse (28,4451), e $D_H$ a menor correlação (0,92).

As estimativas de correlação simples entre as medidas de distância obtidas a partir das 42 populações estão dispostas na Tabela 24. Todas as correlações foram bastante elevadas e significativas ($P < 0,05$), sendo a menor entre $D_H$ e $D_{L73}$ (0,86). Hedrick (1975) ao comparar as distâncias $D_H$, $D_{N72,}$ $D_R$ e $D_{CC}$ verificou altas correlações entre elas, a semelhança deste estudo. O autor ainda afirma que correlações entre $D_H$ e outras medidas podem não ser tão altas quando as populações não estão em EHW. No presente trabalho, provavelmente, as altas correlações entre $D_H$ com as demais medidas de distâncias, se deve a condição de EHW para algumas das populações simuladas. Além disso, as altas associações e padrões de agrupamento semelhantes entre todas as distâncias são atribuídos ao comportamento das populações dentro de um panorama evolutivo, pois estas não sofreram com efeitos sistemáticos de seleção, fluxo gênico e mutação. Além disso, seus tamanhos amostrais ($N_i = 200$), fizeram com que os efeitos da deriva genética fossem insignificantes.

Melchinger et al. (1991) derivaram resultados teóricos, de modo que as estimativas de $D_R$ entre duas linhagens endogâmicas homozigotas apresentam relação linear com o coeficiente de coancestralidade (Malécot, 1948). Assim, $D_R$ foi adequada para avaliar a relação entre a divergência genética de populações endogâmicas, baseadas em marcadores moleculares codominantes e no coeficiente de coancestralidade. Esta linearidade é desejada em estudos que visam organizar e validar coleções núcleo e identificar duplicatas em bancos de germoplasma, bem como descobrir as relações de pedigree entre as populações, como uma necessidade para a detecção de variedades essencialmente derivadas do melhoramento (Reif et al. 2005).

Tabela 24. Estimativas de correlação de Pearson[*] entre as medidas de distância[#] obtidas a partir de 42 observações (populações simuladas)

| | $D_E$ | $D_R$ | $D_{GS}$ | $D_{COS}$ | $D_{CC}$ | $D_{N83}$ | $D_{N72}$ | $D_m$ | $D_{RWC}$ | $D'_{RWC}$ | $D_{L72}$ | $D_{L73}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $D_R$ | 1,00 | 0,00 | | | | | | | | | | |
| $D_{GS}$ | 1,00 | 1,00 | 0,00 | | | | | | | | | |
| $D_{COS}$ | 0,99 | 0,99 | 0,99 | 0,00 | | | | | | | | |
| $D_{CC}$ | 0,99 | 0,99 | 0,99 | 1,00 | 0,00 | | | | | | | |
| $D_{N83}$ | 0,95 | 0,95 | 0,96 | 0,97 | 0,97 | 0,00 | | | | | | |
| $D_{N72}$ | 0,96 | 0,96 | 0,96 | 0,95 | 0,95 | 0,98 | 0,00 | | | | | |
| $D_m$ | 0,96 | 0,96 | 0,97 | 0,96 | 0,96 | 0,99 | 1,00 | 0,00 | | | | |
| $D_{RWC}$ | 1,00 | 1,00 | 1,00 | 0,99 | 0,99 | 0,96 | 0,96 | 0,97 | 0,00 | | | |
| $D'_{RWC}$ | 0,96 | 0,96 | 0,97 | 0,96 | 0,96 | 0,98 | 0,98 | 0,99 | 0,97 | 0,00 | | |
| $D_{L72}$ | 0,96 | 0,96 | 0,97 | 0,96 | 0,96 | 0,98 | 0,98 | 0,99 | 0,97 | 1,00 | 0,00 | |
| $D_{L73}$ | 0,95 | 0,95 | 0,96 | 0,95 | 0,95 | 0,98 | 0,98 | 0,99 | 0,96 | 1,00 | 1,00 | 0,00 |
| $D_H$ | 0,90 | 0,90 | 0,90 | 0,89 | 0,90 | 0,87 | 0,88 | 0,88 | 0,90 | 0,87 | 0,87 | 0,86 |

[*]Todas as correlações foram significativas ao nível de 5% de probabilidade pelo teste de Mantel, com 1000 permutações.

[#]Distâncias: $D_E$: Euclidiana média; $D_R$: Rogers; $D_{GS}$: Rogers modificada; $D_{COS}$: angular complemento aritmético do cosseno; $D_{CC}$: angula comprimento da corda; $D_{N83}$: Nei et al. (1983); $D_{N72}$: genética padronizada de Nei; $D_m$: mínima; $D_{RWC}$: Reynolds; $D'_{RWC}$ Reynolds (ignorando o termo associado ao tamanho amostral); $D_{L72}$: Latter (1972); $D_{L73}$: Latter (1973); $D_H$: genotipica de Hedrick.

Reif et al. (2005) compararam sete medidas de distância baseadas em informações de freqüências alélicas ($D_E$; $D_R$; $D_{GS}$; $D_{CC}$; $D_{N72}$; $D_{N83}$ e $D'_{RWC}$), a partir de sete populações de milho tropical do CIMMYT, submetendo-as a uma análise par a par, denominada *Procrutes*, capaz de comparar diferentes métodos multivariados utilizando o mesmo conjunto de dados. Com a visualização gráfica, os autores observaram que a distância entre $D_E$ e $D_{GS}$ foi igual à zero, dada a relação $D_{GS} = \sqrt{2mD_E}$. Essas duas distâncias foram agrupadas junto a $D_R$ e $D_{CC}$. O coeficiente $D_{N83}$ foi posicionado entre $D_E$, $D_{GS}$, $D_R$ e $D_{CC}$ de um lado e, $D_{N72}$ e $D'_{RWC}$ do outro. Segundo os autores, este acontecimento estava de acordo com o esperado, porque $D_{N72}$ e $D'_{RWC}$ são baseadas em pressuposições similares: uma população ancestral dividida em subpopulações divergindo pela deriva genética ($D'_{RWC}$) ou pela mutação e deriva ($D_{N72}$).

Em um estudo de simulações, Takezaki & Nei (1996) avaliaram a eficiência de diferentes medidas de distância, aplicadas a marcadores microssatélites, quanto à probabilidade de cada uma obter a topologia verdadeira ($T_v$) da árvore filogenética. Foram considerados os modelos de mutação de alelos infinitos (IAM) e o de mutações *stepwise* (SMM). Dentre as medidas de distância avaliadas encontravam-se $D_{N72}$, $D_{N83}$, $D_R$ e $D_{CC}$, tidas como medidas tradicionais. Os resultados mostraram que, em ambos os modelos, as distâncias $D_{CC}$ e $D_{N83}$, no geral, mostraram os mais elevados valores de $T_v$ em relação a outras medidas de distância, com ou sem efeito de afunilamento genético. Por outro lado, a distância $D_{N72}$ foi uma das mais apropriadas para estimar o tempo de divergência evolutiva.

### 4.2.2. Diversidade genética entre e dentro de populações de retrocruzamento

A busca por variabilidade genética e, consequentemente, genótipos de desempenho superior aos parentais, foi sugerida em populações de retrocruzamento (Lorencetti et al., 2006). Basicamente um programa de retrocruzamento assistido por marcadores moleculares visa ao

monitoramento dos genes controladores característica introduzida e redução no tempo necessário para a recuperação do genoma recorrente (Guimarães et al., 2006).

Existem algumas maneiras de analisar como a variabilidade é distribuída ao nível molecular. No presente trabalho a análise da variação genética entre e dentro das populações base $P_1$ ou $P_3$ com gerações de retrocruzamento, cujo genitor recorrente era $P_3$ ($RC_{31mp}$ e $RC_{34mp}$), foi realizada através do método de análise de variância de freqüências gênicas sob um modelo aleatório (Cockerham, 1969, 1973 e Weir & Cockerham, 1984); análise de variância molecular baseado na estruturação da matriz de distância Euclidiana entre todos os pares de indivíduos (Excoffier et al., 1992, Peakall et al., 1995; Maguire et al., 2002) e na relação entre heterozigosidades (Nei, 1973). Nestas análises estimam-se parâmetros análogos aos índices de fixação de níveis hierárquicos, conhecidos como estatísticas F, inicialmente propostas por Wright (1951). Wright definiu as estatísticas $F_{IT}$ e $F_{IS}$ como correlações entre duas unidades gaméticas (haplótipos) para produzir indivíduos aparentados considerando todas as populações e aparentados dentro das populações, respectivamente, enquanto $F_{ST}$ é a correlação entre dois gametas extraídos aleatoriamente de cada população (Nei, 1977). O grau de diferenciação entre populações pode ser medido por $F_{ST}$.

