Joint Identification of Multiple Genetic Variants of Obesity in a Korean Genome-wide Association Study

Sohee Oh¹, Seoae Cho² and Taesung Park¹*

¹Department of Statistics, Seoul National University, Seoul 151-747, Korea, ²Department of Statistics, University of California, Berkeley, USA

Abstract

In recent years, genome-wide association (GWA) studies have successfully led to many discoveries of genetic variants affecting common complex traits, including height, blood pressure, and diabetes. Although GWA studies have made much progress in finding single nucleotide polymorphisms (SNPs) associated with many complex traits, such SNPs have been shown to explain only a very small proportion of the underlying genetic variance of complex traits. This is partly due to the fact that most current GWA studies have relied on single-marker approaches that identify single genetic factors individually and have limitations in considering the joint effects of multiple genetic factors on complex traits. Joint identification of multiple genetic factors would be more powerful and provide a better prediction of complex traits, since it utilizes combined information across variants. Recently, a new statistical method for joint identification of genetic variants for common complex traits via the elastic-net regularization method was proposed. In this study, we applied this joint identification approach to a large-scale GWA dataset (i.e., 8842 samples and 327,872 SNPs) in order to identify genetic variants of obesity for the Korean population. In addition, in order to test for the biological significance of the jointly identified SNPs, gene ontology and pathway enrichment analyses were further conducted.

Keywords: genome-wide association study, elastic-net regularization method, obesity

Introduction

Advances in genotyping technologies have allowed us to move from candidate gene approaches to large-scale genome-wide association (GWA) approaches (Hirschhorn and Daly, 2005; Wang et al., 2005) for identifying genetic variants for common complex diseases. In recent years, GWA studies have led to many discoveries of genetic variants affecting common complex traits, including height, blood pressure, and diabetes (Li et al., 2008; Saxena et al., 2007; Sladek et al., 2007; Thorleifsson et al., 2009; Wallace et al., 2008; Weedon et al., 2008).

Obesity, among many disease-related phenotypes, is becoming more important, because obesity is one of the risk factors for many chronic diseases, such as type 2 diabetes, cardiovascular disease, hypertension, and some types of cancer (Bell et al., 2005; Hfker and Wijmenga, 2009; Hill and Peters, 1998; Ichihara et al., 2008). With the increasing adoption of Western lifestyles, the prevalence of obesity is rapidly catching up in the developing world (Must et al., 1999). As a common complex phenotype, obesity is the result of genes, environmental factors, and interactions between the genes and environmental factors (Farooqi and O’Rahilly, 2007; Feitosa et al., 2002; Scuteri et al., 2007).

Much research has identified genetic factors associated with obesity-related phenotypes using single nucleotide polymorphisms (SNPs). Body mass index (BMI), defined as weight (kg)/height (m²), is the most commonly used measurement of obesity status. The World Health Organization (WHO) recommends the following cutoff points for BMI to classify weight status in adults 20 years of age or older: <18.5 kg/m² (underweight), 18.5-24.9 kg/m² (normal weight), 25.0-29.9 kg/m² (over weight), 30.0-39.9 kg/m² (obese), and ≥40 kg/m² (extremely obese). Scuteri et al., conducted association studies with BMI, hip circumstance, and weight, They reported that common genetic variants in the FTO gene are associated with substantial changes in three obesity-related phenotypes (Scuteri et al., 2007). Feitosa et al., used BMI and performed a linkage study from the National Heart, Lung, and Blood Institute Heart Study. They found significant signals near the leptin gene (chromosome 7q32.3) and chromosome 13q14 (Feitosa et al., 2002). Liu et al., identified the CTNNBL1 gene for obesity from US Caucasian and French replication case-control samples. They also confirmed the previously identified associations between obesity and INSIG2 and PFKP (Liu et al., 2008).

Loos et al., performed association studies in chromosomal region 18q21 and reported that common variants near the MC4R gene influence fat mass, weight, and obesity risk (Loos et al., 2008), Thorleifsson et al., con-
ducted genome-wide association studies for BMI and weight. Using 11 significant genome-wide loci, they also conducted association tests with obesity (obese for BMI ≥ 30, normal 18.5 < BMI < 25) and type 2 diabetes and reported significant results in chromosomal region 12q13 with both obesity and type 2 diabetes (Thorleifsson et al., 2009).

