Annals of human genetics

doi: 10.1111/j.1469-1809.2010.00597.x

Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis

Seoae Cho[1]§, Kyunga Kim[2]§, Young Jin Kim[1,3], Jong-Keuk Lee[4], Yoon Shin Cho[3], Jong-Young Lee[3], Bok-Ghee Han[3], Heebal Kim[5], Jurg Ott[6] and Taesung Park[1,7]*

[1]*Interdisciplinary Program in Bioinformatics, Seoul National University, South Korea, 151-747*
[2]*Department of Statistics, Sookmyung Women's University, South Korea, 140-742*
[3]*Center for Genome Science, National Institute of Health, South Korea, 122-701*
[4]*Asan Institute for Life Sciences, University of Ulsan College of Medicine, South Korea, 138-736*
[5]*Department of Agricultural Biotechnology, Seoul National University, South Korea, 151-921*
[6]*Beijing Institute of Genomics, No. 7 Bei Tu Cheng West Road, Beijing 100029, China*
[7]*Department of Statistics, Seoul National University, South Korea, 151-747*

Summary

Unraveling the genetic background of common complex traits is a major goal in modern genetics. In recent years, genome-wide association (GWA) studies have been conducted with large-scale data sets of genetic variants. Most of those studies have relied on single-marker approaches that identify single genetic factors individually and can be limited in considering fully the joint effects of multiple genetic factors on complex traits. Joint identification of multiple genetic factors would be more powerful and would provide better prediction on complex traits since it utilizes combined information across variants. Here we propose a multi-stage approach for GWA analysis: (1) prescreening, (2) joint identification of putative SNPs based on elastic-net variable selection, and (3) empirical replication using bootstrap samples. Our approach enables an efficient joint search for genetic associations in GWA analysis. The suggested empirical replication method can be beneficial in GWA studies because one can avoid a costly, independent replication study while eliminating false-positive associations and focusing on a smaller number of replicable variants. We applied the proposed approach to a GWA analysis, and jointly identified 129 genetic variants having an association with adult height in a Korean population.

Keywords: Genome-wide association, multiple regression, elastic-net variable selection, empirical replication, adult height

Introduction

In genome-wide association (GWA) studies, hundreds of thousands of genetic variants (e.g., single-nucleotide polymorphisms or SNPs) selected across the entire genome are usually measured on a high-density genotyping platform and evaluated to identify genomic regions where the true trait-related genes may lie. In recent years, GWA studies have led to many discoveries of genetic variants affecting common complex traits, including height, blood pressure and diabetes (Saxena et al., 2007; Sladek et al., 2007; Li et al., 2008; Wallace et al., 2008; Weedon et al., 2008; Thorleifsson et al., 2009). Although GWA studies have made progress in finding SNPs associated with many complex traits (Hindorff et al., 2009), such SNPs have been shown to explain only a very small proportion of the underlying genetic variance of most complex traits (Goldstein, 2009; Kraft & Hunter, 2009). This conclusion was supported by two pieces of evidence. First, the relative risks that are found to be conferred by common risk genotypes account for only a small proportion of the sibling recurrence risk. Second, in several GWA studies of human diseases, some loci have been detected more than once, but each study has identified multiple loci that were not

§S. Cho and K. Kim made equal contributions to this work.
*Corresponding author: T. Park, Department of Statistics, Seoul National University, 56-1 Shillim-Dong, Kwanak-Gu, Seoul, South Korea, 151-747. Tel: +82 (0)2-880-8924; Fax: +82 (0)2-883-6144; E-mail: tspark@stats.snu.ac.kr

identified in other studies. This would suggest that many more loci remain to be discovered.

Most current GWA studies have relied on single-marker approaches that test individual associations between each SNP and the trait because individual identification is simple and readily applicable. While single-marker approaches can be optimal when a single genetic factor is responsible for a trait, they may not be appropriate for investigating a complex polygenic trait due to the following reasons. First, it can be hard to examine 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 analysis, a certain multiple testing correction procedure is required to reduce the increased probability of finding false positives. Most multiple testing correction procedures currently employed in GWA studies, such as Bonferroni correction and false discovery rate (Benjamini & Hochberg, 1995), assume independence among the tests. These procedures might be too conservative to identify causal SNPs with small to moderate effects because the data in GWA studies are likely to contain linked SNPs; thus, the individual association tests for those linked SNPs cannot be independent. A permutation-based multiple testing correction procedure can be used to account for such dependencies among SNPs (Westfall & Young, 1993), but it significantly increases the computational burden that GWA studies already bear in practice.

When multiple genetic factors exist for a complex trait, the joint identification of such factors would be more powerful and provide better prediction of the trait since it utilizes combined information across multiple genetic variants. Traditional approaches for such joint identification would be multiple linear/logistic regression methods with a variable selection procedure (e.g., stepwise selection). However, there are certain challenges in applying multiple regression methods to GWA analysis. First, the large number of predictor variables (i.e., SNPs) and relatively small sample sizes would render multiple regression methods ill-defined and induce a heavy computational burden in selection of variables. In addition, it can be often assumed that only a relatively small number of SNPs contribute to the trait. With such sparsity, a dimension reduction can be more beneficial both for computational efficiency and accuracy of variable selection (Fan & Lv, 2008). Second, multicollinearity is likely to exist because of the linkage among SNPs. Multiple linear regressions are very sensitive to multicollinearity, with which model parameters become unstable with large variances.

In order to address these challenges, we propose here a multi-stage strategy that exploits the elastic-net regression method to identify a joint effect of multiple genetic variants. Many regularization methods, including the least absolute shrinkage and selection operator (LASSO), ridge and elastic-net, have been proposed for model fitting and variable selection in ill-defined multiple regressions (Tibshirani, 1996; Le Cessie & Van Houwelingen, 1992; Zou & Hastie, 2005; Zou, 2006; Wu & Lange, 2008), but only a few have been applied to GWA analysis (Wu et al., 2009; Shi et al., 2008). While ridge regularization induces shrinkage of predictors, and thus makes parameter estimation more stable (Le Cessie & Van Houwelingen, 1992), LASSO regularization leads many regression coefficients to become exactly zero and hence facilitates automatic variable selection in which only one predictor is selected among the correlated predictors (Tibshirani, 1996). Elastic-net regularization uses ridge and LASSO penalties simultaneously to take advantages of both regularization methods (Zou & Hastie, 2005). Thus, it provides shrinkage and automatic variable selection and can deal more efficiently with the severe multicollinearity that often exists in GWA analysis. As shown in general context (Tibshirani, 1996; Le Cessie & Van Houwelingen, 1992; Zou & Hastie, 2005; Zou, 2006; Wu & Lange, 2008), the elastic-net regularization would perform better than LASSO in GWA analysis, in which multicollinearity persistently exists due to linkage disequilibrium among nearby SNPs, and this was supported by our preliminary investigation via simulations (data not shown).

For additional improvement in the selection of variables, we employ pre-screening in which SNPs having weak correlations with the trait are eliminated beforehand. This correlation screening method was suggested by Fan and Lv (Fan & Lv, 2008) to reduce an ultrahigh dimensionality (e.g., a large number of SNPs and relatively small sample size in GWA studies) while keeping all the important SNPs, and was shown to improve variable selection with regard to accuracy and speed.

