G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(4):181-186
https://doi.org/10.5808/GI.2016.14.4.181

ORIGINAL ARTICLE

Genetic Association Analysis of Fasting and 1- and 2-Hour Glucose Tolerance Test Data Using a Generalized Index of Dissimilarity Measure for the Korean Population

Jaeyong Yee[1], Yongkang Kim[2], Taesung Park[2], Mira Park[3]*

[1]Department of Physiology and Biophysics, Eulji University, Daejeon 35233, Korea,
[2]Department of Statistics, Seoul National University, Seoul 08826, Korea,
[3]Department of Preventive Medicine, Eulji University, Daejeon 34824, Korea

Glucose tolerance tests have been devised to determine the speed of blood glucose clearance. Diabetes is often tested with the standard oral glucose tolerance test (OGTT), along with fasting glucose level. However, no single test may be sufficient for the diagnosis, and the World Health Organization (WHO)/International Diabetes Federation (IDF) has suggested composite criteria. Accordingly, a single multi-class trait was constructed with three of the fasting phenotypes and 1- and 2-hour OGTT phenotypes from the Korean Association Resource (KARE) project, and the genetic association was investigated. All of the 18 possible combinations made out of the 3 sets of classification for the individual phenotypes were taken into our analysis. These were possible due to a method that was recently developed by us for estimating genomic associations using a generalized index of dissimilarity. Eight single-nucleotide polymorphisms (SNPs) that were found to have the strongest main effect are reported with the corresponding genes. Four of them conform to previous reports, located in the *CDKAL1* gene, while the other 4 SNPs are new findings. Two-order interacting SNP pairs of are also presented. One pair (rs2328549 and rs6486740) has a prominent association, where the two single-nucleotide polymorphism locations are *CDKAL1* and *GLT1D1*. The latter has not been found to have a strong main effect. New findings may result from the proper construction and analysis of a composite trait.

Keywords: gene-gene interaction, genome wide association, glucose tolerance test

Introduction

Genome-wide association studies have been aiming to find the association between a single-nucleotide polymorphism (SNP) and complex traits. It started as a single-locus approach that tested a single SNP at a time and selected the top SNPs. However, it has become clearer that most complex diseases are associated with multiple genes and their interactions. Therefore, a multi-locus approach is now regarded as a necessity [1, 2]. Multifactor dimensionality reduction is one of the widely accepted methods for addressing this issue [3]. Meanwhile, it has been customary to categorize a phenotype into a binary trait to divide the observed outcomes into either a case or control (affected-unaffected). However, some diseases, such as obesity, are to be classified with several levels of affectedness [4]. In that case, it would be appropriate to perform a multi-class phenotype analysis. A disease, such as hypertension, may not be fully characterized by a single phenotype but by 2 or more observed phenotypes [5]. Then, a composite phenotype should be constructed and analyzed with genomic data to estimate the genome-wide association of the particular disease. Because the individual observables should have 2 or more classes, a composite phenotype would be a multi-class trait, with an indefinite number of classes. Combining the 2 aspects above, a genomic association study needs a method that is able to take gene-gene interactions and multi-class

Received October 21, 2016; Accepted November 16, 2016

*Corresponding author: Tel: +82-42-259-1615, Fax: +82-42-259-1689, E-mail: mira@eulji.ac.kr

and multivariate traits into consideration simultaneously.

In this paper, we will apply a recently introduced generalized index of dissimilarity (GIDS) [5] to demonstrate a multivariate multi-class trait genomic association study with diabetes. A composite phenotype with the complete set of classes, which was constructed with 3 individual phenotypes obtained by measuring the blood glucose levels using different protocols [6, 7], was analyzed with genomic data to estimate the genome-wide association. Gene-gene interactions and the main effect were investigated. Among the strongly associated SNPs, some were found to be consistent with previous results, while the rest of them were new findings. SNP pairs that were found to have a strong association are also presented.

Methods

Index of dissimilarity as a measure of association

Introduced originally as a segregation measure [8], the index of dissimilarity (IDS) has been shown to be effective in measuring genomic associations, as well [5]. A generalized version of the IDS (GIDS) has been reported to be able to yield the association strength with a phenotype having an arbitrary number of classes [5]. Fig. 1 has the schematics for the association of 2-order gene-gene interactions with a *J*-class phenotype. GIDS is defined as below following the notation in Fig. 1.