Willer et al. performed genome-wide studies with quantitative BMI phenotype and binary traits (overweight (BMI ≥ 25) or obese (BMI ≥ 30)). They confirmed a strong association of FTO and MC4R with BMI and identified candidate causal loci: the TMEM18, KCTD15, GNPDA2, SH2B1, MTCH2, and NEGR1 genes. They also analyzed the impact of gene expression and copy number polymorphisms (CNPs) on BMI (Willer et al., 2008).

However, most genome-wide association studies have relied on single-marker approaches that test individual associations between single SNPs and the trait. However, single-SNP approaches can be optimal for Mendelian traits and may not be appropriate for investigating a common complex polygenic trait because of the following reasons. First, it would be hard to test and predict the accumulated and/or joint effects of multiple genetic factors on the trait. Second, due to the large number of individual tests in GWA analyses, a certain multiple testing correction procedure is required to reduce the testing correction procedures. Only a small number of markers pass the genome-wide significance level, and they do not fully explain the etiology of common complex traits.

When multiple genetic factors exist for a common complex trait, the joint identification of such factors would be more powerful and provide a better prediction of the trait, since it utilizes combined information across multiple genetic variants. For example, traditional approaches, such as multiple linear/logistic regression methods with a variable selection procedure, can use joint identification. However, there are some difficulties in applying multiple regression methods to GWA analysis. Since the number of predictor variables is relatively larger than the sample sizes, multiple regression methods are ill-defined and induce a computational burden for variable selection procedures. In addition, multicollinearity problems may exist due to linkage disequilibrium (LD) among SNPs. To improve power in identifying multiple genetic variants associated with obesity, more powerful statistical methods need to be developed for genome-wide association studies.

To capture multiple putative genetic variants for common complex diseases, Cho et al. (2010) proposed new statistical methods using an elastic-net (EN) regularization method along with some consistency measures based on bootstrap sampling. The EN regularization method was originally proposed for model fitting and variable selection in ill-defined multiple regressions (Tibshirani, 1996; Zou and Hastie, 2005) and has been applied to GWA analyses very recently (Shi et al., 2008; Wu et al., 2009).

In this paper, we applied the EN regularization method to a large-scale GWA dataset (i.e., 8842 samples and 327,872 SNPs) in order to identify genetic variants of obesity in the Korean population. In addition, in order to test for the biological significance of the jointly identified SNPs, gene ontology and pathway enrichment analyses were further conducted.

Methods

KARE data

The Korea Association Resource (KARE) project, established as part of the Korean Genome Epidemiology Study (KoGES) in 2001, was started to undertake a large-scale genome-wide association analysis in Korean population-based cohorts. These cohorts consisted of 10,038 participants in the urban Ansan (n=5020) and rural Ansung (n=5018) communities. In this analysis, we focused on BMI, one of more than 260 lifestyle-related traits.

DNA samples were isolated from the peripheral blood of all participants and were genotyped with the Affymetrix Genome-Wide Human SNP array 5.0. Genotypes were called using Bayesian Robust Linear Modeling and the Mahalanobis Distance (BRLMM) algorithm (Rabbee and Speed, 2006). We performed quality control processes for the samples and genotypes as previously described in Cho et al. (2009). In order to increase coverage of the common variants and capture additional association signals, we performed SNP imputation with PLINK using the JPT/CHB component of HapMap as the reference. After sample and genotype quality control, 327,872 SNPs for 8842 individuals were available in the KARE data. We applied our proposed regularization method to Korean genome-wide data.

Statistical methods

In order to identify multiple causal SNPs among a huge number of SNPs, we conducted a multi-stage procedure, proposed by Cho et al. (2010). In the first stage, we selected SNPs having an association with the trait via single-SNP association tests. In the second stage, the multiple-SNP associations were searched based on penalized regression with the EN regularization method.
Stage 1: Prescreening SNPs for dimensionality reduction: Using linear regression with adjustments for sex, age, and area (Ansan and Ansung), single SNP association was performed for each SNP. In this analysis, an additive genetic mode based on a minor allele for a variant was considered. From the results, a subset of SNPs that had a strong association with the trait were selected for dimensionality reduction. In our analysis, we chose 1000 SNPs.

