보건학박사 학위논문

Type 2 diabetes and DNA methylation Changes as Risk Factors of Occurrence and Prognosis

제2형 당뇨 관련 후성유전학 지표 발굴 연구

2021년 8월

서울대학교 보건대학원

보건학과 유전체 & 건강 빅데이터 전공

김 하 경

Type 2 diabetes and DNA methylation Changes as Risk Factors of Occurrence and Prognosis

제2형 당뇨 관련 후성유전학 지표 발굴 연구

지도교수 성 주 헌

이 논문을 보건학 박사 학위논문으로 제출함

2021년 5월

서울대학교 보건대학원

보건학과 유전체 & 건강 빅데이터 전공

김 하 경

김하경의 박사 학위논문을 인준함

2021년 5월

위 원 장 조 성 일

부위원장 곽 수 헌

위 원 원 성 호

위 원 최 승 아

위 원 성 주 헌

Type 2 diabetes and DNA methylation Changes as Risk Factors of Occurrence and Prognosis

제2형 당뇨 관련 후성유전학 지표 발굴 연구

지도교수 성 주 헌

이 논문을 보건학 박사 학위논문으로 제출함

2021년 5월

서울대학교 보건대학원

보건학과 유전체 & 건강 빅데이터 전공

김 하 경

김하경의 박사 학위논문을 인준함

2021년 5월

위 원 장 조 성 일

부위원장 곽 수 헌

위 원 원 성 호

위 원 최 승 아

위 원 성 주 헌

Abstracts

Objective: There is growing body of evidence that epigenetic changes including DNA methylation (DNAm) influence the risk of type 2 diabetes and its microvascular complications. We conducted a methylome-wide association study (MWAS) to identify differentially methylated regions (DMRs) of type 2 diabetes and diabetic kidney disease (DKD) in Korean population.

Methods: We performed an initial MWAS in 232 participants in a case-control study design with type 2 diabetes and 197 non-diabetic controls with Illumina EPIC bead chip using peripheral blood leukocytes. Type 2 diabetes group was subdivided to 87 DKD cases and 80 non-DKD controls. Additional 819 individuals from two population-based cohorts were used to investigate the association of the identified DMRs with quantitative metabolic traits. We developed a DNAm score using identified DMRs to predict the occurrence of type 2 diabetes. To examine the causal relationship between differentially methylated CpGs and type 2 diabetes/DKD and between the metabolic traits and differentially methylated status, we performed Mendelian randomization (MR) analyses.

Results: We identified eight DMRs (each at *BMP8A*, *NBPF20*, *STX18*, *ZNF365*, *CPT1A*, and *TRIM37*, and two at *TXNIP*) which were significantly associated with risk of type 2 diabetes ($P < 9.0\times10^{-8}$), including three that were previously known (DMRs in *TXNIP* and *CPT1A*), in 429 type 2 diabetes cases and controls. DNAm score consisted of these DMRs differentiated the risk of developing type 2 diabetes

in an independent prospective cohort with a relative risk of 2.86 (95% CI 1.10-7.44) between the lowest and highest deciles of DNAm score. DMRs in *CPT1A* and *TXNIP* were associated with quantitative metabolic traits, including fasting glucose, HbA1c, and body mass index. DMRs of type 2 diabetes have little methylation quantitative loci (mQTL), and not causally associated with type 2 diabetes. Three DMSs of type 2 diabetes (cg26823705, cg08867893, and cg17082373) affect the increase of HbA1c level. We also identified three DMRs (on *COMMD1*, *TMOD1*, and *FHOD1*) associated with DKD in 167 DKD cases and controls. DMRs of DKD did not show meaningful overlap with those of type 2 diabetes. DKD was causally affected by methylation changes of DMRs of DKD, which was highly influenced by genetics. In MR analysis, the estimated glomerular filtration rate was causally associated with DNAm of these three DMRs.

Conclusions: In an East Asian population, we identified eight DMRs, including five novel ones, associated with type 2 diabetes and three DMRs associated with DKD at methylome-wide statistical significance. Our findings suggest that epigenetic machinery of DKD may be different from that are responsible for the development of T2D.

KEYWORDS: Diabetic kidney disease; DNA methylation; Epigenome; MWAS; type 2 diabetes

Student number: 2018-35969

Contents

List of Tables

List of Figures

I. Introduction: Epigenetics, Epigenetic markers and Type 2 diabetes

1. Overview of epigenetics

1.1. Epigenetics and DNA methylation

DNA methylation (DNAm) is one of most widely studied and well-characterized epigenetic mechanisms, which was first suggested as long term memory function [1]. Epigenetic is heritable changes to the genome expression that do not affect underlying DNA sequences. Recently, development of next-generation sequencing and microarray technologies accelerated expanding the knowledge of epigenetics. Epigenetic mechanisms can be divided into three main categories: DNAm, histone modification and regulation by non-coding RNAs.

DNAm is epigenetic mechanisms involving the covalent transfer of a methyl group to the C-5 position of the cytosine ring of DNA and formation of 5-methylcytosine (5mC). DNAm is mostly (70-80%) found on cytosines followed by guanine residues (CpG) and is much less commonly observed at non-CpG region such as CpA, CpT, and CpC. CpGs tends to concentrate at specific clusters, called CpG islands. Most CpGs in promoter regions are unmethylated to facilitate binding between proteins and promoter DNA. Otherwise, CpGs located at gene body are likely to be hypo-methylated, which mad repetitive DNA elements silence. DNAm is regulated by a DNA methyltransferase (DNMTs) family (DNMT1, DNMT3). DNMT3 performs de-novo methylation during development, whereas DNMT1 acts to maintain methyltransferase during DNA synthesis.

1.2. Environmental exposures trigger epigenetic change

Epigenetic modification is predominantly influenced by environmental and lifestyle exposures. In monozygotic twins, epigenetic variation was smaller than that observed in population studies [2, 3].

There are two important features of epigenetic mechanisms; 'phenotypic plasticity' and 'stability'. Firstly, phenotypic plasticity is the ability of the genome to produce different phenotypes when exposed to environmental stimuli, which are potentially reversible. Several studies demonstrated that DNAm change is triggered by environmental exposure in a lifetime. Air pollution, toxic metabolites from bacteria, heavy metal, smoking, prenatal nutrition, and nonchemical stressor, which of them are strongly related to health outcome, were associated DNAm changes.

Once the DNAm level alters, these changes persist after initial exposures are disappeared. Epigenetic changes by biological, lifestyle and environmental factors are stably remembered by cells across the lifespan and generations. This is called 'stability'. For example, children whose mothers experienced Dutch famine during early pregnancy had a high risk of coronary heart disease and obesity compared to those not exposed to famine during the pregnancy period [4].

Considering these two characteristics, epigenetic markers have emerged as a possible biomarker for several environmental factors and health status. Furthermore, it can be used in risk prediction and early detection or prognosis of the disease.

1.3. DNA methylation profiling

The accelerated development of array and sequencing technologies has dramatically improved DNAm profiling, which made to understand the DNAm landscape at a genome-wide level. There are several DNAm profiling methods: restriction enzyme-based methods, affinity enrichment-based methods, and bisulfite sequencing method. Nevertheless, the bisulfite sequencing method is regarded as a gold standard in DNAm experiments, due to its qualitative, quantitative, and efficient approach [5].

The release of Infinium HumanMethylation Beadchip technology by Illumina [6] enables efficient and comprehensive DNAm profiling in humans. The technology involves the bisulfite conversion of genomic DNA and amplification, followed by the hybridization of the converted DNA to arrays containing predesigned probes to distinguish between methylated and unmethylated Cs.

Three are three different platforms in Illumina assay; HumanMethylation 27 (27K), HumanMethylation450 (450K), Methylation EPIC (EPIC) BeadChips. Each platform measure >27K, >450K, and >850K DNAm sites. For 27K, only Infinium I probe design was employed. Infinium II probe design was added from 450K release. As Infinium II probes show a reduced dynamic range of measured methylation values as compared with the Infinium I probes, it is necessary to apply proper normalization methods to reduce the dye biases.

The Infinium I probe assay uses two types of probes per one CpG locus: methylated allele (M) in the red channel and unmethylated allele (U) in the green channel. The Infinium II probe assay uses a single probe per one CpG locus with dual-color readout. The DNAm level is the ratio of the methylation intensity (M)

to the overall intensities (M+U) at a single CpG (Equation 1). Alpha is the regularization constant, which is usually set to 100.

1.4. Characteristic of DNA methylome data for Methylome-wide association study (MWAS)

A Methylome-wide association study (MWAS) is an examination of a genome-wide set of DNAm markers with a phenotype of interest, in parallel to genome-wide association studies (GWAS) searching the association between genetic variants and a phenotype. In an MWAS, DNAm levels of CpGs are individually tested for association with an outcome of interest like an association test of single nucleotide polymorphisms (SNPs) and interesting traits in GWAS. As the study design of GWAS is at the stage of maturity, MWAS analysis heavily relies on approaches specifically designed for GWAS.

DNA methylome data, however, have considerably different characteristics compared to genome data. Methylation value is quantitative measures range from 0 to 1 which have a high possibility of confounding effects. The main sources of unexpected variability in DNA methylome data are technical bias including batch effects. In MWAS, batch effects are majorly from bisulfite conversion batch, experimental batch, chip, and position on-chip for array studies. To minimize batch effects, a proper study design that assures phenotype of interest is independent of expecting batch effects, optimizing laboratory procedures and statistical approaches to correct observed batch effects in the data is needed [7].

Normalization is the most important step in data preprocessing to remove technical biases. The major focus of the currently developed normalization method

is on Illumina's Infinium platforms, which are the most widely used platforms to profile DNAm. Illumina 450K and EPIC chip have two different probes (Infinium I and II) as previously described. The β-values obtained from Infinium I probes are more stable and reproducible than those from Infinium II probes. Thus, normalizing the data from two array types is needed before incorporating data. There are a few frequently used normalization methods: Peak-based correction (PBC), Subset-quantile within array normalization (SWAN), and beta mixture quantile (BMIQ) normalization. PBC rescales β of Infinium II probes to the same modes for distribution of β of Infinium I probes. SWAN is based on the assumption that the distribution of the β value is the same as the distribution of probes that have the same number of CpGs. BMIQ normalization adjusts the distribution of Infinium II probe's β values, using a three-state beta mixture models.

Another consideration in DNA methylome data is cellular heterogeneity. Differentially methylated signals are highly cell-type specific and can be partitioned into two categories: signals independent of cellular heterogeneity and signals caused by differential mixtures of cell types. The latter one is considered as a confounder to be removed in DNAm analysis. As DNAm profiles extracted from peripheral blood, which is most widely studied because of the convenience of data collection, have diverse subtypes of cells, accounting for cellular heterogeneity is the most important to have unbiased MWAS results. There is a reference-based cell-type deconvolution algorithm including the 'Housman' method and reference-free algorithm such as ReFACTor.

2. Epigenetics and type 2 diabetes

2.1. Epidemiology of Type 2 diabetes

Type 2 diabetes (type 2 diabetes) is the most common complex disease, characterized by hyper-glycemia due to impaired beta-cell function and insulin resistance. The prevalence of type 2 diabetes is estimated to be 9.3% of the adult population worldwide in 2019 and is projected to rise to 10.9% by 2030 [8]. In Korea, 12.4% of adults have type 2 diabetes in 2018 [9]. Therefore, type 2 diabetes has emerged as one of the biggest international health agendas.

Type 2 diabetes is associated with considerable morbidity and mortality. Though diabetes is likely to be underestimated as a cause of death due to the complications that lead to death, diabetes was the sixth leading cause of death both in the US and Korea in 2018 [10, 11]. type 2 diabetes patients have almost twice the risk of death than the disease-free population of similar age [8]. Cardiovascular disease (CVD) accounts for two-thirds of deaths of people with type 2 diabetes. CVD mortality in type 2 diabetes patients is 2 to 4 fold higher than in the normal population. Also, atherosclerosis and hypertension are common complications of type 2 diabetes.

Type 2 diabetes is the consequence of the interaction between unmodifiable and modifiable (lifestyle) factors. Genetics is the most representative unmodifiable factor. The other unmodifiable factors such as age over 45 [12], ethnicity, gestational diabetes, and low birth weights also contribute to an increased risk of type 2 diabetes. Modifiable or lifestyle risk factors include increased body mass index (BMI), sedentary lifestyle, hypertension, nutrition, and smoking. Obesity is one of the strongest risk factors in the development of type 2 diabetes [13].

2.2. DNAm changes in type 2 diabetes

Increasing evidence showed that type 2 diabetes arises from an interaction between genetic and non-genetic factors. To date, the numerous family-based study and genome-wide association studies (GWAS) of type 2 diabetes demonstrated the genetic influence on type 2 diabetes. The estimated heritability of type 2 diabetes range from 25 to 69% [14]. Currently, more than 560 genetic loci have been identified to be associated with type 2 diabetes through studies involving more than 1,400,000 participants [15]. Still, identified variants do not fully explain the expected heritability [16]. The role of epigenetics, thus, emerged to find 'missing heritability'. These epigenetic changes are considered to be a link between genetics and non-genetic factors.

Recent advances in DNAm analysis enabled genome-wide scale studies, investigating DNAm changes associated with the risk of type 2 diabetes.

The MWAS studies have found DNAm change associated with type 2 diabetes in pancreatic islets[17-19], skeletal muscle [20], adipose tissue [21], liver [22, 23], and peripheral blood [24-30]. To study the etiology of type 2 diabetes in principle, DNAm markers should be collected in type 2 diabetes-relevant tissue such as pancreatic islets, liver, adipose tissue, or skeletal muscle. However, disease-relevant tissue has a limitation to collect compared to peripheral blood due to its invasiveness. For this reason, MWAS using peripheral blood samples has an advantage in collecting a larger sample size. Besides, some studies [31, 32] suggest that epigenetic changes in DNA from peripheral blood could reflect the epigenetic change in disease-relevant tissues. There is also evidence that reliable results using peripheral blood could draw by investigating cell type composition

[33].

The recent MWAS of type 2 diabetes recently discovered significant DMR in *TXNIP [24-30]*, *SOC3* [25], *SREBF1 [24, 25]*, *ABCG1[24, 26]*, *PHOSPHO1[25]* and *CPTA1[24, 30]* in peripheral blood **(Table 1)**. Among them, *TXNIP*, *ABCG1*, and *SREBF1* are also reported in MWAS of incident type 2 diabetes. The DMR most commonly reported to be associated with type 2 diabetes, independent of BMI, is *TXNIP* (cg19693031). Hyperglycemia is one of the potential mechanisms driving hypo-methylation of *TXNIP*. In this region, type 2 diabetes cases are consistently hypo-methylated in diverse ethnicity. In addition, methylation of TXNIP is identified to be associated with fasting glucose and HbA1c, which are strongly related metabolic traits of type 2 diabetes. MWAS of type 2 diabetes demonstrated that DMRs of type 2 diabetes are not overlapped with previous GWAS loci of type 2 diabetes, suggesting that DNAm is the biological mechanism that is somewhat different from the genetic effect.

2.3. Metabolic memories of diabetic complication, DKD

The duration and severity of uncontrolled hyperglycemia are major determinants of diabetic microvascular complications. Clinical trials have suggested that uncontrolled hyperglycemia has long-term adverse consequences, which are mediated through metabolic memory [19]. Recent studies provided evidence that epigenetic changes play a key role in metabolic memory [20, 21]. DKD is one of the common microvascular complications that affects about 40% of diabetes patients and could result in end-stage renal disease. There are a few studies that investigated peripheral blood DNAm markers associated with DKD [34-36], but

still in the exploration stage.

2.4. Genetic influences on DMRs

DNAm can be influenced by genetic factors as well as environmental factors, meaning that careful study design in MWAS to discover disease-related DMRs is important to minimize the influence of possible genetic influences. Some studies identified the associations between SNPs and DNAm level across the genome [37-40]. Genetic variants may directly influence DNAm marks, through cis effects, or trans effects. These genetic associations to DNAm markers are called methylation quantitative loci (mQTL). Genetic variation causes about 24% of the variance in DNAm in childhood and 21% of the variance in middle age [41, 42]. DNAm profiles are more similar between families than unrelated individuals. The result of MWAS should be interpreted carefully considering the impact of genetic variation. In the study of genetic influences on type 2 diabetes methylation loci in the UK, 118 associations between known type 2 diabetes SNPs were observed. Also, a further 226 mQTLs were identified between independent non–type 2 diabetes SNPs and CpGs [43].

Table 1. Reported CpG associated with type 2 diabetes

Study	Population	Differentially methylated site
Chambers et al (2015)	Indian-Asians Europeans	ABCG1(cg06500161), SREBF1 (cg11024682), TXNIP (cg19693031), PHOSPHO1 (cg02650017), SOCS3 (cg18181703)
Cardona et al (2019)	Europeans (UK)	TXNIP (cg19693031), ABCG1 (cg06500161), PHGDH (cg14476101), SLC9A3R1 (cg14020176), SREBF1 (cg11024682), CPT1A (cg00574958), SLC1A5 (cg02711608)
Kulkarini et al (2015)	Mexican-Americans	ABCG1(cg06500161), SAMD12 (cg07960624) , TXNIP (cg19693031)
Soriano-Tarraga et al (2016)	Spanish	TXNIP (cg19693031), POR (cg01676795), PFKFB3 (cg08994060)
Florath et al (2016)	German	RNU5E (cg13612055), NFE2L3 (cg10536999), TXNIP (cg19693031)
Al Muftah et al (2016)	Arab Caucasian	DQX1 (cg06721411), TXNIP (cg19693031)
Meeks et al (2018)	African (Ghana)	TXNIP (19693031), C7orf50 (cg04816311), CPT1A (cg00574958), TPM4 (cg07988171)

3. Imputation of DNA methylation markers

3.1. Review of DNA methylation imputation methods

With the development of the DNAm array technique, there were dramatic changes in the available number of CpG sites over the generation of the platform of the DNAm array. New platforms have replaced the old platforms every few years. As we discussed above, a different platform such as Illumina HumanMethylation 27K, 450K, and EPIC beadchips cover a different number of CpG sites. In this background, researchers have difficulty in integrating data generated from different platforms.

Imputation is the method of replacing missing data, which is broadly used in the process of genotype data quality control. Missing values are imputed by using the information we already know in a given algorithm. Though the imputation method of methylome-data is in the early stage, a few methods are developed based on linear regression and machine learning algorithms. The existing imputation method of DNAm markers is compared in **Table 2.**

'DeepCpG'[44] is a computational approach based on deep neural networks, which combines convolution neural network (CNN) and gated recurrent unit (GRU) algorithm using methylation and genotype data. It was an early model of DNAm markers' imputation and has a limitation with categorical predicted methylation value of 0 (un-methylated) or 1 (fully-methylated).

The rationale of the 'methyLImp'[45] is based on the high degree of inter-sample correlation in methylation data. Methylation level shows inter-sample correlation with both short- and long-range [46, 47]. Imputation accuracy of 'methyLImp' was considerable, but that in disease samples was significantly lower than in

normal samples. In addition, to apply ‘methyLImp’, a reference dataset is essential. ‘MRCNN’[48] is not originally developed for imputation CpG markers, but for predicting CpG markers from genotype data based on the convolutional neural network (CNN) algorithm. Given that the presence of mQTL (association between genotype and methylation), however, ‘MRCNN’ could be utilized as an imputation method as long as genotype data is available.

Table 2. Comparison of DNAm marker imputation method

	deepCpG (2017)	methylImpt (2019)	MRCNN (2019)	MRM (2020)	CUE (2020)
Features	Combination of CNN and GRU using methylation and genotype data	Correlation matrix using methylation data only	CNN using genotype data	Stacked regression using subjects and regional models	Ensemble learning framework which leverages several machine learning algorithms
Accuracy	Accuracy 89%	RMSE 0.035	RMSE 0.1468~0.2815	RMSE 0.1379	RMSE 0.026
Disadvantages	Output: 0 or 1	Accuracy is significantly lower in disease than in normal samples	Deep learning model needs to be trained.	Data preprocessing is difficult.	Landmark CpGs are necessary.

Abbreviation:: CNN, Convolutional neural network; GRU, gate recurrent unit; RMSE, root mean square error

‘MRM (mixture regression model)’[49] uses radial basis function, integrating information of inter-sample correlation and intra-sample correlation between neighboring CpGs. Clustering algorithm between similar regions and similar populations is the main part of the ‘MRM’ method. If methylation data to be imputed has a complex subpopulation, using the ‘MRM’ method would be helpful due to the complex clustering algorithm although preprocessing datasets for clustering is somewhat complicated.

‘CUE (CpG impUtation Ensemble for DNA methylation)’[50] ensembles multiple classical statistical and machine learning methods, including logistic regression, k-nearest neighbors, random forest, penalized functional regression model, and XGBoost. This method is proposed for imputation CpG sites from 450K to EPIC chip and shows remarkable imputation performance. However, there is a strong limitation. If representative CpGs used for training of each ensemble method are lost during the QC procedure of the 450K dataset, the ‘CUE’ method was not applicable.

3.2. Imputation of DNA methylation population datasets

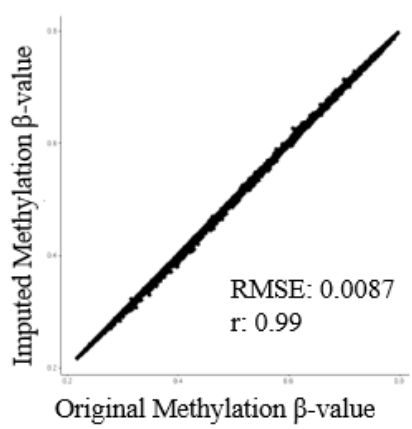
As described below, four types of datasets were available for the study (**Table 2**). Among them, three datasets were assayed by Illumina 450K platform, which needs imputation. As the ‘methyLImp’ method was the most intuitive way we first tried to impute missing markers using ‘methyLImp’ and evaluated imputation performance using HTS 1 dataset.

Table 3. Available datasets for the study

Dataset	N	Platform
Healthy twin study (HTS) 1	149	Illumina EPIC
HTS 2	390	Illumina 450K
Korean Genome and Epidemiology Study (KoGES)	400	Illumina 450K
Seoul National University Hospital Diabetes Clinic (SNUH) cohort	429	Illumina EPIC

First of all, HTS 1 dataset was divided into two parts: reference data and test data. For test data, DNAm markers only included in the EPIC platform were deleted. Next, imputation of deleted markers of test data using reference data was performed. To evaluate the imputation performance of 'methyLImp', the difference of original β-value from EPIC chip and imputed β-value of test set were compared. The root mean square error was 0.0087 and Pearson's correlation was 0.99 (**Figure 1**). As 'methyLImp' had shown remarkable imputation performance, remain datasets are imputed by 'methyLImp' using HTS 1 as reference data.

Figure 1. Scatter plot of methylation beta-value



4. Objectives

The aim of the study is to identify DNAm markers and their characteristics associated with type 2 diabetes and DKD. The workflow of the study is visualized in **Figure 2**.

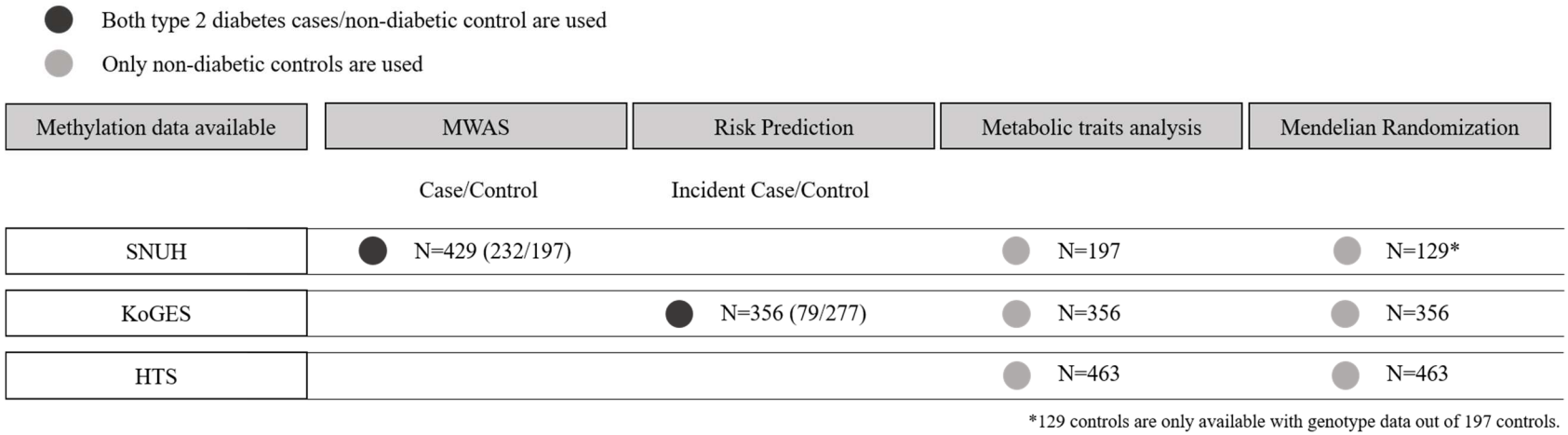
The objectives of the study are:

1) To identify DMSs associated with type 2 diabetes through MWAS and test prediction of type 2 diabetes incidence using the aggregated score of DMSs.