$$GIDS = \frac{1}{2}\frac{\sum_{j=1}^{J}\sum_{i}|n_{ij}-E_{ij}|}{\sum_{j=1}^{J} n \cdot P_{\cdot j} \cdot (1-P_{\cdot j})} \text{ where}$$

$$n=\sum_{j=1}^{J} n_{\cdot j},\ E_{ij}=\frac{n_{j\cdot} \cdot n_{\cdot j}}{n},\ P_{\cdot j}=\frac{n_{\cdot j}}{n} \quad (1)$$

Indices *i* and *j* represent the i^{th} multi-locus genotype and the j^{th} multi-class phenotype, respectively. The numerator of this equation measures the extent of uneven distributions by each phenotype class, where the denominator indicates the maximum possible unevenness [8]. To visualize this concept, let us reduce GIDS to a binary—i.e., $J = 2$, class—as shown in Eq. (2).

$$GIDS(J=2)=\frac{1}{2}\frac{\sum_{i}\left|n_{i1}-\frac{(n_{i1}+n_{i2})n_{\cdot 1}}{n}\right|+\sum_{i}\left|n_{i2}-\frac{(n_{i1}+n_{i2})n_{\cdot 2}}{n}\right|}{n_{\cdot 1}\left(1-\frac{n_{\cdot 1}}{n}\right)+n_{\cdot 2}\left(1-\frac{n_{\cdot 2}}{n}\right)}$$

$$=\frac{1}{2}\sum_{i}\left|\frac{n_{i1}}{n_{\cdot 1}}-\frac{n_{i2}}{n_{\cdot 2}}\right| \quad (2)$$

When there is little association between the genotype and phenotype, there would be a minimal difference between the 2 fractional terms for each *i*. Maximum association occurs when the differences for each *i* in this equation add up to the maximum.

Estimation of the association strength

To estimate the association strength, GIDS is calculated using Eq. (1). The number of samples, n_{ij}, that goe into the formula for GIDS is counted using a 2-way contingency table, constructed for the *k*-locus interactive genotype and *J*-class phenotype, as diagrammed in Fig. 1. The number of elements of this table will be $3^k \times J$. GIDS would be calculated exhaustively for every SNP or combination of SNPs associated with a multi-class categorical phenotype. Any order of a gene-gene interaction associated with a phenotype of an arbitrary number of categories can be estimated in theory, limited only by computing time. GIDS spans from 0 to 1, representing null and maximum association at either of the extremes. p-values that account for multiple comparisons can be obtained by constructing a null distribution common to all of the GIDS values [9]. Permutation of the dataset is performed to provide a non-associated dataset. A single GIDS that has the maximum value among all GIDS values obtained with a permuted dataset is collected. Repeated permutation and collection make the null distribution of GIDS. Now, a p-value should be the probability that the null distribution exceeds a particular GIDS value. Using this null

	Phenotypic outcome classes (j=1,...,J)				
Genotypic multifactor classes (i=1,...,9)	1	2	3	...	J
AABB	n_{11}	n_{12}	n_{13}	...	n_{1J}
AABb	n_{21}				n_{2J}
⋮	⋮		...		⋮
aaBb	n_{81}				n_{8J}
aabb	n_{91}	n_{92}	n_{93}	...	n_{9J}

Fig. 1. Schematics for the 2-order gene-gene interaction and multi-class phenotype. The number of samples, n_{ij}, for a genotypic class *i* and phenotypic class *j*. When $J = 2$, it represents the common binary, or case-control trait.

Table 1. OGTT criteria [6, 7] with KARE [13] result

Blood glucose (mg/dL)		Diagnosis	Proportion (%) (n = 8,371)	OGTT-1h normal[a] (%)
Fasting	OGTT-2h			
<110	<140	Normal	71.28	89.48
<110	≥140 and ≤200	IGT	20.26	52.12
<110	>200	Diabetes	2.99	6.80
≥110 and ≤125	<140	IFG	0.49	46.34
≥110 and ≤125	≥140 and ≤200	IGT	0.54	8.89
≥110 and ≤125	>200	Diabetes	1.74	0.00
>125	<140	Diabetes	0.29	58.33
>125	≥140 and ≤200	Diabetes	0.13	72.73
>125	>200	Diabetes	2.28	1.05

OGTT, oral glucose tolerance test; KARE, Korean Association Resource; OGTT-1h, 1-hour OGTT; OGTT-2h, 2-hour OGTT; IGT, impaired glucose tolerance; IFG, impaired fasting glucose.
[a]OGTT-1h criterion for normal: <180 mg/dL.

distribution, GIDS can be standardized as follows, where $\overline{\text{GIDS}}$ and S_{GIDS} represent the mean and standard deviation of the null distribution, respectively.

$$s\,GIDS = \frac{GIDS - \overline{\text{GIDS}}}{S_{GIDS}} \quad (3)$$

Association strengths from different orders of gene-gene interactions may be compared using the standardized GIDS (sGIDS), defined above [10].

