



Article

# HisCoM-G×E: Hierarchical Structural Component Analysis of Gene-Based Gene–Environment Interactions

Sungkyoung Choi <sup>1</sup> , Sungyoung Lee <sup>2</sup>, Iksoo Huh <sup>3</sup>, Heungsun Hwang <sup>4</sup> and Taesung Park <sup>5,6,\*</sup>

<sup>1</sup> Department of Applied Mathematics, Hanyang University (ERICA), Ansan 15588, Korea; day0413@hanyang.ac.kr

<sup>2</sup> Center for Precision Medicine, Seoul National University Hospital, Seoul 03080, Korea; biznok@snu.ac.kr

<sup>3</sup> Department of nursing, College of Nursing and Research Institute of Nursing Science, Seoul National University, Seoul 03080, Korea; huhixoo@gmail.com

<sup>4</sup> Department of Psychology, McGill University, Montreal, QC H3A 1G1, Canada; heungsun.hwang@mcgill.ca

<sup>5</sup> Department of Statistics, Seoul National University, Seoul 08826, Korea

<sup>6</sup> Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea

\* Correspondence: tspark@stats.snu.ac.kr; Tel.: +82-02-880-8924

Received: 27 July 2020; Accepted: 4 September 2020; Published: 14 September 2020



**Abstract:** Gene–environment interaction (G×E) studies are one of the most important solutions for understanding the “missing heritability” problem in genome-wide association studies (GWAS). Although many statistical methods have been proposed for detecting and identifying G×E, most employ single nucleotide polymorphism (SNP)-level analysis. In this study, we propose a new statistical method, Hierarchical structural CoMponent analysis of gene-based Gene–Environment interactions (HisCoM-G×E). HisCoM-G×E is based on the hierarchical structural relationship among all SNPs within a gene, and can accommodate all possible SNP-level effects into a single latent variable, by imposing a ridge penalty, and thus more efficiently takes into account the latent interaction term of G×E. The performance of the proposed method was evaluated in simulation studies, and we applied the proposed method to investigate gene–alcohol intake interactions affecting systolic blood pressure (SBP), using samples from the Korea Associated REsource (KARE) consortium data.

**Keywords:** gene–environment interactions; generalized structured component analysis (GSCA); genome-wide association study (GWAS)

## 1. Introduction

Common genetic variants, identified by genome-wide association studies (GWAS), generally explain only a very small proportion of the variance of complex diseases or traits [1,2]. This may be because the genetic risk of single nucleotide polymorphisms (SNPs) is modified by environment, gender, age, or other factors. Understanding those interaction effects can help us to explain the “missing heritability” problem of complex diseases or traits [2]. Several studies of gene–environment interactions (G×E) have been performed for a variety of human complex diseases [3–5]. However, these studies have been mainly for candidate disease-associated genes. Moreover, findings from G×E studies based on genome-wide scale investigations [6] are quite limited, due to statistical power issues. An implementation of HisCoM-G×E can be downloaded from the website <http://statgen.snu.ac.kr/software/hiscom-gxe>.

Several methods have been proposed to improve statistical power for identifying G×E. These methods can be broadly categorized into two types. The first, which aims at increasing statistical power by performing an SNP-based G×E test, is the two-stage approach [7,8]. In the first stage, filtering is

conducted at the SNP level. In the next stage, SNP-based G×E tests are conducted, and correctional procedures for multiple testing are performed only on the remaining SNPs.

The second type of method focuses on increasing statistical power by performing gene-based (or set-based) G×E analysis. The gene-based G×E test has several advantages, including aggregating multiple G×E signals within the same gene and reducing multiple-testing burdens. Furthermore, the gene-based G×E test simplifies biologically meaningful interpretation of the disease. Ma et al. proposed that gene-based gene–gene interaction (GGG) tests represent an extension of four existing methods of combining  $p$ -values [9] as follows: (1) gene-based association test using extended Simes (GATES) procedures [10]; (2) truncated tail strength [11]; (3) truncated  $p$ -value product [12]; and (4) minimum  $p$ -value [13]. GGG tests have been introduced for gene–gene interaction assessments at the gene level, and can also be applied to G×E studies.

Additionally, Lin et al. (2013) proposed a gene-based G×E test, i.e., the gene–environment set association test (GESAT), by extending the SNP-set (sequence) kernel association test (SKAT) to a G×E setting [14]. Furthermore, Lin et al. (2016) proposed the interaction sequence kernel association test (iSKAT), for assessing rare variants by environmental interactions [15]. The GESAT and iSKAT methods assume random G×E effects, following a distribution with mean zero and variant  $\tau^2$ ; therefore, testing for G×E effects is equivalent to testing for a zero variance of  $\tau^2$ . The iSKAT method can also be applied to G×E studies, for common variants, using the weight parameter,  $w = \text{Beta}(\text{minor allele frequency (MAF)}, 0.5, 0.5)$  [15,16].

In this study, we present a novel statistical method for G×E analysis, namely the Hierarchical structural CoMponent analysis of gene-based Gene–Environment interactions (HisCoM-G×E). The proposed method is based on generalized structured component analysis (GSCA), which tests hypothesized defined latent variables as components, using the weighted sums of observed variables [17]. Taking that into account, Lee et al. (2016) proposed a pathway-based approach, using a hierarchical structure of collapsed rare variants of high-throughput sequencing data (PHARAOH), by extending GSCA to pathway analysis of rare variants [18]. Furthermore, Choi et al. (2018) proposed the hierarchical structural component analysis of gene–gene interactions (HisCoM-GGI), an extension of the PHARAOH method, for gene–gene interaction analysis [19]. The HisCoM-GGI method can evaluate both gene-level and SNP-level interactions. The HisCoM-G×E method is an extension of the HisCoM-GGI method for performing gene-based G×E analysis. HisCoM-G×E introduces latent variables, such as a gene, and an environmental factor, which are defined as a weighted sum of the observed variables, such as SNPs, and an environment factor. Accordingly, our proposed method can efficiently account for the biological relationship between a gene and an environmental factor (e.g., alcohol intake), within the structured component.

