

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

Sohee Oh¹, Seoae Cho² and Taesung Park^{1*}

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 that 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 the 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 (under weight), 18.5-24.9 kg/m² (normal weight), 25.0-29.9 kg/m² (over weight), 30.0-30.9 kg/m² (obese), and ≥40 kg/m² (extremely obese). Scuteri *et al.* conducted association studies with BMI, hip circumference, 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 Family 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 CTNBL1 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-

*Corresponding author: E-mail tspark@stats.snu.ac.kr
Tel +82-2-880-8924, Fax +82-2-883-6144
Accepted 2 September 2010

ducted genome-wide association studies for BMI and weight. Using 11 significant genome-wide loci, they also conducted association tests with obesity (obese for $BMI \geq 30$, 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 ($BMI \geq 25$) or obese ($BMI \geq 30$)). 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 Ansong ($n=5018$) 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 Ansong), 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):

$$y_j = \gamma_0 + \gamma_1 Sex_j + \gamma_2 Age_j + \gamma_3 Area_j + \sum_{i=1}^p \beta_i SNP_{ij} + \varepsilon_j$$

where y_j , Sex_j , Age_j , $Area_j$, SNP_{ij} , and ε_j represent the trait value, sex, age, area, the number of minor alleles for the i th SNP marker, and the measurement error of the j th individual, respectively; $i=1, \dots, p$, and p is the number of SNPs under consideration; and $j=1, \dots, n$, and n is the number of individuals. γ_0 , γ_s , and β_i s 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 ($n \ll p$). EN regularization solves the following problem:

$$\min_{(\alpha, \lambda, \beta) \in \mathbb{R}^{p+1}} \left[\frac{1}{2n} \sum_{j=1}^n (y_j - \gamma_0 - \gamma_1 Sex_j - \gamma_2 Age_j - \gamma_3 Area_j - \sum_{i=1}^p \beta_i SNP_{ij})^2 + \lambda P_\alpha(\beta) \right]$$

where $SNP_j = \{SNP_{1j}, SNP_{2j}, \dots, SNP_{pj}\}^T$ for the j th individual, $\beta = (\beta_1, \dots, \beta_p)^T$. Here, λ is a tuning parameter, and the EN is defined as follows:

$$p_\alpha(\beta) = (1 - \alpha) \frac{1}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 = \sum_{i=1}^p \left[\frac{1}{2} (1 - \alpha) \beta_i^2 + \alpha |\beta_i| \right]$$

with another tuning parameter, α . 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. The 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 λ 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 λ and α , 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

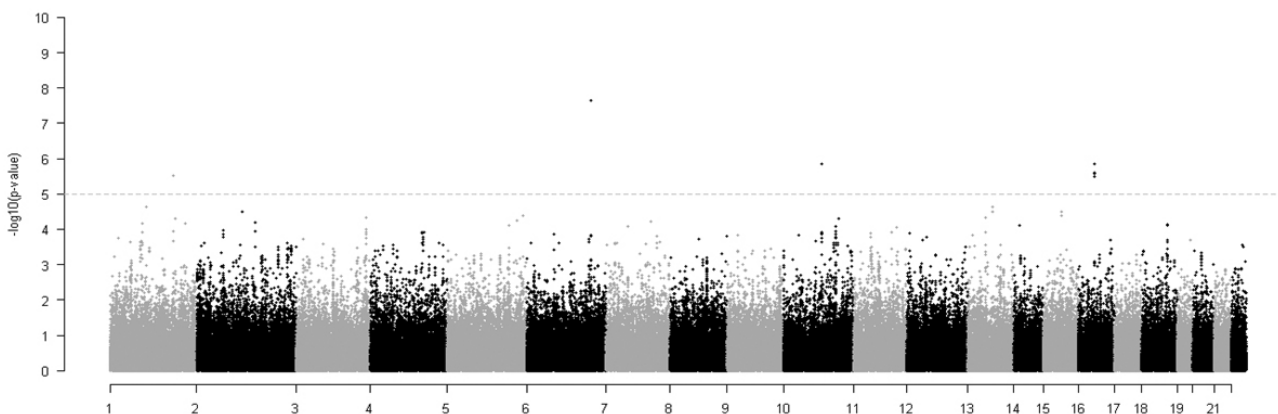
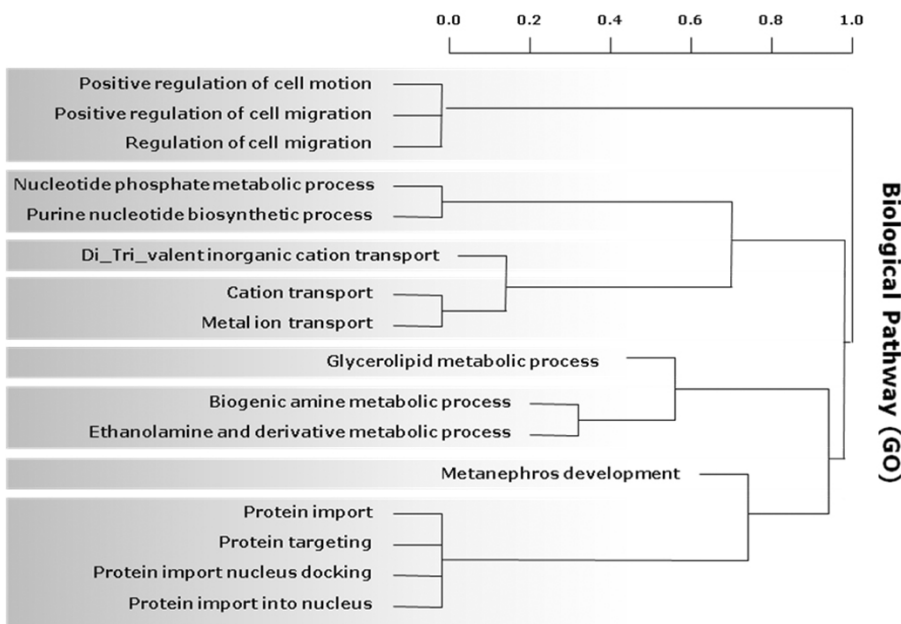