2) To identify DMSs associated with DKD through MWAS in a Korean population.

3) To explore the association and the causation of these DMRs with type 2 diabetes-related metabolic traits in non-diabetic controls.

Figure 2. Workflow of the study



II. MWAS of type 2 diabetes and Verifying markers

1. Material and methods

1.1. Study design and population

This was a case-control MWAS to investigate peripheral leukocyte DNAm markers associated with the development of T2D and DKD in the Korean population, which included 429 participants (232 T2D cases and 197 non-diabetic controls). By following the T2D patients' cohort, we conducted a nested case-control analysis for 87 DKD cases and 80 non-DKD diabetic controls. This study was conducted as part of the Seoul National University Hospital Diabetes Clinic (SNUH) cohort, which was initiated in January 2001 and is still ongoing [51]. To investigate the association of the identified DMSs with diabetes-related metabolic traits, two additional population-based cohorts were used: 1) the Korean Genome and Epidemiology Study (KoGES) consisting of 356 non-diabetic controls [52] and 2) the Healthy Twin Study, Korea (HTS), consisting of 463 non-diabetic controls [53]. To test the predictive performance of the DMSs in the development of incident T2D cases, prospective follow-up information on the KoGES was used (79 incident cases and 277 controls during 14 years of follow-up). KoGES and HTS cohorts were also used for quantitative metabolic traits analysis to examine whether metabolic traits were associated with the DNAm changes in identified DMSs.

Type 2 diabetes was diagnosed according to the 2020 American Diabetes Association guidelines [54]. Non-diabetic controls were enrolled if they were aged ≥ 60 years, had a fasting glucose < 5.5 (mmol/L), and HbA1c < 42 mmol/mol (6.0%). The Institutional Review Board (IRB) of the Biomedical Research Institute at Seoul National University Hospital (IRB No. 1603-079-749) and the

IRB of the Seoul National University (IRB No. 1902-003-015) approved the study protocol and written informed consent was obtained from each participant.

1.2. DNA methylation profiling

Genomic DNA was extracted from the peripheral blood cells of the participants immediately after sampling using the Gentra Puregene Blood Kit (Qiagen, Hilden, Germany). The DNA samples from the SNUH cohort were treated with sodium bisulfite using the Zymo EZ-96 DNA methylaton kit (Zymo Research, Irvine, CA, USA) and applied to the Illumina EPIC BeadChip (EPIC) (Illumina, San Diego, CA, USA). Samples from KoGES were assayed using 450K. The experiments for the SNUH cohort were done by Macrogen Inc. (Seoul, Republic of Korea) in three major batches between the years 2016 and 2018 as shown in **Figure 3**. The raw methylation intensity data were processed using R package RnBeads [55]. Methylation value was quantified using beta value, which represents the proportion of the methylated probe intensity of the total intensity, from 0 (fully un-methylated) to 1 (fully methylated). The procedures for quality control (QC) and normalization are detailed in **Figure 4**. The methylation beta value was normalized using the beta-mixture quantile normalization (BMIQ) method to reduce the difference between the two types of probes [56]. Cell type composition was calculated using the ReFACTor algorithm in the python toolset GLINT [57]. The batch effect adjustment process was done before cell type calculation to avoid misestimating of cell type by PCA from ReFactor algorithm. The analyses were performed using the R software (version 3.5.1) and Python 2.7.

1.3. Removing Batch effect

We thoroughly investigated the experimental batch effects and tried to eliminate any of the confounders. Removal of batch effect was confirmed by principle component analysis (PCA) and MWAS between controls. The detailed process of the experimental batch-effect correction for the SNUH cohort is described in **Figure 5**.

As our samples were recruited and extracted 3 times through 2016-2018, removing batch effect is an important part of quality control. We selected the best batch effect correction method using duplicated samples that experimented in a different year and tried to eliminate any confounders using the R package SVA [58].

We used the 'duplicate correlation' function in R package limma, which was originally designed to estimate the correlation between duplicate spots (regularly spaced replicate spots on the same array) or between technical replicates from a series of arrays. We adjusted each combination of experiment slide, experiment year, and experimented person and investigated duplicated correlation. The duplicate correlation function is mainly designed to find a tissue-specific difference of the individual, where both tissues experimented at the same time. Though getting duplicate correlation value between the samples experimented separately is not fitted for the function and therefore correlation value is not higher, we just used the correlation value as the reference of selecting the best batch effect correction combination. Only adjusting experiment year showed the best duplicate samples correlation. To ascertain that the batch effect was well removed, we verified it in two ways. First, principle component analysis (PCA) was applied to data before adjusted and data after adjusted. As seen in **Figure 6**, no longer batch effect was observed after adjusting batch effect correction.

1.4. MWAS of type 2 diabetes

MWAS was performed to identify the DMSs associated with type 2 diabetes. M-value was used in the statistical analysis to consider the heteroscedasticity and the assumption of Gaussian distribution. The M-value is calculated as the log2 ratio of the intensities of the methylated probe and unmethylated probe, which has known to be more statistically valid compared to the β-value. An empirical Bayes approach in the R package limma [59] was used to evaluate the association between methylation levels and type 2 diabetes status. The M-values were fitted using the model matrix with covariates including the experimental batch, cell-type composition, age, sex, and body mass index (BMI). Pairwise contrasts between cases and controls were made at each DMR after adjusting for confounding factors referred above. For statistical significance, we set a threshold that satisfied two conditions: uncorrected methylome-wide significance of $P < 9\times10^{-8}$ as previously suggested [60] and a false discovery rate (FDR) < 0.05. We tested whether the association of the previously reported DMSs of type 2 diabetes were replicated in our data set ($P < 0.05$). A total of 59 probes from the previous report were available after the QC procedures of DNAm profiling.

Secondly, we tried to find if there are differentially methylated sites between control groups from 2017 recruited samples and from 2018's We found cg09547764 was differentially methylated in 9×10^{-8} significance level. Reported CpG sites were searched in EWAS Atals [61]. cg09547767 was previously reported to be associated with hepatocellular carcinoma [62] not associated with type 2 diabetes or related traits. We concluded that the batch effect was successfully removed to conduct type 2 diabetes and DKD MWAS.

1.5. Prediction of type 2 diabetes using the DNAm score

We calculated the DMS-based methylation scores to predict incident type 2 diabetes as previously described [63]. The DNAm score was calculated as follows: DNAm score of i^{th} DMS = (beta value of i^{th} DMS – the median beta value of i^{th} DMS in the reference group) × coefficients of i^{th} DMS

The sum of the DNAm score consisted of eight significant DMSs, that were identified in our MWAS and used in the prediction model. Independent non-diabetic controls from HTS were used for the reference group. The risk of type 2 diabetes according to the DNAm score was analyzed with Cox proportional-hazards models for one standard deviation (SD) increase in the DNAm score. The risk was also compared between the top 10% and bottom 10% of the DNAm scores.

1.6. Prediction of type 2 diabetes prevalence using the DNA methylation

We also tried to predict the prevalence of type 2 diabetes using DNAm markers. Logistic regression and XGboost (eXtreme Gradient boosting) was implemented in modeling. The odds ratio for type 2 diabetes risk prediction by clinical variables and DNAm score was evaluated by logistic regressions (Model 1: Age, sex and BMI included, Model 2: Age, sex, BMI, and 8 CpG sites included). The performance of prediction evaluated by the area under the curve (AUC) and visualized by receiver operation characteristics (ROC) curve using the R package 'pROC'[64]. To avoid overfitting, we randomly divided our dataset as 80% training set and 20% test set with case-control stratification. Training set was used for model training and test set was used for performance evaluation. To get unbiased result and overcome the power limitation caused by small sample sizes, bootstrap method (1,000 replication) was applied and median AUC was displayed

as a representative value. Confidence interval was calculated through bootstrap method as suggested in elsewhere [65]. The analyses were performed using R software (version 3.6.0). XGboost, the one of supervised machine learning techniques, is a decision tree-based ensemble model. Additive learning with a small subset tree of input features is utilized in the XGboost method. Since logistic regression predicts outcomes based on a linear combination of features, individual features of the model could not be considered independently. As the importance of each CpG site in prediction models is thought to be different, we applied a decision tree-based training model. Integrated analysis of three cohorts in prediction models was not available due to possible batch effect, we only used the SNUH cohort in this analysis. Training and test set was spilt into 8:2. Test accuracy using two different models was compared with increasing depth of XGboost.

1.7. Causal analysis of DNA methylation to Type 2 diabetes

The Mendelian Randomization (MR) approach was implemented to investigate the causal effect of DNAm on type 2 diabetes. MR is a method to estimated unconfounded causal estimates like randomized clinical trials by using genetic components as an IV [66]. First of all, reported mQTLs of 8 CpGs on two public databases (BIOS QTL [67, 68] browser and mQTLdb [69]) were investigated. Next, GWAS was conducted through 13,491,044 genetic variants to find mQTL in the Korean population. As none of the genome-wide significant (5x10-8) me-QTL loci were discovered in GWAS of the Korean population, we investigated SNPs associated with each CpGs and not associated with type 2 diabetes at the nominal significance level ($P< 0.05$) (**Table 4**). For causality analysis, the GRS score, the sum of risk allele divided by the number of mQTLs for each CpGs, was used as an instrumental variable (IV). MR was conducted using a 2-stage least square

regression method (**Figure 7**); Step 1: investigate association from IV to methylation of each CpG, Step 2: investigate causal association from predicted methylation of each CpG by IV to risk of type 2 diabetes.

1.8. The phenotypic variance explained by DNA methylation

We evaluated the proportion of phenotypic variance explained by DNAm. By using the Sequential Oligogenic Linkage Analysis Routines (SOLAR), which is a tool to evaluate the heritability of phenotype implementing the variance component models, phenotypic variance of type 2 diabetes is divided into genetic variance, environmental variance, and covariate induced variance with variance decomposition method. In SOLAR, kinship matrices estimated by pedigree are utilized to calculate heritability. Information of pedigree is available in the HTS dataset, but not in others (KoGES and SNUH). To estimate genetic variance without pedigree data, we calculated identity-by-descent (IBD) based on genotype data using Plink 1.9 [70]. IBD of two cohorts was converted into a kinship matrix using 'IBDkin' software [71]. We investigate the heritability and the proportion of type 2 diabetes variance explained by the following covariates: 1) Age, sex and BMI 2) Sum of DNAm 3) Age, sex, BMI, and the sum of DNAm. A standardized M-value was used to calculate the sum of DNAm to ensure every CpG has an equal variance. For CpGs which showed a negative association with type 2 diabetes, a negative value (-1) was multiplied to make the sum of DNAm higher in type 2 diabetes patients. The analysis was performed in R package 'solarius' [72].

2. Results

2.1. Clinical characteristics of the study participants

The clinical characteristics of the study participants are presented in **Table 5.** A total of 429 individuals (232 type 2 diabetes cases and 197 non-diabetic controls) were included in our analysis to investigate the DMSs associated with type 2 diabetes. Non-diabetic controls were older and had lower levels of fasting glucose, HbA1c, and SBP compared to participants with diabetes. Additional two cohorts (KoGES and HTS) was used in prediction of type 2 diabetes and mQTL MR analysis. Clinical characteristics of two cohorts are presented in **Supplementary Table 1.** Samples from the KoGES, with 79 incident type 2 diabetes cases and 277 controls who were followed for 14 years, were used to evaluate the prediction of type 2 diabetes using the DNAm score.

2.2. Differentially methylated regions of type 2 diabetes

We investigated a total of 749,315 CpG sites for their association with type 2 diabetes in 232 cases and 197 control samples from the SNUH cohort. The volcano and QQ plots are shown in **Figure 8**. There were eight DMSs that were significantly associated with type 2 diabetes at $P < 9\times10\text{-}8$ and FDR < 0.05 after adjusting for age, sex, and BMI (**Table 6,** detailed in **Supplementary Table 2**). Among these, three were previously reported: cg26974062 and cg19693031 in thioredoxin interacting protein (TXNIP) and cg00574958 in carnitine palmitoyltransferase 1A (CPT1A). There were five DMSs that were newly discovered to be significantly associated with type 2 diabetes in our study: cg25139493 at bone morphogenetic protein 8A (BMP8A), cg26823705 at neuroblastoma breakpoint family member 20 (NBPF20), cg14530801 at syntaxin 18 (STX18), cg08867893 at zinc finger protein 365 (ZNF365), and cg17082373

at tripartite motif-containing 37 (TRIM37).

We examined whether previously known type 2 diabetes-associated DMSs were replicated in our Korean samples. A total of ten DMSs out of 59 sites described in the recent literature [24-29, 73] were replicated at nominal significance ($P < 0.05$) **(Table 7 and Supplementary Table 3**). For 7 DMSs out of these ten, the direction of methylation (hyper- or hypo-) in our study agreed with those in previous reports. We compared the previously reported effect size and replicated effect size in our Korean data. cg19693031 was constantly hypo-methylated in type 2 diabetes patients. cg06500161 and cg11024682 were constantly hyper-methylated. Our replication was consistent with previous studies **(Table 8 and Supplementary Figure 1)**. Although the effect of cg00574958 and cg04816311 was controversial in previous studies. The effect size from the EPIC-Norfolk study, which studies sample size was the biggest, was compared with the effect size from our MWAS (**Figure 9**). Even if the DMS from this study was not significant in the EPIC-Norfolk study, the effect size was highly consistent (Pearson's correlation =0.91).

2.3. Prediction of the risk of type 2 diabetes using the DNAm score

We tried to predict type 2 diabetes using the DNAm score in an independent validation set from the KoGES (79 incident type 2 diabetes cases and 277 non-diabetic controls). In the Cox proportional-hazards model, a one SD increase in DNAm score was associated with an increased risk of incident type 2 diabetes (risk ratio [RR] 1.33, 95% confidence interval [CI] 1.08–1.65).

For clinical variables, sex was not associated with a risk of incident type 2 diabetes ($P= 0.22$). As $1kg/m^2$ BMI increases 1.14 times risk of incident type 2 diabetes higher (HR 1.14, 95% CI 1.08-1.21). When participants with the highest decile of

DNAm score were compared with the lowest, the HR for T2D was 2.86 (95% CI 1.10-7.44) during 14 years of follow-up (**Figure 10**). The average time of reaching a 10% cumulative risk of T2D occurrence differed by about 8 years between the two groups (5.5 years and 13.6 years for the highest and lowest decile groups, respectively).

2.4. Prediction of prevalence of type 2 diabetes using the DNA methylation

Prediction accuracy of logistic regression was significantly higher (P<0.01) in the model where DNAm included in prediction variables. Results of logistic regression are presented in **Figure 11-12**. The highest test set accuracy using only clinical variables (Age, sex, and BMI) of XGboost algorithm was 72.09 in-depth 3 and the highest test set accuracy using both clinical variables and methylation profiles was 81.40% in-depth 7 (**Figure 13**). Training models with DNAm showed higher test accuracy during model depth increases from 1 to 14. When evaluating the feature importance of the model using F-score, it was found that the models including cg08867893 and cg17082373 improved the predictive performance more than the model including BMI in predicting diabetes (**Figure 14**). The plotting tree is presented in **Figure 15**. DNAm profiles help to predict type 2 diabetes for people over 59.

2.5. The causal effect of DNA methylation on type 2 diabetes

Some me-QTLs associated with two CpGs of type 2 diabetes (cg19693031 (TXNIP) and cg00574958 (CPT1A)) were found on public databases to be significant, but not replicated in the Korean population. We found some mQTLs with 8 CpGs of type 2 diabetes in nominal significance level (P<0.05). Among

them, mQTLs of 5 CpGs (cg26823705, cg19693031, cg1453080, cg08867893, and cg17082373) included SNPs reported to be associated with type 2 diabetes in previous GWAS **(Table 9)**. Also, some of them has relationship with HbA1c (rs12712928 [74, 75], rs1387153 [75] and rs10830963 [76, 77]), fasting glucose (rs780094 [75, 78-81], rs1387153 [82], rs10830964 [75]), serum metabolite measurement (rs780094 [83], rs1387153 [83, 84] and rs10830963 [83]), triglyceride level and glomerular filtration level (rs1260326 [85, 86] and rs780094[87]). GRS from selected mQTLs was strongly associated with each type 2 diabetes DMSs **(Table 10)**. In MR analysis, none of CpG showed causal association to type 2 diabetes in both reported type 2 diabetes SNPs included and excluded analysis **(Table 11)**.

2.6. The phenotypic variance explained by DNA methylation

Heritability of type 2 diabetes arranges 64-69%, which was the consistent result with previous studies (**Table 12**). The variance of age, sex, and BMI explained 7.39% of type 2 diabetes and the sum of DNAm standardized M-value accounted for 14.31 of the variance of type 2 diabetes. When both clinical variables and DNAm were considered, a total of 16.2% of type 2 diabetes variance was explained.

3. Discussion

In this study, we conducted a MWAS to identify the DMSs associated with type 2 diabetes in a Korean population. We validated the association of seven previously known DMSs with type 2 diabetes and identified five novel DMSs to be associated with type 2 diabetes. To the best of our knowledge, this is the first MWAS to investigate the association of DMSs with the risk of type 2 diabetes in an East

Asian population.

A number of MWASs have been reported for type 2 diabetes since 2015, and DMSs in *ABCG1*, *CPT1A*, *PHOSPHO1*, *SOCS3*, *SREBF1*, and *TXNIP* have been well-replicated in Europeans [24-29, 73, 88]. In line with these findings, we showed that DMSs in *ABCG1*, *CPT1A*, *PHOSPHO1*, *SREBF1*, and *TXNIP* were also associated with type 2 diabetes in the Korean population. This finding suggests that there are common DNAm marks that are consistently associated with type 2 diabetes across different ethnic groups. Although we have identified five novel type 2 diabetes DMSs in Koreans, we were not able to find an independent sample set of East Asians to replicate this finding. However, at least one DMS, cg25139493 at *BMP8A*, showed a nominally significant association with diabetes in the EPIC-Norfolk study (P = 0.033) with the same direction. Lack of an association in our study for some of the DMSs discovered in Europeans might be either due to a lack of a true association or the moderate sample size of this study. Among the commonly identified DMSs, *cg10508317* in *SOCS3* had a P of 0.65 in our study, suggesting that the moderate sample size alone cannot explain the lack of association.

There were two DMSs in *TXNIP* that were significantly associated with type 2 diabetes, and these two associations were independent when the conditional analysis was performed for each other. The DMSs in *TXNIP* have been most extensively investigated and are known to be associated with prevalent [26, 27, 29, 73] and even incident [24, 25] type 2 diabetes.

Among the five novel DMSs associated with type 2 diabetes, none of them were associated with gene expression in whole blood (the Biobank-based Integrative Omics Study database, https://genenetwork.nl/biosqtlbrowser/) [68]. However,

several DMSs in these newly identified genes were reported to have roles in the pathogenesis of diabetes. *BMP8A* encodes a secreted ligand of the TGF-β superfamily of proteins, and it resides in a type 2 diabetes GWAS region. As an intronic variant of this gene, rs61779331 is significantly associated with the risk of diabetes ($P = 6.91\times10^{-25}$) [89]. *BMP8A* is expressed in white adipose tissue and has a role in lipid metabolism in zebrafish [90, 91]. The role of *NBPF20* in diabetes is not understood yet. However, there are reports that its expressions in skeletal muscle and adipose tissue are decreased in women with polycystic ovary syndrome, where insulin resistance is increased [21, 92]. Skeletal muscle expression of *NBFP20* was also correlated with decreased C-peptide levels in these women. *STX18* may be involved in the fusion of Golgi-derived transport vesicles with the endoplasmic reticulum, and an intronic variant, rs532717716 in this gene is significantly associated with the waist-hip ratio ($P = 1.46\times10^{-8}$) [93]. *TRIM* encodes a member of the tripartite motif family with E3 ubiquitin-ligase activity, and a mutation of this gene is known to cause muscle-liver-brain-eye nanism (MIM 253250) [94]. However, the role of these five DMSs in the development of type 2 diabetes remains speculative.

As obesity-induced insulin resistance results in T2D incidence, there is a significant difference in BMI between T2D patients and controls. Several DNAm markers were reported to be associated with BMI [95-97] and there are few overlapped CpGs with DMSs of T2D. Few overlapped CpGs suggest that different mechanisms are affecting DNAm of T2D patients, which is distinguished from the effect of BMI. In this study, we focused on differentially methylated sites caused by glucose variation and tried to rule out the effect of BMI on DNAm by adjusting BMI as a covariate. DMSs from MWAS not adjusted for BMI was as same as in

MWAS adjusted for BMI (**Supplementary Table 4**).

As seen in MR analysis, methylation changes themselves do not affect the incidence of type 2 diabetes. Null results of MR implies that the other risk factor exists which both affect DNAm changes and type 2 diabetes. Also, we explored whether discovered mQTL SNPs are associated with type 2 diabetes. Seven SNPs have a pleiotropic effect on DNAm and type 2 diabetes. All SNPs except rs147538848 have an association with metabolic traits also. Especially, mQTL of cg19693031 (rs1260356 and rs780094) which is the most frequently reported CpG of type 2 diabetes across ethnicity, is broadly associated with diverse metabolic traits (Serum metabolites, triglyceride, fasting glucose, fasting blood insulin, and leptin measurement). This finding suggests that the consistent strong association of cg19693031 and type 2 diabetes is partly due to mQTL with pleiotropy effect or GWAS signals are confounded by mQTL effect on DMSs. As we could not conclude in this analysis, a pathway from SNPs to type 2 diabetes via DNAm need further investigation.

There are several studies about the DNAm difference between ethnic groups [98-100]. As DNAm patterns reflect social and environmental factors, the genetic and cultural differences between diverse ethnic groups might lead to different DNAm profiles. Therefore, discovering East-Asian specific DNAm loci would contribute to understanding the risk factor of type 2 diabetes in the East-Asian population.

This study has several strengths. First, findings from East Asians with different epidemiologic profiles may add to the existing body of evidence that shows the role of epigenetics in the development of type 2 diabetes. Second, this study attempted to compare the DNAm profiles encompassing incident type 2 diabetes cases and to provide insight into the roles in its etiology and complication.

There are several limitations to our study. First, the sample size of this study was modest for MWAS, and we were not able to replicate our novel findings in an

independent sample of East Asians. Still, top signals were consistent with previous reports from Europeans and served as a positive control for our findings. Second, because the sample collections and experiments were performed at three different time points, there is a possibility of batch effects. However, we have tried extensively to adjust for potential batch effects and confirmed that the signals between compatible controls only showed white noise, and the association P-value distribution shows that our results were well-calibrated. It is not likely that our findings stem from the possible batch effects. Third, our DNAm data were measured using peripheral blood and not using disease-relevant tissues, such as the pancreas, skeletal muscle, and adipose tissue. It is well known that DNAm is considerably tissue-dependent [101-103]. Blood-based DNA has low specificity compared to that from disease-relevant tissue because it is a mixture of DNA from various cells [104, 105]. As a result, DNAm analysis with peripheral blood extracted DNA might lead to false-positive results. Nevertheless, collecting DNA from peripheral blood is more accessible due to the better compliance of patients. Other disease-relevant tissues are only able to collect invasively. Some studies suggest that epigenetic changes in DNA from peripheral blood could reflect epigenetic changes in tissues [31, 32]. Besides, there is another research that blood-based DNA could be used as informative biomarkers after considering cell type composition [33]. We calculated cell type composition by statistical method and added it to covariates to minimize the confounding effect from cellular heterogeneity. Forth, the bisulfite conversion method was applied in DNA conversion. Bisulfite conversion deaminates every single unmethylated cytosine to uracil. There is a possibility of misinterpretation of unmethylated cytosine as methylated cytosine due to incomplete conversion, leading to false-positive results for MWAS [106]. Fifth, not enough the number of samples are used in the prediction model of type 2 diabetes prevalence. Also, the same samples are used

as in MWAS analysis which might lead to the winner's curse. For reliable results, the prediction model should be validated in larger samples.