Previous methods such as GGG tests, used for detecting G×E, integrate multiple  $p$ -values from each SNP-level interaction test into a single  $p$ -value of gene-level interactions. By contrast, the HisCoM-G×E method can evaluate G×E effects on the phenotype of interest, all at once. Moreover, our proposed method can account for linkage disequilibrium (LD) between SNPs within a gene, by imposing a ridge penalty.

In this report, using simulation studies, we show that the performance of the HisCoM-G×E method is similar to or better than other traditional approaches. In addition, we applied the HisCoM-G×E method to a GWAS dataset, as associated with systolic blood pressure (SBP), and an environmental factor (alcohol intake frequency), via genotyping 2252 participants from the Korean Association REsource (KARE) cohort study, using the Affymetrix Genome-Wide Human SNP Array 5.0 [20].

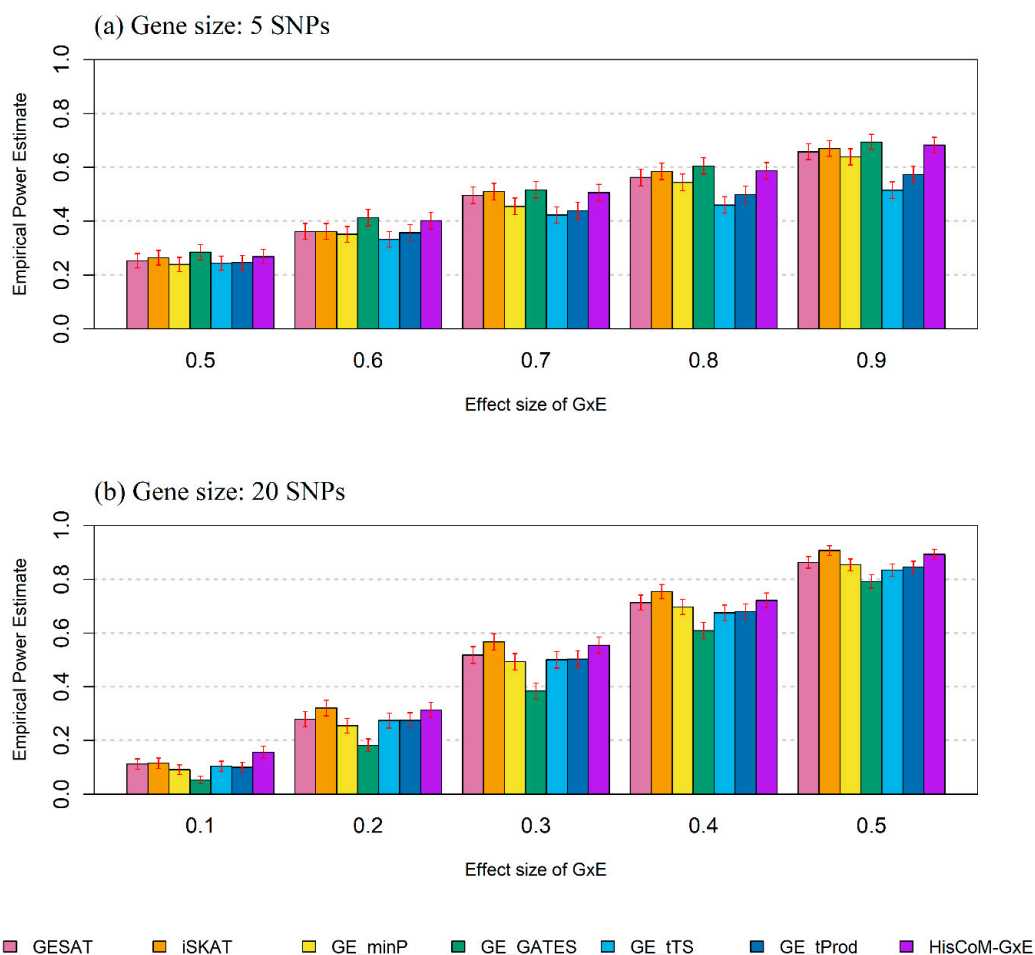
## 2. Results

### 2.1. Type I Error Rate

Table S1 (Supplementary Materials) shows that the empirical type I error rates of the GESAT, iSKAT, GE\_minP, GE\_GATES, GE\_tTS, GE\_tProd, and HisCoM-G×E methods were all well preserved, at nominal significance levels for the three scenarios, when the gene size was 5, 50, and 100 SNPs.

### 2.2. Power Comparison

To evaluate the statistical power of the GESAT, iSKAT, GE\_minP, GE\_GATES, GE\_tTS, GE\_tProd, and HisCoM-G×E methods, we varied gene and effect sizes of G×E, and then calculated the empirical power for each parameter setting. The results in Figure 1a, when the gene size was 5 SNPs, showed that the GE\_GATES method was the most powerful, while the HisCoM-G×E method yielded comparable power. As shown in Figure 1b, when the gene size was 20 SNPs, the iSKAT method was generally the most powerful, and HisCoM-G×E yielded comparable power. In the case of the small effect size of G×E, HisCoM-G×E was more powerful than any of the other methods.



**Figure 1.** Empirical power estimates as effect size of G×E when the gene sizes are (a) 5 SNPs and (b) 20 SNPs. The empirical power estimates were calculated with 1000 replicates. Results for each method can be distinguished by plotting colors. Each bar is the mean and the error bars represent standard deviation (SD).

In summary, we can conclude that the HisCoM-G×E method performs reasonably well, regardless of gene size. In other words, it can yield power that is quite comparable to those of the most powerful methods, in the case of different gene sizes.

