Detecting Gene-by-Environment Interactions in Coronary Artery Disease

Boxiang Liu[1,2] {bliu2@stanford.edu}, Robert S. Kirby[3] {rskirby2@stanford.edu}

Department of Biology[1], Statistics[2] and Electrical Engineering[3], Stanford University

Abstract

Coronary artery disease is one of the leading causes of mortality, resulting in 1 in 4 deaths in the U.S. It is a complex disease influenced by both genetic and environmental risk factors. Numerous genome-wide association and epidemiological studies have investigated the genetic and environmental risk factors separately, but few have studied their interactions. In this paper we use novel statistical methods to jointly analyze genetic and environmental risk factors. Leveraging recent methodological advances in expression quantitative trait loci studies, our method controls false discovery rate while maximizing power. Further, we use simulation to provide a guideline on the minimum sample size for future gene-by-environment studies.

1. Introduction

As one of the leading causes of mortality, coronary artery disease (CAD) is a complex disease influenced by both genetic and environmental risk factors. Numerous genome-wide association and epidemiological studies have shown that environmental factors such as exercise and consumption of fatty food, and genetic makeup both influence the risk of developing CAD. However, little is known about the interactions between genetic and environmental risk factors. In this study, we aim to discover genes and genetic variants that interact with artery inflammation. We use coronary artery smooth muscle cell (CASMC), constituting more than 80% of atherosclerotic plaque mass, to model coronary arterial tissue. CASMCs switch between a synthetic phenotype when stimulated with serum (disease state) and a contractile phenotype when resting (healthy state). We therefore use serum media and serum-free media to model inflamed and normal CASMCs. We use generalized linear models to detect interaction effects. The inputs to our model are two vectors of integers representing total expression and allele-specific expression (either maternal or paternal) of any given gene. In addition, the model requires a vector of binary environmental factors. The model uses a likelihood ratio test to output a binary prediction, which takes value 1 if the given gene interacts with serum stimulation.

2. Related Work

Two methods exist to detect gene-by-environment interactions. Knowles et al. proposed a generalized linear mixed model framework to detect interactions with both discrete and continuous environmental factors[1]. However, since effect sizes for interaction effect are usually smaller than main effects, the model is underpowered with small sample sizes. In fact, the proposed procedure was tested on 900 samples to ensure enough power in detection of moderate interaction effects. An alternative approach is to stratify samples by environmental conditions and analyze each condition separately[2]-[4]. Expression quantitative trait loci (eQTLs, fig. 1) are mapped using total expression information in each condition separately, and the difference between two sets of eQTLs is considered environmentally interactive. This model is well powered for moderate sample sizes (>=60), but it is only able to model binary environmental factors. Moreover, both methods seem to be underpowered for 21 samples we obtained for this study (section 3). A previous attempt used Fisher's exact test on the gene level to increase power (unpublished data). This method successfully detected ~50 environmentally interactive genes, but lost the ability to detect associated interaction eQTLs. In this study, we leverage recent methodological advances in eQTL mapping to jointly models allele-specific

expression and total expression to increase power to detect both genes and genetic variants that interact with serum stimulation.

3. Dataset and preprocessing

We have obtained 21 CASMC samples collected under two environmental conditions. Eleven samples are cultured in serum-free media (SF) to emulate the contractile phenotype. Another ten samples are treated with fetal-bovine serum (FBS) to induce the synthetic phenotype. We performed RNA sequencing (RNAseq) and whole-genome sequencing (WGS) on all samples. We aligned RNAseq data to the human reference genome v19 using STAR, and WGS data using BWA. We called WGS variants using iSAAC and quantified total expression using HTSeq and allele-specific expression using custom scripts. We correct known (age, sex, ancestry, batch) and latent covariate of total expression using PEER. Self-reported ancestry is confirmed using PCA (figure not shown). We imputed missing variants and phased haplotypes using impute2 with 1000 Genomes reference panel. For each gene-variant pair, the input to our model is a 21-by-4 matrix. Each row of the matrix represents a sample, and the 4 columns are paternal allele read count, maternal allele read count, environment variable, and test variant genotype (fig. 1). Since ~3000 genes are testable, and on average each gene associates with 1000 test SNPs, there are ~3 million matrices in total. We do not have a test set because there is no 'ground truth'. However, our statistical test controls the false discovery rate to 0.05.