In an East Asian population, we identified five new DMSs associated with type 2 diabetes DKD with methylome-wide statistical significance. In addition, seven type 2 diabetes DMSs reported predominantly in a European population were replicated in our study. We hope our findings provide a better understanding of the pathophysiology of type 2 diabetes and enhance the prediction of type 2 diabetes.

Figure 3. Visualization of data recruitment of SNUH

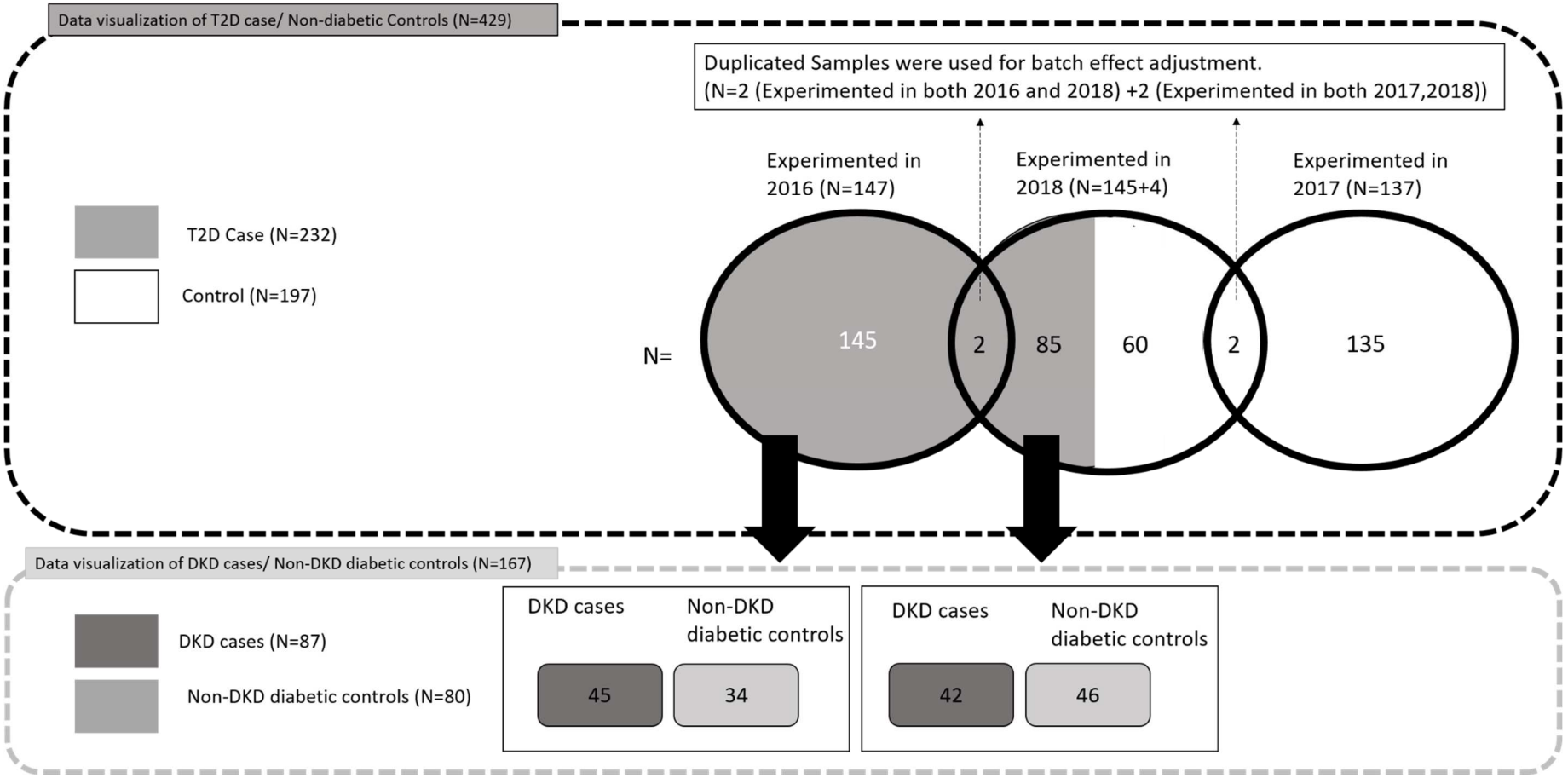


Figure 4. QC procedure of DNAm data for SNUH and KoGES

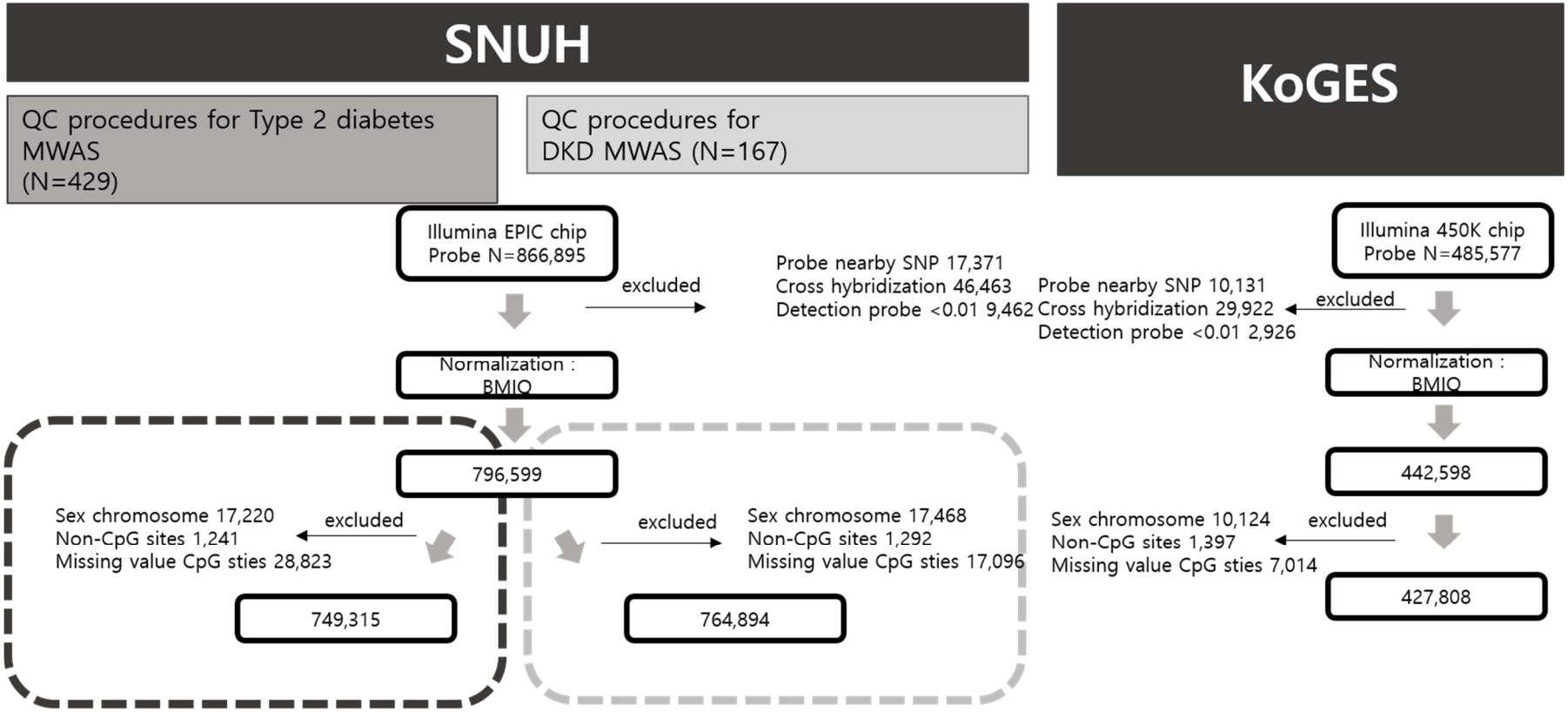
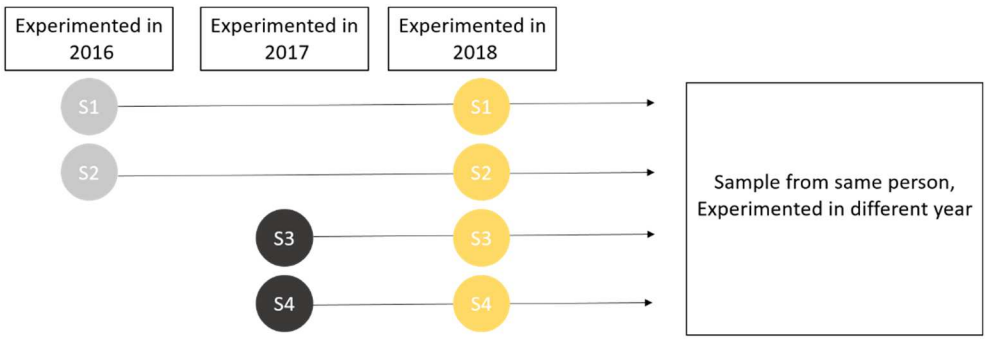


Figure 5. Process of batch effect correction



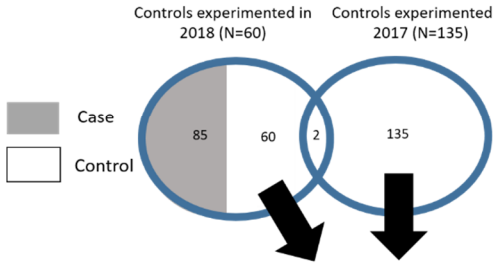
1. Duplicate Correlation analysis
Using different batch correction combination
(Experiment slide number (A), Experiment year (B), person who experimented (C))

Batch Combination	no batch	A	B	A+B	A+B+C	B+A	B+A+C	B+C	B+C+A	C	C+B	C+B+A	C+A	C+A+B
Duplicate correlation	0.213	0.094	0.262	0.116	0.113	0.166	0.163	0.262	0.166	0.226	0.261	0.166	0.132	0.138

2. Selected Best batch effect correction

3. Principle Component Analysis (PCA)

3. EWAS using only controls from different experiment year
Phenotype : experiment year



If there is differentially methylated CpG sites due to experiment batch effect

Cell count, age sex and BMI was used for covariates

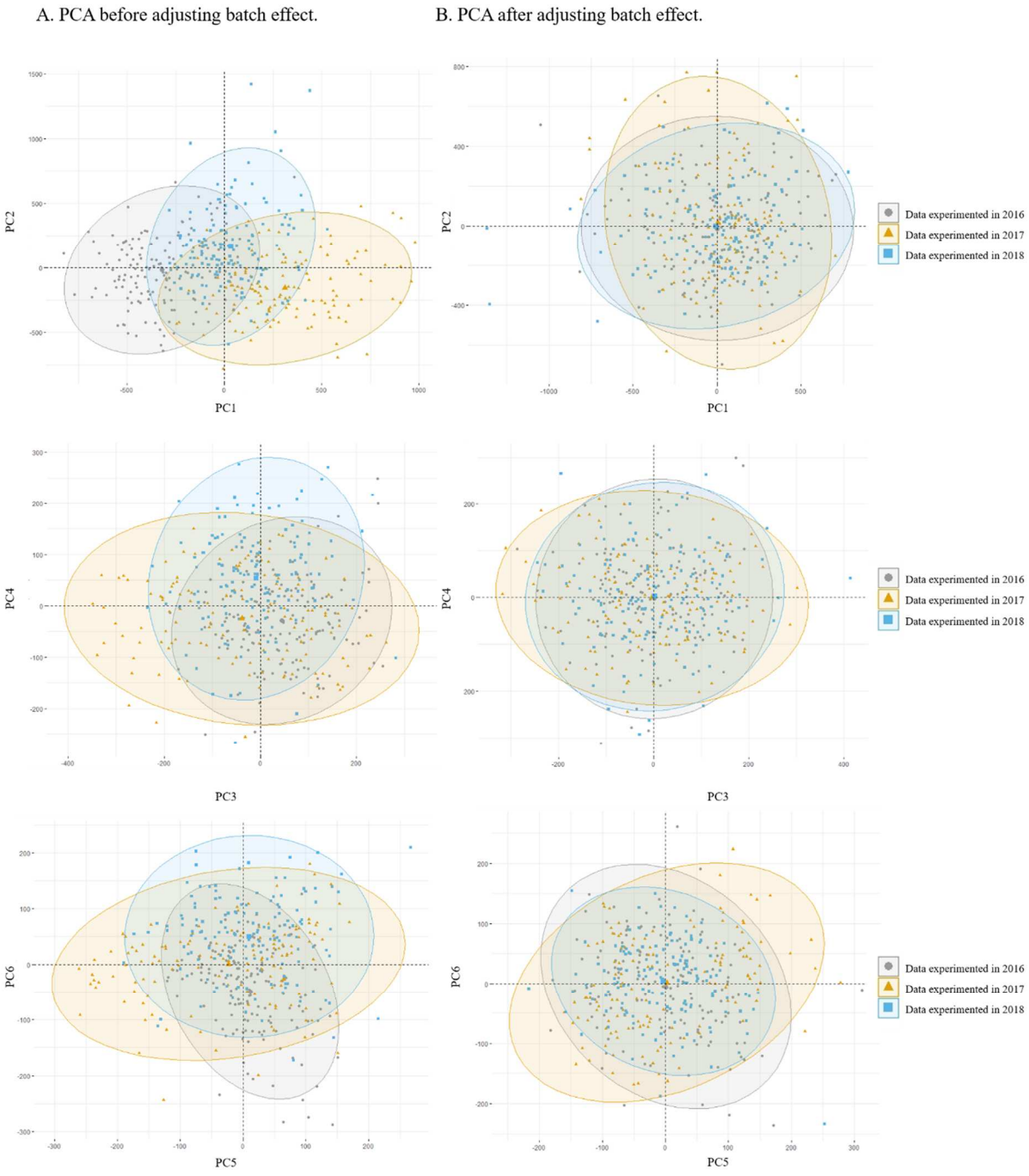
4. Differentially methylated at statistical significant level 9×10^{-8}

	Position	Coefficients	P-value	Nearby gene
cg09547767	Chr 1: 39957387	0.015	8.65E-08	BMP8A

Cg09547764 was reported to be associated with hepatocellular carcinoma according to EWAS Atlas.

Probe ID	Studies	Correlations	Location	Related genes (transcript: location)	CpG islands	Related traits
cg09547767	1 +		chr1: 39957387 +	BMP8A (ENST00000331593: promoter)	Island	hepatocellular carcinoma (HCC)

Figure 6. Principle component analysis of methylation value



*Individual experimented in same year clustered together.

Table 4. mQTL SNPs of type 2 diabetes DMSs

SNP	Position	RA	Effect size	*P*
		cg25139493		
rs7776054	chr6:135097778	G	0.0058	0.023
rs9399137	chr6:135097880	C	0.0058	0.023
rs6975024	chr7:44192287	T	0.00677	0.0413
rs9933309	chr16:88778524	C	0.00509	0.0357
		cg26823705		
rs2779116	chr1:158615625	C	0.0246	0.00382
rs12712928	chr2:44964941	G	0.0163	0.0473
rs7616006	chr3:12226148	G	0.0176	0.0412
rs13134327	chr4:143738642	G	0.0176	0.0332
rs17140821	chr7:19177581	A	0.0166	0.0453
rs6980507	chr8:42527944	G	0.0212	0.00329
rs8067360	chr17:82736950	T	0.0192	0.0299
		cg26974062		
rs9818758	chr1:rs9818758	G	0.0154	0.0332
rs7772603	chr2:rs7772603	A	0.0121	0.0189
rs6980507	chr6:rs6980507	G	0.01	0.0208
rs10823343	chr7:rs10823343	A	0.0112	0.0499
rs12819124	chr9:rs12819124	C	0.0152	0.0381
rs1203936	chr20:22592430	G	2.02	0.046
		cg19693031		
rs1260326	chr2:27508073	C	0.021	0.0165
rs780094	chr2:27518370	C	0.0211	0.00726
rs10513537	chr3:159187070	A	0.0258	0.0303
rs6533530	chr4:110810780	T	0.0274	0.0472
rs7776054	chr6:135097778	A	0.0211	0.0197
rs6980507	chr8:42527944	G	0.0173	0.0455
rs11619319	chr13:27913462	A	0.0197	0.022
		cg14530801		
rs11964178	chr6:109240832	A	0.00174	0.0121

rs17168486	chr7:14858657	T	0.00148	0.0119
rs6980507	chr8:42527944	G	0.00114	0.0377
rs7851507	chr9:136434270	T	0.00135	0.0487
rs12579302	chr12:89656726	A	0.00122	0.0383
cg08867893				
rs17140821	chr7:19177581	G	0.00459	0.0494
rs1387153	chr11:92940662	T	0.00476	0.0325
rs10830963	chr11:92975544	G	0.00488	0.0311
rs1447352	chr11:92989595	A	0.00722	0.00451
rs9933309	chr16:88778524	C	0.0044	0.0364
rs8067360	chr17:82736950	C	0.00534	0.0341
cg00574958				
rs7776054	chr6:135097778	A	0.00408	0.0461
rs10244051	chr7:15024208	T	0.00478	0.0415
rs10486607	chr7:29144873	C	0.0176	0.00824
cg17082373				
rs7851507	chr9:136434270	T	0.00215	0.00594
rs147538848	chr12:31313679	A	0.00301	0.0022
rs11667918	chr19:17121689	T	0.00132	0.0455
rs11086054	chr19:17135927	T	0.00152	0.0225

Figure 7. Diagram of MR using mQTL as an IV

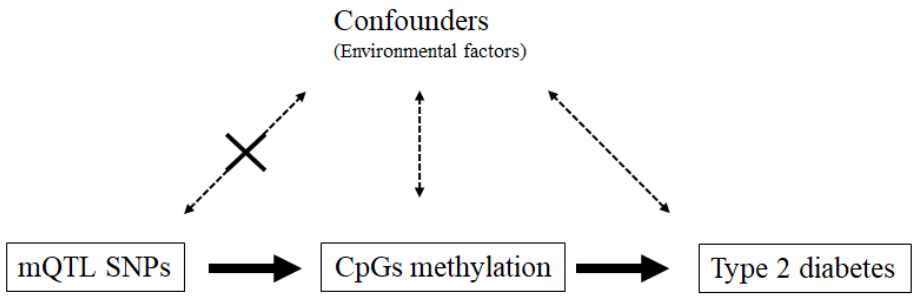


Table 5. Clinical characteristic of study participants (SNUH)

	SNUH (N=429)		
	Type 2 diabetes cases (N=232)	Non-diabetic controls (N=197)	*P* for difference
	Mean±SD/ N (%)	Mean±SD/ N (%)	
Age	61.5±10.1	63.9±3.3	<0.001
Sex			
Male	128 (55.2)	104 (52.8)	0.358
BMI (kg/m2)	24.4±2.9	23.7±2.8	0.011
Fasting glucose (mmol/L)	8.3 ±3.1	5.1±0.44	<0.001
HbA1c (mmol/mol)	66 ±17	38±2.8	<0.001
HbA1c (%)	7.9±1.6	5.4±0.25	<0.001
SBP (mg/dL)	132 ±20	119±16	<0.001
eGFR (ml/min)	71 ±31	86±15	<0.001
CKD Stage			<0.001
Stage 1	69 (29.7)	72 (36.6)	
Stage 2	96 (41.4)	119 (60.4)	
Stage 3	37 (16.0)	5 (2.5)	
Stage 4	14 (6.0)	1 (0.5)	
Stage 5	16 (6.9)	0 (0)	
Duration of type 2 diabetes (years)	13.4 (8.9)	-	-

Data are presented as mean ± standard deviation or n (%)

Statistical difference analysis was done with T-test for continuous variables and Chi-square test

Abbreviation: BMI, Body mass index; eGFR, estimated Glomerular Filtration Rate; HbA1c, Glycated hemoglobin; SBP, Systolic blood pressure

Table 6. Differential methylated CpGs for type 2 diabetes

CpG	Nearby gene	Position (hg19)	CGI information	Gene property	Adjusted M-values differences (β-values)	*P* *
cg25139493	*BMP8A*	chr1:39957400	Island	5'UTR	0.340 (0.013)	6.89×10^{-8}
cg26823705	*NBPF20*	chr1:145435523	Open Sea	Body	-0.277 (-0.047)	5.67×10^{-9}
cg26974062	*TXNIP*	chr1:145440734	Open Sea	Body	-0.500 (-0.031)	9.64×10^{-11}
cg19693031	*TXNIP*	chr1:145441552	Open Sea	3'UTR	-0.465 (-0.061)	2.65×10^{-9}
cg14530801	*STX18*	chr4:4544016	Island	TSS1500	0.755 (0.001)	2.12×10^{-9}
cg08867893	*ZNF365*	chr10:64134160	Island	5'UTR	0.454 (0.009)	2.13×10^{-8}
cg00574958	*CPT1A*	chr11:68607622	North shore	5'UTR	-0.373 (-0.016)	5.40×10^{-10}
cg17082373	*TRIM37*	chr17:57184450	South shore	TSS200	-0.482 (-0.017)	4.56×10^{-8}

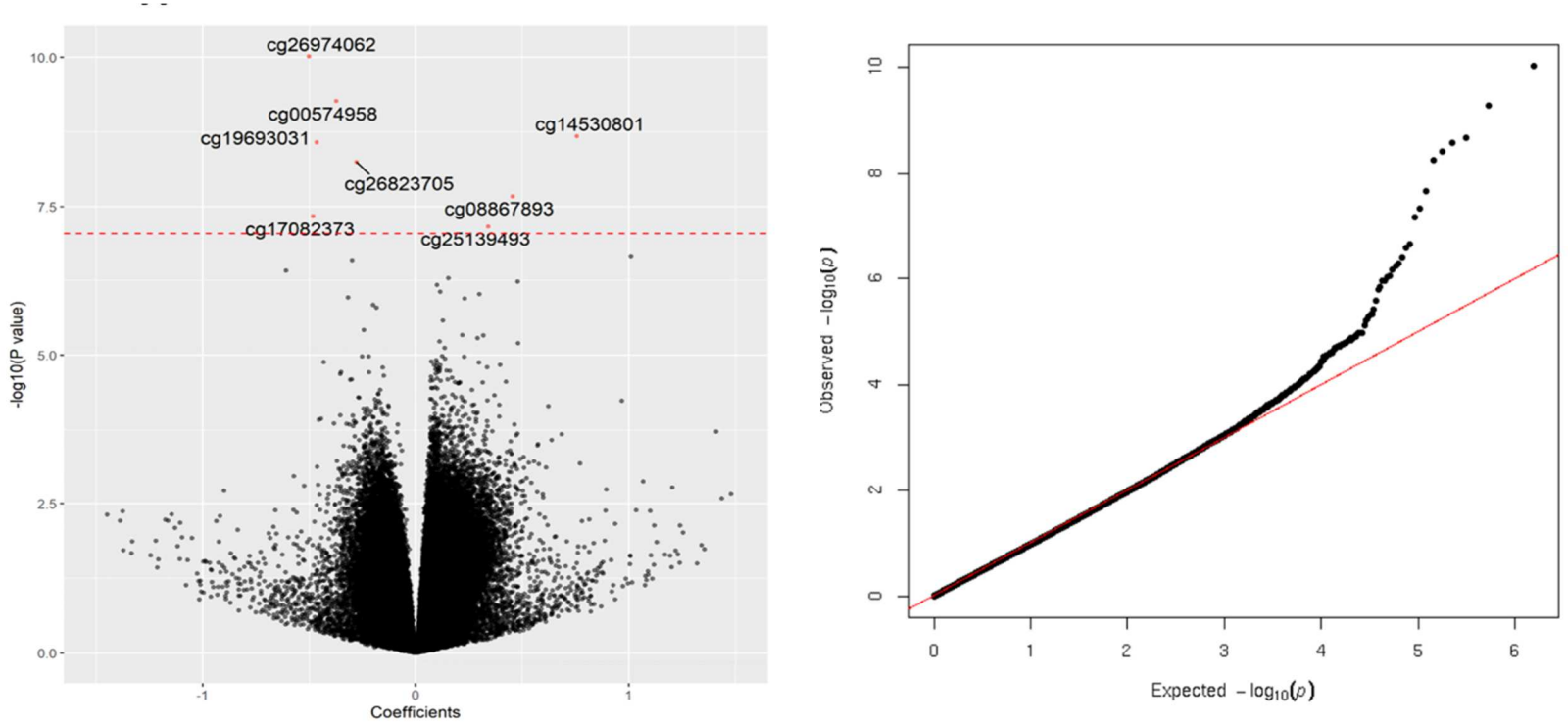
TSS200 is the region from TSS to 200 base pairs upstream of TSS; TSS1500 covers 200 to 1500 base pairs upstream of TSS.