The proposed approach consists of three main stages: (1) prescreening for dimensionality reduction, (2) joint identification of putative SNPs via elastic-net variable selection, and (3) empirical replication study of identified SNPs based on bootstrap samples. At the pre-screening stage, we employ single-SNP association tests to filter out SNPs weakly correlated with the trait. As a result, a set of SNPs showing the highest correlations with the trait are selected for the next stage. At the next stage, putative causal SNPs are jointly identified by penalized multiple regressions with elastic-net regularization. For empirical replication, we propose a new measure called bootstrap selection stability (BSS), which suggests how consistently each identified SNP is replicated via elastic-net regularization in the bootstrap samples. This stage would provide more insights into underlying causal relationships by focusing on a smaller number of important SNPs and would generate more reliable estimates by excluding noisy SNPs.

The proposed multi-stage approach is then applied to a large-scale GWA dataset (i.e., 8,842 samples and 327,872 SNPs) of adult human height in a Korean population. In order

to test for the biological significance of the jointly identified SNPs, gene ontology and pathway enrichment analyses are further conducted.

Materials and Methods

KARE Data

The Korea Association Resource (KARE) project was initiated in 2007 to undertake large-scale GWA analyses. Participants in this project were recruited from two community-based cohorts (i.e., the rural Ansung and urban Ansan cohorts) in the Gyeonggi Province of South Korea. The Ansung and Ansan cohorts consist of 5,018 and 5,020 participants, respectively, ranging in age from 40 to 69 years. More than 260 traits have been extensively examined through epidemiological surveys, physical examinations and laboratory tests. Here, we focus on the GWA analysis of height because human height is a highly heritable polygenic quantitative trait with biomedical importance (Li et al., 2004).

Genomic DNA samples were isolated from peripheral blood drawn from the participants and were genotyped on the Affymetrix Genome-Wide Human SNP array 5.0 containing 500,568 SNPs. Prior to the analysis, we performed genotype calling and quality control processes as previously described in Cho et al. (Cho et al., 2009), and imputed missing genotypes using PLINK software and the JPT/CHB reference panel in HAPMAP. For genotype calling, the Bayesian Robust Linear Modeling using Mahalanobis Distance (BRLMM) Genotyping Algorithm was used. From sample and SNP quality controls, finally a total of 8,842 individuals and 327,872 SNPs were included in the analyses. As noted in Cho et al. (Cho et al., 2009), there was no evidence for possible population stratification in the KARE data.

Statistical Analysis

For the joint identification of putative causal SNPs among a huge number of SNPs, we developed a multi-stage procedure with three main stages as follows. At Stage 1, we eliminate SNPs having a weak correlation with the trait via single-SNP association tests. At Stage 2, the multiple-SNP associations are searched based on penalized multiple regression with elastic-net regularization. At Stage 3, the jointly identified SNPs are evaluated via BSS which we have proposed for the empirical assessment of how consistently a SNP is selected via elastic-net regularization from the bootstrap samples.

Stage 1: Prescreening SNPs for dimensionality reduction
For each SNP, the single-SNP association with height is examined using linear regression with adjustment for gender, age and recruitment area (namely, Ansung and Ansan). An additive model is assumed for the genetic mode. Then, a subset of SNPs showing the strongest association with the trait is chosen for dimensionality reduction. Note that the subset size (i.e., the number of SNPs remaining for Stage 2) can be determined according to computational concerns in practice. In our GWA analysis of KARE data, we chose 1000 such SNPs.

Stage 2: Joint identification of putative causal SNPs via penalized regression with elastic-net variable selection
On the basis of the following multiple linear regression, putative trait-related SNPs are simultaneously identified via elastic-net variable selection:

$$\gamma_j = \beta_0 + \sum_{i=1}^{p} \beta_i SNP_{ij} + \gamma_1 GENDER_j + \gamma_2 AGE_j + \gamma_3 AREA_j + \varepsilon_j \quad (1)$$

where γ_j, SNP_{ij}, $GENDER_j$, AGE_j, $AREA_j$ and ε_j represent the trait value, the number of minor alleles for the ith SNP marker, gender, age, recruitment area (namely, Ansung and Ansan) and measurement error of the jth individual, respectively; $i = 1, 2, \ldots, p$ and p is the number of SNPs under consideration; $j = 1, 2, \ldots, n$ and n is the number of individuals. The β_0, β_is and γs denote overall mean, effect sizes of SNPs and effect sizes of the corresponding covariates, respectively. Note that elastic-net regularization is particularly useful where the number of highly correlated predictor variables (p) is much larger than the sample size (n). Elastic-net regularization solves the following problem:

$$\min_{(\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Re^{p+4}} \left[\frac{1}{2n} \sum_{j=1}^{n} \left(\gamma_j - \beta_0 - \mathbf{SNP}_j^T \boldsymbol{\beta} - \mathbf{COV}_j^T \boldsymbol{\gamma} \right)^2 + \lambda P_\alpha(\boldsymbol{\beta}) \right] \quad (2)$$

where $\mathbf{SNP}_j = (SNP_{1j}, SNP_{2j}, \ldots, SNP_{pj})^{\mathrm{T}}$ and $\mathbf{COV}_j = (GENDER_j, AGE_j, AREA_j)^{\mathrm{T}}$ for the jth individual, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^{\mathrm{T}}$. Here, λ is a tuning parameter and the elastic-net penalty is defined as follows:

$$P_\alpha(\boldsymbol{\beta}) = (1-\alpha)\frac{1}{2} \|\boldsymbol{\beta}\|_{l_2}^2 + \alpha \|\boldsymbol{\beta}\|_{l_1} = \sum_{i=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_i^2 + \alpha|\beta_i| \right] \quad (3)$$

with another tuning parameter α. The elastic-net penalty creates a useful compromise between ridge ($\alpha = 0$) (Tibshirani, 1996) and LASSO ($\alpha = 1$) penalties (Le Cessie & Van Houwelingen, 1992). The elastic-net with $\alpha = 1 - \delta$ for some small $\delta > 0$ performs in a manner similar to LASSO, but is robust to extreme correlations among predictor variables. Since elastic-net regularization performs both shrinkage and automatic variable selection simultaneously, parsimonious model selection is possible. The choice of the tuning parameters is critical to select important variables with accurate estimation. The tuning parameter λ controls the strength of the penalty, which shrinks each coefficient toward the origin and enforces sparse solutions. Cross validation

(e.g., 10-fold) is generally employed to find the best values of λ and α, which minimize mean-squared prediction error.

Stage 3: Empirical replication study of identified SNPs based on bootstrap samples

For validation of the jointly identified SNPs via elastic-net variable selection, we propose an empirical replication study using a large number (e.g., $B = 1000$) of *in silico* samples. The bootstrap technique is employed to generate *in silico* samples, each of which has same population structure and contains the same number of individuals as does the original data. Bootstrapping is a resampling technique by which a bootstrap sample is constructed by random sampling with replacement from the original dataset. Usually, each bootstrap dataset has equal sample size to the original dataset. For each bootstrap dataset, phenotype-associated SNPs are simultaneously detected via elastic-net variable selection. Note that we use the fixed optimal value of λ, which is chosen at Stage 2, to reduce the computational burden. Then, bootstrap selection stability (BSS) is defined for ith SNP as follows:

$$BSS_i = \frac{1}{B}\sum_{b=1}^{B} I_i^b, \quad \text{where } I_i^b = \begin{cases} 1 & \text{if replicated in } b^{th} \text{ bootstrap sample} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

BSS indicates how consistently a SNP is replicated in B bootstrap datasets and would imply replication likeliness. The idea for BSS arose from the bootstrap analysis (Park & Hastie, 2008) that was conducted to measure the statistical relevance of selected variables by the forward stepwise procedure. Among the SNPs identified at Stage 2, we can further eliminate SNPs with low BSS and focus on SNPs with high BSS (e.g., SNPs with BSS $\geq$ 95%).

