scientific reports

OPEN

Identification of genetic loci affecting body mass index through interaction with multiple environmental factors using structured linear mixed model

Hae-Un Jung[1], Won Jun Lee[1], Tae-Woong Ha[1], Ji-One Kang[2], Jihye Kim[3], Mi Kyung Kim[3,4], Sungho Won[5], Taesung Park[6], Ji Eun Lim[2]✉ & Bermseok Oh[2]✉

Multiple environmental factors could interact with a single genetic factor to affect disease phenotypes. We used Struct-LMM to identify genetic variants that interacted with environmental factors related to body mass index (BMI) using data from the Korea Association Resource. The following factors were investigated: alcohol consumption, education, physical activity metabolic equivalent of task (PAMET), income, total calorie intake, protein intake, carbohydrate intake, and smoking status. Initial analysis identified 7 potential single nucleotide polymorphisms (SNPs) that interacted with the environmental factors (*P* value $< 5.00 \times 10^{-6}$). Of the 8 environmental factors, PAMET score was excluded for further analysis since it had an average Bayes Factor (BF) value < 1 (BF = 0.88). Interaction analysis using 7 environmental factors identified 11 SNPs (*P* value $< 5.00 \times 10^{-6}$). Of these, rs2391331 had the most significant interaction (*P* value $= 7.27 \times 10^{-9}$) and was located within the intron of *EFNB2* (Chr 13). In addition, the gene-based genome-wide association study verified *EFNB2* gene significantly interacting with 7 environmental factors (*P* value $= 5.03 \times 10^{-10}$). BF analysis indicated that most environmental factors, except carbohydrate intake, contributed to the interaction of rs2391331 on BMI. Although the replication of the results in other cohorts is warranted, these findings proved the usefulness of Struct-LMM to identify the gene–environment interaction affecting disease.

Gene-environment interaction (GEI) studies evaluate the extent to which the phenotype was affected by the interaction[1]. These studies help in our understanding of the complex human traits that are determined through the interaction of gene and environment[2]. In addition, the gene identified through this analysis can help in discovering the biological pathways underlying the phenotype[3] and in increasing the accuracy of prediction of the disease incidence, calculated by the main genetic and environmental effects[2,4]. Therefore, series of genome-wide interaction studies (GWIS) have been progressed to identify the genetic loci interacted with environmental factors in diverse traits[5–7].

Obesity is a serious environment-related disease that affects many people across the world. Severely obese people have been reported to die 8 to 10 years earlier than those with normal weight. Moreover, obesity is strongly associated with diseases such as cardiovascular diseases, type 2 diabetes, and cancer. In the past 2 decades, the global incident rate of obesity has rapidly increased, and the prevalence of obesity is estimated to double in the next decade[8]. Studies have demonstrated that body mass index (BMI), an index for obesity, is influenced by environments such as physical activity, nutrient intake, and diverse environmental exposure. Likewise, it is well known that BMI is affected by multiple genetic factors, which have been identified by series of genome-wide

[1]Department of Biomedical Science, School of Medicine, Kyung Hee University, Seoul, South Korea. [2]Department of Biochemistry and Molecular Biology, School of Medicine, Kyung Hee University, Seoul, South Korea. [3]Department of Preventive Medicine, College of Medicine, Hanyang University, Seoul, South Korea. [4]Department of Public Health Science, College of Medicine, Hanyang University, Seoul, South Korea. [5]Department of Public Health Science, Seoul National University, Seoul, South Korea. [6]Department of Statistics, Seoul National University, Seoul, South Korea. ✉email: jelim@khu.ac.kr; ohbs@khu.ac.kr

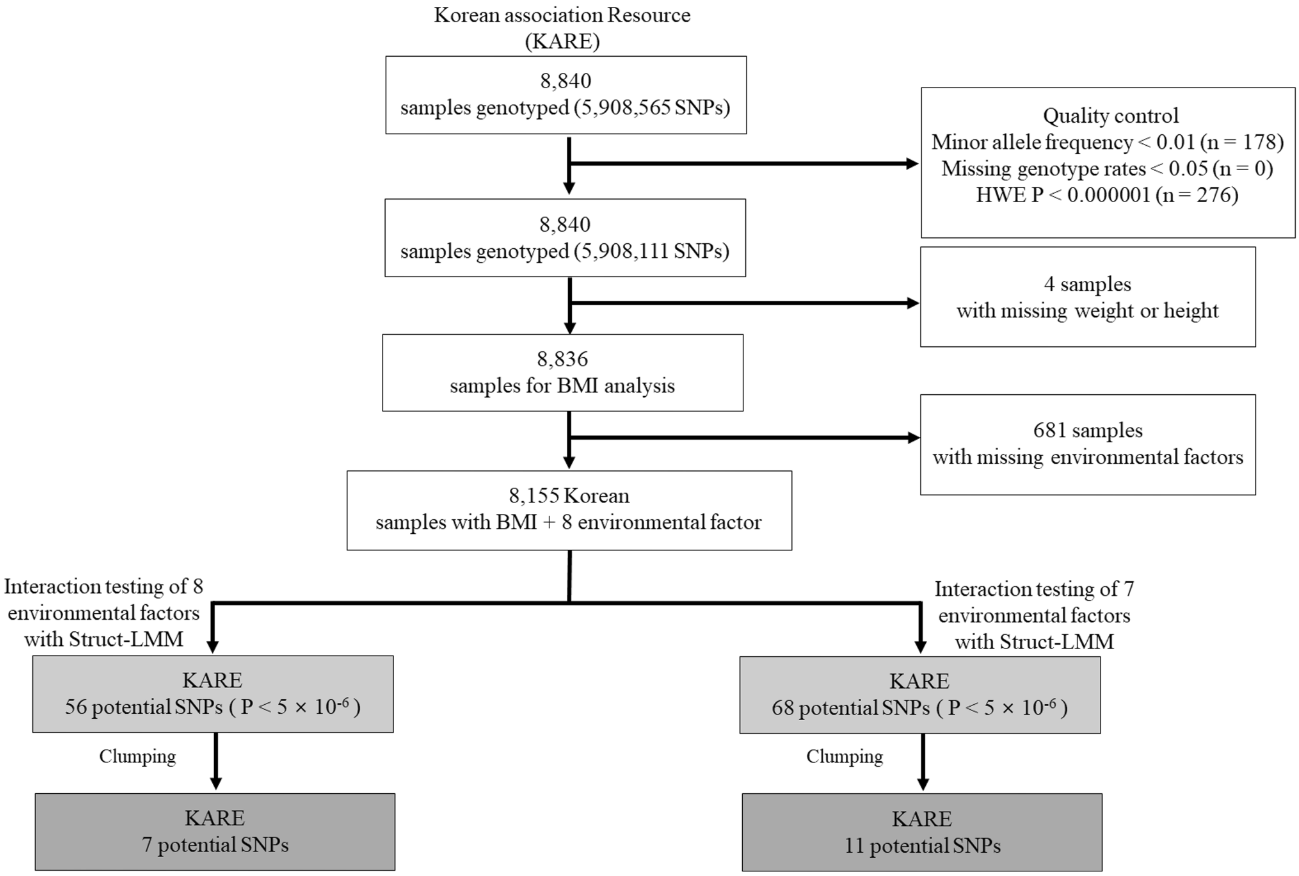
Korean association Resource (KARE)
8,840 samples genotyped (5,908,565 SNPs)
Quality control
Minor allele frequency < 0.01 (n = 178)
Missing genotype rates < 0.05 (n = 0)
HWE P < 0.000001 (n = 276)
8,840 samples genotyped (5,908,111 SNPs)
4 samples with missing weight or height
8,836 samples for BMI analysis
681 samples with missing environmental factors
8,155 Korean samples with BMI + 8 environmental factor
Interaction testing of 8 environmental factors with Struct-LMM
Interaction testing of 7 environmental factors with Struct-LMM
KARE 56 potential SNPs (P < 5 × 10-6)
KARE 68 potential SNPs (P < 5 × 10-6)
Clumping
Clumping
KARE 7 potential SNPs
KARE 11 potential SNPs

