Original Article
Genetics

Diabetes Metab J 2021;45:231-240
https://doi.org/10.4093/dmj.2019.0163
pISSN 2233-6079 · eISSN 2233-6087

dmj
DIABETES & METABOLISM JOURNAL

Exome Chip Analysis of 14,026 Koreans Reveals Known and Newly Discovered Genetic Loci Associated with Type 2 Diabetes Mellitus

Seong Beom Cho*, Jin Hwa Jang, Myung Guen Chung, Sang Cheol Kim

Division of Biomedical Informatics, Center for Genome Science, National Institute of Health, Korea Center for Disease Control and Prevention, Cheongju, Korea

Background: Most loci associated with type 2 diabetes mellitus (T2DM) discovered to date are within noncoding regions of unknown functional significance. By contrast, exonic regions have advantages for biological interpretation.
Methods: We analyzed the association of exome array data from 14,026 Koreans to identify susceptible exonic loci for T2DM. We used genotype information of 50,543 variants using the Illumina exome array platform.
Results: In total, 7 loci were significant with a Bonferroni adjusted $P=1.03\times10^{-6}$. rs2233580 in paired box gene 4 (*PAX4*) showed the highest odds ratio of 1.48 ($P=1.60\times10^{-10}$). rs11960799 in membrane associated ring-CH-type finger 3 (*MARCH3*) and rs75680863 in transcobalamin 2 (*TCN2*) were newly identified loci. When we built a model to predict the incidence of diabetes with the 7 loci and clinical variables, area under the curve (AUC) of the model improved significantly (AUC=0.72, $P<0.05$), but marginally in its magnitude, compared with the model using clinical variables (AUC=0.71, $P<0.05$). When we divided the entire population into three groups—normal body mass index (BMI; <25 kg/m^2), overweight (25≤ BMI <30 kg/m^2), and obese (BMI ≥30 kg/m^2) individuals—the predictive performance of the 7 loci was greatest in the group of obese individuals, where the net reclassification improvement was highly significant (0.51; $P=8.00\times10^{-5}$).
Conclusion: We found exonic loci having a susceptibility for T2DM. We found that such genetic information is advantageous for predicting T2DM in a subgroup of obese individuals.

Keywords: Diabetes mellitus, type 2; Exome; Genetic predisposition to disease; Microarray analysis

INTRODUCTION

Diabetes is an irreversible disease, identification of risk factors and disease prevention are therefore important for its management. Previous research has revealed risk factors for type 2 diabetes mellitus (T2DM) including body mass index (BMI), waist circumference, blood pressure, family history, and genetic background [1,2].

Many investigators have assumed that identification of genetic factors will improve the prediction of T2DM, so many genetic studies have been performed [3]. With the advent of high-throughput genotyping technologies, genome-wide association studies (GWASs) have become prevalent [4]. Several hundreds of thousands of tag single nucleotide polymorphisms (SNPs), which represent genotype variation of neighboring SNPs, are selected and probes for the tag SNPs are used in SNP microarray chips. Usually, more than 10,000 samples are used and imputed with a reference panel from a population sequencing project such as the 1000 Genome Project [5]. These GWASs have been actively applied to the study of T2DM, and

Corresponding author: Seong Beom Cho https://orcid.org/0000-0002-5734-1645
Division of Biomedical Informatics, Center for Genome Science, National Institute of Health, KCDC, 187 Osongsaengmyeong 2-ro, Osong-eup, Heungdeok-gu, Cheongju 28159, Korea
E-mail: sbcho1749@gachon.ac.kr

*Current affiliation: Department of Biomedical Informatics, Gachon University College of Medicine, 38 Dokjeom-ro 3beon-gil, Namdong-gu, Incheon 21565, Korea

Received: Aug. 31, 2019; Accepted: Feb. 10, 2020

meta-analysis of the individual GWASs identifies genome-wide loci associated significantly with the development of T2DM [3,6]. For example, Unoki et al. [7] reported that rs2283228, which is located in potassium voltage-gated channel subfamily Q member 1 (*KCNQ1*), shows genome-wide significance and the result was replicated in a different population [7]. In a meta-analysis of GWASs of European populations, 12 novel loci that are enriched in genes involved in cell cycle regulation have been identified [8]. A trans-ethnic meta-analysis of several consortia for a GWAS of diabetes revealed a directional consistency of the risk alleles across different ethnic groups and seven newly identified susceptible loci for T2DM [9]. While these studies used SNP microarray data, a sequencing-based GWAS and combination with the chip-based data were applied recently [10]. The results showed that variants associated with T2DM are located within the region that identified with previous GWASs.

Although the exome covers only about 1% of the entire human genome, it is easier to find functional implications for the significant loci because there is a lot of information about gene function. In genetic studies of T2DM and related traits, the exome-wide association studies have been performed with an exome array or exome sequencing to identify susceptible loci [10-14].

In the present research, using exome microarray data from 14,026 Koreans, we identified coding variants having susceptibility for T2DM. We also applied the genotype information of the variants into a model to predict T2DM and tested the performance of the model in the entire population and subgroups stratified by BMI.

METHODS

Study population

We used epidemiological data from the Ansan/Ansung (ASAS), Cardio-Vascular disease Association Study (CAVAS) and Health Examinee (HEXA) cohort, which are part of the Korean Genome and Epidemiology (KoGES) project [15]. The present study was approved by the Institutional Review Board of Korea Centers for Disease Control & Prevention (2017-03-10-1C-A) and written consent was obtained from all participants. The participants in the ASAS cohort underwent an oral glucose tolerance test (OGTT) using a solution of 75 g glucose. Using the OGTT results, we defined T2DM patients according to their past history of diabetes or American Diabetes Association criteria [16], which include (1) fasting glucose ≥126 mg/dL, (2) 2-hour glucose ≥200 mg/dL, and (3) glycosylated hemoglobin ≥6.5%. The OGTT had not been used in participants from the CAVAS and HEXA cohorts, and so the history and fasting glucose level were used to determine diabetes status in these cohorts.

Exome chip genotyping and quality control

We used the exome chip data that had been generated in a previous study [17]. Detailed genotyping and quality control processes were described previously. In brief, the Illumina HumanExome Chip v1.1 (Illumina Inc., San Diego, CA, USA), which includes probes for 242,901 variants, was used for genotyping. Illumina GenomeStudio version 2011.1 software clustered the probe signals to determine the genotype. After genotype calling, variant filtering was performed based on the following criteria: (1) completely missing in all participants, (2) monomorphic SNP, (3) Hardy-Weinberg equilibrium $P<1.0\times10^{-6}$, and (4) genotype call rate of <95%. Samples with cryptic relatedness and sex discrepancy were also excluded. After filtering, we imputed a missing genotype using the phasing function of the Beagle software [18].

Genome-wide association analysis

We used PLINK 1.9 for GWAS of the exome chip and meta-analysis [19]. The logistic function was applied with covariates of sex, age, and BMI. We used the adjust option to estimate λ for genomic control of GWAS statistics. For meta-analysis, we used the meta function of PLINK, and the result was filtered with Cochrane Q statistics and the I^2 index. Bonferroni's multiple testing correction was applied to select exome-wide significant loci. To search published GWAS results, we used the Korean Reference Genome Database (KRGDB) [20].

Prediction analysis and comparison of performance

Using follow-up information from the ASAS cohort data, we predicted T2DM using clinical and genetic variables. The glm function of the R statistical software package was used to develop the predictive model using follow-up information for T2DM status [21]. To compare different models, we used the area under the curve (AUC) and continuous net reclassification improvement (cNRI). To estimate the AUC, we used the partial receiver operating curve component of the R package [22]. The cNRI was estimated and tested statistically using the reclassification function of the PredictABEL R package [23].

RESULTS

Study population

The age of the study population ranged from 39 to 89 years, and mean±standard deviation (SD) was 54.22±9.23 years (Table 1). The number of female participants (n=7,588) was higher than that of male participants (n=6,438). The mean±SD of the BMI was 24.37±3.09, and the BMI of 5,561 participants (40%) was over the threshold for being overweight (≥25). Among the 14,026 participants to whom the exome array was applied, there were 1,992 (14%) patients with T2DM. The three covariates (sex, age, and BMI) that were used in the GWAS were significantly different between the group with T2DM and the control group (P<0.05) (Supplementary Table 1).