### 2.3. Real Data Analysis: KARE Dataset

We next applied the iSKAT, GE\_GATES, and HisCoM-G×E methods to investigate gene–alcohol intake interactions affecting systolic blood pressure (SBP), using samples from the KARE GWAS dataset [20]. In all, 16,361 genes were included in the analysis. The *p*-values for G×E, using HisCoM-G×E, were calculated via 5000 permutations. In Figure S1 (Supplementary Materials), the quantile–quantile (QQ) plot of GE\_GATES showed some evidence for deflation of *p*-values, whereas iSKAT and HisCoM-G×E did not. Figure S2 (Supplementary Materials) shows the Manhattan plots for the iSKAT, GE\_GATES, and HisCoM-G×E methods, where the horizontal red line denotes the threshold for 0.05 genome-wide significance level by a Bonferroni correction of  $3.06 \times 10^{-6}$  ( $\alpha = 0.05/16,361$  genes). Table 1 summarizes the 35 G×Es with *p*-values less than the nominal significance level of 0.001. The two genes, *NUCB2* and *ACE*, identified by iSKAT, associate with SBP [21–23], and in fact, the angiotensin-converting enzyme (*ACE*) is a long-held target of numerous antihypertensives [24]. Notably, the HisCoM-G×E method also successfully identified the gene *UGDH*, well known to relate to peripheral arterial disease (PAD) [25]. Additionally, it has been reported that SBP associates with significantly higher risks for a PAD event [26], and that *FCAMR* and *PIGR*, identified by HisCoM-G×E, both significantly associate with coronary atherosclerosis disease [27]. It also has been reported that elevated SBP can predict coronary atherosclerosis, and provides additional information for predicting coronary calcification [28], and is also a risk factor for stroke, myocardial infarction, kidney dysfunction, and aneurysm [29]. Unfortunately, there were no statistically significant interactions detected after Bonferroni correction or the false discovery rate (FDR), using an adjusted *p*-value (*q*-value < 0.05) [30].

**Table 1.** Significant ( $p < 0.001$ ; bold) gene–environment interactions that affect systolic blood pressure (SBP), according to the interaction sequence kernel association test (iSKAT), the gene–environment gene-based association test using extended Simes (GE\_GATES), and HisCoM-G×E, using the Korea Associated REsource (KARE) genome-wide association studies (GWAS) dataset.

No	CHR	GENE	# of SNPs	iSKAT	GE_GATES	HisCoM-G×E
1	1	<i>NBPF14</i>	54	0.5697	0.9902	<b><math>6.00 \times 10^{-5}</math></b>
2	19	<i>TLE</i>	7	0.2587	0.2491	<b><math>1.00 \times 10^{-4}</math></b>
3	4	<i>COMMD8</i>	116	0.0285	0.2064	<b><math>1.40 \times 10^{-4}</math></b>
4	4	<i>UGDH</i>	90	0.0141	0.0518	<b><math>2.00 \times 10^{-4}</math></b>
5	8	<i>ZFAND1</i>	91	0.1227	0.1713	<b><math>2.20 \times 10^{-4}</math></b>
6	2	<i>RY1</i>	8	0.1900	0.9966	<b><math>2.60 \times 10^{-4}</math></b>
7	9	<i>GAPVD1</i>	120	0.4700	0.6905	<b><math>2.80 \times 10^{-4}</math></b>
8	1	<i>FCAMR</i>	107	0.7512	0.9398	<b><math>3.40 \times 10^{-4}</math></b>
9	9	<i>ADEF</i>	78	0.6676	0.4259	<b><math>4.20 \times 10^{-4}</math></b>
10	1	<i>PIGR</i>	119	0.7613	0.9911	<b><math>5.00 \times 10^{-4}</math></b>
11	3	<i>EIF1B</i>	89	0.1935	0.2530	<b><math>8.00 \times 10^{-4}</math></b>
12	19	<i>ZNF321</i>	41	<b><math>6.74 \times 10^{-4}</math></b>	0.0262	0.0182
13	17	<i>ACE</i>	154	<b><math>9.50 \times 10^{-4}</math></b>	0.0051	0.0220
14	19	<i>ZNF566</i>	11	0.3003	<b><math>5.46 \times 10^{-4}</math></b>	0.0670
15	13	<i>SMAD9</i>	313	<b><math>7.25 \times 10^{-4}</math></b>	0.5510	0.0950

Table 1. Cont.

No	CHR	GENE	# of SNPs	iSKAT	GE_GATES	HisCoM-G×E
16	8	LETM2	107	0.0027	$5.51 \times 10^{-4}$	0.0974
17	16	KATNB1	38	$1.57 \times 10^{-4}$	0.0011	0.1555
18	16	PAR6A	60	$7.36 \times 10^{-5}$	0.0010	0.1887
19	16	ACD	69	$1.08 \times 10^{-4}$	0.0014	0.2463
20	11	CALCB	62	0.8119	$2.45 \times 10^{-4}$	0.2515
21	14	RAD51L1	1199	$2.71 \times 10^{-4}$	0.1166	0.2515
22	11	TRIM5	53	$9.48 \times 10^{-4}$	0.0070	0.2633
23	16	C16orf86	68	$3.30 \times 10^{-5}$	0.0015	0.2787
24	4	SORCS2	1270	$5.99 \times 10^{-4}$	$2.38 \times 10^{-5}$	0.3954
25	11	NUCB2	134	$8.16 \times 10^{-5}$	0.0341	0.4232
26	21	DSCR3	230	$3.35 \times 10^{-4}$	0.0060	0.4502
27	16	RLTPR	82	$3.30 \times 10^{-5}$	$8.65 \times 10^{-4}$	0.4794
28	16	KIFC3	131	0.0393	$9.48 \times 10^{-4}$	0.5110
29	1	ST6GALNAC3	1212	$6.51 \times 10^{-4}$	0.2535	0.5270
30	7	YKT6	61	$9.96 \times 10^{-4}$	0.1369	0.5700
31	19	ZNF99	90	$8.90 \times 10^{-4}$	0.0154	0.5736
32	14	SSTR1	40	$7.67 \times 10^{-4}$	0.0249	0.8029
33	2	C2orf21	50	$5.25 \times 10^{-4}$	0.1123	0.8119
34	8	WHSC1L1	134	$5.79 \times 10^{-4}$	0.0020	0.8379
35	4	TNFRSF1A	57	0.5246	$2.57 \times 10^{-4}$	0.8798