Figure 1. Diagram showing the study design. Left side of the diagram depicts the interaction between 8 environmental factors and 5,908,111 single nucleotide polymorphisms (SNPs), and the right side of the diagram depicts the interaction between 7 environmental factors (without physical activity) and 5,908,111 SNPs. The numbers in the parenthesis indicate the SNPs excluded by the specified quality control. *BMI* body mass index.

association studies (GWAS)[9]. Among the GWAS single nucleotide polymorphisms (SNPs), the *FTO* locus is the strongest genetic variant and is recently reported to affect BMI through interaction with environmental factors, including physical activity, diet, and smoking[7,10–13].

Most of the interaction studies for obesity calculate the interaction between a single genetic factor or multiple genetic factors and a single environment factor[9,14–17]. However, there is a possibility that the interaction of multiple environments and a single genetic factor can also affect the phenotype[7,18]. Recently, the structured linear mixed model (Struct-LMM) analysis was reported to identify the genetic loci, characterized by the interaction of multiple environmental factors[18]. Using this method, high-dimensional environmental data can be used in population cohorts to help understand the effects of genetic factors in a group on complex traits and diseases[18].

In this study, we applied the Struct-LMM method to the Korea Association Resource (KARE), a well-known Korean population GWAS database, comprised of 8,840 participants[19]. A total of 8 obesity-related environmental factors, which could interact with multiple environmental factors for BMI, were examined on genome-wide genetic variants. These environments included estimated daily alcohol consumption, education, physical activity metabolic equivalent of task (PAMET), income, total calorie intake, protein intake, carbohydrate intake, and smoking status.

Materials and methods

Korean association resource (KARE) cohort.

We utilized data from KARE for all the analyses. Participants of KARE cohort were recruited from two regions in South Korea (Ansan and Ansung) from 2009 to 2012 for the Korean Genome and Epidemiology Study. All study participants aged ≥ 40 years provided written informed consent, and approval was obtained from the institutional review board. The exclusion criteria were as follows: history of cancer, gender inconsistencies, cryptic relatedness, low genotype call rate (< 95%), and sample contamination (Fig. 1)[19,20].

From the 8,840 samples, 8,155 samples (with all the 8 environmental factors related to obesity) were selected. The distribution of the variables in the samples has been summarized in Table 1.

Genotype data.

The KARE study utilized the Affymetrix Genome-Wide Human SNP Array GeneChip 5.0[19]. SNP imputation was performed using IMPUTE2 with the 1000 Genomes Project (haplotype phase 1)[19]. At

	Underweight	Normal weight	Overweight	Obese
KARE cohort	BMI < 18.5	(18.5 ≤ BMI < 23)	(23 ≤ BMI < 25)	(BMI ≥ 25)
Number of participants	146	2347	2138	3524
Males (%)	60.96%	51.34%	47.99%	47.21%
age (mean, SD)	56.35(9.74)	52.23(9.38)	51.56(8.73)	52.10(8.54)
Height (cm) (mean, SD)	160.66(7.73)	160.62(8.53)	160.33(8.47)	159.69(8.94)
Weight (kg) (mean, SD)	45.60(5.29)	55.29(6.55)	61.91(6.70)	70.06(8.88)
Body mass index (kg/m^2) (mean, SD)	17.60(0.82)	21.37(1.16)	24.02(0.57)	27.42(2.07)
For estimated day consumption of alcohol (g/day) (mean, SD)	1.84(1.27)	1.78(1.13)	1.79(1.10)	1.78(1.13)
Education				
Less than primary school	71	736	637	1214
Completed middle school	28	530	470	796
Completed high school	30	781	724	1020
Completed college	5	90	80	116
Completed university	12	183	195	319
More than university	0	27	32	59
PAMET (mean, SD)	11,288.22(7820.73)	10,029.26(6481.24)	9519.96(6069.51)	9271.68(6038.71)
Income				
x < 500,000 won	55	465	349	611
500,000 won ≤ x < 1,000,000 won	23	396	340	504
1,000,000 won ≤ x < 1,500,000 won	21	345	318	548
1,500,000 won ≤xx < 2,000,000 won	18	333	310	479
2,000,000 won ≤ x < 3,000,000 won	17	446	382	652
3,000,000 won ≤ x < 4,000,000 won	7	226	251	365
4,000,000 won ≤ x < 6,000,000 won	4	102	144	222
6,000,000 won < x	1	34	44	93
Total calorie intake (kcal) (mean, SD)	1866.09(739.49)	1925.13(686.87)	1940.74(687.59)	1975.49(690.45)
Protein intake (g) (mean, SD)	61.75(30.15)	65.12(31.18)	66.49(27.62)	67.66(29.29)
Carbohydrate intake (g) (mean, SD)	332.10(134.90)	339.01(111.35)	340.36(119.58)	347.41(115.20)
Smoking status				
Never	62	1294	1247	2178
Previous	22	340	343	588
Sometimes	10	64	51	99
Often	10	649	470	659

Table 1. Basic characteristics of Korean Association Resource (KARE) participants included in this study, stratified by weight as a criterion. The numbers with percentile indicate mean and standard deviation (SD), and the numbers without percentile indicate the sample number in the variable.

the baseline, genetic data was available for 6,461,358 SNPs in samples of 8,840 KARE participants. We performed quality control based on the following exclusion criteria: variants with missing genotype call rates > 0.05, minor allele frequency < 0.01, and Hardy–Weinberg equilibrium P value $< 1.00 \times 10^{-6}$. Based on these criteria, 5,908,111 SNPs were included in the study (Fig. 1).

Dataset for phenotype and environmental factors. The methods for the measurement of height and weight have been described in a previous study[20]. BMI was calculated as weight (kg)/height (m^2). For analyses, BMI was transformed to normal distribution using Gaussian function in Struct-LMM.

The environmental factors used in Struct-LMM were selected based on previous studies on BMI[9,14,16,17]. The 8 environmental factors were alcohol consumption[9], education[14], physical activity[9], income[9], total calorie intake[16], protein intake[16], carbohydrate intake[16], and smoking status[17] (Supplementary Table 1).

The estimated daily consumption of alcohol (g/day) was calculated, as described previously[20]. PAMET was obtained from each participant using a structured questionnaire that included four types of physical activities including sleeping, five different sedentary activities, non-sedentary activities, and only leisure-time physical activity[21]. The total calorie intake, protein intake, and carbohydrate intake were calculated from a food frequency questionnaire (FFQ), as described[22,23].