Fig. 1. Results of single-marker association tests for 327,872 SNPs in the first stage.

Table 1. Nine genes detected as underlying human obesity variations in previous GWA studies were identified in our GWA analysis

Nearly genes	rs ID	Class	Locus	Minor allele	MAF	Effect size	Previous GWA studies
SEC16B	Rs623479	3' UTR	1q25.2b	C	0.15	-0.0335	Thorleifsson (rs10913469-C)
RARB	Rs17016781	Intron	3p24.2a	G	0.40	0.0171	Cotsapas (rs1435703-T)
ERC2	Rs9875049	Intron	3p14.3c	C	0.09	0.0276	Kiel (rs2054989)
ADARB2	Rs11250504	Intron	10p15.3c	T	0.13	-0.0014	Lowe (rs6560749-T)
PANK1	Rs10881614	Intron	10q23.31c	C	0.38	0.0607	Sabatti (rs11185790-A)
BDNF	Rs11030104	Intron	11p14.1d	G	0.44	-0.0079	Thoreleifsson (rs6265-G) Thoreleifsson (rs925946-T) Thoreleifsson (rs7481311-T)
FTO	Rs9939609	Intron	16q12.2b	A	0.13	0.1873	Scherag (rs1558902-A) Cotsapas (rs9941349-T) Cho, Willer, Frayling Scuteri (rs9930506-A)
LOC729076	Rs17178527	Unknown	6q24.1	A	0.25	-0.1186	Cho
Unknown	Rs17089410	Unknown	13q21.33	C	0.13	0.0813	Cho

**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-

Table 2. Enrichment KEGG pathway

KEGG_pathway	p-value	Genes	FE
Endocytosis	0,048	FLT1, ERBB4, ARRB1, SH3GLB1, NEDD4L, IQSEC1	2,961
Axon guidance	0,05	ABLIM1, PLXNC1, PLXNA2, SRGAP3, SRGAP1	3,52

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 ($p\text{-value} < 1 \times 10^{-5}$), 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.

