Análise Multiestágio na Identificação de Genes em Estudos de Famílias e Dados de SNP †

Mirian de Souza[1], Suely R. Giolo[2], Mariza de Andrade[3], Júlia P. Soler[1]

[1]Departamento de Estatística, Universidade de São Paulo, Brasil

[2]Departamento de Estatística, Universidade Federal do Paraná, Brasil

[3]Biostatistics Department, Mayo Clinic, USA

† *Apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)*

Resumo

Um dos maiores problemas no mapeamento de genes associados a doenças complexas atualmente é o de como tratar computacionalmente a alta dimensionalidade dos dados genéticos. Neste trabalho é considerado um procedimento multiestágios para análise de dados de famílias e plataformas de marcadores SNPs, proposto por Aulchenko et al. (2007), que permite reduzir consideravelmente o tempo computacional. Sob a formulação de modelos mistos, os autores propõem modelar o efeito de SNPs como fixo e associado com o componente de variância residual. Há controvérsias sobre como deve ser modelado o efeito de marcadores do tipo SNP quando se dispõe de dados de famílias. Neste trabalho serão discutidas limitações e vantagens desta alternativa de análise bem sua implementação computacional usando os recursos do aplicativo estatístico R. Como aplicação são considerados dados de um estudo com famílias brasileiras.

1. Introdução

O estudo de doenças complexas, tais como hipertensão, diabetes e obesidade é de grande importância na área médica, pois essas doenças afetam muitas pessoas em nosso país e no mundo. Acredita-se que o padrão de variação destas doenças envolve componentes ambientais, genéticos e suas possíveis interações. Sobre os componentes genéticos, supõe-se que são constituídos tanto por variações no genoma que são comuns na população quanto variações raras, que ocorrem em poucas famílias (Almasy and Blangero, 2008; Blangero, 2004). Para o mapeamento de genes associados a doenças complexas, os avanços biotecnológicos têm sido inestimáveis no sentido de permitirem a amostragem cada vez mais densa e completa do genoma humano. Contudo, o

desenvolvimento e emprego de metodologias de análise de dados genéticos não têm acompanhado tal velocidade. Neste contexto, destaca-se a genotipagem extremamente densa de variantes comuns do genoma humano via plataformas de SNPs (*Single Nucleotide Polimorphisms*), cuja análise dos dados gerados envolve complicações, principalmente, decorrentes da alta dimensionalidade dos dados e do pequeno efeito individual. Tais desafios analíticos têm sido tratados principalmente no contexto de estudos de associação caso-controle (Ziegler et al., 2008), mas com menor ênfase em estudos com famílias. Neste trabalho, para o mapeamento de determinantes genéticos associados a fatores de risco cardiovascular utilizando dados de famílias e plataformas de SNPs, é considerado o procedimento de análise em multiestágios proposto por Aulchenko et al. (2007). Neste caso, sob a formulação de modelos mistos, os autores propõem modelar o efeito de SNPs como fixo e associado com o componente de variância residual. São discutidas limitações e vantagens desta alternativa de análise. Para aplicação são considerados dados de um estudo com famílias brasileiras.

2. Metodologia

No mapeamento de genes, basicamente, dois processos de amostragem estão envolvidos: amostragem de indivíduos de uma população de interesse e amostragem do genoma destes indivíduos. A amostragem do genoma dos indivíduos é feita por meio de plataformas de marcadores moleculares e, em geral, se destacam duas classes de marcadores que têm sido usados em mapeamento de genes: os do tipo microsatélites e os SNPs (*Single Nucleotide Polimorphisms*). As plataformas de microsatélites são pouco densas (utilizam espaçamentos médios de 5 a 10cM), seus marcadores estratificam a população em muitas categorias genotípicas (são altamente polimórficos) e, em geral, cada marcador corresponde a uma grande região cromossômica (de centenas a milhares de bases). Deste modo, estes marcadores podem conter mais de uma variação na seqüência de DNA, que seja comum na população ou mesmo rara, isto é, que segrega em poucas famílias. Já as plataformas de SNPs são altamente densas (cerca de 1 milhão de marcadores no caso da plataforma Affymetrics 6.0), cada marcador é dialélico, isto é, estratifica a população em até três grupos genotípicos (AA, Aa e aa), e correspondem a variações de uma única base na seqüência de DNA que são comuns na população (prevalência de pelo menos 1%).