Biological Significance: Gene Ontology and Pathway Enrichment Analyses

In order to demonstrate the biological significance of the genetic variants jointly identified from the proposed multi-stage approach based on elastic-net variable selection, we map the identified SNPs to exon/intron or within the 5-kb upstream/0.5-kb downstream regions of known genes, and perform Gene Ontology (GO) and pathway enrichment analyses. We used GO/pathway analysis for exploratory purposes (i.e., summarizing SNP-level results and exploring possible height-relevance) rather than for rigorous analysis of height-related SNPs. First, all GO terms related to the identified genes are searched via the *DAVID* functional annotation tool. Then, the EASE software tool (Hosack et al., 2003) is used to calculate the over-representation statistic and its EASE score for each of the searched GO terms. Over-represented GO terms are significantly detected with p-value < 0.05, and presented via hierarchical clustering with a dissimilarity matrix defined similar to Kosiol et al. (Kosiol et al., 2008) to reflect the fact that genes associated with GO terms are not mutually exclusive. Similarly, all pathways related to the identified genes are surveyed via the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2004) and evaluated by the over-representation statistic and its EASE score. The "significantly regulated pathways" are identified if they contain at least two genes and have p-value < 0.1. In addition, the *DAVID* functional annotation tool is used to find functional clusters related to the identified genes.

Simulation Analysis: Cross-Validation for Selecting the Tuning Parameter

The 10-fold cross-validation method has been popularly used to choose tuning parameters in many regularization-based variable selection methods, including the elastic-net, and its validity and effectiveness has been proven via simulations in general context (Tibshirani, 1996; Zou & Hastie, 2005; Zou, 2006; Friedman et al., 2008; Wu et al., 2009). However, it is also known that cross-validation typically works well in problems where the true regression function is sparse and the signals are large. Since SNPs do not seem to have large signals in GWA studies, we conducted simulations to investigate the performance of cross-validation in a GWA context.

For each simulation dataset, the sample size was 200, and a total of 1000 SNPs were simulated with three different values (0.3, 0.5 and 0.8) of the correlation among SNPs in two stages which were adopted from Wu et al.'s work (Wu et al., 2009). First, a predictor vector $X_j = (x_{1j}, \ldots, x_{pj})$ for the jth individual was simulated from a p-dimensional multivariate normal distribution with zero mean and covariance matrix S whose (k,l)-th entry was $S_{kl} = 0.3, 0.5, 0.8$ if $k \neq l$, and $S_{kk} = 1$ otherwise. Then, assuming SNPs have equal allele frequencies, we set the genotype of the ith SNP for the jth individual equal to -1, 0 or 1 according to whether $x_{ij} < -c$, $-c \leq x_{ij} \leq c$, or $x_{ij} > c$. The cutoff c was the first quartile of a standard normal distribution.

Assuming that the first 20 SNPs are true causal, the trait value y_j was generated based on the following multiple regression setting:

$$y_j = \beta_0 + \sum_{i=1}^{20} \beta_i SNP_{ij} + \varepsilon_j \qquad \text{where } \varepsilon_j \sim N(0, \sigma^2) \tag{5}$$

In order to see how well the 10-fold cross-validation works when predictive power is not strong, we set all βs $= 1$ with various levels of signal-to-noise (5, 10, . . . , and 100), according to which the standard deviation σ was chosen. For each simulation setting, 100 datasets were generated. The proposed approach was performed while selecting the tuning parameter via the 10-fold cross validation for each simulated dataset.

Results

The proposed multi-stage analysis was applied to KARE data for a GWA study of adult height in Koreans. At the pre-screening stage, the individual association between each of

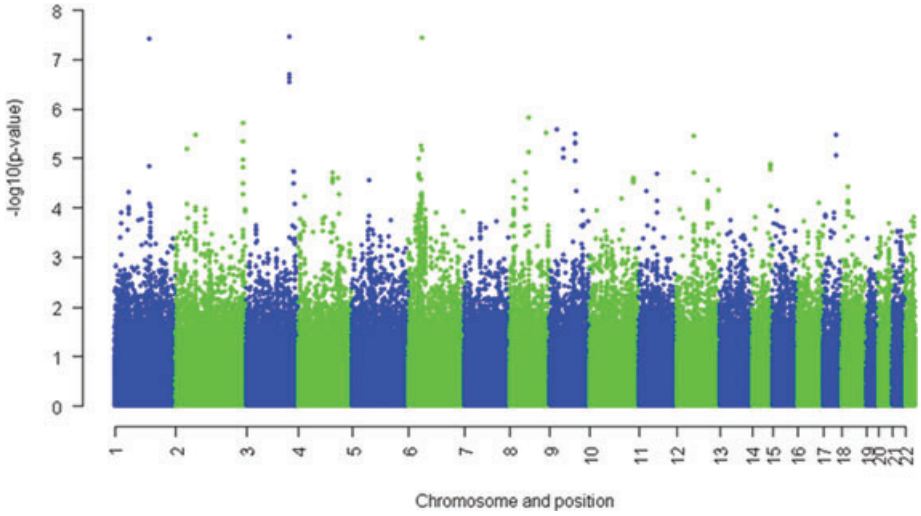

Figure 1. Results of single-marker association tests for 327,872 SNPs in the pre-screening stage. The y-axis represents p-values in minus logarithmic scale. The x-axis stands for genomic locations of SNPs across genome.

327,872 SNPs and height was evaluated via linear regression adjusted for gender, age and recruitment area (Fig. 1). We chose 1000 SNPs showing the strongest individual associations which could be handled by our computational resource.

At the next stage, a total of 516 SNPs were jointly identified as putative height-related genetic variants by penalized multiple regression with elastic-net regularization. Among them, 257 SNPs were mapped to 208 genes. Of these genes, 19 genes that had been detected to affect human height variation in previous GWA studies (Weedon et al., 2008; Cho et al., 2009; Gudbjartsson et al., 2008; Soranzo et al., 2009; Sovio et al., 2009; Lettre et al., 2008) were also identified in our GWA analysis (Table 1). Our identification included all eight SNPs that have been previously reported with height associations in the Korean population (Cho et al., 2009). Another observation was that two intron SNPs (rs17647719 and rs17523270) of cytochrome P450 19 A1 (*CYP19A1*) were associated with height in Koreans. The *CYP19* gene has been known to be associated with height variations in Caucasians from a family-based association study (Yang et al., 2006). We also observed a significant association with height in an intron SNP of the insulin-like growth factor 1 gene (*IGF1*) as well as three SNPs in the upstream region of the insulin-like growth factor binding protein 4 gene (*IGFBP4*). A polymorphic CA repeat in *IGF1* is known to be associated with gender-specific differences in body height (Rietveld et al., 2004) but others have suggested that the common variation in the growth hormone (GH)/IGF1 axis is not a major determinant of adult height (Lettre et al., 2007).