Result of the exome-wide association study

From the AS, HEXA, and CAVAS cohorts, exome chip data of 7,461, 3,456, and 4,566 participants were used in the analysis. After filtering, 50,840 variants were tested for their genetic susceptibility to T2DM. The genomic inflation factor was 1.00, which indicated that there was no inflation of the GWAS result. In the quantile-quantile plot, most of the variants showed null distribution, and a few loci that had a significant association with T2DM showed a marked deviation from the diagonal line (Supplementary Fig. 1).

When we applied Bonferroni's multiple testing correction to the results of GWAS using logistic regression with the additive genetic model, 15 variants showed exome-wide significance with cut-off $P=9.83\times10^{-7}$ (0.05/50,840). We used the PLINK clump function to filter the variants based on linkage disequilibrium (LD) and distance between variants. We applied default parameters with LD=0.2 and distance=250K. After the clumping, we obtained six variants (Table 1 and Fig. 1). Among them, rs2233580 showed the highest significance ($P=1.60\times10^{-10}$) with an odds ratio (OR) of 1.48 (95% confidence interval [CI], 1.31 to 1.67), which is located on paired box gene 4 (*PAX4*). The rs2206734, which is located on cyclin-dependent kinase like 1 (*CDKL1*), showed the second highest significance ($P=1.60\times10^{-10}$). The rs2237895 of *KCNQ1* gene was identified as susceptible loci only in the GWAS of an East Asian population, and it was replicated in our result. The rs11960799 has not been reported as a susceptible locus for T2DM, although it has shown nominal significances for triglyceride level, rheumatoid arthritis, and birth weight previously [24]. The OR of the variant was less than 1 (=0.82), which indicated a protective effect for T2DM. The rs75680863 also showed a protective effect (OR, 0.61; 95% CI, 0.51 to 0.74) for T2DM. This variant was a novel locus that had never been reported for its association with the development of T2DM. rs2237892 also showed a protective effect against the development of T2DM (OR, 0.83; 95% CI, 0.77 to 0.89). It is also located in *KCNQ1*, and previous studies reported that this variant had genome-wide significance in its association with T2DM [7,25]. When nominal significance (P<0.05) is considered, there are a lot of significant associations between T2DM and the other traits and the rs2237892 variant (Supplementary Fig. 2). Interestingly, the ORs of the previous studies of GWAS for T2DM are >1, which indicates increasing risk of disease. rs10965250, having the least significant P value, appears to have a protective effect against T2DM. The variant has been reported to have a genome-wide association with T2DM in a European population [8].

We also performed a meta-analysis using our results and summary statistics from the GWAS catalog database. In the GWAS catalog, we extracted 1,222 variants having genome-

Table 1. Result of exome-wide association analysis of 14,026 Koreans

Chr	BP	SNP	Gene	MA	MAF	OR (95% CI)	*P* value
7	127253550	rs2233580	*PAX4*	A	0.07	1.48 (1.31–1.67)	1.60×10^{-10}
6	20694884	rs2206734	*CDKAL1*	A	0.46	1.25 (1.17–1.34)	2.03×10^{-10}
11	2857194	rs2237895	*KCNQ1*	C	0.31	1.25 (1.16–1.34)	3.73×10^{-9}
5	126356042	rs11960799	*MARCH3*	G	0.35	0.82 (0.76–0.88)	1.19×10^{-7}
22	31007023	rs75680863	*TCN2*	T	0.05	0.62 (0.51–0.74)	2.08×10^{-7}
11	2839751	rs2237892	*KCNQ1*	A	0.38	0.83 (0.77–0.89)	4.83×10^{-7}
9	22133284	rs10965250	–	A	0.43	0.84 (0.71–0.90)	6.73×10^{-7}

Chr, chromosome; BP, base position; SNP, single nucleotide polymorphism; MA, minor allele; MAF, minor allele frequency; OR, odds ratio; CI, confidence interval; *PAX4*, paired box gene 4; *CDKAL1*, cyclin-dependent kinase 5 regulatory associated protein 1-like 1; *KCNQ1*, potassium voltage-gated channel subfamily Q member 1; *MARCH3*, membrane associated ring-CH-type finger 3; *TCN2*, transcobalamin 2.

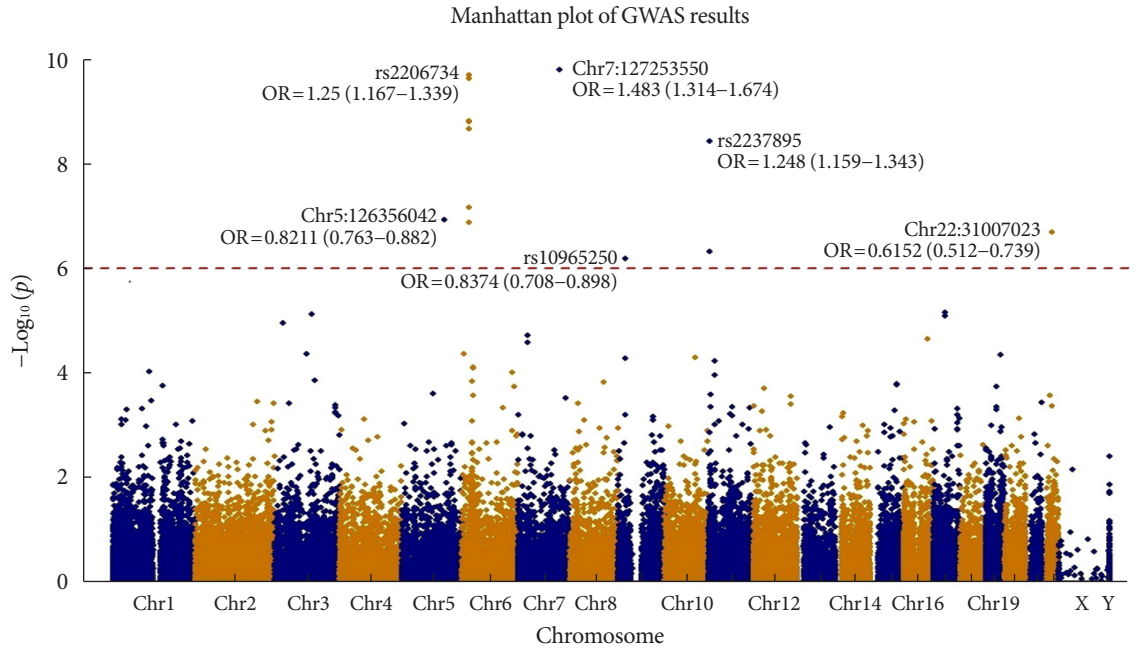


Fig. 1. Manhattan plot of *P* values from exome-wide association analysis (GWAS). The 6 points that are placed over the red dotted line indicates significant loci with Bonferroni's multiple testing correction (adjusted $P=9.83\times10^{-7}$). OR, odds ratio.

wide significance ($P<5\times10^{-8}$) in susceptibility for T2DM. Among them, 124 variants were identified in our exome array data. We used the variants in a meta-analysis, and the summary statistics of the variants from our results and the GWAS catalog were applied to the meta-analysis. The *P* values from a fixed effects model were used to test significance. We used a *P* threshold of 9.83×10^{-7}, which was used to identify significant exome-wide results in our data. To exclude a heterogeneous effect of the variants under the fixed effects assumption, we selected results with a threshold of $I^2<50\%$ and *P* of Cochrane Q statistics >0.05. After filtering, 23 variants were found to be significant (Fig. 2 and Supplementary Table 2). rs10440833, which lies in *CDKL1*, showed the most significant result ($P=3.27\times10^{-27}$). Interestingly, of the exome-wide significant loci in our data, only rs2237895 showed significance in the meta-analysis ($P=1.25\times10^{-14}$) (Supplementary Table 2).

Prediction of future diabetes using clinical and genetic markers

Using significant loci in the exome-wide association analysis, we performed prediction analysis. In the prediction analysis, we used clinical variables including fasting glucose, BMI, high-density lipoprotein, sex, parental history of diabetes, triglyceride, systolic pressure, diastolic pressure, and treatment status of hypertension. These have been frequently used to develop models to predict T2DM [26]. The incident diabetes was identified using 10-year follow-up information of the ASAS cohort. The number of baseline participants who did not have T2DM was 7,889 and 6,317 participants had exome chip data. In total, 1,054 participants having exome chip data developed T2DM during follow-up.