For the Struct-LMM interaction analysis of alcohol consumption, the amount of alcohol consumption additionally increased in each group by 20 g/day from the 1st group (less than 20 g/day) up to the 6th group (> 100 g/day)[24]. For education analysis, participants were divided into two groups, one with education no more than secondary school and another with education more than secondary school. The physical activity was analyzed from PAMET score as a continuous variable. For income analysis, participants were divided into groups based on various levels of income, as shown in Supplementary Table 1. For the analysis of total calorie intake, protein intake, and carbohydrate intake, participants were divided into four groups, as shown in Supplementary Table 1. For smoking analysis, participants were divided into two groups, one with no or previous history of smoking and the other with current smoking[17].

Statistical analysis. SNP quality control was performed in PLINK v.1.9.0, as described previously[25]. The gene-environment interaction on BMI for the eight environmental factors was analyzed by Struct-LMM v.0.3.1[18], adjusted for age, sex, and recruitment area[26].

We performed Manhattan plot drawing, box plot drawing, association analysis, residual value calculation, bar graph drawing, and correlation in R stats package version 3.5.1 (www.r-project.org). For Manhattan plot drawing, qqman package was used for residual value calculation, while for association analysis, lme4 package in R stats package was used. We calculated the Bayes Factor (BF) for each environmental factor to explore the most relevant environments for GEI in Struct-LMM v.0.3.1. The Bayes Factor method used in Struct-LMM is a statistical method comparing two models, one with the environmental factors and another without the environmental factors, in order to assess which model is better by quantifying the power of each model.

We performed a gene-based genome-wide association analysis using the MAGMA[27] tool provided by FUMA[28] through GWIS results calculated by Struct-LMM in this study.

Ethics approval and consent to participate. This study was performed in accordance with the World Medical Association Declaration of Helsinki. All participants provided written informed consent to participate in the study[19]. Approval for the study was obtained from the Institutional Review Board (IRB) of Kyung Hee University (KHSIRB-19-387(EA), KHSIRB-20-077(EA)).

Results

Association of environmental factors on BMI. The environmental characteristics of the 8155 participants, included in this Struct-LMM analysis, have been summarized in Table 1. The participants were categorized into four obesity groups, based on the BMI values stated by the World Health Organization (Asia–Pacific region) and the Korean Obesity Society[29]. The average BMI of the participants was 24.61 (standard deviation = 3.12).

The relationship between the environmental factors was assessed using a correlogram (Supplementary Fig. 1). It was observed that PAMET and carbohydrate intake had relatively low correlation with other environmental factors, compared to the other 6 factors.

All the 8 environmental factors showed an association with BMI (P-value < 0.05). However, the associations of alcohol consumption (P-value $= 2.83 \times 10^{-2}$), education (P-value $= 7.02 \times 10^{-3}$), protein intake (P-value $= 9.94 \times 10^{-3}$), and carbohydrate intake (P-value $= 1.02 \times 10^{-2}$) with BMI, were not valid after the Bonferroni correction for multiple comparisons (P-value $< 6.2 \times 10^{-3}$) (Supplementary Table 1). Alcohol consumption, income, total calorie intake, and protein intake were positively correlated with BMI, while education, PAMET, carbohydrate intake, and smoking were negatively correlated with BMI. The order of effect size of the 8 environmental factors, regardless of positive or negative correlation with BMI, was as follows: protein intake classified as quartile ($\beta = 0.81$), alcohol consumption classified as 6 groups ($\beta = 0.75$), smoking status classified as smoker and non-smoker ($\beta = -0.65$), education classified as 2 groups ($\beta = -0.21$), total calorie intake classified as quartile ($\beta = 0.15$), carbohydrate intake classified as quartile ($\beta = 0.08$), income classified as 8 groups ($\beta = -0.07$), and PAMET as a continuous variable ($\beta = -2.81 \times 10^{-5}$) (Supplementary Table 1).

Analysis of gene–environment interaction on BMI using Struct-LMM. We performed the GEI test between the 5,908,111 SNPs and 8 environmental factors on BMI, using Struct-LMM adjusted for age, sex, and recruitment area. The study design has been depicted in Fig. 1. Based on the results, we did not find an association of genome-wide significance with P-value $< 5 \times 10^{-8}$, as shown in the Manhattan plot (Fig. 2). However, we found 7 potential associations with P-value $< 5 \times 10^{-6}$ (Table 2).

For each potential genetic variant, we calculated the BF value of each environmental factor to examine the interaction reliability of the respective environmental factor on BMI, using the tool included in the Struct-LMM program[18]. For rs2391333, showing the most significant interaction (P-value $= 9.17 \times 10^{-8}$), the BF value of total calorie intake was the highest (BF = 8.43) among the 8 environmental factors, while the BF values of both carbohydrate intake and PAMET were the lowest (BF < 0.01 each, Fig. 3). When the BF values of the environmental factors were added from all the 7 potential SNPs, the BF value of PAMET was the lowest among the 8 environmental factors (Fig. 3, Supplementary Table 2). To confirm whether this was true for other SNPs as well, we recalculated the BF value of the 8 environmental factors from 66 independent lead SNPs that showed an interaction association with P-value $< 5 \times 10^{-5}$ (Supplementary Table 3). A BF value < 1 indicates that the environmental factors, through the GEI, does not improve the test power of Struct-LMM model. In another word, the environmental factor does not help in identifying genetic factors affecting BMI through the interaction. When the BF from all 66 SNPs were added, the BF value of PAMET was found to be the lowest compared to other environmental factors ($BF_{PAMET} = 57.95$). In addition, the average BF value of PAMET was < 1. (BF_{PAMET} average = 0.88, $BF_{Alcohol\ consumption}$ average = 1.30, $BF_{Education}$ average = 2.78, BF_{Income} average = 3.34, $BF_{Smoking\ status}$ average = 1.17, $BF_{Total\ calorie\ intake}$ average = 1.41, $BF_{Protein\ intake}$ average = 2.87, and $BF_{Carbohydrate\ intake}$ average = 1.50) (Supplementary Table 3). Therefore, we analyzed the GEI again by Struct-LMM, after excluding PAMET.

Second analysis of gene–environment interaction using 7 environmental factors. We performed the GEI test between the 5,908,111 SNPs and 7 environmental factors (without PAMET) on BMI, using Struct-LMM adjusted for age, sex, and recruitment area (Fig. 1). We discovered 11 independent potential genetic