The GO enrichment analysis of the 208 genes showed that several biological processes were significantly enriched; these included lipid metabolism, transport, sensory perception, muscle system processes, development and morphogenesis, signal transduction and nucleotide metabolic process (Fig. 2). Among these processes, lipid metabolism, muscle system processes, and development and morphogenesis are likely to be directly involved in human height or body mass. Based on the KEGG pathway database, we found four pathways enriched in these 208 genes (Table 2): ErbB signaling, olfactory transduction, long-term depression and calcium signaling pathways. The first three pathways are related to the neurosensory system in that ErbB signaling is known to be associated with a partial lack of sensory neurons and the development of neurodegenerative diseases (Yarden & Sliwkowski, 2001; Corfas et al., 2004); olfactory transduction is a series of events that sends nerve signals to the brain where they are perceived as smells; and long-term depression is the weakening of a neuronal synapse. Although no significant association of these neuron sensory pathways with human height has been reported (as far as we know), our results can be supported by the following suggestive evidence and hence would be worthy of further investigation. First, a large amount of recent data indicates that excessive ErbB signaling in humans is associated with the development of various types of tumors (Corfas et al., 2004; Bublil & Yarden, 2007; Cho & Leahy, 2002; Yang et al., 2006). Previous GWA studies have reported that height-associated SNPs were in or near genes involved in pathways related to cancer (Visscher, 2008; Lettre et al., 2008). Second, the long-term depression pathway is known to include the *IGF1* gene that is related to human height (Obrepalska-Steplowska et al., 2003; Rietveld et al., 2004; Sweeney et al., 2005) as well as bone density (Rivadeneira et al., 2006;

Table 1 Nineteen genes detected as underlying human height variation in previous GWA studies were identified in our GWA analysis. (C: Cho et al. (Cho et al., 2009), G: Gudbjartsson et al. (Gudbjartsson et al., 2008), L: Lettre et al. (Lettre et al., 2008), S: Soranzo et al. (Soranzo et al., 2009), W: Weedon et al. (Weedon et al., 2008)).

Nearly genes	RS ID	Class	Locus	BSS	Minor allele	MAF	Effect size	Previous GWA studies
PLAG1	rs13273123	intron	8q12.1b	99.6	G	0.066	−0.2465	C, G(rs10958476)
HMGA1	rs6918981	intron	6p21.31e	98.6	G	0.209	0.1598	C, G(rs1776897)
ETV6	rs17818498	intron	12p13.2a	98.2	C	0.295	0.1648	G(rs2187642)
FUBP3	rs11243976	intron	9q34.11e	96.7	A	0.206	0.1290	L(rs7466269)
LCORL	rs16859571	intron	4p15.32b	95.7	C	0.214	−0.1397	G(rs6830062), S(rs6830062), W(rs16896068)
LTBP1	rs41464348	intron	2p22.3d	94.5	A	0.354	−0.1265	C
HHIP	rs6812389	intron	4q31.22a	91.8	G	0.191	0.1118	G(rs1812175), L(rs1492820), S(rs1812175), W(rs6854783)
Unknown	rs17038182		1p12b	91.2		0.420	−0.1077	C
FBP2	rs600130	intron	9q22.32a	89.7	C	0.147	−0.1597	C
UQCC	rs1570004	intron	20q11.22b	84.7	A	0.271	0.1739	G(rs6088792), L(rs6060369), S(rs6088813)
ZBTB38	rs10513137	intron	3q23c	84.3	A	0.261	0.1165	C, W(rs6440003)
CRADD	rs10859569	intron	12q22b	82.6	T	0.403	0.0914	G(rs3825199)
BCAS3, TBX2	rs2079795		17q23.2b	81.8	T	0.328	0.1882	C, G(rs757608)
RUNX2	rs2677101	intron	6p12.3f	66.4	C	0.463	−0.0646	G(rs9395066)
CABLES1	rs12455718	intron	18q11.2b	54.6	C	0.233	−0.0909	G(rs4800148)
ANKS1A	rs13210323	intron	6p21.31d	51.9	A	0.431	0.0727	G(rs4913858)
EFEMP1	rs3791675	intron	2p16.1d	49.4	C	0.225	0.0119	C, W
ADCY3	rs6545800	intron	2p23.3c	49.2	T	0.440	−0.0329	G(rs6733301)
SUPT3H	rs10948197	intron	6p21.1a	38.3	C	0.337	−0.0432	G(rs9395066)

Rivadeneira et al., 2003) and body mass (Sweeney et al., 2005; Voorhoeve et al., 2006). Another pathway enriched in our analysis was the calcium signaling pathway, which has typical physiological roles– including muscle contraction and cellular motility– as well as biochemical roles such as regulating enzyme activity, ion pumping, and components of the cytoskeleton. In addition, calcium signaling has been known to be important in a variety of developmental events in the vertebrate embryo because it affects cell fate specification and morphogenesis (Slusarski & Pelegri, 2007).

At the validation stage, the replication likeliness of the jointly identified 516 SNPs was further evaluated by computing the proposed BSS for each SNP. While 60 SNPs had BSS $< 50\%$, 129 SNPs were selected with BSS $\geq 95\%$ (Supporting Table 1). These 129 SNPs would be considered empirically replicated SNPs related to the height trait in the Korean population. Among the 129 SNPs, 64 SNPs were mapped to the known genes, and included three nonsynonymous SNPs in *PKD1L2*, *UNC84B* and *SLBP* as well as four upstream/downstream SNPs of *UCHL3*, *ZGPAT*, *LOC100130828* and *SLFN12L*. Some previously reported genes (*PLAG1*, *HMGA1*, *ETV6*, *FUBP3* and *LCORL*) were identified with high BSS $\geq 95\%$ in our analysis. A total of 11 GO terms were over-represented in the 64 genes, including the biological processes related to transcriptional regulation, signaling, proteolysis, cell structure and other metabolisms (Supporting Fig. 1). Using the DAVID classification tool with default clustering options and median classification stringency, we found that 16 such known genes were clustered into three groups (Table 3). For example, four genes in the glycosyltransferase group are involved in glycan structures biosysthesis 1 (*EXT1*, *XYLT1*, *GALNTL4*) or 2 (*ST3GAL5*) KEGG pathways. Recent studies have shown that the glycosylation, or the attachment of glycans to proteins, provides the functional diversity needed to generate extensive phenotypes from a limited genotype (Raman et al., 2005). Many lines of evidence indicate the functional roles of glycans in cell growth and development, tumor growth and metastasis, anticoagulation, immune recognition/response, cell-cell communication and microbial pathogenesis (Raman et al., 2005). On the basis of OMIM data searching (http://www.ncbi.nlm.nih.gov/omim/), EXT1 is known to be associated with chondrosarcoma and multiple hereditary exostoses (MHE; an autosomal dominant disorder characterized by multiple projections of bone capped by cartilage); *ST3GAL5* is associated with Amish infantile epilepsy syndrome; and *XYLT1* is a candidate modifier in the severity of pseudoxanthoma elasticum. Our results of the GO and biological function analyses implicated that a clustering of the genes exhibiting genetic association with human height