Os índices $F_{IT}$, $F_{IS}$ e $F_{ST}$ são estimados quando se consideram os níveis hierárquicos de populações, indivíduos dentro de populações e genes dentro de indivíduos. Outros parâmetros podem também ser estimados a partir destas análises, quando outros níveis são incluídos (Nei, 1973; Excoffier et al., 1992; Weir, 1996).

Pelas Tabelas 25, 26, 27 e 28 verificou-se que com os avanços nas gerações de retrocruzamento, aumentou-se a recuperação do genoma do parental recorrente ($P_3$) e, consequentemente, maior diferenciação genética com o genitor doador ($P_1$), refletidos na diminuição do percentual de variação entre as populações $P_3$-$RC_{31mp}$ para $P_3$-$RC_{34mp}$ e aumento do percentual de variação entre $P_1$-$RC_{31mp}$ para $P_1$-$RC_{34mp}$. O avanço nas gerações de retrocruzamento permite a recuperação progressiva das características do genótipo (elite) recorrente, mantendo aquela que foi

introduzida pelo genótipo doador. Em teoria, a recuperação média a cada geração é a metade da constituição genética do genitor recorrente, em relação a anterior. No primeiro cruzamento entre genótipo recorrente x genótipo (não-adaptado) doador, o híbrido tem 50% do genótipo recorrente. Na primeira geração de retrocruzamento ($RC_1F_1$) a recuperação média é de 75% e na quarta geração ($RC_4F_1$) já se tem 96,87% do genótipo recorrente.

Tabela 24. Percentual das fontes de variação genética sobre todos os 20 locos simulados para a população base $P_1$ e de retrocruzamento $RC_{31mp}$, segundo o método de Cockerham (1969, 1973), AMOVA (Excoffier et al., 1992) e Nei (1973)

| | Percentual de variação | | |
|---|---|---|---|
| Fonte de Variação | Cockerham | AMOVA | Nei |
| Entre Populações | 12,15 | 21,48 | 12,15 |
| Dentro de Populações | 43,39 | -0,75 | 87,85 |
| Dentro de Indivíduos | 44,45 | 79,27 | - |

Tabela 25. Percentual das fontes de variação genética sobre todos os 20 locos simulados para a população base $P_1$ e de retrocruzamento $RC_{34mp}$, segundo o método de Cockerham (1969, 1973), AMOVA (Excoffier et al., 1992) e Nei (1973)

| | Percentual de variação | | |
|---|---|---|---|
| Fonte de Variação | Cockerham | AMOVA | Nei |
| Entre Populações | 18,54 | 31,10 | 18,54 |
| Dentro de Populações | 41,90 | 2,16 | 81,46 |
| Dentro de Indivíduos | 39,56 | 66,74 | - |

Tabela 26. Percentual das fontes de variação genética sobre todos os 20 locos simulados para a população base $P_3$ e de retrocruzamento $RC_{31mp}$, segundo o método de Cockerham (1969, 1973), AMOVA (Excoffier et al., 1992) e Nei (1973)

| | Percentual de variação | | |
|---|---|---|---|
| Fonte de Variação | Cockerham | AMOVA | Nei |
| Entre Populações | 1,10 | 1,94 | 1,10 |
| Dentro de Populações | 47,13 | -4,36 | 98,90 |
| Dentro de Indivíduos | 51,77 | 102,42 | - |

Tabela 27. Percentual das fontes de variação genética sobre todos os 20 locos simulados para a população base $P_3$ e de retrocruzamento $RC_{34mp}$, segundo o método de Cockerham (1969, 1973), AMOVA (Excoffier et al., 1992) e Nei (1973)

| | Percentual de variação | | |
|---|---|---|---|
| Fonte de Variação | Cockerham | AMOVA | Nei |
| Entre Populações | 0,09 | -0,07 | 0,09 |
| Dentro de Populações | 49,28 | -1,10 | 99,91 |
| Dentro de Indivíduos | 50,63 | 101,17 | - |

Há de se considerar que no presente estudo as gerações de retrocruzamento não foram submetidas à seleção. Bernardo et al. (1997) afirmam que a seleção e deriva genética durante as autofecundações podem causar diferenças entre a contribuição parental observada e esperada para progênies endogâmicas. No entanto, os autores mostraram que a seleção durante o retrocruzamento geralmente favoreceu o parental recorrente em relação ao doador. Utilizando marcadores RFLP distribuídos em dez grupos de ligação do milho, os autores verificaram que, em média, a contribuição do parental recorrente em populações de $RC_1F_1$ são da ordem de 0,765, próximo ao valor esperado de 0,75.

As estimativas das estatísticas F estão presentes nas Tabelas 29, 30, 31 e 32. Segundo Gao & Hong (2000), $F_{IS}$ representa os desvios das proporções de EHW esperadas dentro das populações (aproximadamente

igual à média do coeficiente de endogamia *f* das populações); $F_{ST}$ mede a fixação de diferentes alelos em diferentes populações e $F_{IT}$ mede os desvios das proporções de EHW no sistema das populações como um todo.

Em todas as comparações, os valores de $F_{IS}$ foram não significativos, dado o valor de *p* das AMOVAs e a inclusão nos intervalos de confiança (I.C 95%) do valor zero (Tabelas 29, 30, 31 e 32). Os pequenos valores de $F_{IS}$ se justificam, dado os baixos valores dos coeficientes de fixação/endogamia *f* ($F_{IS}$) nas populações individualmente (Tabela 10), não caracterizando excesso nem deficiência de heterozigotos, a exceção de $RC_{31mp}$ (Tabela 15). Na comparação entre $P_1$ com $RC_{31mp}$ e $RC_{34mp}$, nas Tabelas 29 e 30, respectivamente, $F_{IT}$ foi alto e significativo, representando que as populações, na média dos locos, desviaram das proporções de EHW, dada a significância da AMOVA e a não inclusão no intervalo de confiança (I.C 95%) do valor zero, ao contrário da comparação entre $P_3$ e as populações de retrocruzamento. Valores de $F_{ST}$ indicam o percentual da variação genética total existente entre populações, sendo eles iguais as proporções de variação das AMOVAs (Tabelas 25, 26, 27 e 28). Embora a maioria da diversidade genética tenha sido atribuída a diferenças dentro das populações, valores significativos de $F_{ST}$ sugerem a existência de diferenciação fenotípica, a exemplo da comparação $P_3$ e $RC_{31mp}$ (Tabela 31).

Tabela 29. Estimativas das estatísticas F de Wright, para a população base $P_1$ e de retrocruzamento $RC_{31mp}$, segundo método de Cockerham (1969 e 1973), AMOVA (Excoffier et al, 1992) e Nei (1973). Intervalos de confiança (I.C) a 95% de probabilidade e valor de *p* baseados em 1000 estimativas de *bootstrap* sobre os locos

| Análise | | Sobre todos os locos | | |
|---|---|---|---|---|
| | | $F_{IS}$ | $F_{ST}$ | $F_{IT}$ |
| Cockerham | Estimativa | -0,0095 | 0,2148 | 0,2073 |
| | I.C inferior (95%) | -0,0353 | 0,1334 | 0,1397 |
| | I.C superior (95%) | 0,0144 | 0,3055 | 0,2849 |
| AMOVA | Estimativa | -0,0095 | 0,2148 | 0,2073 |
| | Valor de *p* | $0,8400^{ns}$ | $0,0000^{*}$ | $0,0000^{*}$ |
| Nei | Estimativa | -0,1764 | 0,1403 | -0,0114 |

$^{*}$ ($P < 0,05$); $^{ns}$ ($P > 0,05$)

Tabela 30. Estimativas das estatísticas F de Wright, para a população base $P_1$ e de retrocruzamento $RC_{34mp}$, segundo método de Cockerham (1969 e 1973), AMOVA (Excoffier et al, 1992) e Nei (1973). Intervalos de confiança (I.C) a 95% de probabilidade e valor de *p* baseados em 1000 estimativas de *bootstrap* sobre os locos

| Análise | | Sobre todos os locos | | |
|---|---|---|---|---|
| | | $F_{IS}$ | $F_{ST}$ | $F_{IT}$ |
| Cockerham | Estimativa | 0,0313 | 0,3111 | 0,3326 |
| | I.C inferior (95%) | 0,0067 | 0,2101 | 0,2349 |
| | I.C superior (95%) | 0,0570 | 0,4223 | 0,4393 |
| AMOVA | Estimativa | 0,0313 | 0,3111 | 0,3326 |
| | Valor de *p* | $0,0020^{ns}$ | $0,0000^{*}$ | $0,0000^{*}$ |
| Nei | Estimativa | -0,1665 | 0,2206 | 0,0908 |

$^{*}$ (P < 0,05); $^{ns}$ (P > 0,05)

