Harvard University
Harvard University Biostatistics Working Paper Series

Year 2012 *Paper* 147

C2BAT: A Novel Method for Association Between Ge- netic Markers and Multiple Phenotypes

Melissa Naylor* Christoph Lange†

*

†Harvard School of Public Health, clange@hsph.harvard.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper147

Copyright ©2012 by the authors.

C2BAT: A Novel Method for Association Between Ge- netic Markers and Multiple Phenotypes

Melissa Naylor and Christoph Lange

Abstract

The purpose of this technical report is to describe a novel method developed to detect association between a genetic marker and multiple phenotypes. In order to obtain a one-degree of freedom test, a generalized principal component approach is suggested that aggregates the information about the genetic effect in the ?rst prin- cipal component, while the remain principal components contain only environment noise. A limited simulation study is done validating the method. For scenarios in which the genetic effect is constant across all measurements and there is no envi- ronmental correlation between the measurements, preliminary results suggest that this method has similar power to standard methods of analysis. Since such setups favor the standard approach, the new method is expected to be ef?cient for many realistic alternative hypotheses, including those with varying genetic effect sizes across measurements and those accounting for environmental correlation.

C2BAT: A Novel Method for Association Between Genetic Markers and Multiple Phenotypes

Melissa G. Naylor and Christoph Lange

Pritzker School of Medicine, University of Chicago, Chicago, Illinois, USA
Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA

1 Introduction

The purpose of this technical report is to describe a novel method developed to detect association between a genetic marker and multiple phenotypes. In order to obtain a one-degree of freedom test, a generalized principal component approach is suggested that aggregates the information about the genetic effect in the first principal component, while the remain principal components contain only environment noise. A limited simulation study is done validating the method. For scenarios in which the genetic effect is constant across all measurements and there is no environmental correlation between the measurements, preliminary results suggest that this method has similar power to standard methods of analysis. Since such setups favor the standard approach, the new method is expected to be efficient for many realistic alternative hypotheses, including those with varying genetic effect sizes across measurements and those accounting for environmental correlation.

2 Methods

Several phenotypes are tested simultaneously for association with one genetic marker. Here we consider a single SNP with genotypes 0, 1, and 2. First, the data is split into two smaller datasets (A and B), each containing a roughly half of the total subjects with their corresponding genotypes. Using only Dataset A, a MANOVA is performed to test whether the phenotypes differ across genotypes, and Wilks' test is used to obtain a p-value, p_A. Dataset A is also used to perform a multivariate regression of the phenotypes on genotype in order to estimate the regression coefficients. The estimated coefficients from regression analysis of dataset A are then used to calculate a linear combination of the phenotypes in dataset B. Thus, information derived from dataset A is used to create a combined phenotype. The combined phenotype is regressed on genotype, producing a second p-value, p_B. The two p-values, p_A and p_B are then combined into an overall p-value using Fisher's method. Since

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

the resulting p-value can be highly dependent on how the data was initially split into datasets A and B, we repeat all of the above steps multiple times and take the median of the resulting p-values. Note that the median of the p-values is not a valid p-value itself.

In order to obtain a valid p-value for the entire method, we use permutation tests, in which the genotypes are randomly reassigned. The permutation p-value is the proportion of times the observed median is less than the median observed from the unpermuted data. The permutation test also allows us to relax assumptions. For example, when regressing the combined phenotype on genotype in the second stage of the analysis, we assume genotype is a continuous variable even though it is categorical. By using permutations we can be sure that our method is valid despite this.

3 Simulations

To examine the power of this method, we compare the power of our method to that of a standard MANOVA and MANOVA with permutation p-values. For each simulation, we generated data from 100 subjects, each with a genotype from a SNP with allele frequency of 0.1 and three continuous phenotypes. Two of the phenotypes are associated with the genotype and one is not. Both of the associated phenotypes had a heritability of 0.025, 0.05, 0.075, or 0.1. Simulated data in which no phenotypes were associated with the SNP were used to verify the validity of the method, i.e., that the type 1 error is 0.05.

For each simulated dataset, we obtained three p-values: from MANOVA, from MANOVA with permutations, and from C2BAT with permutations. When applying the C2BAT method to simulated data, the median was taken after splitting the data 10 times. All permutation p-values were obtained by randomly permuting the genotypes 1000 times and reporting the fraction of times the p-values were less than the p-value obtained from the orginal, non-permutated data. Power was calculated as the fraction of times the p-value was less than 0.05 when 1000 sets of simulated data were analyzed.