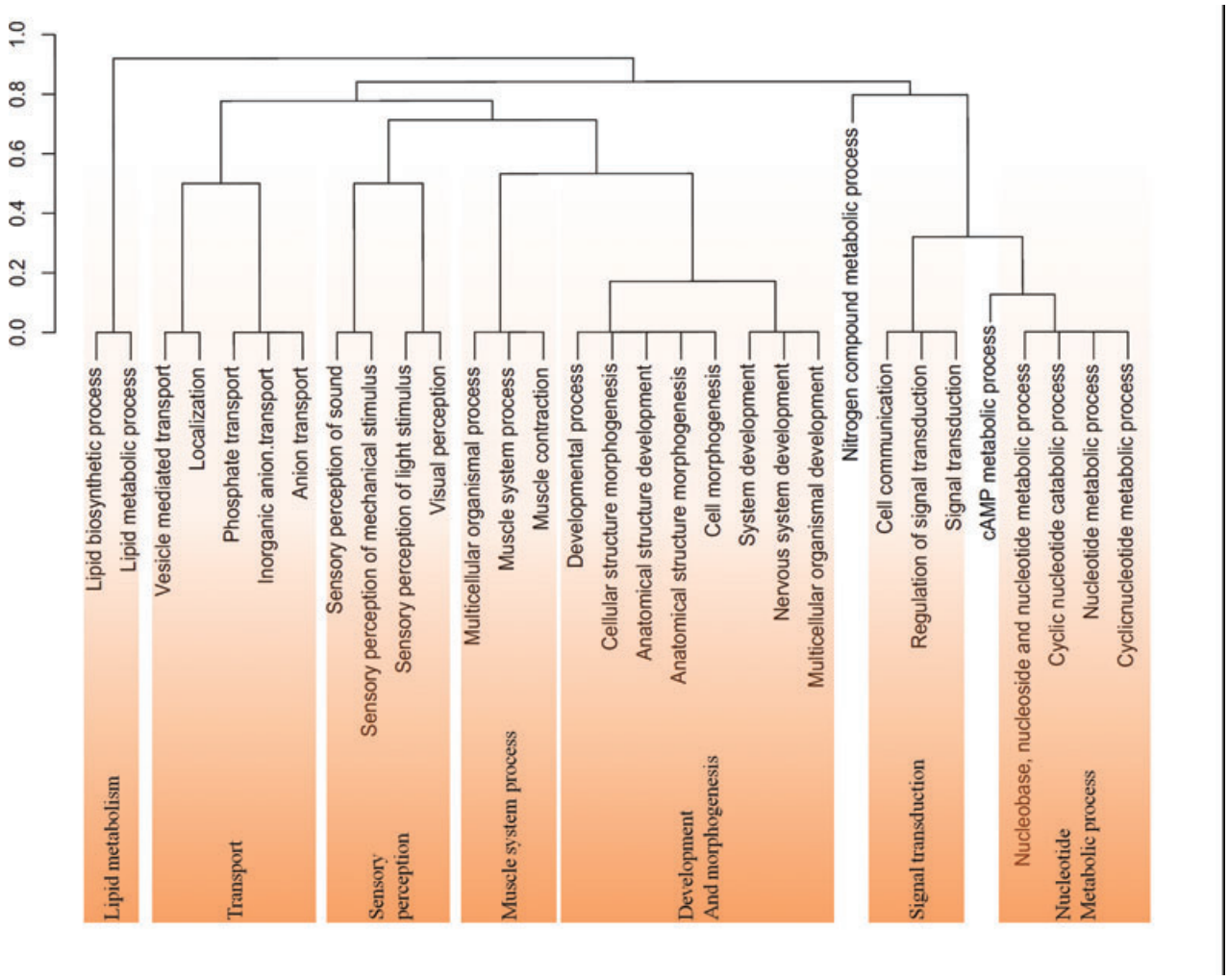

Figure 2. Hierarchical clustering of over-represented GO terms enriched in the 208 genes jointly associated with Korean adult height. Note that dissimilarity between two GO terms, X and Y, was defined as $d_{XY} = 1 - |\mathcal{N}(X) \cap \mathcal{N}(Y)| / \min\{|\mathcal{N}(X)|, |\mathcal{N}(X)|\}$ where N(C) denotes the set of genes to GO category C (Kosiol et al. 2008).

Table 2 KEGG pathway terms enriched in the 208 genes associated with Korean adult height.

KEGG pathway	Gene	FE[a]	P-value[b]
ErbB signaling	CAMK2D, ERBB3, NRG3, PLCG2	3.606	0.094
Olfactory transduction	ADCY3, CAMK2D, PRKG1	7.926	0.052
Long-term depression	IGF1, ITPR3, PRKG1, PPP2R2C	4.086	0.070
Calcium signaling	ADCY3, CAMK2D, ERBB3,ITPR3, PLCG2, SLC8A1	2.642	0.070

[a]FE stands for fold enrichment
[b]P-values were calculated based on EASE statistics

are likely related to a series of processes: (1) signal transduction including membrane receptor activation via glycosylation regulated by a group of glycosyltransferases; (2) then kinase activity; and (3) finally transcriptional regulation in the nucleus. This speculation might be worth further biological verification in that it would provide an insight into the genetic orchestration of a group of the height-associated genes in humans.

Among 129 height-related SNPs, eight SNPs showed the highest replication likeliness (i.e., BSS = 100%), and were

Table 3 Functional clusters identified in height-associated genes with BSS ≥ 95%.

Representative term in each group	Gene
Glycosyltransferase	EXT1, GALNTL4, ST3GAL5, XYLT1
Protein kinase	CDK10, HERC2, KALRN, PRKCH, PRKG1
Regulation of transcription in nucleus	ETV6, EYA4, FUBP3, LCORL, MYT1L, PLAG1, RORA

located in or near five known genes (*ADAMSL1*, *UCHL3*, *CSMD1*, *ST3GAL5*, *FREM1*). Based on these eight SNPs, we investigated the cumulative effect of the jointly identified genetic variants via the proposed approach on the trait. The cumulative predictive value of a set of trait-related variants can be assessed by counting the number of height-increasing alleles in each individual (i.e., "height score" (Lettre et al., 2008)).

As an alternative to the height score, we propose a "genotypic score" that incorporates the number of height-increasing alleles and their relative effect sizes. For the jth individual, the genotypic score GS_j is defined as below:

$$GS_j = \sum_{i=1}^{8} w_i x_{ij} \qquad \text{with } w_i = \frac{8|\hat{\beta}_i|}{\sum_{k=1}^{8}|\hat{\beta}_k|} \tag{6}$$

where x_{ij} is the number of height-increasing alleles, w_i is the weight, and $\hat{\beta}_i$ is the estimated effect size for the ith SNP. In fact, the genotypic score is a weighted version of the height score and indicates the weighted sum of height-increasing alleles. We classified all individuals with complete genotypes for the eight SNPs into nine groups according to their genotypic scores. Among the individuals in this population, 1.5% of people have a genotypic score less than 5, and 1.2% of people have a genotypic score more than 12 (Fig. 3). To examine the relationship between the genotypic score and height, the average height was computed for each group and regressed on the genotypic score (Fig. 3). We observed a linear increase in the average height with an increasing genotypic score ($R^2 = 0.917$). For a unit increase in the genotypic score, the average height increases by 0.469 cm. The average height difference between the smallest and largest genotypic groups was 4.55 cm and 4.12 cm, respectively, for males and females.

The explanatory power of the SNPs identified via the proposed approach was investigated by fitting a multiple regression model with the identified SNPs and examining the adjusted R^2 value. For each of various BSS cut-off values (i.e., 80 ∼ 100%), we calculated the adjusted R^2 value from a multiple regression model with the corresponding number of the significant SNPs (Supporting Table 2). For the purpose of comparison, we also looked at the explanatory power of the SNPs identified via the conventional approach using single-SNP association tests. Note that multiple regression models were fitted with the same numbers of SNPs from the conventional approach as from the proposed approach. In Supporting Table 2, we present the adjusted R^2 for the 290 BSS > 80% SNPs and the top 290 SNPs. Based on these values, we would conclude that 73% of phenotypic variation is explained by the 290 BSS > 80% SNP variants while 67% is explained by the top 290 SNPs. Overall, the observations in Figure 4 would indicate that our approach using a BSS cut-off can identify a set of SNPs providing a better explanation of phenotypic variation than the single marker test approach. Furthermore, this power difference increased as the number of SNPs in multiple regression models increased.

