Silvana Schneider

Estimação de proporções alélicas e genotípicas individuais de dados CNVs (Copy Number Variations)

Belo Horizonte, fevereiro de 2013

Silvana Schneider

Orientadora: Prof.ª Dra. Denise Duarte
Co-orientadora: Dra. Maíra Rodrigues

Estimação de proporções alélicas e genotípicas individuais de dados CNVs (Copy Number Variations)

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais para a obtenção do título de Mestre em Estatística.

Programa de Pós-Graduação em Estatística
Departamento de Estatística
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Belo Horizonte, fevereiro de 2013.

Agradecimentos

Primeiramente, quero agradecer a Deus pelo cuidado com minha vida e pela oportunidade de fazer um mestrado em uma instituição pública e de qualidade.

A Jesus Cristo, pela minha vida.

A minha família, a qual tenho uma gratidão especial; que apesar da distância, sempre esteve em meu coração. A minha mãe, Irene, ao meu pai, Orlando e a minha irmã, Roberta, exemplos de honestidade e humildade, sobre os quais posso dizer que até o seu silêncio me ensina.

Ao meu namorado, Antunes, por todo carinho, paciência e incentivo. Por me proporcionar muitos momentos felizes.

À professora Denise, por ter aceitado orientar-me neste trabalho. Por toda paciência durante as horas de orientação, por te me ensinado com simplicidade e humildade, se tornando uma amiga.

À Dra Maíra, pela co-orientação neste trabalho. Por todo o suporte computacional, não medindo esforços e tempo.

Ao professor Eduardo Tarazona Santos, pela bolsa de iniciação, a qual foi um grande meio para que eu desenvolvesse minha aprendizagem. E aos demais integrantes do Laboratório, por todas as explicações sobre genética.

A todos os professores do Departamento de Estatística, pelo ensino de qualidade.

Aos colegas, Talita, Mariana, Nívea, Wecsley, Paulo e Fabrícia, pelos estudos e companheirismo durante o mestrado. Aos amigos Bernardo, pela ajuda na revisão ortográfica, e Diego, pelos helps nos códigos do R.

Resumo

Copy Number Variation (CNVs) são segmentos de DNA que apresentam variações do número de cópias de uma sequência (gene), em relação ao número usual de duas cópias por indivíduo, podendo variar de um kilobase a três megabases de tamanho. E, inúmeras pesquisas já identificaram a associação de variações no número de cópias de alguns genes com diversas doenças genéticas complexas, tais como o lúpus, artrite reumatoide e diabetes do tipo 1, infecções por HIV e malária. A maioria dos métodos disponíveis para a identificação de CNVs são capazes de descrever apenas o número total de cópias de um gene ou segmento de DNA por indivíduo, deixando subjacente o número de cópias por cromossomo. Gaunt *et al*. (2010) desenvolveram um programa chamado CoNVEM, baseado no Algoritmo EM (*Expectation-Maximization*) para determinar a proporção alélica em dados haploides de CNV, supondo que os dados estão em Equilíbrio de Hardy-Weinberg (EHW). No entanto, quando os dados estão em Desequilíbrio de Hardy-Weinberg não há ferramentas estatísticas para estimarmos as proporções alélicas e genotípicas de dados de CNVs. Propomos um método capaz de estimar essas proporções quando existe endogamia na população (desequilíbrio), baseado no Algoritmo EM e na abordagem da função de verossimilhança perfilada. Os cálculos foram implementados em um programa que denominamos por CNVice (*Inbreeding Coefficients Estimation for CNV data*), utilizando linguagem de programação R, realizado em parceria com o Laboratório de Diversidade Genética Humana, http://ldgh.com.br/, da Universidade Federal de Minas Gerais.

Abstract

Copy Number Variation (CNV) are DNA segments that exhibit variations in the number of copies of a sequence (gene), in relation to the usual number of two copies per individual, ranging from one kilobase to three megabases in size. Numerous studies have identified an association of the variations in the number of copies of some genes with various complex genetic diseases such as lupus, rheumatoid arthritis and type 1 diabetes, HIV infection and malaria. Most methods available for identifying CNVs are able to describe only the total number of copies of a gene or DNA segment per individual, leaving behind the number of copies per chromosome. Gaunt *et al*. (2010) developed a program called CoNVEM, which is based on the EM algorithm (*Expectation-Maximization*) to determine the allele frequencies in haploids of CNV data, assuming the data are in Hardy-Weinberg equilibrium (HWE). However, when the data are in Hardy-Weinberg Disequilibrium there aren´t any statistical tools to estimate the proportions of allelic and genotypic data CNVs. We propose a method to estimate these proportions when there is inbreeding in the population (imbalance), based on the EM algorithm and the approach of the profiled likelihood function. The calculations have been implemented in a program that we call by CNVice (*Inbreeding Coefficients Estimation for CNV data*), using R programming language, developed in partnership with the Laboratory of Human Genome Diversity, http://ldgh.com.br/, of Federal University of Minas Gerais.

Sumário

1. Introdução

O início das pesquisas em larga escala sobre *Copy Number Variation* (CNVs) se deu em 2004, porém as descobertas iniciais datam desde 1925 (Sebat *et al*, 2004 e Sturtevant, 1925). CNVs são segmentos de DNA que apresentam variações do número de cópias de uma sequência (gene), em relação ao número usual de duas cópias por indivíduo (ou seja, uma cópia em cada cromossomo, sendo uma herdada do pai e outra da mãe). Tais segmentos podem variar de um kilobase a três megabases de tamanho (Feuk *et al.*, 2006). Um indivíduo que apresenta variação no número de cópias de um gene ou segmento de DNA pode apresentar tanto um número maior do que o usual de duas cópias (que são duplicações), ou um número menor, de uma ou zero cópias (que são deleções). A forma mais simples para representar a deleção é 0 vs. 1 e a duplicação 1 vs. 2, porém alguns lócus possuem uma maior variação do número de cópias, como, por exemplo, o gene *CCL3L1*, cujo número de cópias pode variar de 0 a 14 cópias (Gonzalez *et al.*, 2005).

Inúmeras pesquisas já identificaram a associação de variações no número de cópias de alguns genes com diversas doenças genéticas complexas, tais como o lúpus, artrite reumatoide e diabetes do tipo 1 (revisão encontrada em Mills *et al*. (2011) e Santhosh *et al*. (2011). Além de doenças genéticas, é também conhecido pelos cientistas que alguns CNVs protegem contra infecções por HIV e malária. Tais descobertas demonstram a relevância dos estudos de CNVs para a comunidade científica mundial.

Por ser uma área recente de estudo, as análises estatísticas dos dados de CNVs são bastante limitadas. Por exemplo, os métodos disponíveis para a identificação de CNVs são capazes de descrever apenas o número total de cópias de um gene ou segmento de DNA por indivíduo, deixando subjacente o número de cópias por cromossomo (chamado de alelo). Dessa forma, se um indivíduo possui 3 cópias de um determinado gene, não é possível determinar se essas cópias resultam da combinação: 3 cópias em um cromossomo e 0 no outro (ou alelos 3 e 0, e genótipo (3,0)), ou 2 cópias em um cromossomo e 1 no outro (ou alelos 2 e 1, e genótipo (2,1)).

Essa limitação se estende para estudos populacionais de CNVs (Zuccherato, 2012), os quais precisam determinar as distribuições de número de cópias total de determinado gene em uma população, bem como as distribuições das proporções alélicas e genotípicas desconhecidas (Gaunt *et al*., 2010). Especificamente, proporção alélica no caso de CNV é o número de vezes que o alelo aparece em um cromossomo (ou, mais genericamente,

em uma das fitas do DNA), e proporção genotípica é a combinação de seu aparecimento nas duas fitas. Essa situação é ilustrada na Figura 1, onde podemos ver em (b) que o cromossomo 2 apresenta uma deleção que leva à uma segmentação haploide, e o cromossomo 3 apresenta uma duplicação que leva à uma segmentação triploide, porém não se sabe em qual cromossomo está a duplicação localizada.

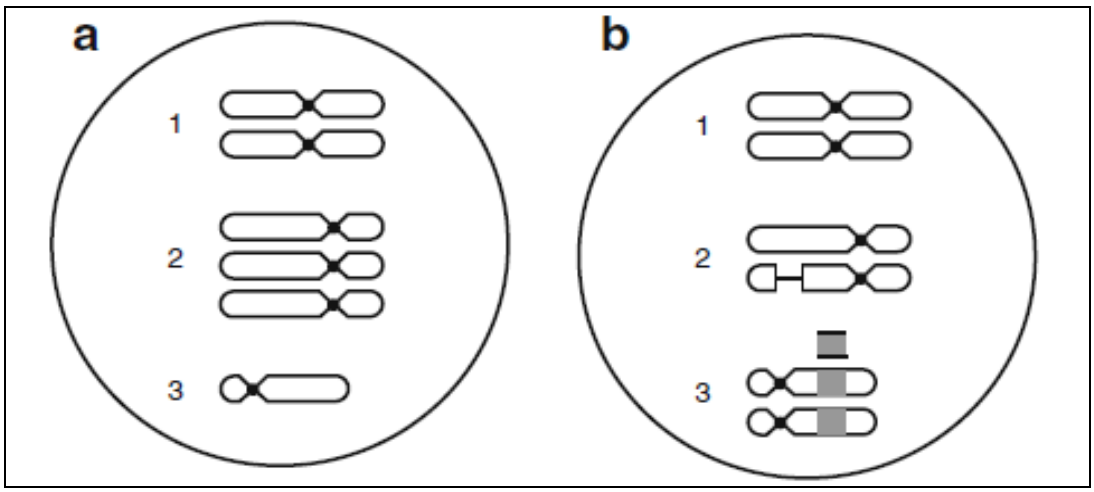


Figura 1: Ilustração esquemática de CNV, os estados dos cromossomos 1, 2 e 3 são diploide, trissômico e monossômico, respectivamente, Stenberg P. and Larsson J. (2011).

Para inferir as proporções alélica e genotípica em dados do número de cópias de determinado gene, foi desenvolvido por Gaunt *et al*. (2010) um programa chamado CoNVEM *(CNV Allele Frequency Estimation by Expectation Maximization)*, baseado no Algoritmo EM (*Expectation-Maximization*) para determinar a proporção alélica em dados haploides de CNV, supondo que os dados estão em Equilíbrio de Hardy-Weinberg (EHW). Esse programa está disponível em http://apps.biocompute.org.uk/convem/, publicado em linguagem *python* (Gaunt *et al*., 2010). O CoNVEM permite a estimação da proporção alélica e genotípica populacional, baseado na frequência do CNV observada em uma amostra.

O Algoritmo EM permite encontrar estimativas de máxima verossimilhança quando as observações podem ser vistas como dados incompletos, os dados podem ser parâmetros desconhecidos ou observações faltantes, neste caso o problema tratado são parâmetros com falta de informação para podermos estimá-los. Cada iteração do algoritmo consiste no passo da esperança (valor esperado) seguido pelo passo da maximização. Este método tem como base a ideia de substituir uma maximização de verossimilhança muito difícil por uma sequência de maximizações mais fáceis, cujo limite é a resposta para o problema original, ou seja, a estimação dos parâmetros desconhecidos (neste caso serão as proporções alélicas) converge para as estimativas de máxima verossimilhança (Dempster, 1977 e Casella, 2010).

O princípio de Hardy-Weinberg (H-W), também conhecido como Equilíbrio de Hardy-Weinberg (EHW), lida com a genética mendeliana no contexto de populações diploides, indivíduos que se reproduzem sexualmente. O princípio de H-W afirma que, se as proporções alélicas esperadas na população com dois alelos são p e q, então as proporções genotípicas esperadas serão p^2, q^2 e $2pq$, e essas proporções não sofrem alteração de uma geração para outra geração (Hardy 1908; Kempthorne, 1957).

Neste trabalho, propomos a estimação da proporção genotípica individual, com base no número de cópias do indivíduo e na amostra de indivíduos com os respectivos números de cópias. Além disso, conhecendo-se o número de cópias gênicas dos pais, a proporção genotípica dos filhos pode ser estimada com o auxílio do teorema de Hardy-Weinberg e do Teorema de Bayes (ver, por exemplo, em Shiryaev (2000)). Por exemplo, sabendo-se que o pai tem duas cópias de determinado gene, então o pai poderá ter o genótipo (0,2) ou (1,1), a mãe tem duas cópias, podendo ter o genótipo (0,2) ou (1,1), e o filho tem três cópias, porém como nem o pai nem a mãe tem o alelo 3 este filho só poderá ter o genótipo (1,2) (Figura 2).

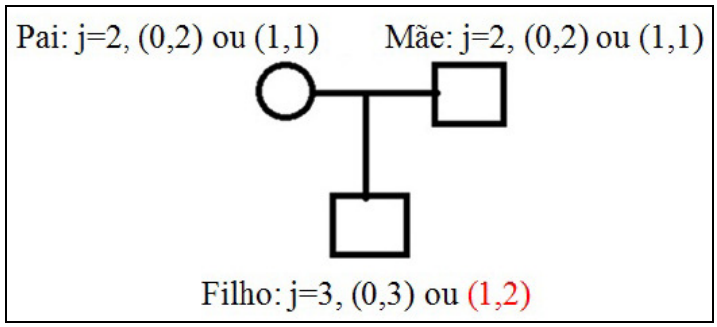


Figura 2: Figura ilustrativa com a combinação alélica dos pais e do filho.

Também propomos a inclusão do Coeficiente de Endogamia quando os dados não estão em Equilíbrio de Hardy-Weinberg. Esse Coeficiente é formalmente definido como a probabilidade de dois genes de um indivíduo qualquer, em um determinado lócus, serem cópias de um mesmo gene ancestral (Lange, 2002). A endogamia é o método de acasalamento de indivíduos cujo grau de parentesco é superior ao existente na população. O fato dos pais de um indivíduo serem geneticamente semelhantes aumenta a probabilidade de que ele receba de seus pais genes idênticos, que representam cópias de um mesmo gene presente em um ancestral comum, aumentando assim a probabilidade de genótipo homozigoto. Homozigoto é o genótipo formado por dois alelos iguais, e heterozigoto por alelos diferentes.

Na estimação do Coeficiente de Endogamia, será necessário usarmos a abordagem da função de verossimilhança perfilada, pois o objetivo primário é estimar o parâmetro de interesse, que são as proporções alélicas, e o Coeficiente de Endogamia é tratado como

parâmetro de perturbação. A ideia é substituir o parâmetro de perturbação por uma estimativa de máxima verossimilhança na verossimilhança original, considerando-se valores fixos das proporções alélicas.

Essa função de verossimilhança perfilada não é genuína, pois algumas propriedades básicas não são válidas. Por exemplo, a função escore não tem necessariamente média zero e a informação de Fisher pode apresentar vício, pode também levar à inconsistência e ineficiência dos estimadores de máxima verossimilhança. A função de verossimilhança pode levar à uma precisão excessiva. No entanto, a função de verossimilhança perfilada tem algumas propriedades interessantes, não apenas observadas na família exponencial. A estimativa de máxima verossimilhança é igual à estimativa de máxima verossimilhança perfilada. Ao se testar a hipótese sobre um parâmetro, por exemplo, p, a estatística da razão de verossimilhança baseada em $l_f(p)$, é igual à estatística da razão de verossimilhança baseada em $l(p,f)$ (Silva, 2005).

Os cálculos apresentados neste trabalho foram implementados em um programa que estende o algoritmo proposto na ferramenta CoNVEM (Gaunt *et al*. 2010), através da implementação na linguagem R. Ao programa foram adicionados o cálculo das proporções genotípicas individuais, a estimação da proporção genotípica do filho dados os genótipos dos pais, a estimação da proporção genotípica dos pais dado o genótipo do filho, o teste de aderência (Teste Kolmogorov-Smirnov, ver por exemplo, Siegel e Castellan, 2006) e a estimação do Coeficiente de Endogamia. O programa foi feito na plataforma estatística R, que tem código aberto e, portanto, pode ser facilmente expandido para incluir demais cálculos pertinentes.

Utilizamos o programa na obtenção da proporção alélica do número de cópias do gene *CCL3L1*, um dos genes com maior variação observada entre as diferentes populações mundiais, atualmente foco de pesquisa devido à sua correlação positiva entre o ácido ribonucleico mensageiro e níveis de proteína e diversas doenças (Gonzalez *et al.*, 2005; McKinney *et al.*, 2010, Zuccherato, 2012). O gene foi pesquisado na amostra que é composta por três populações nativas americanas estabelecidas entre os Andes e a região amazônica do Peru: Shimaa e Monte Carmelo, do grupo étnico Matsiguenga, e Ashaninka do grupo étnico Ashaninka, a amostra também é composta pelos Quechua, e pelos Europeus tipados por Field *et al.,* (2009).

Este trabalho foi realizado em parceria com o laboratório do programa de pós-graduação em genética do departamento de biologia da Universidade Federal de Minas Gerais, chamado Laboratório de Diversidade Genética Humana, http://ldgh.com.br/. Sendo

que o professor responsável pelo Laboratório expôs o quê existia na literatura até o momento (programa CoNVEM) e o seu interesse em obter estimativas sobre CNVs quando a população não estivesse em Equilíbrio de Hardy-Weinberg, bem como estimativas individuais e que a informação sobre os pais pudesse ser utilizada para melhorar a estimativa realizada para o filho.