4. Methods

4.1 Binomial generalized linear model

We model the allele-specific count given the total count as a binomial random variable. This modeling choice naturally leads to a binomial generalized linear model. The binomial distribution is a member of the exponential family.

$$p(y \mid n;\eta) = b(y)(\exp(\eta T(y) - a(\eta)))$$

Where $b(y) = \frac{n!}{y!(n-y)!}$, $\eta = \log\frac{p}{1-p}$, $a(\eta) = -n\log(1-p)$, and $T(y) = y$. From now on we use σ to represent the logistic function.

4.2 BGLM to detect gene-by-environment interaction

Our null hypothesis states that environment factors do not influence allele-specific expression. The ratio between allele-specific read count and the total read count is captured by an intercept μ.

$$H_0: y \mid n;\mu \sim Binom(n, \sigma(\mu))$$

Our alternative hypothesis states that environmental factors will influence the ratio, or p = $\sigma(\mu + \beta^e e)$.

$$H_0: y \mid n;\mu \sim Binom(n, \sigma(\mu + \beta^e e))$$

In this model, y denotes allele-specific expression (arbitrarily chosen as the reference allele), and n denotes the total expression. The intercept μ accounts for mapping bias and other unknown global factors, and e represents the environmental variable. We use likelihood ratios to test the significance of the environmental term.

$$\Lambda = -2\log(\frac{\sup\{L(\beta^e,\mu): \beta^e = 0\}}{\sup\{L(\beta^e,\mu): \beta^e \in \mathbb{R}\}})$$

Asymptotically, Λ follows a χ_1^2 distribution, from which we obtain the p-values.

4.3 Generalized linear mixed model (GLMM)

The generalized linear mixed model is an extension of the generalized linear model. In this model, we decompose the mean effect into a fixed component μ and a random component ε.

$$H_0: y \mid n; \mu \sim Binom(n, \sigma(\mu + \varepsilon))$$

$$H_0: y \mid n; \mu \sim Binom(n, \sigma(\mu + \beta^e e + \varepsilon))$$

We model the random effect as a Gaussian random variable.

$$\varepsilon \mid v \sim N(0, v); v \sim IG(a, b)$$

The variance of the random effect is chosen to be conjugate distribution inverse Gamma. The hyper-parameters a and b are estimated by pooling information across all genes.

4.4 Linear model to test variant-by-environment interaction

Interactions between genes and environment are usually driven by genetic variants. We use a linear model to test for significant variant-environment interactions.

$$H_0: n = \beta^g g + \beta^e e + \mu + \varepsilon$$

$$H_1: n = \beta^g g + \beta^e e + \beta^{g \times e} ge + \mu + \varepsilon$$

In the above models, n is the total expression, and g and e are the genetic and environmental influences on gene expression, respectively. The null hypothesis states that the genetic and environmental effects are additive, whereas the alternative hypothesis states that they can be multiplicative. We again use likelihood ratios to test the significance of the interaction term.

4.5 Jointly model total expression and allele-specific expression to increase power

We use the software package WASP to map eQTLs[5]. WASP gains power by combining total expression and allele-specific expression. In brief, WASP models the total expression using a BetaNegativeBinomial distribution and allele-specific expression using BetaBinomial distributions. BetaNegativeBinomial is an overdispersed Poisson with two hyper-parameters, and BetaBinomial is an overdispersed Binomial distribution with two hyper-parameters. The use of overdispersed distributions effectively controls type I error rate.

4.6 Simulation

To quantify the number of samples needed to identify true positives, we simulated total (n) and allele-specific (y) read counts using overdispersed Poisson and Binomial models, respectively. All hyper-parameters are estimated from real data using the generalized linear mixed model.

$$n \mid \lambda \sim Pois(\lambda); \lambda \sim Gamma(2,2)$$

$$y \mid n, p \sim Binomial(n, p); p = \sigma(\beta^g g + \beta^e e + \beta^{g \times e} ge + \varepsilon); \varepsilon \sim N(0, IG(1.13, 0.0122))$$

5. Results

5.1 Interaction testing is underpowered for small sample size

We restrict our analysis to chromosome 22 to reduce computational time. After filtering for bi-allelic loci, we found 1798 loci that have at least 1 heterozygous individual. After further filtering for loci with larger than 10 heterozygous samples, we are left with 100 testable loci. We performed likelihood ratio tests using the binomial generalized linear model described in section 4.1 for all testable loci and observed marked enrichment towards 0, indicating presence of environmentally

interactive genes (figure not shown). However, we observed patterns of overdispersion when comparing GLM with permutation test p-values. As shown in fig. 2, the red points lying above the diagonal is indicative of deflation of likelihood ratio p-values and increased type I error rate. We think the overdispersion is caused by latent data structure such as PCR duplicates during the sequencing process. Adding a random effect ε corrected for overdispersion (section 4.2). In fig. 2, the black points lying below the diagonal indicate conservative estimates of p-values of the mixed model. Under the null assumption, the p-value distribution is uniform so we compared GLMM p-values to a uniform distribution (fig. 3). The black points lying close to the diagonal suggests little evidence for presence of environmentally interactive genes. Further, the black horizontal line corresponds to p-values of 1, which is typical of an underpowered study. We hypothesize that setting total read depth threshold would prioritize testing of significance genes. Indeed, the number of 1's decreased as the threshold increased. However, the study is nonetheless underpowered as clearly shown in fig. 3. For completeness, we carried out interaction eQTL (iQTL) mapping on the list of genes with nominal GLMM p-value less than 0.05 using a linear model (section 4.4). Not surprisingly, we did not discover significant iQTLs after multiple hypothesis correction (figure not shown).

5.1 Simulation reveals GLMM requires larger sample size

We hypothesize that increasing the number of samples will salvage the power issue. To test this hypothesis, we performed simulation (section 4.6) using 4 different sample sizes and 5 different interaction effect sizes. With 20 samples, the true positive rate is below 5% even for a large effect size of 0.5 (controlling type I error rate at 0.05). On the other hand, a large sample size of 500 can recover more than 70% of true positives for medium effect size of 0.3 (Fig. 4). We conclude that binomial generalized linear mixed model is underpowered for our study.

5.2 Joint modeling of total and allele-specific counts improves power

A recent study shows that joint modeling of total and allele-specific expression can improve power to detect eQTLs (section 4.5)[5]. We adopted this strategy to map eQTLs for two environmental conditions separately. Concordant with their observation, joint modeling is well powered for small sample sizes of 10 and 11 (Fig. 5). Interaction eQTLs are subsequently defined using a two-step procedure. In step 1, we obtain a list of significant eQTLs in our condition of interest using a stringent FDR of 1%. In step 2, we obtain a list of eQTLs for the other condition using a lenient FDR of 50%. The difference between two sets is taken as interaction eQTLs for the condition of interest. This two-step procedure is conservative as only eQTL with significant difference between two conditions are selected. Figure 6 shows an example of iQTL chr22:42913295. This SNP regulates RRP7A in opposite directions under two environmental conditions. Other genes found includes THOC5, a regulator of cholesterol metabolism and has been implicated in cardiovascular diseases[6], and A4GALT, another metabolic regulator that has been previously associated with obesity and heart diseases[7].

6. Conclusion and future work

Leveraging both total and allele-specific counts, we discovered several environmentally interactive variants and their associated genes despite small sample sizes. Currently, we only analyzed chromosome 22 for the sake of computational efficiency; we hope to scale up to the entire human genome. In addition, we will compare iQTLs with relevant GWAS variants to test for enrichment in overlapping. Last but not least, we will perform weighted tests by incorporating epigenetic information obtained from ATAC-seq assays. The result of this project shows promises for interaction eQTL testing in a small cohort.

