Knowledge-Based Analysis of Genome-Wide SNP Scanning Data

Manhong Dai, Weijian Xuan, Huda Akil, Stanley J. Watson, Fan Meng
Molecular and Behavioral Neuroscience Institute and Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109

Many million dollar-scale genome-wide association (GWA) studies are currently underway for identifying genetic elements involved in complex disorders. However, prevailing GWA data analysis methods are focusing on the association of individual SNP alleles with a complex disease, although multiple alleles are involved by definition. Available data suggest few SNP alleles can pass the genome-wide multiple testing criteria and the false positive rate is very high [1]. Analysis methods suitable for detecting multiple genetic elements underlying complex diseases will be extremely useful.

We believe that the GWA data analysis problem is similar in nature to the expression microarray data analysis: multiple genes may participate in the same pathophysiological process but the contribution from individual genes is small. The elegant gene-set based analysis methods [2, 3] are designed for detecting subtle but coordinated expression change of functionally related genes utilizing existing functional knowledge. It is conceivable that the use of group statistics for functionally related SNPs, in addition to the statistics of individual SNPs, may increase the sensitivity of detecting multiple SNPs involved in the targeted disorder besides directly linking SNPs to known biological functions.

Toward this purpose, we adapt the core gene-set analysis algorithms to GWA data analysis by adding 1) support for SNP genotype data and SNP statistics such as Chi square test 2) weighting mechanisms for correcting the dependence of multiple SNPs in the same linkage disequilibrium region within the same SNP group 3) SNP function group definitions using existing knowledge, such as Entrez Gene, Gene Ontology, KEGG/BioCarta/GenMAPP pathways and cytobands.

We apply such knowledge-based approach to the Mayo Clinic-Perlegen Parkinson's disease GWA data, which is the only set of GWA data currently available in the public domain[4]. In order to objectively judge the biological relevance of the results as well as the influence of various parameter settings, SNP dependency correction methods and SNP-gene relationship definitions, we develop an automated evaluation method utilizing the gene-Parkinson's disease relationship extracted from the full Medline database using in-house natural language processing engines [5].

Both the automated validation and manual literature searches indicate that our approach has dramatic improvement over the individual SNP-based approach in identifying genes known to be related to the Parkinson's disease in top ranked candidate gene lists. Such increases are statistically significant over random chances based on Medline literature. A dozen or so top-ranked genes can also pass the genome-wide multiple testing criteria.

Although the use of knowledge-based approach for GWA data analysis is still a work in progress, results from the Parkinson's disease data set are very encouraging and we strongly recommend researchers having access to GWA data sets to test this approach. The R-package for knowledge-based GWA data analysis is freely available at: http://arrayanalysis.mbni.med.umich.edu/MBNIUM.html .

[1] Thomas DC (2006) Are we ready for genome-wide association studies? Cancer Epidemiol Biomarkers Prev, 15: 595-8.
[2] Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102:15545-50.
[3] Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ (2005) Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A102: 13544-9.
[4] Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA et al (2005) High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77: 685-93.
[5] Xuan W, Watson SJ, and Meng F (2005) GeneInfoMiner--a web server for exploring biomedical literature using batch sequence ID. Bioinformatics, 21: 3452-3.