The Jackson Laboratory

Leading the search for tomorrow's cures

Gregory W. Carter
Assistant Professor
207.288.6025 *t* | 207.288.6150 *f* | greg.carter@jax.org

December 28, 2015

Dear Editors,

My co-authors and I submit this presubmission inquiry for our manuscript entitled “A Bayesian framework for Generalized Linear Mixed Models in Genome-Wide Association Studies”, as an article in *Nature Methods*.

Recent technical and methodological advances have greatly expanded genome-wide association studies (GWAS). The advent of low-cost whole-genome sequencing facilitates high-resolution variant identification, and the development of linear mixed models (LMM) allows improved identification of putatively causal variants. While essential for correcting false positive associations due to population stratification, LMMs have been restricted to numerical variables. However, phenotypic traits in association studies are often categorical, coded as binary case-control or ordered variables describing disease stages. Furthermore, optimally integrating the results of prior studies remains a methodological challenge. To address these issues, we have devised a method for genomic association studies that implements a generalized linear mixed model (GLMM) in a Bayesian framework, called Bayes-GLMM.

Bayes-GLMM has four major features: support of multiple phenotypic data types, including categorical variables; cohesive integration of previous GWAS results for related traits by Bayesian modeling; correction for sample relatedness by mixed modeling; and model estimation by both MCMC sampling and maximal likelihood estimation. For computing efficiency with very large data sets such as whole-genome sequenced populations, the Bayes-GLMM method was implemented in the Stan programming environment. This facilitated the following optimizations: conjugate prior distributions; vectorization of model statements to take advantage of the efficient matrix operations in Stan; parallel computing; and reparameterizion of multivariate normal distributions by Cholesky factoring.

To demonstrate our method, we applied Bayes-GLMM to the whole-genome sequencing cohort in the Alzheimer’s Disease Sequencing Project (ADSP). This study contains 576 individuals distributed across 111 families, each with Alzheimer’s disease diagnosed at four confidence levels. The profound population structure in these data required a mixed model approach, and the categorical trait necessitated a generalized model.

In summary, this work provides the first implementation of a flexible, generalized mixed model approach in a Bayesian framework. We are confident that it will be of broad interest to the genetics and genomics communities. Thank you for your considering this inquiry.

Sincerely,

Gregory W. Carter, PhD