References

- Abdi, H. (2007). The Bonferroni and Sidak corrections for multiple comparisons. In *Encyclopedia of measurement and statistics*, Salkind N, ed. Thousand Oaks, CA: Sage, pp. 1-9.
- Bell, C.G., Walley, A.J., and Froguel, P. (2005). The Genetics of Human Obesity. *Nat. Rev. Genet.* 6, 221-234.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R. Statist. Soc. B* 57, 289-300.
- Cho, S., Kim, K., Kim, Y.J., Lee, J.K., Cho, Y.S., Lee, J.Y., Han, B.G., Kim, H., Ott, J., and Park, T. (2010). Joint identification of multiple genetic variants via EN variable selection in a genome-wide association analysis. *Ann. Hum. Genet.* 74, 416-428.
- Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M., Cha, S.H., Kim, J.W., Han, B.K., Min, H., Ahn, Y., Park, M.S., Han, H.R., Jang, H.Y., Cho, E.Y., Lee, J.E., Cho, N.H., Shin, C., Park, T., Park, J.W., Lee, J.K., Cardon, L., Clarke, G., McCarthy, M.I., Lee, J.Y., Lee, J.W., Oh, B., and Kim, H.R. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* 41, 527-534.
- Cotsapas, C., Speliotes, E.K., Hatoum, I.J., Greenawalt, D.M., Dobrin, R., Lum, P.Y., Suver, C., Chudin, E., Kemp, D., Reitman, M., Voight, B.F., Neale, B.M., Schadt, E.E., Hirschhorn, J.N., Kaplan, L.M., Daly, M.J., and the GIANT consortium. (2009). Common body mass index associated variants confer risk of extreme obesity. *Hum. Mol. Genet.* 18, 3502-3507.
- Das, U.N., and Rao, A.A. (2007). Gene expression profile in obesity and type 2 diabetes mellitus. *Lipids in Health and Disease* 6, 35.
- Farooqi, I.S. and O' Rahilly, S. (2007). Genetic factors in human obesity. *Obesity Rev.* 8 (Suppl. 1) 37-40.
- Feitosa, M.F., Borecki, I.B., Rich, S.S., Arnett, D.K., Sholinsky, P., Myers, R.H., Leppert, M., and Province, M.A. (2002). Quantitative-trait loci influencing body-mass index reside on chromosome 7 and 13: The National Heart, Lung, and Blood Institute Family Heart Study. *Am. J. Hum. Genet.* 70, 72-82.
- Heard-Costa, N.L., Zillikens, M.C., Monda, K.L., Johansson, Å., Harris, T.B., Fu, M., Haritunians, T., Feitosa, M., Aspelund, T., Eiriksdottir, G., Garcia, M., Launer, L.J., Smith, A.V., Mitchell, B.D., McArdle, P.F., Shuldiner, A.R., Bielikski, S.J., Boerwinkle, E., Brancati, F., Demerath, E.W., Pankow, J.S., Arnold, A.M., Chen, Y.D.I., Glazer, N.L., McKnight, B., Psaty, B.M., Rotter, J.I., Amin, N., Campbell, H., Gyllenstein, U., Pattaro, C., Pramstaller, P.P., Rudan, I., Struchalin, M., Vitart, V., Gao, X., Kraja, A., Province, M.A., Zhang, Q., Atwood, L.D., Dupuis, J., Hirschhorn, J.N., Jaquish, C.E., O' Donnell, C.J., Vasani, R.S., White, C.C., Aulchenko, Y.S., Estrada, K., Hofman, A., Rivadeneira, F., Uitterlinden, A.G., Witteman, J.C.M., Oostra, B.A., Kaplan, R.C., Gudnason, V., O' Connell, J.R., Borecki, I.B., van Duijn, C.M., Cupples, L.A., Fox, C.S., and North, K.E. (2009). NRXN3 is novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS One* 5, e1000539.