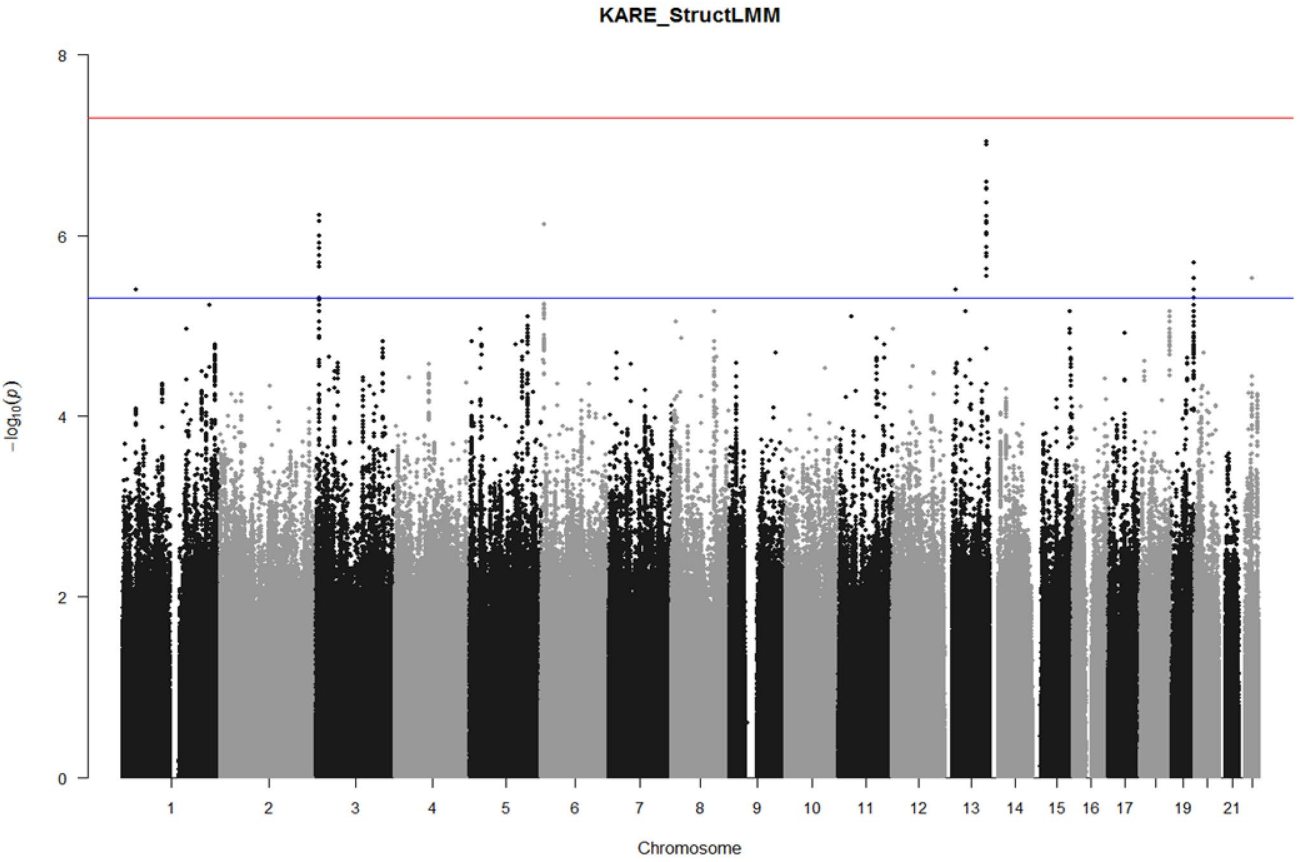

Figure 2. Manhattan plots for genome-wide interaction study using eight environmental factors. The red line indicates the suggestive threshold (P value < 5.00 × 10^{-6}). The blue line indicates the genome-wide significance threshold (P value < 5.00 × 10^{-8}).

Chromosome	Nearest gene	SNP ID	Position[a]	Minor allele	Major allele	MAF[b] (%)	*P* value[c]
1	*GJB5*	rs12402440	107,166,694	G	T	20.91	4.03×10^{-6}
3	*LOC100288428*	rs59756727	7,656,037	A	T	44.08	6.02×10^{-7}
6	*SNRNP48*	rs7760212	8,242,302	A	G	32.02	7.47×10^{-7}
13	*EFNB2**	rs2391333	34,968,257	T	C	41.42	9.17×10^{-8}
13	*MTIF3*	rs9512706	28,041,615	A	C	19.87	3.94×10^{-6}
19	*ZNF787*	rs668056	56,595,733	T	C	28.74	2.22×10^{-6}
22	*ISX*	rs5755279	35,186,350	G	A	23.91	3.21×10^{-6}

Table 2. Interactions between individual single nucleotide polymorphisms (SNPs) and 8 environmental factors. [a]Chromosomal positions are based on the 1000 Genomes Project's haplotype phase 1 in NCBI build 37 (hg19). [b]MAF, minor allele frequency. [c]The P value for the effects of interaction between genotypes and environmental factors on BMI was assessed by using a Struct-LMM with adjustment for age, sex, and recruitment area. The nearest gene to which the SNP is located.

variants with P-value < 5 × 10^{-6} (Table 3, Fig. 4). A variant, rs2391331, located within the intron of *EFNB2* gene (Chr 13), showed a genome-wide significance (P-value = 7.27 × 10^{-9}).

We analyzed the interaction association of the other 6 potential variants identified from the analysis with the 8 environmental factors, including PAMET. While three SNPs (rs12402440, rs59756727, and rs9512706) had lower P-values in the interaction test involving 7 environmental factors, the other 3 SNPs (rs7760212, rs668056, and rs5755279) had higher P-values (Supplementary Table 4).

Since waist and hip circumferences were also used as indices for obesity, we analyzed the interaction of rs2391331 with 7 environmental factors on both these traits by Struct-LMM analysis. Both the correlation between BMI and waist circumference was 0.76 (P value < 5.00 × 10^{-16}), and the correlation between BMI and hip circumference was also 0.76 (P value < 5.00 × 10^{-16}) in this study population. Results showed that rs2391331