Esta dissertação está dividida em cinco capítulos e dois apêndices. No primeiro capítulo, é apresentada uma breve introdução dos conceitos genéticos e estatísticos, bem como a motivação para nosso estudo. No segundo capítulo, são expostos os objetivos deste trabalho, sendo o objetivo principal a estimação das proporções alélicas para populações que estão em desequilíbrio de Hardy-Weinberg. No terceiro capítulo, são apresentados os cálculos do algoritmo utilizado no programa que denominamos de CNVice (*Inbreeding Coefficients Estimation for CNV data*), as implementações para estimar as proporções genotípicas condicionais e as extensões dos cálculos para incluir o parâmetro de Endogamia. No quarto capítulo, são apresentadas algumas simulações, uma breve descrição da amostra, bem como os resultados encontrados com a aplicação do programa em dados reais. No quinto capítulo, são apresentadas as considerações finais. E, no apêndice 1, é apresentado um fluxograma dos cálculos utilizados no programa CNVice, no apêndice 2, é exposto o cálculo do Teste da Razão de Verossimilhanças utilizado neste trabalho.

2. Objetivos

2.1. Objetivo Principal

Este trabalho tem como objetivo principal obter a estimativa da proporção alélica para dados de CNV através do número de cópias de uma sequência (gene), quando a população não está em Equilíbrio de Hardy-Weinberg, bem como a estimação do Coeficiente de Endogamia.

2.2. Objetivos Secundários

Temos como objetivos secundários os seguintes:

i. obter a estimativa da proporção alélica de uma sequência (gene) utilizando dados de CNV da amostra em estudo, supondo que a população esteja em Equilíbrio de Hardy-Weinberg;
ii. obter a estimativa da proporção genotípica individual, dado que o número de cópias do indivíduo seja conhecido;
iii. utilizar a informação sobre o número de cópias dos pais para melhorar a estimativa da proporção genotípica do filho, e utilizar a informação sobre o número de cópias do filho para melhorar a estimativa da proporção genotípica dos pais.

3. Metodologia

3.1. Formulação do Problema

A presente técnica pode ser utilizada para conjuntos de dados que são constituídos pelo número observado de cópias de uma sequência (gene), em uma amostra de indivíduos. O genótipo multilócus é definido pela combinação de dois multilócus haplotípicos, que será chamado de genótipo, e as características observáveis desse genótipo serão denominadas de fenótipo. O número de genótipos (c_j) para o j-ésimo fenótipo é uma função do número de lócus heterozigotos (s_j)(Polanska, 2003):

$$c_j = \begin{cases} 2^{s_j - 1}, s_j > 0 \\ 1, s_j = 0 \end{cases} \quad (1)$$

Por exemplo, considere o genótipo afásico (genótipo para o qual não há definido um conjunto de haplótipos), $TGTC\begin{matrix} GCA \\ TGC \end{matrix}G$, com três lócus heterozigotos, s_j=3, então a lista de todas as possíveis fases desse genótipo é ($c_j = 2^2 = 4$):

TGTCGCAG *TGTCTGCG*	*TGTCGGAG* *TGTCTCCG*
TGTCTCAG *TGTCGGCG*	*TGTCTGAG* *TGTCGCCG*

Figura 3: Lista das possíveis fases do genótipo afásico, Polanska (2003).

A técnica tem por objetivo saber qual dos quatro genótipos é o mais provável de ocorrer. Uma das possibilidades para se resolver esse problema é usando a informação da genealogia familiar, porém pode haver falta de alguns membros ou/e ter um alto custo. Sendo assim, o objetivo é achar a melhor estimativa para as proporções haplotípicas, usando somente a informação incluída em dados de genótipos afásicos.

As proporções haplotípicas podem ser estimadas via estimação de máxima verossimilhança, sob a suposição de Equilíbrio de Hardy-Weinberg e acasalamento ao acaso.

Seguindo a notação proposta na literatura da área genética, sejam os alelos haplotípicos h_k e h_l, onde k e l são números naturais que denotam o número de cópias de cada alelo e h é apenas uma letra que se refere a alelo haploide. Embora k e l não sejam

observados diretamente em um indivíduo, o total, $k+l=j$, é observado, ou seja, somente o número total de cópias (fenótipo) é observado.

O número total de cópias (fenótipo) será representado por j, ou seja, cada indivíduo pertence a uma classe j. Contudo, os membros da classe j não serão todos iguais, pois h_k varia de h_o até h_j, e h_j varia de h_o até h_k, seguindo a condição de que $k+l=j$. Por exemplo, se um indivíduo possui 5 cópias de um determinado gene (j=5), ele poderá ter os alelos 0, 1, 2, 3, 4 ou 5 pois poderá ter as seguintes composições genotípicas: (0, 5), (1, 4), (2, 3).

Sob EHW, $P(h_k,h_l)$ é a proporção do i-ésimo genótipo, composta pelos haplótipos k e l, dada por:

$$P_i(h_k, h_l) = \begin{cases} p_k^2, k = l \\ 2p_k p_l, k \neq l \end{cases} \tag{2}$$

e p_i é a proporção do i-ésimo haplótipo.

Especificamente para dados de CNV, a classe j é composta por todos os genótipos (h_k, h_l) com $k+l=j$. Então, sob EHW, a proporção da j-ésima classe, P_j, é dada pela soma de todos os possíveis genótipos que constituem a classe j,

$$P_j = \begin{cases} p_{j/2}^2, j = 0 \\ p_{j/2}^2 + \sum_{k=0}^{j/2-1} 2p_k p_{j-k}, j : par \\ \sum_{k=0}^{\frac{j-1}{2}} 2p_k p_{j-k}, j : ímpar \end{cases} \tag{3}$$

A proporção genotípica dada pela equação (2) supõe Equilíbrio de Hardy-Weinberg. Esse modelo é baseado nas suposições que seguem: (a) tamanho da população infinito; (b) gerações discretas; (c) acasalamento ao acaso; (d) ausência de seleção; (e) ausência de migração; (f) ausência de mutação; (g) proporção genotípica inicial semelhante para ambos os sexos.

O Equilíbrio de Hardy-Weinberg postula que os alelos ocorrem de forma independente, e se temos dois alelos, A_1 e A_2, os genótipos resultantes do cruzamento desses dois alelos serão A_1/A_1, A_1/A_2 e A_2/A_2. Agora consideremos o resultado do cruzamento dos genótipos A_1/A_1 e A_1/A_2, o primeiro genótipo produz somente gametas A_1, e o segundo genótipo produz gametas A_1 e A_2 na mesma proporção. Então, o cruzamento A_1/A_1 x A_1/A_2 resultará nas proporções genotípica para os filhos de ½(A_1/A_1) e ½(A_1/A_2), os demais

cruzamentos com suas respectivas proporções podem ser vistos na Tabela 1. Estas proporções, para valores resultantes de vários cruzamentos, são conhecidas como Taxa de Segregação (Lange, 2002).

Supomos que a proporção inicial para o genótipo A_1/A_1 seja u, para o genótipo A_1/A_2 seja v e para o genótipo A_2/A_2 seja w. Sob as suposições descritas acima, a próxima geração será composta pelos genótipos nas proporções apresentadas na terceira coluna da Tabela 1.

Tabela 1: Proporções genotípicas segundo Equilíbrio de Hardy-Weinberg

Tipo de Acasalamento	Natureza dos Filhos	Proporção
A_1/A_1 x A_1/A_1	A_1/A_1	u^2
A_1/A_1 x A_1/A_2	½(A1/A1 + A_1/A_2)	$2uv$
A_1/A_1 x A_2/A_2	A_1/A_2	$2uw$
A_1/A_2 x A_1/A_2	¼(A_1/A_1)+ ½(A_1/A_2) + ¼(A_2/A_2)	v^2
A_1/A_2 x A_2/A_2	½(A_1/A_2 + A_2/A_2)	$2vw$
A_1/A_2 x A_2/A_2	A_2/A_2	w^2

Fonte: Lange K. (2002).

A proporção para uma amostra com n indivíduos, condicionada nas proporções fenotípicas $P_0, P_1, ..., P_m$, é dada pela distribuição Multinomial,

$$P = \frac{n!}{n_0!n_1!...n_m!} \times P_0^{n_0} \times P_1^{n_1} \times ... \times P_m^{n_m} \tag{4}$$

onde m é o número máximo de fenótipos e n_j é o número de indivíduos que apresentam o j-ésimo fenótipo, $j \in [0; m]$ e $\sum_{j=0}^{m} n_j = n.$

Portanto, a função de verossimilhança das proporções haplotípicas, dado o número de indivíduos com cada fenótipo é:

$$f(n, p_0, p_1, ..., p_m) = c\prod_{j=0}^{m}\left(\sum_{i=0}^{c_j} P(h_{ik}, h_{il})\right)^{n_j} \tag{5}$$

c é um constante para representar $\frac{n!}{n_0!n_1!...n_m!}$.

As estimativas de máxima verossimilhança para as proporções haplotípicas poderiam, em princípio, ser encontradas analiticamente ou numericamente, resolvendo-se o conjunto de equações resultantes de (m-1) equações de derivadas parciais igualadas à zero, obtidas por:

$$\frac{\partial \log f(n, p_0, p_1, ..., p_m)}{\partial p_k} = \sum_{j=0}^{m} \frac{n_j}{P_j} \frac{\partial P_j}{\partial p_k} \tag{6}$$

3.2. Algoritmo EM

Muitas vezes, as proporções alélicas, p_k e p_l são desconhecidas, assim, não é possível encontrar as soluções analíticas para a equação (6). Mesmo sendo conhecido, o problema se torna intratável se o número máximo de alelos haplotípicos, *m,* for grande. Esse problema pode ser solucionado com a aplicação do algoritmo EM (*Expectation Maximization*), formulado por Dempster *et al.* (1977), neste trabalho ele é aplicado com o objetivo de estimar proporções haplotípicas.

Esse método tem como base a ideia de substituir uma maximização de verossimilhança muito difícil por uma sequência de maximizações mais fáceis, cujo limite é a resposta para o problema original, ou seja, a estimação dos parâmetros desconhecidos (neste caso as proporções alélicas) converge para o EMV. O processo EM é notável em parte pela simplicidade e generalização da teoria a ele associada, e em parte pela vasta amostra de exemplos que ele engloba. Quando os dados completos subjacentes vêm de uma família exponencial, cujo estimador de máxima verossimilhança é facilmente computado, então cada passo de maximização do algoritmo EM é da mesma forma facilmente computado (Dempster, 1977).

A observação de apenas um indivíduo pode produzir pouca informação sobre seus alelos h_k e h_l, mas dada uma população de indivíduos, inferências sobre as proporções alélicas, p_k e p_l, podem ser realizadas.

Assumindo-se que a população está em EHW (Hardy, 1908), a proporção genotípica é dada por:

$$P_i(h_k, h_l) = \begin{cases} p_k^2, k = l \\ 2 p_k p_l, k \neq l \end{cases} \tag{7}$$

onde , p_k e p_l são as proporções alélicas.

As proporções alélicas $p_0, p_1, ..., p_h$ são desconhecidas e, portanto, faz-se necessário o uso de um procedimento iterativo para estimar essas proporções. Começamos com os valores $p_0^{(0)}, p_1^{(0)}, p_2^{(0)}, ..., p_h^{(0)}$. A proporção genotípica na g-ésima iteração pode ser expressa por:

$$\tilde{P}_i^{(g)}(h_k, h_l) = \begin{cases} p_k^{(g)2}, k = l \\ 2p_k^{(g)} p_l^{(g)}, k \neq l \end{cases} \quad (8)$$

Os valores iniciais de $p_0^{(0)}, p_1^{(0)}, p_2^{(0)}, ..., p_h^{(0)}$ são usados para estimar as proporções genotípicas, como se fossem as reais proporções alélicas para estimar as proporções genotípicas desconhecidas.

3.2.1. Passo E

No passo E do algoritmo EM, tomamos a esperança de $f(N \mid p)$, onde N é um vetor de variáveis aleatórias que representam os valores observados do número de indivíduos para cada fenótipo, n_j e a densidade é condicionada no vetor de parâmetros p $(p_0, p_1, p_2, ..., p_m)$.

Para entendermos o Passo E, vamos considerar um exemplo: Sejam os fenótipos j=0, j=1 e j=2, e $n_0 = n_{0,0}$, $n_1 = n_{0,1}$, $n_2 = n_{1,1} + n_{0,2}$. Conhece-se apenas o número de indivíduos com cada fenótipo, ou seja, n_0, n_1, n_2 são conhecidos (Soler, 2011).

Dado que,

$$(n_{0,0}, n_{0,1}, n_{0,2} + n_{1,1}) \sim Multinomial(n, p_{0,0}, p_{0,1}, p_{0,2} + p_{1,1}),$$

tem-se que,

$$P(n_{0,2} \mid (n_{0,2} + n_{1,1}), n_{0,0}, n_{0,1}) = P(n_{0,2} = x \mid (n_{0,2} + n_{1,1}) = y, n_{0,0} = z, n_{0,1} = w)$$

$$= \frac{P(n_{0,2} = x, (n_{0,2} + n_{1,1}) = y, n_{0,0} = z, n_{0,1} = w)}{P((n_{0,2} + n_{1,1}) = y, n_{0,0} = z, n_{0,1} = w)} = \frac{P(n_{0,2} = x, n_{1,1} = y - x, n_{0,0} = z, n_{0,1} = w)}{P(n_{0,2} + n_{1,1} = y, n_{0,0} = z, n_{0,1} = w)}.$$

$$= \frac{\frac{n!}{x!(y-x)!z!w!(n-x-y+x-z-w)!} p_{0,2}^x p_{1,1}^{y-x} p_{0,0}^z p_{0,1}^w (1 - p_{0,2} - p_{1,1} - p_{0,0} - p_{0,1})^{n-x-y+x-z-w}}{\frac{n!}{y!z!w!(n-y-z-w)!} (p_{0,2} + p_{1,1})^y p_{0,0}^z p_{0,1}^w (1 - p_{0,2} - p_{1,1} - p_{0,0} - p_{0,1})^{n-x-y+x-z-w}}$$

$$= \frac{y!}{x!(y-x)!} \frac{p_{0,2}^x p_{1,1}^{y-x}}{(p_{0,2} + p_{1,1})^y} = \binom{y}{x} \left(\frac{p_{0,2}}{p_{0,2} + p_{1,1}} \right)^x \left(1 - \frac{p_{0,2}}{p_{0,2} + p_{1,1}} \right)^{y-x}.$$

Logo,

$$n_{0,2} \mid (n_{0,2} + n_{1,1}), n_{0,0}, n_{0,1}, p \sim Binomial\left((n_{0,2} + n_{1,1}), \frac{p_{0,2}}{p_{0,2} + p_{1,1}} \right).$$

Portanto,

$$\mathrm{E}(n_{0,2} \mid (n_{0,2} + n_{1,1}), n_{0,0}, n_{0,1}, p) = (n_{0,2} + n_{1,1}) \times \frac{p_{0,2}}{p_{0,2} + p_{1,1}} = n_2 \times \frac{p_{0,2}}{p_{0,2} + p_{1,1}} = n_2 \times \frac{p_{0,2}}{P_2} = n \times p_{0,2}.$$

A proporção $p_{0,2}$ é dada por $\tilde{P}_i(h_0, h_2)$, esta por sua vez segue o Equilíbrio de Hardy Weinberg através de:

$$\tilde{P}_i(h_k, h_l)^{(g)} = \begin{cases} p_k^{(g)2}, k = l \\ 2 p_k^{(g)} p_l^{(g)}, k \neq l \end{cases}.$$

Portanto, a esperança do número de indivíduos para cada genótipo é obtida por:

$$\mathrm{E}(n_{k,l} \mid Y, p^{(g)}) = n_j \times \frac{P_i^{(g)}(h_k, h_l)}{P_j^{(g)}} = n \times P_i^{(g)}(h_k, h_l) \tag{9}$$

Y denota o vetor com todos os demais $n_{k,l}$, ou seja, $Y = n - n_{k,l}$.

O número de indivíduos esperados em cada classe *j* é obtido pela multiplicação de $n \times P_j^{(g)}$, onde $P_j^{(g)}$ é a proporção fenotípica estimada na *g*-ésima iteração e *n* é o número total de indivíduos da amostra.

3.2.2. Passo M

É assumido que os dados completos possuem densidade de probabilidade $f(N \mid p)$, que é uma função do vetor de parâmetro *p* e do vetor de variáveis aleatórias N ($n_j \in [0; m]$, sendo *m* o valor do maior fenótipo. No passo E do algoritmo EM calcula-se a esperança condicional, dada por:

$$Q(p \mid p^{(g)}) = \mathrm{E}(f(N \mid p) \mid Y, p^{(g)})$$

onde $p^{(g)}$ são as estimativas de *p* na *g*-ésima iteração.

No passo M do algoritmo, maximiza-se $Q(p \mid p^{(g)})$ com respeito à p. Estas serão as novas estimativas para os parâmetros, $p^{(g+1)}$.

Voltando ao exemplo, vamos maximizar a função $Q(p \mid p^{(g)})$, que é derivada da densidade $(n_{0,0}, n_{0,1}, n_{0,2} + n_{1,1}) \sim Mult(n, p_{0,0}, p_{0,1}, p_{0,2} + p_{1,1})$, substituindo $n_{0,2}$ por $n_{0,2}^{(g)}$, e fazendo essa substituição para todos os demais números de cópias dos genótipos. A maximização pode ser realizada com a introdução do multiplicador de Lagrange.

Dado o logaritmo da função densidade

$$\ln f(N \mid p) = \ln\left(\frac{n!}{n_{0,0}!n_{0,1}!n_{0,2}!n_{1,1}!}\right) + n_{0,0}\ln(p_0^2) + n_{0,1}\ln(2p_0p_1) + n_{0,2}\ln(2p_0p_2) + n_{1,1}\ln(p_1^2),$$

as derivadas parciais são:

$$\frac{\partial \ln f(N \mid p)}{\partial p_0} = \frac{2n_{0,0}^{(g)} + n_{0,1}^{(g)} + n_{0,2}^{(g)}}{p_0} + \lambda$$

$$\frac{\partial \ln f(N \mid p)}{\partial p_1} = \frac{2n_{1,1}^{(g)} + n_{0,1}^{(g)}}{p_1} + \lambda$$

$$\frac{\partial \ln f(N \mid p)}{\partial p_2} = \frac{n_{0,2}^{(g)}}{p_2} + \lambda$$

$$\frac{\partial \ln f(N \mid p)}{\partial \lambda} = p_0 + p_1 + p_2 - 1$$

onde λ denota o multiplicador de Lagrange.