Figure 1 depicts the results of the simulation study. The power estimates plotted are also displayed in table below.

Heritability	C2BAT (permuations)	MANOVA (permutatons)	MANOVA
0.0	0.054	0.048	0.051
0.0025	0.436	0.442	0.446
0.005	0.734	0.748	0.749
0.075	0.913	0.916	0.915
0.01	0.964	0.969	0.970

Collection of Biostatistics
Research Archive

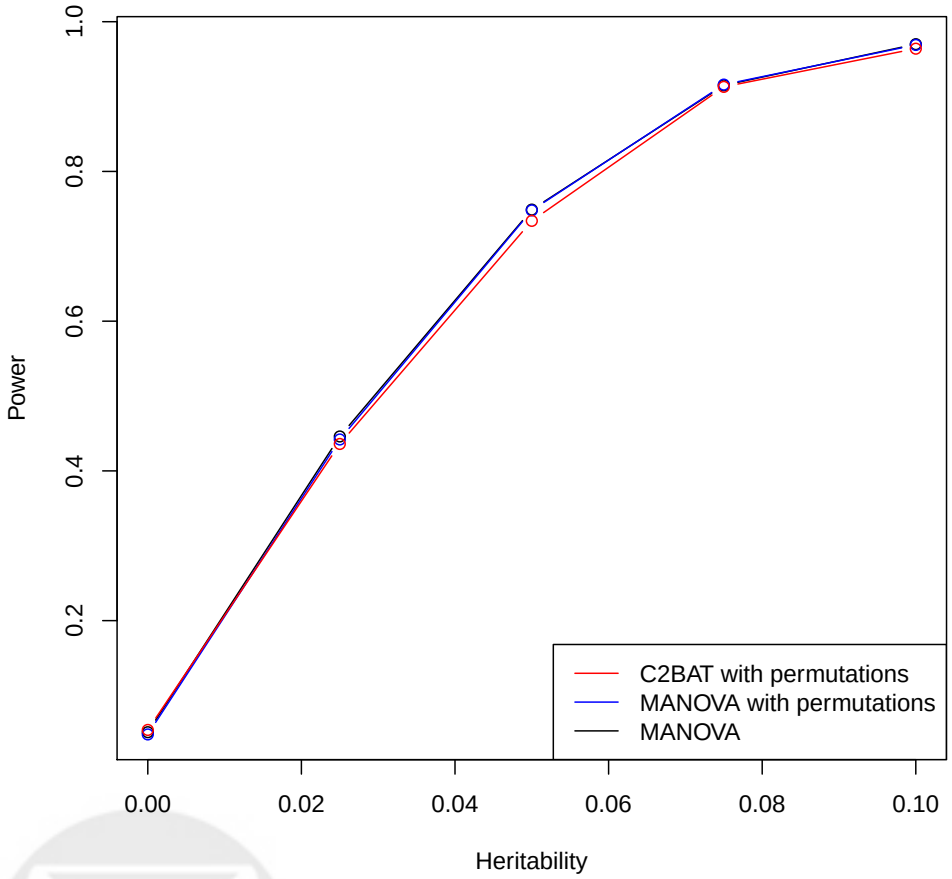

Figure 1: Estimated power of C2BAT (red) as compared to MANOVA with (blue) and without (black) permutations. The x-axis refers to the heritability of each of two associated phenotypes. (The third phenotype is not associated with genotype.)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

4 Conclusion

C2BAT is a novel method for detecting association between genetic markers and multiple phenotypes. The limited simulation study presented here shows that the method is valid (maintains a type one error of 0.05 under the null hypothesis) and has power nearly equivalent to a standard MANOVA. Permutation p-values had negligible effect on the power of MANOVA, but are a necessary feature of the C2BAT method. Since the simulation study was performed under the assumption of constant genetic effect sizes across measurements and in the absence of environmental correlation, the simulation studies do not cater to the strengths of C2BAT. There are many aspects of this method that could be varied. For example, the proportion of data allocated to dataset A versus dataset B could be optimized or other methods for combing p-values could be used (instead of Fisher's). Additional work to refine this method or explore its power in different settings is warranted.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive