Genetic association analysis of coronary heart disease by profiling gene-environment interaction based on latent components in longitudinal endophenotypes

C Charles Gu  
Washington University School of Medicine in St. Louis

Wei (Will) Yang  
Washington University School of Medicine in St. Louis

Aldi T. Kraja  
Washington University School of Medicine in St. Louis

Lisa de las Fuentes  
Washington University School of Medicine in St. Louis

Victor G. Davila-Roman  
Washington University School of Medicine in St. Louis

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Part of the Medicine and Health Sciences Commons

Recommended Citation
https://digitalcommons.wustl.edu/open_access_pubs/329

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
**Genetic association analysis of coronary heart disease by profiling gene-environment interaction based on latent components in longitudinal endophenotypes**

C Charles Gu*1,2, Wei (Will) Yang1, Aldi T Kraja3, Lisa de las Fuentes4 and Victor G Dávila-Román4

Addresses: 1Division of Biostatistics, Washington University School of Medicine, 660 South Euclid Avenue, Box 8067, St. Louis, Missouri 63110, USA, 2Department of Genetics, Washington University School of Medicine, 660 South Euclid Avenue, Box 8067, St. Louis, Missouri 63110-1093, USA, 3Department of Statistical Genomics, Department of Genetics, 4444 Forest Park Boulevard, Box 8506, St. Louis, MO 63108, USA and 4Cardiovascular Imaging and Clinical Research Core Laboratory, Cardiovascular Division, Department of Medicine, Washington University School of Medicine, 660 South Euclid Avenue, Campus Box 8086, St. Louis, Missouri 63110, USA

E-mail: C Charles Gu* - gc@wubios.wustl.edu; Wei (Will) Yang - will@wubios.wustl.edu; Aldi T Kraja - aldi@dsgmail.wustl.edu; Lisa de las Fuentes - lfuentes@wustl.edu; Victor G Dávila-Román - vdamila@wustl.edu

*Corresponding author

from Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009


This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S86

© 2009 Gu et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract**

Studies of complex diseases collect panels of disease-related traits, also known as secondary phenotypes or endophenotypes. They reflect intermediate responses to environment exposures, and as such, are likely to contain hidden information of gene-environment (G × E) interactions. The information can be extracted and used in genetic association studies via latent-components analysis. We present such a method that extracts G × E information in longitudinal data of endophenotypes, and apply the method to repeated measures of multiple phenotypes related to coronary heart disease in Genetic Analysis Workshop 16 Problem 2. The new method identified many genes, including SCN1B (sodium channel nonvoltage-gated 1 beta) and PKP2 (plakophilin 2), with potential time-dependent G × E interactions; and several others including a novel cardiac-specific kinase gene (TNNI3K), with potential G × E interactions independent of time and marginal effects.

**Background**

"Endophenotypes" refer to the host of measurements representing physiologic indicators, biochemical assays, and responses to challenges, or the latent components extracted from such data [1]. When derived properly, the latent traits lay more proximal to the causal genotypes than do clinical phenotypes, and thus, provide potentially meaningful but otherwise unobserved context of gene-environment (G × E) interaction. Several recent studies report positive findings with endophenotypes in genetic analysis of complex diseases [2,3]. Our group recently developed a supervised statistical learning...
approach for multivariate analysis (SLAM) that uses latent component methods to extract meaningful latent traits for association studies. We have applied this method to the study of hypertension and hypertensive heart disease [4]. The method worked well to identify meaningful latent traits of hypertensive heart disease that led to detection of significant genotype × phenotype associations that were missed by analyses of measured clinical phenotypes.

The repeated measures of multiple coronary heart disease (CHD)-related phenotypes from Genetic Analysis Workshop (GAW) 16 Problem 2 are ideal for testing this approach to identify genetic variants that interact with the environment in the development of CHD. For example, systolic and diastolic blood pressure represent continuous and independent risk factors for CHD events [5], and increased blood pressure was associated with pathologic remodeling of the left ventricle [6]. Studies have identified dyslipidemia as a major cause of CHD; therapies that lower serum low-density lipoprotein cholesterol reduce CHD risk [7]. Indeed, metabolic syndrome, which represents a constellation of major risk factors including abdominal obesity, atherogenic dyslipidemia, elevated blood pressure, insulin resistance, and prothrombotic and proinflammatory states, has been shown to increase risk of CHD [8]. Repeated measures on these CHD endophenotypes and their environmental risks (e.g., cigarette smoking) contain valuable information about underlying mechanisms of G × E interactions that is biologically relevant to the development and/or modulation of CHD.

In the present study, we explore such underlying mechanisms using latent components (referred to as “G × E context”) extracted by an extension of the SLAM approach to analyze the multivariate longitudinal data in GAW 16 Problem 2.

**Methods**

**Data adjustment and quality control**

Samples in the “Offspring Cohort” (of the Framingham Heart Study) and data from Visits 1, 3, 5, and 7 were used in this study. The primary phenotype of interest was CHD event, and the data on ten variables including CHD endophenotypes (body mass index, three lipids, blood pressures, and glucose) and environmental covariates (age at visit, cigarette smoking, and alcohol use) were used for latent component analysis. The endophenotypes were checked for normality and outliers, and log-transformed when necessary; this was followed by centering the variables by sex to remove unwanted confounding. The residuals were used as input for all downstream analyses. We used the genome-wide dense single-nucleotide polymorphisms (SNPs) dataset provided for Problem 2 (~550,000 SNPs typed by Affymetrix GeneChip® Human Mapping 500 k Array Set). Quality of each SNP array was checked first for low call rate (<95%) and/or abnormal heterozygosity (<0.25 or >0.3); then each SNP on the array was checked for its minor allele frequency (MAF < 0.05), missing rate in the sample (>5%), and deviation from Hardy-Weinberg equilibrium (p < 10^-6). Problematic individuals and SNPs were moved from further analysis.

**Identifying latent G × E context**

Latent component analysis (also known as factor analysis) aims to effectively reduce the number of dimensions (variables) for analysis while minimizing the loss of information. Conventional factor analysis, of which the well-known principal-component analysis is a special case, seeks to reduce the dimensionality by expressing the original variables as linear combinations of a smaller number of independent, Gaussian, latent variables (components). However, it tends to neglect meaningful structural information such as clustering in data, which often requires non-Gaussian components and proper treatment of higher order of moments than covariance and correlations. The method of independent-component analysis (ICA) overcomes this problem by treating observed traits as a mixture of underlying components that are more likely independent, non-Gaussian, and with less complexity than observed ones. It identifies such latent components by maximizing a measure of multivariate non-Gaussianity of linear combinations of original variables [9]. The SLAM approach is built on ICA using supervised validation of extracted components followed by consensus analysis of validated components to identify robust and biologically-meaningful latent traits [4]. For the present study, we extend SLAM by applying multi-level ICA to facilitate analysis of longitudinal multivariate data.

**Time-dependent longitudinal G × E context**

We first applied a four-component ICA at each visit on the ten selected variables to identify latent components that define potentially meaningful underlying G × E context for CHD. Then, correlations between the derived independent components (ICs) at consecutive visits were examined and those with strongest correlations were concatenated to form four longitudinal latent components (LLCs). We hypothesized that each (most) of the derived LLCs represents a particular G × E mechanism (context) for the development of the disease, and can be used as a derived “environment” variable for teasing out potential G × E interactions. This was verified by logistic regression of CHD on each extracted LLC at every time point (adjusted for age at visit) to evaluate LLCs as a
predictor of CHD, followed by regression of each LLC on genome-wide SNP genotypes to assess its genetic content.

Summary time-independent G × E context
A second-level two-component ICA was then performed on each LLC over the four time points (i.e., visits). For each LLC, we anticipated that one such derived component will capture the main signal of the time-course, while the other absorbs remaining signal and noise. Identification of the signal component was assisted by a clinical expert knowledgeable in the component’s capacity in predicting CHD risks. In the end, this procedure derives time-independent components (TICs) to capture the time course of G × E interactions in the context defined by each LLC. Note that familial relationship is ignored during ICA extraction and is accounted for in the downstream association analysis.

Profiling G × E interactions
The identified latent components provide different G × E context useful for teasing out potential G × E interactions. The LLCs have repeated measures at each visit, and can be used to facilitate a focused search for SNPs with potential time-dependent G × E interactions. The LLCs have repeated measures at each visit, and can be used to facilitate a focused search for SNPs with potential time-dependent G × E interactions. Logistic regression of CHD status was carried out first against SNP genotypes and LLC values, then by an expanded model including the term for SNP × LLC interaction. These analyses were done for each visit, and results were aggregated to spot trends of time-dependent interactions. Finally, the TICs represent summary profiling of time-course of G × E interactions for CHD. The detection of SNPs with potential time-independent G × E interactions for CHD was then achieved by testing for significant SNP × TIC interactions using logistic regression. In all regression analyses, the generalized estimating equation approach was used to adjust for correlations among family members in the sample.

Results
Samples of 2584 individuals in the “Offspring Cohort” were used in this study. After removing samples missing at least 1 visit (n=403) and those without GWAS data (n=187), we performed genotype quality control and excluded 33 samples due to low call rate (n=22), abnormal heterozygosity (n=5) and population outliers (n=6). A total of 1961 individuals were used in the analyses reported below. To achieve normality, log-transformations were applied to triglyceride, blood glucose, cigarettes smoked per day, and alcohol use per week.

Time-dependent G × E interactions
Repeated measures from Visits 1, 3, 5, and 7 were used to extract time-dependent G × E context, the LLCs for CHD. The extended SLAM procedure was used to extract four LLCs as described previously.

We first tested each LLC as a predictor of CHD. Table 1 shows the results by logistic regression of CHD on LLC at each visit. With the exception of LLC1, the derived components LLC2-LLC4 were all significant predictors of CHD. Both LLC2 and LLC3 are highly significant predictors of CHD at the early two visits, while LLC2 was also fairly significant at Visit 7. LLC4 seemed to have captured a complementary axis that became significant predictor of CHD at later visits when the average age of the Offspring Cohort reaches 53 to 60 years old. Note that LLC1 absorbs remaining variation in the data, and may still contain some G × E information (as confirmed below). We then examined genetic association of SNPs with each LLC by longitudinal regression. There were a number of SNPs with p ≤ 0.05, but few achieved high significance. Most notable ones were all associated with LLC3 (21 SNPs in 9 genes with p ≤ 1 × 10⁻³, data not shown). These results supported the idea that the LLCs may be used as a derived “environment” variable for teasing out potential G × E interactions.

We then tested for SNP × LLC interactions in an expanded model including the interaction terms (CHD=age+LLC+SNP+SNP*LLC). A total of 76 SNPs in 59 genes were found having 96 significant interactions with various LLCs at different time points, at a significance level of α = 1 × 10⁻⁶. A majority of these SNP × LLC interactions (63/96 = 66%) were detected in the middle two visits, and close to half were interacting with LLC1. Many of the genes detected are relevant to CHD. In Table 2, we displayed some representatives in details. Some of them, including SCNN1B and PKP2, are well known candidate genes of CHD.

Time-independent G × E Interactions
As described previously, we then performed second-level ICA on each LLC to extract the component that captures main signal of the time-course underlying the LLC. These components are denoted TIC1 to TIC4. Note that we

<table>
<thead>
<tr>
<th>LLC</th>
<th>Visit 1</th>
<th>Visit 3</th>
<th>Visit 5</th>
<th>Visit 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.093742</td>
<td>0.128905</td>
<td>0.189618</td>
<td>0.073004</td>
</tr>
<tr>
<td>2</td>
<td>0.000056*</td>
<td>0.000023</td>
<td>0.375683</td>
<td>0.010654</td>
</tr>
<tr>
<td>3</td>
<td>0.000115</td>
<td>0.007251</td>
<td>0.140016</td>
<td>0.466008</td>
</tr>
<tr>
<td>4</td>
<td>0.333980</td>
<td>0.040160</td>
<td>0.000080</td>
<td>0.007423</td>
</tr>
</tbody>
</table>

*Bold font indicates p-values ≤ 0.05.
include LLC1 in the second-level ICA analysis, assuming that there may still be G × E information hidden in this "noise" component. Results of validation analyses of the derived TICs are shown in Figure 1, where their capacity in predicting CHD risks are evident in eight clinical indicators of CHD including blood pressures, lipids, blood sugar, body mass index, and age. It is interesting to note that both of the two components (denoted as TIC3-1 and TIC3-2 in Figure 1) derived from LLC3 may qualify as a "signal" component, although TIC3-2 did better in predicting CHD events.

To study effects of potential interactions between SNPs and so-derived TICs, we carried out logistic regression of CHD status on SNP genotypes and an interaction term for SNP × TIC, using two types of models: ones that include main effect of TICs and ones without. We hypothesized that SNPs for which the two models produced similar significance levels imply possibility of "pure" SNP × TIC effects. At a significance level of \( \alpha = 10^{-5} \), we displayed in Table 3 genes containing such SNPs, for which the inclusion of main effects of TICs did not substantially change the significance level of detected interactions between the SNPs and the time-independent latent components. The maximum of the \( p \)-values of the two models were shown for the most significant SNPs in each gene. Among these, the cardiac-specific kinase TNNI3K interacts specifically with cardiac troponin I and has been found to protect the myocardium from ischemic injury [10].

**Discussion**

In this study, we showed that the SLAM approach can be extended by including a novel method of two-level latent component analysis to address the challenge of analyzing multivariate longitudinal data of the correlated phenotypes, endophenotypes, covariates, and environmental factors typically found in genome-wide association studies of complex diseases such as CHD. Repeated measures from Visits 1, 3, 5, and 7 from GAW16 Problem 2 were used to extract LLCs (longitudinal latent components) that represents differential age- (visit-) dependent risks. The second-level analysis of LLCs extracted time-independent components (TICs) and captured variants with "pure" SNP × TIC interactions. The method seemed to have worked well in teasing out variants with promising G × E interactions in CHD, by analyses in derived context that potentially homogenize samples according underlying G × E mechanisms. Note that medication uses were not directly modeled because their effects should be reflected in the measured endophenotypes. The derivation of the longitudinal latent components (LLC) may benefit from

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Chr</th>
<th>MAF</th>
<th>Gene</th>
<th>Visit</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP_A-4199078</td>
<td>2</td>
<td>0.07</td>
<td>IL1R N</td>
<td>LLC1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-1788738</td>
<td>4</td>
<td>0.06</td>
<td>KLF3</td>
<td>LLC4</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-1978322</td>
<td>4</td>
<td>0.06</td>
<td>TACR3</td>
<td>LLC3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-2260338</td>
<td>5</td>
<td>0.06</td>
<td>FTMT</td>
<td>LLC2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-2031704</td>
<td>5</td>
<td>0.15</td>
<td>NDS1</td>
<td>LLC4</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-1987480</td>
<td>6</td>
<td>0.15</td>
<td>TRDN</td>
<td>LLC3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-2090526</td>
<td>6</td>
<td>0.06</td>
<td>CITED2</td>
<td>LLC2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-4217972</td>
<td>10</td>
<td>0.07</td>
<td>KCNMA1</td>
<td>LLC2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-4272586</td>
<td>12</td>
<td>0.05</td>
<td>PIK3CG</td>
<td>LLC1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-1961226</td>
<td>12</td>
<td>0.10</td>
<td>PKP2</td>
<td>LLC1 LLC1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-4222134</td>
<td>13</td>
<td>0.15</td>
<td>IRS2</td>
<td>LLC2 LLC2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-2306682</td>
<td>16</td>
<td>0.07</td>
<td>SONNI1B</td>
<td>LLC4 LLC4</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-2019383</td>
<td>21</td>
<td>0.07</td>
<td>PSMGI1</td>
<td>LLC1 LLC1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SNP_A-2051756</td>
<td>22</td>
<td>0.07</td>
<td>MB</td>
<td>LLC3 LLC3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

**Table 2: Representative candidate genes and SNPs that were detected with significant SNP × LLC interactions (\( p \leq 10^{-6} \))**
more rigorous mathematical treatment, e.g., survival analysis of time to events of CHD. Finally, further characterization of G × E mechanisms will require identifying the right environment variables after the extended SLAM analysis, followed by explicit modeling of G × E interactions, and will be the topic of our future studies.

List of abbreviations used
CHD: Coronary heart disease; G × E: Gene-environment interaction; GAW: Genetic Analysis Workshop; IC: Independent components; LLCs: Longitudinal latent components; MAF: Minor allele frequency; SLAM: Supervised statistical learning approach for multivariate analysis; SNP: Single-nucleotide polymorphism; TIC: Time-independent component.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
CCG developed the concept and method, participated in analysis, drafted and revised the manuscript, and gave final approval for publication; WY performed analysis and help drafted the manuscript; ATK participated in analysis and manuscript development; LdlF revised the manuscript critically.

Acknowledgements
This research is supported in part by NIH grants HL091028, HL071782, and an AHA grant 0855626G. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575.

This article has been published as part of BMC Proceedings Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3/S7/S86.

Table 3: Genes containing ≥1 SNPs with “pure” SNP × TIC interactions for CHD (α = 10^-5), and p-values for most significant SNPs and their MAFs

<table>
<thead>
<tr>
<th>Chr</th>
<th>MAF</th>
<th>Gene</th>
<th>TIC1</th>
<th>TIC2</th>
<th>TIC3</th>
<th>TIC4</th>
<th>No. significant SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.052</td>
<td>TNNI3K</td>
<td></td>
<td></td>
<td>1.59 × 10^-6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.113</td>
<td>CNBP</td>
<td></td>
<td></td>
<td>8.73 × 10^-6</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0.060</td>
<td>PCDH7</td>
<td>4.76 × 10^-7</td>
<td></td>
<td>8.27 × 10^-6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.066</td>
<td>PARP8</td>
<td></td>
<td></td>
<td>3.83 × 10^-6</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>0.160</td>
<td>TRDN</td>
<td></td>
<td></td>
<td>1.78 × 10^-6</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>0.237</td>
<td>EIF3H</td>
<td></td>
<td></td>
<td>1.16 × 10^-7</td>
<td>4.58 × 10^-6</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0.085</td>
<td>FREQ</td>
<td></td>
<td></td>
<td>3.00 × 10^-6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.053</td>
<td>FAS</td>
<td></td>
<td></td>
<td>6.01 × 10^-6</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>0.061</td>
<td>GABRG3</td>
<td></td>
<td></td>
<td>8.73 × 10^-6</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>0.085</td>
<td>TJP1</td>
<td></td>
<td></td>
<td>2.25 × 10^-4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0.313</td>
<td>ARHGAP17</td>
<td>9.18 × 10^-6</td>
<td></td>
<td>1.59 × 10^-6</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

References