- Hfker, M. and Wijmenga, C. (2009). A supersized list of obesity genes. *Nat. Genet.* 41, 139-140.
- Hill, J.O. and Peters, J.C. (1998). Environmental contributions to the obesity epidemic. *Science* 280, 1371-1374.
- Hinney, A., Nguyen, T.T., Scherag, A., Friedel, S., Brönner, G., Müller, T.D., Grallert, H., Illig, T., Wichmann, H.E., Rief, W., Schäfer, G., Gebebrand, J. (2007). Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS One* 2, e1361.
- Hirschhorn, J.N. and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95-108.
- Hosack, M., Dennis, G., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4, R70.
- Ichihara, S. and Yamada, Y. (2008). Genetic factors for human obesity. *Cell. Mol. Life. Sci.* 65, 1086-1098.
- Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, M.B., Ferris, F.L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 335-389.
- Kosiol, C., Vinař, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., Siepel, A. (2008). Patterns of positive selection in six mammalian genomes. *PLoS Genet* 4, e1000144.
- Le Cessie, S. and Van Houwelingen, J. (1992). Ridge estimators in logistic regression. *App. Statist.* 41, 191-201.
- Li, H., Wetten, S., Li, L., St Jean, P.L., Upmanyu, R., Surh, L., Hosford, D., Barnes, M.R., Briley, J.D., Borrie, M., Coletta, N., Delisle, R., Dhalla, D., Ehm, M.G., Feldman, H.H., Fornazzari, L., Gauthier, S., Goodgame, N., Guzman, D., Hammond, S., Hollingworth, P., Hsiung, G.Y., Johnson, J., Kelly, D.D., Keren, R., Kertesz, A., King, K.S., Lovestone, S., Loy-English, I., Matthews, P.M., Owen, M.J., Plumpton, M., Pryse-Phillips, W., Prinjha, R.K., Richardson, J.C., Saunders, A., Slater, A.J., St George-Hyslop, P.H., Stinnett, S.W., Swartz, J.E., Taylor, R.L., Wherrett, J., Williams, J., Yarnall, D.P., Gibson, R.A., Irizarry, M.C., Middleton, L.T., and Roses, A.D. (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch. Neurol.* 65, 45
- Liu, Y.J., Liu, X.G., Wang, L., Dina, C., Yan, H., Liu, J.F., Levy, S., Papanicolaou, C.J., Drees, B.M., Hamilton, J.J., Meyre, D., Delplanque, J., Pei, Y.F., Zhang, L., Recker, R.R., Froguel, P., and Deng, H.W. (2008). Genome-wide association scans identified CTNBL1 as a novel gene for obesity. *Hum. Mol. Genet.* 17, 1803-1813.
- 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., Berndt, S.I., Prostate, Lung, Colorectal,