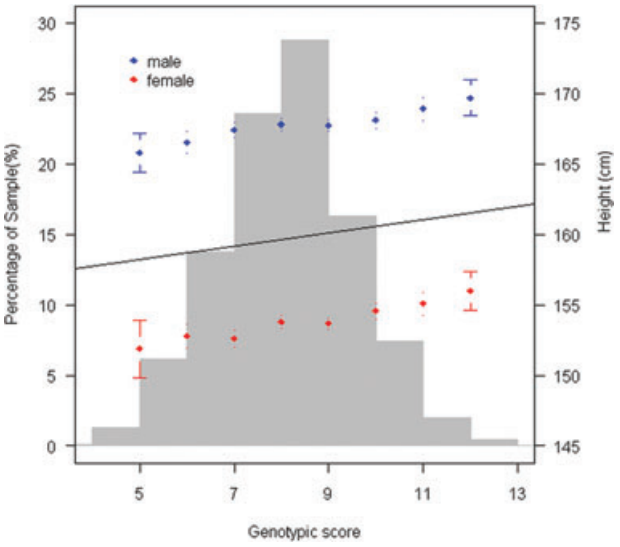

Figure 3. Cumulative effects of eight SNPs identified with 100% BSS on Korean adult height. The genotypic score was computed as the weighted number of height-increasing alleles for each of the samples having complete genotypes at the eight SNPs. According to the genotypic scores, the individuals were classified into nine groups. The light grey histogram in the background represents the relative fraction of individuals in each genotypic score group. The black line indicates the regression of the average height of each group on the genotypic score. For each group, the mean ± 95% confidence intervals of height were plotted separately for males and females.

Simulations were performed to investigate the validity of using the 10-fold cross-validation method to select the tuning parameter λ in the elastic-net stage (Supporting Table 3). The performance was described with the number of true positives (TP), sensitivity and specificity. Under mild and low

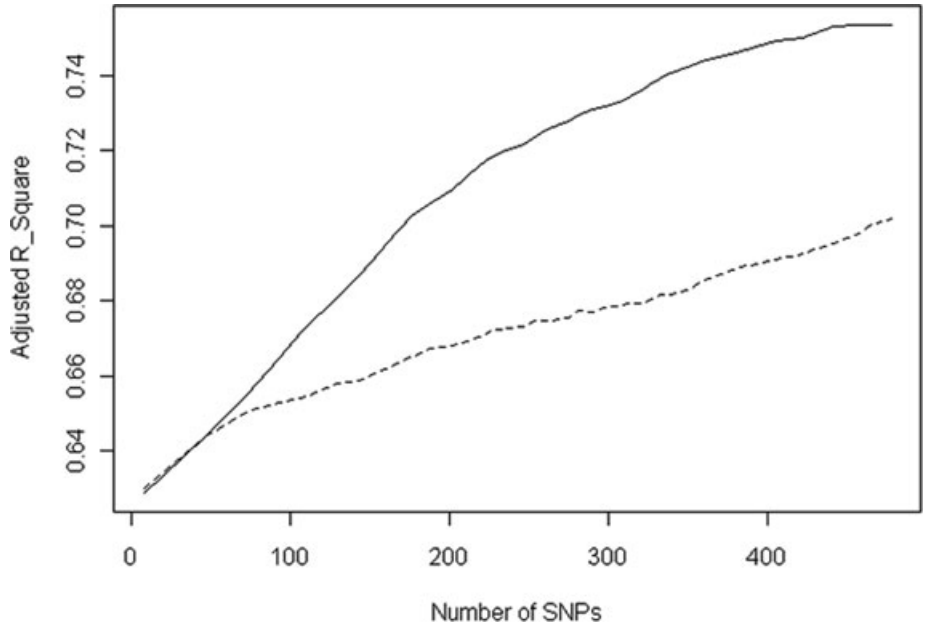

Figure 4. Explanatory power comparison between the proposed approach (solid line) and the conventional single-marker approach (dashed line) via adjusted R^2.

correlations, TP was close to 20 overall, even when signal-to-noise is small. Sensitivity and specificity were also fairly good. Although TP suffered from high correlation, sensitivity and specificity were not especially bad. Based on these limited simulations, the cross-validation method seems to work well in selecting true causal SNPs via the elastic-net variable selection method.

Discussion

The single-marker association test has been employed as a conventional approach for GWA studies of common complex traits. Despite many successes, it can be expected to be inefficient for identifying genetic factors of complex multifactorial traits because individual SNPs are independently searched for phenotypic association. In this study, we propose a multi-stage approach based on elastic-net regression, which facilitates joint detection of multiple genetic factors underlying phenotypic variation and hence is suited for GWA analysis of complex traits. Furthermore, our approach enables an efficient search for potential associations in the extremely large number of SNPs in GWA studies by reducing high dimensionality at the pre-screening stage. Using a specific data-set based on Korean adult height, our strategy was demonstrated to be useful for GWA analysis of quantitative traits. By employing multiple logistic regression, our strategy can also be applicable to case-control traits.

The proposed approach employs the elastic-net penalty (i.e., a combination of LASSO and ridge penalties) in multiple regression to take advantage of regularization properties, such as automatic variable selection and stable estimation in the presence of multicollinearity (i.e., correlation structure among genetic factors or SNPs), and hence has the following two advantages against ordinary multiple regression as well as the conventional single-marker approach. First, elastic-net regression tends to provide more reliable identification than ordinary multiple regression which is very sensitive to multicollinearity resulting from highly correlated predictors. In the existence of such multicollinearity, the estimates of regression coefficients (e.g., effect sizes of SNPs) become unstable possibly with large variance, particularly when there exist groups of many highly correlated SNPs (e.g., SNPs in high linkage disequilibrium) (Zou & Hastie, 2005; Friedman et al., 2008). Because a ridge penalty induces shrinkage of predictors, these coefficients can be stably estimated. Furthermore, a LASSO penalty provides an effective variable selection for high-dimensional problems, especially in sparse cases (i.e., only a relatively small number of SNPs are related to the trait). By combining these two penalties, elastic-net regression obtains an estimation stability based on which the joint identification of phenotype-related SNPs can be more reliably conducted.

In GWA analysis, linkage disequilibrium (LD) among nearby SNPs is an apparently existing phenomenon which can provide valuable information in understanding genetic

structure. Thus, it would be important and very interesting to take LD into account in a direct manner at the elastic-net stage. Although LD was not taken into account in this study, the elastic-net uses information on multiple correlated variables in the selection process in an indirect manner (Zou & Hastie, 2005). This feature might make the elastic-net superior to other regularization-based variable selection techniques (e.g., Lasso).

Replication has been a standard procedure to separate true associations from false positives before biological interpretation in GWA studies (Ioannidis et al., 2001). However, it cannot always be done in practice; furthermore, its success depends on the comparability of the original and replication datasets (e.g., differences in population structure and study design). Because the bootstrap samples are resampled from the original data and are likely to share population structure and study design with the original data, they can be used as intra-replication studies that are comparable to the original data. In other words, the BSS value for a specific SNP would indicate how likely it is that the identification of the SNP will be reproducible in replication studies. This feature of predicting the replicability can be very beneficial in GWA studies because researchers can eliminate false-positive associations and focus on a smaller number of replicable SNPs without a costly, independent replication study. To demonstrate the validity of the proposed empirical replication procedure, we used two sub-samples of the original data according to recruitment areas (Ansan and Ansung), and conducted replication studies. While only ~50% of all 516 SNPs identified at Stage 2 were replicated, ~90% of 129 SNPs having BSS $\geq$ 95% were replicated in two sub-samples (Fig. 5). This indicates that SNPs with higher BSS were more likely to be replicated in two sub-samples.