Analysis

Construction of multi-class phenotype

Among the procedures to measure blood glucose levels, fasting and the oral glucose tolerance test (OGTT) are commonly conducted [11, 12]. Standard time intervals between the intake and measurement are 1 h and 2 h, denoted as OGTT-1h and OGTT-2h [7]. Diagnostic criteria are recommended as a combination of the measurements [6]. Listed in Table 1, the diagnosis is determined by considering fasting and OGTT-2h simultaneously [7]. Each of them was categorized into 3 distinct ranges, resulting in 9 composite ranges, which gave the criteria for the 4 diagnostic categories. Genomic data from the Korean Association Resource (KARE) project [13] were analyzed with the phenotypes of fasting and OGTT-2h and -1h. With the number of valid samples (n = 8371), the proportions that fell into each of the 9 composite categories are shown in the fourth column of Table 1. Note that neither of fasting and OGTT-2h can determine the diagnostic result independently of each other. An additional diagnostic criterion may be given by OGTT-1h [7]. The last column of Table 1 shows the percentage of the samples that were found to be normal in OGTT-1h within each of the 9 composite categories. A need for additional categorization seems to be apparent, because most of the percentages are away from 0% or 100%. To make such a multi-class phenotype, 3 of the individual KARE phenotypes—fasting, OGTT-2h, and OGTT-1h—were first categorized into 3 classes for the first 2 phenotypes and into 2 classes for the third one, following the reference criteria [6, 7]. A single composite phenotype of 3×3×2 classes was constructed as such. The resulting 18-class composite phenotype was analysed with the genotype part using GIDS to identify the most associated single- and two-locus models.

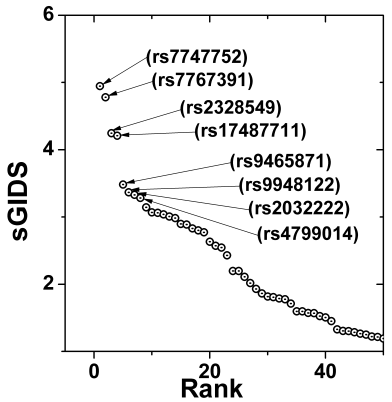

Fig. 2. Most strongly associated single-nucleotide polymorphisms identified by sGIDS. sGIDS reveals strongly associated single loci with fasting, OGTT-1h, and OGTT-2h blood glucose values categorized as a single multi-class phenotype. sGIDS, standardized generalized index of dissimilarity; OGTT-1h, 1-hour oral glucose tolerance test; OGTT-2h, 2-hour oral glucose tolerance test.

Single- and two-locus models

All of the available 327,872 SNPs in the KARE dataset were thoroughly examined for the association with the

Table 2. Top associated SNPs by sGIDS in the single-locus model

Single-locus model					
SNP	Chromosome	Gene	sGIDS	p-value	Previous report
rs7747752	6	*CDKAL1*	4.9451	0.0007	[14-16]
rs7767391	6	*CDKAL1*	4.7792	0.0008	[17]
rs2328549	6	*CDKAL1*	4.2449	0.0021	[18]
rs17487711	8	*LOC105379297*	4.2079	0.0021	
rs9465871	6	*CDKAL1*	3.4821	0.0060	[20-22]
rs9948122	18	*ATP9B*	3.3667	0.0068	
rs2032222	18	*ATP9B*	3.3274	0.0068	
rs4799014	18	*ATP9B*	3.2819	0.0072	

SNP, single-nucleotide polymorphism; sGIDS, standardized generalized index of dissimilarity.

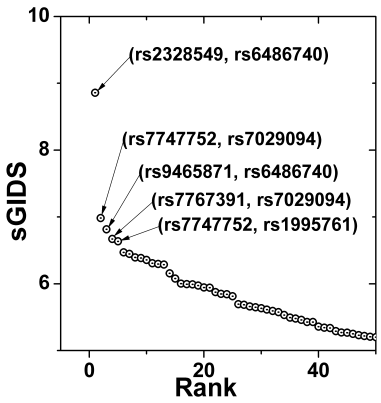

Fig. 3. Most strongly associated single-nucleotide polymorphism pairs identified by sGIDS. sGIDS reveals strongly associated 2-order interactions between 2 loci with fasting, OGTT-1h, and OGTT-2h blood glucose values categorized as a single multi-class phenotype. sGIDS, standardized generalized index of dissimilarity; OGTT-1h, 1-hour oral glucose tolerance test; OGTT-2h, 2-hour oral glucose tolerance test.

multi-class phenotype constructed above. Ranked by the sGIDS, the top 50 SNPs are plotted in Fig. 2, representing the main effects. Prominent SNPs may be seen in groups of 2, 2, and 4 showing the strongest association strengths. Those SNPs are listed in Table 2 with detailed information.