Bold style font indicates novel CpG sites, which were not reported to be associated with type 2 diabetes previously.

**P* corresponding to M-values

Abbreviation: chr, chromosome; CGI, CpG island; hg19, human genome build 19; TSS, Transcription start site; UTR, Untranslated Region

Figure 8. Volcano plot and QQplot of type 2 diabetes MWAS



* Red dashed line indicates the significant level ($P < 9x10^{-8}$)

Table 7. Replication of reported type 2 diabetes associated differential methylated CpGs

CpG	Nearby gene	Position (hg19)	Adjusted M-values differences (β-values)	*P* *	Reported effect size † (*P*) ‡	Reported population	Reference
cg19693031	*TXNIP*	chr1:145441552	-0.464 (-0.061)	2.65×10^{-9}	-0.08 (1.00×10^{-13})	AMR, SAS, EUR,	[24,25,26,27,28]
cg19266329	*POLR3GL (-108)* [a]	chr1:145456128	-0.136 (-0.023)	0.001	0.50 (9.98×10^{-11})	AMR	[**26**]
cg00574958	*CPT1A*	chr11:68607622	-0.373 (-0.016)	5.40×10^{-10}	-0.37 (5.20×10^{-9})	AMR, EUR	[**24**,26]
cg17058475	*CPT1A*	chr11:68607737	-0.318 (-0.018)	1.10×10^{-6}	-0.37 (3.89×10^{-9})	AMR	[**26**]
cg14597545	*ADPGK*	chr15:73074210	0.068 (0.012)	0.013	-0.69 (1.50×10^{-8})	AMR	[**26**]
cg11024682	*SREBF1*	chr17:17730094	0.095 (0.016)	2.89×10^{-4}	0.06 (8.40×10^{-9})	SAS, EUR	[24,**25**]
cg02650017	*PHOSPHO1*	chr17:47301614	-0.175 (-0.005)	2.83×10^{-4}	-0.06 (2.10×10^{-9})	SAS, EUR	[**25**]
cg26836479	*DEDD2*	chr19:42706353	-0.121 (-0.004)	0.007	0.56 (1.05×10^{-7})	AMR	[**26**]
cg06500161	*ABCG1*	chr21:43656587	0.140 (0.024)	2.65×10^{-9}	0.08 (2.20×10^{-13})	AMR, SAS, EUR	[24,**25**,26]
cg04816311	*C7orf50*	chr7:1066650	0.091 (0.016)	0.018	0.41 ($1.70\text{x}10^{-8}$)	AMR, EUR	[**24**,26]

Population: AMR (Ad mixed American), SAS (South Asian), EUR (European)
Bold style font indicates the reference of reported effect size column
**P* corresponding to M-values

†Reported odds ratio (or risk ratio) was transformed to effect size.

‡Reported effect sizes and *P* were selected on the reference with the largest sample sizes.

[a] Unlike other CpG loci, this locus is not located in the gene. The nearest gene was displayed with the distance. Distance is expressed in kb.
Abbreviation: chr, chromosome; hg19, human genome build 19

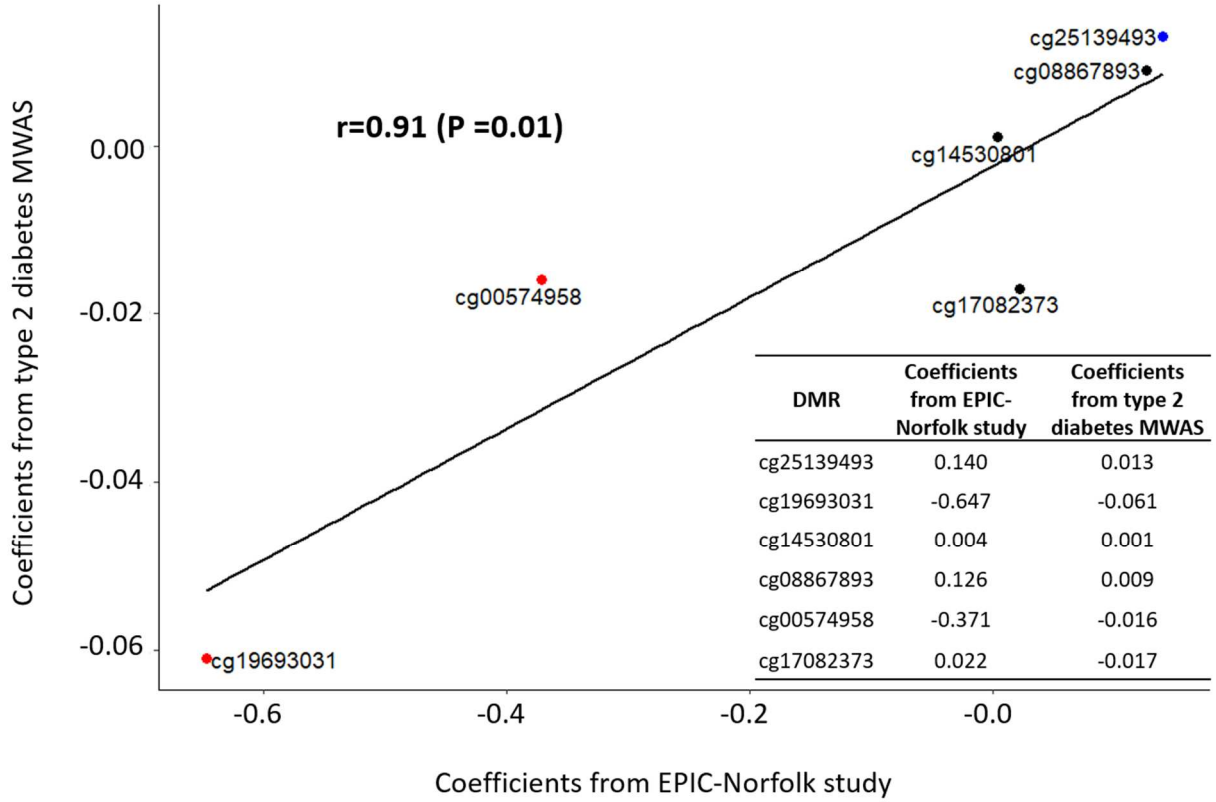
Table 8. Comparison of reported and replicated effect size of CpGs reported multiple times

CpG	Reference [a]	Reported effect size [b]	Reported *P*	Effect size in this study (β-values)	*P* in this study
cg19693031	[26]	-0.58	$1.53x10-^{19}$	-0.061	$2.65x10^{-9}$
cg19693031	[25]	-0.08	$1.00x10^{-13}$	-0.061	$2.65x10^{-9}$
cg19693031	[24]	-0.65	$2.70x10^{-21}$	-0.061	$2.65x10^{-9}$
cg19693031	[28]	-0.03	$4.50x10^{-7}$	-0.061	$2.65x10^{-9}$
cg19693031	[27]	-0.13	$7.30x10-^{16}$	-0.061	$2.65x10^{-9}$
cg19266329	[26]	0.50	$9.98x10^{-11}$	-0.023	0.001
cg00574958	[26]	0.38	$4.29x10^{-9}$	-0.016	$5.40x10^{-10}$
cg00574958	[24]	-0.37	$5.20x10^{-9}$	-0.016	$5.40x10-^{10}$
cg17058475	[26]	-0.37	$3.89x10^{-9}$	-0.018	$1.10x10^{-6}$
cg14597545	[26]	-0.69	$1.50x10^{-8}$	0.012	0.013
cg11024682	[25]	0.06	$8.40x10^{-9}$	0.016	$2.89x10^{-4}$
cg11024682	[24]	0.44	$6.00x10^{-10}$	0.016	$2.89x10^{-4}$
cg02650017	[25]	-0.06	$2.10x10^{-9}$	-0.005	$2.83x10^{-4}$
cg26836479	[26]	0.56	$1.05x10^{-7}$	-0.004	0.007
cg06500161	[26]	0.39	$9.43x10^{-10}$	0.024	$2.65x10^{-9}$
cg06500161	[25]	0.08	$2.20x10^{-13}$	0.024	$2.65x10^{-9}$
cg06500161	[24]	0.50	$6.40x10^{-14}$	0.024	$2.65x10^{-9}$
cg04816311	[26]	-0.50	$5.47x10^{-8}$	0.016	0.018
cg04816311	[24]	0.41	$1.70x10^{-8}$	0.016	0.018

[a] Reference number is based on the main text.

[b] Reported odds ratio (or risk ratio) was transformed to effect size.

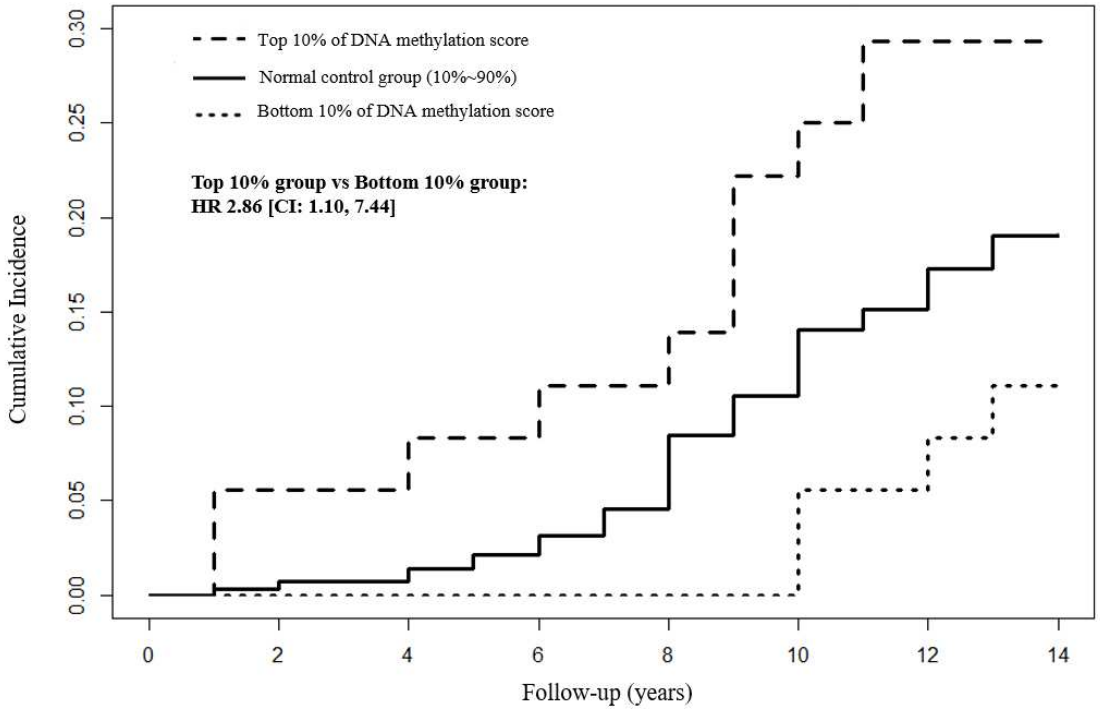
Figure 9. Correlation between results from EPIC-Norfolk study and type 2 diabetes MWAS



DMR	Coefficients from EPIC-Norfolk study	Coefficients from type 2 diabetes MWAS
cg25139493	0.140	0.013
cg19693031	-0.647	-0.061
cg14530801	0.004	0.001
cg08867893	0.126	0.009
cg00574958	-0.371	-0.016
cg17082373	0.022	-0.017

The red dot indicates that P from the EPIC-Norfolk study is under methylome-wide significance level ($P < 9x10^{-8}$); the blue dot indicates that P from EPIC-Norfolk study is under nominal significance level ($P < 0.05$) and the Black dot indicates that P from EPIC-Norfolk study is over nominal significance level.
Two DMRs (cg26823705 and cg26974062) were not available in the EPIC-Norfolk study.

Figure 10. Cumulative incidence of predictive type 2 diabetes



The risk of type 2 diabetes according to DNAm score was analyzed with Cox proportional-hazards models.
The solid line represents the bottom 10% group; long dashed line represents the top 10% group, and the short dashed line represents the normal control group (10-90% of DNAm score)
Abbreviation: CI, confidence interval; SD, standard deviation

Table 9. Previously reported type 2 diabetes associated mQTLs

SNP	mQTL associated CpG	Position	Gene	RA	Reported GWAS traits
rs12712928	cg26823705	Chr2:44964941	AC012354.1	G	HbA1c, type 2 diabetes
rs1260326	cg19693031	Chr 2:27508073	GCKR	C	Serum metabolites, Triglyceride, Total cholesterol, type 2 diabetes, BMI-adjusted leptin measurement, glomerular filtration rate
rs780094	cg19693031	Chr 2:27518370	GCKR	C	Serum metabolites, Triglyceride,, fasting blood sugar, type 2 diabetes, fasting blood insulin
rs17168486	cg14530801	Chr 7:14858657	DGKB	T	Type 2 diabetes
rs1387153	cg08867893	Chr 11:92940662	AP003171.1, SNRPGP16	T	Fasting blood sugar, metabolic syndrome, HbA1c, type 2 diabetes
rs10830963	cg08867893	Chr 11:92975544	MTNR1B	G	Type 2 diabetes, fasting blood sugar, HbA1c, serum metabolite measurement
rs147538848	cg17082373	Chr 12:31313679	SINHCAF	A	Type 2 diabetes

Abbreviation: Chr, Chromosome; GWAS, Genome-wide association study; mQTL, methylation quantitative trait loci; RA, Risk allele

Table 10. Association of mQTLs scores as an IV and type 2 diabetes DMS

CpG	Coefficient (SD)	P-value	F-statistics	R-square
cg25139493	2.53 (0.80)	**0.002**	9.9	0.06
cg26823705	-7.02 (2.77)	**0.012**	6.43	0.04
cg26974062	10.91 (2.16)	**<0.001**	25.51	0.15
cg19693031	15.11 (3.66)	**<0.001**	17.05	0.11
cg14530801	0.95 (0.22)	**<0.001**	17.97	0.11
cg08867893	2.76 (2.76)	**<0.001**	14.14	0.09
cg00574958	3.12 (0.88)	**0.001**	12.44	0.09
cg17082373	0.52 (0.19)	**0.006**	7.95	0.06

Bold style font indicates P-value under 0.05

Table 11. The causal association between DNAm and type 2 diabetes

CpG	OR [95% C.I]	P
cg25139493	1.08 [0.98-1.19]	0.103
cg26823705	1.00 [0.95-1.05]	0.982
cg26974062	0.99 [0.94-1.04]	0.582
cg19693031	0.99 [0.97-1.02]	0.682
cg14530801	0.74 [0.51-1.07]	0.121
cg08867893	1.00 [0.88-1.13]	0.983
cg00574958	0.90 [0.81-1.00]	0.052
cg17082373	0.76 [0.51-1.13]	0.173

Abbreviation: OR, Odds ratio

Table 12. The proportion of phenotypic variance(T2D) explained by DNAm

	Family data (HTS only)		Population data (HTS, KoGES, SNUH)	
Model	**Heritability**	**The proportion of phenotypic variance explained by covariates (Kullback-Leibler R-squared)**	**Heritability**	**The proportion of phenotypic variance explained by covariates (Kullback-Leibler R-squared)**
Age+Sex	47.30	12.02	69.84	7.39
Standardized DNAm score (M-value)	49.84	2.05	66.08	14.31
Age+Sex+Standardized DNAm score (M-value)	45.77	13.88	64.01	16.20

Figure 11. ROC curve using logistic regression algorithm prediction

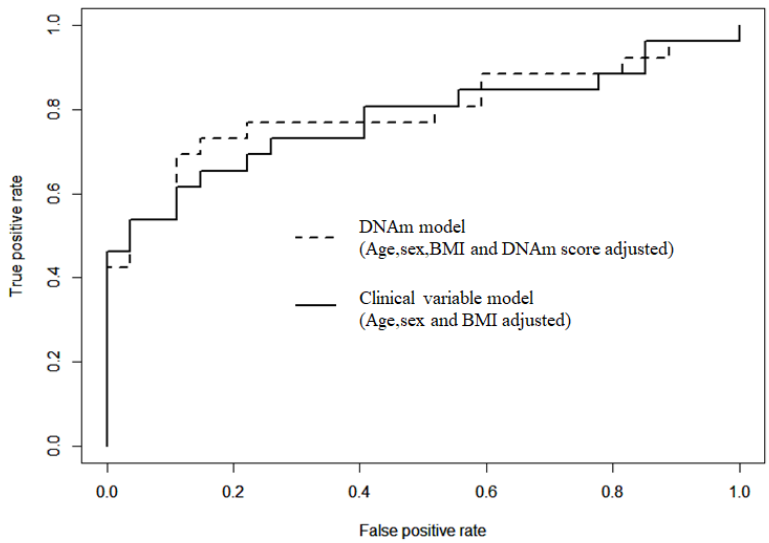


Figure 12. AUC of two logistic models after 1000 replication

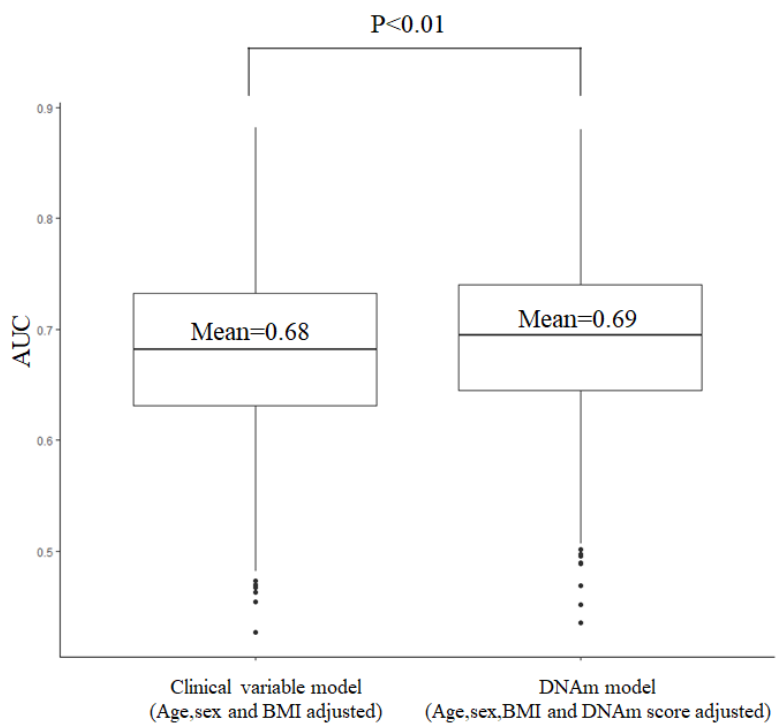
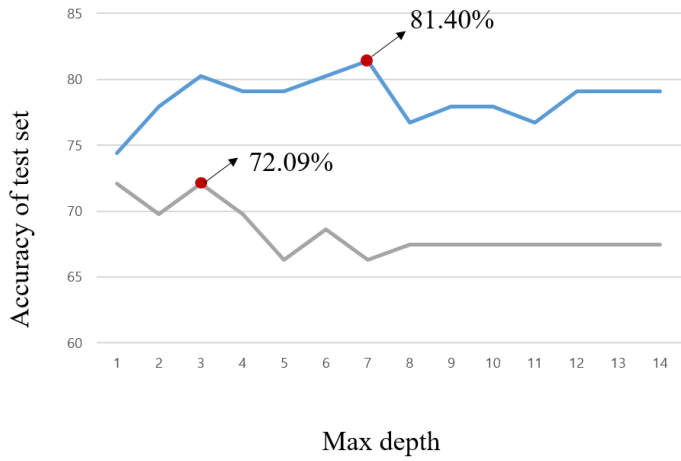


Figure 13. Test set accuracy of XGboost prediction model by increasing max depth



Blue line: 8 CpG markers + age+ sex+BMI model
Gray line: age+sex+BMI model

Figure 14. Feature importance in XGboost prediction model

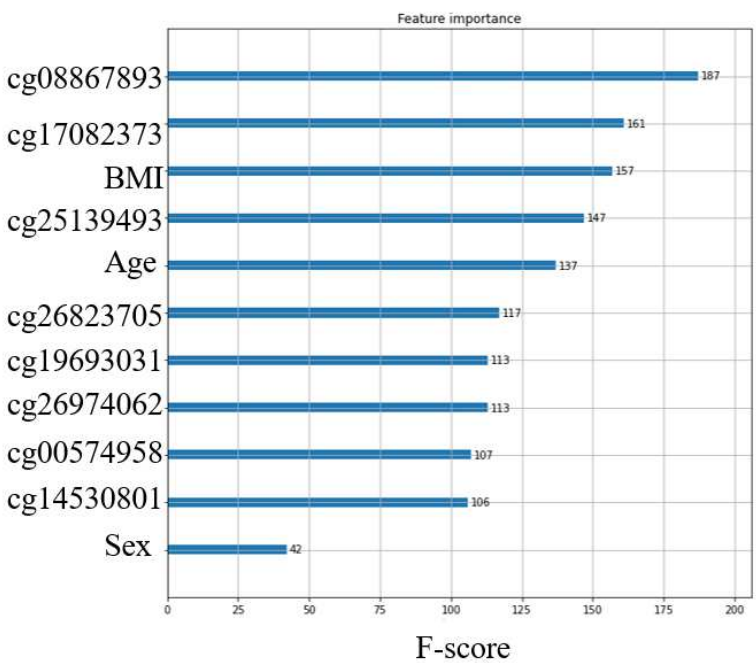
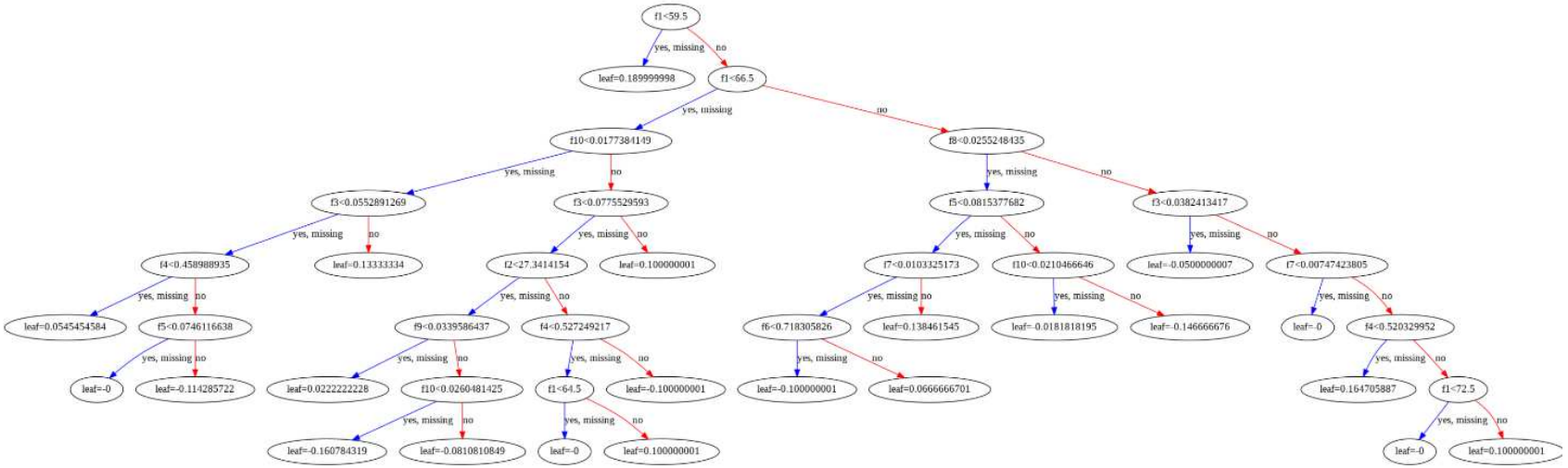