To measure the predictive performance, we estimated the AUC and the net reclassification index (NRI) was used to compare the prediction models with or without genetic information (Fig. 3). When the prediction model was built using the clinical variables, the AUC was 0.71 (95% CI, 0.70 to 0.73). The AUC increased to 0.72 (95% CI, 0.70 to 0.74) with the genotype information of the 7 loci. The difference in the AUCs was significant ($P=1.04\times10^{-3}$) and the NRI (0.11) was also significant ($P=9.80\times10^{-4}$).

We then compared the prediction performances of the clinical variables and genetic markers between BMI subgroups. We divided the total population into three BMI subgroups: normal (BMI <25, $n=3,727$), overweight (25≤ BMI <30, $n=2,314$), and obese (BMI ≥30, $n=267$) individuals. We applied logistic regression to each subgroup and compared the prediction results with the analysis of the entire population. We found that there was a substantial difference in the prediction perfor-

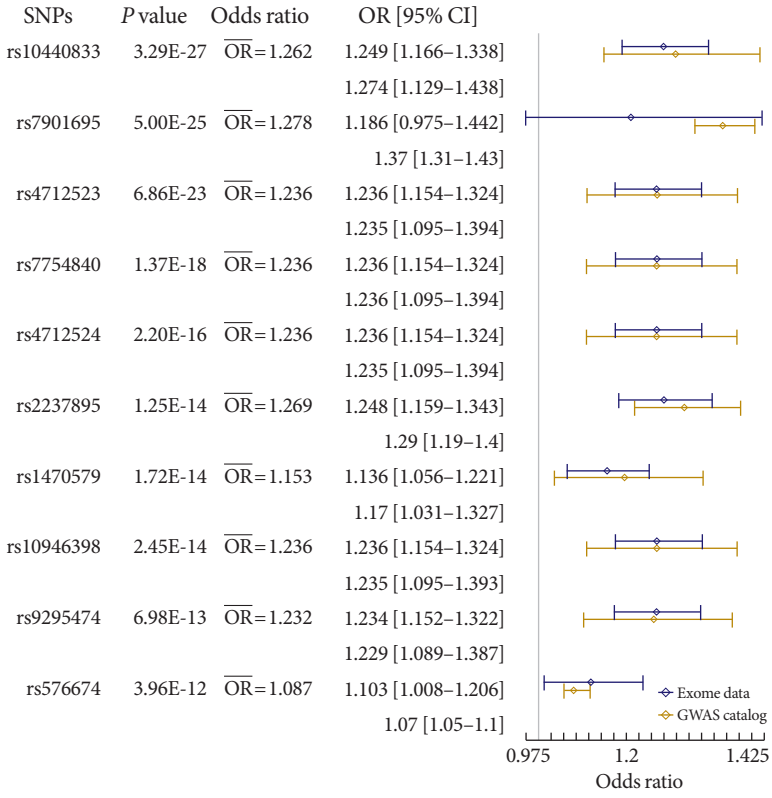


Fig. 2. Forest plot of meta-analysis with result from Korean exome data and genome-wide association stud (GWAS) catalog summary data. *P* value indicates *P* values of meta-analysis with fixed effect model. In the presentation of odds ratio (OR) and confidence interval (CI), upper one is from Korean exome data, and the lower one is from GWAS catalog. Blue line of the forest plot indicates OR and CI of Korean exome data. The whole results were presented in Supplementary Table 2. SNP, single nucleotide polymorphism.

mance for future T2DM between the BMI subgroups (Fig. 3). In the normal group, AUC of the clinical model and combined model (clinical+genetic information) was 0.70 (95% CI, 0.68 to 0.73) and 0.71 (95% CI, 0.69 to 0.74), respectively. The difference in AUC was significant ($P=0.03$). The NRI was 0.14 (95% CI, 0.05 to 0.23), which was also significant ($P=3.58\times10^{-3}$). In the overweight group, there was no significant difference in AUC between clinical model and combined model (0.72 and 0.73, respectively), although the NRI was significant (NRI=0.12; 95% CI, 0.02 to 0.22; $P=0.02$). The difference in AUC between the clinical model and combined model was highest for the obese group (Fig. 3). Although not significant ($P=0.16$), the AUC of the combined model increased to 0.75 (95% CI, 0.69 to 0.81), compared with that of the clinical model in the obese group, which was 0.72 (95% CI, 0.65 to 0.79). Moreover, the NRI was 0.52 ($P=8.00\times10^{-5}$); that is, 4.3-fold that of the overweight group, which had the lowest NRI value. To exclude the possibility that the greater AUC in the combined model in the obese group resulted from a small sample size, we randomly selected the same number of participants as those in the obese group ($n=267$) and estimated the AUC. We repeated this process 1,000 times. No AUC from random selection of the participants showed a greater AUC than that of the obese group (Supplementary Fig. 3).

DISCUSSION

Here we identified the exome-wide loci having susceptibility for T2DM in a Korean population. When we applied the loci to the predictive model for T2DM, we found that while performance of the model was marginal in the total population, the performance was more obvious in the obese group.

rs2233580 had the highest OR (1.45; 95% CI, 1.31 to 1.67; $P=1.60\times10^{-10}$). This variant had been reported in a recent exome-wide association study using exome sequencing data, and the current data were used to replicate the analysis [27].

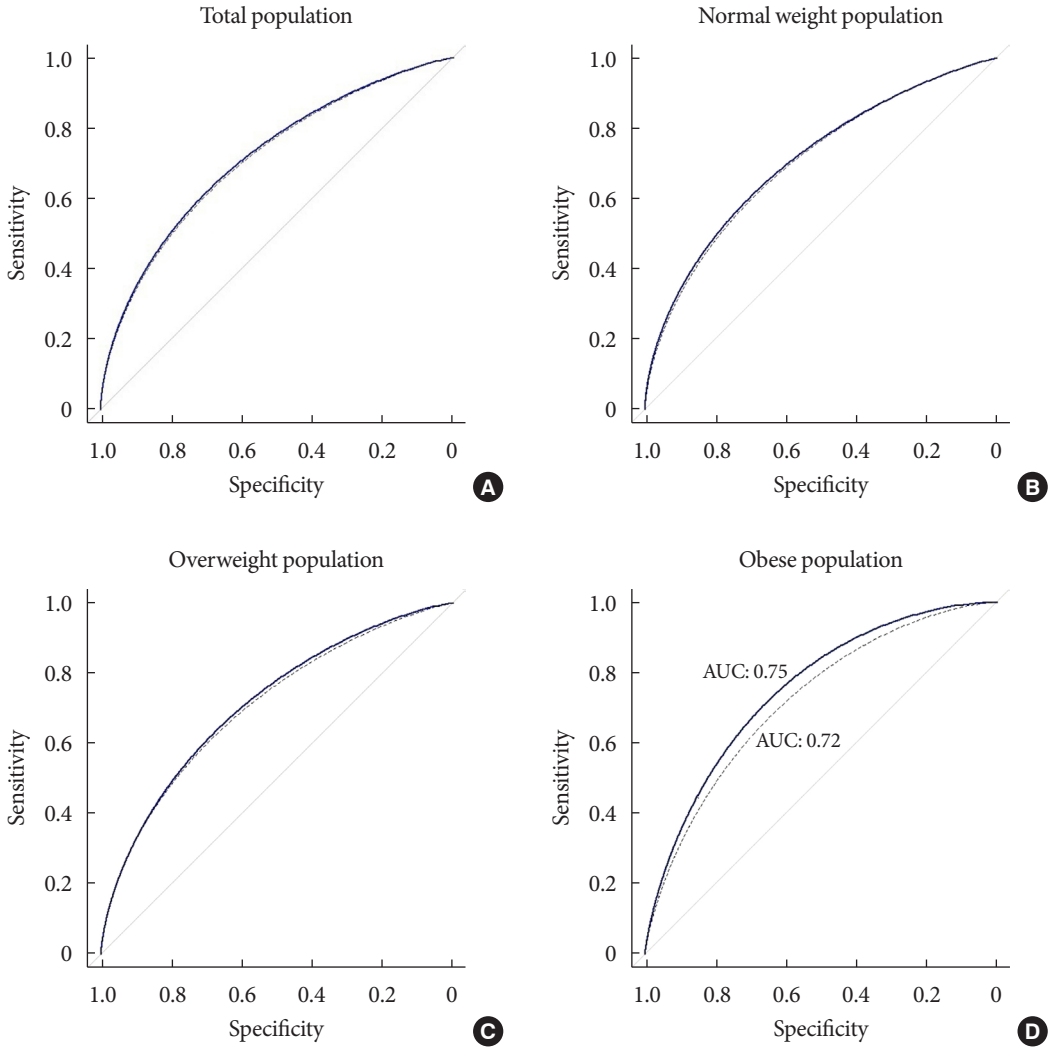


Fig. 3. Receiver operating curve (ROC) plots of prediction model with or without genotype information in total, normal, overweight, and obese population. The thick line indicated ROC of combined model (clinical variable+genetic marker). The dashed line shows ROC of clinical model that integrated clinical variables in the prediction of incident type 2 diabetes mellitus. (A) There was a small gap between the two ROCs of total population. (B) In the normal weight group, there was a slight gap between the lines, which showed no substantially wider than that of total population. (C) Compared with the plots of normal weight population, there was wider gap between the plots, although still the magnitude is small. (D) In the obese group, area under the curve (AUC) of combined model (0.75) is greater than that of clinical model (0.72), which showed a clear gap between the ROC curves.

