Biostatistics (2022), **0**, 0, *pp.* 1–2
doi:10.1093/biostatistics/multiglmm

Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data

ELIZABETH A. HERON*, COLM O'DUSHLAINE, RICARDO SEGURADO,

LOUISE GALLAGHER, MICHAEL GILL

Neuropsychiatric Genetics Research Group and Department of Psychiatry, Trinity College Dublin, Trinity Centre for Health Sciences, James's Street, Dublin 8, Ireland

eaheron@tcd.ie

Summary

In the analysis of genome-wide association (GWA) data, the aim is to detect statistical associations between single nucleotide polymorphisms (SNPs) and the disease or trait of interest. These SNPs, or the particular regions of the genome they implicate, are then considered for further study. We demonstrate through a comprehensive simulation study that the inclusion of additional, biologically relevant information through a 2-level empirical Bayes hierarchical model framework offers a more robust method of detecting associated SNPs. The empirical Bayes approach is an objective means of analyzing the data without the need for the setting of subjective parameter estimates. This framework gives more stable estimates of effects through a reduction of the variability in the usual effect estimates. We also demonstrate the consequences of including additional information that is not informative and examine power and false-positive rates. We apply the methodology to a number of genome-wide association (GWA) data sets with the

*To whom correspondence should be addressed.

inclusion of additional biological information. Our results agree with previous findings and in the case of one data set (Crohn's disease) suggest an additional region of interest.

Key words: Coronary artery disease; Crohn's disease; Multilevel model; Rheumatoid arthritis; Semi-Bayes; Type 2 diabetes.

1. Introduction

2. Methods

In our development of the methodology for the hierarchical model, we will concentrate on the case–control study design. Here, we present a 2-level EB-HM. The first level comprises a logistic regression model, and the second level comprises a Gaussian regression incorporating additional biologically relevant information.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]