- and Ovarian (PLCO) Cancer Screening Trial, Jacobs, K.B., Chanock, S.J., Hayes, R.B., Bergmann, S., Bennett, A.J., Bingham, S.A., Bochud, M., Brown, M., Cauchi, S., Connell, J.M., Cooper, C., Smith, G.D., Day, I., Dina, C., De, S., Dermitzakis, E.T., Doney, A.S., Elliott, K.S., Elliott, P., Evans, D.M., Sadaf Farooqi, I., Froguel, P., Ghorji, J., Groves, C.J., Gwilliam, R., Hadley, D., Hall, A.S., Hattersley, A.T., Hebebrand, J., Heid, I.M., KORA, Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H.E., Herrera, B., Hinney, A., Hunt, S.E., Jarvelin, M.R., Johnson, T., Jolley, J.D., Karpe, F., Keniry, A., Khaw, K.T., Luben, R.N., Mangino, M., Marchini, J., McArdle, W.L., McGinnis, R., Meyre, D., Munroe, P.B., Morris, A.D., Ness, A.R., Neville, M.J., Nica, A.C., Ong, K.K., O'Rahilly, S., Owen, K.R., Palmer, C.N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Nurses' Health Study, Randall, J.C., Rayner, N.W., Ring, S.M., Sandhu, M.S., Scherag, A., Sims, M.A., Song, K., Soranzo, N., Speliotes, E.K., Diabetes Genetics Initiative, Syddall, H.E., Teichmann, S.A., Timpson, N.J., Tobias, J.H., Uda, M., SardiNIA Study, Vogel, C.I., Wallace, C., Waterworth, D.M., Weedon, M.N., Wellcome Trust Case Control Consortium, Willer, C.J., FUSION, Wraight, Yuan, X., Zeggini, E., Hirschhorn, J.N., Strachan, D.P., Ouwehand, W.H., Caulfield, M.J., Samani, N.J., Frayling, T.M., Vollenweider, P., Waeber, G., Mooser, V., Deloukas, P., McCarthy, M.I., Wareham, N.J., Barroso, I., Jacobs, K.B., Chanock, S.J., Hayes, R.B., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H.E., Kraft, P., Hankinson, S.E., Hunter, D.J., Hu, F.B., Lyon, H.N., Voight, B.F., Ridderstrale, M., Groop, L., Scheet, P., Sanna, S., Abecasis, G.R., Albai, G., Nagaraja, R., Schlessinger, D., Jackson, A.U., Tuomilehto, J., Collins, F.S., Boehnke, M., and Mohlke, K.L. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.* 40, 768-775.
- Lowe, J.K., Maller, J.B., Pe'er, I., Neale, B.M., Salit, J., Kenny, E.E., Shea, J.L., Burkhardt, R., Smith, J.G., Ji, W., Noel, M., Foo, J.N., Blundell, M.L., Skilling, V., Garcia, L., Sullivan, M.L., Lee, H.E., Labek, A., Ferdowsian, H., Auerbach, S.B., Lifton, R.P., Newton-Cheh, C., Breslow, J.L., Stoffel, M., Daly, M.J., Altshuler, D.M., and Friedman, J.M. (2009). Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PLoS Genet.* 5, e1000365.
- Marsh, M. (2001). Endocytosis. Oxford university press. P.vii.
- Meyre, D., Delplanque, J., Chèvre, J.C., Lecoœur, C., Lobbens, S., Gallina, S., Durand, E., Vatin, V., Degraeve, F., Proença, C., Gaget, S., Körner, A., Kovacs, P., Kiess, W., Tichet, J., Marre, M., Hartikainen, A.L., Horber, F., Potoczna, N., Hercberg, S., Levy-Marchal, C., Pattou, F., Heude, B., Tauber, M., McCarthy, M.I., Blakemore, A.J., Montpetit, A., Polychronakos, C., Weill, J., Coin, L.J., Asher, J., Elliott, P., Jarvelin, M.R., Visvikis-Siest, S., Balkau, B., Sladek, R., Balding, D., Walley, A., Dina, C., and Froguel, P. (2009). Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* 41, 157-159.
- Miaczynska, M. and Stenmark, H. (2008). Mechanisms and functions of endocytosis. *J. Cell Biol.* 180, 7-11.
- Must, A., Spadano, J., Coakley, E.H., Field, A.E., Colditz, G., and Dietz, W.H. (1999). The disease burden associated with overweight and obesity. *JAMA* 282, 1523-1029.
- Rabbee, N. and Speed, T.A. (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* 22, 7-12.
- Sabatti, C., Service, S.K., Hartikainen, A.L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C.G., Zaitlen, N.A., Varilo, T., Kaakinen, M., Sovio, U., Ruukonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., McCarthy, M.I., Daly, M.J., Jarvelin, M.R., Freimer, N.B., and Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41, 35-46.
- Saxena, R., Voight, B.F., Lyssenko, V., Burt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J., Hughes, T.E., Groop, L., Altshuler, D., Almgren, P., Florez, J.C., Meyer, J., Ardlie, K., Bengtsson Boström, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H.N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Råstam, L., Speliotes, E.K., Taskinen, M.R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjögren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S.B., Chirn, G.W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316, 1331-1336.
- Scherag, A., Dina, C., Hinney, A., Vatin, V., Scherag, S., Vogel, C.I., Müller, T.D., Grallert, H., Wichmann, H.E., Balkau, B., Heude, B., Jarvelin, M.R., Hartikainen, A.L., Levy-Marchal, C., Weill, J., Delplanque, J., Körner, A., Kiess, W., Kovacs, P., Rayner, N.W., Prokopenko, I., McCarthy, M.I., Schäfer, H., Jarick, I., Boeing, H., Fisher, E., Reinehr, T., Heinrich, J., Rzehak, P., Berdel, D., Borte, M., Biebermann, H., Krude, H., Rosskopf, D., Rimbach, C., Rief, W., Fromme, T., Klingenspor, M., Schürmann, A., Schulz, N., Nöthen, M.M., Mühleisen, T.W., Erbel, R., Jöckel, K.H., Moebus, S., Boes, T., Illig, T., Froguel, P., Hebebrand, J., and Meyre, D. (2010). Two new loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. *PLoS One* 6, e1000916.
- Scuteri, A., Sanna, S., Chen, W.M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrù, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G.B., Fink, A.A., Weder, A.B., Cooper, R.S., Galan,