Igualando-se à zero as equações parciais acima, encontramos as soluções:

$$\begin{aligned} p_0 &= \frac{n_0}{n}\frac{P(h_0,h_0)^{(g)}}{P_0^{(g)}} + \frac{1}{2}\frac{n_0}{n}\frac{P(h_0,h_1)^{(g)}}{P_0^{(g)}} + \frac{1}{2}\frac{n_0}{n}\frac{P(h_0,h_2)^{(g)}}{P_0^{(g)}} \\ p_1 &= \frac{n_1}{n}\frac{P(h_1,h_1)}{P_1} + \frac{1}{2}\frac{n_0}{n}\frac{P(h_0,h_1)^{(g)}}{P_1^{(g)}} \\ p_2 &= \frac{1}{2}\frac{n_0}{n}\frac{P(h_0,h_2)^{(g)}}{P_2^{(g)}}. \end{aligned} \tag{10}$$

Podemos perceber que a equação (7) estima a proporção genotípica baseada nos valores iniciais das proporções alélicas, portanto é necessário fazer uma atualização da

informação, $\tilde{P}_i(h_k, h_l)$, utilizando-se os dados provenientes da amostra, $\frac{n_j}{n}$. Essa atualização ocorre no momento que substituímos $n_{0,2}$ por $n_{0,2}^{(g)}$.

Reescrevendo as soluções dadas em (10), encontramos o ponto estacionário da função. A solução resulta nas equações:

$$p_0^{(g+1)} = \frac{2n_{0,0} + n_{0,1} + n_{0,2}}{2n} = \frac{2nP^{(g)}(h_0, h_0) + nP^{(g)}(h_0, h_1) + nP^{(g)}(h_0, h_2)}{2n}$$

$$= P^{(g)}(h_0, h_0) + \frac{1}{2}P^{(g)}(h_0, h_1) + \frac{1}{2}P^{(g)}(h_0, h_2)$$

$$p_1^{(g+1)} = \frac{n_{0,1} + n_{1,1}}{2n} = \frac{nP^{(g)}(h_0, h_1) + 2nP^{(g)}(h_1, h_1)}{2n} = \frac{1}{2}P^{(g)}(h_0, h_1) + P^{(g)}(h_1, h_1)$$

$$p_2^{(g+1)} = \frac{n_{0,2}}{2n} = \frac{nP(h_0, h_2)}{2n} = \frac{1}{2}P^{(g)}(h_0, h_2).$$

Logo, generalizando as equações acima, a proporção alélica estimada para a próxima iteração é calculada por:

$$\hat{p}_t^{(g+1)} = \frac{1}{2}\sum_{i=0}^{m}\sum_{k=0}^{j}\delta_{it}P_i(h_k, h_l)^{(g)} \qquad (11)$$

onde δ_{it} indica o número de vezes que o alelo *t* aparece no genótipo $i = (h_k, h_l)$ e *m* é o valor da classe máxima. Esses valores são armazenados nos valores iniciais da equação (2) e o algoritmo se reinicia até haver convergência, quando a diferença entre a proporção alélica estimada atual e a anterior for inferior a 0,00000001 ou haver no máximo 10.000 iterações.

Na figura 4 podemos ver o fluxograma do algoritmo publicado por Gaunt *et al.* (2010), onde o passo E (Expectation step) é a equação $\tilde{P}(h_k, h_l)^{(g)}$ e as demais equações são o passo M (Maximization step).

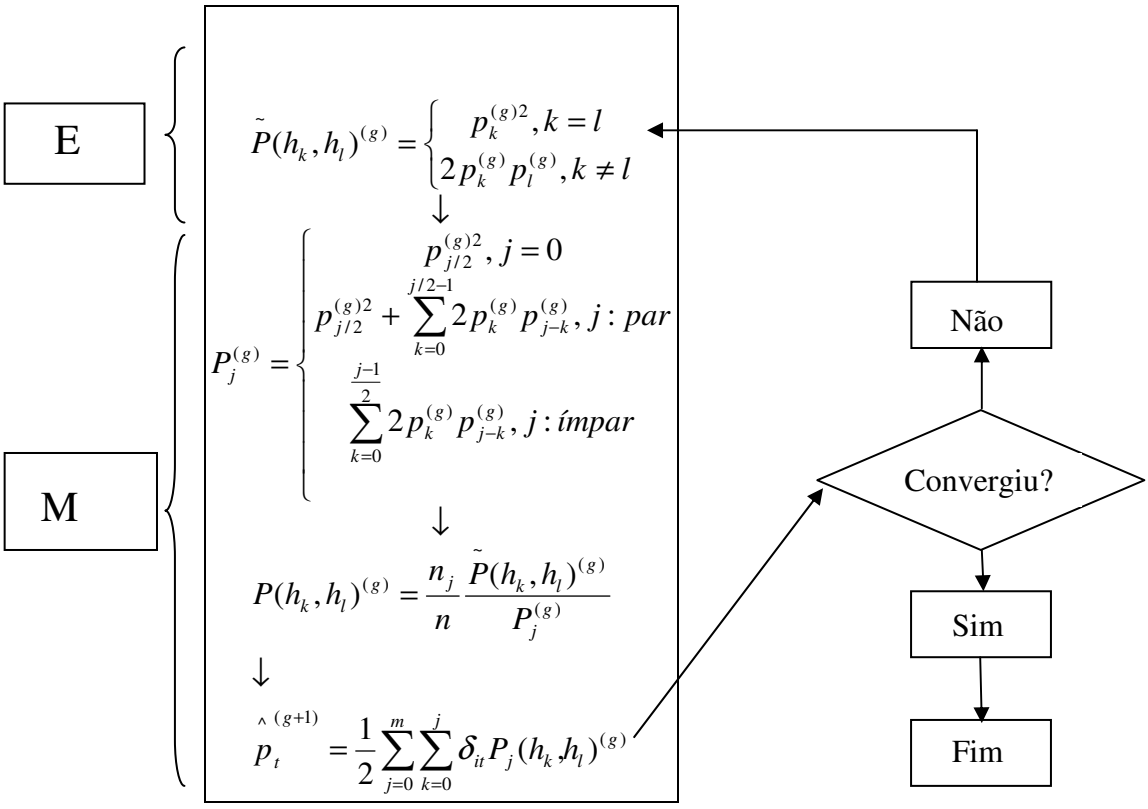


Figura 4: Fluxograma extraído de Gaunt *et al*. (2010), passos do Algoritmo EM para dados de CNV utilizado no programa CoNVEM.

A equação (12) pode ser pensada como um passo de atualização, onde $\frac{n_j}{n}$ são as proporções amostrais que atualizam $\tilde{P}(h_k,h_l)$,

$$P(h_k,h_l)^{(g)} = \frac{n_j}{n}\frac{\tilde{P}(h_k,h_l)^{(g)}}{P_j^{(g)}} \quad . \tag{12}$$

3.3. Estimação da proporção genotípica individual e das proporções condicionais

Os cálculos do algoritmo CoNVEM estimam somente as proporções populacionais. Porém, levando-se em consideração o fato de que o fenótipo, *j*, de cada indivíduo seja conhecido, a estimativa da proporção genotípica individual pode ser obtida fazendo-se uma padronização da proporção genotípica pela soma de todas as proporções genotípicas possíveis para o fenótipo *j*. Logo, a proporção genotípica individual é dada por:

$$P_{ind}(h_k, h_l \mid j)^{(g)} = \frac{2P(h_k, h_l)^{(g)}}{\sum_{k=0}^{m}\sum_{l=0}^{m} P(h_k, h_l)^{(g)}}, j = k + l \tag{13}$$

A estimativa da proporção genotípica dos filhos pode ser melhorada utilizando-se a informação do fenótipo e da proporção genotípica dos pais. Essa proporção pode ser obtida pelo uso do Teorema de Bayes, onde a proporção genotípica dos pais pode ser interpretada como a probabilidade *à priori* e a proporção genotípica do filho condicionada na proporção genotípica dos pais como sendo a probabilidade *à posteriori*. Portanto a probabilidade genotípica para um determinado filho dado a probabilidade genotípica dos pais é obtida por:

$$P(h_{kf}, h_{lf}) = \frac{P(h_{kf}, h_{lf} \mid h_{km}h_{lm}, h_{kp}h_{lp})P(h_{km}h_{lm} \mid j_m)P(h_{kp}h_{lp} \mid j_p)}{\sum_{km=0}^{m}\sum_{lm=0}^{m}\sum_{kp=0}^{m}\sum_{lp=0}^{m} P(h_{kf}, h_{lf} \mid h_{km}h_{lm}, h_{kp}h_{lp})P(h_{km}h_{lm} \mid j_m)P(h_{kp}h_{lp} \mid j_p)} \tag{14}$$

onde h_{kf} e h_{lf} denotam os alelos dos filhos, h_{km} e h_{lm} denotam os alelos da mãe e h_{kp} e h_{lp} denotam os alelos dos pais, j_m denota o fenótipo da mãe e j_p o fenótipo do pai. E $kf + lf = jf$, também é necessário que $(h_{kf} = h_{km})$ ou $(h_{kf} = h_{lm})$ ou $(h_{kf} = h_{kp})$ ou $(h_{kf} = h_{lp})$; e $(h_{lf} = h_{km})$ ou $(h_{lf} = h_{lm})$ ou $(h_{lf} = h_{kp})$ ou $(h_{lf} = h_{lp})$.

A estimativa da proporção genotípica dos pais também pode ser melhorada utilizando-se a informação do fenótipo e da proporção genotípica do filho. Neste caso a proporção genotípica do filho pode ser interpretada como sendo a probabilidade *à priori*, e a proporção genotípica dos pais dado a proporção genotípica do filho como sendo a probabilidade *à posteriori*. Logo, pelo uso do Teorema de Bayes a probabilidade genotípica dos pais condicionada na informação dos filhos é obtida por:

$$P(h_{km}, h_{lm}; h_{kp}, h_{lp} \mid h_{kf}, h_{lf}) = \frac{P(h_{km}, h_{lm}; h_{kp}, h_{lp} \mid h_{kf}, h_{lf})P(h_{kf}h_{lf} \mid j_f)}{\sum_{kf=0}^{m}\sum_{lf=0}^{m} P(h_{km}, h_{lm}; h_{kp}, h_{lp} \mid h_{kf}, h_{lf})P(h_{kf}h_{lf} \mid j_f)} \tag{15}$$

onde, $km + lm = jm$ e $kp + lp = jp$, também é necessário que $(h_{km} = h_{kf})$ ou $(h_{km} = h_{lf})$ ou $(h_{lm} = h_{kf})$ ou $(h_{lm} = h_{lf})$; e $(h_{kp} = h_{kf})$ ou $(h_{kp} = h_{lf})$ ou $(h_{lp} = h_{kf})$ ou $(h_{lp} = h_{lf})$.

As proporções genotípicas individuais $P(h_{km}h_{lm} \mid j_m)$, $P(h_{kp}h_{lp} \mid j_p)$ e $P(h_{kf}h_{lf} \mid j_f)$ das equações (14) e (15) são obtidas através da equação (13).

Para entendermos como as probabilidades condicionais das equações acima ($P(h_{kf}, h_{lf} \mid h_{km}h_{lm}, h_{kp}h_{lp})$ e $P(h_{km}, h_{lm}; h_{kp}, h_{lp} \mid h_{kf}, h_{lf})$) são obtidas vamos analisar um exemplo.

Suponha que a mãe de determinado indivíduo possua o fenótipo 5, ou seja, ela possui 5 cópias de determinado gene, que o seu pai possua o fenótipo 4, ou seja, ele possui 4 cópias de determinado gene, e o filho possui o fenótipo 3, ou seja, 3 cópias de determinado gene. A mãe poderá ter os genótipos (0,5), (1,4) e (2,3), o pai poderá ter os genótipos (0,4), (1,3) e (2,2). Com base nessas informações fazem-se todos os cruzamentos possíveis dos genótipos dos pais (Tabela 2).

Tabela 2: Genótipos possíveis para o filho, sendo o pai com 4 cópias e a mãe com 5 cópias.

		Genótipos possíveis para o pai		
		(0,4)	(1,3)	(2,2)
Genótipos possíveis para a mãe	(0,5)	(0,0)	(0,1)	(0,2)
		(0,4)	(0,3)	(0,2)
		(5,0)	(5,1)	(5,2)
		(5,4)	(5,3)	(5,2)
	(1,4)	(1,0)	(1,1)	(1,2)
		(1,4)	(1,3)	(1,2)
		(4,0)	(4,1)	(4,2)
		(4,4)	(4,3)	(4,2)
	(2,3)	(2,0)	(2,1)	(2,2)
		(2,4)	(2,3)	(2,2)
		(3,0)	(3,1)	(3,2)
		(3,4)	(3,3)	(3,2)

Sabendo-se o fenótipo do filho, que neste caso é 3 cópias, ele poderá ter os genótipos (0,3), (3,0), (1,2) e (2,1). Portanto, a probabilidade de o filho ter o genótipo (3,0) é 0,25, pois ele só poderá ter esse genótipo se a mãe possuir o genótipo (2,3) e o pai possuir o genótipo (0,4), dos quatro possíveis genótipos resultantes desse cruzamento, (2,3) x (0,4), o genótipo (3,0) é um deles, ou seja, a probabilidade será ¼. A probabilidade de o filho possuir o genótipo (3,0) também será 0,25. Já a probabilidade dele possuir o genótipo (1,2) é 0,50, pois dos quatro possíveis genótipos resultantes do cruzamento (1,4) x (2,2), o genótipo (1,2) aparece duas vezes, ou seja, a probabilidade será ½. E, finalmente, a probabilidade de o filho possuir o genótipo (2,1) também será 0,25.

As probabilidades condicionais são obtidas de forma semelhante ao exemplo acima, basta fazer todos os possíveis genótipos para os pais, depois fazer o cruzamento de todos os genótipos e ver quais deles podem ser o genótipo do filho (que resultam no fenótipo do filho).

3.4. Estimação do Coeficiente de Endogamia.

Sob Equilíbrio de Hardy-Weinberg os cálculos apresentados na Seção 3.2 estimam perfeitamente as proporções alélicas e genotípicas, porém quando a população não está em equilíbrio não é aconselhado o uso do algoritmo CoNVEM descrito em Gaunt *et al.* (2010), pois os resultados podem não ser fidedignos ou pode não haver convergência. Neste caso, quando os dados não estão em Equilíbrio de Hardy-Weinberg, esta a nossa maior contribuição neste trabalho.

Propomos uma alternativa para solucionar esse caso, incluir um parâmetro, que denotamos por *f,* capaz de captar o mínimo excesso de homozigotos quando os dados sugerem que há endogamia. Esse parâmetro mede o grau de endogamia, conhecido como Coeficiente de Endogamia, formalmente definido como a probabilidade de que dois genes de um indivíduo qualquer, em um determinado lócus, sejam cópias de um mesmo gene ancestral. A endogamia é um termo genérico para se referir a acasalamentos onde os indivíduos estão acoplados (acasalados) a indivíduos mais estreitamente relacionados do que com membros aleatórios de toda a população. A noção da proximidade do relacionamento entre duas pessoas é facilmente visualizada em um simples caso, por exemplo, pais e filhos são mais próximos relacionados do que avós e netos, da mesma forma o acasalamento de duplos primos de primeiro grau é considerado em genética como sendo endogamia (consanguinidade). O objetivo é fazer com que seja possível obter uma medida quantitativa da intensidade entre consanguinidade e grau da relação (Kempthorne, 1957).

O genótipo homozigoto (h_k, h_k) tem probabilidade $fp_k + (1-f)p_k^2$ (Lange, 2002), onde p_k é a probabilidade de dois genes serem cópias de um mesmo gene ancestral h_k e p_k^2 é a probabilidade de dois genes não serem cópias de um mesmo gene ancestral, sendo cópias independentes de h_k. O genótipo heterozigoto (h_k, h_l) tem probabilidade $(1-f)2p_k p_l$, pois é impossível que os genes sejam cópias do mesmo gene ancestral (Lange, 2002). Sendo assim, a proporção genotípica pode ser obtida pelo cálculo:

$$\tilde{P}_f(h_k, h_l) = \begin{cases} fp_k + (1-f)p_k^2, k = l \\ (1-f)2p_k p_l, k \neq l \end{cases}. \tag{16}$$

Da mesma forma que na equação (7), propomos um método iterativo para estimar as proporções, porém agora as proporções alélicas p_0, p_1,..., p_h e o parâmetro f são

desconhecidos. Começamos com os valores $p_0^{(0)}, p_1^{(0)}, p_2^{(0)},..., p_m^{(0)}$ para as proporções alélicas, e, podemos obter as proporções genotípicas na *g*-ésima iteração através do uso do algoritmo EM.

3.4.1. Passo E

Da mesma forma que na seção 3.2 o valor esperado de $f(N \mid p, f)$ será dada por:

$$\mathrm{E}(n_{k,l} \mid Y, p^{(g)}) = n_j \times \frac{P_f^{(g)}(h_k, h_l)}{P_{f_j}^{(g)}} = n \times P_f^{(g)}(h_k, h_l). \tag{17}$$

Y denota o vetor com todos os demais $n_{k,l}$, ou seja, $Y = n - n_{k,l}$.

