Análise de associação genômica ampla baseada em conjunto de genes: implementação em R

Aline Taise Guerreiro[1]
Roberto Hiroshi Higa[2]

Variações genéticas presentes em uma população podem estar associadas a muitas características como susceptibilidade a doenças em humanos (ex: diabetes, câncer, e doenças psiquiátricas). Atualmente, tecnologias de genotipagem de baixo custo, baseadas em marcadores moleculares do tipo polimorfismo de base única *Single Nucleotide Polymorphism* (SNP) são utilizados para identificar variações desse tipo associadas com doenças. Tais estudos, são denominados, estudos de associação genômica ampla, *Genome Wide Association Studies* (GWAS). No caso de espécies de interesse agropecuário, essas variações genéticas estão relacionadas a características que podem impactar ganhos de qualidade e produção. Portanto, é de extrema importância a utilização de novos métodos computacionais para identificação desses marcadores, já que isto pode contribuir para a seleção de indivíduos superiores, considerando os traços fenotípicos de interesse em espécies animais utilizadas em programas de melhoramento coordenados pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa).

Uma das estratégias para GWAS, ainda não explorada pela Embrapa, é a análise de enriquecimento de conjuntos de genes com função biológica similar. Originalmente, proposto no contexto de análise de expressão gênica, a análise de enriquecimento de um conjunto de genes *Gene Set Enrichment Analysis* (GSEA), é um método que analisa conjuntos de genes que compartilham a mesma função biológica, localização cromossômica ou via regula-

[1] Universidade Estadual de Campinas (Unicamp)
[2] Embrapa Informática Agropecuária

tória , procurando identificar aqueles que apresentam diferenças de expressão entre as situações analisadas, apesar dos genes individualmente não apresentarem diferenças de expressão altamente significativas (CURTIS et al., 2005; SUBRAMANIAN et al., 2005). O objetivo deste trabalho é a criação de um pacote R (R CORE TEAM, 2014) que implemente quatro diferentes métodos de GSEA no contexto de GWAS, considerando adaptações para aplicação em espécies animais de interesse para a agricultura. Os métodos implementados foram: a) GSEA-SNP (HOLDEN et al., 2008); b) SRT (O'DUSHLAINE et al., 2009); c) twoStage-RF (CHANG et al., 2008); c) twoStage-RF; d) modelado por modelo linear misto (WANG et al., 2011).

O método GSEA-SNP é uma adaptação do método GSEA para o contexto de GWAS, onde o pressuposto é que um fenótipo está associado a variações em genes que compartilham a mesma função biológica, via de regulação ou localização cromossômica. Da mesma forma, a ideia é identificar vias (conjuntos de SNPs) associadas com o fenótipo de interesse, apesar de o nível de significância ao se analisar a associação de cada SNP individualmente não ser tão alto. O teste da razão de SNPs, em inglês *SNP Ratio Test* (SRT), compara a proporção de SNPs considerados significativos com o total de SNPs nos genes pertencentes a uma determinada via de regulação, cromossomo ou função biológica. Um p-valor empírico é utilizado para testar a hipótese de que vias altamente associados com o fenótipo são enriquecidos em SNPs significativos. Já a metodologia twoStage-RF, que utiliza Florestas Aleatórias, em inglês *Random Forest* (RF), aborda o problema por meio de análise discriminante, onde na análise de cada conjunto de genes, os SNPs são utilizados como uma variável preditora e o fenótipo como uma variável resposta. Por fim, o mesmo problema é abordado por modelos lineares mistos (MLM, na sigla em inglês *Mixed Linear Model*), sendo a associação de um conjunto de genes modelada como o efeito fixo no modelo linear. O teste de associação para um conjunto de genes consiste em testar se este efeito fixo é diferente de zero.

Os quatro métodos mencionados foram implementados utilizando a linguagem R e o ambiente Rstudio e, no momento, encontram-se em fase de testes e organização na forma de um pacote R, enquanto que para manipulação dos dados de genótipos foi utilizado o pacote R snpStats. Além disso, foram considerados duas categorias de conjuntos de genes, um baseado no banco de dados e vias biológicas KEGG (KANEHISA et al., 2014) e o outro baseado em anotações funcionais GO (ASHBURNER et al., 2000) . Essas

categorias foram utilizadas para os testes durante a implementação dos métodos. Tanto para a implementação do método GSEA-SNP, quanto para a do método SRT não foram utilizados pacotes específicos do R. Em ambos os casos foram criadas funções R para cálculo das estatísticas utilizadas pelo métodos Kolmogorov-Smirnov e SRT, e obtenção dos p-valores. Para implementação do método twoStage-RF utilizou-se o pacote R randomForest, que calcula a associação dos conjuntos de genes com o fenótipo com base em árvores aleatórias. Finalmente, no caso da implementação do método baseado em modelo linear misto, utilizou-se o pacote R nlme para ajuste do modelo linear e o pacote locfdr para estimação dos parâmetros de associação.

Os trabalhos futuros incluem a finalização dos testes com dados simulados, a documentação dos scripts desenvolvidos e sua organização como um pacote R.

Palavras-chave: GWAS, SNPs, Random Forests, mixed models.

Referências

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene Ontology: tool for the unification of biology. **Nature Genetics**. v. 24, n. 1, p. 25-29, May 2000.

CHANG, R. F.; WIENCKE, J. K; WIEMELS, J. L.; SMIRNOV, I.; PICO, A. R.; TIHAN T.; PATOKA, J.; MIIKE, R.; SISON, J. D.; RICE, T.; WRENSCH, M. R. Pathway analysis of single-nucleotide polymorphisms potentially associated with glioblastoma multiforme susceptibility using random forests. **Cancer Epidemiology Biomarker & Revention**, v. 17, n. 6, p. 1368-1373, June 2008.

CURTIS, R. K.; ORESIC, M.; VIDAL-PUIG, A. Pathways to the analysis of microarray data. **Trends In Biotechnology**, v. 23, n. 8, p. 429-435, Aug. 2005.

HOLDEN, M.; DENG, S.; WOJNOWSKI, L.; KULLE, B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. **Bioinformatics**, v. 24, n. 23, Oct. 2008.

KANEHISA, J.; GOTO, S.; SATO, Y.; KAWASHIMA, M.; FURUMICHI, M.; TANABE, M. Data, information, knowledge and principle: back to metabolism in KEGG. **Nucleic Acids Research**, v. 42, n. D1, p. D199-D205, Jan. 2014.

O'DUSHLAINE, C.; KENNY, E.; HERON, E. A.; SEGURADO, R.; GILL, M.; MORRIS, D. W.; CORVIN, A. The SNP ratio test: pathway analysis of genome-wide association datasets. **Bioinformatics**, v. 25, n. 20, p. 2762-2763, 2009.

R CORE TEAM. R: a language and environment for statistical computing. Disponível em: <http://www.R-project.org/>. Acesso em: 26 set. 2014.

SUBRAMANIAN, A.; TAMAYO, P.; K. MOOTHA, V. K.; MUKHERJEE, S.; EBERT, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDE, E. S.; MESIRO, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, v. 102, n. 43, Oct. 2005, p. 15545-15550.

WANG, L.; JIA, P.; WOLFINGER, R. D.; CHEN, X.; GRAYSON, B. L.; AUNE, T. M.; ZHAO, Z. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. **Bioinformatics**, v. 27, n. 5, p. 686–692, Mar. 2011.