Desenvolvimento de pacotes estatísticos em linguagem R para análise de associação genômica ampla baseada em conjuntos de genes

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

Variações genéticas presentes em uma população podem estar associadas a muitas características como suceptibilidade 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 (SNP, na sigla em inglês Single Nucleotide Polymorphism) são utilizados para identificar variações desse tipo associadas com doenças. Tais estudos, são denominados, estudos de associação genômica amplo (GWAS, na sigla em inglês Genome Wide Association Studies). 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, que ainda não foi explorada pela Embrapa, é a análise de enriquecimento de conjuntos de genes com função biológica simlar (CURTIS et al., 2005). Originalmente, proposto no contexto de análise de expressão gênica, a análise de enriquecimento de um conjuntos de genes (GSEA), é um método que analisa conjuntos de genes que

[1] Universidade Estadual de Campinas - aline.guerreiro@colaborador.embrapa.br
[2] Embrapa Informática Agropecuária - roberto.higa@embrapa.br

compartilham a mesma função biológica, localização cromossômica ou via regulatória (SUBRAMANIAN et al., 2005), procurando identificar conjuntos de genes que apresentam diferenças de expressão entre as situações analisadas, apesar de os genes individualmente não apresentarem diferenças de expressão altamente significativas. O método GSEA-SNP (HOLDEN et al., 2008) é uma adaptação do método GSEA para o contexto de GWAS, onde o pressuposto é que um fenótipo que indica estado de doença 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) associados com o fenótipo de interesse, apesar do nível de significância ao se analisar a associação de cada SNP individualmente não ser tão alto. Também existem outras proposições para realização deste tipo de análise, tais como o teste da razão de SNPs (SRT, na sigla em inglês SNP Ratio Test) (O'DUSHLAINE et al., 2009) que compara a proporção de SNP's considerados significativos com o total de SNP's nos genes pertencentes a uma determinada via de regulação, cromossomo ou função biológica. Calcula-se, então, um p-valor empírico que testará a hipótese de que vias altamente associados com o fenótipo são enriquecidos em SNPs significativos. Há ainda, outras metodologias como a análise discriminante que utiliza Florestas Aleatórias (RF, na sigla em inglês Random Forest) (CHANG et al., 2008), onde na análise de cada conjunto de genes, os SNP's são utilizados como uma variável preditora e o estado de doença como uma variável resposta. Já em modelos lineares mistos (MLM, na sigla em inglês Mixed Linear Model) (WANG et al., 2011), a associação de um conjunto de genes com mesma função biológica, via de regulação ou localização cromossômica é modelada como um 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.

O objetivo deste trabalho é a criação de um pacote R (R CORE TEAM, 2013) que implemente cada um dos métodos mencionados, considerando adaptações para aplicação em espécies animais de interesse para a agricultura. O processo de criação de pacotes R segue um protocolo descrito no manual do próprio software R, denominado "Writing R Extension", e é apoiado por funções implementadas no próprio R. Por isso, a metodologia de desenvolvimento deste trabalho consiste em primeiro estudar e entender os métodos estatísticos mencionados acima; estudar o protoloco de produção de pacotes do R e as funções de apoio existentes; e, então, implemen-

tar os métodos estatísticos com base no protocolo de produção de pacotes R, ambos estudados nas etapas anteriores.

Atualmente, o trabalho encontra-se em fase inicial de estudos para entendimento dos métodos estatísticos e planejamento do pacote a ser implementado.

Referências

CHANG, J. S.; YEH, 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 & Prevention**, Philadelphia, v. 17, n. 6, p. 1368-1373, Jun. 2008.

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

HOLDEN, M.; DENG, S.; WOJNOWSKI, L.; KULE, B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. **Bioinformatics**, Oxford, v. 24, n. 23, 2008. Doi:10.1093/bioinformatics/btn516.

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 Applications Note**, Oxford, v. 25, n. 20, p. 2762–2763, July 2009. Doi: 10.1093/bioinformatics/btp448.

R CORE TEAM. **R**: a language and environment for statistical computing. Disponível em: <http://www.R-project.org/>. Acesso em: 18 set. 2013.

SUBRAMANIAN, A.; TAMAYO, P.; MOOTHA, V. K.; MUKHERJEE, S.; EBERT, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDER, E. S.; MESIROV, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 43, p. 15545-15550, 2005. Disponível em: <http://www.pnas.org/content/102/43/15545.full.pdf+html?with-ds=yes>. Acesso em: 30 set. 2013.

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