Onde $P_f^{(g)}(h_k, h_l)$ é dado pela equação (16), e $P_{f_j}^{(g)}$ é dado pela equação abaixo:

$$P_{f_j}^{(g)} = \begin{cases} f^{(g)} p_{j/2}^{(g)} + (1 - f^{(g)}) p_{j/2}^{(g)2}, j = 0 \\ f^{(g)} p_{j/2}^{(g)} + (1 - f^{(g)}) p_{j/2}^{(g)2} + \sum_{k=0}^{j/2-1} (1 - f^{(g)}) 2 p_k^{(g)} p_{j-k}^{(g)}, j : par \\ \sum_{k=0}^{\frac{j-1}{2}} (1 - f^{(g)}) 2 p_k^{(g)} p_{j-k}^{(g)}, j : \acute{i}mpar \end{cases}. \tag{18}$$

A função de verossimilhança conjunta de *f* e de *p*, sendo *p* o vetor de todas as proporções p_k e p_l com $k+l=j$, é dada por:

$$L(f, p) = \prod_{j=0}^{jm\acute{a}ximo} (P_{f_j}^{(g)})^{n_j} \tag{19}$$

onde n_j é o número de indivíduos com a classe *j*, P_j é a proporção da classe *j* com todas as proporções genotípicas possíveis, e *jmáximo* é o valor da classe mais alta.

Para resolvermos a função de verossimilhança conjunta de *f* e de *p*, dada pela equação (19), vamos utilizar a verossimilhança perfilada. Essa função possui um parâmetro de perturbação, *f*, e ele é substituído por uma estimativa consistente na verossimilhança original, e estimado com valores fixados de *p*.

Essa função de verossimilhança não é uma verossimilhança genuína, pois algumas propriedades básicas não são válidas. Por exemplo, a função escore não tem necessariamente média zero e a informação de Fisher pode apresentar vício, pode também levar a inconsistência e ineficiência dos estimadores de máxima verossimilhança. A função de verossimilhança perfilada tipicamente induz "excesso de precisão", dado que trata do parâmetro de incômodo como conhecido, uma função dos dados e do parâmetro de interesse, e assim, despreza a incerteza inerente à estimação dos parâmetros de perturbação (Silva, 2005).

Porém, a função de verossimilhança perfilada tem algumas propriedades interessantes, não apenas observadas na família exponencial, e são:

i) a estimativa de máxima verossimilhança é igual à estimativa de máxima verossimilhança perfilada, ou seja, se θ é o vetor paramétrico θ=(p, f) e a verossimilhança perfilada é dada por $L_f(p)=L(p, \hat{f}_p)$, $\hat{f}_p$ é estimativa de máxima verossimilhança de f para p fixado. Então $\hat{p}_p = \hat{p}$, onde $\hat{p}$ é a estimativa de máxima verossimilhança de p e $\hat{p}_p$, a estimativa de máxima verossimilhança perfilada;

ii) Ao se testar a hipótese sobre p, a estatística da razão de verossimilhança baseada em $l_f(p)$, $l_f(p) = \log L_f(p)$, é igual à estatística da razão de verossimilhança baseada em $l(p,f)$, ou seja, $RV = 2\left[l(\hat{p},\hat{f}) - l(p,\hat{f})\right] = 2\left[l_p(\hat{p}) - l_p(p)\right]$, onde $\hat{f}$ é a estimativa de máxima verossimilhança de f ;

iii) A função escore perfilada, denotada por $u_p(p)$, é dada por:

$$l_p(p,\hat{f}_p) = \left.\frac{\partial l(p,f)}{\partial p}\right|_{(p,f)=(p,\hat{f}_p)} = u_p(p)$$

iv) A informação de Fisher perfilada observada é igual à informação observada, ou seja:

$\left(j_p(p)\right)^{-1} = j^{pp}(p,\hat{f}_p)$, onde $j_p(p) = -\partial^2 l_p(p)/\partial p^T \partial p$ é a informação de Fisher perfilada observada e $j^{pp}(p,f)$ é o bloco superior esquerdo do inverso da matriz de informação de Fisher observada $j(p,f) = -\partial^2 l(p,f)/\partial(p,f)^T \partial(p,f)$;

v) As estatísticas de Wald e escore de Rao são respectivamente:

$$W = (\hat{p} - p)^T (i^{pp}(\hat{p}, \hat{f}))^{-1} (\hat{p} - p) \text{ e } S_R = l_p^T(p, \hat{f}_p) i^{pp}(p, \hat{f}_p) l_p(p, \hat{f}_p)$$

onde $i^{pp}(p, f) = \mathrm{E}\{j^{pp}(p, f)\}$. E, assintoticamente, estas estatísticas têm distribuição qui-quadrado com número de graus de liberdade igual à dimensão do vetor *p*. Em tais expressões, a informação esperada pode ser substituída pela informação observada sem modificação da distribuição assintótica, e consequentemente por *iii*) e *iv*),

$$W = (\hat{p} - p)^T j_p(\hat{p})(\hat{p} - p) \text{ e } S_R = u_p^T(p)(j_p(p))^{-1} u_p(p).$$

Dada *L(f, p)* , a função de verossimilhança perfilada de *p* pode ser obtida por (Pawitan, 2001):

$$L(p) = \max_f L(f, p) \tag{20}$$

onde a maximização é realizada para valores fixos de *p*.

O logaritmo da verossimilhança dado por:

$$\log L(f, p) = \sum_{j=0}^{jmáximo} n_j \log(P_j^{(g)}). \tag{21}$$

E, por fim, a função escore perfilada é dada por:

$$\frac{\partial}{\partial f} \log L(f, p) = \sum_{j=0}^{jmáximo} n_j \frac{\frac{\partial}{\partial_f} P_j^{(g)}}{P_j^{(g)}}$$

$$= \sum n_j \frac{[p_{j/2} - p_{j/2}^{(g)2}] I_{j=0} + [p_{j/2} - p_{j/2}^{(g)2} - \sum_{k=0} 2p_k p_{j=k}] I_{j2n} + [-\sum 2p_k p_{j-k}] I_{j=2n+1}}{\{[fp_{j/2} + (1-f)p_{j/2}^{(g)2}] I_{j=0} + [fp_{j/2} + (1-f)p_{j/2}^{(g)2} + \sum_{k=0}^{j/2-1} (1-f)2p_k^{(g)} p_{j-k}^{(g)}] I_{j=2n} + [\sum_{k=0}^{\frac{j-1}{2}} (1-f)2p_k^{(g)} p_{j-k}^{(g)}] I_{j=2n+1}\}}.$$

Deve-se enfatizar que para valores fixos de p a Estimativa de Máxima Verossimilhança de f é geralmente uma função de p, assim pode ser escrito como:

$$L(p) = L(\hat{f}_p, p). \tag{22}$$

A verossimilhança perfilada é então tratada como uma verossimilhança normal, e tendo-se o valor da estimativa de f, volta-se nas equações iniciais para calcular a proporção genotípica:

$$\tilde{P}_{\hat{f}}(h_k, h_l) = \begin{cases} \hat{f}p_k^{(g)} + (1-\hat{f})\, p_k^{(g)2}, k = l \\ (1-\hat{f})2\, p_k^{(g)} p_l^{(g)}, k \neq l \end{cases} \tag{23}$$

Posteriormente a proporção da classe j considerando $k+l=j$ é obtida por:

$$P_{\hat{f}\,j}^{(g)} = \begin{cases} \hat{f}p_{j/2} + (1-\hat{f})\, p_{j/2}^{(g)2}, j = 0 \\ \hat{f}p_{j/2} + (1-\hat{f})\, p_{j/2}^{(g)2} + \sum_{k=0}^{j/2-1} (1-\hat{f})2\, p_k^{(g)} p_{j-k}^{(g)}, j : par \\ \sum_{k=0}^{\frac{j-1}{2}} (1-\hat{f})2\, p_k^{(g)} p_{j-k}^{(g)}, j : \acute{i}mpar \end{cases} \tag{24}$$

E, tem-se então:

$$\mathrm{E}(n_{k,l} \mid Y, p^{(g)}) = n_j \times \frac{P_{\hat{f}}^{(g)}(h_k, h_l)}{P_{\hat{f}_j}^{(g)}} = n \times P_{\hat{f}}^{(g)}(h_k, h_l).$$

3.4.2. Passo M

No passo M tem-se que os dados completos possuem densidade de probabilidade $f(N \mid p, f)$, e, portanto

$$Q(p \mid p^{(g)}) = \mathrm{E}(f(X \mid p) \mid Y, p^{(g)}, f^{(g)})$$

As derivadas parciais de $Q(p|p^{(g)})$ são igualadas a zero e os $n_{k,l}$ são substituídos por $n_{k,l}^{(g)}$, considerando-se as proporções genotípicas obtidas por:

$$P_{\hat{f}}(h_k, h_l) = \frac{n_j}{n} \frac{\tilde{P}_{\hat{f}}(h_k, h_l)^{(g)}}{P_{\hat{f}_j}^{(g)}}. \tag{25}$$

E, a proporção alélica é então obtida por:

$$\hat{p}_{\hat{f}_t}^{(g+1)} = \frac{1}{2} \sum_{j=0}^{m} \sum_{k=0}^{j} \delta_{it} P_{\hat{f}_j}(h_k, h_l)^{(g)}. \tag{26}$$

onde δ_{it} indica o número de vezes que o alelo *t* aparece no genótipo $i = (h_k h_l)$ e *m* é o valor da classe máxima. Esses valores são armazenados nos valores iniciais da equação (14) e o algoritmo se reinicia até haver convergência, quando a diferença entre a proporção alélica estimada atual e anterior for inferior a 0,00000001 ou haver no máximo 10.000 iterações.

O fluxograma dos cálculos implementados no programa, com o objetivo de estimar o Coeficiente de Endogamia e as proporções alélicas, pode ser analisado no apêndice 1.

4. Simulações e Resultados

4.1. Simulações

Os cálculos expostos na seção 3 foram implementados em um programa que chamamos de CNVice (*Inbreeding Coefficients Estimation for CNV data*), utilizando linguagem de programação R. No momento, os códigos podem ser disponibilizados mediante solicitação, através do *e-mail* silvanaschneider@hotmail.com.

Para testarmos o programa, criamos um algoritmo capaz de gerar amostras, supondo o parâmetro f (Coeficiente de Endogamia) e o vetor de parâmetros p (proporções alélicas) conhecidos. De posse da amostra, roda-se o programa, que deve encontrar os valores estimados de f e p iguais, ou aproximados, aos valores utilizados para gerar as amostras.

Foram feitas simulações de Monte Carlo, com 1000 repetições e considerando-se a distribuição Multinomial. Geramos diferentes proporções alélicas, com tamanho amostral (N) igual a 100 e Coeficiente de Endogamia (f) iguais a 0,05; 0,1; 0,2; 0,3 e 0,4. Todas as análises foram realizadas através do uso do *software* R 2.12.2, que está disponível em http://cran.r-project.org/.

Nas tabelas 3, 4 e 5 são apresentadas algumas das simulações realizadas. Elas estão organizadas da seguinte maneira: na primeira coluna, está exposto o alelo de um gene hipotético; na segunda coluna, as supostas proporções alélicas conhecidas para este gene hipotético; e nas demais colunas estão expostas as proporções estimadas (valor esperado e Intervalo de Confiança), obtidas através das simulações de Monte Carlo.

Com o objetivo de varrer todo o espaço paramétrico para as proporções alélicas, escolhemos diferentes cenários para estas proporções. Algumas proporções estão distribuídas de maneira mais uniforme entre os alelos, outras possuem uma proporção maior para os alelos zero, 1 e 2, e, outras possuem uma proporção maior para os alelos 8 e 9. Juntamente com cada vetor de proporções alélicas, foi incluído o coeficiente de endogamia, no cálculo para gerar as amostras.

Considerando-se o vetor de proporções alélicas (0,1; 0,1; 0,1; 0,1; 0,2; 0,1; 0,1; 0,1; 0,1), podemos ver que os valores esperados para as proporções alélicas são próximos das proporções verdadeiras para todos os alelos com exceção do alelo 4, onde o valor verdadeiro é 0,2 e a estimativa é 0,14993 quando se supõem coeficiente de endogamia f=0,05, e considerando-se f=0,4 a estimativa para a proporção do alelo 4 é 0,18271 [0,14838; 0,21705], esta se aproxima mais do verdadeiro valor, 0,2, ver Tabela 3.

Ainda na Tabela 3, notamos que as estimativas para o vetor de proporções (0,2; 0,6; 0,2; 0; 0; 0; 0; 0; 0) são mais precisas. Por exemplo, considerando o coeficiente de endogamia

f=0,1 obtivemos a proporção estimada para o alelo zero igual a 0,20173 [0,19835; 0,20511], lembrando que a proporção verdadeira é 0,2.

Analisando-se a Tabela 4, podemos notar que, considerando-se o vetor de proporções alélicas (0,00; 0,05; 0,05; 0,10; 0,60; 0,10; 0,05; 0,05; 0,00) a estimativa para estas proporções melhora com o aumento do Coeficiente de Endogamia. Por exemplo, para o alelo 4 com proporção verdadeira igual a 0,60, o seu valor estimado considerando-se o Coeficiente *f*=0,05 foi 0,44226 [0,02572; 0,85880], e considerando-se *f*=0,4 foi 0,48290 [0,0000; 0,84406]. Nesta mesma tabela, o vetor de proporções alélicas (0,3; 0,4; 0,3; 0,0; 0,0; 0,0; 0,0; 0,00; 0,00) possui valor estimados mais precisos que as proporções anteriores, por exemplo: para o alelo 2 a estimativa é 0,28642 [0,27266; 0,30019] com *f*=0,05 e 0,31282 [0,28344; 0,34220] com *f*=0,4.

Para o vetor de proporções (0,01; 0,01; 0,01; 0,01; 0,01; 0,01; 0,02; 0,02; 0,90) as estimativas não são muito boas, principalmente para o alelo 9, onde a proporção verdadeira é 0,90 é a estimada é 0,43021 [0,000; 1,000] quando *f*=0,05. Isso pode ser devido ao vetor ter uma proporção quase zero para todos os alelos exceto para o alelo 9, que possui uma proporção quase igual a 1, ver Tabela 5.

O vetor de proporções alélicas (0,000; 0,080; 0,770; 0,110; 0,029; 0,000), com alelos de zero até 5, foi assim escolhido para representar os dados de uma população com baixos número de cópias. Podemos notar que suas estimativas são próximas das verdadeiras proporções, porém não tão precisas, por exemplo, para o alelo 2 temos valor esperado igual a 0,6869 e IC [0,3647; 1,000].

Através dos histogramas podemos analisar a distribuição das proporções e do Coeficiente de Endogamia estimados. Na Figura 5 temos os histogramas para o vetor de proporções (0,2; 0,6; 0,2; 0; 0; 0; 0; 0; 0), e notamos que para o Coeficiente de Endogamia *f*=0,05 há uma alta frequência próximo de zero, porém com uma suave cauda à direita, o que faz com que a sua estimativa não seja tão precisa. Para as demais proporções a distribuição é simétrica em torno do verdadeiro valor. O mesmo ocorre para o para o vetor de proporções (0,3; 0,4; 0,3; 0; 0; 0; 0; 0; 0), ver Figura 6.

Na Tabela 6, estão as estimativas quando consideramos apenas o valor esperado das proporções para gerarmos as amostras, ou seja, geramos apenas uma amostra para cada vetor de proporções alélicas e rodamos o programa. Notamos que os valores estimados são bem próximos dos verdadeiros valores. Por exemplo, para o alelo 2 do vetor de proporções (0,2; 0,6; 0,2; 0; 0; 0; 0; 0; 0), o valor estimado é 0,2000019 com *f*=0,05, 0,2000003 com *f*=0,1, 0,1999998 com *f*=0,2, 0,1999996 com *f*=0,3 e 0,1999990 com *f*=0,4. Ainda neste mesmo caso, podemos notar que os *f*´s estimados são bem próximos dos verdadeiros valores.

Tabela 3: Proporções Alélicas e Coeficiente de Endogamia estimados, valor esperado e Intervalo de Confiança.