- P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E., and Abecasis, G.R. (2007). Genome-wide association scan shows genetic variants in the FTO Gene are associated with obesity-related traits. *PLoS Genet.* 3, 1200-1210.
- Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R., and Klein, B. (2007). Lasso-patternsearch algorithm with application ophthalmology and genomic data. *Stat. 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., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C., and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445, 881-885.
- Thorleifsson, G., Walters, G.B., Gudbjartsson, D.F., Steinthorsdottir, V., Sulem, P., Helgadóttir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., Jonsdottir, T., Olafsdottir, E.J., Olafsdottir, G.H., Jonsson, T., Jonsson, F., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Lauritzen, T., Aben, K.K., Verbeek, A.L., Roeleveld, N., Kampman, E., Yanek, L.R., Becker, L.C., Tryggvadottir, L., Rafnar, T., Becker, D.M., Gulcher, J., Kiemenev, L.A., Pedersen, O., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (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. R. Statist. Soc. B.* 58, 267-288.
- Wallace, C., Newhouse, S.J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R.J., Marçano, A.C., Hajat, C., Burton, P., Deloukas, P., Brown, M., Connell, J.M., Dominiczak, A., Lathrop, G.M., Webster, J., Farrall, M., Spector, T., Samani, N.J., Caulfield, M.J., and Munroe, P.B. (2008). Genome-wide association analysis study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am. J. Hum. Genet.* 82,139-149.
- Wang, W.Y., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet.* 6, 109-118.
- Weedon, M. N., Lango, H., Lindgren, C.M., Wallace, C., Evans, D.M., Mangino, M., Freathy, R.M., Perry, J.R., Stevens, S., Hall, A.S., Samani, N.J., Shields, B., Prokopenko, I., Farrall, M., Dominiczak, A., Diabetes Genetics Initiative, Wellcome Trust Case Control Consortium, Johnson, T., Bergmann, S., Beckmann, J.S., Vollenweider, P., Waterworth, D.M., Mooser, V., Palmer, C.N., Morris, A.D., Ouwehand, W.H., Cambridge GEM Consortium, Zhao, J.H., Li, S., Loos, R.J., Barroso, I., Deloukas, P., Sandhu, M.S., Wheeler, E., Soranzo, N., Inouye, M., Wareham, N.J., Caulfield, M., Munroe, P.B., Hattersley, A.T., McCarthy, M.I., and Frayling, T.M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* 40, 575-583.
- 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.
- Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C., Lettre, G., Lim, N., Lyon, H.N., McCarroll, S.A., Papadakis, K., Qi, L., Randall, J.C., Roccascocca, R.M., Sanna, S., Scheet, P., Weedon, M.N., Wheeler, E., Zhao, J.H., Jacobs, L.C., Prokopenko, I., Soranzo, N., Tanaka, T., Timpson, N.J., Almgren, P., Bennett, A., Bergman, R.N., Bingham, S.A., Bonnycastle, L.L., Brown, M., Burt, N.P., Chines, P., Coin, L., Collins, F.S., Connell, J.M., Cooper, C., Smith, G.D., Dennison, E.M., Deodhar, P., Elliott, P., Erdos, M.R., Estrada, K., Evans, D.M., Gianniny, L., Gieger, C., Gillson, C.J., Guiducci, C., Hackett, R., Hadley, D., Hall, A.S., Havulinna, A.S., Hebebrand, J., Hofman, A., Isomaa, B., Jacobs, K.B., Johnson, T., Jousilahti, P., Jovanovic, Z., Khaw, K.T., Kraft, P., Kuokkanen, M., Kuusisto, J., Laitinen, J., Lakatta, E.G., Luan, J., Luben, R.N., Mangino, M., McArdle, W.L., Meitinger, T., Mulas, A., Munroe, P.B., Narisu, N., Ness, A.R., Northstone, K., O'Rahilly, S., Purmann, C., Rees, M.G., Ridderstråle, M., Ring, S.M., Rivadeneira, F., Ruukonen, A., Sandhu, M.S., Saramies, J., Scott, L.J., Scuteri, A., Silander, K., Sims, M.A., Song, K., Stephens, J., Stevens, S., Stringham, H.M., Tung, Y.C., Valle, T.T., van Duijn, C.M., Vimalaswaran, K.S., Vollenweider, P., Waeber, G., Wallace, C., Watanabe, R.M., Waterworth, D.M., Watkins, N., Wellcome Trust Case Control Consortium, Witteman, J.C., Zeggini, E., Zhai, G., Zillikens, M.C., Altshuler, D., Caulfield, M.J., Chanock, S.J., Farooqi, I.S., Ferrucci, L., Guralnik, J.M., Hattersley, A.T., Hu, F.B., Jarvelin, M.R., Laakso, M., Mooser, V., Ong, K.K., Ouwehand, W.H., Salomaa, V., Samani, N.J., Spector, T.D., Tuomi, T., Tuomilehto, J., Uda, M., Uitterlinden, A.G., Wareham, N.J., Deloukas, P., Frayling, T.M., Groop, L.C., Hayes, R.B., Hunter, D.J., Mohlke, K.L., Peltonen, L., Schlessinger, D., Strachan, D.P., Wichmann, H.E., McCarthy, M.I., Boehnke, M., Barroso, I., Abecasis, G.R., Hirschhorn, J.N., and Genetic Investigation of ANthropometric Traits Consortium. (2008). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* 41, 25-34.
- Wu, T., Chen, Y., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25, 714.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J.R. Statist. Soc. B.* 67, 301-320.