In addition, the proposed BSS provides a statistical means to evaluate the statistical relevance of each SNP and can be used as an alternative to the statistical significance (i.e., *p*-values). Testing-based methods, including the conventional single-marker tests, identify the phenotypic association of each SNP based on its *p*-value. In contrast, regularization-based methods, including elastic-net regression, detect phenotypic associations by penalising regression coefficients and thus are unable to provide the statistical significance of identified SNPs (Kyung et al., 2009). There have been efforts to develop an explicit estimate of statistical significance for the SNPs identified via regularization methods. For example, bootstrapping approaches have been proposed to estimate standard errors of regression coefficients in Lasso and ridge penalized regressions (Tibshirani, 1996; Crivelli et al., 1995). Recently, Kyung et al. (Kyung et al., 2009) showed that the bootstrap approach does not provide valid estimates of standard errors for the coefficients that are shrunk to zero in Lasso regression, and suggested a Bayesian formulation for producing valid standard error estimates, which does not appear to be applicable to real GWA data. In this study, we developed BSS as an alternative to the conventional *p*-values to evaluate the statistical importance of the identified SNPs. A higher value of BSS implies a larger chance of being replicated in bootstrap samples.

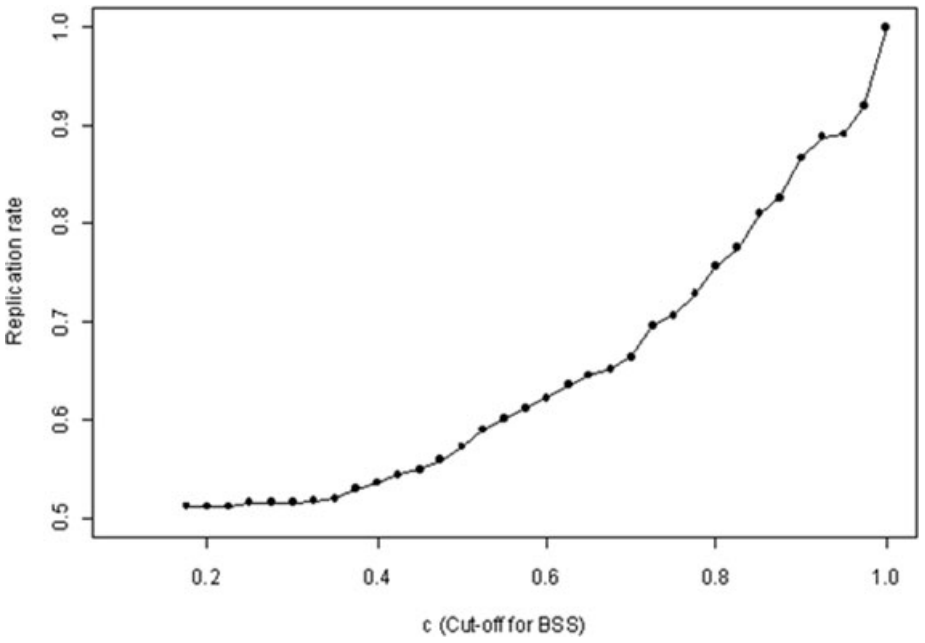

Figure 5. Replication rate in two sub-samples for identified SNPs with BSS $\geq$ c.

While the pre-screening stage is known to be beneficial to both computation and accuracy, it has a practical limitation. Currently, there is no gold standard on how many SNPs having a weak correlation with the phenotype have to be eliminated. When too many SNPs are screened out, one may miss the chance to investigate some true causal SNPs (i.e., false negatives) that can be kept if a smaller number of SNPs are screened out. On the other hand, with regard to the computation aspect, it can be inefficient to keep too many SNPs. One option would be to keep the maximum number of SNPs that can be handled by computation at the main stage to avoid possible false negatives. In order to examine the effects of thresholds at the pre-screening stage in our real data analysis, we considered three pre-screened datasets via different thresholds (i.e., top 1000, top 2000 and top 3000 SNPs with the strongest marginal association). Because the elastic-net reflects correlations among SNPs in variable selection, it is not surprising that the selection results were somewhat different among the pre-screened datasets (data not shown). However, we observed that the SNPs which were commonly selected across the different pre-screened datasets have higher BSS (namely, 15% higher on average) than the SNPs which were selected from only one pre-screened dataset. This observation may indicate that SNPs with high BSS tend to be selected commonly across different pre-screened datasets, and thus the threshold effect might be not severe. Also, note that the false discovery rate (FDR) for the top 1000 SNPs was about 0.52.

The proposed approach can be extended in several ways. First, the power of our approach can be improved by incorporating prior biological knowledge. Recently, a network-constrained regularization procedure was proposed to incorporate information from biological network graphs into linear regression, where the network is represented as a graph and its corresponding Laplacian matrix (Li & Li, 2008). Similarly, to incorporate prior knowledge in the form of biological pathways or networks, Pan et al. (Pan et al., 2009) proposed a grouped penalty that smoothes the regression coefficients of the predictors over the network using microarray data. Another method that transforms literature-based gene association scores to prior probabilities of networks has been applied to learning gene sub-networks (Steele et al., 2009). These kinds of penalties can be additionally introduced into our approach to increase the power of the penalized regression for GWA studies. Second, our approach can be used to investigate genetic interactions and biological processes, such as metabolic pathways and transcriptional programs, which underlie complex traits in GWA studies. Elastic-net variable selection enables the effective identification of groups of (even highly) correlated SNPs associated with the trait. Thus, our approach can be desirable for GWA analysis, even when pathway analysis is subsequently conducted for a more biologically meaningful interpretation of the results. Furthermore, pathway-level analysis can be facilitated by applying our approach to grouped SNPs in biological pathways, and the identification of pathways and processes influencing the trait would ease the interpretation of a large-scale experiment.

Acknowledgements

This work was supported by the Consortium for Large Scale Genome Wide Association Study (2008-E00355-00), 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), Jong-Eun Lee (DNA Link, 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 this valuable data. The National Science Foundation of China (30730057) supported Jurg Ott in joining the project.

References

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statistical Society: Series B* **57**, 289–300.

Bublil, E. & Yarden, Y. (2007) The EGF receptor family: spearheading a merger of signaling and therapeutics. *Curr Opin Cell Biol* **19**, 124–134.

Cho, H. & Leahy, D. (2002) Structure of the extracellular region of HER3 reveals an interdomain tether. *Science* **297**, 1330.

Cho, Y., Go, M., Kim, Y., Heo, J., Oh, J., Ban, H., Yoon, D., Lee, M., Kim, D. & Park, M. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* **41**, 527–534.

Corfas, G., Roy, K. & Buxbaum, J. (2004) Neuregulin 1-erbB signaling and the molecular/cellular basis of schizophrenia. *Nat Neurosci* **7**, 575–580.

Crivelli, A., Firinguetti, L., Monta, O. R. & Mu, Z. M. (1995) Confidence intervals in ridge regression by bootstrapping the dependent variable: A simulation study. *Commun Stat Simul C* **24**, 631–652.

Fan, J. & Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statistical Society: Series B* **70**, 849–911.

Friedman, J., Hastie, T. & Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Software* **33**(1).

Goldstein, D. (2009) Common genetic variation and human traits. *New Engl J Med* **360**, 1696.

