G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(4):160-165
https://doi.org/10.5808/GI.2016.14.4.160

ORIGINAL ARTICLE

Efficient Strategy to Identify Gene-Gene Interactions and Its Application to Type 2 Diabetes

Donghe Li[1], Sungho Won[1,2]*

[1]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul 08826, Korea,
[2]Department of Public Health Science, Seoul National University, Seoul 08826, Korea

Over the past decade, the detection of gene-gene interactions has become more and more popular in the field of genome-wide association studies (GWASs). The goal of the GWAS is to identify genetic susceptibility to complex diseases by assaying and analyzing hundreds of thousands of single-nucleotide polymorphisms. However, such tests are computationally demanding and methodologically challenging. Recently, a simple but powerful method, named "BOolean Operation-based Screening and Testing" (BOOST), was proposed for genome-wide gene-gene interaction analyses. BOOST was designed with a Boolean representation of genotype data and is approximately equivalent to the log-linear model. It is extremely fast, and genome-wide gene-gene interaction analyses can be completed within a few hours. However, BOOST can not adjust for covariate effects, and its type-1 error control is not correct. Thus, we considered two-step approaches for gene-gene interaction analyses. First, we selected gene-gene interactions with BOOST and applied logistic regression with covariate adjustments to select gene-gene interactions. We applied the two-step approach to type 2 diabetes (T2D) in the Korea Association Resource (KARE) cohort and identified some promising pairs of single-nucleotide polymorphisms associated with T2D.

Keywords: epistasis, gene-gene interaction, genome-wide association study, type 2 diabetes mellitus

Introduction

The concept of epistasis, generally defined as interactions among different genes, was first introduced in 1909 by William Bateson to describe the latent effect of one locus over another locus. A quantitative definition to the interaction was proposed in 1918 by R.A. Fisher as a statistical deviation from the additive effects of two loci on a phenotype. This definition enabled interaction analyses by testing whether products of multiple genotypes are statistically associated with phenotypes. More definitions about the gene-gene interaction have been proposed, but some are still not clearly understood. The statistical gene-gene interaction has often been confused with a biological gene-gene interaction. Particularly, the inference on a biological mechanism is complicated because of the lack of direct correspondence between statistical and biological interactions [1]. In general, statisticians define a statistical interaction as a departure from additivity in a linear model using a selected measurement scale [2]. However, as was pointed by Wang *et al*. [3], if one aims to infer biological interactions, statistically modeled interactions and main effect terms should not be interpreted separately [2]. In this paper, we detected gene-gene interactions with a likelihood ratio test. Genotype scores for single-nucleotide polymorphism (SNP) pairs were considered nominal variables, and nine different levels were assumed for a full model. For a reduced model, we considered three levels for each SNP, and thus, our likelihood ratio tests followed a chi-square distribution with 4 degrees of freedom. Therefore, the proposed method can detect biological interactions. It should be noted that biological interactions include statistical interactions.

The method of detecting gene-gene interactions has attracted much attention in genome-wide association studies (GWASs). Including logistic regression analysis for detecting gene-gene interactions, new methods, like comparing linkage disequilibrium (LD) in case and control

Received August 29, 2016; Revised November 8, 2016; Accepted November 20, 2016
***Corresponding author:** Tel: +82-2-880-2714, Fax: +82-303-0942-2714, E-mail: won1@snu.ac.kr

groups, have been recently proposed [4]. However, analyzing a large number of SNPs in a GWAS is computationally very intensive, and various approaches, such as MDR [5], BEAM [6], Random Jungle [7], PLINK [8], and BOolean Operation-based Screening and Testing (BOOST), have been proposed to enable gene-gene interactions on a genome-wide scale.

In this paper, we considered BOOST method, proposed by Wan *et al*. [9]. BOOST uses a non-iteratively estimated measure that is approximately equal to the maximum likelihood estimators for a log-linear model, and it is used to select pairs of SNPs with a specified threshold. There is an updated method, graphical processing units BOOST (GBOOST), which is a BOOST method implemented for a graphical processing units framework for enabling parallel computing to achieve massive assignment in a fast manner [10]. GBOOST achieves a 40-fold speedup compared with BOOST. However, in the algorithm of BOOST, covariates other than SNPs can not be considered. We need a more flexible approach to improve the power of the model.