### 3. Discussion

In this study, we proposed a new statistical model, namely the Hierarchical structural CoMponent analysis of gene-based Gene–Environment interaction (HisCoM-G×E), to identify gene–environment interactions. Like other hierarchical component models, HisCoM-G×E has several advantages. First, HisCoM-G×E can easily incorporate the biological relationship between a gene and environmental factor. Second, HisCoM-G×E can effectively summarize the interaction effect of a gene and its environment, using structured components. Third, it greatly reduces the dimension of SNPs, in a gene, using a latent variable, defined as a weighted sum of observed variables. In this study, we developed the software for HisCoM-G×E (<http://statgen.snu.ac.kr/software/hiscom-gxe>). The basic framework of the HisCoM-G×E software is based on the “Workbench for Integrated Superfast Association study with Related Data” (WISARD) [31]. The HisCoM-G×E software was implemented in C++, and was developed for Linux. When our simulation study was executed on a server having two Intel Xeon E5-2620 processors with 128GB of RAM, the HisCoM-G×E program took 30–40 s and 50–60 s when the total number of SNPs in a gene ( $K$ ) was 5 and 20, respectively. Furthermore, HisCoM-G×E can be extended to a statistical method that can take into account various distribution types (i.e., binomial, Poisson, gamma, and inverse Gaussian distributions) by using the framework of generalized linear models (GLMs) applied in the PHARAOH method [18].

One weakness of the study is that the method does not consider epigenetic events (e.g., DNA and histone modifications). Such phenomena are well known to be affected by the environment and exert gene regulation at an additional level, compared to gene sequence alone (e.g., SNPs, indels, copy number, etc.) [32]. To address this shortcoming, future models will be able to consider both genomics and epigenomics in gene regulation/dysregulation.

### 4. Materials and Methods

#### 4.1. HisCoM-G×E Method

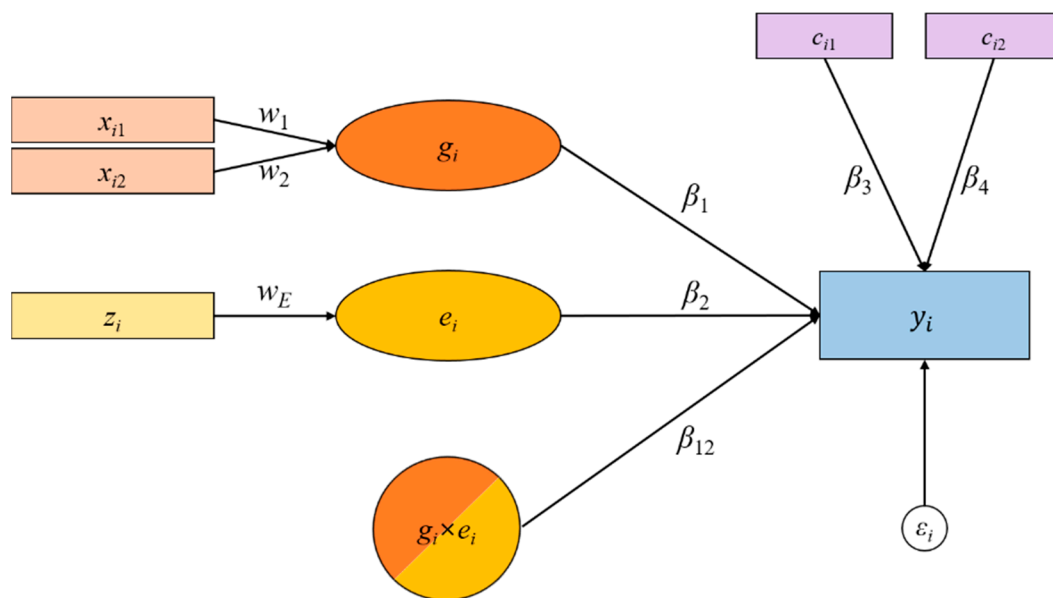
Let  $y_i$  denote the phenotype of the  $i$ -th subject ( $i = 1, \dots, N$ ). Let  $c_{ip}$  denote the  $p$ -th covariate of the  $i$ -th subject ( $p = 1, \dots, P$ ). Let  $K$  be the number of SNPs in a gene. Let  $x_{ik}$  denote the  $k$ -th SNP in a gene, for the  $i$ -th subject ( $k = 1, \dots, K$ ). Let  $w_k$  denote a weight assigned to  $x_{ik}$ . Let  $g_i$  be a latent variable representing the main effect of a gene, which is defined as a weighted sum of  $K$  SNPs, such that  $g_i = \sum_{k=1}^K w_k x_{ik}$ . Let  $z_i$  denote the environment of the  $i$ -th subject, and  $e_i$  be a latent variable representing the environmental effect, which is defined as a weighted sum of environmental effects, such that  $e_i = w_E z_i$ . Then, the latent interaction term between the gene and environment term is represented by another latent variable,  $r_i$ , which is obtained as the product of interacting latent variables, such that  $r_i = (\sum_{k=1}^K w_k x_{ik}) \cdot (w_E z_i) = g_i \cdot e_i$ .