The variant is located in *PAX4*, which is a transcription factor involved in pancreatic islet development [28]. In previous studies, variants of *PAX4* were found to be associated with T2DM [16,29,30]. In particular, recent genome-wide or exome-wide association studies with East Asian populations revealed *PAX4* variants are associated with T2DM [19,21,32]. In the present study and in an exome chip analysis of T2DM in a Chinese population, rs2233580 had the highest OR for significant variants in each study (1.45 and 1.38 in Korean and Chinese populations, respectively) [12]. *PAX4* has a potential role in the management of T2DM, especially by preserving or replenishing β-cells [28]. It was notable that no significant results for rs2233580 have been reported in GWAS of non-Asian populations. rs2237892 and rs2237895, which are located in *KCNQ1*, are also well-known for their genome-wide significant association with the development of T2DM (Supplementary Fig. 4). Interestingly, the significance of rs2237895 has been reported only for an East Asian population, while rs2237892 and other

variants of *KCNQ1*, such as rs2237896 and rs2237897, were significant in other ethnic groups (Supplementary Table 3). The rs2206734 lies in cyclin-dependent kinase 5 regulatory subunit associated protein 1 like 1 (*CDKAL1*), and the SNP has been reported in many GWASs of T2DM or related traits such as fasting glucose level and BMI. When we searched the variant on the KRGDB, we found many associations with metabolic traits including T2DM (Supplementary Fig. 5). rs10965250 is located in the intergenic region between *AL157937.1* and *CDKN2B-AS1* and has been reported to be associated with T2DM and fasting glucose, 2-hour glucose level, waist-hip ratio, and birth weight in various ethnic groups (Supplementary Fig. 6). The two variants (rs2237892 and rs10965250) had OR <1, indicating the protective effect of the variants in the development of T2DM. However, the previously reported ORs were >1, indicating the ethnic-specific effect of the variant. This finding should be confirmed in future studies.

rs11960799 within membrane associated ring-CH-type finger 3 (*MARCH3*) located at 126,356,042 of chromosome 3 had never been reported for its association with T2DM. In previous studies, rs11960799 showed a significant association with triglyceride level, rheumatoid arthritis, and birth weight. However, these variants were found to have nominal significance ($P<0.05$), and did not reach genome-wide significance. Considering a previous report that showed the significant transcript change of *MARCH3* in human islet cells [31], it seems possible that rs11960799 is at least a variant associated with the development of T2DM. rs75680863 is another variant newly discovered in the present study. No previous GWASs have ever reported the significant association of this variant and T2DM. The OR of rs75680863 was 0.63, which is equivalent to an OR of 1.58 if the direction of the coefficient of the variant is positive, and the size of the OR is highest among those of the significant variants. rs75680863 lies within transcobalamin 2 (*TCN2*), the gene for transcobalamin 2, which is a member of the vitamin B12 binding protein family. Because vitamin B12 deficiency is closely related to T2DM [32], there is a possible pathway involving vitamin B12 metabolism and the development of T2DM.

In the present meta-analysis, we identified 32 loci showing significance, and the levels were sufficiently high to reach genome-wide significance ($P=5\times10^{-8}$). In a fixed effects model, more variants ($n=64$) among the 124 that were used in the meta-analysis showed significance. However, after filtering based on Cochrane Q statistics and the I^2 heterogeneity index, the number decreased to 32. Moreover, only a single significant variant from our results was included in the significant result from the meta-analysis. These might imply substantial heterogeneity in the genetic effect for development of T2DM in Koreans.

In this study, we performed prediction analysis using 7 significant loci. While many studies used genetic risk score (GRS) by summing the risk scores of individual loci, we used the 7 loci directly because the number of loci was not so high. When we applied the 7 significant loci to the prediction analysis, we obtained a significant, but marginal, increase in the AUC. This result was consistent with those of previous studies [33]. These results were consistent regardless of the number of genetic markers that were used in the prediction models. For example, Vassy et al. [34] used different numbers of common genetic variants for predicting incident T2DM. Even when genetic information was integrated into the prediction model, the performance showed only marginal improvement. Talmud et al. [35] reported similar results with weighted scoring of 65 common genetic variants. They showed a significant but marginal change of receiver operating curve (ROC) (from 0.75 to 0.76, $P=0.0003$) with integrating a GRS into their prediction model. Lall et al. [36] applied a greater number of genetic variants into a prediction model. They used 1,000 SNPs for developing a prediction model using a doubly weighted GRS method. In the results, the prediction model with the GRS showed a slight increase of AUC (from 0.77 to 0.79), but highly significant *P* value (2.01×10^{-11}) of the difference of AUC between the model with or without GRS [36]. The NRI was mildly but significantly increased in our results; this compares to previous studies that also reported a mild increase of continuous NRI in a model with genotype information [34-36]. This seems to indicate that one clinical implication of a genetic prediction model of incident T2DM lies with the reclassification of the at-risk population, not with lifting overall prediction performance.

When we divided the total population into three groups (normal, overweight, and obese), the results of prediction analysis for incident T2DM differed in the obese group compared with the other groups. The AUC and NRI changed slightly in the normal and overweight groups, but the performance measures showed greater changes in the obese group. Although the AUC of the prediction model with genetic information showed nonsignificant change, the magnitude of the change of the AUC was greater than for the other subgroups (Fig. 2). Moreover, the NRI was highly significant and far

greater than the NRI of the other subgroups. The change of prediction performance in different subgroups of a population have been reported in many GWASs. In particular, such results were found in BMI subgroups. Previous studies reported similar findings in terms of different association and prediction performance in BMI subgroups. Morris et al. [38] found that rs4506565 near transcription factor 7-like 2 (*TCF7L2*) had different associations in a nonobese group (BMI ≥30; OR, 1.54) and an obese group (BMI >30; OR, 1.20). In another GWAS for T2DM using BMI subgroups [39], rs8090011 in *LAMA1* had genome-wide significance in lean individuals (BMI <25; $P=8.4\times10^{-9}$) and not in obese individuals (BMI ≥30; $P=0.04$). Timpson et al. [40] also reported that rs8050136 of fat mass and obesity-associated protein (*FTO*) gene was detectable in the obese group (BMI ≥30.2; $P=1.3\times10^{-13}$), while the significance disappeared in the nonobese group (BMI <30.2; $P=0.19$). Besides significance, prediction performance differs according to the BMI subgroups. In the prediction analysis of T2DM with GRS of 65 common variants, samples of the lowest tertile (BMI <24.5) had an NRI of 27.6%, while those of the upper tertile (BMI >27.5) had an NRI of 2.6% [35]. Previous studies showed that the prediction model with GRS outperformed in the lean body group than in the obese group, but our result was different. The difference of AUC and magnitude of reclassification was greater in the obese group. While GWASs have been identifying susceptible loci for diabetes, previous studies have shown that integration of genetic information or GRS from GWAS results increased the performance of diabetes prediction marginally [33]. However, we found that the prediction performance increased in obese group more than in whole population. In the obese group, the clinic-genetic prediction model increased AUC marginally, but cNRI significantly ($P=8.00\times10^{-5}$). Moreover, permutation test revealed that the AUC of prediction model of obese group is significant (Supplementary method). These results indicate that the clinic-genetic model of obese group increased overall prediction performance marginally, but re-assign risk probability to patients more accurately. Clinically, it would be helpful when patients need to be matched to a preventive protocol because risk assessment is more accurate. Considering obese group had limited number of population in general, it might be cost-effective to apply such model to the specific subgroup of population.