Here, we propose an efficient strategy that combines the BOOST screening stage and logistic regression method. Logistic regression generally shows good statistical power for a wide spectrum of epistasis. Screening with BOOST is a computationally efficient screening method, and a genome-wide search can be completed within a few hours. A follow-up stage of logistic regression with covariates would improve the statistical power of the model. In this paper, we first review the BOOST method and apply the proposed two-stage approach to type 2 diabetes (T2D) in a Korean population. This analysis of gene-gene interactions on a genome-wide scale with BOOST was completed within 42 hours, and we also identified several pairs of SNPs associated with T2D.

T2D is the most common form of diabetes, and unlike people with type 1 diabetes, T2D patients make insulin. However, either their pancreas does not make enough insulin or the body cannot use the insulin well enough. The prevalence of T2D has increased rapidly in recent years. The prevalence of T2D in Korea was estimated to be 7.3% (in people over 20 years of age) in 2005, and the rate of patients with T2D is expected to increase dramatically from 7.08% in 2010 to about 10.85% by 2030 [11]. Even more and more children are being diagnosed with T2D. Environmental effects, like obesity and lack of physical activity, are two of the most common causes of T2D, and the increasing prevalence may be related with them. It is also known that T2D has considerable heritability, which indicates a genetic effect. Until now, genetic variants in nearly 70 loci have been identified for T2D. Some variants from genes, such as KCNQ1 [12], KCNJ11, PPARG, NRF1, IDE, TCF7L2, CDKAL1, HHEX, IGF2BP2, CDKN2A/B, and SLC30A8 [13], were reported as significant susceptible genes to T2D in the East Asian population. To detect the relation between gene-gene interactions and T2D phenotype, we performed a gene-gene interaction analysis with the proposed two-stage analysis. We analyzed 8,842 participants (4,183 males and 4,659 females) with 352,228 autosomal SNPs collected from the Korea Association Resource project (KARE). We found some promising gene-gene interactions with the proposed method, and they will be investigated further in our future follow-up studies.

Methods

The KARE cohort

The KARE project started in 2007 and recruited 10,000 participants aged between 49 to 60 years from Ansung and Ansan, in the Gyeonggi Province of South Korea. About 50 million autosomal SNPs were genotyped with the Affymetrix Genome-Wide Human SNP arrary 5.0 [14]. In total, 8,842 individuals with 352,228 SNPs are available. In our GWASs, we discarded SNPs for which the Hardy-Weinberg equilibrium p-values were less than 10^{-5}, the genotype calling rates were less than 95%, and the minor allele frequencies were less than 0.05. We also eliminated subjects with gender inconsistencies, those whose identity by state was more than 0.8, and those whose calling rates were less than 95%. As a result, we analyzed 8,773 participants (4,117 males and 4,656 females) with 304,245 SNPs.

Definition of T2D

An individual was coded as a T2D patient if the condition satisfied the World Health Organization (WHO) diabetes diagnostic criteria: fasting plasma glucose (glu0) $\geq$ 126 mg/dL, plasma glucose (glu120) $\geq$ 200 mg/dL 2 h after an oral dose, or glycated hemoglobin (HbA1c) $\geq$ 6.5%. A total of 1,169 subjects were diagnosed as cases, and the other individuals were considered controls.

Notations

We assume that there are L SNPs and n subjects. Genotypes at SNP l are denoted by X_l, where $l = 1, \cdots, L$, and Y indicates the disease status: 1 for case and 2 for control. We assume that SNPs are bi-allelic, and capital and lowercase letters always indicate the major and minor alleles, respectively. For instance, AA indicates a homozygous reference genotype, Aa indicates the heterozygous genotype, and aa indicates the homozygous variant genotype. For simplicity, we denote the homozygous reference genotype, heterozygous genotype, and homozygous variant genotype as 1, 2, and 3, respectively.

Review of BOOST

The logistic regression model with only a main effect for two SNPs, p and q, can be modeled by the following form:

$$\log\frac{P(Y=1|X_p=i,X_p=j)}{P(Y=2|X_p=i,X_p=j)}=\beta_0+\beta_i^{x_p}+\beta_j^{x_q},$$

and denote its log-likelihood as L_M. The logistic regression model with main effects and interaction terms is

$$\log\frac{P(Y=1|X_p=i,X_q=j)}{P(Y=2|X_p=i,X_q=j)}=\beta_0+\beta_i^{X_p}+\beta_j^{x_q}+\beta_{ij}^{x_p x_q}.$$

We denote it log-likelihood value by L_F. Then, the interaction effects can be detected by the difference of the maximum log-likelihoods (MLEs) of these two models—i.e., $\hat{L}_F-\hat{L}_M$.