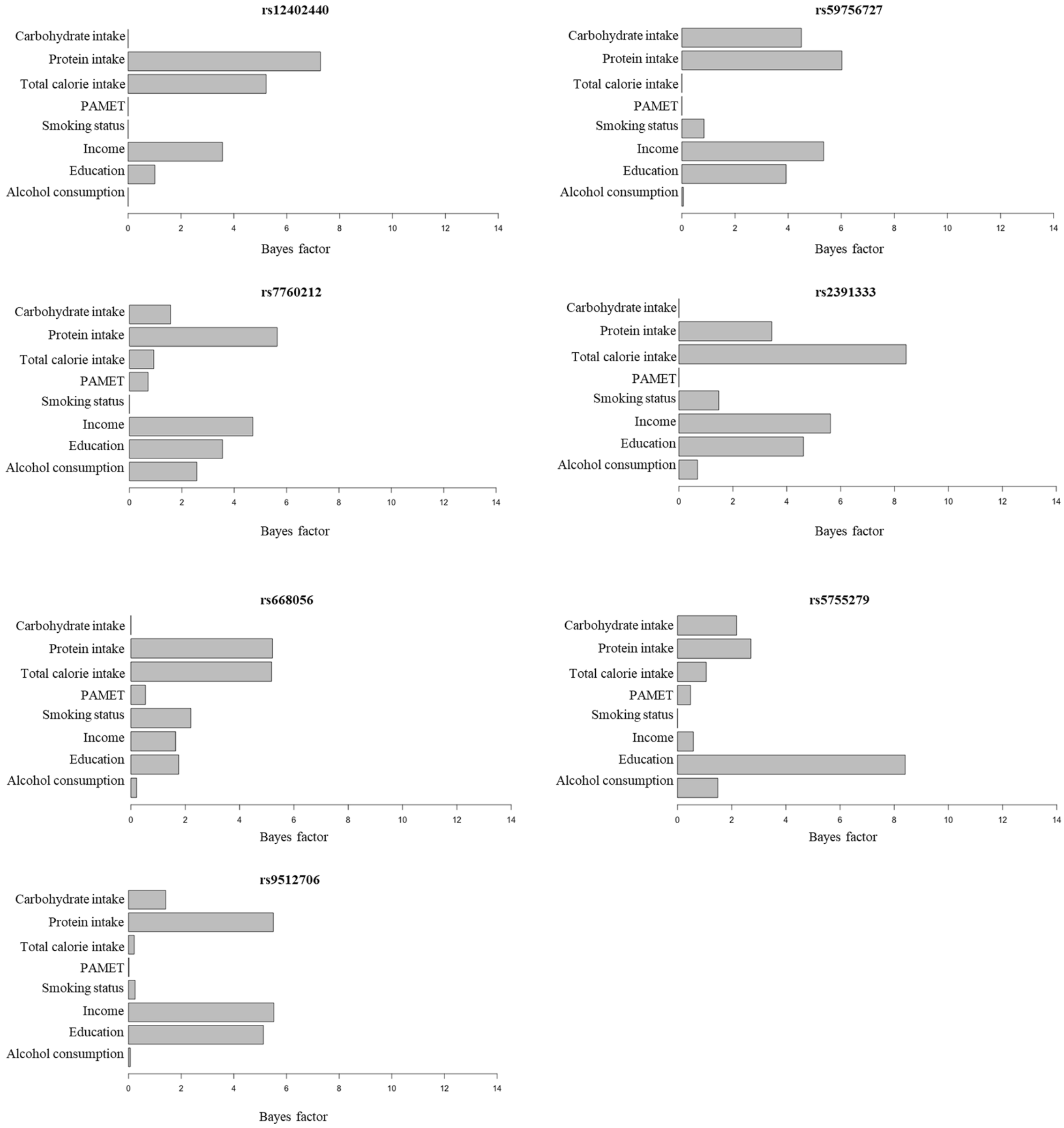

Figure 3. Interaction Bayes factor values of environmental factors for 7 potential single nucleotide polymorphisms (SNPs). Bayes Factor shows evidence of environmental factors that explain GEI at potential SNP.

interacted with the 7 environmental factors on waist circumference (P value = 1.50×10^{-5}) and hip circumference (P value = 6.00×10^{-6}) (Supplementary Table 5).

The Struct-LMM analysis examined the interaction of a genetic variant with multiple environment factors as a whole. Hence, we investigated the interaction of individual environmental factors with rs2391331 using a fixed effect model of linear regression. As shown in Supplementary Table 6, rs2391331 showed interaction P value with protein intake (P-value = 1.64×10^{-4}), income (P value = 3.62×10^{-4}), total calorie intake (P-value = 6.27×10^{-4}), alcohol consumption (P-value = 2.90×10^{-3}), smoking status (P-value = 3.77×10^{-3}), education (P value = 1.88×10^{-2}), and carbohydrate intake (P-value = 5.28×10^{-1}).

Gene-based genome-wide association analysis. In order to validate the previous results of Struct-LMM and also identify the causative genes, we performed a gene-based genome-wide association analysis using the MAGMA[26] tool through GWIS results calculated from Struct-LMM (Supplementary Fig. 3). SNPs were mapped to 17,535 protein-coding genes, making the genome-wide significance level as defined at P-value = 2.85×10^{-6} (0.05/17,535). Three genes of *EFNB2* (Chr 13), *DOCK4* (Chr 7), and *ZNF787* (Chr 19) met the genome-wide significance level, and the P-value of 3 genes were as follows: *EFNB2* P-value = 5.03×10^{-10}, *DOCK4* P-value = 6.79×10^{-7}, and *ZNF787* P-value = 2.34×10^{-6} (Supplementary Table 7). The SNPs with the highest GEI P-value in the gene locus were as follows: [*DOCK4*—rs3801778 (GEI P-value = 8.90×10^{-5})],

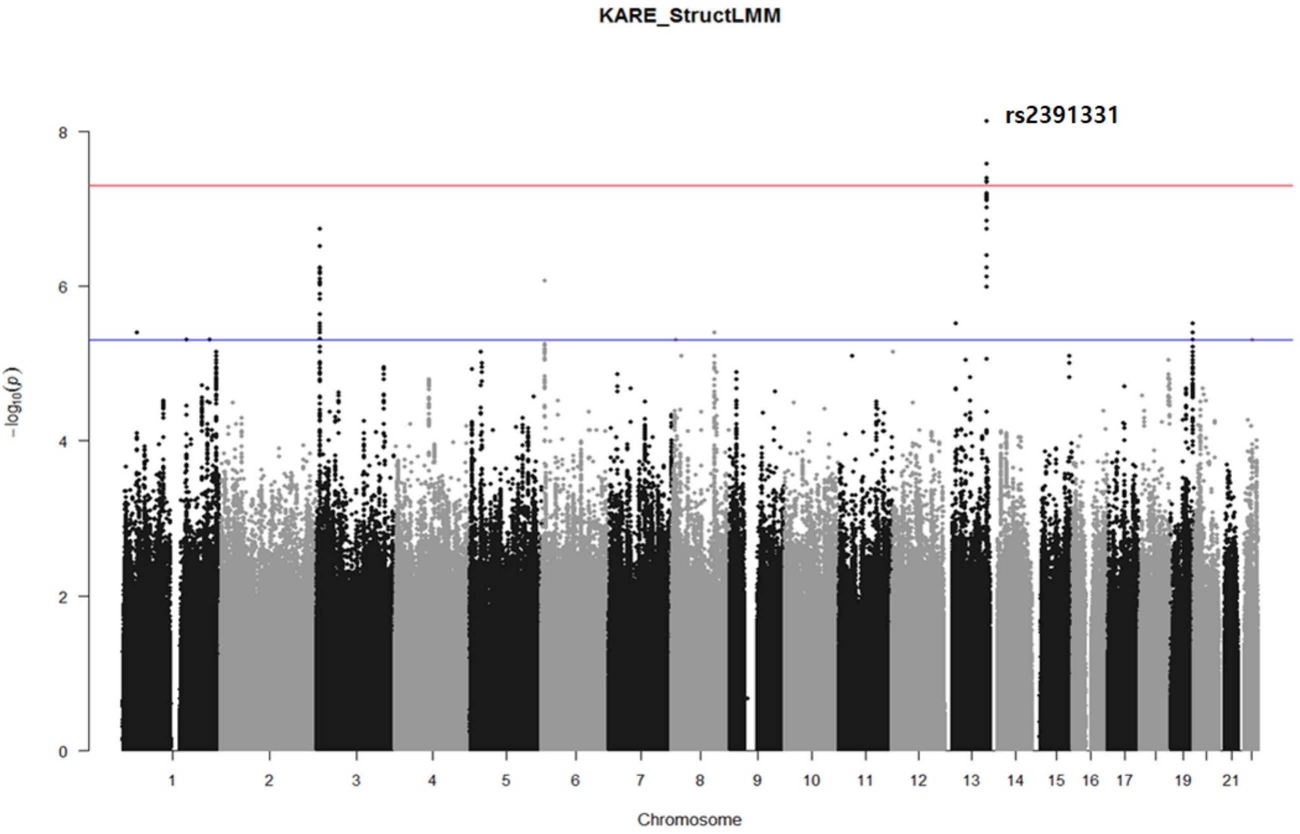

Figure 4. Manhattan plots for genome-wide interaction study using 7 environmental factors. The red line indicates the suggestive threshold (P value < 5.00×10^{-6}). The blue line indicates the genome-wide significance threshold (P value < 5.00×10^{-8}).