Using the calculations for the single-locus model, 1,000 SNPs were selected by sorting them with their respective GIDS values. Then, all of the possible pairs of those candidate SNPs were examined for their association strengths by evaluating sGIDS. The 50 most strongly associated pairs are plotted in Fig. 3, 5 of which are identified with their rs numbers.

Results

Two SNPs, rs7747752 and rs7767391, were found to have the strongest association. They are located in an intron of the cyclin-dependent-like kinase 5 (*CDK5*) regulatory subunit associated protein 1-like 1 (*CDKAL1*) gene located in chromosome 6. This gene has been reported to make contributions to type II diabetes. Both of the SNPs identified in this paper have been also reported to have susceptibility to type II diabetes [14-17]. Note that among the second-tier SNPs—rs2328549 and rs17487711—only the first one is in the intron of *CDKAL1* [18]. rs17487711 has been newly identified to have strong association comparable with that of rs2328549, while it is located in the intron of *LOC105379297* in chromosome 8. The National Center for Biotechnology Information (NCBI) still describes this gene as uncharacterized. In Table 2, 3 out of 4 among third-tier SNPS have been listed to be in the intron of ATPase, class II, type 9B (*ATP9B*) located in chromosome 18 [19], whereas rs9465871 can be found in the intron of *CDKAL1* [20-22].

A single prominent pair (rs2328549, rs6486740) revealed itself. In Table 3, detailed information is provided for the top 5 pairs. One of the SNPs in the pair that showed the strongest association, rs6486740, is located in an intron of the glycosyltransferase 1 domain-containing 1 (*GLT1D1*) gene on chromosome 12, while the other SNP, rs2328549, is located in an intron of *CDKAL1*, which was mentioned in the previous subsection. *GLT1D1* was reported to be related with renal sinus fat [23] and the transfers of glycosyl groups [24]. Among the second-tier SNP pairs, rs7029094 and rs1995761 are located on chromosome 9, but there exists little information about the gene in which they are located.

Discussion

A composite trait of 18 classes, constructed with 3 observables, each of which had 3 or 2 classes, was analyzed with genomic data to estimate the genome-wide association. Gene-gene interactions, as well as the main effect, were investigated. It may be essential to take a multi-class composite phenotype into consideration when performing a genomic association study for the susceptibility to a more

Table 3. Top associated interacting SNP pairs by sGIDS in the 2-locus model

Two-locus model						
SNP 1	SNP 2	Chromosomes	Gene 1	Gene 2	sGIDS	p-value
rs2328549	rs6486740	6, 12	*CDKAL1*	*GLT1D1*	8.8592	0.0001
rs7747752	rs7029094	6, 9	*CDKAL1*	-	6.9825	0.0002
rs9465871	rs6486740	6, 12	*CDKAL1*	*GLT1D1*	6.8159	0.0002
rs7767391	rs7029094	6, 9	*CDKAL1*	-	6.6694	0.0002
rs7747752	rs1995761	6, 9	*CDKAL1*	-	6.6355	0.0002

SNP, single-nucleotide polymorphism; sGIDS, standardized generalized index of dissimilarity.

complex disease, such as diabetes, as presented in this paper. OGTT-2h alone has 3 categories that exceed the commonly used dichotomous classes. However, the intrinsic need of a composite phenotype comes from the fact that diabetes may not be fully diagnosed with a single observable. If a disease may be diagnosed by multiple variables, it would be logical to analyze the susceptibility to it with multiple observables concurrently. A composite phenotype can be expected to have many classes, and the number of them could grow rapidly as the number of required variables increases. Therefore, it would be essential to have a methodology capable of analyzing the genomic association with a phenotype having an indefinite number of classes. GIDS has been demonstrated, in this paper, as a reliable candidate for this purpose. It showed consistency by identifying strongly associated SNPs, in chromosome 6, agreeing with previously reported ones in a single-locus model. The *CDKAL1* gene, in which they are located, has been found to be responsible for the SNPs that cause the susceptibility to type II diabetes. Moreover, we found new SNPs that have not been reported to have a strong association. They were found in chromosomes 8 and 18. The gene-gene interaction result detected a single prominent SNP pair that was noticeably stronger than others. Although 1 of the 2 SNPs in that pair, rs6486740 in the *GLT1D1* gene, was not found to have a strong main effect, it showed a very strong association when it interacted with the counterpart SNP, rs2328549.