However, the difference in the log-likelihood needs very intensive computation for hundreds of billions of pairs of SNPs. Alternatively, there exists a one-to-one correspondence between a logistic regression model and a log-linear model in categorical data analysis [15], and BOOST considers interaction models based on log-linear models [9].

On the basis of the equivalence between the log-linear model and its corresponding logistic regression model, BOOST constructed its test statistic using the homogeneous association model M_H and saturated model M_S and denotes their log-likelihood as L_H and L_S, respectively. Then, we denote the observed genotype count of disease status k with $X_p = i$ and $X_q = j$ by n_{ijk} and the expected genotype count by μ_{ijk}, where $k = 1$ or 2, $i = 1$, 2, or 3, and $j = 1$, 2, or 3. Then, the maximum log-likelihood of both models will be

$$\hat{L}_S=\sum_{i,j,k}[n_{ijk}\log(n_{ijk})-n_{ijk}-\log(n_{ijk}!)].$$

If we let $\hat{\mu}_{ijk}^H$ be the MLE of μ_{ijk} for M_H, we have

$$\hat{L}_H=L_H(\hat{\mu}_{ijk}^H)=\max_{\mu_{ijk}}\sum_{i,j,k}[n_{ijk}\log(\mu_{ijk})-\mu_{ijk}-\log(n_{ijk}!).$$

The interaction effects based on the likelihood ratio test can be calculated by the following forms:

$$\hat{L}_S-\hat{L}_H=\sum_{i,j,k}[n_{ijk}\log\frac{n_{ijk}}{\hat{\mu}_{ijk}^H}-n_{ijk}+\hat{\mu}_{ijk}^H].$$

If we let n be the total sample size, $\hat{\pi}_{ijk}=\frac{n_{ijk}}{n}$ and $\hat{P}_{ijk}=\frac{\hat{\mu}_{ijk}^H}{n}$, it can be further simplified as

$$n\sum_{i,j,k}[\hat{\pi}_{ijk}\log\frac{\hat{\pi}_{ijk}}{\hat{P}_{ijk}}],$$

and it can also be denoted by Kullback-Leibler's [9] form as

$$n\bullet D_{KL}(\hat{\pi}_{ijk}\parallel\hat{P}_{ijk}).$$

This provides another interpretation of interactions, in that the difference of two log-likelihoods is proportional to the Kullback-Leibler divergence of the joint distribution obtained under the saturated model M_S and the distribution obtained under homogeneous association model M_H.

In particular, there is no closed form solution for homogenous association model M_H, and iterative methods are needed to calculate the likelihood ratio tests. Likelihood ratio tests are computationally intensive when facing hundreds of billions of SNP pairs in the interaction analysis.

To address this issue, BOOST uses Kirkwood superposition approximation (KSA) [16] to estimate under M_H as

$$\hat{p}_{ijk}^K=\frac{1}{\eta}\frac{\pi_{ij.}\pi_{i.k}\pi_{.jk}}{\pi_{i..}\pi_{.j.}\pi_{..k}},\text{ where }\eta=\sum_{i,j,k}\frac{\pi_{ij.}\pi_{i.k}\pi_{.jk}}{\pi_{i..}\pi_{.j.}\pi_{..k}}.$$

Therefore, it can be utilized to approximate L_H. If we let L_{KSA} be the likelihood based on KSA, then we get following equation:

$$2(\hat{L}_S-\hat{L}_{KSA})=2n\bullet D_{KL}(\hat{\pi}_{ijk}\parallel\hat{p}_{ijk}^K),$$

and it can be calculated easily on the basis of the contingency table. Through simulation analysis by Wan *et al.* [9], $2(\hat{L}_S-\hat{L}_{KSA})$ is known to be an upper bound of $2(\hat{L}_S-\hat{L}_H)$, and they are almost identical if they are larger than 25. The default value of the threshold τ in BOOST is 30, and its corresponding p-value is 4.89×10^{-6} [9].

In the screening stage of BOOST, all pairwise interactions will be evaluated by using KSA. If $2(\hat{L}_S-\hat{L}_{KSA})>\tau$, the interaction will be considered in the next testing stage; otherwise, it will be discarded from the analyses. All of the non-significant SNP pairs would be filtered out in this manner in the screening stage.