Tabela 31. Estimativas das estatísticas F de Wright, para a população base $P_3$ e de retrocruzamento $RC_{31mp}$, segundo método de Cockerham (1969 e 1973), AMOVA (Excoffier et al, 1992) e Nei (1973). Intervalos de confiança (I.C) a 95% de probabilidade e valor de *p* baseados em 1000 estimativas de *bootstrap* sobre os locos

| Análise | | Sobre todos os locos | | |
|---|---|---|---|---|
| | | $F_{IS}$ | $F_{ST}$ | $F_{IT}$ |
| Cockerham | Estimativa | -0,0445 | 0,0194 | -0,0242 |
| | I.C inferior (95%) | -0,0627 | 0,0113 | -0,0374 |
| | I.C superior (95%) | -0,0261 | 0,0314 | -0,0082 |
| AMOVA | Estimativa | -0,0445 | 0,0194 | -0,0242 |
| | Valor de *p* | $1,0000^{ns}$ | $0,0000^{*}$ | $1,0000^{ns}$ |
| Nei | Estimativa | -0,1704 | 0,0118 | -0,1566 |

$^{*}$ (P < 0,05); $^{ns}$ (P > 0,05)

Tabela 32. Estimativas das estatísticas F de Wright, para a população base $P_3$ e de retrocruzamento $RC_{34mp}$, segundo método de Weir & Cockerham (1984), AMOVA (Excoffier et al, 1992) e Nei (1973). Intervalos de confiança (I.C) a 95% de probabilidade e valor de *p* baseados em 1000 estimativas de *bootstrap* sobre os locos

| Análise | | Sobre todos os locos | | |
|---|---|---|---|---|
| | | $F_{IS}$ | $F_{ST}$ | $F_{IT}$ |
| Weir & Cockerham | Estimativa | -0,0110 | -0,0007 | -0,0117 |
| | I.C inferior (95%) | -0,0319 | -0,0016 | -0,0331 |
| | I.C superior (95%) | 0,0150 | 0,0004 | 0,0147 |
| AMOVA | Estimativa | -0,0110 | -0,0007 | -0,0117 |
| | Valor de *p* | $0,8573^{ns}$ | $1,0000^{ns}$ | $0,8798^{ns}$ |
| Nei | Estimativa | -0,1247 | 0,0008 | -0,1239 |

$^{ns}$ (P > 0,05)

Para espécies arbóreas, uma divergência genética alta entre populações ocorre quando $F_{ST}$ varia entre 15 e 25% da variação total, moderada entre 5 a 15% e baixa quando $\leq$ 5%.(Kageyama et al., 2003). Espécies alógamas ou de sistema misto, mas predominantemente alógamas, com eficiente sistema de dispersão de pólen, apresentam maior parte da diversidade genética distribuída dentro das populações. As habilidades dos indivíduos trocarem genes, aliado ao fluxo gênico entre populações, reduzem as diferenças entre populações por deriva genética e seleção, reduzindo a diversidade genética entre populações.

A análise de variância de Cockerham (1969, 1973) promoveu uma partição mais eqüitativa da variação de indivíduos dentro de populações e de genes dentro de indivíduos, ao passo que a análise molecular de variância (AMOVA) atribuiu grande parte da variação a genes dentro de indivíduos (Tabelas 25, 26, 27, e 28). A variação entre populações e a soma da variação dentro de populações e dentro de indivíduos, foi quantificada de maneira idêntica pelos métodos de Cockerham e Nei (Tabelas 25, 26, 27, e 28). Valores das estatísticas F foram iguais pelos métodos de Cockerham e AMOVA.

Cockerham (1969, 1973) e Weir & Cockerham (1984) assumem que as populações sob investigação são derivadas de um ancestral comum num mesmo tempo e que todas as populações são igualmente relacionadas, existindo ou não migração entre elas. No entanto, Nei (1986) e Nei & Kumar (2000) argumentam que a aplicação desta estrutura se assemelha a grupos de populações experimentais. Em populações naturais este modelo quase nunca se aplica, pois estas são relacionadas do ponto de vista evolutivo. Nei (1977) mostra que as estatísticas F de Wright podem ser definidas como razões entre heterozigosidades, ou estatísticas H (definidas por Nei, 1973), ao invés de correlações entre unidades gaméticas. Estas definições, segundo Nei (1977), são independentes do número de alelos envolvidos, da atuação de forças evolutivas e do sistema reprodutivo da espécie.

A análise de dados moleculares via estatísticas F no entendimento de como se distribui a variação entre populações parentais e gerações segregantes é uma alternativa interessante. Desvios significativos nas proporções teóricas de recuperação do genoma do genitor recorrente

ocorrem em razão do pequeno número de plantas que são selecionadas a cada geração de retrocruzamento e da falta de controle sobre o tamanho da região genômica do parental doador próxima ao gene-alvo que é introgredido por arraste genético (*linkage drag*) (Young, et al., 1989). Caracteres como rendimento de grãos e seus componentes primários, tais como número de panículas por planta e peso de panícula principal, são determinados por grande número de genes, sendo fortemente influenciados pelo ambiente e, portanto, de difícil estimativa e seleção. Neste sentido, todo conhecimento que possa contribuir para incrementar a eficiência na seleção de caracteres governados por um grande número de genes é de suma importância nos programas de melhoramento (Lorencetti et al., 2006).

### 4.3. Comparações entre os programas computacionais

A literatura constantemente referencia os programas Arlequin, GDA, GENEPOP, GENES, POPGENE, PowerMarker e TFPGA na execução de análises qde dados genéticos. Cada um apresenta uma formatação peculiar de inserção do arquivo de dados, conforme descrita nos manuais.

Todos estes programas computacionais atendem as necessidades básicas para um estudo de estrutura e diversidade genética de unidades taxonômicas (espécies, populações e cultivares, acessos etc) a partir de dados codominantes. A limitação maior de alguns programas está na capacidade de analisar diferentes tipos de dados genéticos, sejam eles haplotípicos ou diplóides. O tempo de processamento das análises foi bastante rápido, considerando inclusive aquelas análises cujo volume de dados foi expressivo (42 populações, 8400 indivíduos e 20 locos, variando de 1 a 5 alelos). No programa Arlequin, a análise de desequilíbrio gamético, mesmo sendo realizada com apenas 5 populações apresentou uma ligeira demora durante o processamento das 1000 permutações que foram definidas. Não houve dificuldades durante o manuseio dos programas. Mas o Arlequin e PowerMarker necessitam da leitura de seus respectivos manuais, pois os dados são inseridos a partir da criação de projetos.

Evidentemente, que a leitura dos manuais é uma necessidade, uma vez que a confecção do arquivo de dados é peculiar a cada programa.

Ao GENEPOP atribuíu-se um único inconveniente que é a dependência de uma conexão à *internet* para realização das análises. Em algumas tentativas de análise o *site* do GENEPOP (*on the Web*) encontrava-se fora do ar. A Tabela 33 fornece um panorama das análises biométricas realizadas pelos programas.

O aplicativo GENES foi o que apresentou maior número de medidas descritivas, fundamentais aos estudos de diversidade genética dentro de populações, grupos ou espécies. Medidas importantes como heterozigosidades e índices de endogamia/fixação estão presentes em todos os *software*s. Embora o Arlequin não estime o índice de fixação como estatística descritiva, a AMOVA permite o cálculo deste índice para cada população avaliada. O PIC, outro coeficiente importante, encontra-se apenas no GENES e PowerMarker. O índice Shannon-Wiener apresentou variações nos programas GENES e POPGENE. Como discutido anteriormente, a base logarítmica do índice pode variar. O GENES utiliza a base *e* (ln). No entanto, foi testada a base "*e"* e as bases de 2 a 10, para o programa POPGENE. Em nenhuma delas os valores foram concordantes com a saída do programa. A variação dos programas para os estimadores de heterozigosidade esperada e índices de fixação permitem análises comparativas, embora elas tenham apresentados valores bastante concordantes, porém com filosofias diferentes.