Let  $\beta_1, \beta_2, \beta_p$ , and  $\beta_{12}$  denote the coefficients of the gene, the environment, the  $p$ -th covariate, and latent interaction effects on  $y_i$ , respectively. Then, the relationships between the phenotypes and latent variables are established, such that:

$$y_i = \beta_1 g_i + \beta_2 e_i + \beta_{12} r_i + \sum_{p=1}^P \beta_p c_{ip} + \varepsilon_i \tag{1}$$

where  $\varepsilon_i$  is the error term for subject  $i$ .

Figure 2 shows a simple HisCoM-G×E model with two covariates (i.e.,  $P = 2$ ). Rectangles and circles represent observed and latent variables, respectively. For illustrative purposes, we assume that the gene consists of two SNPs (i.e.,  $K = 2$ ). The gene, then, is defined as a latent variable constructed by a weighted sum of its SNPs, which in turn influence a phenotype, signified by single-headed arrows. A G×E is similarly defined as a latent variable, constructed via the products of interacting latent variables.



**Figure 2.** A schematic diagram of Hierarchical structural CoMponent analysis of gene-based Gene-Environment interactions (HisCoM-G×E). The exemplary model is described with the number of single nucleotide polymorphisms (SNPs) ( $x_{ik}$ )  $K = 2$  and the number of covariates ( $c_{ip}$ )  $P = 2$ . The variables  $w$ s denote the weights assigned to the latent variables and  $\beta$ s are coefficients of the latent variables ( $g_i$  and  $e_i$ ). The  $g_i \times e_i$  term represents a latent interaction term (or effect) on the phenotype ( $y_i$ ) and  $\varepsilon_i$  is the error term.

Let  $\mathbf{W}$  be a matrix of  $w_1, w_2, \dots, w_k, w_E$ , and  $\mathbf{B}$  be a matrix of  $\beta_1, \beta_2, \beta_p$ , and  $\beta_{12}$ . To estimate the unknown parameters for HisCoM-G×E, we adopted the alternating regulated least-squares (ALS) algorithm [18,33]. To estimate unknown parameters in  $\mathbf{W}$  and  $\mathbf{B}$ , we sought to minimize the following Equation (2), subject to the standardization constraint on each latent variable,  $\sum_{i=1}^N r_i^2 = N$  [34]. Therefore, the penalized least-squares equation is given as follows:

$$\phi = SS(\mathbf{y} - \hat{\mathbf{y}}) + \lambda_S SS(\mathbf{W}) + SS(\mathbf{B}), \quad (2)$$

where  $\lambda_S$  is the ridge parameter for the gene. Before applying the algorithm, we used  $k$ -fold cross-validation (CV) to select the values of  $\lambda_S$  automatically. In our analysis, the ridge parameter,  $\lambda_S$ , is based on five-fold CV, using 15 different starting points of ridge parameters, ranging from 0 to 0.5.

Testing for interactions is equivalent to testing the null hypothesis for the G×E effect, as shown by the following:

$$H_0: \beta_{12} = 0. \quad (3)$$

We also adopted a permutation procedure, used in our previously developed PHARAOH method [18], to test the statistical significance of the estimated effects of genetics, environment, and the G×E gene–environment. By permuting the phenotype, the HisCoM-G×E method generates the null distributions of weights and path coefficients. Thus, we can obtain empirical  $p$ -values of interactions between gene and environmental factors.

#### 4.2. Simulation Study

To investigate the performance of HisCoM-G×E, we conducted simulation studies to generate human genomic data, with realistic linkage disequilibrium (LD) patterns. We employed a similar simulation strategy that was suggested by Choi et al. [19] for the gene–gene interactions study. In the gene–environment interaction analysis, various factors, i.e., minor allele frequency, haplotype block, and many other factors, can affect the results of a simulation study. In order to reflect various factors, we performed a simulation study by randomly extracting genome data from a real GWAS dataset. In each simulation, we assumed that the phenotype data arose from the following Equation (4), defined as a function of the sum of the genetic value of SNPs, under an additive model, environmental factor, interaction effect, and a random error, following a standard normal distribution, as follows:

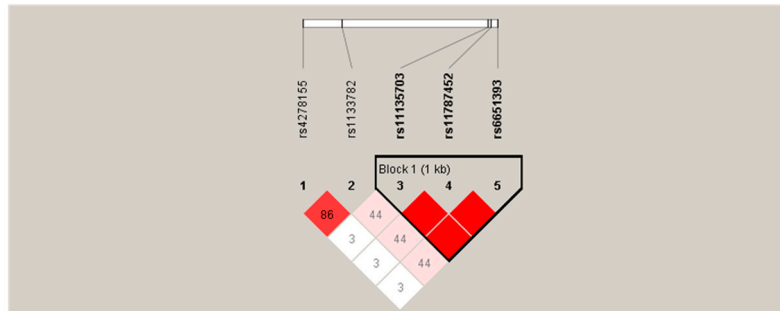
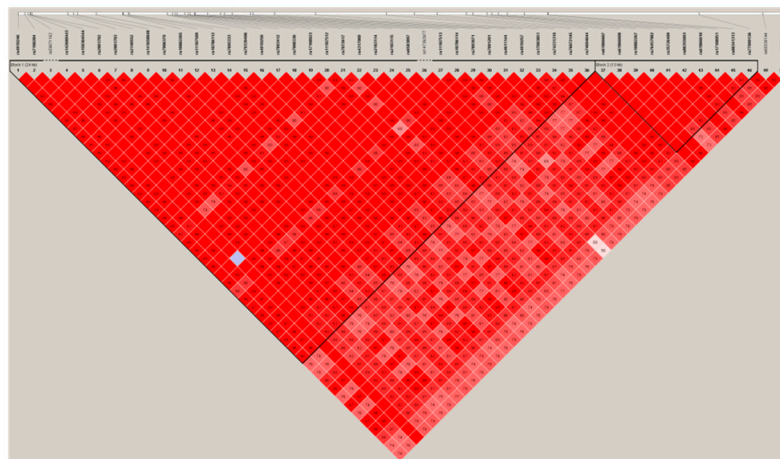
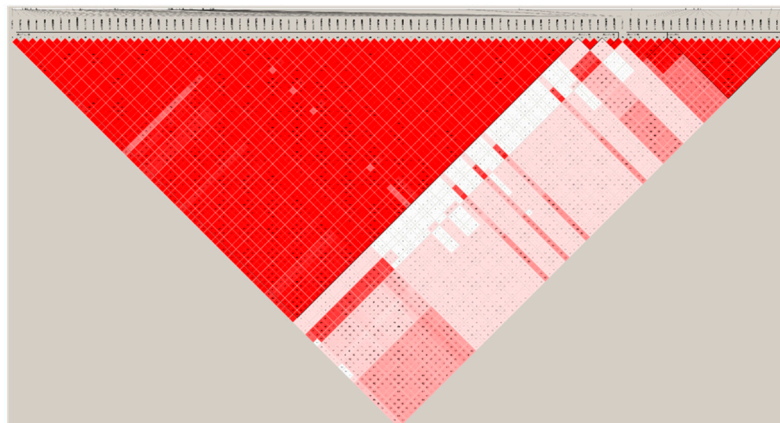
$$y_i = \beta_0 + \beta_1 \left( \sum_{k=1}^K w_k x_{ik} \right) + \beta_2 e_i + \beta_{12} \left( \sum_{k=1}^K w_k x_{ik} e_i \right) + \varepsilon_i, \quad (4)$$