These seemed to indicate that the genetic information may be more valuable predictor of diabetes in a specific subpopulation. The identification of such groups should be investigated in further research. In addition, each loci showed the same tendency of high performance of prediction in obese group (Supplementary Table 4). This might result from our use of exonic variants in the prediction model. We thought that the change of prediction accuracy between BMI subgroups results from epigenetic change of the loci. It is reported that the *KCNQ1* gene of blood cells is hyper-methylated in obese population [37]. If so, it is also possible that *KCNQ1* gene of diabetes-related cells can be hyper-methylated, which might result in change of *KCNQ1* gene expression, and prediction accuracy of the loci eventually. Considering that *KCNQ1* contains two loci of the significant results in this analysis, it is possible that higher AUC is associated with methylation change, and the other loci might also have the methylation changes nearby the loci, which affect the prediction accuracy in obese group.

In summary, we found significant exonic variants for T2DM, which were previously identified or newly detected. Genetic prediction using the loci showed a mild but significant increase of AUC and NRI. We found that the variants had higher performance in the obese group than in the normal and overweight groups. We believe that our results provide clues to reveal the genetic architecture of T2DM in the Korean population.

SUPPLEMENTARY MATERIALS

Supplementary materials related to this article can be found online at https://doi.org/10.4093/dmj.2019.0163

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

AUTHOR CONTRIBUTIONS

Conception or design: S.B.C.
Acquisition, analysis, or interpretation of data: S.B.C., J.H.J., M.G.C., S.C.K.
Drafting the work or revising: S.B.C., J.H.J.
Final approval of the manuscript: S.B.C., J.H.J., M.G.C., S.C.K.

ORCID

Seong Beom Cho *https://orcid.org/0000-0002-5784-1645*

FUNDING

This work was supported by the Post-Genome Multi-Ministerial Project (grant number: 3000-3031-405:2017-NI72001-00, 3000-3031-405:2019-NI-093-00).

ACKNOWLEDGMENTS

None

REFERENCES

1. Bellou V, Belbasis L, Tzoulaki I, Evangelou E. Risk factors for type 2 diabetes mellitus: an exposure-wide umbrella review of meta-analyses. PLoS One 2018;13:e0194127.
2. Moore AF, Florez JC. Genetic susceptibility to type 2 diabetes and implications for antidiabetic therapy. Annu Rev Med 2008;59:95-111.
3. Billings LK, Florez JC. The genetics of type 2 diabetes: what have we learned from GWAS? Ann N Y Acad Sci 2010;1212: 59-77.
4. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet 2017;101:5-22.
5. Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. Annu Rev Genomics Hum Genet 2018; 19:73-96.
6. McCarthy MI, Zeggini E. Genome-wide association studies in type 2 diabetes. Curr Diab Rep 2009;9:164-71.
7. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 2008;40:1098-102.
8. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 2010; 42:579-89.
9. Cook JP, Morris AP. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. Version 2. Eur J Hum Genet 2016;24:1175-80.
10. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature 2016;536:41-7.
11. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A, Stringham HM, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 2013;45: 197-201.
12. Cheung CY, Tang CS, Xu A, Lee CH, Au KW, Xu L, et al. Exome-chip association analysis reveals an Asian-specific missense variant in PAX4 associated with type 2 diabetes in Chinese individuals. Diabetologia 2017;60:107-15.
13. Yasukochi Y, Sakuma J, Takeuchi I, Kato K, Oguri M, Fujimaki T, et al. Two novel susceptibility loci for type 2 diabetes mellitus identified by longitudinal exome-wide association studies in a Japanese population. Genomics 2019;111: 34-42.
14. Mahajan A, Wessel J, Willems SM, Zhao W, Robertson NR, Chu AY, et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. Nat Genet 2018;50:559-71.
15. Kim Y, Han BG; KoGES group. Cohort profile: the Korean Genome and Epidemiology Study (KoGES) consortium. Int J Epidemiol 2017;46:e20.
16. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes: 2018. Diabetes Care 2018;41(Suppl 1):S13-27.
17. Kim YK, Hwang MY, Kim YJ, Moon S, Han S, Kim BJ. Evaluation of pleiotropic effects among common genetic loci identified for cardio-metabolic traits in a Korean population. Cardiovasc Diabetol 2016;15:20.
18. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing- data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084-97.
19. 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;81:559-75.
20. Jung KS, Hong KW, Jo HY, Choi J, Ban HJ, Cho SB, et al. KRG-DB: the large-scale variant database of 1722 Koreans based on whole genome sequencing. Database (Oxford) 2020; 2020: baz146.
21. R Core Team: R: A language and environment for statistical computing 2018. Available from: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing (cited 2020 May 5).
22. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12: 77.
23. Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction mod-

els. Eur J Epidemiol 2011;26:261-4.
24. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 2009;41:56-65.
25. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 2008;40:1092-7.
26. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011;9:103.
27. Kwak SH, Chae J, Lee S, Choi S, Koo BK, Yoon JW, et al. Nonsynonymous variants in PAX4 and GLP1R are associated with type 2 diabetes in an East Asian population. Diabetes 2018;67:1892-902.
28. Lorenzo PI, Juarez-Vicente F, Cobo-Vuilleumier N, Garcia-Dominguez M, Gauthier BR. The diabetes-linked transcription factor PAX4: from gene to functional consequences. Genes (Basel) 2017;8:101.
29. Ma RC, Hu C, Tam CH, Zhang R, Kwan P, Leung TF, et al. Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4. Diabetologia 2013;56: 1291-305.
30. Sujjitjoon J, Kooptiwut S, Chongjaroen N, Semprasert N, Hanchang W, Chanprasert K, et al. PAX4 R192H and P321H polymorphisms in type 2 diabetes and their functional defects. J Hum Genet 2016;61:943-9.
31. Ottosson-Laakso E, Krus U, Storm P, Prasad RB, Oskolkov N, Ahlqvist E, et al. Glucose-induced changes in gene expression in human pancreatic islets: causes or consequences of chronic hyperglycemia. Diabetes 2017;66:3013-28.
32. Pflipsen MC, Oh RC, Saguil A, Seehusen DA, Seaquist D, Topolski R. The prevalence of vitamin B(12) deficiency in patients with type 2 diabetes: a cross-sectional study. J Am Board Fam Med 2009;22:528-34.
33. Vassy JL, Meigs JB. Is genetic testing useful to predict type 2 diabetes? Best Pract Res Clin Endocrinol Metab 2012;26:189-201.
34. Vassy JL, Hivert MF, Porneala B, Dauriz M, Florez JC, Dupuis J, et al. Polygenic type 2 diabetes prediction at the limit of common variant detection. Diabetes 2014;63:2172-82.
35. Talmud PJ, Cooper JA, Morris RW, Dudbridge F, Shah T, Engmann J, et al. Sixty-five common genetic variants and prediction of type 2 diabetes. Diabetes 2015;64:1830-40.
36. Lall K, Magi R, Morris A, Metspalu A, Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. Genet Med 2017;19:322-9.
37. Gomez-Uriz AM, Milagro FI, Mansego ML, Cordero P, Abete I, De Arce A, et al. Obesity and ischemic stroke modulate the methylation levels of KCNQ1 in white blood cells. Hum Mol Genet 2015;24:1432-40.
38. Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genet Epidemiol 2010;34:335-43.
39. Perry JR, Voight BF, Yengo L, Amin N, Dupuis J, Ganser M, et al. Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. PLoS Genet 2012;8:e1002741.
40. Timpson NJ, Lindgren CM, Weedon MN, Randall J, Ouwehand WH, Strachan DP, et al. Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. Diabetes 2009;58:505-10.

Supplementary Method

Permutation test of area under the curve

We performed permutation test of area under the curve (AUC) of clinic-genetic model in obese group to identify whether the AUC is valid or not. First, we permutated diabetes status of obese group and estimated AUC of the data with the permutated disease status 1,000 times. Second, we randomly sampled the same number of obese group from normal and overweight group, and permutated diabetes status. The *P* value of the permutation test is

$$P=\frac{n(permAUC>oriAUC)}{nPerm}$$

where *permAUC* is a AUC with permutated diabetes status, and *oriAUC* is AUC of clinico-genetic model in obese group. *n()* is number of *permAUC*>*oriAUC*, and *nPerm* is number of permutation (1,000).

Supplementary Table 1. Comparison of sex, age, and BMI between normal and diabetes group

	Total	Control	DM	*P* value
Sex	6,438/7,588	5,382/6,652	1,056/936	7.27E-12
Age, yr	54.22±9.23	53.77±9.2	56.95±8.31	2.57E-52
BMI, kg/m^2	24.35±3.09	24.18±3.03	25.38±3.22	1.53E-52

BMI, body mass index; DM, diabetes mellitus.