Todos os programas são capazes de testar o equilíbrio de Hardy-Weinberg (EHW), mas por metodologias diferentes. O PowerMarker juntamente com o TFPGA são os únicos que testam o EHW via teste de qui-quadrado, razão de verossimilhança e teste exato. O GDA no teste exato pede apenas ao usuário que forneça o número de permutações para gerar as possíveis tabelas de contingência, enquanto que os outros se baseiam no processo de permutação de Guo & Thompson (1992), em que o usuário deve fornecer o número de dememorizações e *batches* para a obtenção das tabelas de contingência e, consequentemente, estimação dos valores de *p* e significância do teste. O GENEPOP, através do teste exato, possibilita inferir a respeito do excesso e deficiência de heterozigotos, por meio do teste U

Tabela 33. Análises biométricas executadas pelos programas utilizados no estudo da diversidade genética de 42 populações de melhoramento simuladas com 20 locos codominantes e multialélicos

| | Programas | | | | | | |
|---|---|---|---|---|---|---|---|
| Análise | Arlequin | GDA | GENEPOP | GENES | POPGENE | PowerMarker | TFPGA |
| Medidas descritivas | | | | | | | |
| Nº médio alelos/loco | X | X | X | X | X | X | X |
| Nº total de alelo na população | | | | X | | | |
| Nº de alelos raros | | | | X | | | |
| Nº efetivo de alelos | | | | | X | | |
| Proporção de alelos na população | | | | X | | | |
| Proporção de locos polimórficos | X | X | | X | X | | X |
| Nº médio de alelos/loco polimórfico | | X | | X | | | |
| Nº médio de genótipos/loco | | | | | X | X | |
| PIC[#] | | | | X | | X | |
| Índice Shannon-Wiener | | | X | X | | | |
| Heterozigosidade observada | X | X | X | X | X | X | X |
| Heterozigosidade esperada | X | X | X | X | X | X | X |
| Índice de fixação | | X | X | X | X | X | |
| Equilíbrio | | | | | | | |
| Hardy-Weinberg | X | X | X | X | X | X | X |
| Pares de locos | X | | | X | | X | |
| Excesso/Deficiência de heterozigotos | | | X | | | | |
| Distância “Genética” | | | | | | | |
| Euclidiana média | | | | X | | X | |
| Rogers | | | | X | | X | |
| Rogers modificada | | | | X | | | X |
| Angular | | | | X | | X | |
| Nei et al. (1983) | | | | | | X | |
| Padronizada de Nei | | X | | X | X | X | X |
| Mínima | | | | | | X | X |

Tabela 33. Continuação...

| Análise | Programas | | | | | | |
|---|---|---|---|---|---|---|---|
| | Arlequin | GDA | GENEPOP | GENES | POPGENE | PowerMarker | TFPGA |
| Distância "Genética" [##] | | | | | | | |
| Reynolds | | X | | | | X | X |
| Latter's | | | | | | X | |
| Genotípica de Hedrick | | | | X | | | |
| Agrupamento | | | | | | | |
| Otimização de Tocher | | | | X | | | |
| UPGMA | | X | | X | X | X | X |
| Projeção gráfica bidimensional | | | | X | | | |
| Estatísticas F e análogas | | | | | | | |
| ANOVA - freqüências alélicas | | X | | | | X | X |
| AMOVA | X | | | | | | |
| Estatísticas H de Nei | | | | X | X | | |
| Outros Procedimentos | | | | | | | |
| Reamostragem | X | X | X | X | | X | X |
| Teste de Mantel | X | | | X | | X | X |

[#]Conteúdo de informação polimórfica

[##]Termo generalizado a todas medidas de distância

O programa PowerMarker além de testar o equilíbrio gamético, fornece vários coeficientes de desequilíbrio, enquanto o GENES fornece somente o coeficiente de desequilíbrio clássico (D). Neste tipo de análise o usuário deve estar atento. Esta talvez tenha sido a análise mais controversa. Por exemplo, o GDA executa diferentes análises de desequilíbrio (gamético e genotípico), via teste exato, mas a aplicação deste ou daquele teste vai de encontro ao tipo de informação disponível, ou seja, são dados haplotípicos ou genotípicos? A fase gamética é conhecida ou desconhecida? No manual do Arlequin (Excoffier et al., 2006), os autores afirmam que quando a fase gamética é desconhecida, como no presente estudo, deve-se usar um procedimento diferente do teste exato, pois a composição haplotípica do conjunto de dados é desconhecida, ou melhor, apenas estimada com base nas proporções de EHW. Assim, o desequilíbrio entre um par de locos é um teste para dados genotípicos usando uma razão de verossimilhança, em que a verossimilhança da amostra avaliada sobre a hipótese de não associação entre locos (equilíbrio de ligação) forma um quociente com a verossimilhança da amostra quando a associação entre os locos existe (veja Slatikin & Excoffier, 1996). A distribuição empírica é obtida por um processo de permutação, conforme descrito na seção 3.4.1.3a. Daí a preferência pelos programas Arlequin, GENES e Powermarker, ao invés do GDA. Os aplicativos computacionais PowerMarker e GDA permitem investigações sobre o equilíbrio multiloco e multialélico. Relembrando que no presente estudo considerou-se apenas dois alelos por loco.

A avaliação da diversidade genética interpopulacional contou com medidas de distância genética ou genotípica. O PowerMarker é o aplicativo que exibe maior número de medidas de distância, inclusive com medidas não incluídas neste estudo. Este é um aspecto importante, pois as medidas de distância são baseadas em diferentes pressuposições genético-evolutivas, apesar de algumas delas serem correlacionadas e até estimarem valores muito próximos ou iguais, como relatado anteriormente (Tabela 16 a 22 e Figuras 4 a 14).

O método de agrupamento UPGMA é mais comum entre os softwares (Tabela 33). Entretanto, optou-se apenas pela saída gráfica do

programa GENES que permite definir um ponto de corte no dendrograma para definição de grupos por meio de um critério estatístico.

Os programas GDA e TFPGA permitem na análise de variância de freqüências alélicas dois a quatro níveis hierárquicos, enquanto o PowerMarker até três níveis. O Arlequin é o programa que realiza a análise de variância molecular (AMOVA) de dois a quatro níveis hierárquicos, com variações na análise conforme o tipo de dado utilizado. O GENES também realiza a AMOVA para dados haplotípicos e, juntamente com o POPGENE, estimam as estatísticas H de Nei (1973; 1977) que formam quocientes para a estimação dos índices de fixação F de Wright. O GENES ainda calcula a estatística $G_{ST}$, apesar de considerar neste tipo de análise apenas dois alelos: o de maior freqüência e os restantes agrupados.

Procedimentos de reamostragem como *bootstrap* e *jackknife* permitem testar os parâmetros F e componentes de variância correspondentes. Já o teste de Mantel tem sido a alternativa mais apropriada para comparar duas matrizes de distância.

Estes programas foram desenvolvidos por esforços individuais de pesquisadores que se preocupam em criar, distribuir e aperfeiçoá-los sem qualquer ônus ao usuário. O usuário deve testá-los e ter sensibilidade para decidir qual dele(s) é o mais apropriado as suas investigações e que apresentae, na sua concepção, uma interface mais amigável.

## 5. CONCLUSÕES

1) Populações de acasalamento ao acaso, sem sofrer efeitos de seleção, migração e mutação, mantiveram o polimorfismo genético, as proporções do equilíbrio de Hardy-Weinberg e com os avanços das gerações tenderam ao equilíbrio gamético;

2) Populações híbridas ($F_1$) foram detentoras de alelos raros e polimorfismo genético em alto grau. A alta diversidade genética de híbridos foi atribuída à contribuição gênica dos parentais. Tais populações apresentaram os locos em desequilíbrio para genes que possuiam diferenças de freqüências alélicas entre as populações cruzadas;

3) Populações de autofecundação podem exibir diversidade gênica, em função da presença continua de homozigotos na população. A proporção do polimorfismo genético foi comparável a de acasalamento ao acaso, pois autofecundações não eliminam a variação genética e sim a reorganizam em genótipos homozigotos. No entanto, a riqueza genotípica de populações submetidas à autofecundação foi reduzida ao longo das gerações. Populações oriundas de autofecundação apresentaram desequilíbrio dentro dos locos;

4) Populações $F_2$ foram detentoras de maior variabilidade genética;

5) Os tamanhos efetivos das populações foram inversamente relacionados ao nível de endogamia;

6) O desenvolvimento de populações de retrocruzamento é uma alternativa promissora na busca por genótipos superiores. O processo de hibridação no retrocruzamento pode levar um loco ao equilíbrio de Hardy-Weinberg ou não, dependendo das

combinações gênicas e genotípicas entre o hibrido $F_1$, ou a geração antecedente de retrocruzamento, e o genitor recorrente;

7) No geral as medidas descritivas foram bem correlacionadas entre si, à exceção da heterozigosidade observada e coeficiente de fixação/endogamia. Destaca-se o conteúdo de informação polimórfica, o número efetivo de alelos, índice de fixação/endogamia e as heterozigosidades esperada e observada na caracterização da diversidade dentro de populações.

8) O índice Shannon-Winner e a distância de Hedrick permitiram, respectivamente, quantificar a riqueza e a divergência genotípica das populações, muitas vezes não detectada por medidas gênicas.