Figure 15. Plotting tree from XGbosst prediction model



Feature: f1, age; f2, BMI, f3, cg25139493; f4, cg26823705; f5, cg26974062; f6, cg19693031; f7, cg14530801; f8, cg08867893, f9, cg00574958; f10, cg17082373

Supplementary Table 1. Clinical characterristics of KoGES and SNUH

	KoGES (N=400)			HTS (N=510)		
	Type 2 diabetes (N=44)	Non-diabetic controls (N=356)	P for difference	Type 2 diabetes (N=47)	Non-diabetic controls (N=463)	P for difference
	Mean±SD/ N (%)	Mean±SD/ N (%)		Mean±SD/ N (%)	Mean±SD/ N (%)	
Age	56.0±6.79	52.7±8.42	0.005	55.8±13.1	42.8±13.0	<0.001
Sex						
Male	23 (52.7)	177 (49.7)	0.873	25 (53.2)	226 (48.8)	0.675
BMI (kg/m2)	25.5±2.86	24.6±3.5	0.061	26.1±3.44	23.8±3.5	<0.001
Fasting glucose (mmol/L)	9.6±3.3	5.3±1.5	<0.001	130±35.6	5.0±0.63	<0.001
HbA1c (mmol/mol)	77.0±18.7	39.0±11.1	<0.001	59.9±19.8	34.6±4.18	<0.001
HbA1c (%)	9.33±1.55	5.72±1.02	<0.001	7.63±1.81	5.35±0.49	<0.001
SBP (mg/dL)	78.5±	125±17.5	0.045	123±20.7	116±17.7	0.035
eGFR (ml/min)	98.4±24.6	94.7±19.8	0.339	73.1±9.17	76.6±10.9	0.017

Data are presented as mean ± standard deviation or n (%)

Statistical difference analysis was done with T-test for continuous variables and Chi-square test

Abbreviation: BMI, Body mass index; eGFR, estimated Glomerular Filtration Rate; HbA1c, Glycated hemoglobin; SBP, Systolic blood pressure

Supplementary Table 2. The entire information of differential methylated CpGs for type 2 diabetes

CpG	Nearby gene	CGI Information	Gene property	Position (hg 19)	Adjusted Differences (M-values)	Adjusted Differences (β-values)	Raw β differences	*P* *	FDR *
cg25139493	*BMP8A*	Island	5'UTR	chr1:39957400	0.340	0.013	0.005	6.89E-08	6.03E-03
cg26823705	*NBPF20*	Open Sea	Body	chr1:145435523	-0.277	-0.047	-0.015	5.67E-09	7.45E-04
cg26974062	*TXNIP*	Open Sea	Body	chr1:145440734	-0.500	-0.031	-0.011	9.64E-11	7.59E-05
cg19693031	*TXNIP*	Open Sea	3'UTR	chr1:145441552	-0.465	-0.061	-0.024	2.65E-09	5.22E-04
cg14530801	*STX18*	Island	TSS1500	chr4:4544016	0.755	0.001	0.000	2.12E-09	5.22E-04
cg08867893	*ZNF365*	Island	5'UTR	chr10:64134160	0.454	0.009	0.003	2.13E-08	2.40E-03
cg00574958	*CPT1A*	North shore	5'UTR	chr11:68607622	-0.373	-0.016	-0.005	5.40E-10	2.13E-04
cg17082373	*TRIM37*	South shore	TSS200	chr17: 57184450	-0.482	-0.017	0.001	4.56E-08	4.49E-03

TSS200 is the region from TSS to − 200 nt upstream of TSS; TSS1500 covers − 200 to − 1500 nt upstream of TSS;

Bold style font indicates novel CpG sites that are not previously reported to be associated with type 2 diabetes

*P and FDR corresponding to M-values

Abbreviation: chr, chromosome; CGI, CpG island; FDR, False discovery rate; FDR, hg19, human genome version 19; TSS, Transcription start site; UTR, Untranslated Region

Supplementary Table 3. Replication information of known type 2 diabetes DMRs

CpG	Position (hg 19)	Nearby gene	Adjusted Differences (M-values)	*P* *	Adjusted Differences (β-values)
cg00076653	Chr4:15341878	*C1QTNF7*	0.061	0.096	0.009
cg00277397	Chr7:71800412	*CALN1*	-0.031	0.559	-0.006
cg19693031	Chr1:145441552	*TXNIP*	-0.465	2.65.E-09	-0.061
cg01657995	Chr6:31804883	*C6orf48*	-0.018	0.777	-0.001
cg01676795	Chr7:75586348	*POR*	0.068	0.097	0.013
cg02560388	Chr2:11969958		0.06	0.150	0.004
cg19266329	Chr1:145456128		-0.136	0.001	-0.023
cg02711608	Chr19:47287964	*SLC1A5*	-0.031	0.425	-0.003
cg03497652	Chr16:4751569	*ANKS3*	0.032	0.428	0.007
cg03699074	Chr16:88849875	*FAM38A*	-0.1	0.067	-0.006
cg03725309	Chr1:109757585	*SARS*	-0.033	0.401	-0.002
cg04344749	Chr1:25871814	*LDLRAP1*	0.018	0.478	0.003
cg04645070	Chr11:34393106		-0.017	0.636	-0.001
cg04727071	Chr19:4061359	*ZBTB7A*	-0.039	0.308	-0.006
cg00574958	Chr11:68607622	*CPT1A*	-0.373	5.40.E-10	-0.016
cg04973995	Chr10:74057977		0.004	0.928	0.0003
cg04992150	Chr3:13457267	*NUP210*	0.01	0.641	-0.0001
cg05400498	Chr10:32049926		0.01	0.756	0.002
cg05778424	Chr17:55169508	*AKAP1*	0.049	0.093	0.008

cg06007201	Chr16:88850218	*FAM38A*	-0.048	0.321	-0.002
cg06178887	Chr19:42599631	*POU2F2*	-0.022	0.436	-0.003
cg06235429	Chr11:67373114	*NDUFV1*	0.037	0.334	0.005
cg17058475	Chr11:68607737	*CPT1A*	-0.317	1.10.E-06	-0.018
cg06721411	Chr2:74753759	*DQX1*	0.034	0.291	0.007
cg07092212	Chr11:46382544	*DGKZ*	-0.022	0.511	-0.002
cg07960624	Chr8:119208486	*SAMD12*	-0.029	0.596	-0.004
cg08309687	Chr21:35320596		-0.09	0.105	-0.014
cg08788930	Chr8:142201685	*DENND3*	0.074	0.077	0.011
cg08994060	Chr10:6214026	*PFKFB3*	-0.024	0.609	-0.004
cg09247619	Chr1:198648849	*PTPRC*	-0.009	0.747	-0.001
cg10508317	Chr17:76355146	*SOCS3*	0.02	0.651	0.001
cg10919522	Chr14:74227441	*C14orf43*	-0.073	0.133	-0.007
cg14597545	Chr15:73074210	*ADPGK*	0.068	0.013	0.012
cg11183227	Chr15:91455407	*MAN2A2*	0.063	0.165	0.008
cg11376147	Chr11:57261198	*SLC43A1*	0.028	0.328	0.002
cg13199639	Chr6:33360495	*KIFC1*	0.011	0.816	0.001
cg13514042	Chr7:1192202		-0.028	0.558	-0.003
cg13640297	Chr6:170099238	*WDR27*	0.032	0.264	0.007
cg14020176	Chr17:72764985	*SLC9A3R1*	0.037	0.173	0.006
cg14476101	Chr1:120255992	*PHGDH*	0.013	0.840	0.002
cg11024682	Chr17:17730094	*SREBF1*	0.095	2.89.E-04	0.016
cg15585213	Chr1:14146635	*PRDM2*	-0.019	0.755	-5.6E-05

cg15962267	Chr5:138612986	*SNHG4,SNORA74A,MATR3*	-0.038	0.230	-0.007
cg16809457	Chr6:90399677	*MDN1*	0.031	0.439	0.006
cg02650017	Chr17:47301614	*PHOSPHO1*	-0.175	2.83.E-04	-0.005
cg17315426	Chr14:23527629	*CDH24*	0.036	0.419	0.001
cg17666418	Chr6:106612002		-0.051	0.339	-0.005
cg18181703	Chr17:76354621	*SOCS3*	0.019	0.664	0.003
cg26836479	Chr19:42706353	*DEDD2*	-0.121	0.007	-0.004
cg06500161	Chr21:43656587	*ABCG1*	0.135	7.58.E-06	0.023
cg21699330	Chr7:26193032	*NFE2L3*	-0.037	0.545	-0.001
cg21766592	Chr19:47288066	*SLC1A5*	-0.024	0.695	-0.001
cg22909677	Chr6:109172312	*ARMC2*	-0.086	0.101	-0.005
cg24531955	Chr8:23154691	*LOXL2*	0.015	0.734	0.001
cg25130381	Chr1:27440721	*SLC9A1*	0.006	0.875	0.001
cg26546155	Chr16:57269635	*RSPRY1*	0.016	0.547	0.003
cg26712428	Chr10:105218771	*CALHM1*	-0.027	0.786	-0.003
cg26804423	Chr7:8201134	*ICA1*	0.015	0.609	0.004
cg04816311	Chr7:1066650	*C7orf50*	0.091	0.018	0.016

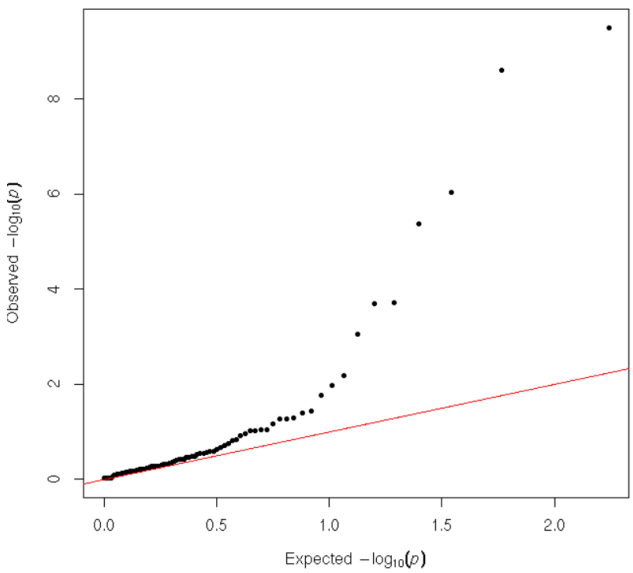
Abbreviation: chr, chromosome; hg19, human genome version 19
*P-value corresponds to M-value.

Supplementary Table 4. Comparison MWAS result of discovered T2D DMSs between BMI adjusted model and BMI not adjusted model

CpG	BMI not adjusted		BMI Adjusted	
	Adjusted M-values differences (β-values)	*P* *	Adjusted M-values differences (β-values)	*P* *
cg25139493	0.337 (0.012)	$8.74x10^{-08}$	0.340 (0.013)	6.89×10^{-8}
cg26823705	-0.275 (-0.047)	$6.78\ x10^{-09}$	-0.277 (-0.047)	5.67×10^{-9}
cg26974062	-0.497 (-0.031)	$1.17\ x10^{-10}$	-0.500 (-0.031)	9.64×10^{-11}
cg19693031	-0.464 (-0.061)	$2.42\ x10^{-09}$	-0.465 (-0.061)	2.65×10^{-9}
cg14530801	0.755 (0.001)	$1.72\ x10^{-09}$	0.755 (0.001)	2.12×10^{-9}
cg08867893	0.454 (0.009)	$2.04\ x10^{-08}$	0.454 (0.009)	2.13×10^{-8}
cg00574958	-0.379 (-0.016)	$3.07\ x10^{-10}$	-0.373 (-0.016)	5.40×10^{-10}
cg17082373	-0.488 (-0.018)	$3.15\ x10^{-08}$	-0.482 (-0.017)	4.56×10^{-8}

**P* corresponding to M-values

Supplementary Figure 1. QQ plot of reported type 2 diabetes associated CpGs



III. MWAS of Diabetic Complication, DKD

1. Method and materials

1.1. Study design and participants

By following the type 2 diabetes patients' cohort described in chapter II, we conducted a nested case-control analysis for 87 DKD cases and 80 non-DKD diabetic controls. Among the type 2 diabetes participants, DKD was defined as the presence of either albuminuria (urine albumin-to-creatinine ratio [ACR] of ≥ 30 mg/g) or estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73 m² calculated using the Modification of Diet in Renal Disease (MDRD) equation [107]. The control for the DKD cases were type 2 diabetes patients with a disease duration of 10 years or longer unlike DKD cases, eGFR ≥ 60 mL/min/1.73 m², and normoalbuminuria (urine ACR of <30 mg/g). The Institutional Review Board (IRB) of the Biomedical Research Institute at Seoul National University Hospital (IRB No. 1603-079-749) and the IRB of the Seoul National University (IRB No. 1902-003-015) approved the study protocol and written informed consent was obtained from each participant. The process of DNAm profiling is detailed in **Chapter II.**

1.2. MWAS of DKD

MWAS for DKD was done similarly to that for type 2 diabetes. M-value (log2 transformed β-value) to consider statistical assumption was fitted using a model matrix with empirical Bayes approach. Cell-type composition, age, sex, HbA1c,

duration of diabetes, and BMI were used as covariates in pairwise contrast. In the DKD analysis, the genomic control approach [108] as used in the genome-wide association study (GWAS) was implemented to remove false-positive signals. For statistical significance, the uncorrected methylome-wide significance of $P < 9\times10^{-8}$ and a false discovery rate (FDR) < 0.05 were used. The overlap of DMSs of DKD and DMSs of type 2 diabetes discovered in **Chapter II** was investigated. We calculated the DNAm score of type 2 diabetes /DKD as suggested in **Chapter II** to explore the differences between DKD, type 2 diabetes patients, and non-diabetic normal controls. MWAS comparing DKD patients and non-diabetic controls was done to verify DMSs of type 2 diabetes and DKD. The analyses were performed using the R software (version 3.5.1)

1.3. The causal effect of DNA methylation on DKD

To investigate the causal association between DNAm and DKD, the MR method was applied as in type 2 diabetes in **Chapter II.** To select IVs, we first searched reported mQTL in the public database (BIOS QTL [67, 68] browser and mQTLdb [69]) but there was none. Next, mQTL of DKD DMSs in the Korean population was investigated through GWAS. SNPs with P-value from GWAS under genome-wide significance level ($P < 5x10^{-8}$) were selected for IV of CpGs (**Table 13**). The Sum of risk alleles of mQTLs divided by the number of mQTLs was used as an IV. THE two-stage MR method was implemented using 2-stage least square regression method. To support the results from two-stage MR and to consider a violation of IV assumption, a robust MR method using summary data estimation by the R package 'Mendelian Randomization' [109] was applied. In MR analysis, there are three of IV assumption: 1) IV is associated with the exposure, 2) there are no associations between the IV and potential confounders and 3) IV have no direct effect on the

outcome. Pleiotropy, which means that one genetic variant is associated with multiple traits, is one of the reasons for IV assumption #3 violation. However, the presence of pleiotropy does not mean that the result of MR is unreliable. If pleiotropic effects are independent of genetic associations with exposures, reliable estimates can still be obtained. This is referred to InSIDE (INstrument Strength Independent of Direct Effect) assumption. However, the InSIDE assumption is difficult to test. Several robust MR methods are suggested to solve the problems due to pleiotropy using a different strategy. The InSIDE assumption is necessary for the MR-Egger method [110]. Weighted median [111] and mode-based method [112] try to overcome pleiotropy by assuming that majority of IVs are valid. MR-PRESSO [113] and MR-Lasso [114] tries to escape assumption violation by the shrinking effect of outlier IVs.

The weighted median method is a robust estimator that provides consistent estimates even if up to 50% of IVs are invalid [111]. The mode-based method utilizes a normal density for each genetic variant and constructs a smoothed density by summing normal densities [112]. The MR-Egger method is based on Egger's regression, which is used for adjusting publication bias in meta-analyses [110]. MR-Lasso is a transformation of the inverse-variance method by adding a lasso penalty term. In MR-Lasso regression, intercept term for each genetic variants and Lasso-penalty term is added in the model. If pleiotropic effect exists in the genetic variant, intercept term is different from zero. After calculating intercept term of each variant, MR-Lasso exclude the invalid genetic variants, which of intercept term is not zero, from the model and performs IVW method. Using multiple robust MR methods using different strategies is recommended to researchers. We selected the best one among diverse robust MR methods referring to elsewhere [115]. Slob et al. tested the performance of robust MR methods in diverse scenarios. We summarized results (for number of IVs=10) in **Figure 16**. First, pleiotropy was tested by MR-Egger intercept.

In each scenario, we selected the best method, one for each of the different strategy categories. Weighted median, Mode based and MR-Lasso were selected.

2. Results

2.1. Clinical characteristics of the study participants

Among the participants with diabetes, we were able to classify 87 and 80 as DKD cases with a diabetes duration of 10 years or more and non-DKD controls, respectively (**Table 14**). Between DKD cases and non-DKD controls, statistical differences were not observed in age, the composition of sex, BMI, glucose, and HbA1c level. On contrary, DKD cases had higher SBP, lower eGFR, and higher urine ACR compared to non-DKD controls

2.2. Differentially methylated regions of DKD

In our nested case-control cohort of DKD (87 cases and 80 controls), we explored whether there were DMSs related to DKD. Three DMSs were significantly associated with DKD ($P < 9\times10^{-8}$ and FDR < 0.05). These include cg16944159 at *copper metabolism domain containing 1* (*COMMD1*), cg16079347 at *tropomodulin 1* (*TMOD1*), and cg11530914 at *formin homology 2 domain containing 1* (*FHOD1*) (**Table 15 and Supplementary Table 5**). QQ and volcano plot is visualized in **Figure 17**.

There was little overlap in the DMSs between type 2 diabetes and DKD (**Table 16**). One of the type 2 diabetes DMSs (cg00574958 in *CPT1A*) showed an association with DKD at a nominal significance of $P < 0.05$, and none of the three DKD DMSs were associated with type 2 diabetes. For the 59 previously reported type 2 diabetes DMSs, four DMSs (cg01676795, cg03497652, cg14020176, and

cg08309687) were associated with DKD at nominal significance (**Supplementary Table 6).** There was no difference between type 2 diabetes and DKD patients in DNAm score of type 2 diabetes, where scores of both groups were significantly different from that of non-diabetic normal controls (**Figure 18**). DNAm score of DKD was the highest among three groups: DKD, type 2 diabetes, and controls (**Figure 19**). In MWAS comparing DKD and normal non-diabetic controls, all of DMSs of DKD were significant and some DMSs of type 2 diabetes were significant in nominal significance range (P<0.05) **(Table 16)**. DNAm score of DKD was not associated with eGFR (P=0.14) (**Figure 20**).

For sensitivity analysis, MWAS with DKD patients and non-diabetic controls was performed (**Table 17 and Figure 21**). Two DMSs of type 2 diabetes was not significantly different between DKD patients and non-diabetic controls.

2.3. The causal effect of DNA methylation on DKD

There were 7 mQTLs for cg16944159, 40 for cg16079347, and 4 for cg11530914 which satisfies genome-wide significance ($P<5x10^{-8}$). All of the mQTLs were strongly associated with each DMSs considering F-statistics and R-square, where even R-square of cg16944159 and cg16079347 exceed 0.20 (**Table 18**). Two DKD DMSs (cg16944159 and cg11530914) causally affected DKD (**Table 19**). Increasing 1% of cg16944159 makes DKD risks 1.15 times higher and increasing 1% of cg11530914 makes 1.19 times higher. We implemented robust MR methods including weighted median, mode-based, and MR-Lasso. As positive pleiotropy was observed (P for MR-Egger intercept test=0.042) and InSIDE assumption could not be tested, weighted-median and MR-LASSO were selected as a representative method. Robust MR methods could not be applied to cg11530914 due to a lack of the number of mQTLs to get a reliable result. All of the methods were still significant (**Table 20 and Figure 22**).

3. Discussion

This study aimed to identify DMSs associated with DKD in the Korean population. We identified 3 novel DMSs to be associated with DKD. Additionally, the overlap between DMSs of type 2 diabetes and that of DKD was investigated to explore the pathogenesis of DKD.

DKD, a highly prevalent complication of type 2 diabetes, progresses to ESRD and finally contributes to the morbidity of type 2 diabetes patients [34]. DKD is characterized by the accumulation of extracellular matrix proteins such as collagen and fibronectin in renal, hypertrophy, and fibrosis in glomerular and tubular cells [116]. Currently, controlling glucose level, angiotensin-receptor blocker, angiotensin-converting enzyme inhibitors are used as a therapy to manage DKD [117]. However, the DKD patients who received intensive glycemic control continue to experience the complication of diabetes even after glucose level had has normalized, suggesting the 'metabolic memory' phenomena [118-121]. The persistent effect of hyperglycemia through metabolic memory suggests epigenetic mechanisms play an important role in the progression and prevention of DKD.

One of the interesting findings of this study is the identification of DMRs of DKD. As far as we are aware, this is the first MWAS to investigate DMRs of DKD in peripheral leukocytes. We identified three DMRs each in *COMMD1*, *TMOD1*, *FHOD1* to be significantly associated with DKD after adjustment for multiple covariates including age, sex, BMI, duration of diabetes, HbA1c, and SBP. Although how three DMRs play a role in DKD requires further investigation, there are possible mechanisms. *COMMD1* is coexpressed with epithelial sodium channel (ENaC) in the renal medulla and plays an important role in the down-regulation of ENaC and Na^+ reabsorption [122]. *TMOD1* is expressed in renal

distal tubules and collecting ducts, both of which are critical in water balance. Knockout mice model of *Tmod1* revealed oliguria, hyperosmolar urine, and increased blood pressure [123].

DNAm can affect gene expression and phenotype related to gene expression without changing DNA itself and can mediate the interaction between genes and the environment. Several studies support the evidence that DNAm level change has a major role in the pathogenesis of DKD through altered gene expression. Increased expression of TGF-β1 leads to renal fibrosis through hyper-methylation of Rasal1 and following procedure of increased Ras-GTP signaling [124]. In another study, mice with deletion of the *Sirt1* gene showed worse symptoms in albuminuria with hypo-methylation of claudin-1. In the db/db mice (mice with gene mutation producing leptin receptor) experiment, transcriptional repression of the KLF4 associated with hyper-methylation at the promoter region of the gene which encodes nephrin. These changes lead to podocyte apoptosis and proteinuria [125]. Also, the promoter region of AGT which is an associated gene with chronic kidney disease showed hypo methylation of H3K9 and increased expression in DKD. This DNAm change was not reversible even though DKD treatment was applied [126]. Although mechanisms of three DMSs discovered in DKD MWAS need further investigation, the evidence that DNAm affects the pathogenesis of DKD through gene expression helps to understand the role of DNAm in DKD.

Best of our knowledge, there was no DKD specific MWAS in the population. Some genes such as *UNC13B [127], TAMM41, PMPCB, TSFM, AUH [128], DNMT1 [129],* and others [130] were discovered at study design with DKD patients in type 1 diabetes. Also, there were MWAS about CKD patients including some DKD patients. *COL4A1, COL4A2, SIX2, HNF [131], PTPN6-PHB2,*

ANKRD11, TNRC18 [132], and others were found to be associated with CKD. As DKD specific DMSs located genes discovered from this study are not included in previously reported genes with CKD or DKD in type 1 diabetes patients, we can infer that DNAm profiles of DKD in type 2 diabetes are different from CKD and DKD in type 1 diabetes patients.