Gudbjartsson, D., Walters, G., Thorleifsson, G., Stefansson, H., Halldorsson, B., Zusmanovich, P., Sulem, P., Thorlacius, S.,

Gylfason, A. & Steinberg, S. (2008) Many sequence variants affecting diversity of adult human height. *Nat Genet* **40**, 609–615.
Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F. & Manolio, T. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* **106**, 9362.
Hosack, D. A., Dennis, G., Jr., Sherman, B. T., Lane, H. C. & Lempicki, R. A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol* **4**, R70.
Ioannidis, J., Ntzani, E., Trikalinos, T. & Contopoulos-Ioannidis, D. (2001) Replication validity of genetic association studies. *Nat Genet* **29**, 306–309.
Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277.
Kosiol, C., Vina, T., Da Fonseca, R., Hubisz, M., Bustamante, C., Nielsen, R. & Siepel, A. (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet* **4**, e1000144.
Kraft, P. & Hunter, D. (2009) Genetic Risk Prediction–Are We There Yet? *New Engl J Med* **360**, 1701.
Kyung, M., Gill, J., Ghosh, M. & Casella, G. (2010) Penalized Regression, Standard Errors, and Bayesian Lassos. *BayesAn* **5**, 1–44.
Le Cessie, S. & Van Houwelingen, J. (1992) Ridge estimators in logistic regression. *Appl Statist* **41**, 191–201.
Lettre, G., Butler, J. L., Ardlie, K. G. & Hirschhorn, J. N. (2007) Common genetic variation in eight genes of the GH/IGF1 axis does not contribute to adult height variation. *Hum Genet* **122**, 129–39.
Lettre, G., Jackson, A., Gieger, C., Schumacher, F., Berndt, S., Sanna, S., Eyheramendy, S., Voight, B., Butler, J. & Guiducci, C. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40**, 584–591.
Li, C. & Li, H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175–82.
Li, H., Wetten, S., Li, L., St Jean, P., Upmanyu, R., Surh, L., Hosford, D., Barnes, M., Briley, J. & Borrie, M. (2008) Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol* **65**, 45.
Li, M., Liu, P., Li, Y., Qin, Y., Liu, Y. & Deng, H. (2004) A major gene model of adult height is suggested in Chinese. *J Hum Genet* **49**, 148–153.
Obrepalska-Steplowska, A., Kedzia, A., Trojan, J. & Gozdzicka-Jozefiak, A. (2003) Analysis of coding and promoter sequences of the IGF-I gene in children with growth disorders presenting with normal level of growth hormone. *J Pediatr Endocrinol Metab* **16**, 1267–75.
Pan, W., Xie, B. & Shen, X. (2009) Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, in press.
Park, M. & Hastie, T. (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30.
Raman, R., Raguram, S., Venkataraman, G., Paulson, J. C. & Sasisekharan, R. (2005) Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat Methods* **2**, 817–824.
Rietveld, I., Janssen, J. A., Van Rossum, E. F., Houwing-Duistermaat, J. J., Rivadeneira, F., Hofman, A., Pols, H. A., Van Duijn, C. M. & Lamberts, S. W. (2004) A polymorphic CA repeat in the IGF-I gene is associated with gender-specific differences in body height, but has no effect on the secular trend in body height. *Clin Endocrinol (Oxf)* **61**, 195–203.
Rivadeneira, F., Houwing-Duistermaat, J. J., Vaessen, N., Vergeer-Drop, J. M., Hofman, A., Pols, H. A., Van Duijn, C. M. & Uitterlinden, A. G. (2003) Association between an insulin-like growth factor I gene promoter polymorphism and bone mineral density in the elderly: the Rotterdam Study. *J Clin Endocrinol Metab* **88**, 3878–84.
Rivadeneira, F., Van Meurs, J. B., Kant, J., Zillikens, M. C., Stolk, L., Beck, T. J., Arp, P., Schuit, S. C., Hofman, A., Houwing-Duistermaat, J. J., Van Duijn, C. M., Van Leeuwen, J. P., Pols, H. A. & Uitterlinden, A. G. (2006) Estrogen receptor beta (ESR2) polymorphisms in interaction with estrogen receptor alpha (ESR1) and insulin-like growth factor I (IGF1) variants influence the risk of fracture in postmenopausal women. *J Bone Miner Res* **21**, 1443–56.
Saxena, R., Voight, B., Lyssenko, V., Burtt, N., De Bakker, P., Chen, H., Roix, J., Kathiresan, S., Hirschhorn, J. & Daly, M. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336.
Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R. & Klein, B. (2008) Lasso-patternsearch algorithm with application to ophthalmology and genomic data. *Statistics and its interface* **1**, 137.
Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A. & Hadjadj, S. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885.
Slusarski, D. C. & Pelegri, F. (2007) Calcium signaling in vertebrate embryonic patterning and morphogenesis. *Dev Biol* **307**, 1–13.
Soranzo, N., Rivadeneira, F., Chinappen-Horsley, U., Malkina, I., Richards, J., Hammond, N., Stolk, L., Nica, A., Inouye, M. & Hofman, A. (2009) Meta-Analysis of Genome-Wide Scans for Human Adult Stature Identifies Novel Loci and Associations with Measures of Skeletal Frame Size. *PLoS Genet* **5**, e1000445.
Sovio, U., Bennett, A., Millwood, I., Molitor, J., O'reilly, P., Timpson, N., Kaakinen, M., Laitinen, J., Haukka, J. & Pillas, D. (2009) Genetic Determinants of Height Growth Assessed Longitudinally from Infancy to Adulthood in the Northern Finland Birth Cohort 1966. *PLoS Genet* **5**, e1000409.
Steele, E., Tucker, A. T., Hoen, P. A. & Schuemie, M. J. (2009) Literature-based priors for gene regulatory networks. *Bioinformatics* **25**, 1768–74.
Sweeney, C., Murtaugh, M. A., Baumgartner, K. B., Byers, T., Giuliano, A. R., Herrick, J. S., Wolff, R., Caan, B. J. & Slattery, M. L. (2005) Insulin-like growth factor pathway polymorphisms associated with body size in Hispanic and non-Hispanic white women. *Cancer Epidemiol Biomarkers Prev* **14**, 1802–9.
Thorleifsson, G., Walters, G., Gudbjartsson, D., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S. & Jonsdottir, I. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet* **41**, 18–24.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statistical Society: Series B* **58**, 267–288.
Visscher, P. M. (2008) Sizing up human height variation. *Nat Genet* **40**, 489–90.
Voorhoeve, P. G., Van Rossum, E. F., Te Velde, S. J., Koper, J. W., Kemper, H. C., Lamberts, S. W. & De Waal, H. A. (2006) Association between an IGF-I gene polymorphism and body fatness: differences between generations. *Eur J Endocrinol* **154**, 379–88.

Wallace, C., Newhouse, S., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R., Mar Ano, A. & Hajat, C. (2008) Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet* **82**, 139–149.

Weedon, M., Lango, H., Lindgren, C., Wallace, C., Evans, D., Mangino, M., Freathy, R., Perry, J., Stevens, S. & Hall, A. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40**, 575–583.

Westfall, P. & Young, S. (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, New York. pp. 340.

Wu, T., Chen, Y., Hastie, T., Sobel, E. & Lange, K. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714.

Wu, T. & Lange, K. (2008) Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat* **2**, 224–244.

Yang, T., Xiong, D., Guo, Y., Recker, R. & Deng, H. (2006) Association analyses of CYP19 gene polymorphisms with height variation in a large sample of Caucasian nuclear families. *Hum Genet* **120**, 119–125.

Yarden, Y. & Sliwkowski, M. (2001) Untangling the ErbB signalling network. *Nature Rev Mol Cell Biol* **2**, 127–137.

Zou, H. (2006) The adaptive lasso and its oracle properties. *JASA* **101**, 1418–1429.

Zou, H. & Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. Roy. Statistical Society: Series B* **67**, 301–320.

Supporting Information

Additional supporting information may be found in the online version of this article:

Figure S1. GO terms over-represented in the height-associated genes with BSS $\geq$ 95%.

Table S1. Height-associated SNPs identified with BSS $\geq$ 95%

Table S2. Adjusted R^2 in multiple regressions with the SNPs identified via the proposed elastic-net variable selection vs the common single-marker test.

Table S3. Simulation results: performance of 10-fold cross-validation.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organised for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received: 5 February 2010
Accepted: 28 April 2010