7. Figures

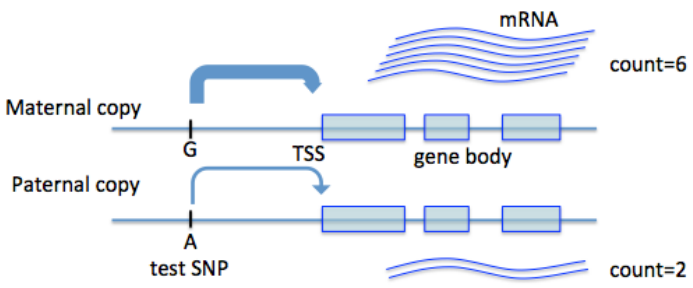

Fig. 1 An eQTL diagram. Test SNPs G and A at eQTL locus differentially regulate downstream genes

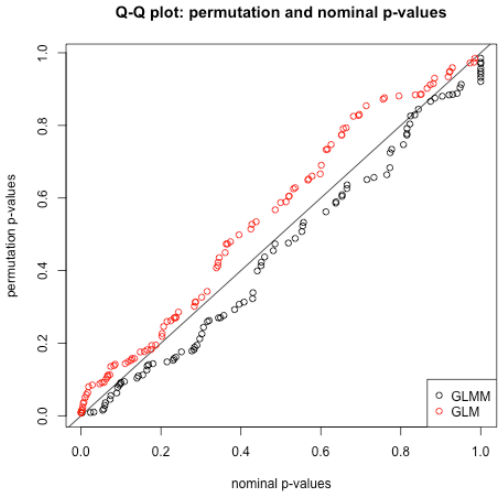

Fig. 2 Quantile-quantile plot comparing permutation and parametric test p-values

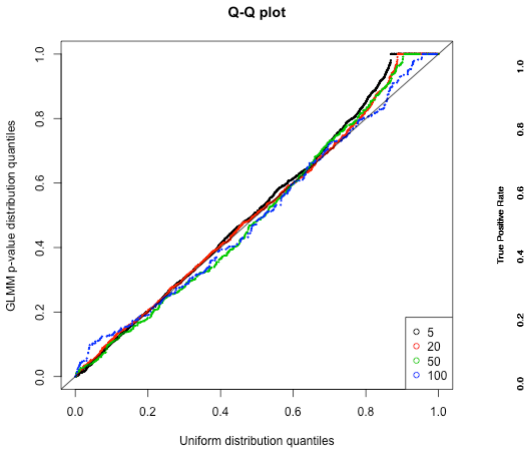

Fig. 3 Quantile-quantile plot comparing GLMM p-values to a normal distribution

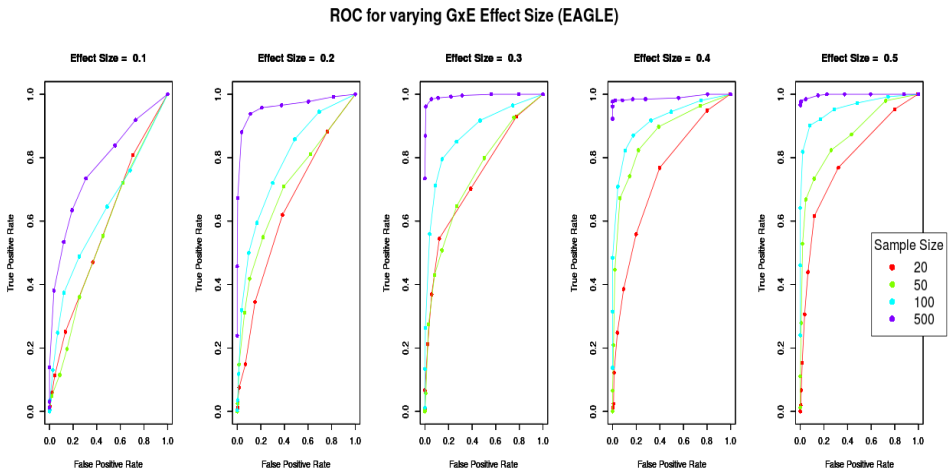

Fig. 4 ROC curve for GLMM using simulated data, faceted by interaction effect size