It is interesting to note that there was little overlap between DMRs of type 2 diabetes and DKD. Among eight DMRs, just one DMR of type 2 diabetes was marginally associated with the risk of DKD at one-to-one comparison with nominal P = 0.046 for DKD. No significant difference of type 2 diabetes DNAm score and significant difference between their groups of DKD DNAm score suggests that there is a common profile among all type 2 diabetes and something different characteristics exist in DKD DNAm profiles. Given that the DNAm score of DKD was the lowest in type diabetes, even more than the score in the control group, we can conclude that DNAm profiles of DKD are clearly distinguished by that of type 2 diabetes.

Presently, it is unclear that why some type 2 diabetes patients progress to DKD and some are not. Although recent genetic studies demonstrated numerous genetic components associated with DKD [133, 134], there was still a large discrepancy between observed heritability and variability explained by discovered genetic variants, which is called 'missing heritability' phenomena. 'Missing heritability' raises the possibility of the environmental difference between DKD patients and non-DKD diabetic controls. Furthermore, the metabolic and hemodynamic stimulus of uncontrolled diabetes is one of the causes of renal cellular hypertrophy and hyper filtration which progress to DKD [135], suggesting the environmental stimuli effect on a gene. Our discovery of little overlap between DMRs of type 2 diabetes and DKD supports the evidence that DKD is not just a prolonged

complication of type 2 diabetes.

The causal association between DMSs of DKD and DKD is also an interesting part of this study. Causal associations of eGFR on DKD DMSs were still significant with a larger effect size after adjusting diabetes duration (**Supplementary Table 7**). An increase of effect size after adjusting diabetes duration suggests that diabetes duration is associated with DKD incidence.

Given the number of mQTLs of DKD DMSs, robust estimates could be achieved by assuming that two IV assumption is violated: InSIDE assumption violated and majority IV is invalid. In this scenario, we selected MR-Lasso as the best performing method. The causal association was still significant in the MR-Lasso result. The causal association was still significant in the MR-Lasso result.

Significant causal association achieved by the MR method suggests two important things: First, genetic factors contribute a large part of the variation of DKD DMSs up to 20%. Second, the risk of DKD is increased by the DNAm of DKD DMSs changes. This two fact suggests that genetics have a larger effect on DKD than we expected.

This study has the strength in the DKD cases were collected from a well-described clinical cohort and the definitions of DKD cases and controls were reliable with sufficient clinical data. There were a few limitations in this study. First, these results would be interpreted with caution as the sample size for DKD analysis was limited. Second, we could not replicate our results in different populations due to the data availability. Replication and verification of three novel DMSs should be progressed in further research. Third, MR analysis and discovery mQTLs were performed in one sample due to data limitation.

In conclusion, we identified three novel DMSs associated with DKD with methylome-wide statistical significance in the East-Asian population. We hope

our findings provide a better understanding of the pathophysiology of DKD and contribute to developing a better intervention in type 2 diabetes for not to progress to DKD.

Table 13. mQTLs of DKD DMSs

SNP	Position	RA	Effect size	P
cg16944159				
rs111063605	chr1:1891009	G	0.057	1.70E-08
rs79841299	chr2:392873	T	0.121	2.56E-08
rs4453859	chr3:8648555	C	0.060	1.00E-08
chr8:110621585	chr6:621585	G	0.346	3.90E-08
chr12:122723521	chr9:2723521	C	0.260	2.40E-08
rs369577967	chr10:4915293	G	0.263	2.50E-08
rs9610398	chr11:6493933	T	0.249	1.00E-08
cg16079347				
chr1:222630177:T:G	chr1:2630177	G	0.319	2.00E-08
rs1765791	chr1:1730121	T	0.049	2.91E-08
rs139182578	chr2:904748	G	0.319	2.00E-08
rs144220295	chr2:1003233	G	0.319	2.00E-08
chr2:211589950:G:A	chr2:1589950	A	0.319	2.00E-08
chr3:26766625:A:G	chr3:6766625	G	0.319	2.00E-08
chr3:26813453:G:A	chr3:6813453	A	0.319	2.00E-08
chr3:26826951:G:A	chr3:6826951	A	0.319	2.00E-08
chr3:60065457:A:G	chr3:65457	G	0.319	2.00E-08
chr4:73908419:T:C	chr4:3908419	C	0.319	2.00E-08
chr4:74040957:G:T	chr4:4040957	T	0.319	2.00E-08
chr4:74182605:C:T	chr4:4182605	T	0.319	2.00E-08
chr4:74305656:A:G	chr4:4305656	G	0.319	2.00E-08
chr4:74305660:G:A	chr4:4305660	A	0.319	2.00E-08
chr4:137133667:A:G	chr4:7133667	G	0.319	2.00E-08
chr5:9908189:G:A	chr5:9908189	G	0.252	2.00E-08
chr5:20322264:A:G	chr5:322264	G	0.319	2.00E-08
chr5:20518672:G:A	chr5:518672	A	0.319	2.00E-08

chr5:107887702:T:C	chr5:7887702	C	0.319	2.00E-08
chr5:107971539:G:A	chr5:7971539	A	0.319	2.00E-08
chr7:151381144:C:T	chr7:1381144	T	0.319	2.00E-08
chr9:111425306:A:G	chr9:1425306	G	0.319	2.00E-08
chr10:82036407:A:G	chr10:2036407	G	0.319	2.00E-08
chr10:82044946:G:C	chr10:2044946	C	0.319	2.00E-08
chr10:82138329:A:G	chr10:2138329	G	0.319	2.00E-08
chr10:82225780:C:T	chr10:2225780	T	0.319	2.00E-08
chr10:82248756:A:G	chr10:2248756	G	0.319	2.00E-08
chr11:38927568:T:G	chr11:8927568	T	0.335	8.60E-09
chr11:38930499:C:T	chr11:8930499	C	0.335	8.60E-09
chr11:38988164:T:C	chr11:8988164	T	0.335	8.60E-09
chr11:38993936:T:C	chr11:8993936	T	0.335	8.60E-09
chr11:99537805:T:G	chr11:9537805	G	0.319	2.00E-08
rs9574598	chr13:789120	T	0.118	1.00E-08
chr13:94812439:A:G	chr13:4812439	G	0.319	2.00E-08
chr14:76185901:A:C	chr14:6185901	C	0.319	2.00E-08
chr14:76559400:C:T	chr14:6559400	T	0.319	2.00E-08
chr17:31281610:G:A	chr17:1281610	G	0.156	1.00E-08
chr17:31285968:G:T	chr17:1285968	G	0.156	1.00E-08
chr18:51431843:C:G	chr18:1431843	G	0.319	2.00E-08
chr19:7668674:G:T	chr19:7668674	T	0.319	2.00E-08
cg11530914				
rs4540208	chr5:8030376	T	0.168	2.02E-09
rs4243437	chr5:8031661	C	0.168	2.02E-09
rs4132382	chr5:8042467	C	0.168	2.02E-09
chr9:38383736	chr9:8383736	A	0.093	7.30E-09

Abbreviation: chr, chromosome; RA, Risk alelle

Figure 16. Recommended robust MR methods by scenarios of MR assumption (N of IV =10)

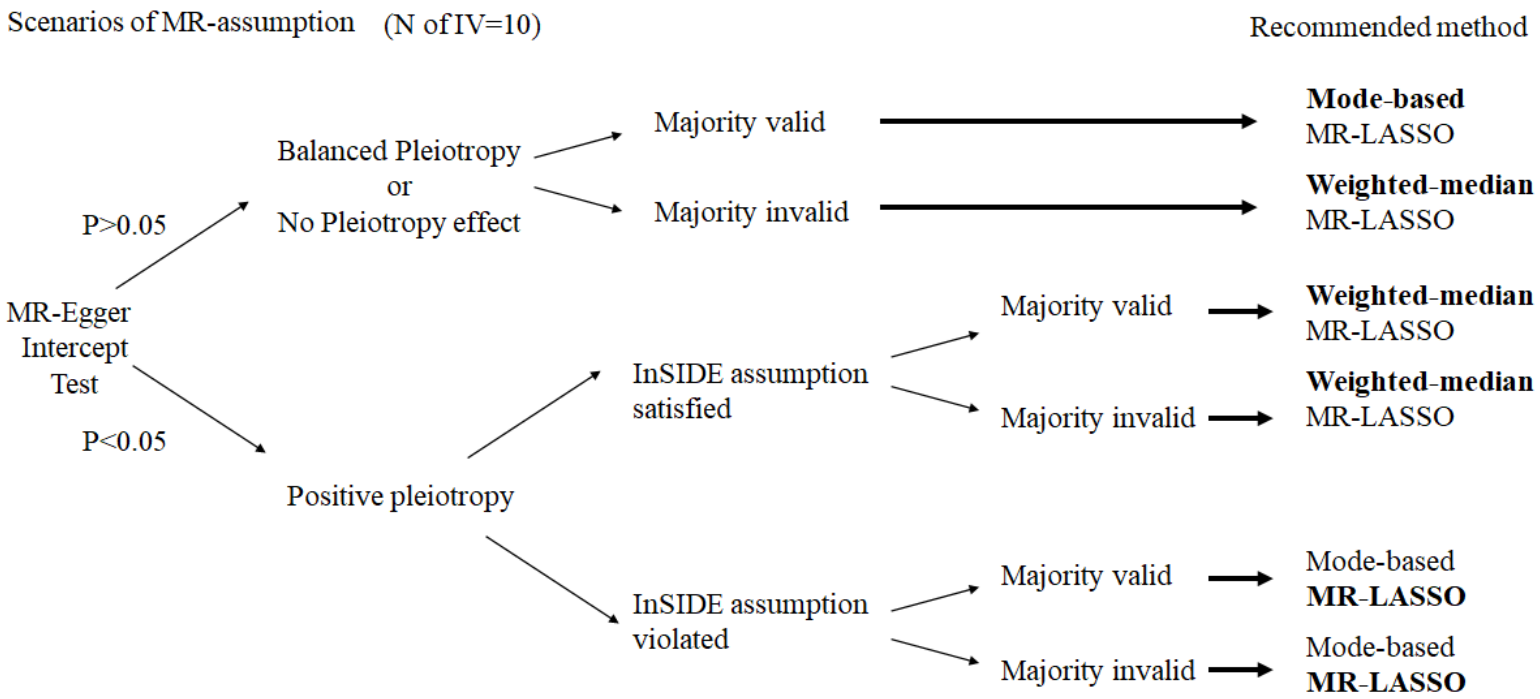


Table 14. Clinical characteristic of study participants of DKD MWAS

	Sub-cohort of SNUH		
	DKD cases (N=87)	**Diabetic controls (type 2 diabetes duration ≥10) (N=80)**	***P* for difference**
Age	63.01±10.76	63.43±8.61)	0.78
Sex			
Male	48 (55.17)	33 (41.25)	0.10
BMI (kg/m2)	24.6±3.23	23.89±2.64	0.11
Glucose (mmol/L)	8.49±3.80	8.48±2.94	0.35
HbA1c (mmol/mol)	66.61±18.31	67.38±16.72	0.79
HbA1c (%)	7.97±1.67	8.04±1.53	0.79
SBP (mg/dL)	138±18.6	127±20.3	<0.01
eGFR (ml/min)	46.2±30.9	87.9±19.7	<0.01
Urine ACR (mg/g)	1210±1756	13.58±7.72	<0.01
CKD Stage			
Stage 1	6 (6.90)	43 (53.75)	<0.01
Stage 2	22 (25.29)	33 (41.25)	
Stage 3	29 (33.33)	4 (5.00)	
Stage 4	14 (16.09)	0 (0)	

Stage 5	16 (18.39)	0 (0)	
Duration of type 2 diabetes (years)	17.90±9.16	16.18±5.58	0.14

Data are presented as mean ± standard deviation or n (%)
Statistical difference analysis was done with T-test for continuous variables and Chi-square test for categorical variables
Abbreviation: BMI, Body mass index; eGFR, estimated Glomerular Filtration Rate; HbA1c, Glycated hemoglobin; SBP, Systolic blood pressure

Table 15. Differential methylated CpGs for DKD

CpG	Nearby gene	Position (hg19)	CGI information	Gene property	Adjusted M-values differences (β-values)	*P* *
DKD						
cg16944159	*COMMD1*	chr2:62132759	Open sea	TSS200	0.272 (0.008)	8.20×10^{-8}
cg16079347	*TMOD1*	chr9:100265181	Island	5'UTR	0.237 (0.015)	2.87×10^{-9}
cg11530914	*FHOD1*	chr16:67281528	Island	-	0.261 (0.008)	7.97×10^{-8}

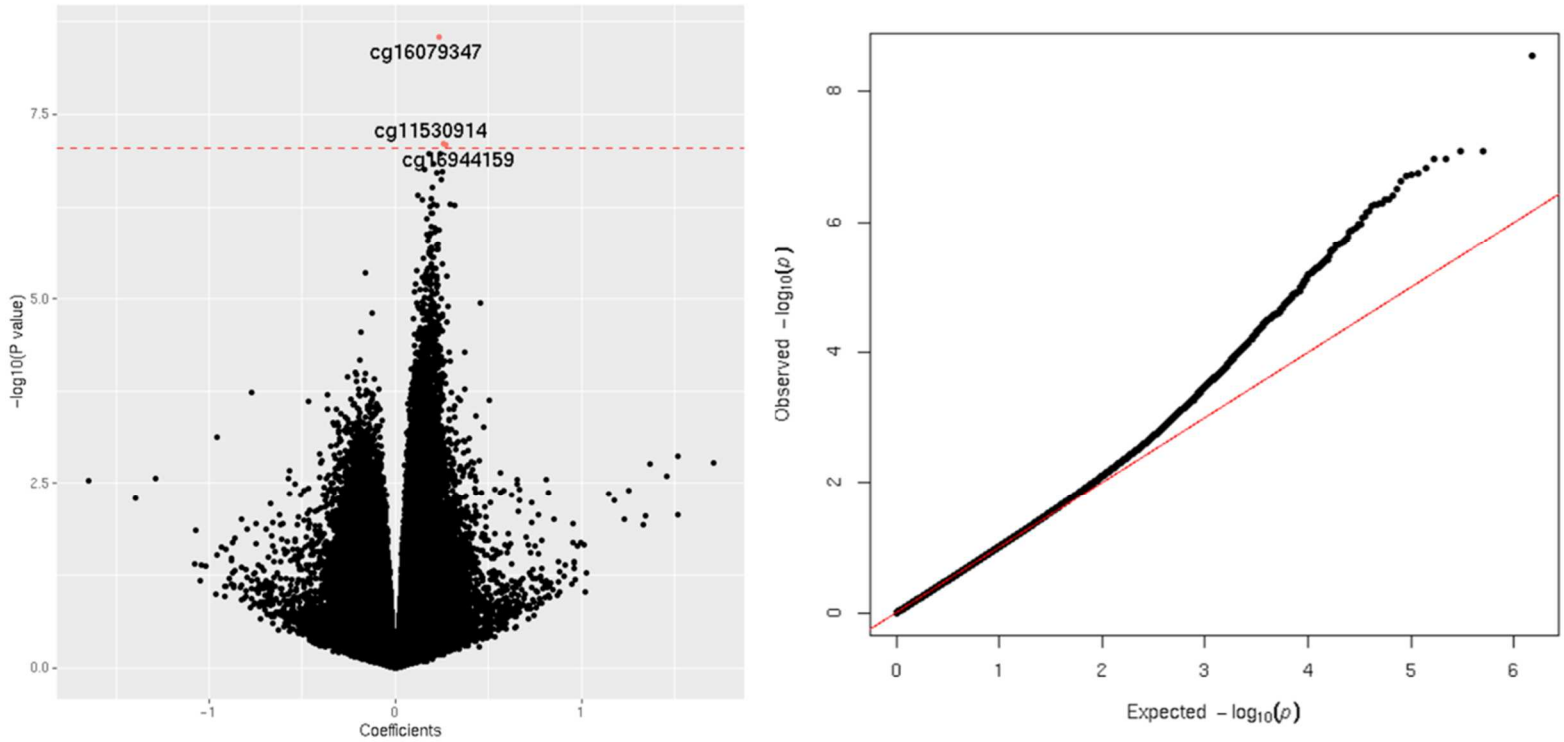
TSS200 is the region from TSS to 200 base pairs upstream of TSS; TSS1500 covers 200 to 1500 base pairs upstream of TSS.

Bold style font indicates novel CpG sites, which were not reported to be associated with type 2 diabetes previously.

**P* corresponding to M-values

Abbreviation: chr, chromosome; CGI, CpG island; hg19, human genome build 19; TSS, Transcription start site; UTR, Untranslated Region

Figure 17. Volcano plot and QQ-plot of DKD MWAS



* Red dashed line indicates the significant level ($P < 9x10^{-8}$)

Table 16. Overlap between type 2 diabetes MWAS and DKD MWAS

CpG	Position (hg19)	Nearby gene	Adjusted M-values Differences (β-values)	*P* *
Type 2 diabetes DMRs in DKD MWAS				
cg25139493	chr1:39957400	*BMP8A*	0.070	0.377
cg26823705	chr1:145435523	*NBPF20*	0.067	0.170
cg26974062	chr1:145440734	*TXNIP*	0.030	0.713
cg19693031	chr1:145441552	*TXNIP*	0.103	0.189
cg14530801	chr4:4544016	*STX18*	0.012	0.938
cg08867893	chr10:64134160	*ZNF365*	0.055	0.601
cg00574958	chr11:68607622	*CPT1A*	0.141	**0.046**
cg17082373	chr17:57184450	*TRIM37*	0.104	0.426
DKD DMRs in type 2 diabetes MWAS				
cg16944159	chr 2:62132759	*COMMD1*	-0.001	0.603
cg16079347	chr 9:100265181	*TMOD1*	-0.001	0.756
cg11530914	chr 16:67281528	*FHOD1*	-0.001	0.473

DMRs discovered in type 2 diabetes/DKD MWAS were investigated for the effect significance in DKD/type 2 diabetes MWAS.

**P* corresponding to M-values

Bold style font indicates *P* below nominal significance ($P < 0.05$).

Figure 18. Boxplot of DNAm score of type 2 diabetes

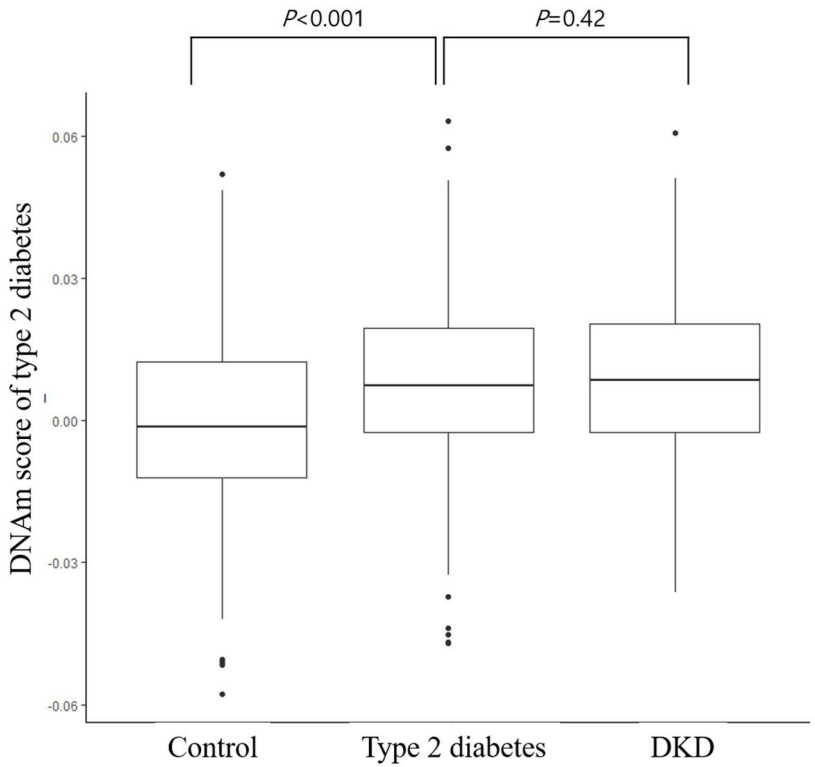


Figure 19. Boxplot of DNAm score of DKD

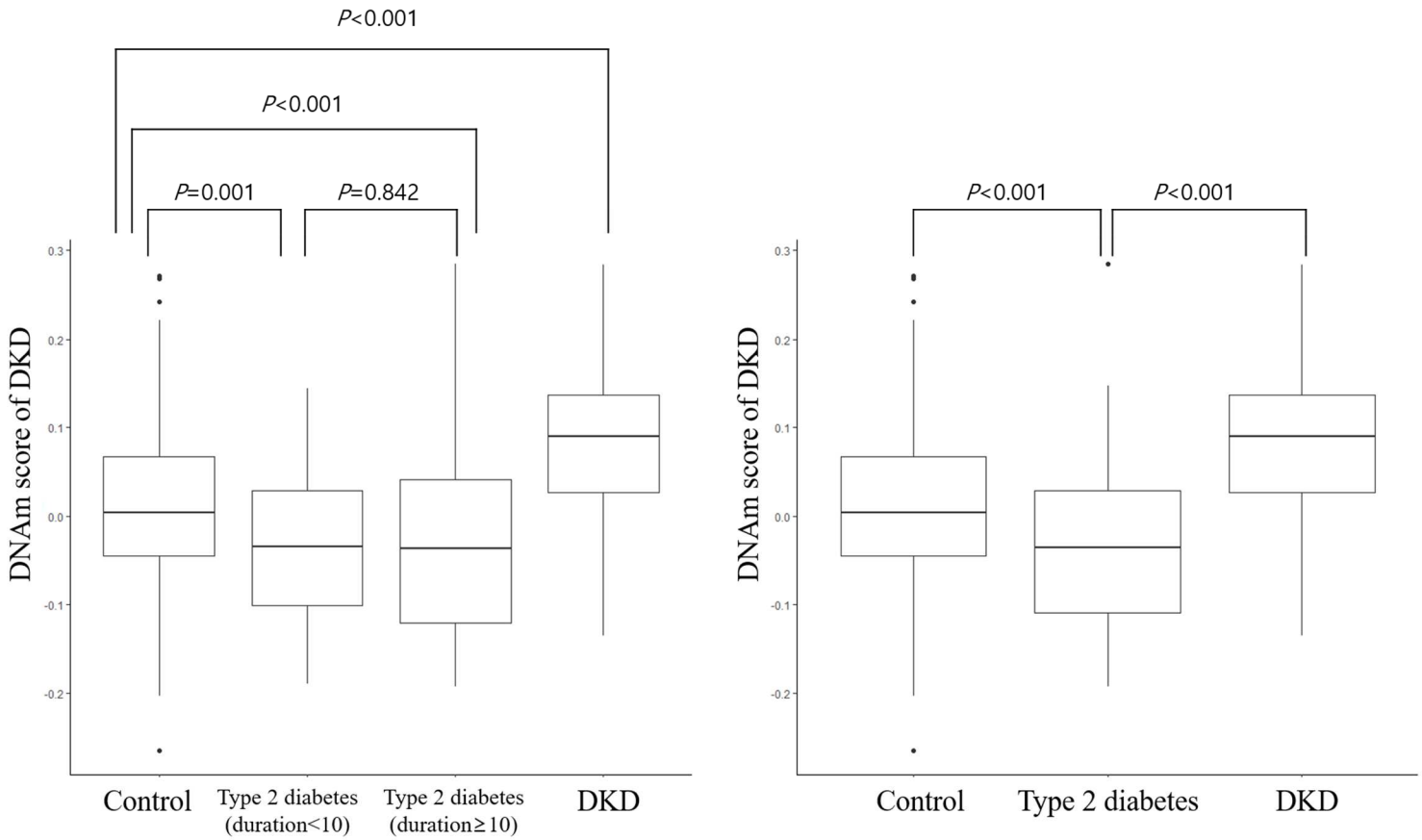


Figure 20. DNAm score of DKD DMSs by eGFR

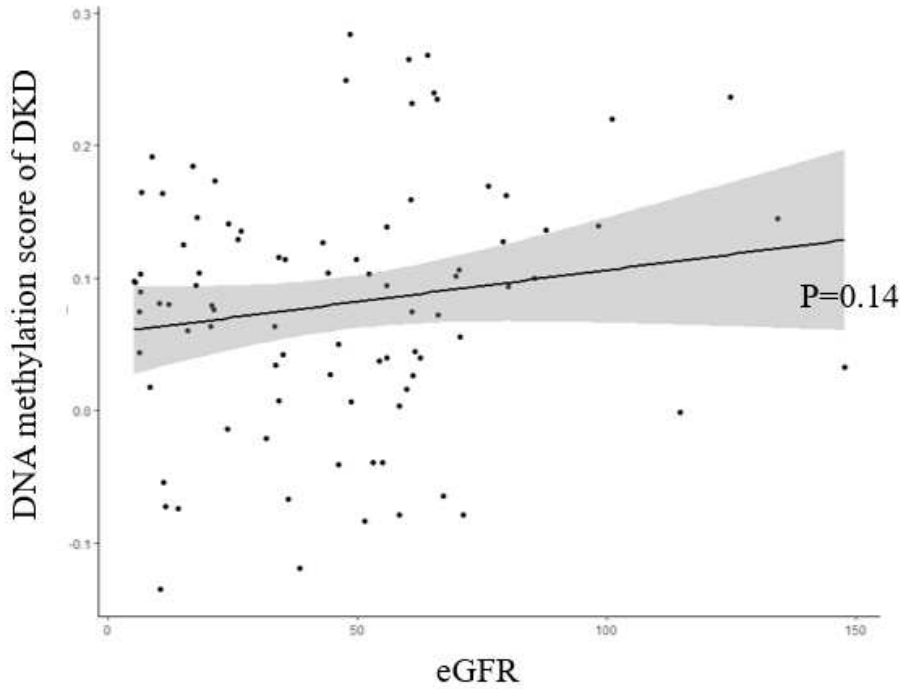


Table 17. DNAm score difference of type 2 diabetes/DKD from MWAS of comparing DKD and non-diabetic controls