In summary, we have confirmed previous results and at the same time found new strong genomic associations in both single- and 2-locus models by applying GIDS to a composite trait with 18 classes.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A2062848). It was also supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (No. 2016M3A9B694241).

References

1. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, *et al*. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010;11:446-450.
2. Evans DM, Marchini J, Morris AP, Cardon LR. Two-stage two-locus models in genome-wide association. *PLoS Genet* 2006;2:e157.
3. 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.
4. Kim K, Kwon MS, Oh S, Park T. Identification of multiple gene-gene interactions for ordinal phenotypes. *BMC Med Genomics* 2013;6 Suppl 2:S9.
5. Yee J, Kim Y, Park T, Park M. Using the generalized index of dissimilarity to detect gene-gene interactions in multi-class phenotypes. *PLoS One* 2016;11:e0158668.
6. World Health Organization. *Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia: Report of a WHO/IDF Consultation*. Geneva: World Health Organization, 2006.
7. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2005;28 Suppl 1:S37-S42.
8. Sakoda JM. A generalized index of dissimilarity. *Demography* 1981;18:245-250.
9. Jensen DD, Cohen PR. Multiple comparisons in induction algorithms. *Mach Learn* 2000;38:309-338.
10. Yee J, Kwon MS, Park T, Park M. A modified entropy-based approach for identifying gene-gene interactions in case-control study. *PLoS One* 2013;8:e69321.
11. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, *et al*. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* 2010;42:142-148.
12. Zheng C, Dalla Man C, Cobelli C, Groop L, Zhao H, Bale AE, *et al*. A common variant in the MTNR1b gene is associated with increased risk of impaired fasting glucose (IFG) in youth with obesity. *Obesity (Silver Spring)* 2015;23:1022-1029.
13. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, *et al*. A large-scale genome-wide association study of Asian pop-

ulations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41:527-534.
14. Chen P, Takeuchi F, Lee JY, Li H, Wu JY, Liang J, *et al*. Multiple nonglycemic genomic loci are newly associated with blood level of glycated hemoglobin in East Asians. *Diabetes* 2014; 63:2551-2562.
15. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, *et al*. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet* 2010;18:111-117.
16. Ryu J, Lee C. Association of glycosylated hemoglobin with the gene encoding *CDKAL1* in the Korean Association Resource (KARE) study. *Hum Mutat* 2012;33:655-659.
17. Quaranta M, Burden AD, Griffiths CE, Worthington J, Barker JN, Trembath RC, *et al*. Differential contribution of *CDKAL1* variants to psoriasis, Crohn's disease and type II diabetes. *Genes Immun* 2009;10:654-658.
18. Ng MC, Saxena R, Li J, Palmer ND, Dimitrov L, Xu J, *et al*. Transferability and fine mapping of type 2 diabetes loci in African Americans: the Candidate Gene Association Resource Plus Study. *Diabetes* 2013;62:965-976.
19. Takatsu H, Baba K, Shima T, Umino H, Kato U, Umeda M, *et al*. ATP9B, a P4-ATPase (a putative aminophospholipid translocase), localizes to the trans-Golgi network in a CDC50 protein-independent manner. *J Biol Chem* 2011;286:38159-38167.
20. Miyaki K, Oo T, Song Y, Lwin H, Tomita Y, Hoshino H, *et al*. Association of a cyclin-dependent kinase 5 regulatory subunit-associated protein 1-like 1 (CDKAL1) polymorphism with elevated hemoglobin A(1)(c) levels and the prevalence of metabolic syndrome in Japanese men: interaction with dietary energy intake. *Am J Epidemiol* 2010;172:985-991.
21. Ryoo H, Woo J, Kim Y, Lee C. Heterogeneity of genetic associations of *CDKAL1* and *HHEX* with susceptibility of type 2 diabetes mellitus by gender. *Eur J Hum Genet* 2011;19:672-675.
22. Wu Y, Li H, Loos RJ, Yu Z, Ye X, Chen L, *et al*. Common variants in *CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8*, and *HHEX/IDE* genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population. *Diabetes* 2008;57: 2834-2842.
23. Foster MC, Yang Q, Hwang SJ, Hoffmann U, Fox CS. Heritability and genome-wide association analysis of renal sinus fat accumulation in the Framingham Heart Study. *BMC Med Genet* 2011;12:148.
24. Dahlin A, Litonjua A, Irvin CG, Peters SP, Lima JJ, Kubo M, *et al*. Genome-wide association study of leukotriene modifier response in asthma. *Pharmacogenomics J* 2016;16:151-157.