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,05	[IC 95%]	Porp.est.,f verd.= 0,1	[IC 95%]	Porp.est.,f verd.= 0,2	[IC 95%]
0	0,1	0,09668	[0,06342; 0,12995]	0,10153	[0,06921; 0,13385]	0,10533	[0,07330; 0,13736]
1	0,1	0,10119	[0,04658; 0,15579]	0,09917	[0,04737; 0,15096]	0,09639	[0,05346; 0,13933]
2	0,1	0,10510	[0,01483; 0,19537]	0,10560	[0,02386; 0,18734]	0,10907	[0,03987; 0,17827]
3	0,1	0,11663	[0,03224; 0,20101]	0,11251	[0,03677; 0,18824]	0,10986	[0,05351; 0,16620]
4	0,2	0,14993	[0,02492; 0,27493]	0,15322	[0,04541; 0,26103]	0,15986	[0,07601; 0,24372]
5	0,1	0,13284	[0,03620; 0,22948]	0,12743	[0,04409; 0,21077]	0,11495	[0,05831; 0,17159]
6	0,1	0,11939	[0,00761; 0,23116]	0,11230	[0,01498; 0,20962]	0,10664	[0,03176; 0,18151]
7	0,1	0,09576	[0,01298; 0,17853]	0,09792	[0,02667; 0,16918]	0,09792	[0,04308; 0,15277]
8	0,1	0,06488	[0,00229; 0,12747]	0,07563	[0,01478; 0,13648]	0,09290	[0,04445; 0,14135]
	f estimado	0,05316	[0,00000; 0,20805]	0,08369	[0,00000; 0,27313]	0,17085	[0,00000; 0,41368]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,3	[IC 95%]	Porp.est.,f verd.= 0,4	[IC 95%]
0	0,1	0,10695	[0,08316; 0,13073]	0,10469	[0,08582; 0,12357]
1	0,1	0,09595	[0,06867; 0,12323]	0,09801	[0,08029; 0,11573]
2	0,1	0,10273	[0,06042; 0,14504]	0,10602	[0,08004; 0,13201]
3	0,1	0,10573	[0,07270; 0,13875]	0,10296	[0,08443; 0,12148]
4	0,2	0,17757	[0,12695; 0,22818]	0,18271	[0,14838; 0,21705]
5	0,1	0,10925	[0,07557; 0,14294]	0,09770	[0,07862; 0,11677]
6	0,1	0,10182	[0,05611; 0,14752]	0,05661	[0,07876; 0,13256]
7	0,1	0,09343	[0,06155; 0,12531]	0,09383	[0,07465; 0,11301]
8	0,1	0,10310	[0,07330; 0,13290]	0,10700	[0,08641; 0,12759]
	f estimado	0,26574	[0,00000; 0,53719]	0,36587	[0,10027; 0,63147]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,05	[IC 95%]	Porp.est.,f verd.= 0,1	[IC 95%]	Porp.est.,f verd.= 0,2	[IC 95%]
0	0,2	0,19705	[0,19576; 0,19834]	0,20173	[0,19835; 0,20511]	0,20424	[0,18876; 0,21971]
1	0,6	0,60595	[0,60215; 0,60974]	0,60081	[0,59176; 0,60986]	0,59166	[0,55854; 0,62477]
2	0,2	0,19529	[0,19122; 0,19936]	0,19516	[0,18646; 0,20385]	0,20286	[0,18197; 0,22375]
3	0,0	0,00171	[0,00025; 0,00318]	0,00231	[0,00000; 0,00511]	0,00125	[0,00000; 0,00420]
4	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
5	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
6	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
7	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
8	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
	f estimado	0,048583	[0,00000; 0,16891]	0,08628	[0,00000; 0,23701]	0,17951	[0,00000; 0,37278]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,3	[IC 95%]	Porp.est.,f verd.= 0,4	[IC 95%]
0	0,2	0,20942	[0,17134; 0,24750]	0,21180	[0,15266; 0,27093]
1	0,6	0,58163	[0,50387; 0,65938]	0,57427	[0,45552; 0,69302]
2	0,2	0,20824	[0,16622; 0,25026]	0,21376	[0,15324; 0,27428]
3	0,0	0,00071	[0,00000; 0,00290]	0,00018	[0,00000; 0,00106]
4	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
5	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
6	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
7	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
8	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
	f estimado	0,26637	[0,04253; 0,49021]	0,36834	[0,11791; 0,61876]

* Prop. Verd.: Proporção alélica conhecida, utilizada para gerar as amostras; Porp.est.: proporção estimada através dos cálculos; [IC 95%]: Intervalo com 95% de Confiança; f verd.: coeficiente de endogamia conhecido, utilizado para gerar as amostras; f estimado: coeficiente estimado através dos cálculos.

Tabela 4: Proporções Alélicas e Coeficiente de Endogamia estimados, valor esperado e Intervalo de Confiança.

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,05	[IC 95%]	Porp.est.,f verd.= 0,1	[IC 95%]	Porp.est.,f verd.= 0,2	[IC 95%]
0	0,00	0,00807	[0,00000; 0,02756]	0,00783	[0,00000; 0,02551]	0,00894	[0,00000; 0,03293]
1	0,05	0,03549	[0,00000; 0,07619]	0,03735	[0,00000; 0,07782]	0,04205	[0,00000; 0,08779]
2	0,05	0,06317	[0,00307; 0,12327]	0,06112	[0,00000; 0,12665]	0,06254	[0,00000; 0,14237]
3	0,10	0,17317	[0,00000; 0,36011]	0,16563	[0,00000; 0,35098]	0,15173	[0,00000; 0,31855]
4	0,60	0,44226	[0,02572; 0,85880]	0,45457	[0,03759; 0,87155]	0,46907	[0,00000; 0,86922]
5	0,1	0,17316	[0,00000; 0,36623]	0,17036	[0,00000; 0,36078]	0,15507	[0,00000; 0,33001]
6	0,05	0,06204	[0,00336; 0,12071]	0,06012	[0,00000; 0,12202]	0,06236	[0,00000; 0,13825]
7	0,05	0,03331	[0,00000; 0,07310]	0,03334	[0,00000; 0,07187]	0,03798	[0,00000; 0,08389]
8	0,00	0,00652	[0,00000; 0,02452]	0,00586	[0,00000; 0,02107]	0,00633	[0,00000; 0,02579]
	f estimado	0,03502	[0,00000; 0,19410]	0,05859	[0,00000; 0,26130]	0,12928	[0,00000; 0,42511]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,3	[IC 95%]	Porp.est.,f verd.= 0,4	[IC 95%]
0	0,00	0,00894	[0,00000; 0,03728]	0,00701	[0,00000; 0,03046]
1	0,05	0,04691	[0,00000; 0,09855]	0,05064	[0,00000; 0,10296]
2	0,05	0,06836	[0,00000; 0,16006]	0,07152	[0,00000; 0,17308]
3	0,10	0,14114	[0,00000; 0,29646]	0,13037	[0,00000; 0,26363]
4	0,60	0,47290	[0,00000; 0,86296]	0,48290	[0,00000; 0,84406]
5	0,1	0,14379	[0,00000; 0,30226]	0,13253	[0,00000; 0,27015]
6	0,05	0,06404	[0,00000; 0,15358]	0,07020	[0,00000; 0,16951]
7	0,05	0,04290	[0,00000; 0,09613]	0,04565	[0,00000; 0,10029]
8	0,00	0,00721	[0,00000; 0,03127]	0,00568	[0,00000; 0,02573]
	f estimado	0,20061	[0,00000; 0,55620]	0,29874	[0,00000; 0,71412]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,05	[IC 95%]	Porp.est.,f verd.= 0,1	[IC 95%]	Porp.est.,f verd.= 0,2	[IC 95%]
0	0,3	0,28713	[0,28056; 0,29369]	0,29496	[0,28353; 0,30638]	0,30609	[0,28831; 0,32388]
1	0,4	0,42505	[0,40838; 0,44173]	0,41098	[0,38400; 0,43796]	0,39151	[0,35117; 0,43184]
2	0,3	0,28642	[0,27266; 0,30019]	0,29251	[0,27270; 0,31232]	0,30106	[0,27329; 0,32883]
3	0,0	0,00140	[0,00000; 0,00503]	0,00155	[0,00000; 0,00579]	0,00134	[0,00000; 0,00645]
4	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
5	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
6	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
7	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
8	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
	f estimado	0,07150	[0,00000; 0,23972]	0,10221	[0,00000; 0,29497]	0,17170	[0,00000; 0,40610]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,3	[IC 95%]	Porp.est.,f verd.= 0,4	[IC 95%]
0	0,3	0,30849	[0,28490; 0,33208]	0,30247	[0,27713; 0,32782]
1	0,4	0,38492	[0,33397; 0,43587]	0,38438	[0,33168; 0,43708]
2	0,3	0,30583	[0,27449; 0,33717]	0,31282	[0,28344; 0,34220]
3	0,0	0,00076	[0,00000; 0,00472]	0,00033	[0,00000; 0,00235]
4	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
5	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
6	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
7	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
8	0,0	0,00000	[0,00000; 0,00000]	0,00000	[0,00000; 0,00000]
	f estimado	0,26364	[0,00000; 0,53580]	0,36072	[0,05970; 0,66173]

* Prop. Verd.: Proporção alélica conhecida, utilizada para gerar as amostras; Porp.est.: proporção estimada através dos cálculos; [IC 95%]: Intervalo com 95% de Confiança; f verd.: coeficiente de endogamia conhecido, utilizado para gerar as amostras; f estimado: coeficiente estimado através dos cálculos.

Tabela 5: Proporções Alélicas e Coeficiente de Endogamia estimados, valor esperado e Intervalo de Confiança.

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,05	[IC 95%]	Porp.est.,f verd.= 0,1	[IC 95%]	Porp.est.,f verd.= 0,2	[IC 95%]
0	0,01	0,02717	[0,00000; 0,12428]	0,02864	[0,00000; 0,13154]	0,03087	[0,00000; 0,13896]
1	0,01	0,03333	[0,00000; 0,14768]	0,03248	[0,00000; 0,14279]	0,03380	[0,00000; 0,14386]
2	0,01	0,03307	[0,00000; 0,15328]	0,03064	[0,00000; 0,14352]	0,03763	[0,00000; 0,17213]
3	0,01	0,03903	[0,00000; 0,17152]	0,03861	[0,00000; 0,16426]	0,03600	[0,00000; 0,15410]
4	0,01	0,04718	[0,00000; 0,19029]	0,05116	[0,00000; 0,20754]	0,04932	[0,00000; 0,20232]
5	0,01	0,04880	[0,00000; 0,19295]	0,04874	[0,00000; 0,19558]	0,04513	[0,00000; 0,17916]
6	0,02	0,04614	[0,00000; 0,19394]	0,04417	[0,00000; 0,18752]	0,04211	[0,00000; 0,17709]
7	0,02	0,05944	[0,00000; 0,22024]	0,06025	[0,00000; 0,22737]	0,05568	[0,00000; 0,20933]
8	0,9	0,43021	[0,00000; 1,00000]	0,42857	[0,00000; 1,00000]	0,43405	[0,00000; 1,00000]
	f estimado	0,17217	[0,00000; 0,69289]	0,18816	[0,00000; 0,72492]	0,23368	[0,00000; 0,81062]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,3	[IC 95%]	Porp.est.,f verd.= 0,4	[IC 95%]
0	0,01	0,02894	[0,00000; 0,12571]	0,03307	[0,00000; 0,14001]
1	0,01	0,03436	[0,00000; 0,14632]	0,03261	[0,00000; 0,13718]
2	0,01	0,03338	[0,00000; 0,14871]	0,03701	[0,00000; 0,16681]
3	0,01	0,03689	[0,00000; 0,15859]	0,03408	[0,00000; 0,14659]
4	0,01	0,04092	[0,00000; 0,16398]	0,03980	[0,00000; 0,16212]
5	0,01	0,04200	[0,00000; 0,17153]	0,04154	[0,00000; 0,16973]
6	0,02	0,04056	[0,00000; 0,17215]	0,04068	[0,00000; 0,17082]
7	0,02	0,05543	[0,00000; 0,20846]	0,04800	[0,00000; 0,18117]
8	0,9	0,46964	[0,00000; 1,00000]	0,47831	[0,00000; 1,00000]
	f estimado	0,27749	[0,00000; 0,88693]	0,30859	[0,00000; 0,93814]

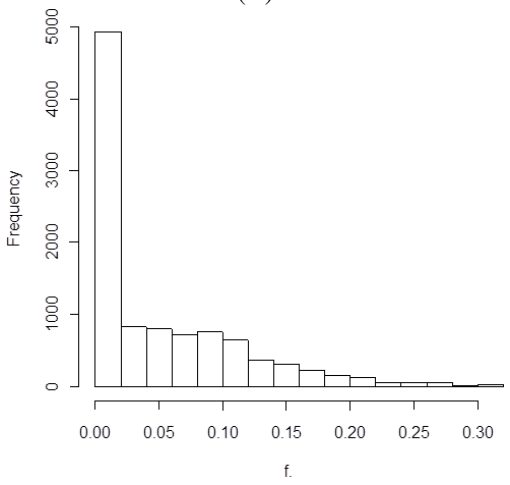
Alelo	Prop. Verd.	Porp.est.,f verd.= 0,05	[IC 95%]	Porp.est.,f verd.= 0,1	[IC 95%]	Porp.est.,f verd.= 0,2	[IC 95%]
0	0,000	0,03495	[0,00000; 0,15910]	0,03342	[0,00000; 0,15267]	0,03933	[0,00000; 0,17683]
1	0,080	0,09168	[0,03458; 0,14878]	0,09731	[0,01434; 0,18029]	0,10845	[0,00000; 0,23587]
2	0,770	0,68696	[0,36475; 1,00000]	0,67723	[0,33026; 1,00000]	0,64648	[0,21399; 1,00000]
3	0,110	0,12701	[0,05353; 0,20050]	0,13378	[0,02924; 0,23833]	0,14005	[0,00000; 0,29228]
4	0,029	0,05708	[0,00000; 0,19746]	0,05563	[0,00000; 0,19444]	0,06254	[0,00000; 0,22044]
5	0,000	0,00169	[0,00000; 0,00331]	0,00157	[0,00000; 0,00332]	0,00184	[0,00000; 0,00469]
	f estimado	0,03901	[0,00000; 0,21126]	0,06026	[0,00000; 0,27732]	0,10469	[0,00000; 0,39309]

Alelo	Prop. Verd.	Porp.est.,f verd.= 0,3	[IC 95%]	Porp.est.,f verd.= 0,4	[IC 95%]
0	0,000	0,03475	[0,00000; 0,15882]	0,03380	[0,00000; 0,15475]
1	0,080	0,11595	[0,00000; 0,27515]	0,11538	[0,00000; 0,28257]
2	0,770	0,63989	[0,17975; 1,00000]	0,64244	[0,18955; 1,00000]
3	0,110	0,14823	[0,00000; 0,33479]	0,14673	[0,00000; 0,34013]
4	0,029	0,05825	[0,00000; 0,20325]	0,05901	[0,00000; 0,19975]
5	0,000	0,00177	[0,00000; 0,00489]	0,00183	[0,00000; 0,00573]
	f estimado	0,18097	[0,00000; 0,55335]	0,26009	[0,00000; 0,69883]

* Prop. Verd.: Proporção alélica conhecida, utilizada para gerar as amostras; Porp.est.: proporção estimada através dos cálculos; [IC 95%]: Intervalo com 95% de Confiança; f verd.: coeficiente de endogamia conhecido, utilizado para gerar as amostras; f estimado: coeficiente estimado através dos cálculos.

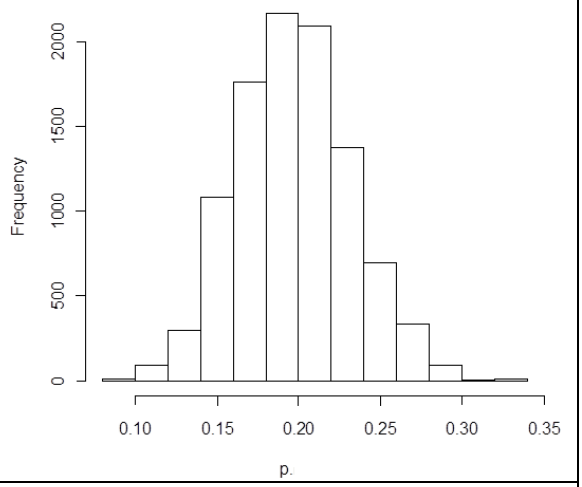
f verdadeiro = 0,05
f estimado = 0,048583

(A)



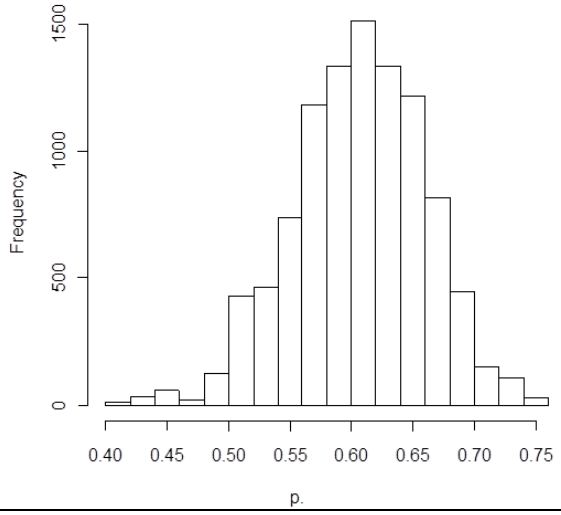
Proporção verdadeira para o alelo 0 = 0,2
Proporção estimada para o alelo 0 = 0,19705

(B)



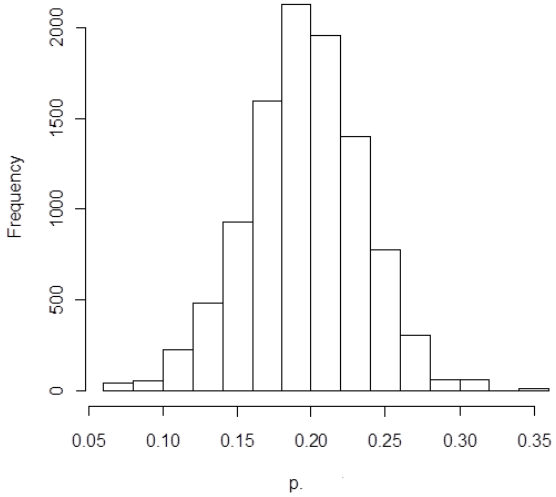
Proporção verdadeira para o alelo 1 = 0,6
Proporção estimada para o alelo 1 = 0,58163

(C)



Proporção verdadeira para o alelo 2 = 0,2
Proporção estimada para o alelo 2 = 0,19529

(D)



Proporção verdadeira para o alelo 3 = 0,0
Proporção estimada para o alelo 3 = 0,00171

(E)

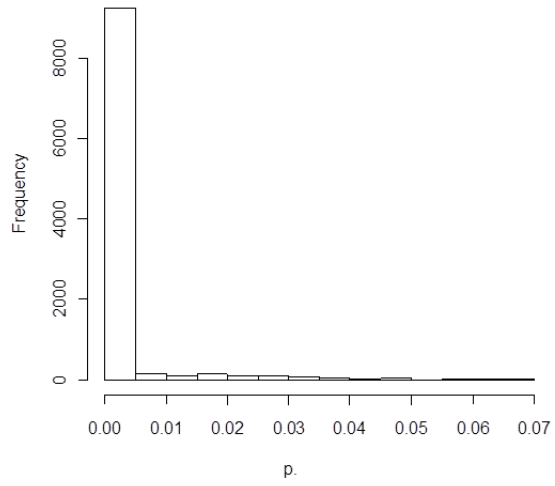
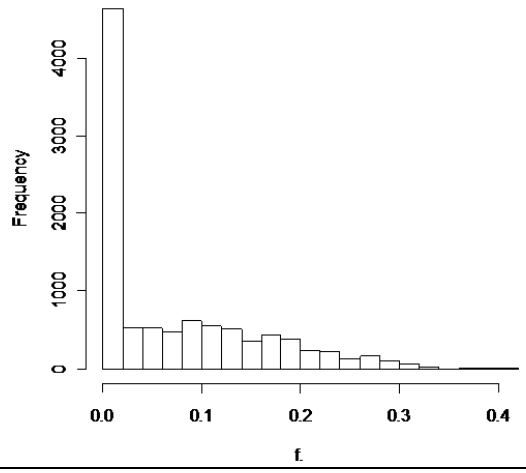


Figura 5: Estimativas para o vetor de proporções (0,2; 0,6; 0,2; 0; 0; 0; 0; 0; 0).