Supplementary Table 2. Result of meat-analysis with our data and genome-wide association study catalog

Chr	BP	SNP	A1	A2	P.FE	OR.FE	Q	I
6	20688121	rs10440833	A	T	3.29E-27	1.2497	0.9853	0
10	114754088	rs7901695	G	A	5.00E-25	1.3531	0.1672	47.59
6	20657564	rs4712523	G	A	6.86E-23	1.2552	0.5603	0
6	20661250	rs7754840	C	G	1.37E-18	1.1991	0.2851	12.49
6	20657865	rs4712524	G	A	2.20E-16	1.2281	0.7947	0
11	2857194	rs2237895	C	A	1.25E-14	1.2626	0.6013	0
3	185529080	rs1470579	C	A	1.72E-14	1.1644	0.4303	0
6	20661034	rs10946398	C	A	2.45E-14	1.1923	0.1729	46.17
6	20652717	rs9295474	G	C	6.98E-13	1.197	0.2171	34.36
13	33554302	rs576674	G	A	3.96E-12	1.0715	0.5165	0
2	161171454	rs7593730	A	G	8.85E-12	1.1041	0.2747	16.19
11	17409572	rs5219	A	G	1.16E-11	1.1277	0.2851	12.47
10	94465559	rs5015480	G	A	6.70E-11	1.1767	0.7479	0
11	92673828	rs1387153	A	G	1.27E-09	1.0891	0.8863	0
17[a]	36098040	rs4430796	G	A	3.65E-09	1.132	0.1743	45.82
15[a]	80432222	rs11634397	G	A	3.95E-09	1.061	0.5825	0
2[a]	227093745	rs2943641	A	G	1.20E-08	1.1753	0.2832	13.18
9[a]	136155000	rs635634	A	G	1.34E-08	1.0843	0.4732	0
3[a]	185511687	rs4402960	A	C	3.40E-08	1.1404	0.9856	0
9[a]	4292083	rs10758593	A	G	5.40E-08	1.0795	0.9431	0
12[a]	55098996	rs1153188	A	T	4.87E-07	1.0794	0.762	0
11[a]	17418477	rs757110	C	A	5.15E-07	1.0731	0.6318	0
15[a]	77832762	rs7177055	A	G	6.74E-07	1.0724	0.2404	27.45

Chr, chromosome; BP, base position; SNP, single nucleotide polymorphism; A1, minor allele; A2, major allele; P.FE, *P* value of fixed effect model; OR.FE, odds ratio of fixed effect model; Q, *P* value of Cochrane Q statistic; I, I^2 heterogeneity index.
[a]Chromosome indicates that the variant is significant in exome-wide association analysis.

Supplementary Table 3. Genome-wide association study catalogue data of rs2237892, rs2237895, rs2237896, and rs2237897

PubMed ID	Initial sample size	Replication sample size	Chr	POS	Strongest SNP-risk allele	Risk allele frequency	*P* value	OR or beta	95% CI (text)	Platform [SNPs passing QC]	Mapped_ trait
28869590	6,353 South Asian ancestry cases, 7,179 South Asian ancestry controls, 3,871 European ancestry cases, 16,427 European ancestry controls, 34,840 cases, 114,981 controls	7,888 South Asian ancestry cases, 20,679 South Asian ancestry controls, 387 European ancestry cases, 2,092 European ancestry controls, 19,998 East Asian ancestry cases, 30,983 East Asian ancestry controls	11	2818521	rs2237892-T	0.0626	2.00E-63	0.215	0.19–0.24 unit decrease	Affymetrix, Illumina [NR] (imputed)	Type 2 diabetes mellitus
28869590	6,353 South Asian ancestry cases, 7,179 South Asian ancestry controls, 3,871 European ancestry cases, 16,427 European ancestry controls, 34,840 cases, 114,981 controls	7,888 South Asian ancestry cases, 20,679 South Asian ancestry controls, 387 European ancestry cases, 2,092 European ancestry controls, 19,998 East Asian ancestry cases, 30,983 East Asian ancestry controls	11	2818521	rs2237892-T	0.0626	1.00E-09	0.1389	0.094–0.183 unit decrease	Affymetrix, Illumina [NR] (imputed)	Type 2 diabetes mellitus
21573907	1,804 Hispanic cases, 780 Hispanic controls	European ancestry individuals	11	2818521	rs2237892-C	NR	4.00E-06	1.2	1.11–1.29	Affymetrix [315658]	Type 2 diabetes mellitus
18711367	187 Japanese ancestry cases, 1,504 Japanese ancestry controls	6,552 East Asian ancestry cases, 6,621 East Asian ancestry controls, 2,830 European ancestry cases, 3,740 European ancestry controls	11	2818521	rs2237892-C	0.61	2.00E-42	1.45	1.34–1.47	NR [82343]	Type 2 diabetes mellitus
19401414	519 Japanese ancestry cases, 503 Japanese ancestry controls	5,110 Japanese ancestry cases, 6,867 Japanese ancestry controls	11	2818521	rs2237892-C	0.59	1.00E-26	1.33	1.27–1.41	Illumina [482625]	Type 2 diabetes mellitus
23945395	5,976 Japanese ancestry cases, 20,829 Japanese ancestry controls	24,416 East Asian ancestry cases, 13,985 East Asian ancestry controls	11	2818521	rs2237892-C	0.61	4.00E-29	1.3	1.24–1.36	Illumina [6209637] (imputed)	Type 2 diabetes mellitus
22961080	1,839 Han Chinese ancestry cases, 1,873 Han Chinese ancestry controls	15,979 East Asian ancestry cases, 19,360 East Asian ancestry controls, 794 Malay ancestry cases, 1,240 Malay ancestry controls, 159 Filipino ancestry cases, 1,624 Filipino ancestry controls	11	2818521	rs2237892-C	0.657	1.00E-07	1.32	1.19–1.46	Illumina [2234194] (imputed)	Type 2 diabetes mellitus
30054458	61,714 European ancestry cases, 1,178 Pakistani ancestry cases, 593,952 European ancestry controls, 2,472 Pakistani ancestry controls	NA	11	2818521	rs2237892-T	0.062489595	9.00E-10	0.096	0.065–0.127 unit decrease	NR [5035015] (imputed)	Type 2 diabetes mellitus

(*Continued to the next page*)

Supplementary Table 3. Continued

PubMed ID	Initial sample size	Replication sample size	Chr	POS	Strongest SNP-risk allele	Risk allele frequency	*P* value	OR or beta	95% CI (text)	Platform [SNPs passing QC]	Mapped_ trait
20174558	995 Han Chinese ancestry cases, 894 Han Chinese ancestry controls	1,803 Han Chinese ancestry cases, 1,473 Han Chinese ancestry controls	11	2835964	rs2237895-C	0.33	1.00E-09	1.29	1.19–1.40	Illumina [516737]	Type 2 diabetes mellitus
26818947	15,463 Japanese ancestry cases, 26,183 Japanese ancestry controls	20,490 East Asian ancestry cases, 22,922 East Asian ancestry controls, 38,947 European ancestry cases, 121,903 European ancestry controls, 10,587 South Asian ancestry cases, 14,378 South Asian ancestry controls, 3,848 Mexican/Latino cases, 4,366 Mexican/ Latino controls	11	2837210	rs2237896-G	NR	3.00E-70	1.32546	1.28479–1.36741	Illumina [7521072] (imputed)	Type 2 diabetes mellitus
18711366	194 Japanese ancestry cases, 1,558 Japanese ancestry controls	5,324 East Asian ancestry cases, 7,037 East Asian ancestry controls, 4,085 European ancestry cases, 5,302 European ancestry controls	11	2837316	rs2237897-C	0.34	1.00E-16	1.33	1.24–1.41	Affymetrix [207097]	Type 2 diabetes mellitus
24390345	3,848 Mexican and other Latin American ancestry cases, 4,366 Mexican and other Latin American ancestry controls	896 Mexican American cases, 832 Mexican American controls, 2,768 East Asian ancestry cases, 2,880 East Asian ancestry controls, 1,082 South Asian ancestry cases, 1,105 South Asian ancestry controls, 1,874 European ancestry cases, 1,333 European ancestry controls, 2,127 African American cases, 2,103 African American controls, 698 Native Hawaiian ancestry cases, 579 Native Hawaiian ancestry controls, 2,009 Singaporean ancestry cases, 1,959 Singaporean ancestry controls	11	2837316	rs2237897-C	NR	9.00E-15	1.31	1.22–1.41	Illumina [9200000] (imputed)	Type 2 diabetes mellitus