where  $\varepsilon_i$  is the error term for subject  $i$ .

#### 4.3. Type I Error Rate

To check whether the type I error rate is well controlled, we generated datasets under the null hypothesis of no interaction between a gene and its environment. A dataset of  $N$  samples was generated from the KARE GWAS dataset samples [20]. We considered the following three scenarios, with varying number of SNPs, in a gene: (1) five SNPs ( $K = 5$ ) in *TNFRSF10C*, (2) fifty SNPs ( $K = 50$ ) in *GPR120*, or (3) one hundred SNPs ( $K = 100$ ) in *PLA2G4C*. In Figure 3, the LD pattern plots of those genes were generated using Haploview v4.2 software (Broad Institute, Cambridge, MA, USA) [35].

For each simulated dataset, we randomly selected 500, 1000, and 2000 samples from the KARE GWAS dataset, and generated 1000 simulated datasets, for each scenario. The empirical type I error was estimated based on the significance levels  $\alpha = 0.05, 0.01$ , and  $0.005$ .

(a) *TNFRSF10C* gene ( $K = 5$ )(b) *TNFRSF10C* gene ( $K = 50$ )(c) *PLA2G4C* gene ( $K = 100$ )

**Figure 3.** Linkage disequilibrium (LD) patterns of genes used in simulation studies: (a) *TNFRSG10C* gene ( $K = 5$ ); (b) *TNFRSF10C* gene ( $K = 50$ ); (c) *PLA2G4C* gene ( $K = 100$ ). Numbers indicate the  $D'$  values expressed as percentiles. A standard color scheme is used to display the LD pattern, with red for perfect LD ( $r^2 = 1$ ), white for no LD ( $r^2 = 0$ ) and shades of pink/red for intermediate LD ( $0 < r^2 < 1$ ).

#### 4.4. Power Comparison

For power analysis, we varied the number of causal SNPs and the effect sizes of  $G \times E$ . The performance of HisCoM- $G \times E$  was compared to that of GESAT, iSKAT, and the following four GGG methods: GE\_minP, GE\_GATES, GE\_tTS, and GE\_tProd [9,14,15]. Two scenarios were considered, with different numbers of SNPs, in various genes as follows: (1) five SNPs ( $K = 5$ ) in *TNFRSF10C* and (2) twenty SNPs ( $K = 20$ ) in *ZNF616*. For each simulated dataset, 500 samples were randomly selected from the KARE GWAS dataset. The parameters were set as follows: total SNPs in a gene ( $K$ ) = 5 and 20, sample size ( $N$ ) = 500, the proportion of causal SNPs = 20%, an effect size, for each SNP ( $w_k$ ) =



0.5, and the effect size of G×E was as follows: (1) five SNPs ( $K = 5$ ) with  $\beta_{12} = 0.5, 0.6, 0.7, 0.8, 0.9$ , or (2) twenty SNPs ( $K = 20$ ) with  $\beta_{12} = 0.1, 0.2, 0.3, 0.4, 0.5$ . Since the phenotype was generated by both the effect of SNP ( $w_k$ ) and the effect of interaction terms ( $\beta_{12}$ ) together, we tested situations in which the effect size was smaller than 1. When the effect size was larger than 1, no difference in the power between statistical methods was observed. The empirical power was estimated by counting the number of associations below a specified significance level (e.g.,  $\alpha = 0.05$ ), based on the total number of tests within each set of 1000 simulated datasets.

#### 4.5. Real Data Analysis: KARE Dataset

The participants of the KARE project were recruited from Ansong (rural) and Ansan (urban) cohorts in South Korea [20]. Among 8840 subjects, we excluded the subjects with a history of cancer (104 individuals), anti-hypertensive therapy (1019 individuals), drug treatments regulating blood pressure (1055 individuals), and individuals missing values for either body mass index (BMI) (4 individuals) or systolic blood pressure (SBP) (1 individual). Among the female participants (4658 individuals), the alcohol intake frequency was very low, but was relatively high for males, so only 2252 males were included in this study. This study was approved by the Institutional Review Board of the Korean National Institute of Health (IRB No. E1908/001-004). All subjects provided written informed consent to participate in this study, and were then anonymized. We used a detailed genotyping and quality control process previously described [20], and missing genotypes were imputed by IMPUTE2, using Phase 1 of the 1000 Genomes Project as a reference panel [36]. The genetic variants with low IMPUTE2 INFO scores ( $< 0.8$ ), high missing call rates ( $> 5\%$ ), low minor allele frequencies ( $\leq 0.01$ ), or of insignificant  $p$ -values by the Hardy–Weinberg equilibrium test ( $\leq 1 \times 10^{-6}$ ) were excluded from the study. Then, we assigned SNPs to genes if they were located within the gene body, or its 20 kb upstream or downstream sequences. As a result, 228,3615 SNPs were mapped to a total of 1,6361 genes. The gene–alcohol intake interaction was analyzed by the HisCoM-G×E method, assuming an additive genetic effect, adjusted for age, recruitment area, and BMI.