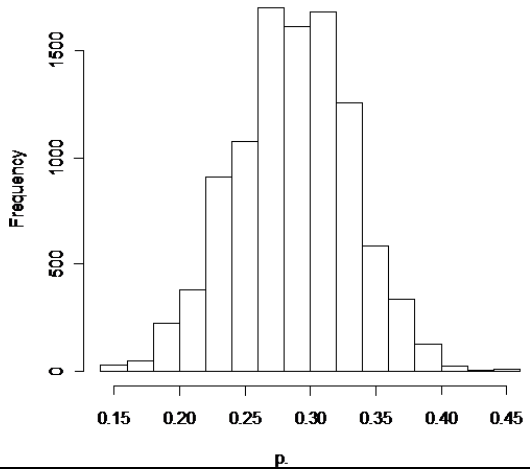
f verdadeiro = 0,05
f estimado = 0,07150

(A)



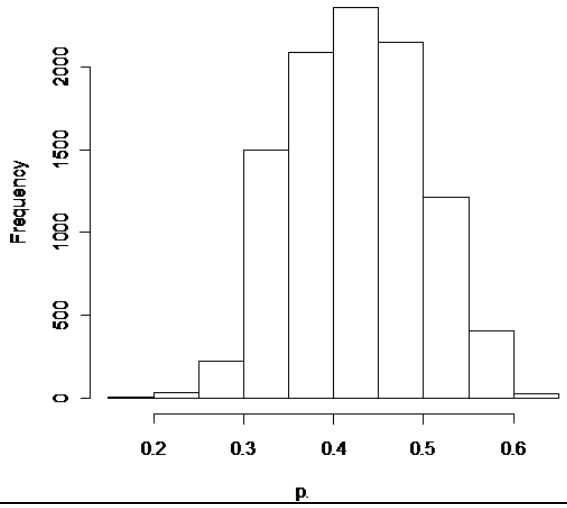
Proporção verdadeira para o alelo 0 = 0,3
Proporção estimada para o alelo 0 = 0,28713

(B)



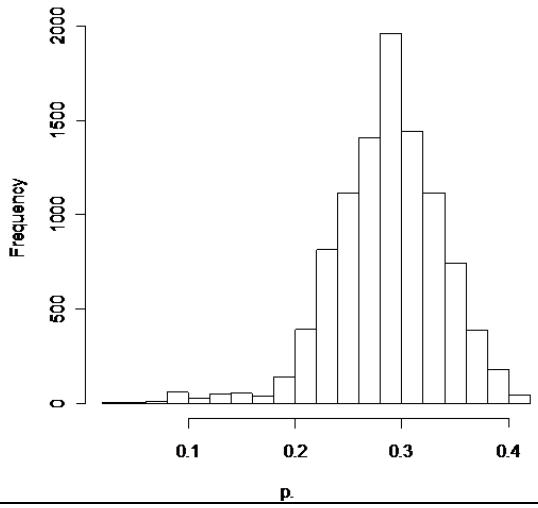
Proporção verdadeira para o alelo 1 = 0,4
Proporção estimada para o alelo 1 = 0,42505

(C)



Proporção verdadeira para o alelo 2 = 0,3
Proporção estimada para o alelo 2 = 0,28642

(D)



Proporção verdadeira para o alelo 3 = 0,0
Proporção estimada para o alelo 3 = 0,00140

(E)

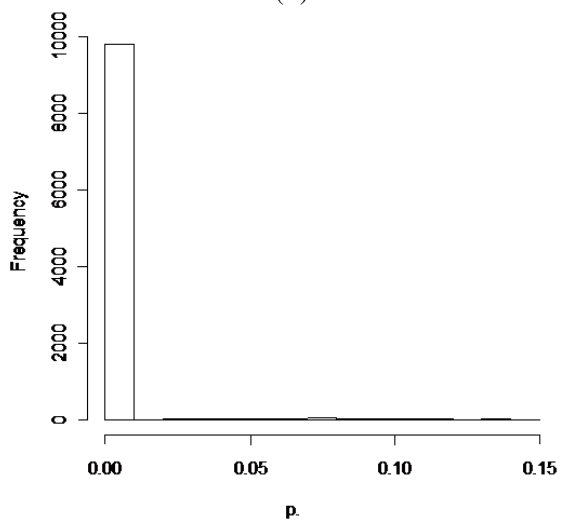
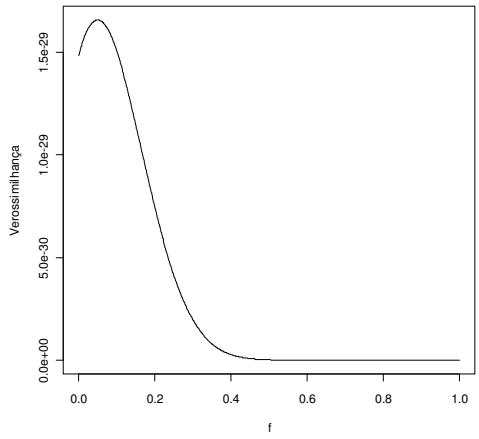


Figura 6: Estimativas para o vetor de proporções (0,3; 0,4; 0,3; 0; 0; 0; 0; 0; 0).

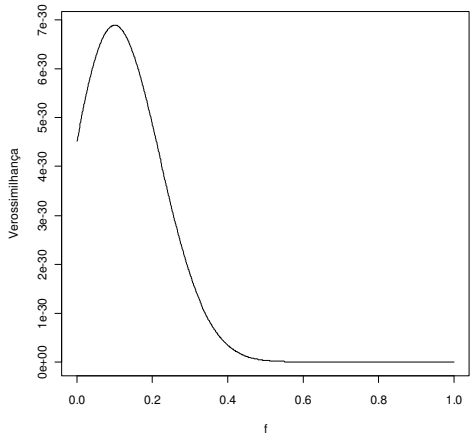
Tabela 6: Proporções alélicas e Coeficiente de Endogamia estimados.

Alelo	Prop. Verd.	Prop. Est., f=0,05	Prop. Est., f=0,1	Prop. Est., f=0,2	Prop. Est., f=0,3	Prop. Est., f=0,4
0	0,2	0,2000019	0,2000003	0,1999998	0,1999996	0,1999990
1	0,6	0,5999961	0,5999993	0,6000004	0,6000008	0,6000021
2	0,2	0,2000019	0,2000003	0,1999998	0,1999996	0,1999990
3	0	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
4	0	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
5	0	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
6	0	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
7	0	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
8	0	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
9	0	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
	f estimado	0,049983	0,099997	0,200002	0,300005	0,4000148
0	0,1	0,1000223	0,100009	0,0999972	0,0999963	0,0999986
1	0,1	0,0999803	0,0999930	0,1000013	0,1000008	0,1000001
2	0,1	0,1000168	0,1000079	0,0999984	0,0999994	0,1000001
3	0,1	0,0999737	0,0999867	0,1000040	0,1000044	0,1000015
4	0,2	0,2000139	0,2000067	0,1999981	0,199998	0,1999994
5	0,1	0,0999737	0,0999867	0,1000040	0,1000044	0,1000015
6	0,1	0,1000168	0,1000079	0,0999984	0,0999994	0,1000001
7	0,1	0,0999803	0,099993	0,1000013	0,1000008	0,1000001
8	0,1	0,1000223	0,1000090	0,0999972	0,0999963	0,0999986
	f estimado	0,0499264	0,0999642	0,2000148	0,3000245	0,4000109
0	0,01	0,0100002	0,0099999	0,0099998	0,0100002	0,0099998
1	0,01	0,0100000	0,0100000	0,0100000	0,0100000	0,0100000
2	0,01	0,0100001	0,0099999	0,0099998	0,0100002	0,0099998
3	0,01	0,0099999	0,0100000	0,0100001	0,0100000	0,0100000
4	0,01	0,0100000	0,0100000	0,0100001	0,0099999	0,0100003
5	0,01	0,0099997	0,0100001	0,0100004	0,0099996	0,0100004
6	0,02	0,0199998	0,0200001	0,0200001	0,0200000	0,0199997
7	0,02	0,0199995	0,0200002	0,0200007	0,0199994	0,2000070
8	0,90	0,9000008	0,8999996	0,8999991	0,9000008	0,8999992
	f estimado	0,0499770	0,1000104	0,2000245	0,2999818	0,4000160
0	0,00	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
1	0,05	0,0500000	0,0500005	0,0499999	0,0499998	0,0499996
2	0,05	0,0500000	0,0500012	0,0499991	0,0499985	0,0499977
3	0,10	0,0999999	0,0999971	0,1000017	0,1000027	0,1000043
4	0,60	0,6000001	0,6000024	0,5999987	0,5999979	0,5999967
5	0,10	0,0999999	0,0999971	0,1000017	0,10000270	0,1000043
6	0,05	0,0500000	0,0500012	0,0499991	0,0499985	0,0499977
7	0,05	0,0500000	0,0500005	0,0499999	0,0499998	0,0499996
8	0,00	0,0000000	0,0000000	0,0000000	0,000000	0,000000
	f estimado	0,04999888	0,09997962	0,2000095	0,3000149	0,4000266
0	0,000	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000
1	0,080	0,0808897	0,0808900	0,0499999	0,0808896	0,0808895
2	0,770	0,7785639	0,7785665	0,0499991	0,7785630	0,7785627
3	0,110	0,1112243	0,1112181	0,1000017	0,1112263	0,1112272
4	0,029	0,0293221	0,0293254	0,5999987	0,0293211	0,0293206
5	0,000	0,0000000	0,0000000	0,1000017	0,0000000	0,0000000
	f estimado	0,05053314	0,1009682	0,2000095	0,3023457	0,4026848

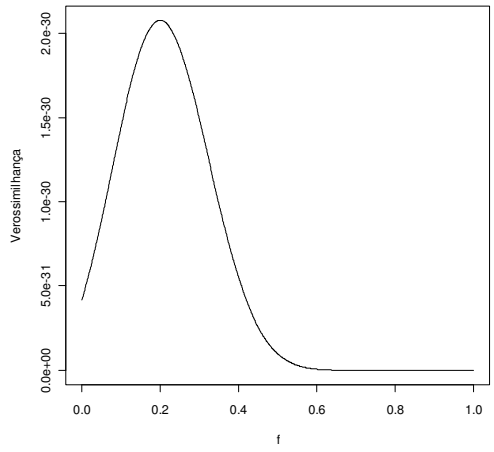
F verdadeiro = 0,05
F estimado = 0,04998433
(A)



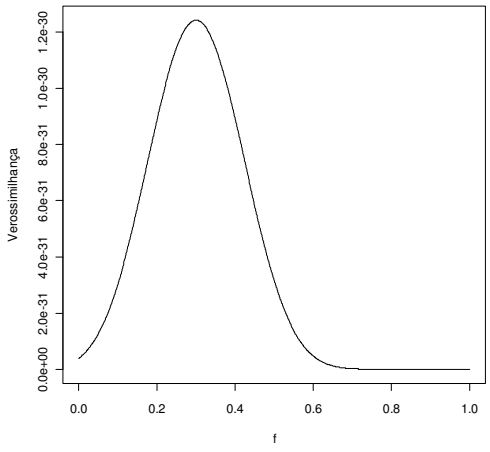
F verdadeiro = 0,1
F estimado = 0,09999705
(B)



F verdadeiro = 0,2
F estimado = 0,2000021
(C)



F verdadeiro = 0,3
F estimado = 0,3000048
(D)



F verdadeiro = 0,4
F estimado = 0,4000144
(E)

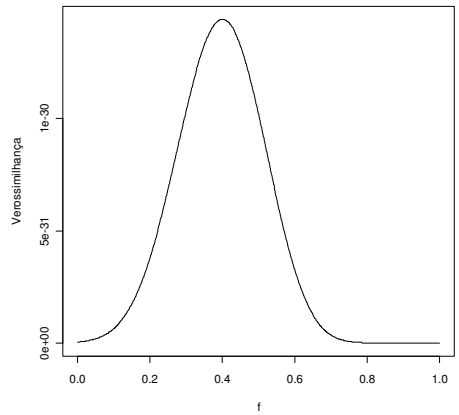


Figura 7: Gráficos para a função de verossimilhança considerando-se as proporções alélicas (0.2, 0.6, 0.2, 0, 0, 0, 0, 0, 0).

Na Figura 7, estão apresentados gráficos da função de verossimilhança considerando o vetor de proporções alélicas (0.2, 0.6, 0.2, 0, 0, 0, 0, 0, 0) e tamanho amostral igual a 100. Supondo-se as proporções e o número de indivíduos para cada número de cópia, n_j, conhecidos, pode-se perceber que o ponto de máximo é aproximadamente o valor do parâmetro verdadeiro. No gráfico (A) se encontra a curva de máxima verossimilhança para as proporções alélicas e Coeficiente de Endogamia 0,05. No gráfico (B), o Coeficiente de Endogamia considerado é 0,1, no gráfico (C) o Coeficiente de Endogamia considerado é 0,2, no gráfico (D) o Coeficiente de Endogamia considerado é 0,3 e no gráfico (E) o Coeficiente de Endogamia é 0,4.

Os gráficos da função de verossimilhança para as proporções alélicas (0.3, 0.4, 0.3, 0, 0, 0, 0, 0, 0) estão no apêndice 3. Os gráficos estão expostos na mesma sequência que os gráficos da Figura 7, ou seja, em cada gráfico está a verossimilhança com o respectivo Coeficiente de Endogamia estimado.

Nas Tabelas 3, 4 e 5 podemos analisar, através das simulações de Monte Carlo, que algumas estimativas não são muito precisas, isso pode ser devido ao comportamento da distribuição das proporções alélicas, ou devido ao número de réplicas utilizadas na simulação. As estimativas são melhores para as proporções alélicas (0.3, 0.4, 0.3, 0, 0, 0, 0, 0, 0) e (0.2, 0.6, 0.2, 0, 0, 0, 0, 0, 0), sendo mais precisas. Seria bom analisarmos um número maior de réplicas, porém devido ao tempo foram feitas apenas 1000 para cada vetor.

As estimativas não sofrem alterações com o tamanho da amostra utilizado. Realizamos algumas simulações com tamanho de amostra igual a 50, 100 e 1000, e os valores não sofriam alterações, isso se deve ao fato de utilizarmos uma distribuição Multinomial onde os parâmetros são os P_j´s e sendo P_j estimado por N_j/N, ou seja, a proporção sempre permanecerá a mesma.

Ao analisarmos a Figura 7 e o apêndice 3 podemos ver que a verossimilhança é bem comportada, e as estimativas de máxima verossimilhança são aproximadamente iguais ao verdadeiro valor. Na Tabela 6 podemos verificar que as proporções alélicas estimadas são aproximadamente iguais ao valor para as proporções verdadeiras, considerando-se apenas uma amostra gerada a partir do valor esperado para as proporções alélicas.

Tendo em vista que o método utilizado nas simulações apresentou um bom desempenho, o utilizamos em dados reais, e os resultados se encontram na próxima seção.

4.2. Aplicações a dados reais

Pelo fato das populações nativas serem sub-representadas nos estudos genéticos, o objetivo do Laboratório de Diversidade Genética Humana da Universidade Federal de Minas Gerais, foi estudar a variação do número de cópias de populações nativas americanas. Para obter dados sobre populações nativas amostras de DNA humano foram extraídas do sangue periférico utilizando o *Gentra Puregene Blood Kit* (Qiagen®), coletado em indivíduos Ashaninka (n=288), Monte Carmelo (n=24), Shimaa (n=89) e Quéchua (n=120).

As amostras originárias do grupo étnico Ashaninka foram coletadas no Departamento de Junin/Peru, as de Monte Carmelo e Shimaa (do grupo étnico Machiguenga) foram coletadas no Departamento de Cusco/Peru. Essas três populações estão localizadas na região amazônica peruana chamada de "Selva Alta", fazem parte de comunidades rurais pequenas e isoladas. Todas as coletas dos dados foram realizadas em colaboração com o grupo do Dr. Robert Gilman, da *Universidad Peruana Cayetano Heredia*, Lima/Peru. Mais detalhes sobre o tamanho amostral e a localização das populações analisadas estão apresentados na Tabela 7 (Zuccherato, 2012).

Tabela 7: Descrição das populações nativas Peruanas utilizadas no presente estudo.