Chromosome	Nearest gene	SNP ID	Position[a]	Minor allele	Major allele	MAF[b] (%)	P value[c]
1	*GJB5*	rs12402440	34,968,257	G	T	20.91	4.02×10^{-6}
1	*C1orf110*	rs12405096	162,926,260	T	C	41.78	4.60×10^{-6}
1	*RAB3GAP2**	rs74565497	220,415,930	G	A	8.12	4.82×10^{-6}
3	*LOC100288428*	rs59756727	8,242,302	A	T	44.08	1.81×10^{-7}
6	*SNRNP48*	rs7760212	7,656,037	A	G	32.02	8.62×10^{-7}
8	*MSRA**	rs117301188	10,070,612	A	C	9.62	4.70×10^{-6}
8	*RSPO2**	rs7833349	108,966,199	T	C	24.70	3.81×10^{-6}
13	*MTIF3*	rs9512706	28,041,615	A	C	19.87	2.91×10^{-6}
13	*EFNB2**	rs2391331	107,157,709	C	T	42.90	7.27×10^{-9}
19	*ZNF787*	rs668056	56,595,733	T	C	28.74	2.87×10^{-6}
22	*ISX*	rs5755279	35,186,350	G	A	23.91	4.65×10^{-6}

Table 3. Interactions between individual single nucleotide polymorphisms (SNPs) and 7 environmental factors. [a]Chromosomal positions are based on the 1000 Genomes Project's haplotype phase 1 in NCBI build 37 (hg19). [b]MAF is minor allele frequency. [c]The P value for the effects of interaction between genotypes and environmental factors on BMI were assessed by using a Struct-LMM with adjustment for age, sex, and recruitment area. The nearest gene to which the SNP is located.

[*EFNB2*—rs2391331 (GEI P-value = 7.27×10^{-9})], and [*ZNF787*—rs642776 (GEI P-value = 3.00×10^{-6})]. We summarized the BF values of each environmental factor for rs3801778 and rs642776 in Supplementary Table 8. Further, we investigated the expression quantitative trait loci (eQTL) information using Genotype-Tissue Expression (GTEx) version 8[30]. We could find eQTLs of LD-linked proxy SNPs ($r^2 > 0.8$) from rs2391331 and rs642776, and summarized the results in Supplementary Table 9. The proxy SNPs, rs10508174, rs11069646, rs7983579, and rs7327929 for rs2391331 are associated with *EFNB2* gene in the pituitary or testis. The proxy SNPs, rs7250351, rs1007851, rs6509982, rs493717, and rs35766803 for rs642776 are associated with *ZNF787* gene in fibroblast cells, esophagus, or skin, and the proxy SNPs, rs6509982 and rs35766803 for rs642776 are associated with *TMEM190* gene in the lung.

BF of gene–environment associated variant. The BF of environmental factors for the 11 potential SNPs were calculated (Fig. 5, Supplementary Table 10). For the most significant variant, rs2391331, all the environmental factors (except carbohydrate intake), showed BF value > 1, thus indicating the evidence of interaction. The interaction BF values of rs2391331 were as follows: protein intake (BF = 5.28), income (BF = 4.59), total calorie intake (BF = 4.12), alcohol consumption (BF = 2.85), smoking status (BF = 2.63), and education (BF = 1.41) (Supplementary Table 10). Additionally, we investigated the interaction of individual environmental factors with the 10 potential SNPs using a fixed effect model of linear regression (Supplementary Table 11–20). As a result, the most significant *P*-value ($P\text{-value} = 7.65 \times 10^{-7}$) was observed with total calorie intake interacted with rs11730118, which also showed the highest BF value (BF = 10.12) among the 11 potential SNPs of Struct-LMM analysis (Fig. 5, Supplementary Table 10, and Supplementary Table 16).

Box plots showed distribution of the in sample estimated allelic effect size on BMI considering environmental factors for 11 potential variants with GEI (n = 8155, unrelated individuals of Korean population) (Supplementary Fig. 2).

Discussion

In this study, we performed an interaction test between multiple environmental factors and genetic variant, using Struct-LMM at a genome-wide level[18]. The identified genetic variant, rs2391331 was significantly associated with the interaction of 7 environmental factors in a Korean population cohort, KARE ($P\text{-value} = 7.27 \times 10^{-9}$). Of the 7 environmental factors, protein intake was the most influential environmental factor for rs2391331, and the least influential environmental factor was carbohydrate intake (Fig. 5).

A certain genetic variant may affect a trait through multiple environmental factors, as shown in *FTO* and *MC4R*[31–33]. The genetic variant of *FTO* showed interactions with diverse environmental exposures such as physical activity, diet, and alcohol consumption. The effect size of *FTO* variants on BMI was reduced by the increased physical activity but increased by the decreased physical activity[10]. Struct-LMM was modelled to identify such interactions with multiple environmental factors, which may not be identified otherwise. As shown in Supplementary Table 6, rs2391331 could not be identified as a genome-wide significant variant through the conventional fixed effect model approach, which used a single environmental factor. While protein intake, income, total calorie intake, alcohol consumption, and smoking status showed statistical significance after multiple correction, none of these environmental factors showed a genome-wide significance for the interaction with rs2391331 ($P\text{-value} < 5.00 \times 10^{-8}$). Therefore, unless the Struct-LMM model was applied, rs2391331 could not be identified through the genome-wide interaction study.

Another advantage of the Struct-LMM analysis is that it provides a rank about which environmental factor is more reliable for the interaction with the genetic variant, using BF values. As shown in Fig. 5 and Supplementary Table 10, the BF values for rs2391331 indicated that the ranking was in the order of protein intake (BF value = 5.28), income (BF = 4.59), total calorie intake (BF = 4.12), alcohol consumption (BF = 2.85), smoking status (BF = 2.63), education (BF = 1.41), and carbohydrate intake (BF < 0.01). When we investigated the interaction of individual environmental factors with rs2391331 using a fixed effect model of linear regression, the *P*-values of environmental factors were in the order of *P*-value, as follows: protein intake ($P\text{-value} = 1.64 \times 10^{-4}$), income ($P\text{-value} = 3.62 \times 10^{-4}$), total calorie intake ($P\text{-value} = 6.27 \times 10^{-4}$), alcohol consumption ($P\text{-value} = 2.90 \times 10^{-3}$), smoking status ($P\text{-value} = 3.77 \times 10^{-3}$), education ($P\text{-value} = 1.88 \times 10^{-2}$), and carbohydrate intake ($P\text{-value} = 5.28 \times 10^{-1}$) as shown in Supplementary Table 6. The order of *P*-value of each environmental factor was the same as the one of BF value obtained by Struct-LMM. In case of rs2391331, all 6 environmental factors (except carbohydrate intake) had a rather consistent effect on BMI through the interaction.