## 5. Conclusions

Through the simulation study, we showed that HisCoM-G×E controlled type I error rates well in various scenarios. In power comparison studies, there was only one of the most powerful methods available for the cases of varying gene size. Only HisCoM-G×E yielded power comparable to that of the most powerful methods, regardless of gene size. In the case of the small interaction effect size, HisCoM-G×E yielded the most powerful statistic. Therefore, we recommend HisCoM-G×E as the best choice for analyzing smaller-scale gene–environment interaction studies.

The real data analysis of the KARE GWAS dataset demonstrated that HisCoM-G×E successfully identified genes previously reported for G×E interactions for SBP. Therefore, we fully expect that HisCoM-G×E will help researchers understand environmental contributions to human complex diseases.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/1422-0067/21/18/6724/s1>, Figure S1: Quantile-quantile (QQ) plots of the G×E analysis for systolic blood pressure (SBP), Figure S2: Manhattan plots of the G×E analysis for SBP, Table S1: Empirical type I error estimates calculated, with 1000 replicates, at 0.05, 0.01, and 0.005 significance levels ( $\alpha$ ).

**Author Contributions:** Conceptualization, S.C.; methodology, S.C., H.H., and T.P.; software, S.L.; investigation, S.C. and I.H.; writing—original draft preparation, S.C. and T.P.; writing—review and editing, S.C., S.L., I.H., H.H., and T.P.; visualization, S.C.; supervision, T.P.; funding acquisition, T.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare (No. HI16C2037), the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation (NRF).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

G×E	Gene–environment interaction
GWAS	Genome-wide association studies
SNPs	Single nucleotide polymorphisms
HisCoM-G×E	Hierarchical structural CoMponent analysis of gene-based Gene–Environment interactions
KARE	Korea Associated REsource
GGG	Gene-based gene–gene interaction
GESAT	The gene–environment set association test
iSKAT	The interaction sequence kernel association test
PHARAOH	Pathway-based approach, using a hierarchical structure of collapsed rare variants of high-throughput sequencing data
HisCoM-GGI	The hierarchical structural component analysis of gene–gene interactions
LD	Linkage disequilibrium
SBP	Systolic blood pressure
BMI	Body mass index
KNIH	Korean National Institute of Health
QQ plot	Quantile–quantile plot
FDR	False discovery rate
SD	Standard deviation

## References

1. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **2008**, *456*, 18–21. [[CrossRef](#)] [[PubMed](#)]
2. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753. [[CrossRef](#)] [[PubMed](#)]
3. Hunter, D.J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **2005**, *6*, 287–298. [[CrossRef](#)] [[PubMed](#)]
4. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [[CrossRef](#)] [[PubMed](#)]
5. Thomas, D. Gene-environment-wide association studies: Emerging approaches. *Nat. Rev. Genet.* **2010**, *11*, 259–272. [[CrossRef](#)] [[PubMed](#)]
6. Rothman, N.; Garcia-Closas, M.; Chatterjee, N.; Malats, N.; Wu, X.F.; Figueroa, J.D.; Real, F.X.; Van den Berg, D.; Matullo, G.; Baris, D.; et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat. Genet.* **2010**, *42*, 978–998. [[CrossRef](#)]
7. Kooperberg, C.; LeBlanc, M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet. Epidemiol.* **2008**, *32*, 255–263. [[CrossRef](#)]
8. Murcray, C.E.; Lewinger, J.P.; Gauderman, W.J. Gene-Environment Interaction in Genome-Wide Association Studies. *Am. J. Epidemiol.* **2009**, *169*, 219–226. [[CrossRef](#)]
9. Ma, L.; Clark, A.G.; Keinan, A. Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. *PLoS Genet.* **2013**, *9*, 1–12. [[CrossRef](#)]
10. Li, M.X.; Gui, H.S.; Kwan, J.S.H.; Sham, P.C. GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *Am. J. Hum. Gen.* **2011**, *88*, 283–293. [[CrossRef](#)]
11. Jiang, B.; Zhang, X.A.; Zuo, Y.J.; Kang, G.L. A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *J. Theor. Biol.* **2011**, *277*, 67–73. [[CrossRef](#)] [[PubMed](#)]
12. Zaykin, D.V.; Zhivotovsky, L.A.; Westfall, P.H.; Weir, B.S. Truncated product method for combining P-values. *Genet. Epidemiol.* **2002**, *22*, 170–185. [[CrossRef](#)] [[PubMed](#)]
13. Liu, J.Z.; Mcrae, A.F.; Nyholt, D.R.; Medland, S.E.; Wray, N.R.; Brown, K.M.; Hayward, N.K.; Montgomery, G.W.; Visscher, P.M.; Martin, N.G.; et al. Versatile Gene-Based Test for Genome-wide Association Studies. *A. J. Hum. Gen.* **2010**, *87*, 139–145. [[CrossRef](#)] [[PubMed](#)]