9) Para grande tamanho amostral, os diferentes estimadores do coeficiente de fixação/endogamia pouco diferiram em relação a suas estimativas, assim como para as medidas viesadas e não viesadas de heterozigosidade esperada;

10) Os testes de qui-quadrado, razão de verossimilhança e teste exato foramo concordantes na detecção ou não do EHW;

11) Os testes de qui-quadrado ($\chi^2$) e razão de verossimilhança ($G^2$) foram mais concordantes em relação à detecção ou não de desequilíbrio gamético. O teste de razão de verossimilhança via LOD escore, foi mais discordante dos resultados obtidos de $\chi^2$ e $G^2$;

12) Medidas de distância genética, geométrica e genotípica, na ausência de forças evolutivas, foram eficientes no estudo de diversidade e promoveram agrupamentos semelhantes. Medidas geométricas foram mais eficientes à projeção gráfica.

13) No estudo da variação genética em níveis hieráquicos, os métodos de análise de variância de freqüências alélicas, AMOVA e pelas estatísticas H de Nei quantificaram a variabilidade de maneira semelhante, em condições de ausência de forças evolutivas. A análise via estatísticas F permitiu compreender como se distribuía a variação entre populações parentais e gerações segregantes.

14) Os programas GENES e PowerMarker apresentaram maior variedade de técnicas de análise para este estudo de diversidade genética, com base em locos multialélicos.

15) O estudo simulado possibilitou realizar comparações metodológicas e de programas computacionais de maneira satisfatória, sendo este um referencial teórico que pode auxiliar pesquisadores nas inferências sobre a genética de populações.

## 6. REFERÊNCIAS BIBLIOGRÁFICAS


Alfenas, A. C.; et al. **Eletroforese de proteínas e isoenzimas de fungos de essências florestais**. Viçosa: UFV, 1991. 242p.

Almeida, C. M. C. V.; et al. Variability in genetic resources of cacao in Rondônia, Brazil. **Crop Breeding and Applied Biotechnology**. v. 5, p. 317-323, 2005.

Ando, T.; et al. Phylogenetic Analysis of Petunia sensu Jussieu (Solanaceae) using Chloroplast DNA RFLP. **Annals of Botany**. v. 96, p. 289–297, 2005.

Barcelos, E.; et al. Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers. **Pesquisa Agropecuária Brasileira**. v. 37, n. 8, p. 1105-1114, 2002.

Barkley, N. A.; et al. Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). **Theoretical and Applied Genetics**. v. 112, p. 1519-1531, 2006.

Belaj, A.; et al. Comparative study of the discriminating capacity of RAPD, AFLP and SSR markers and of their effectiveness in establishing genetic relationships in olive. **Theoretical and Applied Genetics**. v. 107, p. 736–744, 2003.

Bernardo, R.; et al. Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. **Theoretical and Applied Genetics**. v. 100, p. 552–556, 2000.

Bernardo, R.; et al. RFLP-based estimates of parental contribution to $F_2$- and $BC_1$-derived maize inbres. **Theoretical and Applied Genetics**. v. 94, p. 652-656, 1997.

Bertini, C. H. C. M.; et al; Analysis of cotton genetic diversity by microsatellites and pedigree. **Crop Breeding and Applied Biotechnology**. v. 5, p. 369-378, 2005.

Botstein, D.; et al. Construction of a genetic linkage map in man using restriction fragment lenght polymorphisn. **American Journal of Human Genetics.** v.32, p.314-331, 1980.

Borém, A. **Melhoramento de plantas**. 2. ed. Viçosa: UFV, 1998. 453 p.

Brown-Guidera, G. L.; et al. Evalution of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. **Crop Science**. v. 40, p.815-823, 2000.

Bryan, G. J.; et al. Isolation and characterisation of microsatellites from hexaploid bread wheat. **Theoretical and Applied Genetics.** v. 94, p. 557-563, 1997.

Cardle, L. et al.; Computational and Experimental Characterization of Physically Clustered Simple Sequence Repeats in Plants. **Genetics**. v. 156, p. 847–854, 2000.

Caixeta, E. T.; Borém, A.; Kelly, J. D. Development of microsatellite markers base don BAC common bean clones. **Crop Breeding and Molecular Biotechnology**. v. 5, n. 2, p. 125-133, 2005.

Caixeta, E. T.; et al. Tipos de marcadores moleculares. In: Borém, A.; Caixeta, E. T. (Eds). **Marcadores moleculares**. Viçosa: UFV, 2006. cap. 1, 9-78.

Carlini-Garcia, L. A.; Vencovsky, R.; Coelho, A.S.G. Factorial analysis of bootstrap variances of population genetic parameter estimates. **Genetics and Molecular Biology**. v. 29, n. 2, p. 308-313, 2006.

Carlini-Garcia, L. A.; Vencovsky, R.; Coelho, A.S.G. Métodos bootstrap aplicados em níveis de reamostragem na estimação de parâmetros genéticos populacionais. **Scientia Agrícola**. v. 58, n. 4, p. 785-793, 2001.

Carlini-Garcia, L. A.; Vencovsky, R.; Coelho, A.S.G. Variance additivity of genetic populational parameter estimates obtained through bootstrapping. **Scientia Agricola**. v. 60, n. 1, p. 97-103, 2003.

Cavalli-Sforza L.; Edwards A. W. F. Phylogenetic analysis models and estimation procedure. **Evolution.** 21: 550-570, 1967.

Cockerham, C.C. Analysis of gene frequencies. **Genetics**. n.74, p. 679-700, 1973.

Cockerham, C. C. Variance of gene frequencies. **Evolution**. v. 23, p. 72-84, 1969.

Cole, C.T. Genetic variation in rare and common plants. **Annual Reviews Ecology Systems.** v. 34, p. 213-237, 2003.

Cordeiro, G. M.; et al. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. **Plant Science.** v. 160, p. 1115–1123, 2001.

Constantine, C. C.; Hobbs, R. P.; Lymbery, A. J. FOTRAN programs for analyzing population structure from multilocus genotypic data. **Journal of Heredity**. v. 85, p. 336-337.

Cruz, C.D. A informática no melhoramento genético. In: Nass, L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 34, p. 1085-1118.

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

Cruz, C. D. **Princípios de genética quantitativa**. Viçosa: UFV, 2005. 394 p.

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

Cruz, C.D. **Programa Genes**: biometria. Viçosa: UFV, 2006b. 382 p.

Cruz, C.D. **Programa Genes**: estatística experimental e matrizes. Viçosa: UFV, 2006c. 285p.

Cruz, C. D.; Viana, J. M. S. A methodology of genetic divergence analysis based on sample unit projection on two-dimensional space. **Revista Brasileira de Genética**. v.17, n.1, p. 69-73, 1994.

Dachs, N. **Estatística computacional**. Rio de Janeiro: Livros Técnicos e Científicos, 1998. 236 p.

Dias, L. A. S. Análises multidimensionais. In: Alfenas, A. C. (Ed.) **Eletroforese de isoenzimas e proteínas afins**: fundamentos e aplicações em plantas e microrganismos. Viçosa: UFV, 1998. cap. 9, p. 405-475.

Dias, L. A. S.; et al. A priori choice of hybrid parents in plants. **Genetics and Molecular Research.** v. 3, p. 356-368, 2004.

Dreisigacker, S.; et al. SSR and pedigree analyses of genetic diversity among cimmyt wheat lines targeted to different megaenvironments. **Crop Science**. v 44, p. 381-388, 2004.

Duarte, M. C.; Santos J. B.; Melo, L. C. Comparison of similarity coefficients based on RAPD markers in the common bean. **Genetics and Molecular Biology.** v. 22, p. 427-432, 1999.

Efron, B.; Tibshirani, R. J. **An introduction to the bootstrap**. New York: Chapman & Hall, 1993. 436 p

Emygdio, B. M.; Antunes, I. F.; Choel, E.; Nedel, J. L. Eficiência de coeficientes de similaridade em genótipos de feijão mediante marcadores RAPD. **Pesquisa Agropecuária. Brasileira.** v. 38, p. 243-250, 2003.

Excoffier, L.; Laval, G.; Schneider, S. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. **Evolutionary Bioinformatics Online**. v.1, p. 47-50, 2005.

Excoffier, L.; Laval, G.; Schneider, S. **Arlequin ver. 3.1**: an integrated software package for population genetics data analysis. 2006. Disponível em: < http://cmpg.unibe.ch/software/arlequin3/ >. Acesso em: 26 nov. 2006.

Excoffier, L; Slatkin, Maximum-Likelihood estimation of molecular haplotype frequencies in a diploid population. **Molecular Biology and** Evolution. v. 12, p. 921-927, 1995.1995

Excoffier, L.; Smouse, P. E.; Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. **Genetics**. v. 131, p. 479-491, 1992.

*Falconer, D. S.* ***Introdução à genética quantitativa****. Viçosa: UFV, 1987. 279 p.*

Falconer, D. S.; Mackay T. F. C. **Introduction to quantitative genetics**. 4th ed. New York: Longman, 1996. 464 p.