The genetic variant, rs2391331 was located in the first intron of *EFNB2* gene (Chr 13). EFNB2 is the ligand of erythropoietin-producing hepatocellular kinases (EPH), the largest family of receptor tyrosine kinases[34, 35]. We performed the gene-based genome-wide association analysis and verified that at the gene level *EFNB2* was also statistically significant for the interaction with the 7 environmental factors (Gene-set $P\text{-value} = 5.03 \times 10^{-10}$) (Supplementary Table 7). Further, we found that LD-linked proxy SNPs for rs2391331 were associated with *EFNB2* gene in the pituitary, or testis (Supplementary Table 9). *EFNB2* has been reported to be associated with schizophrenia and ankle injury through the genome-wide association study. Also, several studies have shown the association of this gene hypertension and type 2 diabetes; however, there have been no reports regarding the connection of *EFNB2* with obesity-related traits[36–38]. Recently, it was reported that smooth muscle-specific deletion of *EFNB2*, using SM22Δ-Cre, resulted in lower body weight, reduced vascular smooth muscle cell proliferation, wall, and enlarged arterial diameter[39]. More recently, it has been reported that the expression of *EFNB2* is enriched in proopiomelanocortin (POMC) neurons, a major regulator of energy balance and glucose homeostasis and the loss of *EFNB2* in POMC-expressing progenitors mildly impairs gluconeogenesis and food intake in mice[40]. We do not know yet whether how this report would be related to the findings in this study. Further studies are required to investigate the functional analysis of *EFNB2* related to obesity.

Additionally, we found 4 publications related to the GEI on BMI in which studies 96 SNPs were reported[18,41–43]. We tested the 96 SNPs for replication using the Struct-LMM in this study population, however we could not find any significant interaction under the multiple correction criteria ($P\text{-value} < 0.05 / 96$) (Supplementary Table 21). We may provide the reasons for the failure of replication as follows. First, our sample size (N = 8,155) may not be enough for replication. It is generally accepted that for GEI studies, the bigger sample size is recommended than the size for GWAS[2]. Second, the ethnicity of our sample (East Asian population) is different from the ethnicity of previously mentioned 4 studies (3 studies in European population[18,41,43]; 1 study in African and Hispanic

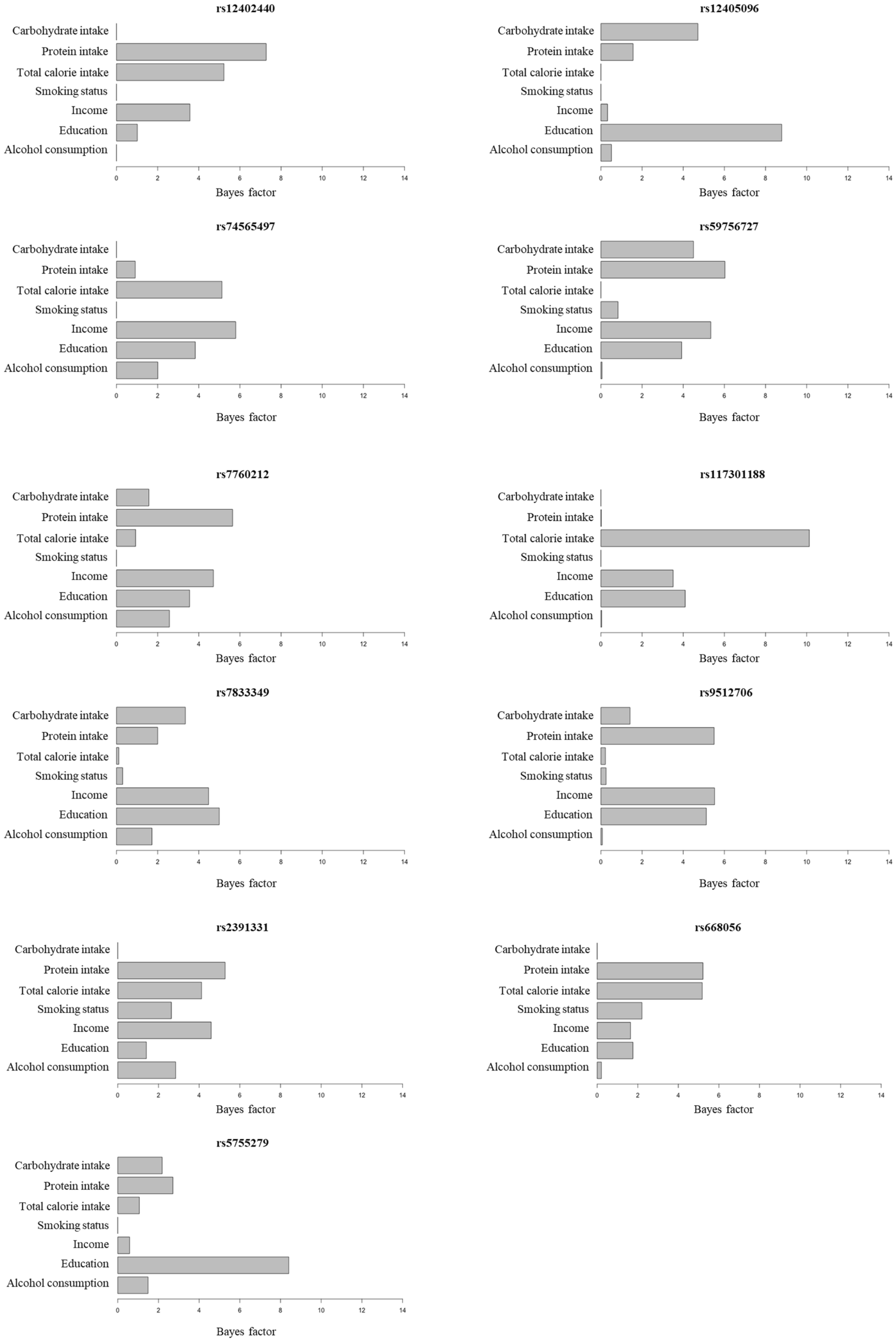

Figure 5. Interaction Bayes factor values of environmental factors for 11 potential single nucleotide polymorphisms (SNPs). Bayes Factor shows evidence of environmental factors that explain GEI at potential SNP.

population[42]). The effects of environmental factors are supposed to be discordant between different ethnical backgrounds, making it hard to produce a replication. Third, there may be a limitation in obtaining the precise environmental data. The environmental data acquired from self-reported questionnaire (i.e., dietary intakes, alcohol, smoking, sociodemographic factors, and physical activities) may be prone to responder bias[9].

There are several limitations to our study. First, as mentioned above, we analyzed the GEI in obesity with a small sample size, which can affect the statistical power and lead to imprecise or incorrect estimates of the magnitude of observed effects[44]. In addition, the number of environmental factors investigated was small. While Struct-LMM analysis has good detection power, even if more than 10 environmental factors are analyzed, a smaller number of environmental factors reduces the detection power of this method[18]. Lastly, the results of rs2391331 were not validated in other cohorts or ethnicities. Although it is not easy to find cohorts with all the diverse environmental factors similar to this study, replication of the results in other cohorts is warranted.

In conclusion, we performed multiple environments-gene interaction analysis to identify potential SNPs of BMI in a Korean cohort. A genome-wide significant interaction of rs2391331, located in the *EFNB2* locus, was identified and the interaction on BMI was influenced by 6 environmental factors, namely protein intake, income, total calorie intake, alcohol consumption, smoking status, and education.