14. Lin, X.Y.; Lee, S.; Christiani, D.C.; Lin, X.H. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **2013**, *14*, 667–681. [[CrossRef](#)]
15. Lin, X.Y.; Lee, S.; Wu, M.C.; Wang, C.L.; Chen, H.; Li, Z.L.; Lin, X.H. Test for Rare Variants by Environment Interactions in Sequencing Association Studies. *Biometrics* **2016**, *72*, 156–164. [[CrossRef](#)]
16. Madsen, B.E.; Browning, S.R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet.* **2009**, *5*, e1000384. [[CrossRef](#)]
17. Hwang, H.S.; Takane, Y. Generalized structured component analysis. *Psychometrika* **2004**, *69*, 81–99. [[CrossRef](#)]
18. Lee, S.; Choi, S.; Kim, Y.J.; Kim, B.J.; Consortium, T.d.-G.; Hwang, H.; Park, T. Pathway-based approach using hierarchical components of collapsed rare variants. *Bioinformatics* **2016**, *32*, i586–i594. [[CrossRef](#)]
19. Choi, S.; Lee, S.; Kim, Y.; Hwang, H.; Park, T. HisCoM-GGI: Hierarchical structural component analysis of gene-gene interactions. *J. Bioinform. Comput. Biol.* **2018**, *16*, e1000384. [[CrossRef](#)]
20. Cho, Y.S.; Go, M.J.; Kim, Y.J.; Heo, J.Y.; Oh, J.H.; Ban, H.J.; Yoon, D.; Lee, M.H.; Kim, D.J.; Park, M. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* **2009**, *41*, 527–534. [[CrossRef](#)]
21. Hoffmann, T.J.; Ehret, G.B.; Nandakumar, P.; Ranatunga, D.; Schaefer, C.; Kwok, P.Y.; Iribarren, C.; Chakravarti, A.; Risch, N. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* **2017**, *49*, 54–64. [[CrossRef](#)] [[PubMed](#)]
22. Wain, L.V.; Vaez, A.; Jansen, R.; Joehanes, R.; van der Most, P.J.; Erzurumluoglu, A.M.; O'Reilly, P.F.; Cabrera, C.P.; Warren, H.R.; Rose, L.M.; et al. Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets From Blood and the Kidney. *Hypertension* **2017**, *70*, e4–e19. [[CrossRef](#)] [[PubMed](#)]
23. Giri, A.; Hellwege, J.N.; Keaton, J.M.; Park, J.; Qiu, C.X.; Warren, H.R.; Torstenson, E.S.; Kovesdy, C.P.; Sun, Y.V.; Wilson, O.D. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **2019**, *51*, 51–62. [[CrossRef](#)]
24. Verme-Giboney, C. Oral angiotensin-converting-enzyme inhibitors. *Am. J. Health Syst. Pharm.* **1997**, *54*, 2689–2703. [[CrossRef](#)] [[PubMed](#)]
25. Ward-Caviness, C.K.; Neas, L.M.; Blach, C.; Haynes, C.S.; LaRocque-Abramson, K.; Grass, E.; Dowdy, E.; Devlin, R.B.; Diaz-Sanchez, D.; Cascio, W.E. Genetic Variants in the Bone Morphogenic Protein Gene Family Modify the Association between Residential Exposure to Traffic and Peripheral Arterial Disease. *PLoS ONE* **2016**, *11*, e0152670. [[CrossRef](#)]
26. Itoga, N.K.; Tawfik, D.S.; Lee, C.K.; Maruyama, S.; Leeper, N.J.; Chang, T.I. Association of Blood Pressure Measurements With Peripheral Artery Disease Events. *Circulation* **2018**, *138*, 1805–1814. [[CrossRef](#)]
27. Ward-Caviness, C.K.; Neas, L.M.; Blach, C.; Haynes, C.S.; LaRocque-Abramson, K.; Grass, E.; Dowdy, Z.E.; Devlin, R.B.; Diaz-Sanchez, D.; Cascio, W.E.; et al. A genome-wide trans-ethnic interaction study links the PIGR-FCAMR locus to coronary atherosclerosis via interactions between genetic variants and residential exposure to traffic. *PLoS ONE* **2017**, *12*, e0173880. [[CrossRef](#)]
28. Her, A.Y.; Cho, K.I.; Garg, S.; Kim, Y.H.; Shin, E.S. Association of Inter-Arm Systolic Blood Pressure Difference with Coronary Atherosclerotic Disease Burden Using Calcium Scoring. *Yonsei Med. J.* **2017**, *58*, 954–958. [[CrossRef](#)]
29. Bavishi, C.; Goel, S.; Messerli, F.H. Isolated Systolic Hypertension: An Update After SPRINT. *Am. J. Med.* **2016**, *129*, 1251–1258. [[CrossRef](#)]
30. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate—A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B-Stat. Methodol.* **1995**, *57*, 289–300. [[CrossRef](#)]
31. Lee, S.; Choi, S.; Qiao, D.; Cho, M.; Silverman, E.K.; Park, T.; Won, S. WISARD: Workbench for integrated superfast association studies for related datasets. *BMC Med. Genom.* **2018**, *11* (Suppl. 2), 39. [[CrossRef](#)] [[PubMed](#)]
32. Trevino, L.S.; Dong, J.; Kaushal, A.; Katz, T.A.; Jangid, R.K.; Robertson, M.J.; Grimm, S.L.; Ambati, C.S.R.; Putluri, V.; Cox, A.R.; et al. Epigenome environment interactions accelerate epigenomic aging and unlock metabolically restricted epigenetic reprogramming in adulthood. *Nat. Commun.* **2020**, *11*, 2316. [[CrossRef](#)] [[PubMed](#)]
33. Hwang, H. Regularized Generalized Structured Component Analysis. *Psychometrika* **2009**, *74*, 517–530. [[CrossRef](#)]

34. Takane, Y.; Hwang, H. An extended redundancy analysis and its applications to two practical examples. *Comput. Stat. Data Anal.* **2005**, *49*, 785–808. [[CrossRef](#)]
35. Barrett, J.C.; Fry, B.; Maller, J.; Daly, M.J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **2005**, *21*, 263–265. [[CrossRef](#)]
36. Howie, B.N.; Donnelly, P.; Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **2009**, *5*, e1000529. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).