Stage 2: Joint identification of putative causal SNPs via EN regularization method: Under the multiple linear regression frameworks, putative trait-related SNPs were identified via the EN variable selection method proposed by Zou and Hastie (2010):

$y_j = \gamma_0 + \gamma_1 \text{Sex}_j + \gamma_2 \text{Age}_j + \gamma_3 \text{Area}_j + \sum_{i=1}^{p} \beta_i \text{SNP}_{ij} + \epsilon_j$

where $y_j$, Sex$_j$, Age$_j$, Area$_j$, SNP$_{ij}$, and $\epsilon_j$ represent the trait value, sex, age, area, the number of minor alleles for the ith SNP marker, and the measurement error of the jth individual, respectively; $i=1, \ldots, p$, and $p$ is the number of SNPs under consideration; and $j=1, \ldots, n$, and $n$ is the number of individuals. $\gamma_0$, $\gamma_1$, $\gamma_2$, and $\gamma_3$ denote the overall mean, the effect sizes of the corresponding covariates, and effect sizes of the SNPs, respectively. EN regularization is particularly useful where the number of highly correlated predictor variables is much larger than the sample size ($n < < p$), EN regularization solves the following problem:

$$\min_{\beta_{i}, \lambda, \alpha} \left[ \frac{1}{2n} \sum_{j=1}^{n} (y_j - \gamma_0 - \gamma_1 \text{Sex}_j - \gamma_2 \text{Age}_j - \gamma_3 \text{Area}_j - \sum_{i=1}^{p} \beta_i \text{SNP}_{ij})^2 + \lambda \left( \frac{1}{2} \sum_{i=1}^{p} |\beta_i|^{1-\delta} + \alpha |\beta_i|^{\delta} \right) \right]$$

with another tuning parameter, $\alpha$. The EN penalty creates a useful compromise between the ridge ($\alpha=0$) (Le Cessi and Van Houwelingen, 1992) and LASSO ($\alpha=1$) (Tibshirani, 1996) penalties. EN with $\alpha=1-\delta$ for some small $\delta>0$ performs in a manner similar to LASSO but is robust for extreme correlations among predictor variables. Since EN regularization performs both shrinkage and automatic variable selection simultaneously, parsimonious model selection is possible. The choice of the tuning parameters is critical for selecting important variables with accurate estimation. The tuning parameter $\lambda$ controls the strength of the penalty, which shrinks each coefficient toward the origin and enforces a sparse solution. Cross-validation (e.g., 10-fold) is generally employed to find the best values of $\lambda$ and $\alpha$, which minimize the mean-squared prediction error.

To investigate the biological significance of the genetic variants jointly identified from the EN regularization method, we mapped the identified SNPs to an exon/intron or within the 5-kb upstream/0.5-kb downstream regions of known genes and performed gene ontology (GO) and pathway enrichment analyses.

