CÁLCULO DE PROBABILIDADES ALÉLICAS E GENOTÍPICAS INDIVIDUAIS EM DADOS DE CNVS (COPY NUMBER VARIATIONS)

Silvana Schneider[1,3], Luciana Zuccherato[2,3], Eduardo Tarazona-Santos[2,3], Maíra Ribeiro Rodrigues[2,3], Denise Duarte[1,3]

Resumo: *Usualmente, os métodos disponíveis para a tipagem de CNVs são capazes de descrever apenas o número total de cópias por genoma diplóide, deixando subjacente a distribuição desconhecida das frequências alélicas e genotípicas. Para inferir estas frequências individuais através de dados do número de cópias de determinado gene, desenvolvemos um programa baseado no algoritmo CoNVEM, supondo que os dados estão em Equilíbrio de Hardy-Weinberg. A implementação do algoritmo foi feita na plataforma estatística R. O programa foi utilizado na obtenção das frequências do gene CCL3L1, um dos genes com maior variação observada entre as diferentes populações mundiais, e atualmente foco de intensa pesquisa devido à sua correlação positiva com o RNAm.*

1 Introdução

Copy Number Variation (CNV) corresponde a segmentos de DNA que variam de um kilobase a vários Megabases de tamanho e apresentam variações do número de cópias em relação a uma sequência referência, onde o número de cópias usual é 2 (Feuk L. et al. 2006). 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 lócus CCL3L1, cujo número de cópias pode variar de 0 a 14 cópias (Gonzalez E. et al. 2005). Usualmente, os métodos disponíveis para a tipagem de CNVs são capazes de descrever apenas o número total de cópias por genoma diplóide, deixando subjacente a distribuição desconhecida das frequências alélicas e genotípicas.

Para inferir as frequências alélica e genotípica de dados do número de cópias de determinado gene, desenvolvemos um programa que aplica Maximização da Esperança, EM, para determinar a frequência alélica de dados haplóides de CNV, supondo que os dados estão em Equilíbrio de Hardy-Weinberg. A implementação do algoritmo foi feita na plataforma estatística R. O programa recebe uma lista do número de indivíduos com o respectivo número de cópias, retorna as frequências alélicas e genotípicas e teste de aderência para comparar as frequências observadas com as esperadas. Ele estende o algoritmo proposto na ferramenta CoNVEM (Gaunt T., 2010), implementado na linguagem python, com a adição do cálculo das frequências genotípicas individuais. O programa tem código aberto e, portanto, pode ser facilmente expandido para incluir demais cálculos pertinentes.

O programa foi utilizado na obtenção da frequência alélica do número de cópias do gene CCL3L1, um dos genes com maior variação observada entre as diferentes populações mundiais, e atualmente foco de intensa pesquisa devido à sua correlação positiva com o RNAm, níveis de proteína e doenças, especialmente para patogênese do HIV e suscetibilidade a doenças autoimunes(Gonzalez E. et al, 2005; McKinney C. et al, 2008 e 2010), em três populações nativas americanas de etnias Ashaninka, Matsiguenga e Quechua localizadas na região andina e amazônica

[1]Departamento de Estatística, Universidade de Federal de Minas Gerais, MG
[2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, MG
[3]Agradecimentos: FAPEMIG, EMBO, CAPES, Medical Research Council (UK).

do Peru. Quando comparadas com dados da população Europeia, estas populações autóctones apresentam um aumento na duplicação de número de cópias, o que pode levar a um aumento nas interações de vários receptores de citosinas pré-inflamatórias e consequentemente ativação do processo imunológico.

2 Metodologia

O algoritmo será descrito fazendo-se uso da mesma notação utilizada no CoNVEM (Gaunt et al., 2010), pois as equações (1) à (5) pertencem a sua autoria. Sejam os alelos h_k e h_l, onde k e l são números naturais que denotam o número de cópias de cada alelo. Embora h_k e h_l não sejam observados diretamente em um indivíduo, o total, $k + l = j$, é observado.

O número total de cópias será representado por j, cada indivíduo pertence à classe j. Porém os membros da classe j não serão todos iguais, pois h_k varia de h_0 até h_j, idem para h_l, seguindo a condição de que $k + l = j$. A observação de apenas um indivíduo pode produzir pouca informação sobre seus alelos hk e hl, porém dada uma população de indivíduos inferências sobre as frequências alélicas, p_k e p_l, podem ser realizadas.

Assumindo-se que a população está em Equilíbrio de Hardy-Weinberg (EHW) [Hardy, 1908], a probabilidade genotípica é dada por:

$$\tilde{P}(h_k, h_l) = \begin{cases} p_k^2, & k = l \\ \\ 2p_k p_l, & k \neq l \end{cases} \tag{1}$$

As frequências alélicas $p_0, p_1, \ldots, p_h$ são inicializadas no primeiro ciclo da iteração com os valores $p_0^{(0)}, p_1^{(0)}, \ldots, p_h^{(0)}$, logo a frequência genotípica pode ser expressa por:

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

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

$$P_j^{(g)} = \begin{cases} p_{j/2}^{(g)2} + \displaystyle\sum_{k=0}^{j/2-1} 2p_k^{(g)} p_{j-k}^{(g)}, & j: \text{ par} \\ \\ \displaystyle\sum_{k=0}^{(j-1)/2} 2p_k^{(g)} p_{j-k}^{(g)}, & j: \text{ ímpar} \end{cases} \tag{3}$$

As equações acima estimam a frequência alélica baseada nos valores iniciais, logo é necessário fazer uma normalização com os dados da amostra. E, para o ajuste da estimativa das classes dos CNV´s para o próximo passo da iteração, baseado nas atuais estimativas e o número observado de cada classe j.

$$P(h_k, h_l) = \frac{n_j}{n} \frac{\tilde{P}(h_k, h_l)^{(g)}}{P_j^{(g)}} \tag{4}$$

onde n_j é o número de indivíduos observados na classe j, e n é o número total de indivíduos.

A frequência alélica estimada para a próxima iteração é calculada por:

$$\hat{p}_t^{(g+1)} = \frac{1}{2} \sum_{j=0}^{m} \sum_{k=0}^{j} \delta_{it} P_j(h_k, h_l)^{(g)} \tag{5}$$

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 frequência alélica estimada atual e a inferior for inferior que 0,00000001 ou haver no máximo 10.000 iterações.

Levando-se em consideração o fato de que a classe, j, de cada indivíduo é conhecida, a estimativa da frequência genotípica individual pode ser estimada por:

$$P_{ind}(h_k, h_l) = \frac{2P(h_k, h_l)^{(g)}}{\sum_{k=1}^{m}\sum_{l=1}^{m} P(h_k, h_l)^{(g)}}, \quad j = k + l \tag{6}$$

3 Resultados

O número de cópias do gene CCL3L1 foi medido em três populações nativas americanas localizadas na região andina do Peru: Quechua ($n = 120$), e na região de transição (amazônica): Ashaninka ($n = 142$) do grupo étnico Ashaninka e Monte Carmelo ($n = 24$), do grupo étnico Matsiguenga. A amostra também é composta pelos Shima ($n = 89$) e pelos Europeus ($n = 4266$) tipados por Field et al 2009.

O número de cópias esperado no grupo dos Ashaninka foi de 3,419, para os Shima foi de 3,528, para Monte Carmelo de 3,292, para os Europeus de 1,992 e para os Quechua de 3,575.

A distribuição da frequência do número de cópias dos Europeus difere das demais populações, em geral eles possuem classes mais baixas que os demais grupos, Tabela 1.

Tabela 1: Frequência Estimada do 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

A frequência alélica se distribui de forma semelhante entre os grupos Ashaninka, Shimaa e Monte Carmelo, como podemos ver na Tabela 2.

Tabela 2: Frequência alélica estimada em cada população.

	Frequência alélica das populações				
Alelo	Ashaninka	Shimaa	Monte Carmelo	Quechua	Europeus
0	0,189	0,298	0,219	0,000	0,035
1	0,670	0,618	0,746	0,448	0,860
2	0,124	0,068	0,035	0,265	0,105
3	0,009	0,004	0,000	0,094	0,000
4	0,000	0,012	0,000	0,193	0,000
5	0,000	0,000	0,000	0,000	0,000
6	0,008	0,000	0,000	0,000	0,000

A frequência genotípica individual estimada para cada classe, para a população Ashaninka, se encontra na Figura 1. Por exemplo: dado que o individuo possui 5 cópias, ele poderá ter o genótipo (1,4) com 0,444 de probabilidade ou o genótipo (2,3) com 0,555 de probabilidade.

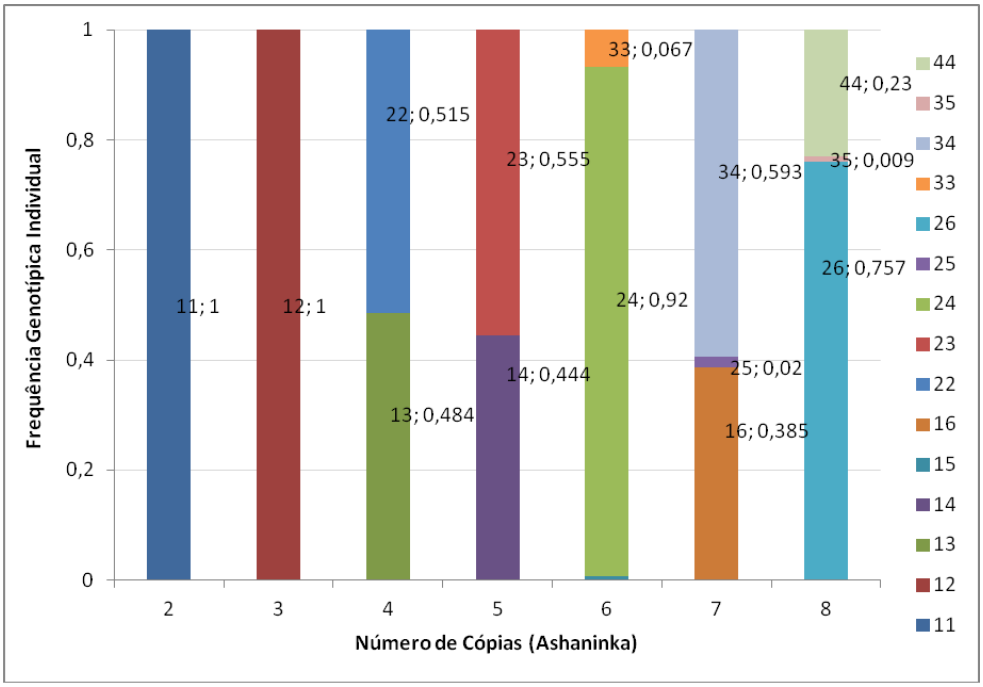

Figura 1: Frequência Genotípica Individual para a população Ashaninka.

4 Considerações Finais

Sob Equilíbrio de Hardy-Weinberg os cálculos estimam perfeitamente as frequências, porém quando a população não está em equilíbrio não é aconselhado o uso do algoritmo, pois os resultados podem não serem fidedignos ou pode não haver convergência. Uma das alternativas para solucionar esse caso é incluir um parâmetro, f, esse parâmetro é o Coeficiente de Endogamia, para captar o quanto a população desvia do Equilíbrio.

A estimativa frequência genotípica dos filhos pode ser melhorada utilizando-se a informação da frequência genotípica dos pais. E, pode ser obtida pelo cálculo:

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

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. 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 $(h_{lf} = h_{lp})$.

Como trabalho futuro, esse cálculo será incorporado ao programa original, juntamente com o cálculo do parâmetro f.

Referências

[1] Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm, *J R Stat Soc, Series B*, **39**, 1-38, 1977.

[2] Feuk, L., Carson, A. R., Scherer, S. W., Structural variation in the human genome, *Nat Rev Genet*, **7(2)**, 85-97, 2006.

[3] Field, F. et al., Experimental aspects of copy number variant assays at CCL3L1, *Nature America*, **15**, 10, 2009.

[4] Gaunt, T. et al., An Expectation-Maximization Program for Determining Allelic Spectrum from CNV Data (CoNVEM): Insights into Population Allelic Architecture and Its Mutational History ConVEM, *Human Mutation*, **31 4**, 414-420, 2010.

[5] Gonzalez, E. et al., The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility, *Science*, **307 5714**, 1434-1440, 2005.

[6] Hardy G. H., Mendelian proportions in a mixed population. *Science*, **28**, 49-50, 1908.

[7] Mckinney, C. et al., Association of variation in Fcgamma receptor 3B gene copy number with rheumatoid arthritis in Caucasian samples, *Ann Rheum Dis*, **69 9**, 1711-1716, 2010.