A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants

Wei Pan,[1,*] Il-Youp Kwak,[1] and Peng Wei[2,*]

In spite of the success of genome-wide association studies (GWASs), only a small proportion of heritability for each complex trait has been explained by identified genetic variants, mainly SNPs. Likely reasons include genetic heterogeneity (i.e., multiple causal genetic variants) and small effect sizes of causal variants, for which pathway analysis has been proposed as a promising alternative to the standard single-SNP-based analysis. A pathway contains a set of functionally related genes, each of which includes multiple SNPs. Here we propose a pathway-based test that is adaptive at both the gene and SNP levels, thus maintaining high power across a wide range of situations with varying numbers of the genes and SNPs associated with a trait. The proposed method is applicable to both common variants and rare variants and can incorporate biological knowledge on SNPs and genes to boost statistical power. We use extensively simulated data and a WTCCC GWAS dataset to compare our proposal with several existing pathway-based and SNP-set-based tests, demonstrating its promising performance and its potential use in practice.

Introduction

Genome-wide association studies (GWASs) have been successful in identifying many genetic variants, mainly SNPs, associated with complex and common disease (see, for example, the online Catalog of Published Genome-Wide Association Studies). However, only a small proportion of the estimated heritability for most human complex traits can be explained by the identified genetic variants. One possible reason is that, due to small effect sizes and genetic heterogeneity (i.e., multiple causal variants), the standard single-SNP-based analysis might not have enough power to identify many causal variants. Although many human genetic diseases are caused by variants in multiple genes, it has been increasingly recognized that, because genomic variants of these genes lead to the same or similar phenotypes, these genes are likely to be functionally related, and such functional relatedness can be exploited to identify novel genes containing variants related to disease. One way to organize functionally related genes is through biological pathways, such as annotated in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.[1] Association analysis of multiple genes with related functions is here generically called pathway analysis (or gene set analysis), which might improve power over testing on single SNPs or single genes one by one. One convincing source of evidence is from tumor sequencing studies, e.g., The Cancer Genome Atlas (TCGA).[2] Although a few genes (e.g., *TP53* [MIM: 191170]) harbor many mutations related to cancer, most harbor few mutations in a tumor-dependent way. For example, a tumor might contain mutations in *PTEN* (MIM: 601728), not in *NF1* (MIM: 613113), whereas another tumor contains mutations in *NF1*, not in *PTEN*. Individually, each of the genes in a related pathway has only a low mutation frequency, but collectively, they have a much higher mutation frequency. Hence, for a disease (e.g., cancer) involving a few pathways, a pathway analysis by aggregating information across multiple genes in a relevant pathway will boost statistical power, and thus is preferred. For example, among the 316 ovarian cancer (MIM: 167000) tumors studied by TCGA, 45% of them had genomic alterations (somatic mutations and DNA copy-number changes) in the PI3K/RAS signaling pathway. This pathway contains seven genes—*PTEN*, *PIK3CA* (MIM: 171834), *AKT1* (MIM: 164730), *AKT2* (MIM: 164731), *NF1*, *KRAS* (MIM: 190070), and *BRAF* (MIM: 164757)—each with only low to moderate genomic alterations in 7%, 18%, 3%, 6%, 12%, 11%, and 0.5% of the tumors, respectively; hence, it should be more powerful to detect genomic alterations at the pathway level than at the individual gene level.

The importance of pathway analysis and many existing approaches have been reviewed by several authors.[3–5] Many pathway-based analysis methods for GWAS data are evolved from those for gene expression data;[6,7] however, higher-dimensional data are involved in the former with up to hundreds to thousands of SNPs, compared to only tens to hundreds of genes in the latter. On the other hand, because it is known that not all the SNPs in any gene or any pathway are related to a disease, statistically it is most important and challenging to adaptively aggregate information over multiple unknown causal SNPs while minimizing the effects of non-causal SNPs. Existing approaches have some limitations. For example, a popular approach[8] used the minimum p value of the multiple SNPs in a gene to summarize association information for the gene, which is not efficient if there are multiple weakly associated SNPs inside the gene. Two other methods, GATES-Simes[9] and HYST,[10] combine gene-level p values based on GATES,[11] a gene-based test using an extended

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA; [2]Division of Biostatistics and Human Genetics Center, University of Texas School of Public Health, Houston, TX 77030, USA
*Correspondence: weip@biostat.umn.edu (W.P.), peng.wei@uth.tmc.edu (P.W.)
http://dx.doi.org/10.1016/j.ajhg.2015.05.018.