December 23, 2015

Dear Editors,

My co-authors and I submit a presubmission inquiry for our manuscript entitled “A Bayesian approach of Generalized Linear Mixed Models for Genome-Wide Association studies”, as a research article in *Nature Methods*.

Recent technical and methodological advances have greatly expanded the 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 stratifications and sample relatedness, 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. Further, while common diseases were studied in multiple GWAS, to integrate previous results of related traits remains an open question in terms of methodology. To address these issues, we built a method, Bayes-GWAS, for genomic association studies that implemented generalized linear mixed model (GLMM) in the Bayesian framework.

Our method comes with four major features: 1, support multiple phenotypic data type; 2, integrate previous GWAS results on the related traits cohesively by Bayesian modeling; 3, implement sample relatedness correction by mixed models, and 4, support model estimations by both MCMC sampling and maximal likelihood estimation.

For the sake of computing efficiency, Bayes-GWAS method was implemented in the Stan programming environment, and optimized in the following ways: (1) conjugate prior distributions, (2) vectorization of model statements to take advantage of the efficient matrix operations in Stan, (3) parallel computing, (4) reparameterize multivariate normal distribution by Cholesky factoring.

To demonstrate our method in the GWAS context, we applied it 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 short, this work provides the first implementation of a flexible, generalized mixed model approach in the Bayesian framework.

Thank you for your considering this inquiry.

Sincerely,



Gregory W. Carter, PhD