Two-stage gene-gene interaction analyses

We first apply the BOOST approach to select the promising pairs of SNPs with the a priori chosen τ. τ can be selected based on the available computing facility, and we set $\tau = 30$ for our analyses. After filtering SNP pairs from the

first stage with BOOST, we apply the logistic regression. BOOST can not adjust for the effects of covariates, and we applied logistic regression analysis with adjustments for sex, age, body mass index (BMI), and the top 10 principal component (PC) scores to the selected pairs of SNPs with BOOST. The logistic regression analysis was performed by using the glm function in R software. To calculate the p-values of the interaction term, we used the ANOVA function by comparing two fitted models in R.

Results

We have carried out an interaction analysis on T2D in the KARE cohort on the genome-wide scale; 8,773 subjects with 304,245 SNPs were considered for detecting gene-gene interactions. Missing genotypes were imputed with Impute2 [17] software.

We applied the proposed two-stage approach to identify the genome-wide significant gene-gene interactions. The analyses were completed within 42 h with an Intel Core i3-4130 CPU 3.40 GHz desktop. A total of 46,282,357,890 interactions were executed, and the Bonferroni-adjusted 0.05 genome-wide significance level is 1.08e-12. Promising pairs of SNPs were selected with BOOST, and 229,965 pairs of SNPs were selected with BOOST; then 229,965 pairs of SNPs were analyzed with logistic regression.

To perform adjustments for the population substructure between individuals, we used the EIGENSTRAT [18] method. EIGENSTRAT calculates the genetic similarities among subjects by using a genetic relationship matrix and applies PC analysis. The generated PC scores are then utilized as covariates for genetic association analyses, and this approach guarantees robustness against a population substructure. Here, we calculated the first 10 PC scores, and they were

Table 1. Results for top 10 highest interaction p-values between two SNPs in the KARE dataset

SNP 1	Gene 2	CHR 1	Position 1	Minor allele 1	Major allele 1	SNP 2	Gene 2	CHR 2	Position 2	Minor allele 2	Major allele 2	Interaction logistic R	p-value
rs1402142	-	4	64970948	C	A	rs8012584	-	14	38826243	A	G	55.09085092	3.11E-11
rs1402142	-	4	64970948	C	A	rs2183235	-	14	38828344	G	C	54.914105	3.39E-11
rs1402142	-	4	64970948	C	A	rs980010	-	14	38822190	A	G	54.89919559	3.41E-11
rs1402142	-	4	64970948	C	A	rs1958459	-	14	38811157	G	A	54.78768732	3.60E-11
rs7652843	-	3	194554885	A	C	rs224110	-	10	64551577	A	T	54.43510545	4.27E-11
rs1402142	-	4	64970948	C	A	rs7145965	-	14	38804433	G	A	53.5692555	6.48E-11
rs1402142	-	4	64970948	C	A	rs1475516	-	14	38799740	G	A	53.17278157	7.84E-11
rs1463367	-	4	48968037	T	C	rs10899912	-	10	44296893	G	A	51.89018414	1.45E-10
rs872234	*BTBD9*	6	38289804	C	T	rs10816769	*TMEM245*	9	111857440	C	G	51.60050802	1.67E-10
rs1864433	-	2	38007984	T	A	rs1110144	*CNTNAP2, MIR548T*	7	148001291	A	G	51.3563867	1.88E-10

SNP, single-nucleotide polymorphism; KARE, Korea Association Resource; CHR, chromosome.

Table 2. The association genes of the SNPs of the top 10 highest p-value interaction pairs

SNP	Gene	Associated genes
rs1402142	-	*HTRA3, AREG, TEC, NRAS*
rs1463367	-	*HTRA3, AREG, TEC, NRAS*
rs872234	*BTBD9*	*BTBD9, DHFRP2, SEMA3D, PHTF2PKD1L1, C7orf44, CAP2, NRAS, CALN1*
rs1864433	-	*FAM82A, CLIP4,VIT, AFF3, TPO*
rs8012584	-	*ANG, ABCB1*
rs2183235	-	*ANG, ABCB1*
rs980010	-	*ANG, ABCB1*
rs1958459	-	*ANG, ABCB1*
rs224110	-	*ANK3, ZNF32, RRET*
rs7145965	-	*ANG, ABCB1*
rs1475516	-	*ANG, ABCB1*
rs10899912	-	*ZNF32, RET*
rs10816769	*TMEM245*	*C9orf5*
rs1110144	*CNTNAP2*	*CNTNAP2, MLL3*

SNP, single-nucleotide polymorphism.

included as covariates for the logistic method in R. Sex, age, and BMI were also included as covariates in the analysis.