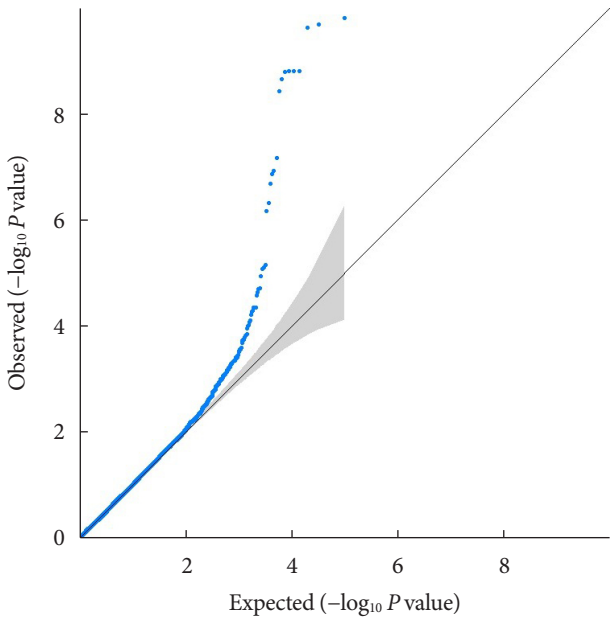
Note that the rs2237895 was identified as susceptible loci of type 2 diabetes mellitus only in East Asian population. The other trans-ethnic genome-wide association studies were not able to identify the rs2237895 as susceptible loci to type 2 diabetes mellitus. We curated necessary information from original data.

Chr, chromosome; POS, position; SNP, single nucleotide polymorphism; OR, odds ratio; CI, confidence interval; NR, not reported; NA, not available.

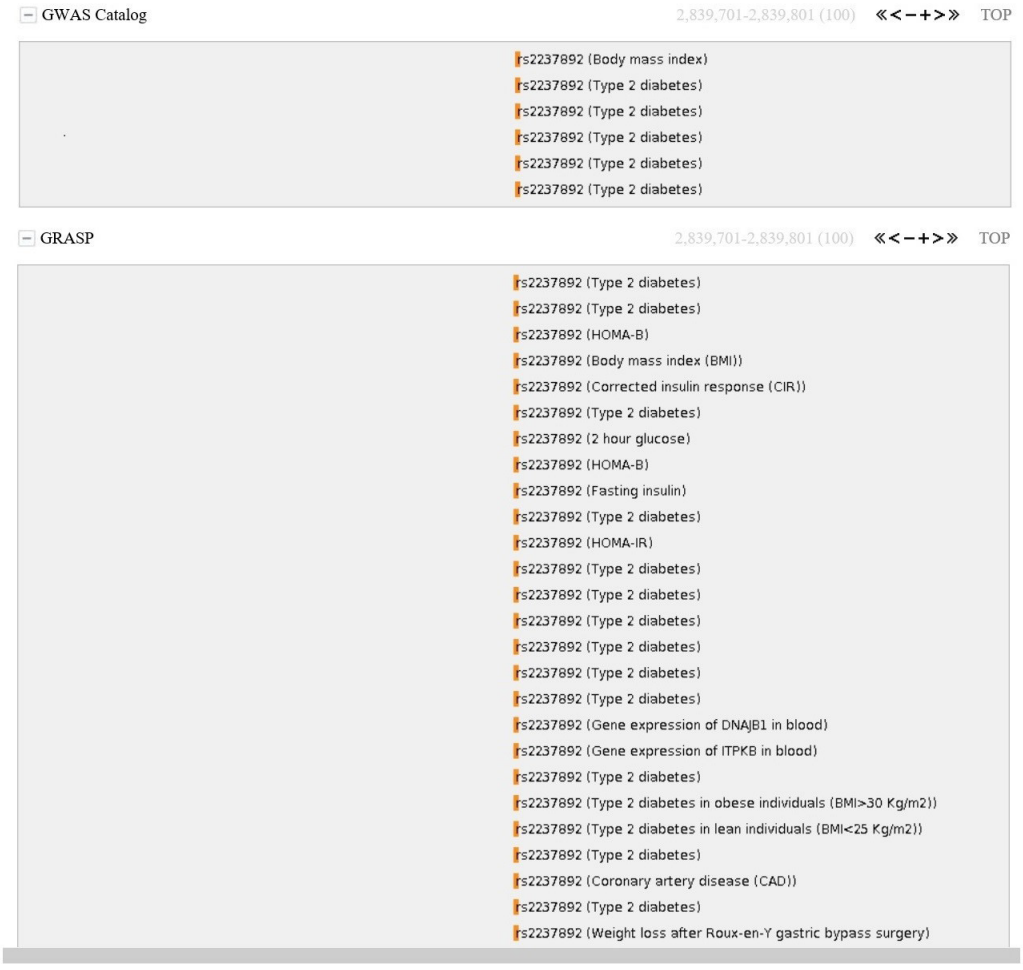
Supplementary Table 4. Predictive performances in each subgroup

SNP	Normal group	Overweight group	Obesity group
rs10965250	0.70 (0.68–0.73)	0.71 (0.68–0.73)	0.72 (0.65–0.78)
rs2206734	0.71 (0.68–0.73)	0.71 (0.68–0.73)	0.72 (0.65–0.78)
rs11960799	0.71 (0.68–0.73)	0.71 (0.68–0.73)	0.72 (0.65–0.79)
rs2237895	0.71 (0.68–0.73)	0.71 (0.68–0.73)	0.73 (0.66–0.79)
rs75680863	0.70 (0.68–0.73)	0.71 (0.68–0.73)	0.73 (0.66–0.79)
rs2237892	0.71 (0.68–0.73)	0.71 (0.68–0.73)	0.73 (0.66–0.80)
rs2233580	0.71 (0.68–0.73)	0.71 (0.68–0.74)	0.73 (0.67–0.79)

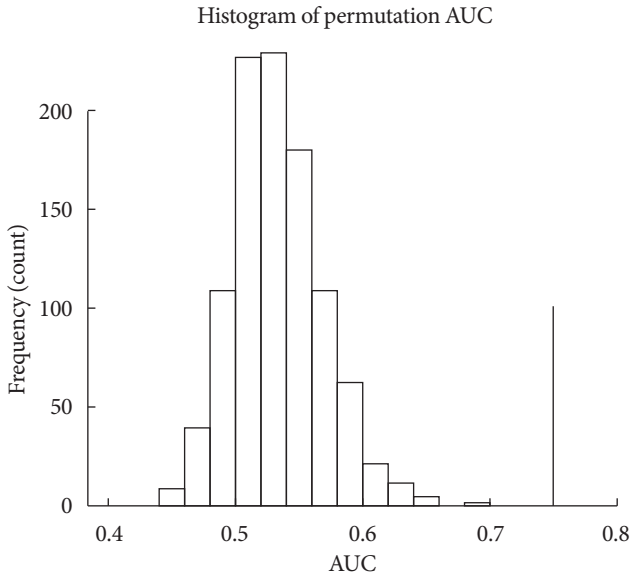
Values are presented as area under curve (AUC).
SNP, single nucleotide polymorphism.



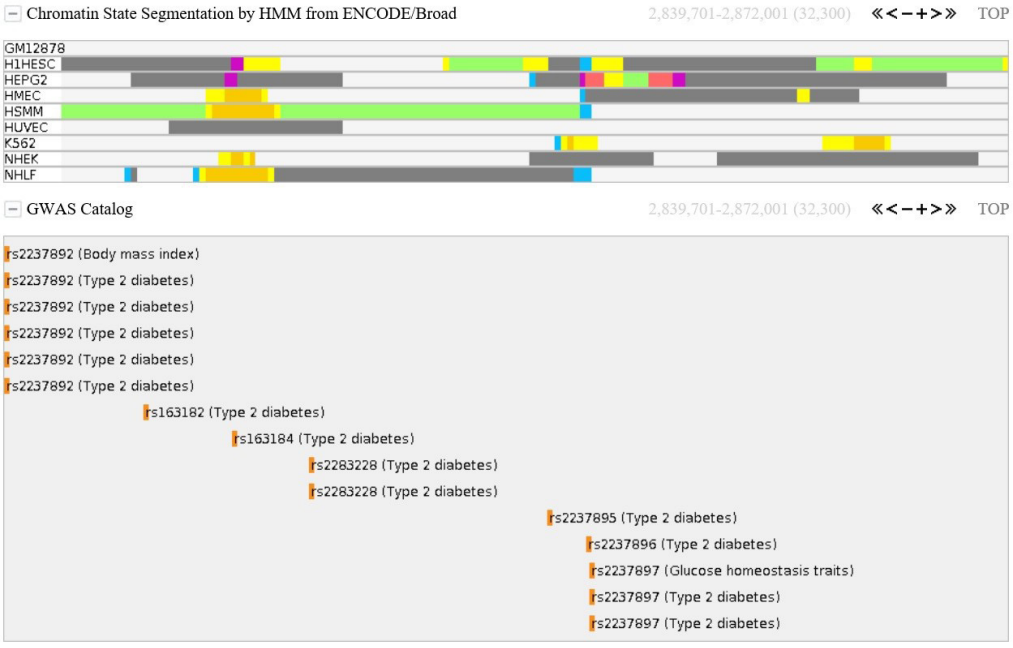
Supplementary Fig. 1. Quantile-quantile plot of *P* values from genome-wide association study.