Faleiro, F. G.; et al. Genetic diversity of cação accessions selected for resistance to witches" broom base don RAPD markers. **Crop Breeding and Applied Biotechnology.** v. 4, p. 12-17, 2004.

Fanizza, G.; Colonna, G.; Resta, P.; Ferrara, G. The effect of the number of RAPD markers on the evaluation of genotypic distances in Vitis *vinifera*. **Euphytica**. 107: 45-50, 1999.

Felsenstein, J. **Phylogeny inference Package (PHYLIP). Ver. 3.5**. Univ. Washington, Seattle, WA. 1993.

Ferreira, D. F. **Estatística Básica.** Lavras: UFLA, 2005. 664 p.

Ferreira, D. F. Uso da simulação no melhoramento. In: Nass, L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 35, p. 1119-1141.

Ferreira, M. E.; Grattapaglia, D. **Introdução ao uso de marcadores moleculares em análise genética**. 3. ed. Brasília: EMBRAPA-CENARGEN, 1998. 220 p.

Flint-Garcia, S. A.; Thornsberry, J . M.; Buckler IV, E. S. Structure of linkage disequilibrium in plants. **Annual Reviews of Plant Biology**. v. 54, p. 357–74, 2003.

Frisch, M.; Bohn, M.; Melchinger, A. E. Comparison of Selection Strategies for Marker-Assisted Backcrossing of a Gene. **Crop Science**. v. 39, p. 1295–1301, 1999.

Fukunaga, K.; et al. Genetic diversity and population structure of teosinte. **Genetics**, v. 169, p. 2241–2254, 2005.

Gaia, J.M.D.; Mota, M. G. C.; Conceição, C. C C. Similaridade genética de populações naturais de pimenta-de-macaco por análise de RAPD. **Horticultura Brasileira**. v.22, p. 686-689, 2004.

Gao, L.; et al. Analysis of microsatellites in major crops assessed by computational and experimental approaches. **Molecular Breeding. 12:** 245–261, 2003

Gao, L-z.; Hong, S. G. De-y. Allozyme variation and population genetic structure of common wild rice *Oryza rufipogon* Griff. in China. **Theoretical and Applied Genetics**. v. 101, p. 494-592, 2000.

Garnier-Gere, P.; Dillimann, C. A computer program for testing parwise linkage disequilibria in subdivided populations. **Journal of Heredity**. v. 83, p. 239, 1992.

Garris, A. J.; et al. Genetic Structure and Diversity in *Oryza sativa* L. **Genetics**. v.169, p. 1631–1638, 2005.

Gimenes, M. A.; Lopes, C. R. Isoenzymatic variation in the germoplasm of Brazilian races of maize (*Zea mays* L.). **Genetics and Molecular Biology**. v. 23, n. 2., 375-380, 2000.

*Goldstein D.B.; et al. Genetic absolute dating based on microsatellites and the origin of modern humans.* ***Proceedings of the National Academy of Sciences****. v. 92, p. 6723-6727, 1995.*

Goodman, M. M. Distance analysis in biology. **Systematic Zoology**. v. 21, p. 174-286, 1972.

Goodman, M. M; Stuber, C. W. Races of maize: VI. Isozyme variation among races of maize in Bolivia. **Maydica**. v.28: p.169-187, 1983.

Goudet, J. FSTAT (version 1.2): a computer program to calculate *F*-statistics. **Journal of Heredity**. v. 86, p. 485-486, 1995.

Grodzicker, T.; et al. Physical mapping of temperature-sensitive mutations adenoviruses. **Cold Spring Harbor Symposia on Quantitative Biology.** v. 39, p439-446, 1974.

Guimarães, C. T.; et al. Marcadores moleculares no melhoramento de plantas. In: Borém, A.; Caixeta, E. T. (Eds). **Marcadores moleculares**. Viçosa: UFV, 2006. cap. 4, 107-144.

Guo, S. W, Thompson, E. A. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. **Biometrics.** v. 48, p. 361-372, 1992.

Haldane, J. B. S. An exact test for randomness of mating. **Journal of Genetics**. v. 52, p. 631-635, 1954.

Hallauer, A. R.; Miranda Filho, J. B. **Quantitative genetics in maize breeding**. 2. ed. Ames: Iowa State University Press, 1988. 468 p.

Hardy, G. H.; Littlewood, J. E.; Pólya, G. Some theorems concerning monotonic functions. In: **Inequalities.** 2nd ed. Cambridge, England: Cambridge University Press, 1988, p. 83-84.

Hartl, D. L.; Clark, A. G. **Principles of Population Genetics**. 3rd edition. Sunderland (MA): Sinauer Associates, 1997. 542 p.

Hedrick, P. W. A new approach to measuring genetic similarity. **Evolution**. v. 25, p. 276-280, 1971.

Hedrick, P. W. Genetic similarity and distance: comments and comparisons. **Evolution**. v. 29, p. 362-366, 1975.

Holcomb, J.; Tolbert, D. M.; Jain, K. A diversity analysis of genetic resources in rice. **Euphytica**. v. 26, p. 441-449, 1977.

Hoyt, E. **Conservação dos parentes silvestres das plantas cultivadas.** Tradução por Coradin, L. Delaware: Addison-Wesley Iberoamericana, 1992. 52 p. (Apoio IBPGR, IUCN, WWF e Embrapa/Cenargen).

Jaaska, V. Isozyme variation and phylogenetic relationships in *Vicia* subgenus *Cracca* (Fabaceae). **Annals of Botany**. 2005. Disponível em < www.aob.oxfordjournals.org/doi:10.1093/aob/mei/260 >. Acesso em: 22 abr. de 2006.

Kageyama, P. Y.; et al., Conservação *in situ* de espécies arbóreas tropicais In: Nass, L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 7, p. 149-158.

Kageyama, P. Y.; et al. Diversidade genética em espécies tropicais de diferentes estágios sucessionais por marcadores genéticos. **Scientia Forestalis**. v. 64, p. 93-107, 2003.

Kumar, S.; Tamura, K.; Nei, M. **MEGA:** molecular evolutionary genetics analysis. Version 1.01. The Pennsylvania State University, University Park, PA. 1994.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. **Psychometrika**. v. 29, n. 1, p. 1-27, 1964.

Labate, A. J. Software for population genetic analyses of molecular marker data. **Crop Science**. v. 40, p. 1521-1528, 2000.

La Rosa, R.; et al. A first linkage map of olive (Olea europaea L.) cultivars using RAPD, AFLP, RFLP and SSR markers. **Theoretical and Applied Genetics**. v. 106, p 1273–1282, 2003.

Laval, G.; SanCristobal, M.; Chevalet, C. Measuring genetic distances between breeds: use of some distances in various short term evolution models. **Genetics Selection Evolution**. v. 34, p. 481-507, 2002

Latter, B. D. H. The island model of differentiation: a general solution. **Genetics.** v. 73, p. 147-157, 1973.

Latter, B. D. H. Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. **Genetics.** v. 70, p. 475-490, 1972.

Levene H. On a matching problem arising in genetics. **Annals of mathematical statistics**. v. 21, p. 91–94, 1949.

Lewis, P. O.; Whitkus, R. GENESTAT for microcomputers. **American Society of Plant Taxonomists.** v. 2, p. 15-16, 1989.

Lewis, P. O; Zaykin, D. **GDA - Genetic Data Analysis**: version 1.1 for Windows 95/NT. 2002. Disponível em: < http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php >. Acesso em: 26 nov. 2006.

Lewontin, R. C. The interaction of selection and linkage. I. General considerations; heterotic models. **Genetics**. v. 49, p. 49-67, 1964.

Lewontin, R. C.; Kojima, K. The evolutionary dynamics of complex polymorphisms. **Evolution**. v. 14, p. 458-472, 1960.

Liu, B.H. **Statistical genomics*:* linkage, mapping, and QTL analysis**. Boca Raton: CRC Press, 1997. 611 p.

Liu, K.; Muse, S. V. PowerMarker: An integrated analysis environment for genetic marker analysis. **Bioinformatics**. v. 21, p. 2128-2129, 2005.

Liu, K.; et al. Genetic structure and diversity among maize inbred lines as Inferred from DNA microsatellites. **Genetics**. v. 165, p. 2117–2128, 2003.

Linn, S.; Arber, W. Host specificity of DNA produced by *Escherichia coli*, X. *In vitro* restriction of phage fd replicative form. **Proceeding of the Nacional Academy of Sciences**. v. 93, p. 1300-1306, 1968

Lorencetti, C.; et al. Retrocruzamento como uma estratégia de identificar genótipos e desenvolver populações segregantes promissoras em aveia. **Ciência Rural.** v.36, n4, jul-ago, 2006.

Macaubaus, C.; et al. The Complex Mutation Pattern of a Microsatellite. **Genome Research**. v. 7, p. 635–641, 1997.

