Avaliação de ferramentas para detecção de interações epistáticas para fenótipos quantitativos em estudos de associação genômica ampla

Augusto Renan Soares[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).

Estudos de GWAS baseiam-se na utilização de plataformas de genotipagem com centenas de milhares de SNPs e conjuntos de dados contendo milhares de indivíduos, sendo caracterizados por um p >> n, onde p é número de SNPs e n é o número de indivíduos. A maioria das metodologias para análise de dados de GWAS baseia-se na realização de testes estatísticos para cada SNP independentemente, seguido por procedimento para correção de múltiplos testes (ZIEGLER et al., 2008). Sabe-se, contudo, que fatores genéticos

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

que afetam a manifestação de um fenótipo pode envolver a participação de muitos genes, bem como a interação entre eles (epistasia), o que agrava a situação de p >> n ao se analisar esse tipo de dado, pois p agora pode ser uma combinação de dois SNPs. Por isso, a maioria dos métodos computacionais para análise de epistasia em GWAS é aplicável a conjuntos de dados contendo apenas poucos milhares de SNPs (ex: MDR e logic regression (FOULKES, 2009).

Ferramentas e procedimentos para detecção de espistasia em conjuntos de dados de GWAS é um aspecto ainda não explorado por projetos da Embrapa. Considerando que os fenótipos de maior interesse para a agropecuária são quantitativos, este trabalho tem por objetivo avaliar ferramentas para detecção de interações epistáticas para fenótipos quantitativos em conjuntos de dados de GWAS.

Inicialmente foi selecionada para avaliação um conjunto de ferramentas reportadas na literatura como adequadas para tratamento de conjuntos de dados baseados em centenas de milhares de SNPs e fenótipos quantitativos. São eles: EPISNP (MA et al., 2008), FastEpistasis (SCHÜPBACH et al., 2010), Athena (TURNER et al., 2010a), GENN (TURNER et al., 2010b), AntEpiSeeker (WANG et al., 2010) e SNPInterForest (YOSHIDA; KOIKE, 2011). A partir da disponibilidade pública dessas ferramentas, facilidade de instalação e operação, tipo de licença e a existência de módulos para execução em paralelo, foi realizada uma triagem, das quais restaram apenas as ferramentas EPISNP e FastEpistasis. Posteriormente, foi acrescentada, a essa lista, a ferramenta EPIQ (ARKIN et al., 2014), que apareceu na literatura apenas recentemente.

Para realizar a avaliação das ferramentas selecionadas, está sendo utilizado um conjunto de dados simulados QTLMAS (ELSEN et al., 2012) A linguagem Python e a ferramenta Plink (PURCELL et al., 2007) foram utilizados para realizar a conversão de formatos de arquivos, de acordo com as exigências de cada ferramenta.

No momento, os testes com as ferramentas selecionadas encontram-se em andamento, utilizando uma máquina com processador de quatro núcleos. Após a fase de comparação dos métodos quanto à acurácia, pretende-se realizar testes distribuindo o processamento entre várias máquinas para avaliar a escalabilidade de cada método.

Palavras-chave: Epistasia, GWAS, fenótipos quantitativos, paralelismo.

Referências

ARKIN, Y.; RAHMANI, E.; KLEBER, M. E.; LAAKSONEN, R.; MÄRZ, W.; HALPERIN, E. EPIQ—efficient detection of SNP–SNP epistatic interactions for quantitative traits. **Bioinformatics**, v. 30, p. i19-i25, June 2014.

ELSEN, J.-M.; TESSEYDRE, S.; FILANGI, O.; LE ROY, P.; DEMEURE, O. XV th QTLMAS: simulated dataset. **BMC Proceedings**, v. 6, p. S1, 2012. Sup. 2.

FOULKES, A. S. **Applied statistical genetics with R**: for population based association studies. New York: Springer. 2009. 252 p. ill.

MA, L.; H.; RUNESHA, H. B.; DVORKIN, D.; GARBE, J. R.; DA, Y. Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. **BMC Bioinformatics**, v. 9, p. 315, 2008.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I; DALY, M. J.; SHAM, P. C. PLINK: a toolset for whole-genome association and population-based linkage analysis. **American Journal of Human Genetics**, v. 81, n. 3, p. 559-575, 2007. Disponível em: <http://pngu.mgh.harvard.edu/purcell/plink/>. Acesso em: 26 set. 2014.

SCHÜPBACH, T.; XENARIOS, I.; BERGMANN, S.; KAPUR, K. FastEpistasis: a high performance computing solution for quantitative trait epistasis. **Bioinformatics**, v. 26, n.11, p. 1468–1469, Apr. 2010.

TURNER, S. D.; DUDEK, S. M.; RITCHIE, M. D. ATHENA: a knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. **BioData Mining**, v. 3, n. 1, p. 5, Sept. 2010a.

TURNER, S. D.; DUDEK, S. M.; RITCHIE, M. D. Grammatical evolution of neural networks for discovering epistasis among quantitative trait Loci. **Lecture Notes in Computer Science**, v. 6023, p. 86–97. 2010b.

WANG, Y.; LIU, X.; ROBBINS, K.; REKAYA, R. AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. **BMC Research Notes**, v. 3, p. 117, 2010.

YOSHIDA, M.; KOIKE, A. SNPInterForest: A new method for detecting epistatic interactions. **BMC Bioinformatics**, v. 12, p. 469, 2011.

ZIEGLER, A.; KÖNIG, I. R.; THOMPSON, J. R. Biostatiscal aspects of genome-wiede association studies. **Biometrical Journal**, v. 50. n. 1, p. 8-28, 2008.