	Adjusted M-values Differences (β-values)	*P*
DMSs of Type 2 diabetes		
cg25139493	0.155 (0.175)	0.006
cg26823705	-0.108 (-0.114)	0.008
cg26974062	-0.248 (-0.249)	$3.7x10^{-4}$
cg19693031	-0.133 (-0.125)	0.049
cg14530801	0.138 (0.137)	0.285
cg08867893	0.167 (0.151)	0.019
cg00574958	-0.151 (-0.149)	0.005
cg17082373	0.011 (0.010)	0.881
DMSs of DKD		
cg16944159	0.102 (0.097)	0.033
cg16079347	0.072 (0.068)	0.036
cg11530914	0.136 (0.130)	0.004

Figure 21. QQ plot of MWAS from comparing DKD and non-diabetic controls

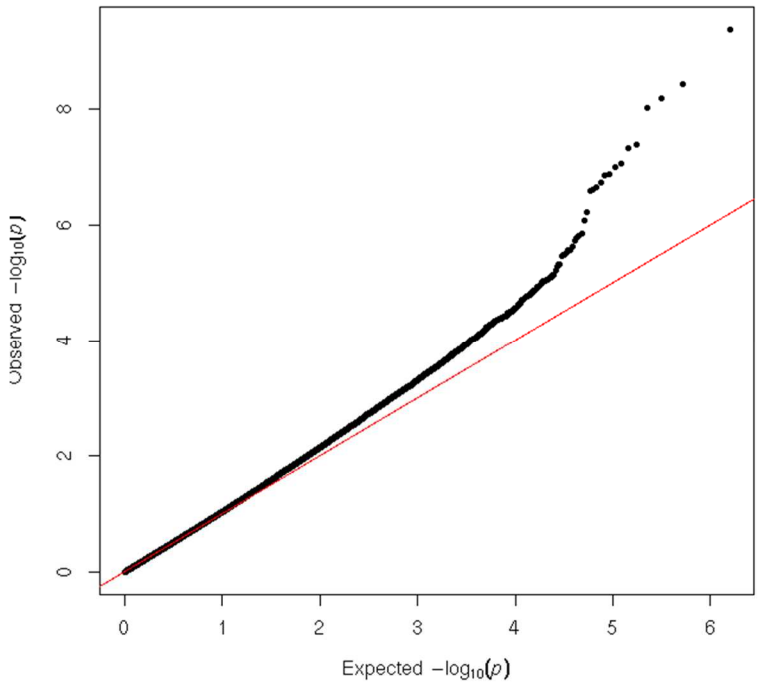


Table 18. IV validation of MR analysis using DKD mQTL as an IV

	Coefficient	P-value	F-statistics	R-square
cg16944159	63.88 (9.39)	<0.001	46.32	0.253
cg16079347	99.35 (15.51)	<0.001	41.03	0.23
cg11530914	22.45 (8.01)	0.006	7.85	0.049

Table 19. Mendelian randomization of the DMSs on DKD

	OR [95% C.I]	P
cg16944159	1.15 [1.07-1.23]	**<0.001**
cg16079347	1.20 [0.87-1.66]	0.263
cg11530914	1.19 [1.07-1.31]	**0.001**

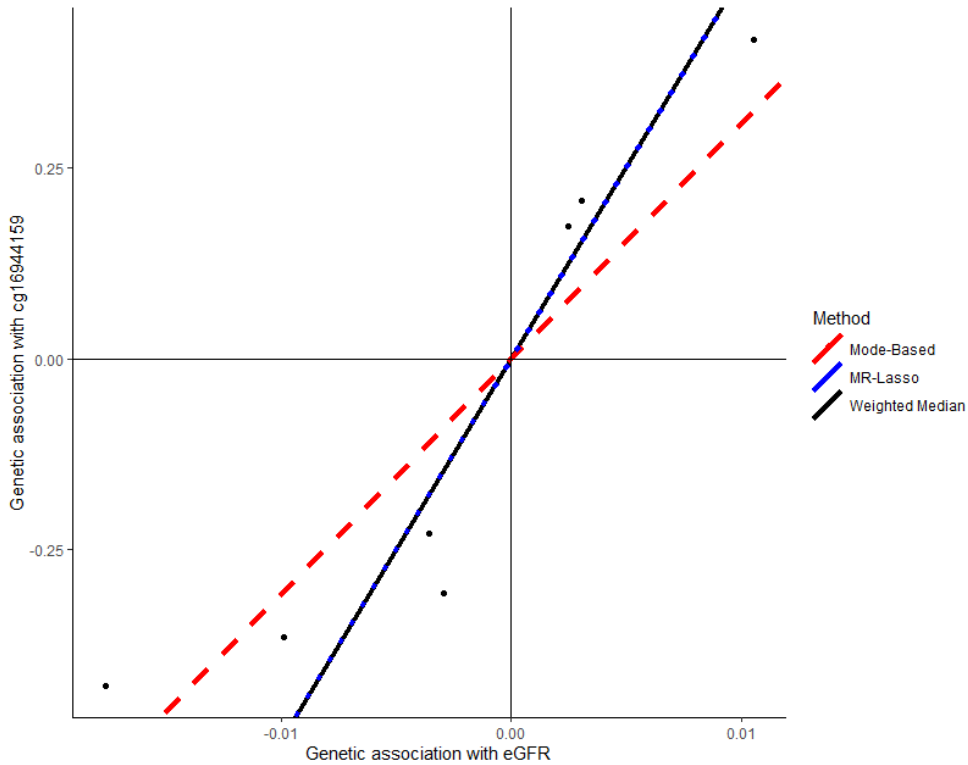
Abbreviation: OR, Odds ratio

Table 20. Causal association of DNAm on DKD using adjusted Mendelian randomization of significant results from an individual level MR

	Estimate	SE	P
Pleiotropy test in MR-Egger (MR-Egger intercept)	0.142	0.070	0.042
Weighted-median	50.32	14.74	<0.001
Mode-based	39.92	13.58	<0.001
MR-Lasso	**50.36**	**8.26**	**<0.001**

Bold style font indicates recommended method considering scenarios of MR-assumption.

Figure 22. Plot of adjusted MR analysis of cg16944159 and DKD



Supplementary Table 5. The entire information of differential methylated CpGs for DKD

CpG	Nearby gene	CGI Information	Gene property	Position (hg 19)	Adjusted Differences (M-values)	Adjusted Differences (β-values)	Raw β differences	*P* *	FDR *
DKD									
cg16944159	*COMMD1*	Opensea	TSS200	chr 2:62132759	0.272	0.008	0.013	8.20E-08	1.63E-02
cg16079347	*TMOD1*	Island	5'UTR	chr 9:100265181	0.237	0.015	0.021	2.87E-09	2.19E-03
cg11530914	*FHOD1*	Island	-	chr 16:67281528	0.261	0.008	0.015	7.97E-08	1.63E-02

TSS200 is the region from TSS to − 200 nt upstream of TSS; TSS1500 covers − 200 to − 1500 nt upstream of TSS;
Bold style font indicates novel CpG sites which are not previously reported to be associated with type 2 diabetes
*P and FDR corresponding to M-values
Abbreviation: chr, chromosome; CGI, CpG island; FDR, False discovery rate; FDR, hg19, human genome version 19; TSS, Transcription start site; UTR, Untranslated Region

Supplementary Table 6. Effect of known type 2 diabetes DMRs found in DKD MWAS

CpG	Position (hg 19)	Nearby gene	Adjusted Differences (M-values)	*P* *	Adjusted Differences (β-values)
cg15585213	chr4:15341878	*C1QTNF7*	-2.36E-03	0.160	-6.94E-02
cg04344749	chr7:71800412	*CALN1*	2.42E-03	0.542	1.39E-02
cg25130381	chr11:68607622	*CPT1A*	-6.82E-03	0.173	-5.58E-02
cg03725309	chr6:31804883	*C6orf48*	-3.99E-04	0.873	5.00E-03
cg14476101	chr7:75586348	*POR*	9.33E-03	0.371	5.79E-02
cg19693031	chr2:11969958		8.60E-03	0.374	5.93E-02
cg19266329	chr17:47301614	*PHOSPHO1*	1.35E-02	0.018	7.91E-02
cg09247619	chr19:47287964	*SLC1A5*	-2.20E-05	0.995	1.82E-04
cg02560388	chr16:4751569	*ANKS3*	-1.80E-04	0.965	-1.44E-02
cg06721411	chr16:88849875	*FAM38A*	-4.76E-04	0.926	-3.53E-03
cg04992150	chr1:109757585	*SARS*	2.36E-04	0.927	4.22E-03
cg00076653	chr1:25871814	*LDLRAP1*	2.26E-03	0.706	1.26E-02
cg15962267	chr11:34393106		2.55E-03	0.588	1.49E-02
cg01657995	chr19:4061359	*ZBTB7A*	-6.30E-03	0.112	-1.09E-01
cg13199639	chr7:1066650	*C7orf50*	7.87E-04	0.507	2.98E-02
cg16809457	chr10:74057977		-7.22E-03	0.267	-4.81E-02
cg17666418	chr3:13457267	*NUP210*	-5.29E-03	0.163	-1.01E-01
cg22909677	chr10:32049926		-1.58E-03	0.682	-3.16E-02

cg13640297	chr17:55169508	*AKAP1*	-9.36E-03	0.084	-5.83E-02
cg04816311	chr16:88850218	*FAM38A*	-9.27E-03	0.208	-5.80E-02
cg13514042	chr19:42599631	*POU2F2*	-3.00E-03	0.501	-3.01E-02
cg26804423	chr11:67373114	*NDUFV1*	-2.91E-04	0.946	-3.82E-03
cg21699330	chr21:43656587	*ABCG1*	-2.38E-03	0.526	-4.23E-02
cg00277397	chr2:74753759	*DQX1*	-5.91E-03	0.430	-4.30E-02
cg01676795	chr11:46382544	*DGKZ*	-2.49E-02	**0.002**	-1.48E-01
cg24531955	chr8:119208486	*SAMD12*	-2.83E-04	0.946	-7.41E-03
cg07960624	chr21:35320596		-1.23E-02	0.123	-8.93E-02
cg08788930	chr8:142201685	*DENND3*	-7.61E-03	0.198	-5.19E-02
cg08994060	chr10:6214026	*PFKFB3*	1.17E-03	0.856	1.27E-02
cg05400498	chr1:198648849	*PTPRC*	2.25E-03	0.638	1.51E-02
cg04973995	chr17:76355146	*SOCS3*	1.40E-04	0.972	-6.36E-03
cg26712428	chr14:74227441	*C14orf43*	2.47E-02	0.068	1.94E-01
cg04645070	chr17:17730094	*SREBF1*	-4.55E-03	0.134	-5.82E-02
cg07092212	chr15:91455407	*MAN2A2*	1.25E-03	0.481	2.44E-02
cg11376147	chr11:57261198	*SLC43A1*	1.63E-04	0.934	7.40E-03
cg06235429	chr6:33360495	*KIFC1*	-4.97E-03	0.269	-4.32E-02
cg00574958	chr7:1192202		4.58E-03	0.065	1.10E-01
cg17058475	chr6:170099238	*WDR27*	3.84E-03	0.264	4.70E-02
cg17315426	chr17:72764985	*SLC9A3R1*	-1.49E-03	0.465	-2.95E-02
cg10919522	chr1:120255992	*PHGDH*	-1.41E-03	0.750	-1.17E-02
cg14597545	chr15:73074210	*ADPGK*	-7.74E-03	0.115	-4.65E-02

cg11183227	chr1:14146635	*PRDM2*	-8.70E-03	0.079	-8.33E-02
cg03497652	chr5:138612986	*SNHG4,SNORA74A, MATR3*	-1.80E-02	**0.005**	-1.15E-01
cg26546155	chr6:90399677	*MDN1*	-1.62E-03	0.736	-1.00E-02
cg03699074	chr11:68607737	*CPT1A*	-4.05E-04	0.917	-2.52E-04
cg06007201	chr14:23527629	*CDH24*	-3.02E-05	0.988	-3.76E-03
cg11024682	chr6:106612002		9.78E-04	0.822	5.97E-03
cg02650017	chr17:76354621	*SOCS3*	-1.29E-03	0.381	-4.72E-02
cg05778424	chr1:145456128		-2.22E-03	0.662	-1.45E-02
cg14020176	chr1:145441552	*TXNIP*	-9.45E-03	**0.029**	-5.51E-02
cg18181703	chr7:26193032	*NFE2L3*	-9.90E-03	0.133	-6.05E-02
cg10508317	chr19:47288066	*SLC1A5*	1.06E-03	0.626	2.09E-02
cg04727071	chr6:109172312	*ARMC2*	-6.37E-03	0.135	-5.12E-02
cg06178887	chr8:23154691	*LOXL2*	-7.39E-03	0.098	-4.70E-02
cg26836479	chr1:27440721	*SLC9A1*	-1.00E-03	0.582	-4.02E-02
cg02711608	chr16:57269635	*RSPRY1*	-3.48E-03	0.273	-4.54E-02
cg21766592	chr10:105218771	*CALHM1*	-3.58E-03	0.141	-1.34E-01
cg08309687	chr7:8201134	*ICA1*	1.60E-02	**0.045**	9.81E-02
cg06500161	chr19:42706353	*DEDD2*	-4.72E-03	0.347	-2.91E-02

Bold style font indicates P under nominal significance (P <0.05)

*P-value corresponds to M-value.

Abbreviation: chr, chromosome; hg19, human genome version 19

Supplementary Table 7. Mendelian randomization of DKD DMSs on metabolic phenotype after adjusting diabetes duration

MR	cg16944159		cg16079347		cg11530914	
	T2D duration adjusted model	Raw model	T2D duration adjusted	Raw model	T2D duration adjusted	Raw model
	Coefficients (*P*)					
BMI	0.29 (0.14)	0.62 (0.63)	0.15 (0.69)	-0.74 (0.73)	0.21 (0.38)	-0.45 (0.68)
Fasting glucose	0.17 (0.82)	0.73 (0.40)	1.70 (0.26)	1.86 (0.26)	1.52 (0.08)	1.32 (0.16)
HbA1c	-0.02 (0.77)	-0.10 (0.35)	0.12 (0.54)	0.12 (0.57)	0.15 (0.16)	0.17 (0.17)
eGFR	**-4.94 (<0.001)**	**-1.08 (0.001)**	**-8.37 (0.003)**	**-1.88 (0.002)**	**-4.11 (0.004)**	**-0.93 (0.006)**
DKD	0.32 (0.20)	0.32 (0.21)	0.20 (0.59)	0.20 (0.60)	0.14 (0.52)	0.14 (0.50)

Abbreviation: BMI, Body mass index; eGFR, estimated Glomerular Filtration Rate; HbA1c, Glycated hemoglobin; SBP, Systolic blood pressure

Bold style font indicates P under nominal significance (P <0.05)

Raw model: not adjusted for diabetes duration

IV. Causal Analysis of Differentially Methylated Regions Associated with type 2 diabetes/DKD

1. Methods and materials

1.1. Study participants

To investigate the association of the identified DMSs with diabetes-related metabolic traits, two additional population-based cohorts were used: 1) KoGES consisting of 356 non-diabetic controls [52] and 2) HTS, consisting of 463 non-diabetic controls [53]. These two cohorts were analyzed together with non-diabetic control of the SNUH cohort.

1.2. DNA profiling

DNA profiling of SNUH and KoGES cohort was described in **Chapter II**. For the HTS samples, 140 and 370 were assayed using EPIC and Illumina 450K BeadChip (450K), respectively.

A similar DNAm QC pipeline for HTS was applied as in **Chapter II**. The raw methylation data were processed using R package RnBeads [55] and quantified using beta value. The procedures for QC and normalization of HTS are detailed in **Figure 23**. The methylation beta value was normalized using the BMIQ method [56]. Cell type composition was calculated using the ReFACTor algorithm in the python toolset GLINT [57]. The analyses were performed using the R software (version 3.5.1) and Python 2.7.

1.3. Genotyping the data

The KoGES and HTS participants were genotyped using Illumina Omni1 (KoGES,

rural), Affymetrix 6.0 (KoGES, urban; HTS), and 5.0 (KoGES, KARE) genotyping arrays. The single nucleotide polymorphisms (SNPs) were filtered by following genotype quality control (QC) criteria: (1) genotyping call rate >0.95 (2) minor allele frequency (MAF) >0.01, (3) P-value in Hardy–Weinberg equilibrium (HWE) testing >10e-6 and individual QC criteria: sample call rate >0.9. Those after QC were imputed using the Korean data from Korean Reference Genome (KRG) which initiated by the Center of Genome Science (CGS) of Korea National Instituted of Health (KNIH) in 2012 [136] and East Asian (EAS) data from the 1,000 Genomes Project Phase 3 (NCBI build 37) as the reference. We used IMPUTE 2.0 software to impute variants that were not directly genotyped [137]. Only SNPs with INFO scores larger than 0.6 were included.

1.4. Association with quantitative metabolic traits

We explored whether DMSs significantly associated with type 2 diabetes or DKD were associated with diabetes-related metabolic traits including BMI, fasting glucose, HbA1C, and eGFR in the non-diabetic controls. The analysis was performed in the combined 1,018 non-diabetic controls from the SNUH, KoGES, and HTS cohorts. Only non-diabetic controls were used for this analysis, as type 2 diabetes cases might have been affected by diabetes-related treatment or medications. MWAS for each metabolic trait was done separately in each of the three control datasets with age, sex, and BMI as covariates. Meta-analysis was performed using the R package meta for a fixed effect model [138].

1.5. MR analysis

The MR approach was implemented similarly in previous studies (**Chapter II and III**). Non-diabetic controls with genotype information were used in the MR analysis (N = 948). The MR was conducted in three ways; Model 1: to investigate the causal effect of the metabolic traits on the methylation of each DMSs, Model 2: to investigate the causal effect of type 2 diabetes/DKD on the methylation of each DMSs, and Model 3: to investigate the causal effect of methylation of each DMSs on metabolic traits (**Figure 24**). Genetic risk scores (GRSs) were built as instrument variables (IVs) for BMI, fasting glucose, HbA1c, eGFR, type 2 diabetes, and DKD. Single nucleotide polymorphisms (SNPs) used to build the GRSs were selected from the GWAS catalogue [139]. The process of selecting genetic markers is summarized in **Supplementary Table 8**. All SNPs reported to be associated with metabolic traits and type 2 diabetes /DKD on the GWAS catalog were extracted (Stage 1). Next, the availability of SNPs in Korean data after QC was checked (Stage 2). We only selected SNPs that replicated with target traits (Stage 3) and were not associated with consequence traits (For metabolic trait, type 2 diabetes/DKD; For type 2 diabetes/DKD, metabolic traits) in nominal significance ($P < 0.05$). Selected SNPs was not showed association with potential confounders (age, sex, and BMI). SNPs that showed nominal significance ($P < 0.05$) for each trait and showed no association with type 2 diabetes and DKD (for metabolic trait SNPs) or metabolic traits (for type 2 diabetes and DKD SNPs) were used to build IV for the MR analysis (**Supplementary Table 9**). GRS was tested whether they violated the assumptions of the MR approach or not, using F-statistic and R-square value. Two-stage least squares (2SLS) regression was conducted in the MR approach. The first stage of MR was regression of the IV on the exposure in non-diabetic controls and the second stage of MR was the regression of

predicted exposure by IV on the outcome trait. Standard errors of the 2SLS regressions were calculated by the bootstrapping method (1,000 replications). MR analysis was carried out using STATA version 14 (StataCorp LP, College Station, TX, USA).

Implementing the MR analysis without considering pleiotropy effects could lead to bias. To confirm that the IV shows a consistent association with the outcome without the condition of exposure (horizontal pleiotropy), robust MR methods using summary data estimation: weighted median, mode-based and MR-Lasso method using the R package 'Mendelian Randomization' [109]. The best performance method is selected by the process as described in **Chapter III**. The best method in different scenarios with the number of IV=30 is presented in **Figure 25**. MR-Lasso overwhelmed other methods in all of the scenarios.

2. Results

2.1. Association of identified DMSs with metabolic traits

To gain further insight into the DMSs of type 2 diabetes and DKD, we performed MWAS for metabolic traits of BMI, fasting glucose, HbA1c, and eGFR (**Supplementary Table 10**) in 1,018 non-diabetic controls from the SNUH, KoGES, and HTS cohorts. Association *P* values are shown in **Figure 26.** Multiple DMSs associated with metabolic traits overlapped between the DMSs of type 2 diabetes. Hypo-methylation of two DMSs (cg19693031 in *TXNIP* and cg00574958 in *CPT1A*), which were also hypo-methylated in type 2 diabetes, were significantly associated with an increase in fasting glucose levels after Bonferroni correction ($P < 0.00625$, 0.05/8). Hypo-methylation of three DMSs (cg26974062 and cg19693031 in *TXNIP*, and cg00574958 in *CPT1A*) of type 2 diabetes were

associated with an increase in HbA1c levels. The DMS of cg00574958 at *CPT1A* was also associated with BMI. None of the DKD DMSs were associated with metabolic traits in non-diabetic controls.

2.2. Causal association of metabolic traits and type 2 diabetes/DKD with CpG methylation

Using the MR method, we further investigated the causal effect of five metabolic traits (BMI, fasting glucose, HbA1c, and eGFR) on the DMSs of type 2 diabetes and DKD in 1,018 non-diabetic controls. The F-statistics of the GRSs, indicating the power of IVs, were large enough except for DKD (**Table 21**). Even though the power of IV for DKD was limited due to the small sample size, the R-square value was 0.04, suggesting that the IVs captured an acceptable level of phenotypic variation. Also, GRSs were not associated with confounding variables such as age or sex (**Supplementary Table 11, Supplementary Figure 2-3**).

The result of the MR is summarized in **Table 22-23 (Detailed in Supplementary Table 12).** First, we studied the causal effect of the metabolic traits on the methylation changes of 11 DMSs. An increase of 1 mmol/L in fasting glucose resulted in 2.04% hypo-methylation of cg00574958 at *CPT1A* (P = 0.008). As eGFR increased by10 ml/min, cg19693031 at *TXNIP* was hypo-methylated by 7.02% (P = 0.045), and three DKD DMSs (cg16944159 at *COMMD1*, cg16079347 at *TMOD1*, cg11530914 and *FHOD1*) were hypo-methylated by 0.93% to 1.88% ($P < 0.05$).