Maddison, D. R.; Swofford, D. L.; Maddison, W. P. NEXUS: an extensible file format for systematic information. **Systematic Biology**. v. 46, p. 590-621, 1997.

Maguire, T. L.; Peakall, R.; Saenger, P. Comparative analysis of genetic diversity in the mangrove species *Avicennia marina* (Forsk.) Vierh. (Avicenniaceae) detected by AFLPs and SSRs. **Theoretical and Applied Genetics**. v. 104, p. 388–398, 2002.

Malécot, G. **Les Mathématiques de l'Hérédite**. Masson et Cie, Paris. 1948.

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

Market, C. L.; Moller, F. Multiple forms of enzymes tissue, ontogenetic and especies specific patterns. **Proceeding of the National Academy of Sciences**. v. 45, p. 453-462, 1959.

Marques, C. et al. Conservation and synteny of SSR loci and QTLs for vegetative propagation in four *Eucalyptus* species. **Theoretical and Applied Genetics**. v. 105, p. 474-478, 2002.

McCouch, S. R.; et al. Microsatellite marker development, mapping and applications in rice genetics and breeding. **Plant Molecular Biology.** v. 35, p. 89–99, 1997.

Meerow, A. W.; et al. Analysis of genetic diversity and population structure within Florida coconut (*Cocos nucifera* L.) germplasm using microsatellite DNA, with special emphasis on the Fiji Dwarf cultivar. **Theoretical and Applied Genetics**. v. 106, p. 715–726, 2003.

Melchinger, A. M.; Messmer, M. M.; Lee, M.; Woodman, W. L.; Lamkey, K. R. Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphism. **Crop Science.** v.31, p.669-678, 1991.

Melo Júnior, A. F.; Carvalho, D. de; Póvoa, J. S. R. Estrutura genética de populações naturais de pequizeiro (*Caryocar brasiliense* Camb.). **Scientia Florestalis.** n. 66, p. 56-65, 2004.

Mello, P C. T.; Miranda, J. E. C.; Costa, C. P. Possibilidades e limitações do uso de híbridos $F_1$ de tomate. **Horticultura brasileira.** v.6, n.2, p. 4-6, nov. 1988.

Meselson, M.; Yuan, R. DNA restriction enzyme from *E. coli*. **Nature**. v. 217, p. 1110-1114, 1968.

Meyer, A. S. et al. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L.). **Genetics and Molecular Biology.** v. 27, p. 83-91, 2004.

Michalakis, Y.; Excoffier. L. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. **Genetics**. v. 142, p. 1061-1064, 1996.

Miller, M. P. **Tools for population genetics analyses (TFPGA) 1.3**: a Windows program for the analysis of allozyme and molecular population

genetic data, 1997. Disponível em: < http://www.marksgeneticsoftware.net/tfpga.htm >. Acesso em: 26 nov. 2006.

Minella, E.; et al. Barley cultivar BRS 224. **Crop Beeding and Applied Biotecnology**. v. 5, n. 3, p. 362-363. 2005a.

Minella, E.; et al. Malting barley cultivar BRS 225. **Crop Beeding and Applied Biotecnology**. v. 5, n. 3, p. 362-363. 2005b.

Miranda J. E. C.; Costa C. P; Cruz C. D. Análise dialélica em pimentão. I. Capacidade combinatória. **Revista Brasileira de Genética**. v. 7, p. 431-440, 1988.

Mohammadi, S. A.; Prasanna, B. M. Analysis of genetic diversity in crop plants – Salient statistical tools and considerations. **Crop Science**. v. 43: p.1235-1248, 2003.

Mojena, R. Hierarchical grouping method and stopping rules: an evaluation. **Computer Journal**, v. 20, p. 359-363, 1977.

Moraes, P. L. R., Monteiro, R. & Vencovsky, R.. Conservação genética de populações de *Cryptocarya moschata* Nees (Lauraceae) na Mata Atlântica do estado de São Paulo. **Revista Brasileira de Botânica**. v. 22, p. 237-248, 1999.

Moraes, P. L. R.; Derbyshire, M. T. V. C. Diferenciação genética e diversidade em populações naturais de *CRYPTOCARYA ASCHERSONIANA* MEZ (LAURACEAE). **Biota Neotropica**. v. 3, n. 1, p. 1-10, 2003. Disponível em: < http://www.biotaneotropica.org.br/v3n1/pt/abstract?article+BN01803012003 > Acesso em: 22 set. 2006.

Moraes, P. L. R.; Derbyshire, M. T. V. C. Estrutura genética de populações naturais de *Cryptocarya Aschersoniana* Mez (lauraceae) através de marcadores isoenzimáticos. **Biota Neotropica**. v. 2, n. 2, p. 1-10, 2002. Disponível em: < http://www.biotaneotropica.org.br/v2n2/pt/abstract?article+BN02402022002> Acesso em: 22 set. 2006.

Moraes, R. M. A. et al. Genetic divergence in soybean parents for backcrossing programs. **Crop Breeding and Applied Biotecnology**. v. 5, n. 3, 2005.

Moraes, R.M.A. **Introgressão de alelos para alto teor de proteína em soja assistida por marcadores moleculares**. 2003. 105 p. Tese (Doutorado em Genética e Melhoramento)-Universidade Federal de Viçosa, Viçosa, MG, Brasil, 2003.

Morgante, M. et al. Genetic mapping and variability of seven soybean simple sequence repeat loci. **Genome**. v. 37, p. 763-769, 1994.

Morgante, M.; Hanafey, M.; Powell, W. Microssatellites are preferentially associated with norepetitive DNA in plant genomes. **Nature Genetics**. v. 30, p. 194-200, 2002.

Nass, L. L. Utilização de recursos genéticos vegetais no melhoramento. In: Nass, L. L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 2, p. 29-55.

Naylor, T. H.; et al. **Técnicas de simulação em computadores.** São Paulo: Vozes, 1971, 401 p.

Nei, M. Analysis of gene diversity in subdivided populations. **Proceedings of the National Academy of Sciences of the United States of America**. Washington, v. 70, p. 3321-3323, 1973.

Nei, M. Definition and estimation of fixation indices. **Evolution**. v. 40, p. 643-645, 1986.

Nei, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. **Genetics**. Pittsburgh, v. 89, p. 583-590, 1978.

Nei, M. Genetic distance between populations. **American Naturalist**. Chicago, v. 106, p. 238-292, 1972.

Nei, M.; Kumar, S. **Molecular evolution and phylogenetics**. New York: Oxford University Press, 2000. 333p.

Nei, M.; Tajima, F.; Tateno, Y. Accuracy of estimated phylogenetic trees from molecular data. **Journal of Molecular Evolution**. v.19, p. 153-170, 1983.

Nei, M. F-statistics and analysis of gene diversity in subdivided populations. **Annual Human Genetics**. v. 41, p.225-233, 1977.

Nordborg M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial selffertilization. **Genetics.** v. 154, p. 923–929, 2000.

Nybom, H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. **Molecular Ecology**. v.13, p.1143-1155.

Okogbenin, E.; Marin, J.; Fregene, M. An SSR-based molecular genetic map of cassava. **Euphytica**. v. 147, p. 433–440, 2006.

Ott, J. Strategies for characterizing highly polymorphic markers in human gene mapping. **American Journal of Human Genetics.** v.51, p.283-290, 1992.

Paiva, J. R.; Kageyama, P. Y. Novo enfoque do melhoramento genético da seringueira para a região amazônica. **Pesquisa Agropecuária Brasileira**. v. 28, n. 12, p. 1391-1398, dez. 1993.

Paiva, J. R.; Kageyama, P. Y.; Vencovsky, R. Genetics of rubber tree (*Hevea brasiliensis* (Willd. ex A. Juss.) Muell. Arg. 2. Mating system. **Silvae Genetica**. v. 34, p. 373-376, 1994.

Paiva, J.R.; Valois, A.C.C. Espécies selvagens e sua utilização no melhoramento. In: Nass, L. L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 4, p. 79-99.

Pal, N.; etal. Development and Characterization of Microsatellite and RFLP-Derived PCR Markers in Oat. **Crop Science**. v. 42, p. 912–918, 2002.

Peakall, R.; Smouse, P. E., Huff, D. R. Evolutionary implications of allozyme and RAPD variation in diploid populations of Buffalograss. (*Buchloë dactyloides* Nutt. Engelm.). **Molecular Ecology**. v.4, p. 135–147, 1995.

Pejic I.; et al. Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. **Theoretical and Applied Genetics**. v. 97, p. 1248–1255, 1998.

Picoli, E. A. T.; et al. Influence of RAPD number of markers and sample size on Eucalyptus genetic distance and diversity. **Crop Breeding and Applied Biotecnhology.** v.4, p. 384-390, 2004.