Etnia	Comunidade	Tamanho Amostral	Distrito	Província	Departamento
Ashaninka	Cushireni	41	Rio Tambo	Satipo	Junin
	Mayapo	78			
	Charahuaja	70			
	Capitiri	35			
	Ivotsote	64			
Machiguenga	Monte Carmelo	24	Echarate	La convencion	Cusco
	Shimma	89			
Quéchua	Jayu Jayu	10	Acora	Puno	Puno
	Ccopamaya	51			
	Laraqueri	8	Pichacan		
	Pichacani	37			
	Camicachi	14	Ilave	El Collao	

Fonte: Zuccherato (2012)

O número de cópias do gene *CCL3L1* foi medido na amostra composta pelos Ashaninka (n=142, pois nos demais indivíduos foram medidos somente outros genes), Shimaa (n=89), Monte Carmelo (n=24), Quechua (n=120). Para avaliar o número de cópias do gene

CCL3L1 também contamos com uma amostra da população Europeia (n=4266), tipados por Field *et al.* (2009).

O gene *CCL3L1* foi escolhido para este trabalho por ser foco de muitas pesquisas científicas devido à variação no número de cópias deste gene ser correlacionada com fisiologia de infecções por HIV, malária, doenças autoimunes, diabetes Tipo 1 e Lúpus. Esta variação se apresenta de forma bem estruturada entre as populações mundiais, sendo que as populações africanas exibem um maior número de cópias (6 cópias em média) e as populações europeias exibem um número menor, isto é, 2 cópias em média (Zuccherato, 2012).

A distribuição de frequência do número de cópias (fenótipo) para as quatro populações da amostra em estudo se encontra na Tabela 7. O número de cópias máximo encontrado foi sete, nas populações Ashaninka e Quechua. Podemos constatar que a distribuição de frequência dos Europeus difere das demais populações, em geral eles possuem número de cópias inferiores aos demais grupos (Tabela 8).

Tabela 8: Frequência estimada do número de indivíduos com determinado número de cópias em cada população.

	Frequência do número de cópias das populações				
Número de Cópias	Ashaninka (%)	Shimaa (%)	Monte Carmelo (%)	Quechua (%)	Europeus (%)
0	0,000	0,000	0,000	0,000	1,828
1	0,000	0,000	0,000	0,000	19,433
2	23,239	7,865	29,167	10,833	58,837
3	37,324	42,697	20,833	36,667	17,722
4	22,535	38,202	37,500	40,000	1,946
5	11,268	11,236	12,500	10,000	0,211
6	4,225	0,000	0,000	1,667	0,023
7	1,408	0,000	0,000	0,833	0,000

As proporções alélicas obtidas através da equação (12), após a convergência e considerando-se Coeficiente de Endogamia f=0, para as cinco populações encontram-se na Tabela 9. E, segundo Zuccherato (2012) as populações Shimaa, Monte Carmelo e Ashaninka, quando comparadas com dados da população europeia, apresentam um aumento na duplicação de número de cópias, o que pode levar a um aumento na ativação do processo imunológico.

Aplicamos o Teste da Razão de Verossimilhança (TRV) para as proporções alélicas, fazendo as comparações duas-a-duas entre as populações, considerando-se a distribuição Multinomial com vetor de parâmetros sendo as proporções fenotípicas (P_j), sendo estas uma

combinação das proporções alélicas de do Coeficiente de Endogamia. Para a hipótese nula consideramos que as populações podem ser consideradas como uma única população, e na hipótese alternativa que as populações não possam ser consideradas iguais. As hipóteses e um exemplo do cálculo do TRV utilizado se encontram no apêndice 2.

O Teste da Razão de Verossimilhança na comparação entre europeus e Quéchuas apresentou um valor igual a aproximadamente 392,11, comparando com $\chi^2_9 = 16{,}919$ (9 graus de liberdade, pois são 8 possíveis números de cópias) rejeita-se a hipótese de que as populações possam ser consideradas iguais. Na comparação entre as populações europeia e Monte Carmelo o valor da estatística de teste foi aproximadamente 61,51, rejeitando a hipótese de que ambas sejam da mesma população. Na comparação entre europeus e Shimaa, a estatística de teste foi aproximadamente 298,11, também rejeitando a hipótese de igualdade entre as populações. E, na comparação entre europeus e Ashaninka o valor da estatística de teste obtido foi aproximadamente 336,47, rejeitando-se a hipótese de que ambas as amostras sejam da mesma população.

Tabela 9: Proporção alélica estimada para cada população, supondo Coeficiente de Endogamia *f*=0.

Proporção alélica estimada para as populações amostradas.					
Alelo	Ashaninka	Shimaa	Monte Carmelo	Quechua	Europeus
h_0	0,00000	0,00000	0,00000	0,00000	0,13162
h_1	0,48501	0,31490	0,54514	0,32204	0,74574
h_2	0,37609	0,60615	0,24306	0,58469	0,11829
h_3	0,09559	0,07895	0,21180	0,08234	0,00348
h_4	0,03984	0,00000	0,00000	0,00558	0,00087
h_5	0,00348	0,00000	0,00000	0,00535	0,00000
h_6	0,00000	0,00000	0,00000	0,00000	0,00000
h_7	0,00000	0,00000	0,00000	0,00000	0,00000

Para apresentar as estimativas da proporção genotípica populacional, obtida por $P(h_k, h_l)$, escolhemos os dados da população europeia, devido ao seu tamanho amostral. Estas estimativas podem ser analisadas na Tabela 10. Podemos ver que os genótipos mais prováveis são (h_1, h_1) com probabilidade 0,557 e (h_0, h_1) com probabilidade 0,194.

A proporção genotípica individual estimada, dado o número de cópias do gene do indivíduo, para a população europeia, se encontra na Tabela 11. Por exemplo: dado que o indivíduo possui 5 cópias, ele poderá ter o genótipo (h_1, h_4) com probabilidade 0,6117 ou o genótipo (h_2, h_3) com probabilidade 0,38828; se o indivíduo possui 6 cópias, ele poderá ter o genótipo (h_2, h_4) com probabilidade 0,9443 ou o genótipo (h_3, h_3) com probabilidade 0,0556.

Tabela 10: Probabilidade genotípica populacional para a pop. europeia, supondo Coeficiente de Endogamia *f*=0.

		Alelo							
		h_0	h_1	h_2	h_3	h_4	h_5	h_6	h_7
Alelo	h_0	0,018284	0,194327	0,031196	0,000917	0,000230	0,000000	0,000000	0,000000
	h_1		0,557177	0,176299	0,005208	0,001291	0,000000	0,000000	0,000000
	h_2			0,014019	0,000819	0,000221	0,000000	0,000000	0,000000
	h_3				0,000013	0,000000	0,000000	0,000000	0,000000
	h_4					0,000000	0,000000	0,000000	0,000000
	h_5						0,000000	0,000000	0,000000
	h_6							0,000000	0,000000
	h_7								0,000000

Tabela 11: Probabilidade Genotípica Individual para a população Europeia.

Número de Cópias	Genótipo	Probabilidade, considerando f=0
0	(h_0, h_0)	1,00000
1	(h_1, h_0)	1,00000
2	(h_0, h_2)	0,05302
	(h_1, h_1)	0,94698
3	(h_0, h_3)	0,00517
	(h_1, h_2)	0,99483
4	(h_0, h_4)	0,01180
	(h_1, h_3)	0,26765
	(h_2, h_2)	0,72054
5	(h_0, h_5)	0,00000
	(h_1, h_4)	0,61172
	(h_2, h_3)	0,38828
6	(h_0, h_6)	0,00000
	(h_1, h_5)	0,00000
	(h_2, h_3)	0,94434
	(h_3, h_3)	0,05566
7	(h_0, h_7)	0,00000
	(h_1, h_6)	0,00000
	(h_2, h_5)	0,00000
	(h_3, h_4)	0,00000

A proporção genotípica de determinado filho pode ser estimada com mais precisão se condicionada no genótipo dos pais, sabendo-se que o número de cópias de ambos são conhecidos. Considerando-se um filho, com *j*=4 cópias, da população europeia, e com pai com *j*=3 cópias e a mãe com *j*=3 cópias, a probabilidade genotípica do filho está apresentada na Tabela 12. Pode-se notar que o filho terá o genótipo (h_2, h_2), com probabilidade 0,989708.

Comparando-se as probabilidades acima, do filho dado os pais, com as probabilidades de um indivíduo com *j*=4 cópias, sem levar em consideração os genótipos dos pais, pode-se notar que o indivíduo poderá ter o genótipo (h_0, h_4) com probabilidade 0,01180, o genótipo (h_1, h_3) com probabilidade 0,26765 e o genótipo (h_2, h_2) com probabilidade 0,72054 (Tabela 10). Porém, considerando-se a informação dos pais ele poderá ter o genótipo (h_1, h_3) com probabilidade 0,010292 e o genótipo (h_2, h_2) com probabilidade 0,989708 (Tabela 12).

Tabela 12: Probabilidade genotípica do filho da pop. europeia, dado o número de cópias e a probabilidade genotípica dos seus pais.

Genótipo Pai	Genótipo Mãe	Genótipo Filho	Prob. Pai	Prob. Mãe	Prob. Filho
(h_1, h_2)	(h_0, h_3)	(h_1, h_3)	0,9948274	0,0051726	0,005146
(h_0, h_3)	(h_1, h_2)	(h_3, h_1)	0,0051726	0,9948274	0,005146
(h_1, h_2)	(h_1, h_2)	(h_2, h_2)	0,9948274	0,9948274	0,989708

Incluindo-se o Coeficiente de Endogamia, *f*, nos cálculos das proporções alélicas e genotípicas, obtivemos as seguintes estimativas para o Coeficiente: na amostra da população Ashaninka o valor da estimativa para *f* foi aproximadamente $6{,}26\text{x}10^{-5}$, e para analisarmos a variabilidade das estimativas encontradas realizamos 300 reamostragens para o número de indivíduos em cada classe, n_j, através do método conhecido como bootstrap, onde encontramos 90% delas entre [4,79e-05; 0,0579]; na população europeia o valor da estimativa foi *f*=0,0072865, sendo 90% das estimativas encontradas entre [6,07e-05; 0.034]; na população Monte Carmelo a estimativa foi *f*=0,2825013, com intervalo [5,79e-05; 0,652]; na população Quéchua a estimativa foi *f*=0,0130045, sendo o intervalo com 90% delas entre [6,41e-05; 0,118]; e na população Shimma a estimativa foi $f=6{,}26\text{x}10^{-5}$, e o intervalo com 90% das estimativas encontras entre [4,89e-05; 6,61e-05].

A estatística do Teste da Razão de Verossimilhanças, para a população Monte Carmelo, entre a verossimilhança sob *f*=0 e a verossimilhança sob *f*=0,2825013 resulta em 5,338, comparando-se com $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = 2{,}42075$ (Self and Liang, 1987) rejeita-se a hipótese de que o Coeficiente de Endogamia seja zero; para a população europeia temos a estatística de teste do TRV igual a 7,861, comparando-se com $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = 2{,}42075$ também rejeita-se a hipótese de nulidade; e para a população Quéchua é igual a 2,516, comparando-se com $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = 2{,}42075$ também rejeitamos a hipótese de que *f* seja zero.

Nas comparações posteriores, vamos considerar a população Monte Carmelo, pois foi onde encontrou-se a maior estimativa para o Coeficiente de Endogamia. Podemos notar que levando-se em consideração o Coeficiente de Endogamia f≅0,28, a proporção dos alelos h_1 e h_3 diminui e a probabilidade do alelo h_2 aumenta (Tabela 13).

Tabela 13: Comparação da probabilidade alélica, com f=0 e com f≅0,28.

Alelos	Prob. Alélica com f=0	Prob. Alélica com f≅0,28
h_0	0,0000000	0,0000000
h_1	0,5451360	0,4248944
h_2	0,2430613	0,4835446
h_3	0,2118027	0,0915611
h_4	0,0000000	0,0000000
h_5	0,0000000	0,0000000

A proporção genotípica estimada, considerando-se o *f* de aproximadamente 0,28, pode ser analisada na Tabela 14. Podemos ver que os genótipos mais prováveis são (h_1,h_1) com probabilidade 0,2916, o genótipo (h_1, h_2) com probabilidade 0,2083 e o genótipo (h_2, h_2) com probabilidade 0,3168. Comparando com as probabilidades considerando-se f=0, podemos notar o genótipo (h_1, h_3) ocorre com uma probabilidade de 0,2986, e com f≅0,28 essa probabilidade é de 0,0581.

Tabela 14: Comparação das proporções genotípicas, considerando-se *f*=0 e *f*≅0,28.

		Alelo, f=0.					
		h_0	h_1	h_2	h_3	h_4	h_5
Alelo	h_0	0	0	0	0	0	0
	h_1		0,2916667	0,2083333	0,2986054	0	0
	h_2			0,0763946	0,1250000	0	0
	h_3				0	0	0
	h_4					0	0
	h_5						0
		Alelo, f≅0,28.					
		h_0	h_1	h_2	h_3	h_4	h_5
Alelo	h_0	0	0	0	0	0	0
	h_1		0,2916667	0,2083333	0,0581221	0	0
	h_2			0,3168779	0,1250000	0	0
	h_3				0	0	0
	h_4					0	0
	h_5						0

A proporção genotípica individual, dado o número de cópias do indivíduo, é apresentada na Tabela 15. Comparando-se os valores, podemos notar que houve uma diferença nas proporções dos genótipos (h_1, h_3) e (h_2, h_2).

Tabela 15: Comparação das proporções genotípicas individuais, com *f*=0 e *f*≅0,28.

Número de Cópias	Genótipo	Probabilidade, considerando f=0	Probabilidade, considerando f≅0,28
0	(h_0, h_0)	1,00000	1,00000
1	(h_1, h_0)	1,00000	1,00000
2	(h_0, h_2)	0,00000	0,00000
	(h_1, h_1)	1,00000	1,00000
3	(h_0, h_3)	0,00000	0,00000
	(h_1, h_2)	1,00000	1,00000
4	(h_0, h_4)	0,00000	0,00000
	(h_1, h_3)	0,79628	0,15499
	(h_2, h_2)	0,203719	0,84501
5	(h_0, h_5)	0,00000	0,00000
	(h_1, h_4)	0,00000	0,00000
	(h_2, h_3)	1,00000	1,00000

A probabilidade genotípica de um filho da população Monte Carmelo, dado o genótipo dos pais, dado que o pai tem 4 cópias, a mãe 3 cópias e o filho 3 cópias é apresentada na Tabela 15. Considerando-se que *f* seja aproximadamente 0,28 a probabilidade do filho ter o genótipo (h_1, h_2) é aproximadamente 0,2496 e o genótipo (h_2, h_1) com probabilidade 0,7503406, que são diferentes das probabilidades genotípicas estimadas com *f*=0 (Tabela 16).

Tabela 16: Comparação das probabilidades condicionais, com f=0 e f≅0,28.

Coeficiente de Endogamia, f=0					
Genótipo Pai	Genótipo Mãe	Genótipo Filho	Prob. Pai	Prob. Mãe	Prob. Filho
(h_0, h_4)	(h_0, h_3)	(h_0, h_3)	0,000000	0,000000	0,000000
(h_1, h_3)	(h_0, h_3)	(h_3, h_0)	0,796281	0,000000	0,000000
(h_1, h_3)	(h_1, h_2)	(h_1, h_2)	0,796281	1,000000	0,796281
(h_2, h_2)	(h_1, h_1)	(h_2, h_1)	0,203719	1,000000	0,203719
Coeficiente de Endogamia, f≅0,28					
Genótipo Pai	Genótipo Mãe	Genótipo Filho	Prob. Pai	Prob. Mãe	Prob. Filho
(h_0, h_4)	(h_0, h_3)	(h_0, h_3)	0,000000	0,000000	0,000000
(h_1, h_3)	(h_0, h_3)	(h_3, h_0)	0,2496594	0,0000000	0,0000000
(h_1, h_3)	(h_1, h_2)	(h_1, h_2)	0,2496594	1,0000000	0,2496594
(h_2, h_2)	(h_1, h_1)	(h_2, h_1)	0,7503406	1,0000000	0,7503406

5. Considerações Finais

Apesar da intensa pesquisa na área da Genética sobre CNVs, ainda há a necessidade de elaborar ferramentas estatísticas para se avaliar estes dados. Tanto em situações onde as proporções alélicas desconhecidas estão em Equilíbrio de Hardy-Weinberg (EHW) quanto em situações onde estão em desequilíbrio. Tendo em vista essa necessidade, propomos um programa, que chamamos de CNVice, capaz de estimar as proporções alélicas e genotípicas quando a população está em EHW e quando não está. Também propomos as estimativas individuais e estimativas condicionadas na informação dos pais.

Por meio da análise das simulações, podemos verificar que o método estima muito bem os parâmetros desconhecidos. Consegue captar o quanto a população se desvia do Equilíbrio de H-W, que é o parâmetro *f* (Coeficiente de Endogamia). Em todos os casos apresentados, o método conseguiu estimar a frequência alélica, resultando em valores bem próximos dos verdadeiros valores.

Ademais podemos constatar a utilidade das proporções condicionais, pois conhecendo-se o número de cópias dos pais, a estimação das proporções genotípicas do filho podem ser mais bem estimadas, restringindo-se às possibilidades genotípicas.

Considerando-se o gene CCL3L1 das populações Shimaa, Monte Carmelo e Ashaninka, quando comparadas com a população Europeia, apresentam uma diferença significativa na distribuição das proporções alélicas, ressaltando assim a diferença entre as populações nativas e a Europeia. Também podemos observar um moderado Coeficiente de Endogamia para a população Monte Carmelo, talvez devido ao fato de ser uma população nativa com baixo índice de miscigenação (Zuccherato, 2012).