Results

The multi-stage EN regularization method was applied to the KARE GWAS data set. In the first stage, the individual association between each of the 327,872 SNPs and BMI was tested via linear regression with adjustments for sex, age, and area. Fig. 1 shows the results of the association tests. Seven SNPs in four loci reached the genome-wide significant threshold ($p$-value $<1 \times 10^{-5}$). One of the SNPs, rs9939609 ($p=1.43 \times 10^{-5}$), corresponded to the widely reported signal in the FTO
Table 1. Nine genes detected as underlying human obesity variations in previous GWA studies were identified in our GWA analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>rs ID</th>
<th>Class</th>
<th>Locus</th>
<th>Minor allele</th>
<th>MAF</th>
<th>Effect size</th>
<th>Previous GWA studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEC16B</td>
<td>Rs623479</td>
<td>3’ UTR</td>
<td>1q25.2b</td>
<td>C</td>
<td>0.15</td>
<td>−0.0335</td>
<td>Thorleifsson (rs10913469-C)</td>
</tr>
<tr>
<td>RARB</td>
<td>Rs17016781</td>
<td>Intron</td>
<td>3p24.2a</td>
<td>G</td>
<td>0.40</td>
<td>0.0171</td>
<td>Cotsapas (rs1435703-T)</td>
</tr>
<tr>
<td>ERC2</td>
<td>Rs9875049</td>
<td>Intron</td>
<td>3p14.3c</td>
<td>C</td>
<td>0.09</td>
<td>0.0276</td>
<td>Kiel (rs2054989)</td>
</tr>
<tr>
<td>ADARB2</td>
<td>Rs11250504</td>
<td>Intron</td>
<td>10q15.3c</td>
<td>T</td>
<td>0.13</td>
<td>−0.0014</td>
<td>Lowe (rs6560749-T)</td>
</tr>
<tr>
<td>PANK1</td>
<td>Rs10881614</td>
<td>Intron</td>
<td>10q23.31c</td>
<td>C</td>
<td>0.38</td>
<td>0.0607</td>
<td>Sabatti (rs11185790-A)</td>
</tr>
<tr>
<td>BDNF</td>
<td>Rs11030104</td>
<td>Intron</td>
<td>11p14.1d</td>
<td>G</td>
<td>0.44</td>
<td>−0.0079</td>
<td>Thoreleifsson (rs8265-G)</td>
</tr>
<tr>
<td>FTO</td>
<td>Rs9939609</td>
<td>Intron</td>
<td>16q12.2b</td>
<td>A</td>
<td>0.13</td>
<td>0.1873</td>
<td>Scherag (rs1558902-A)</td>
</tr>
<tr>
<td>LOC729076</td>
<td>Rs17178527</td>
<td>Unknown</td>
<td>6q24.1</td>
<td>A</td>
<td>0.25</td>
<td>−0.1186</td>
<td>Cho</td>
</tr>
<tr>
<td>Unknown</td>
<td>Rs17089410</td>
<td>Unknown</td>
<td>13q21.33c</td>
<td>C</td>
<td>0.13</td>
<td>0.0813</td>
<td>Cho</td>
</tr>
</tbody>
</table>

Fig. 2. Hierarchical clustering of over-represented GO terms enriched in the 189 genes jointly associated with Korean adult obesity.

gene (Cho et al., 2009; Frayling et al., 2007; Willer et al., 2008). Among the 327,872 SNPs, the top 1000 SNPs showing the strongest associations were selected for the next stage. In the next stage, a total of 511 SNPs were jointly identified as putative BMI-related genetic variants by the EN regularization method.

Among them, 254 SNPs were mapped to 189 genes. Of these genes, 9 genes had been reported to affect human obesity variation, including BMI, weight, and hip circumference, in a previous GWAS (Cho et al., 2009; Cotsapas et al., 2009; Frayling et al., 2007; Heard-Costa et al., 2009; Hinney et al., 2007; Kiel et al., 2007; Loos et al., 2008; Lowe et al., 2009; Meyre et al., 2009; Sabatti et al., 2009; Scherag et al., 2010; Scuteri et al., 2007; Thorleifsson et al., 2008; Willer et al., 2008) (Table 1). Our identification also included three SNPs that had been reported to be associated with BMI and waist-hip ratio (WHR) in the Korean population (Cho et al., 2009).

Finally, a GO enrichment analysis of 189 genes was performed. All GO terms that were related to the identified genes were searched via the DAVID functional annotation tool. Then, the EASE software tool (Hosack et al., 2003) was used to calculate the over-representation statistic and its EASE score for each of the searched GO terms. Over-represented GO terms were significantly detected with a p-value < 0.05 and were presented via hierarchical clustering with a dissimilarity matrix, defined similarly in Kosiol et al. (Kosiol et al., 2008). The GO en-
richment analysis of 189 genes showed that several biological processes were significantly enriched; they included cell migration, metabolic process, transport, and protein processes (Fig. 2).

In addition, based on the KEGG pathway database, two pathways were found to be enriched in these 189 genes: the endocytosis and axon guidance pathways (Table 2). First, the endocytosis pathway regulates many processes, including nutrient uptake, cell adhesion and migration, receptor signaling, pathogen entry, neurotransmission, growth, and differentiation (Miaczynska and Stenmark, 2007). This pathway is related with the receptor-mediated endocytosis of ligands, such as low-density lipoprotein and growth factors. In addition, caveolae, one of four subcategories of this pathway are enriched in membrane proteins such as a cholesterol-binding protein and glycolipids. Second, axon guidance, which represents key steps of formation of the neural network, contains five genes: ABLIM1, PLXNC1, PLXNA2, SRGAP3, and SRGAP1. Das and Rao (2007) reported that genes involved in the axon guidance pathway were upregulated in subjects with type 2 diabetes with a family history of diabetes compared to those who are diabetic (Das and Rao, 2007).