Most significant results for interaction analyses with KARE datasets are listed in Table 1. Only the top 10 significant SNP pairs are listed. The most significant interaction effect with a p-value of 3.11E-11 was found for rs1402142 and rs8012584. The former is associated with the genes *HTRA3, AREG, TEC*, and *NRAS* (Table 2), which are related to metabolic, immune, and hematological diseases, and the latter is associated with the genes *ANG* and *ABCB1*, related to neurological and metabolic diseases. Our results show that even the interactions are significant, but their marginal effects may not be. There are some interesting results that rs872234 is near *DHFRP2*, which is related to diabetes mellitus type 1 [19], and rs1110144 and rs1104853 are both in *CNTNAP2* [20] and *MIR548T*, respectively; they are also known to be associated with diabetes.

Discussion

The analysis of gene-gene interactions on a genome-wide scale is computationally very intensive, and many computational and statistical approaches have been recently proposed to minimize the computational burden. We found that BOOST is highly computationally efficient and can filter out non-significant interaction pairs in a fast manner.

In this study, we proposed an efficient strategy to identify interactions in genome-wide SNP data. We first utilized the screening stage of BOOST to filter out non-significant pairs and then used logistic regression with several covariates, such as age, sex, BMI, and PC scores.

In real data analysis, we used the KARE cohort dataset to detect gene-gene interactions of T2D. The smallest p-value (3.11E-11) of interaction pairs in the KARE data was found for rs1402142 and rs8012584. The Bonferroni-adjusted genome-wide significance level is 1.08e-12, and this SNP pair is not significant genome-wide. This insignificance is partially attributable to the insufficient sample size. With advances in genotyping/sequencing technology, genotyping costs will be much lower, and therefore, in the near future, sufficiently large samples will become available for gene-gene interaction analyses, which may lead us to a better understanding of human diseases.

Acknowledgments

Data for this study was provided with biospecimens from National Biobank of Korea, the Centers for Disease Control and Prevention, Republic of Korea (4845-301, 4845-302 and 307), and this work was supported by Research Resettlement Fund for the new faculty of Seoul National University.

References

1. Ueki M, Cordell HJ. Improved statistics for genome-wide interaction analysis. *PLoS Genet* 2012;8:e1002625.
2. Won S, Kwon MS, Mattheisen M, Park S, Park C, Kihara D, *et al*. Efficient strategy for detecting gene x gene joint action and its application in schizophrenia. *Genet Epidemiol* 2014;38:60-71.
3. Wang X, Elston RC, Zhu X. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nat Rev Genet* 2011;12:74.
4. Hu JK, Wang X, Wang P. Testing gene-gene interactions in genome wide association studies. *Genet Epidemiol* 2014;38:123-134.
5. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138-147.
6. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* 2007;39:1167-1173.
7. Schwarz DF, König IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* 2010;26:1752-1758.
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559-575.
9. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, *et al*. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* 2010;87:325-340.
10. Yung LS, Yang C, Wan X, Yu W. GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics* 2011;27:1309-1310.
11. Ko SH, Kim SR, Kim DJ, Oh SJ, Lee HJ, Shim KH, *et al*. 2011 clinical practice guidelines for type 2 diabetes in Korea. *Diabetes Metab J* 2011;35:431-436.
12. Park SE, Lee WY, Oh KW, Baek KH, Yoon KH, Kang MI, *et al*. Impact of common type 2 diabetes risk gene variants on future type 2 diabetes in the non-diabetic population in Korea. *J Hum Genet* 2012;57:265-268.
13. Park KS. The search for genetic risk factors of type 2 diabetes mellitus. *Diabetes Metab J* 2011;35:12-22.
14. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41:527-534.
15. Agresti A. *Categorical Data Analysis*. 2nd ed. New York: Wiley-Interscience, 2002.
16. Matsuda H. Physical nature of higher-order mutual information: intrinsic correlations and frustration. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 2000;62(3 Pt A):3096-3102.
17. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* 2011;1:457-470.
18. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904-909.

19. Ross KA. Evidence for somatic gene conversion and deletion in bipolar disorder, Crohn's disease, coronary artery disease, hypertension, rheumatoid arthritis, type-1 diabetes, and type-2 diabetes. *BMC Med* 2011;9:12.

20. Giri A, Sanders M, Velez Edwards D, Ikizler T, Roden D, Birdwell K. A genome wide association study of new onset diabetes after transplant in kidney transplantation. *Am J Transplant* 2016;16(Suppl 3):B235.