Póvoa, J. S. R. **Distribuição da variabilidade genética de Cedrela fissilis Vell. em fragmentos florestais no sul de Minas Gerais, por meio de isoenzimas.** Lavras, 2002. 95 p. Dissertação (Mestrado) Universidade federal de Lavras. 2002.

Powell, W. et al. The comparison of RFLP, RAPD, AFLP, and SSR (microsatellite) markers for germplasm analysis. **Molecular Breeding**. v. 2, p. 225–238, 1996.

Prasad, M.; et al. QTL analysis for grain protein content using SSR markers and validation studies using NILs in bread wheat. **Theoretical and Applied Genetics**. v. 106, p. 659–667, 2003.

Pritchard J. K.; Rosenberg N. A. Use of unlinked genetic markers to detect population stratification in association studies. **American Journal of Human Genetic.** v. 65, p. 220–228, 1999.

Querol. **Recursos genéticos, nosso tesouro esquecido:** abordagem técnica e sócio-econômica. Rio de Janeiro: AS-PTA, 1993. 206 p.

Ramalho, M. A. P.; Abreu, A. F. B.; Santos, J. B. Melhoramento de espécies autógamas. In: Nass, L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 9, p. 201-230.

Ranalli, P. Phenotypic recurrent selection in common bean (*Phaseolus vulgaris* L.) base don performance of $S_2$ progenies. **Euphytica.** v. 87 p. 127-132, 1996.

Raymond, M.; Rousset, F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. **Journal of Heredity**. v.86, p. 248-249, 1995.

Reif, J. C.; Melchinger, A. M.; Frisch, M. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. **Crop Science.** v.45: p.1-7, 2005.

Reif, J. C.; et al. Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical, and temperate germplasm by SSR Markers. **Crop Science.** v.44, p. 326–334, 2004.

Reynolds, J.; Weir, B. S.; Cockerham, C. C. Estimation of the coancestry coeffcient: basis for a short-term genetic distance. **Genetics.** 105: 767-779, 1983.

Ridley, M. **Evolução**. 3. ed. Tradução por Ferreira, H. B. Porto Alegre: Artmed, 2006. 752 p.

Robinson, I. P. Aloenzimas na Genética de Populações de Plantas. In: Alfenas, A. C. (Ed.) **Eletroforese de isoenzimas e proteínas afins:** fundamentos e aplicações em plantas e microrganismos. Viçosa: UFV, , 1998, cap. 7, p. 329-380.

Robertson, A.; Hill, W. G. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding breeding coefficients. **Genetics.** v. 107, p. 703–718, 1984.

Rogers, J. S. Measures of genetic similarity and genetic distance. In: **Studies in genetics**. VII. Austin, University of Texas, 1972. p. 145-153.

Rousset, F.; Raymond, M. Testing heterozygote excess and deficiency. **Genetics.** v. 140, p. 1413-1419, 1995.

Santacruz-Varela, A.; et al. Phylogenetic relationships among North American popcorns and their evolutionary links to Mexican and South American popcorns. **Crop Science**. v.4, p.1456-1467, 2004.

Santiago, E., Caballero, A. Effective size and polymorphism of linked neutral loci in populations under directional selection. **Genetics**. v. 149, p. 2105–2117, 1998

SAS. **SAS/STAT user´s guide. Version 6.4**. v. 2, NC, Cary, SAS Institute, 1989.

Sebbenn, A.M.; Seoane, C.E.S.; Kageyama, P.Y.; Vencovsky, R. Conseqüências do manejo florestal no sistema de reprodução de *Tabebuia cassinoides* (Lamarck) A. P. de Candolle. **Revista do Instituto Florestal**. v. 17, n. 2, p. 129-141, 2005.

Schuster, I.; Cruz, C. D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados.** Viçosa: UFV. 2004. 568 p.

Slatkin, M.; Excoffier, L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. **Heredity**. v. 76, p. 377-383, 1996.

Smouse, P. E.; Peakall, R. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. **Heredity**. v. 82, p. 561-573, 1999.

Sneath, P. H., Sokal, R. R. **Numerical taxonomy**: the principles and practice of numerical classification. San Francisco: W. H. Freeman, 1973. 573 p.

Souza, A. P. Biologia molecular aplicada ao melhoramento. In: Nass, L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 29, p. 939-965.

Souza Jr, C. L. Melhoramento de espécies alógamas. In: Nass, L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 8, p. 159-199.

Swofford, D. L.; Selander, R. B. A FORTRAN program for the comprehensive analysis of eletrophoretic data in population genetics and systematics. **Journal of Heredity**. v. 72, p. 281-283, 1981.

Takezaki, N.; Nei, M. Genetic Distances and reconstruction of trees from microsatellite DNA. **Genetics.** v. 144, p. 389-399, 1996.

Tanksley, S.D.; Medina-Filho, H.; Rick, C.M. The effect of isozyme selection on metric characters in an interspecific backcross of tomato: basis of an early screening procedure. **Theoretical and Applied Genetics.** v.60, p.291-296, 1981.

Tardin, F. D; et al.. A. Genetic diversity and determination of the optimum number of RAPD markers in lettuce (*Lactuca sativa* L.). **Acta Scientiarum: Agronomy**. v. 25, p. 1-5, 2003.

Tivang, J. G.; Nienhuis, J.; Smith, O. S. Estimation of samplig variance of molecular marker data using the bootstrap procedure. **Theoretical and Applied Genetics.** v. 89, p. 259-264, 1994.

Valois, A. C. C.; Nass, L. L.; Goes, M. Conservação *ex situ* de recursos genéticos vegetais. In: Nass, L. L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 6, p. 123-147.

Vencovsky, R. Análise de Variância de Freqüências Alélicas. **Revista Brasileira de Genética**. v. 15, p. 53-60, 1992.

Vencovsky, R.; Crossa, J. Variance Effective Population Size under Mixed Self and Random Mating with Applications to Genetic Conservation of Species. **Crop Science**. v. 39, p. 1282–1294, 1999.

Vencovsky, R. Tamanho efetivo populacional na coleta e preservação de germoplasma de espécies alógamas. **Revista IPEF**. v. 35, p. 79-84, 1987.

Waldron, B. I.; et al. RFLP mapping of QTL for fusarium head blight resistance in wheat. **Crop Science**. v. 39, p. 805–811, 1999.

Wadt, L. H.; Kageyama, P. Y. Estrutura genética e sistema de acasalamento de *Piper hispidinervum*. **Pesquisa Agropecuária Brasileira**. v. 39, n.2, p. 151-157, 2004.

Wang, J. Effective size and F-statistics of subdivided populations. I. Monoecious species with partial selfing. **Genetics**. v. 146, p. 1453–1463, 1997.

Weir, B. S.; Cockerham, C. C; Estimating F-statistics for the analysis of population structure. **Evolution**. v. 38, p. 1358-1370, 1984.

Weir, B. S. **Genetic data analysis II**. 2nd edition. Sunderland (MA): Sinauer Associates, 1996. 445p.

Werner, B. K.; Wicox, J. R. recurrente selction for yield in *Glicine max* using genetic male-sterility. **Euphytica.** v.50, p. 19-26, 1990.

Wright, S. The genetical structure of populations. **Annual of Eugenics**. v. 15, p. 313-354, 1951.

Wright, S. **Evolution and the genetics of populations. 4. v Variability within and among natural populations**. Chicago: University of Chicago Press, 1978.

Xavier, A. **Aplicação da Análise Multivariada da Divergência Genética no melhoramento de *Eucalyptus* spp.** 2003. 129 p. Tese (Doutorado Ciências Florestais) – Universidade Federal de Viçosa, Viçosa, MG. 2003.

Zheng, Y. Q.; Ennos, R. A. Genetic variability and structure of natural and domesticated populations of Caribbean pine (*Pinus caribaea* Morelet). **Theoretical and Applied Genetics**. v. 98, p. 765-771, 1999.

Zucchi, M.I.; et al. Genetic struture and gene flow in *Eugenia dysenterica* DC in Brazilian Cerrado utilizing SSR markers. **Genetics and Molecular Biology**. v. 4, n. 26, p. 449-457, 2003.

Yates, F. Contingency tables involving small numbers and the $\chi^2$ test. **Journal of the Royal Statistical Society Supplement.** v. 1, p. 217-235, 1934.

Yeh, F. C.; et al. **POPGENE, the user-friendly shareware for population genetic analysis.** Molecular Biology and Biotechnology Centre, University of Alberta, Canada. 1997. Disponível em < http://www.ualberta.ca/~fyeh/download.htm >. Acesso em: 26 de nov. 2006.

Young, N. D.; TanKsley, S. D. RFLP analysis of the size of chromosomal segments retained around the *Tm-2* locus tomato during backcross breeding. **Theoretical and Applied Genetics**. v. 77, p. 353-359, 1989.