### Discussion

Most genome-wide association studies have reported genetic variants that are associated with common complex traits via single-marker approaches. From this single SNP approach, many genetic variants have been successfully identified. However, it would be hard to identify polygenic factors through a single SNP approach. Additionally, the single SNP approach does not take the LD structure into account. Most multiple comparison procedures, such as Bonferroni correction and false discovery rate (FDR) (Abdi, 2007; Benjamini and Hochberg, 1995), employed in GWA studies assume independence among the tests. However, this assumption does not hold in the presence of an LD. Thus, it is hard to identify variants with moderate effects, even with multiplicity corrections.

In this paper, we conducted association studies for BMI from Korean GWA data via the EN regularization method proposed by Cho et al. (2010). The EN regularization method facilitated the joint detection of multiple genetic factors and hence seemed well suited for the GWA analysis of common complex traits. The EN regularization method has the following advantages. First, it tends to provide more reliable identification of multiple SNPs than ordinary multiple regression, because ordinary regression is very sensitive to multicollinearity due to LD among the SNPs. In the presence of multicollinearity, the coefficients of multiple regression become unstable with large variances. However, by combining ridge and LASSO penalties, EN can select variables more effectively and obtain more stable estimates for high dimensional problems.

In single-marker analysis, seven SNPs in four loci passed the significance level ($p$-value $< 1 \times 10^{-5}$), and two SNPs were reported in Cho et al. (2009). However, through the EN regularization method, we identified 189 genes associated with BMI, of which six genes were reported in other studies (Cho et al., 2009; Cotsapas et al., 2009; Kiel et al., 2007; Lowe et al., 2009; Sabatti et al., 2009; Thorleifsson et al., 2008). From these jointly identified genes, gene ontology and pathway enrichment analyses were further conducted for more meaningful biological interpretation.

Although we conducted an association study with quantitative traits, the EN method can be applied to binary traits, such as hypertension, or data, such as the number of epileptic seizures. Furthermore, the EN method can be applied to identify gene-gene and/or gene-environment interactions.

### Acknowledgments

This work was supported by the Consortium for Large Scale Genome-Wide Association Study, the National Research Foundation (KRF-2008-313-C00086), and the Brain Korea 21 Project of the Ministry of Education, The KARE data analyzed in this study were obtained from the Korean Genome Analysis Project (4845-301), which was funded by a grant from the Korea National Institute of Health (Korea Center for Disease Control, Ministry for Health, Welfare and Family Affairs), Republic of Korea. The authors thank Hyung-Lae Kim (Center for Genome Science, National Institute of Health, South Korea), Nam H. Cho (Department of Preventive Medicine, Ajou University, South Korea), and Chol Shin (Department of Internal Medicine, Korea University Ansan Hospital, South Korea) for their great efforts in generating and providing these valuable data.

### Table 2. Enrichment KEGG pathway

<table>
<thead>
<tr>
<th>KEGG_pathway</th>
<th>p-value</th>
<th>Genes</th>
<th>FE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endocytosis</td>
<td>0.048</td>
<td>FLT1, ERBB4, ARRB1,</td>
<td>2.981</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SH3GLB1, NEDD4L, IOSEC1</td>
<td></td>
</tr>
<tr>
<td>Axon guidance</td>
<td>0.05</td>
<td>ABLIM1, PLXNC1, PLXNA2,</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRGAP3, SRGAP1</td>
<td></td>
</tr>
</tbody>
</table>
References


Loos, R.J.F., Lindgren, C.M., Li, S., Wheeler, E., Zhao, J.H., Prokopenko, I., Inouye, M., Freathy, R.M., Attwood, A.P., Beckmann, J.S., Berntd, S.I., Prostate, Lung, Colorectal,


Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 447, 661-678.