Received: 27 March 2020; Accepted: 5 February 2021
Published online: 02 March 2021

References

1. Wright, A. F., Carothers, A. D. & Campbell, H. Gene–environment interactions–the BioBank UK study. *Pharmacogenomics J.* **2**, 75–82 (2002).
2. Hunter, D. J. Gene–environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
3. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**, 843–854 (2010).
4. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
5. Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene–environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–119 (2007).
6. Watson, N. F. *et al.* Sleep duration and body mass index in twins: a gene-environment interaction. *Sleep* **35**, 597–603 (2012).
7. Bjornland, T., Langaas, M., Grill, V. & Mostad, I. L. Assessing gene-environment interaction effects of FTO, MC4R and lifestyle factors on obesity using an extreme phenotype sampling design: Results from the HUNT study. *PLoS ONE* **12**, e0175071 (2017).
8. Organisation for Economic Co-operation and Development. Directorate for Employment, Labour and Social Affairs. 2014 [Obesity update June 2014]; Available from: http://www.oecd.org/health/Obesity-Update-2014.pdf.
9. Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, A. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLoS Genet.* **13**, e1006977 (2017).
10. Kilpelainen, T. O. *et al.* Physical activity attenuates the influence of FTO variants on obesity risk: A meta-analysis of 218,166 adults and 19,268 children. *PLoS Med.* **8**, e1001116 (2011).
11. Ahmad, S. *et al.* Gene x physical activity interactions in obesity: Combined analysis of 111,421 individuals of European ancestry. *PLoS Genet.* **9**, e1003607 (2013).
12. Young, A. I., Wauthier, F. & Donnelly, P. Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nat.. Commun.* **7**, 12724 (2016).
13. Qi, Q. *et al.* Fried food consumption, genetic risk, and body mass index: Gene-diet interaction analysis in three US cohort studies. *BMJ* **348**, g1610 (2014).
14. Jones-Smith, J. C., Gordon-Larsen, P., Siddiqi, A. & Popkin, B. M. Emerging disparities in overweight by educational attainment in Chinese adults (1989–2006). *Int. J. Obes. (Lond.)* **36**, 866–875 (2012).
15. Silventoinen, K. *et al.* Trends in obesity and energy supply in the WHO MONICA Project. *Int. J. Obes. Relat. Metab. Disord.* **28**, 710–718 (2004).
16. Layman, D. K. *et al.* A moderate-protein diet produces sustained weight loss and long-term changes in body composition and blood lipids in obese adults. *J. Nutr.* **139**, 514–521 (2009).
17. Sun, K., Liu, J. & Ning, G. Active smoking and risk of metabolic syndrome: a meta-analysis of prospective studies. *PLoS ONE* **7**, e47791 (2012).
18. Moore R., Casale F. P., Jan Bonder M., Horta D., Consortium B. *et al.* A linear mixed-model approach to study multivariate gene-environment interactions. *Nat Genet* **51**, 180–6 (2019).
19. Cho, Y. S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* **41**, 527–534 (2009).
20. Kim, Y., Han, B. G. & Ko, G. E. S. G. Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium. *Int J Epidemiol* **46**, 1350 (2017).
21. Kim, K., Shin, Y. J., Nam, J. H., Choi, B. Y. & Kim, M. K. A dose-response relationship between types of physical activity and distress. *J. Korean Med. Sci.* **23**, 218–225 (2008).
22. Ahn, Y. *et al.* Validation and reproducibility of food frequency questionnaire for Korean genome epidemiologic study. *Eur. J. Clin. Nutr.* **61**, 1435–1441 (2007).
23. Kim, J. *et al.* Development of a food frequency questionnaire in Koreans. *Asia Pac. J. Clin. Nutr.* **12**, 243–250 (2003).
24. Cairns, V., Keil, U., Kleinbaum, D., Doering, A. & Stieber, J. Alcohol consumption as a risk factor for high blood pressure Munich Blood Pressure Study. *Hypertension* **6**, 124–131 (1984).
25. Hong, K. W. *et al.* Identification of three novel genetic variations associated with electrocardiographic traits (QRS duration and PR interval) in East Asians. *Hum. Mol. Genet.* **23**, 6659–6667 (2014).
26. Hong, K. W. *et al.* Recapitulation of the association of the Val66Met polymorphism of BDNF gene with BMI in Koreans. *Obesity (Silver Spring)* **20**, 1871–1875 (2012).
27. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
28. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
29. Consultation, W. H. O. E. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet* **363**, 157–163 (2004).
30. Consortium G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

31. Young, K. L. *et al.* Interaction of smoking and obesity susceptibility loci on adolescent BMI: The National Longitudinal Study of Adolescent to Adult Health. *BMC Genet.* **16**, 131 (2015).
32. Demerath, E. W. *et al.* Interaction of FTO and physical activity level on adiposity in African–American and European-American adults The ARIC study. *Obesity (Silver Spring)* **19**, 1866–1872 (2011).
33. Graff, M. *et al.* Correction: Genome-wide physical activity interactions in adiposity—A meta-analysis of 200,452 adults. *PLoS Genet.* **13**, e1006972 (2017).
34. Levy, J. *et al.* EFNB2 haploinsufficiency causes a syndromic neurodevelopmental disorder. *Clin. Genet.* **93**, 1141–1147 (2018).
35. Wang, Y. *et al.* Reduced blood pressure after smooth muscle EFNB2 deletion and the potential association of EFNB2 mutation with human hypertension risk. *Eur. J. Hum. Genet.* **24**, 1817–1825 (2016).
36. Kim, S. K. *et al.* Two genetic loci associated with ankle injury. *PLoS ONE* **12**, e0185355 (2017).
37. Wolthusen, R. P. *et al.* Genetic underpinnings of left superior temporal gyrus thickness in patients with schizophrenia. *World J. Biol. Psychiatry* **16**, 430–440 (2015).
38. Guan, M. *et al.* Genome-wide association study identifies novel loci for type 2 diabetes-attributed end-stage kidney disease in African Americans. *Hum. Genom.* **13**, 21 (2019).
39. Nakayama, A. *et al.* Ephrin-B2 controls PDGFRbeta internalization and signaling. *Genes Dev.* **27**, 2576–2589 (2013).
40. Gervais, M., Labouebe, G., Picard, A., Thorens, B. & Croizier, S. EphrinB1 modulates glutamatergic inputs into POMC-expressing progenitors and controls glucose homeostasis. *PLoS Biol.* **18**, e3000680 (2020).
41. Justice, A. E. *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat. Commun.* **8**, 14977 (2017).
42. Velez Edwards, D. R. *et al.* Gene-environment interactions and obesity traits among postmenopausal African-American and Hispanic women in the Women's Health Initiative SHARe Study. *Hum. Genet.* **132**, 323–336 (2013).
43. Boardman, J. D. *et al.* Is the geneenvironment interaction paradigm relevant to genome-wide studies? The case of education and body mass index. *Demography* **51**, 119–139 (2014).
44. Studies, N.-N. *et al.* Replicating genotype–phenotype associations. *Nature* **447**, 655–660 (2007).

Acknowledgements

This study was conducted with bioresources from the National Biobank of Korea, the Center for Disease Control and Prevention, Republic of Korea (KBN-2019-021). This research was supported by the National Research Foundation (NRF) funded by the Korean Government (MSIT) (2019M3E5D3073365 (the Bio & Medical Technology Development Program) and NRF-2019R1A2C1083980).

Author contributions

J.E.L. and B.O. drafted the research protocol. H.-U.J. performed the statistical analysis. T.-W.H., W.J.L., J.E.L. and B.O. interpreted the data and prepared Table. J.K., M. k.K. and S.W. analyzed and interpreted the result data. T.P. designed Figure. J.-O.K. prepared Supplementary Table and Supplementary Figure. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83684-1.

Correspondence and requests for materials should be addressed to J.E.L. or B.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.