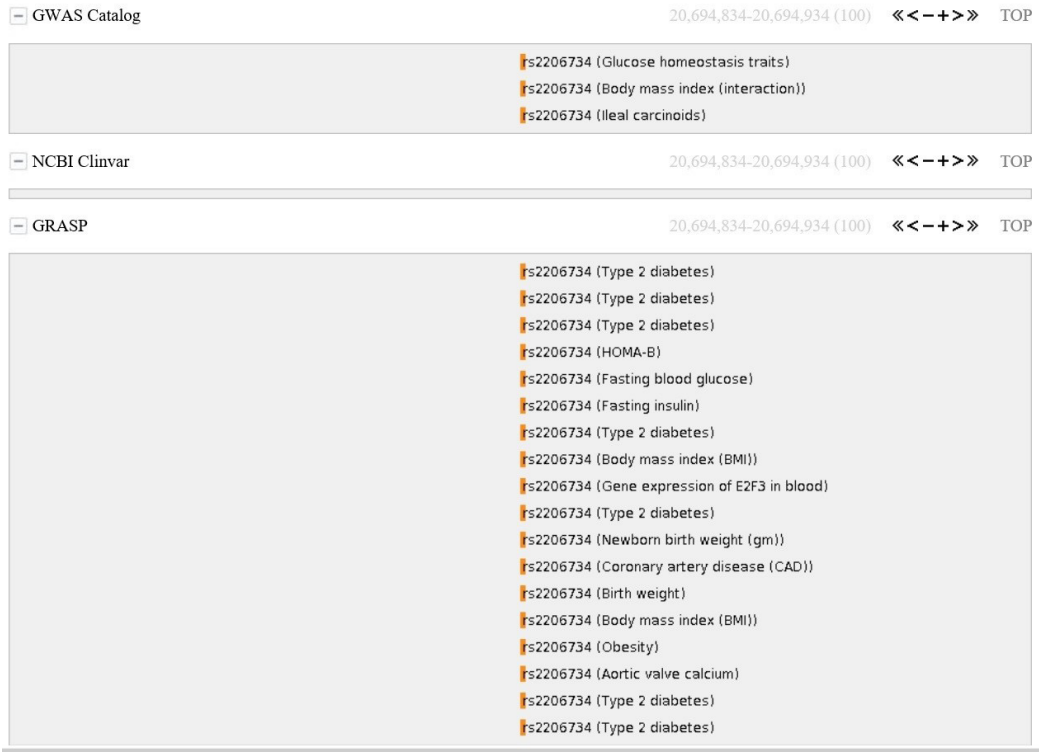
Supplementary Fig. 2. Genome-wide association study (GWAS) results of rs2237892 from Korean Reference Genome Database (KRGDB). Search result from KRGDB. The original data comes from GWAS catalog and genome-wide repository of associations between SNPs and phenotypes (GRASP) database. The results from GWAS catalog had genome-wide significance ($P<5\times10^{-8}$), while those from GRASP had a range of *P* value from nominal significance to genome-wide significance.



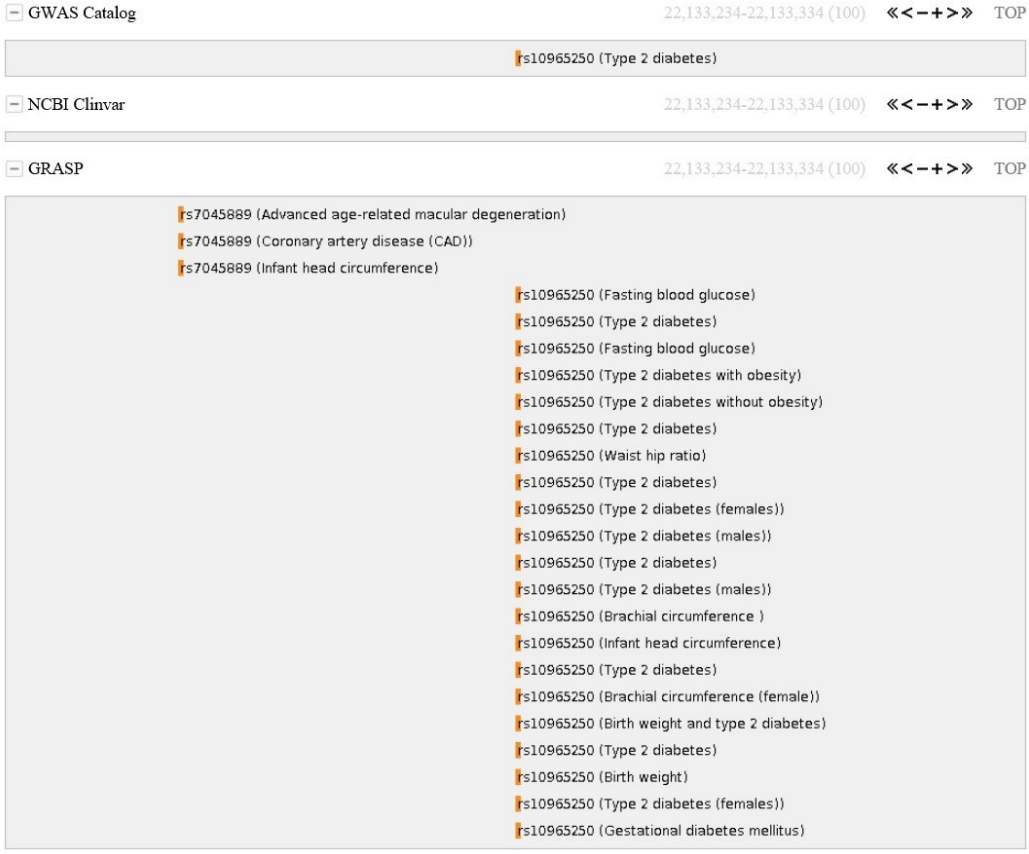
Supplementary Fig. 3. Histogram of permutated area under the curve (AUC) values. We randomly sampled the same number of participants with that of obese population ($n=295$), and estimated AUC with the sampled data. We iterated this process 10,000 times, and there was no permutated AUC value greater than AUC of the obese population (0.75). The thick perpendicular line indicates the AUC of the obese population.



Supplementary Fig. 4. Previous genome-wide association study (GWAS) results in potassium voltage-gated channel subfamily Q member 1 (*KCNQ1*) gene region. As shown, in the *KCNQ1* gene region, there are many genome-wide significant loci for type 2 diabetes mellitus development. HMM, hidden Markov model.



Supplementary Fig. 5. Search result of rs2206734 in Korean Reference Genome Database (KRGDB). In the genome-wide association study (GWAS) catalog, there were no significant results for susceptibility to type 2 diabetes mellitus. However, genome-wide repository of associations between SNPs and phenotypes (GRASP) shows many nominally or genome-widely significant results in the GWAS for type 2 diabetes mellitus and associated trait such as fasting blood glucose and fasting insulin. NCBI, National Center for Biotechnology Information.



Supplementary Fig. 6. Search result of rs10965250 in Korean Reference Genome Database (KRG-DB). On the variant, there is a genome-wide significant result for type 2 diabetes mellitus development in the genome-wide association study (GWAS) catalog. In the genome-wide repository of associations between SNPs and phenotypes (GRASP), many results having nominal significance ($P<0.05$) are shown. NCBI, National Center for Biotechnology Information.