Considerando variáveis quantitativas avaliadas em indivíduos e seus familiares, mesmo sem qualquer informação de marcadores moleculares, é possível obter

estimativas de herdabilidade poligênica que informam se há componente genético envolvido na variação da resposta de interesse. O modelo de componentes de variância tem sido adotado neste caso (Almasy and Blangero, 1998; de Andrade et al., 1999), sendo a herdabilidade definida como a proporção da variância total que é devida a componentes genéticos (coeficiente de correlação intra-classe). Sob tal modelo, a variável fenotípica observada no indivíduo *i* (denotada por Y_i) é definida por:

$$Y_i = \mu + X_i\beta + g_i + e_i \qquad (1)$$

onde, μ é média fenotípica geral, β é o vetor de parâmetros fixos, X_i é uma matriz de especificação dos efeitos fixos, g_i e e_i são variáveis aleatórias definindo o efeito poligênico e o erro, respectivamente. Os efeitos aleatórios, g_i e e_i, são assumidos não correlacionados e normalmente distribuídos, com média zero e variância σ_g^2 e σ_e^2, respectivamente. Em geral, o erro é suposto comum a cada indivíduo, enquanto o componente poligênico é compartilhado entre indivíduos, sendo proporcional ao seu grau de parentesco. A covariância entre as variáveis fenotípicas para os indivíduos *i* e *i'* é dada por:

$$Cov(y_i; y_{i'}) = \begin{cases} \sigma_g^2 + \sigma_e^2 \quad para \quad i = i', \\ 2\phi_{ii'}\sigma_g^2 \quad para \quad i \neq i', mas\ relacionados, \\ 0 \quad para \quad i \neq i' \ \ e \ \ não - relacionados. \end{cases}$$

O parâmetro $2\phi_{ii'}$ é o coeficiente de relacionamento entre os indivíduos *i* e *i'*, sendo dado por $\left(\frac{1}{2}\right)^r$, r representa o grau de relacionamento. A função de verossimilhança considerando os dados dos membros de uma família é, em geral, obtida a partir da distribuição normal multivariada. O teste de herdabilidade poligênica ($H_0 : \sigma_g^2 = 0$ vs $H_0 : \sigma_g^2 > 0$) é feito com base na estatística razão de verossimilhanças que, para o caso de componentes de variância, tem uma distribuição mistura de qui-quadrados.

Quando a informação adicional de marcadores moleculares está disponível, extensões do modelo poligênico (1) podem ser feitas introduzindo um componente aleatório adicional no modelo:

$$Y_i = \mu + X_i\beta + q_i + g_i + e_i \qquad (2)$$

onde, q_i é a variável aleatória que define o efeito de um oligogene ou variante genética de efeito maior sobre o fenótipo, sendo assumida como não correlacionada com as demais variáveis (do poligene e erro) e seguindo uma distribuição Normal com média zero e variância σ_q^2. Neste caso, a covariância entre os indivíduos *i* e *i'* é dada por:

$$Cov(y_i; y_{i'}) = \begin{cases} \sigma_g^2 + \sigma_q^2 + \sigma_e^2 & para \quad i = i', \\ 2\varphi_{ii'}\sigma_g^2 + \pi_{ii'}\sigma_q^2 & para \quad i \neq i', mas\ relacionados, \\ 0 \quad para \quad i \neq i' \ \ e \ \ não - relacionados. \end{cases}$$

o parâmetro $\pi_{ii'}$ é a probabilidade dos indivíduos i e i' compartilharem pelo menos um alelo ibd (idênticos por descendência) em um loco de QTL (Quantitative Trait Locus). O teste do efeito do QTL ($H_0 : \sigma_q^2 = 0$ vs $H_0 : \sigma_q^2 > 0$) é realizado com base na estatística razão de verossimilhanças.