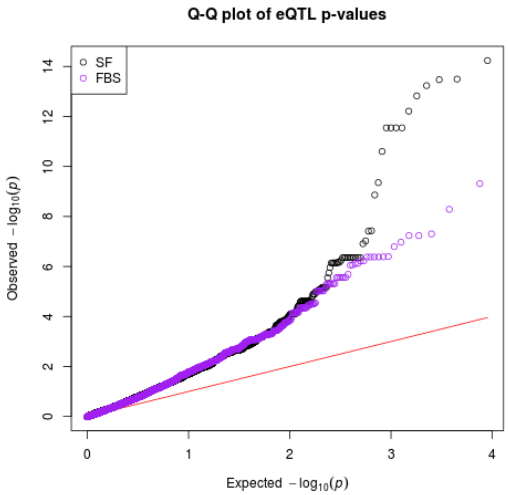

Fig. 5 Quantile-quantile plot compare WASP p-values with uniform distribution

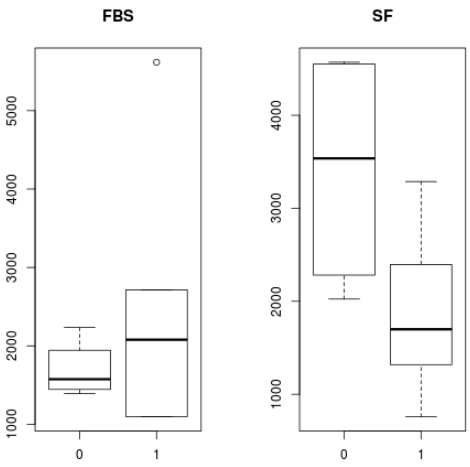

Fig. 6 An example of variant-by-environment interaction

[1] D. A. Knowles, J. R. Davis, A. Raj, X. Zhu, J. B. Potash, M. M. Weissman, J. Shi, D. Levinson, S. Mostafavi, S. B. Montgomery, and A. Battle, "Allele-specific expression reveals interactions between genetic variation and environment," bioRxiv, p. 025874, Sep. 2015.

[2] M. N. Lee, C. Ye, A.-C. Villani, T. Raj, W. Li, T. M. Eisenhaure, S. H. Imboywa, P. I. Chipendo, F. A. Ran, K. Slowikowski, L. D. Ward, K. Raddassi, C. McCabe, M. H. Lee, I. Y. Frohlich, D. A. Hafler, M. Kellis, S. Raychaudhuri, F. Zhang, B. E. Stranger, C. O. Benoist, P. L. De Jager, A. Regev, and N. Hacohen, "Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells," Science, vol. 343, no. 6175, pp. 1246980–1246980, Mar. 2014.

[3] B. P. Fairfax, P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, C. McGee, and J. C. Knight, "Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression," Science, vol. 343, no. 6175, pp. 1246949–1246949, Mar. 2014.

[4] L. B. Barreiro, L. Tailleux, A. A. Pai, B. Gicquel, J. C. Marioni, and Y. Gilad, "Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection.," Proc. Natl. Acad. Sci. U.S.A., vol. 109, no. 4, pp. 1204–1209, Jan. 2012.

[5] B. van de Geijn, G. McVicker, Y. Gilad, and J. Pritchard, "WASP: allele-specific software for robust discovery of molecular quantitative trait loci," bioRxiv, 2014.

[6] M. Keller, D. Schleinitz, J. Förster, A. Tönjes, Y. Böttcher, A. Fischer-Rosinsky, J. Breitfeld, K. Weidle, N. W. Rayner, R. Burkhardt, B. Enigk, I. Müller, J. Halbritter, M. Koriath, A. Pfeiffer, K. Krohn, L. Groop, J. Spranger, M. Stumvoll, and P. Kovacs, "THOC5: a novel gene involved in HDL-cholesterol metabolism," Journal of Lipid Research, vol. 54, no. 11, pp. 3170–3176, Nov. 2013.

[7] T. Hu, C. Darabos, M. E. Cricco, E. Kong, and J. H. Moore, "Genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks.," Pac Symp Biocomput, pp. 207–218, 2015.