Como sugestão de pesquisas futuras, propomos a inclusão da abordagem bayesiana nos cálculos, considerando-se os parâmetros desconhecidos (*f* e *p*) como variáveis aleatórias. Uma vez que os dados possuem distribuição Multinomial, podemos introduzir uma distribuição *à priori* Dirichlet para o vetor de parâmetros *p* e uma distribuição *à priori* Beta o Uniforme(0,1) para o parâmetro *f*. A abordagem bayesiana para cálculos com o Coeficiente de Endogamia foi introduzida por Reis *et al.* (2009), considerando um gene com dois alelos A e B e proporções p_A e $p_B=(1- p_A)$, ou seja, os parâmetros desconhecidos neste caso são apenas dois: o Coeficiente de Endogamia *f* e a proporção alélica p_A.

Apêndice 1

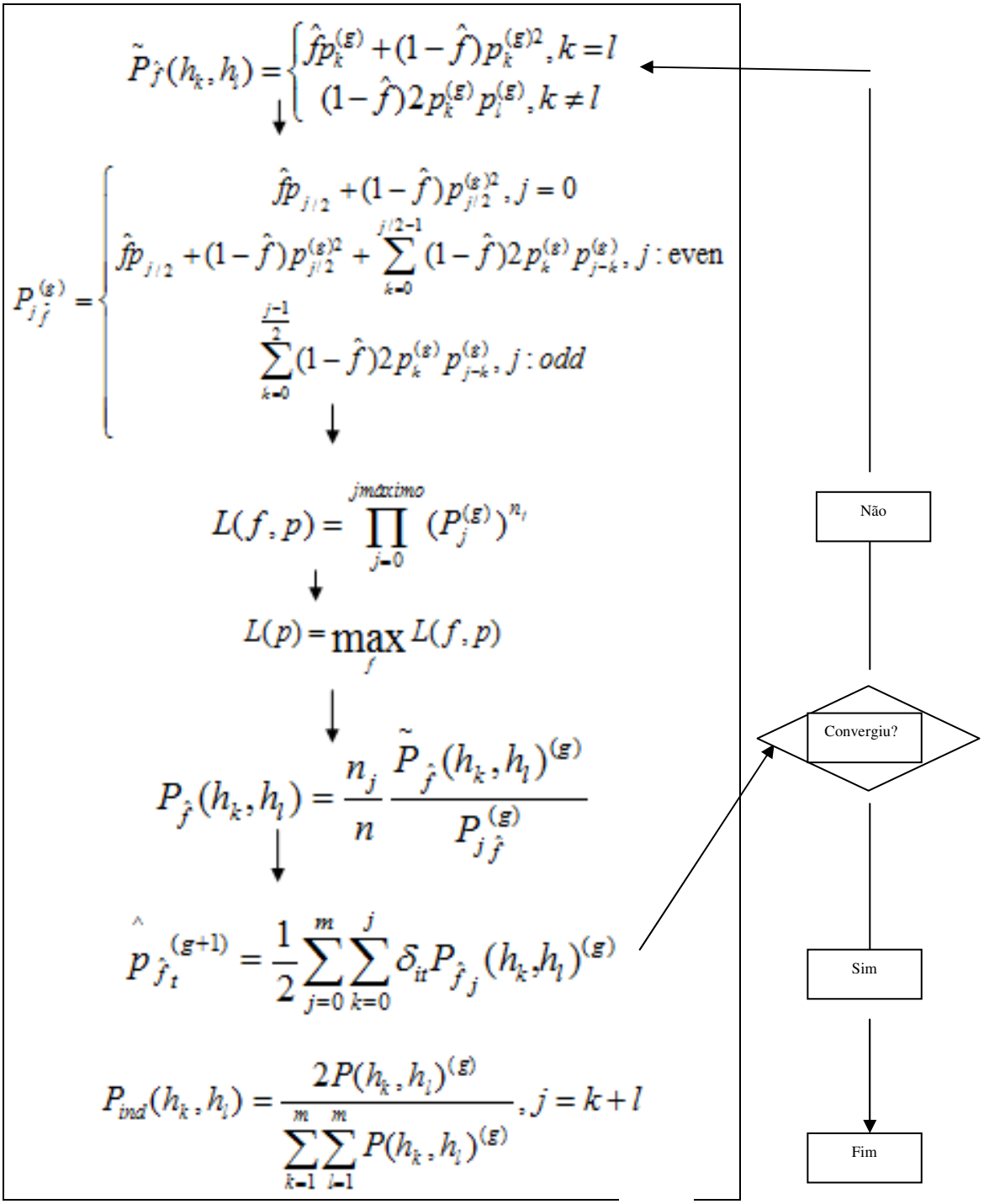


Figura: Fluxograma do código implementado no programa, CNVice.

Apêndice 2

Dado que os dados de CNVs possuem distribuição Multinomial, dada por:

$$f(P_0, P_1,..., P_m \mid n_1, n_2,..., n_m) = \frac{n!}{n_0! n_1! ... n_m!} \times P_0^{n_1} \times P_1^{n_2} \times ... \times P_m^{n_m} \ .$$

O objetivo é analisarmos se as populações podem ser consideradas como provenientes de uma mesma população ou não. Então consideramos como hipótese nula a igualdade entre as proporções fenotípicas de ambas as populações, ou seja:

$$\begin{cases} H_0 : P_{j_i}^X = P_{j_i}^Y, i \in [0, jmáximo] \\ Ha : P_{j_i}^X \neq P_{j_i}^Y, i \in [0, jmáximo] \end{cases}$$

Supondo que ambas as populações, X e Y, sejam iguais, temos que a população W terá o número de indivíduos com o fenótipo *j*, dado pela soma $n_{i,w} = n_{i,x} + n_{i,y}$.

Sendo a função de verossimilhança definida como

$$L(P_0, P_1,..., P_m \mid n_0, n_1,..., n_m) = \prod_{i=1}^{m} f(n_i \mid P_0, P_1,..., P_m) \, .$$

A estatística do Teste da Razão de Verossimilhança para testar H_0 versus H_a é (Casella, 2010)

$$\lambda(x) = \frac{\sup_{\Theta_0} L(P_0, P_1,..., P_m \mid n_0, n_1,..., n_m)}{\sup_{\Theta} L(P_0, P_1,..., P_m \mid n_0, n_1,..., n_m)} \, .$$

A região de rejeição será obtida por

$$D = -2 \ln \lambda(x) \leq \chi_{m+1}^2$$

Logo, a estatística de teste para testarmos se duas populações provém de uma mesma população será dada por:

$$\lambda(x) = \frac{\hat{P}_0^{n_{1,x}} \times \hat{P}_1^{n_{2,x}} \times ... \times \hat{P}_{jmáx}^{n_{jmáx,x}} \times \hat{P}_0^{n_{1,y}} \times \hat{P}_1^{n_{2,y}} \times ... \times \hat{P}_{jmáx}^{n_{jmáx,y}}}{\hat{P}_0^{n_{1,x}} \times \hat{P}_1^{n_{2,x}} \times ... \times \hat{P}_{jmáx}^{n_{jmáx,x}} \times \hat{P}_0^{n_{1,y}} \times \hat{P}_1^{n_{2,y}} \times ... \times \hat{P}_{jmáx}^{n_{jmáx,y}}}$$

onde os $\hat{P}_i$ são obtidos através da maximização de $L(P_0, P_1, ..., P_m \mid n_0, n_1, ..., n_m)$.

Então, sob H_0 temos:

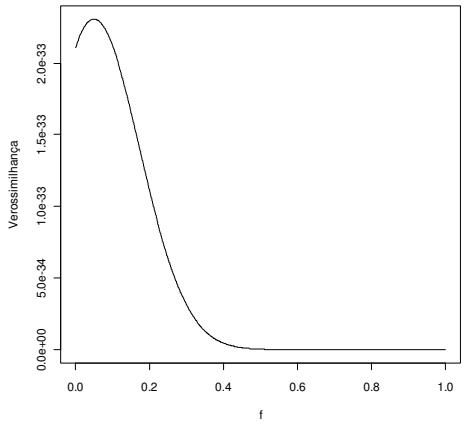
$$\lambda(x) = \frac{\hat{P}_1^{n_{1,x}+n_{1,y}} \times \hat{P}_2^{n_{2,x}+n_{2,y}} \times ... \times \hat{P}_{jmáx}^{n_{jmáx,x}+n_{jmáx,y}}}{\hat{P}_1^{n_{1,x}} \times \hat{P}_2^{n_{2,x}} \times ... \times \hat{P}_{jmáx}^{n_{jmáx,x}} \times \hat{P}_1^{n_{1,y}} \times \hat{P}_2^{n_{2,y}} \times ... \times \hat{P}_{jmáx}^{n_{jmáx,y}}}.$$

Exemplo: Na comparação entre europeus e Shimaa, encontramos as proporções fenotípicas, $\hat{P}_i$, estimadas através da maximização da verossimilhança, como vendo o vetor de proporções (0,1316181, 0,7457365, 0,1182945, 0,0034797, 0, 0) para a populações europeia, e o vetor de proporções (0, 0,3148988, 0,6061574, 0,0789438, 0, 0, 0) para a população Shimaa.

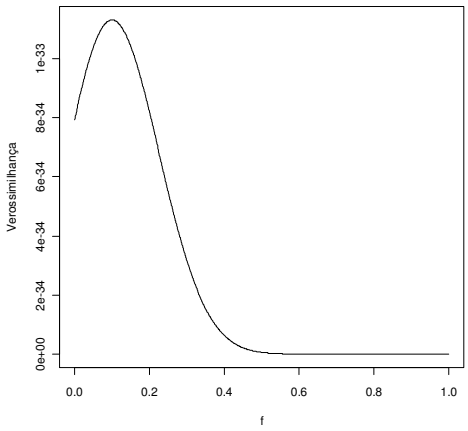
Resultando em $D = -2\ln\lambda(x) = 298,115 \leq \chi_9^2 = 16,919$. Logo, rejeitamos a hipótese de que ambas podem ser consideradas como sendo a mesma população.

Apêndice 3

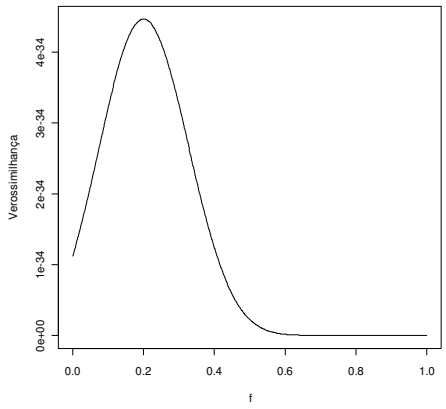
F verdadeiro = 0,05
F estimado = 0,04999928
(A)



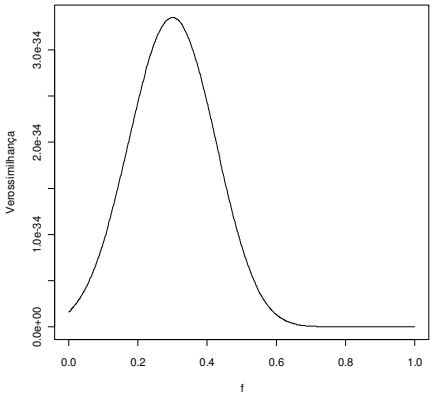
F verdadeiro = 0,1
F estimado = 0,0999872
(B)



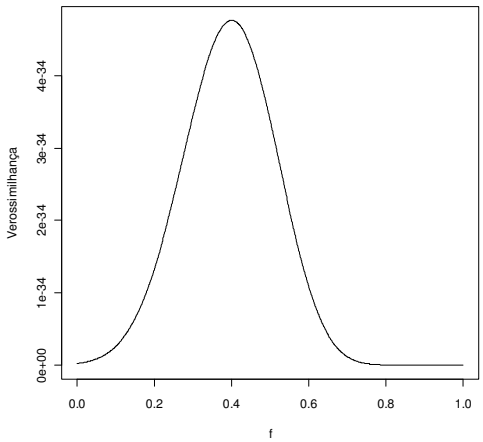
F verdadeiro = 0,2
F estimado = 0,2000007
(C)



F verdadeiro = 0,3
F estimado = 0,2999992
(D)



F verdadeiro = 0,4
F estimado = 0,3999823
(E)



Gráficos para a função de verossimilhança considerando-se as proporções (0.3, 0.4, 0.3, 0, 0, 0, 0, 0, 0)

Referências Bibliográficas

CASELLA G. e BERGER R. Inferência Estatística. Ed. 2ª, p. 291-295, 334-338, Editora: Cengage Learning, 2010.

DEMPSTER, A. P., LAIRD, N. M., RUBIN D.B. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc, Series B*, v. 39, p.1–38, 1977.

FEUK L., CARSON A.R., SCHERER S.W. Structural variation in the human genome. *Nature Rev Genetic*, v. 7, n.2, p.85-97, 2006

FIELD, S. F.; HOWSON, J. M. M.; MAIER, L. M.; WALKER, Susan; WALKER, N. M.; SMYTH, D. J.; ARMOUR, J.; CLAYTON, D.; TODD, J. A.. Experimental aspects of copy number variant assays at CCL3L1. *Nature America*, v. 15, n. 10, 2009.

GAUNT, T. R.; RODRIGUEZ, S.; GUTHRIE, P. A. I. e DAY, Ian N.M.An Expectation–Maximization Program for Determining Allelic Spectrum from CNV Data (CoNVEM): Insights into Population Allelic Architecture and Its Mutational History ConVEM. *Human Mutation*, v. 31, n. 4, p. 414–420, 2010.

GONZALEZ, E.; KULKARNI, H.; BOLIVAR, H.; MANGANO. A,; SANCHEZ, R.; CATANO, G.; NIBBS, R. J.; FREEDMAN, B. I.; QUINONES, M. P.; BAMSHAD, M. J.; MURTHY, K. K.; ROVIN, B. H.; BRADLEY, W.; CLARK, R.A.; ANDERSON, S. A.; O'CONNELL, R.J.; AGAN, B. K.; AHUJA, S. S.; BOLOGNA, R.; SEM, L.; DOLAN, M.; J, AHUJA, S. K.; The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, v. 307, n. 5714, p. 1434-40, 2005.

HARDY G.H. Mendelian proportions in a mixed population. *Science*, v. 28, p.49–50, 1908.

LANGE, K. Mathematical and Statistical Methods for Genetic Analysis. Ed. *Springer*, 2ª ed. Capítulo 3, p. 42-43.

KEMPTHORNE O. An Introduction to Genetic Statistics. Ed. *John Wiley & Sons*. Capítulo 5, p.72-93.

MCKINNEY, C.; FANCIULLIi, M.; MERRIMAN, M. E.; PHIPPS-GREEN, A.; ALIZADEH, B. Z.; KOELEMAN, B.P.; DALBETH, N.; GOW, P. J.; HARRISON, A. A.; HIGHTON, J.; JONES, P.B.; STAMP, L. K.; STEER, S.; BARRERA, P.; COENEN, M. J.; FRANKE, B.; Riel, P.L.; VYSE, T.J.; AITMAN, T. J.; RADSTAKE, T. R.; MERRIMAN, T. R.; Association of variation in Fcgamma receptor 3B gene copy number with rheumatoid arthritis in Caucasian samples. *Ann Rheum Dis*, v. 69, n. 9, p. 1711-6, 2010.

MILLS, R., WALTER K., STERWART C., HANDSAKER R., CHEN K., ALKAN C., ABYZOV A., YOON S., CHEETHAM R. and others. Mapping CN variation by population-scale genome sequencing. *Nature* 470(7332):59-65, 2011.

PAWITAN, Y.. In All Likelihood: Statistical Modelling and Inference Using Likelihood. *Claredon Press, Oxford*, 2001, p.61-62.

POLANSKA J. The EM Algorithm and its implementation for the estimation of frequencies of SNP-haplotypes. *J. Appl. Math. Comput. Sci.,* 2003, Vol. 13, No. 3, 419–429

REIS, R. L.; MUNIZ, J. A.; FONSECA E SILVA, F. SÁFADI, T.; DE AQUINO, L. H. Abordagem bayesiana da sensitividade de modelos para o coeficiente de endogamia. *Ciência Rural,* v.39, n. 6, p.1752-1759, set, 2009.

SANTHOSH G., CAMPBELL C. D. and EICHLER E.E.. Human Copy Number Variation and Complex Genetic Disease. *Annual Review of Genetics*, Vol. 45: 203-226, 2011.

SEBAT J., LAKSHMI B.,TROGE J., ALEXANDER J.,YOUNG J., LUNDIN P., MÄNÉR S., MASSA H., WALKER M. and others. Large-Scale Copy Number Polymorphism in the Human Genome. *Science* 23: Vol. 305 no. 5683 pp. 525-528, July 2004.

SELF S.G. and LIANG K-Y. Asymptotic Properlies of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association*, 82(398) 605□610, 1987.

SHIRYAEV A.N.. Probability. Ed. *Springer*, 2ª ed., 2000, pg 27.

SIEGEL S. e CASTELLAN J. Estatística Não-paramétrica Para Cências Do Cmportamento. Ed. *Artmed*, 2ed., 2006.

SILVA M. F. Estimação e teste de hipótese baseados em verossimilhanças perfiladas. Tese de doutorado, USP, 2005.

SOLER J. M. P. Estudo de simetria na associação genética usando dados de trios. Tese de doutorado, USP, 2011.

STURTEVANT A. H.. The Effects of Unequal Crossing over at the Bar Locus in Drosophila. *Genetics*. 1925 March; V.10(2): 117–147.

ZUCCHERATO L. W.. Estrutura populacional e diversidade de variações em número de cópias (CNVs) de genes do sistema imune em populações nativas da América do Sul. Tese de doutorado, UFMG, 2012.