O modelo (2) tem sido utilizado para o mapeamento de genes em doenças complexas quando dados de marcadores moleculares do tipo microsatélites estão disponíveis. Para dados com marcadores do tipo SNPs a aplicação direta do modelo (2) para identificação de genes parece não ser eficiente e modificações precisam ser pesquisadas. Um dos pontos de questionamento é se o efeito do SNP deve ser modelado estatisticamente como um componente fixo ou aleatório. Neste contexto, já que SNPs são variantes comuns, acredita-se que os dados de SNP não modelam estrutura familiar, isto é, não explicam a correlação entre indivíduos relacionados, sendo mais apropriado modelar seu efeito no componente fixo. Para detectar o efeito do SNP Aulchenko et al. (2007) propõem uma análise de mapeamento de genes em vários estágios.

Nessa proposta de mapeamento o modelo poligênico (1) é estendido para acomodar o efeito dos SNPs X_j, j = 1, 2,...,M, que são modelados como efeitos fixos. Essa proposta tem os seguintes estágios:

Estágio 1. Ajuste do modelo poligênico para se obter os resíduos condicionais, dados por,

$$\hat{e}_i = y - \left(X_i \hat{\beta} + \hat{g}_i \right) \tag{3}$$

em que $\hat{\beta}$ é o coeficiente de regressão associado a covariáveis (não genéticas) e $\hat{g}_i$ é o valor predito do efeito aleatório do poligene.

Estágio 2. Os resíduos obtidos em (3) são usados como uma variável dependente em um modelo de regressão sob premissas clássicas para cada SNP, X_j, tal que,

$$\hat{e}_i = \mu + \beta_j X_{ij} + e_i \quad (4)$$

em que β_j é o correspondente efeito do SNP X_j. Testes simultâneos de $H_o : \beta j = 0$ (j=1, 2, ..., M) são realizados para verificar quais SNPs são significativos.

Estágio 3. Os SNPs com efeitos significativos são selecionados para uma análise final por meio do seguinte modelo poligênico,

$$Y = \mu + X\beta + \beta_j X_j + g + e \quad . \quad (5)$$

Com a utilização deste procedimento em multiestágios, Aulchenko et al. (2007) mostra que é possível reduzir drasticamente o tempo computacional que seria exigido se fossem ajustados diretamente M modelos poligênicos do tipo (5).

3. Resultados

Neste trabalho discutimos as limitações e vantagens do procedimento multiestágio proposto por Aulchenko et al. (2007) e sua implementação computacional usando os recursos do aplicativo estatístico R. A aplicação desta metodologia é também avaliada na análise de um conjunto de dados de famílias brasileiras (Oliveira et al., 2007; Giolo et al. 2009).

4. Referências Bibliográficas

Almasy, L and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Gen*. **62**:1198-211.

Almasy, L. and Blangero, J. (2008). Human QTL linkage mapping. *Genetica*. doi: 10.1007/s10709-008-9305-3

Altshuler, D; Daily, M.J. and Lander, ES. (2008). Genetic Mapping in Human Disease. *Science* **322**: 881-888.

Aulchenko, Y.S., Koning, D., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**:577-585.

Blangero, J. (2004). Localization and identification of human quantitative trait loci: King harvest has surely come. *Current Opinion in Genetics & Development* **14**:233-240.

Blangero J., Goring, H.H.H., Kent, J.A. J., Williams, J.T., Peterson, C.P. Almasy, L. and Dyer, T.D. (2005). Quantitative trait nucleotide analysis using bayesian model selection. *Human Biology* **5**:541-559.

Giolo, S.R., Pereira, A.C., de Andrade, M., Oliveira, C.M., Krieger, J.E. and Soler, J.M.P. (2009). Genetic analysis of age-at-onset for cardiovascular risk factors in a Brazilian family study. *Human Heredity* 68: 131-138.

Oliveira, C.M., Pereira, A.C., de Andrade, M., Soler, J.M.P. and Krieger, J.E. (2008). Heritability of Cardiovascular Risk Factors in a Brazilian Population: Baependi Heart Study. *BMC Medical Genetics* **9**: 32.

Ziegler, A.; Konig, I.R. and Thompson, J.R. (2008). Biostatistics aspects in genome-wide association studies. *Biometrical Journal* **50**: 1-8.