Significant individual level MR results were cross-checked using robust MR methods (Weighted median, mode-based, and MR-Lasso). The causal association of cg00574958-fasting glucose was still statistically significant after heterogeneity

and pleiotropy of SNPs were considered (**Table 24, Figure 25-26**). However, the direction of the cg19693031-eGFR association was opposite in the adjusted-MR method compared to the 2SLS regression result. In addition, the causal effect of eGFR on cg11530914 was nullified when robust MR methods were applied (**Figure 29**).

Next, the causal effect of type 2 diabetes and DKD on the methylation changes of 11 DMSs was investigated**.** The probability of type 2 diabetes and DKD had no effect on the methylation of the 11 DMSs (**Table 25).**

2.3. Causal association of CpG methylation with metabolic traits and type 2 diabetes/DKD

The causal association of CpG methylation with metabolic traits and type 2 diabetes was also investigated. MR was performed using discovered mQTLs in **chapter II-III** as an IV. In individual-level analysis, decrease methylation of cg26823705 and cg17082373 is causally associated with HbA1c increase (effect size: -0.04, -0.31 for each, $P<0.001$ for both). Also increase methylation of cg08867893 increased HbA1c level (effect size: 0.11, $P<0.001$). The effect directions are the same as directions expected from that found in type 2 diabetes MWAS. In the robust MR method, the causal association of two CpGs (cg26823705, cg08867893) on HbA1c remained significant (**Table 26 and Figure 30**). There is a pleiotropic effect on two MR results according to MR-Egger intercept statistics (P=0.002 and P=0.025 for each) although the size of the intercept was small enough to ignore. For cg17082373, robust MR methods were not used as the number of IVs is not enough to apply.

3. Discussion

In this study, we thoroughly investigated 11 DMSs (9 for type 2 diabetes and 3 for DKD) that overlapped between metabolic traits and type 2 diabetes/DKD and the causal effect of metabolic traits on type 2 diabetes/DKD and vice versa by using two-stage MR method. DMSs of type 2 diabetes were highly overlapped with DMSs of metabolic traits such as BMI, fasting glucose, HbA1c, where DMSs of DKD were not. Specifically, fasting glucose level increasing was causally related to hypomethylation of cg0057495. These findings support the hypothesis that glycemic variation influences DNAm in type 2 diabetes. Also, three DMSs of DKD showed a causal association with eGFR. To our knowledge, it is the first study that explored the association between DMSs of type 2 diabetes/DKD and metabolic traits.

MR uses the genetic variation as an IV to examine the causal association between exposure and related outcome. As genetic variants are randomly transmitted to offspring from parents, it is not influenced by confounding factors like in observational studies. As it has been discussed that DNAm changes are simply secondary consequences that occur during disease progression [140], investigating causal relationships using the MR method would help to understand the role of DNAm in type 2 diabetes/DKD.

Metabolic fluctuation plays a major role in DNAm changes [141]. The methyl group attached to CpG during the DNAm process is made through the methionine cycle where methionine, folate, choline, betaine, and vitamins B2, B6, and B12 are required [142]. Also, demethylation is regulated by TET family enzymes which are regulated by the cellular metabolic environment [143]. In this background, metabolic perturbation causes DNAm changes and finally leads altered gene expression that is involved in the pathogenesis of type 2 diabetes such as glucose

intolerance, insulin resistance, β-cell dysfunction, and other conditions [144]. In conclusion, the influence of metabolic conditions on DNAm changes should investigate thoroughly.

There are reports that cg19693031 in *TXNIP* is associated with fasting glucose, insulin resistance [26], and HbA1c [27], which are in accordance with our findings. Hypomethylation of cg00574958 in *CPT1A* was significantly associated with an increased risk of type 2 diabetes and an increase in BMI, fasting glucose, and HbA1c levels. This finding was consistent with previous reports [24, 29, 145]. It is worth noting that the MR analysis revealed evidence of a causal effect of fasting glucose on the hypomethylation of cg00574958 in *CPT1A*. Knockout of *CPT1A* was reported to reduce glucagon secretion in mice experiments and also replicated in human islets [146]. This suggests that hypomethylation of cg00574958 is a marker of hyperglycemia that could proceed with the development of type 2 diabetes. On the other hand, it has been reported that methylation at this DMS was causally associated with type 2 diabetes in an MR study using methylation quantitative loci [24]. The overlap between the DMSs of type 2 diabetes and metabolic traits and the findings from the MR analysis suggest that the metabolic effects may be engraved as epigenetic marks, which further modulate the risk of type 2 diabetes, at least in part.

As eGFR is one of the diagnosis criteria of DKD [147], we could not conclude that eGFR causes DNAm changes associated with DKD. In several studies, eGFR was investigated as a continuous indicator of DKD [132, 148]. As the causal relationship between eGFR and DNAm changes is observed to be null, there is the possibility of an unidentified pleiotropy effect in the observed causality of eGFR and DNAm. Genetic variants of eGFR have a pleiotropic association with 7 phenotypes (diabetes, coronary atherosclerosis, hypertension, essential

hypertension, pulmonary heart disease, phlebitis and thrombophlebitis, and ischemic heart disease) with $P < 5x10^{-8}$ through phenome-wide association study [149].

In opposite direction MR, three CpG sites (cg26823705, cg08867893, and cg17082373) affect the HbA1c level. As an IV, discovered mQTLs in **Chapter II-III** are used. However, as discussed in **Chapter II**, there are no mQTLs of type 2 diabetes reached genome-wide significance level, remaining the possibilities that mQTLs are discovered because they are related to other traits associated with type 2 diabetes. As so, we cannot rule out the possibility of violation of the InSIDE assumption. Therefore robust MR methods, which give unbiased estimation even though the InSIDE assumption is violated such as MR-LASSO, should be applied. Casual association of two CpG sites (cg17082373 was excluded due to lack of the number of SNPs) with HbA1c observed in individual MR analysis remains significant in MR-LASSO.

Unfortunately, the genetic mechanisms of three genes (*NBPF20, ZNF365*, and *TRIM37*) which induce HbA1c change are unknown. In our knowledge, the expression of these genes is associated with cancer or tumor proliferation. NBPF20 is a member of the neuroblastoma breakpoint family. Expression alteration of this gene family is associated with several types of cancer. ZNF35 is an important transcription factor that plays role in colorectal cancer, breast cancer, and hepatocellular carcinoma [150]. Expression of TRIM37 is upregulated in hepatic cancer samples and promotes tumor cell proliferation [151, 152]. There is a lot of evidence that a higher HbA1c level is associated with higher incidence and mortality risk for all cancers [153, 154]. The Discovery of relevance between DNAm of three genes and HbA1c could explain the association of HbA1c level and cancer risk.

This study has several strengths in this study. First, a large number of the study

sample is available for the analysis. We integrated several cohorts to evaluate the association between metabolic trait variations in non-diabetic controls. As the result, a total of 1,016 participants, which is enough number to get an unbiased result, are included in the analysis. Second, we examined the causal association in depth using different two MR models. To clarify whether the DNAm changes are the indicator of type 2 diabetes risk or the consequences of type 2 diabetes, we thoroughly investigated the MR models. Third, it is the first study that investigated metabolic trait effect in the DKD cases from a well-described clinical cohort.

Nevertheless, there are a few limitations to this study. First, a potential pleiotropic effect may have existed between exposure (metabolic traits and type 2 diabetes/DKD) and DNAm via residual association from SNPs. Although we examined the pleiotropic effect in available clinical variables, there could be a rest pleiotropic effect as described above. Second, study cohort heterogeneity between three cohorts could have induced the biased results in overlapped DMSs analysis, though we tried to overcome the heterogeneity problem by applying meta-analysis. Third, the presence of mQTL is one of the confounders in MR analysis. cg00574958 associated with type 2 diabetes showed trans-mQTL and cg16079347 associated with DKD showed cis-mQTL in the Biobank-based Integrative Omics Study database (https://genenetwork.nl/biosqtlbrowser/) [68]. Although, none of the reported SNPs associated with cg00574958 was replicated. Furthermore, among five SNPs in the BIOS database associated with cg16079347, one (rs2149981) was replicated in our data but not causally associated with DKD in MR analysis (P=0.03). Lastly, as the number of DKD patients was limited, the results about DKD should carefully be interpreted.

Figure 23. QC procedure of DNAm data for HTS

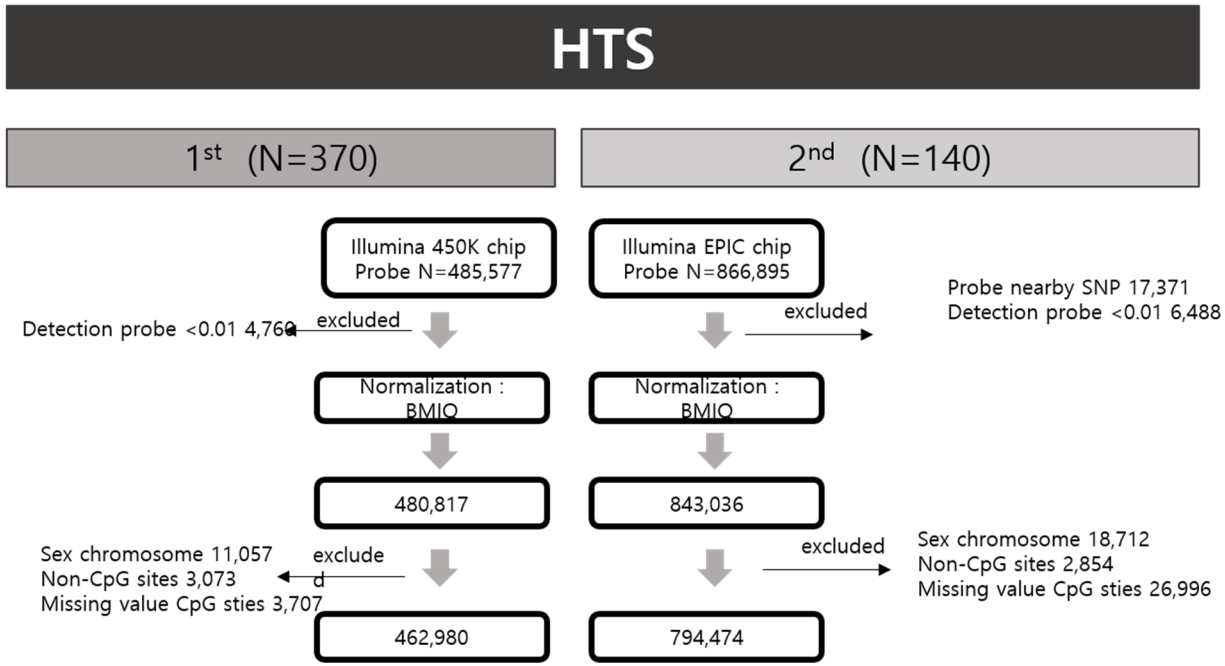


Figure 24. Diagram of Mendelian Randomization

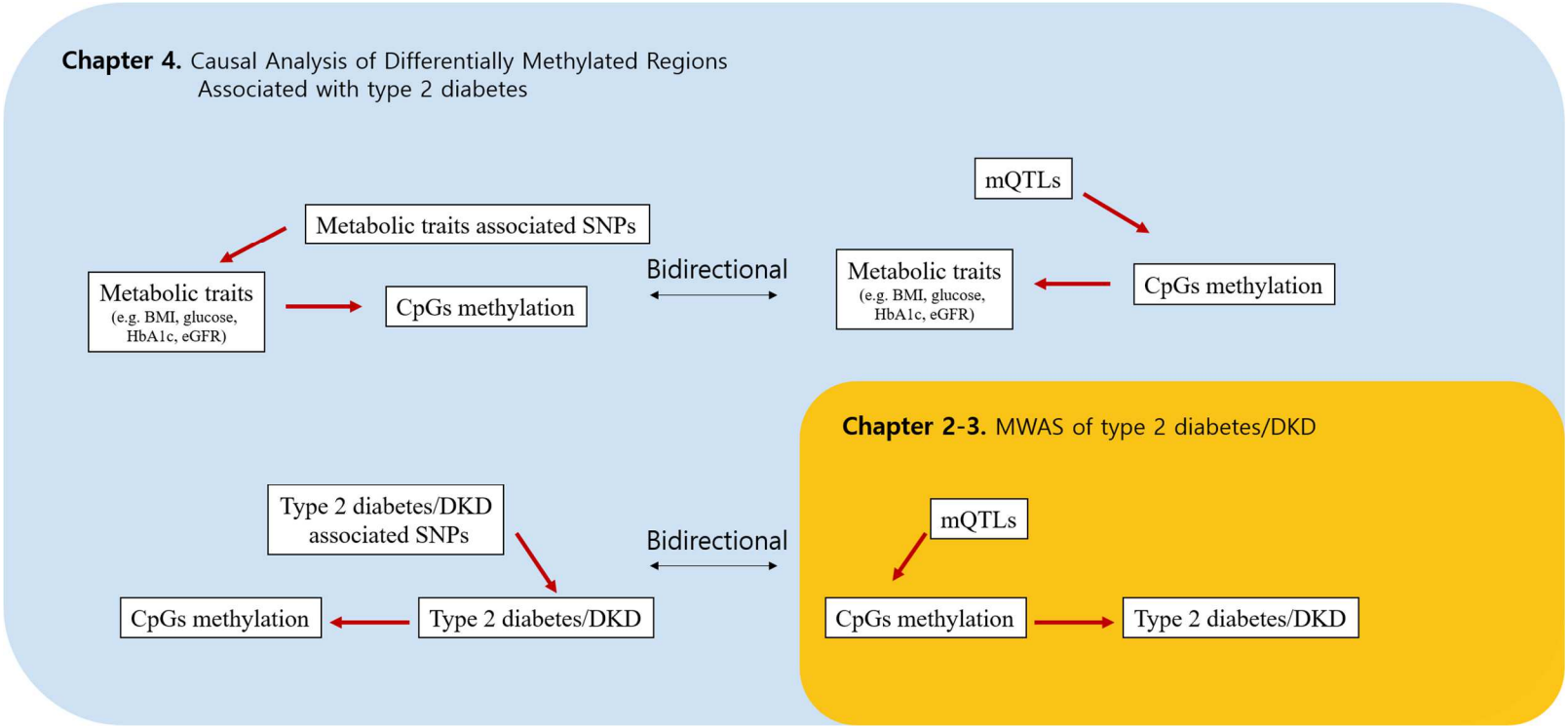


Figure 25. Recommended robust MR methods by scenarios of MR assumption (N of IV=30)

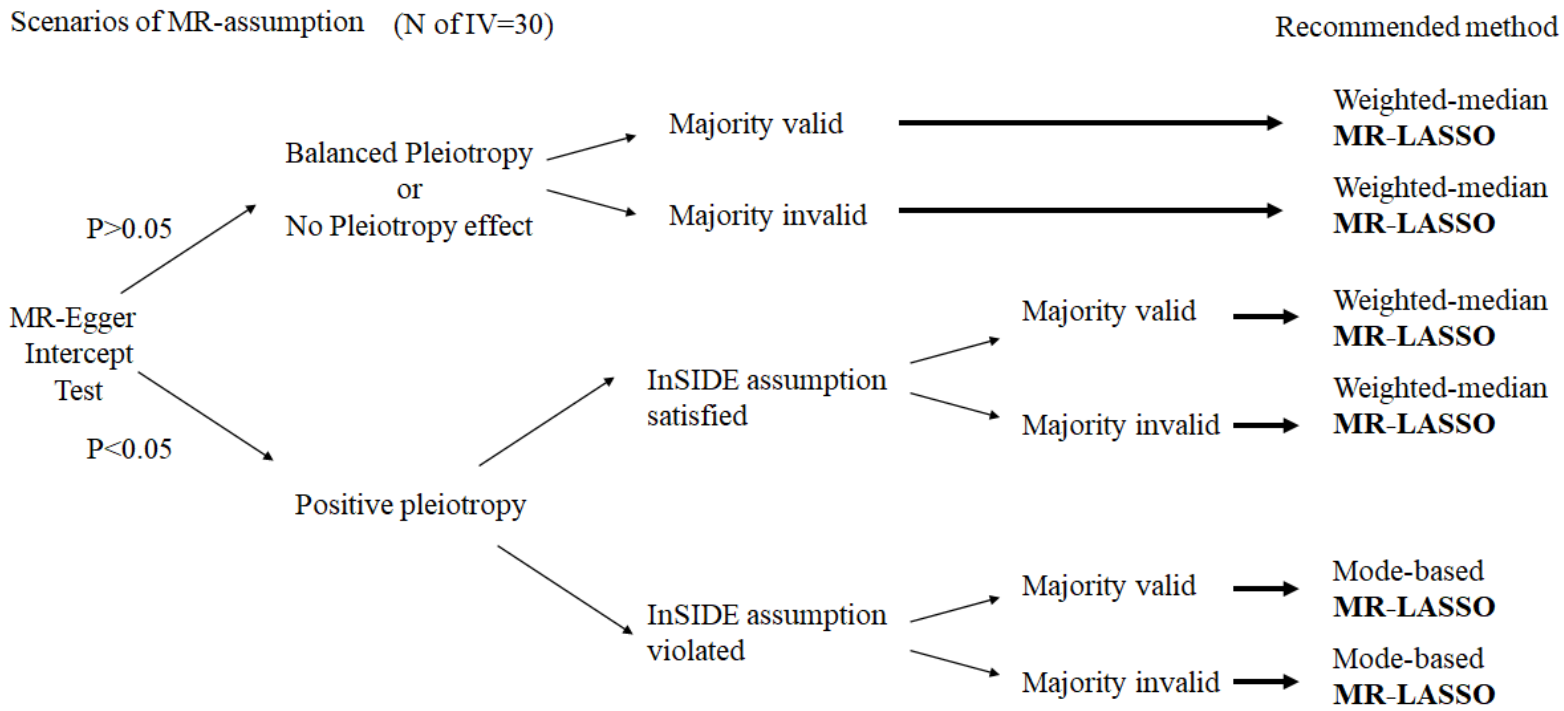
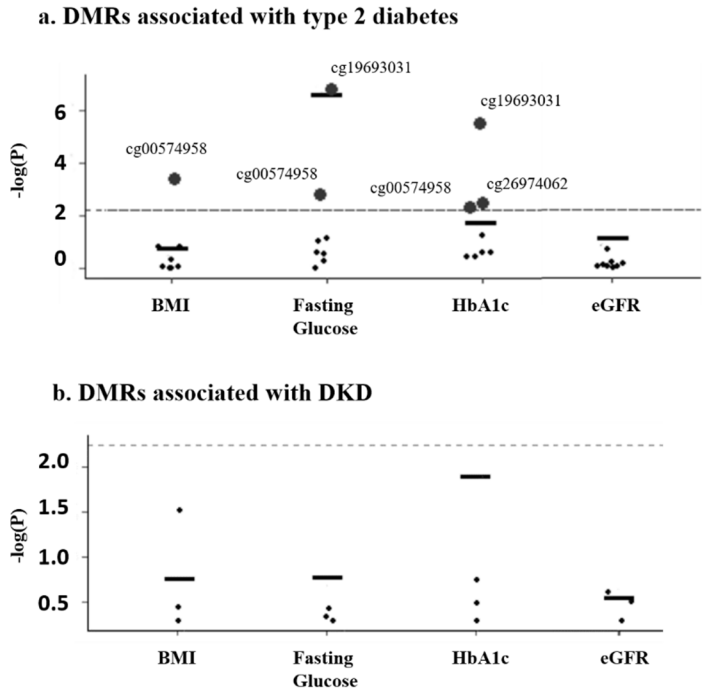


Figure 26. DNAm changes by metabolic traits in type 2 diabetes associated CpGs



Y-axis is the log-transformed *P* of each DMR in metabolic traits MWAS

The dashed line indicates Bonferroni adjusted significance threshold (0.0062, 0.05/8 for type 2 diabetes and 0.0167, 0.05/3 for DKD).

The horizontal line indicates the *P* of the overall SNP significance to metabolic traits derived by meta-analysis. Grey larger dots indicate *P* of the DMRs adjusted for Bonferroni correction ($P < 0.00625$, for type 2 diabetes).

Table 21. Validity of GRS

Variable	N of used SNPs	Coefficients ± SE	*P*	F-statistics/ LR chi-square	R-square
BMI	218	12.43 ± 0.78	<0.001	255.54	0.02
Fasting glucose (mmol/L)	27	1.87 ± 0.14	<0.001	168.73	0.01
HbA1c (mmol/mol)[a]	18	12.11 ± 1.03	<0.001	137.87	0.02
eGFR (ml/min)	33	1.04 ± 0.18	<0.001	33.56	0.01
Type 2 diabetes	64	1.84 ± 0.17	<0.001	115.14	0.01
DKD	1	2.38 ± 1.15	0.040	4.74	0.04

Abbreviation: BMI, Body mass index; eGFR, estimated Glomerular Filtration Rate; HbA1c, Glycated hemoglobin; SBP, Systolic blood pressure; SE, Standard error

Table 22. Mendelian randomization of the metabolic traits on type 2 diabetes

	Type 2 diabetes DMR							
	BMP8A (cg25139493)	***NBPF20*** (cg26823705)	***TXNIP*** (cg26974062)	***TXNIP*** (cg19693031)	***STX18*** (cg14530801)	***ZNF365*** (cg08867893)	***CPT1A*** (cg00574958)	***TRIM37*** (cg17082373)
	Coefficients (%) (*P*)							
BMI (Kg/m^2)	0.05 (0.65)	0.18 (0.83)	-0.32 (0.49)	-1.00 (0.08)	-0.05 (0.50)	0.04 (0.56)	-0.001 (0.99)	-0.10 (0.12)
Fasting Glucose (mmol/L)	-0.13 (0.78)	2.54 (0.41)	-1.29 (0.46)	3.19 (0.06)	0.04 (0.87)	0.13 (0.66)	**-2.04 (0.008)**	0.31 (0.23)
HbA1c (mmol/mol)	0.01 (0.86)	-0.003 (0.99)	-0.01 (0.96)	0.12 (0.47)	-0.004 (0.89)	0.03 (0.28)	-0.07 (0.39)	-0.01 (0.52)
eGFR (10ml/min)	-1.80 (0.05)	-8.19 (0.18)	-5.68 (0.09)	**-7.02 (0.045)**	-0.49 (0.18)	-0.85 (0.18)	-2.72 (0.06)	-0.93 (0.18)
	Coefficients (%) (*P*)							
Probability type 2 diabetes	0.11 (0.84)	4.77 (0.14)	0.44 (0.81)	-0.97 (0.64)	-0.19 (0.48)	-0.55 (0.07)	0.83 (0.35)	-0.39 (0.14)

Bold style font indicates *P* below statistical significance ($P < 0.05$).

The first stage of MR was regression of the IV on the exposure in non-diabetic controls and the second stage of MR was the regression of predicted exposure by IV on the outcome trait.

Abbreviation: BMI, body mass index; eGFR, estimated glomerular filtration rate; SBP, Systolic blood pressure

Table 23. Mendelian randomization of the metabolic traits on DKD

	DKD DMR		
	COMMD1 *(*cg16944159)	***TMOD1*** (cg16079347)	***FHOD1*** (cg11530914)
	Coefficients (%) (*P*)		
BMI (Kg/m^2)	0.62 (0.63)	-0.74 (0.73)	-0.45 (0.68)
Fasting Glucose (mmol/L)	0.73 (0.40)	1.86 (0.26)	1.32 (0.16)
HbA1c (mmol/mol)	-0.10 (0.35)	0.12 (0.57)	0.17 (0.17)
eGFR (10ml/min)	**-1.08 (0.001)**	**-1.88 (0.002)**	**-0.93 (0.006)**
	Coefficients (%) (*P*)		
Probability DKD	0.32 (0.21)	0.20 (0.60)	0.14 (0.50)

Bold style font indicates *P* below statistical significance ($P < 0.05$)**.**

The first stage of MR was regression of the IV on the exposure in non-diabetic controls and the second stage of MR was the regression of predicted exposure by IV on the outcome trait**.**

Abbreviation: BMI, body mass index; eGFR, estimated glomerular filtration rate; SBP, Systolic blood pressure

Table 24. Causal association of metabolic traits on DNAm using adjusted Mendelian randomization of significant results from individual level MR

	Estimate	SE	*P*
cg00574958-glucose			
MR-Egger (intercept)	0.002	0.001	0.264
Weighted-median	-0.002	0.0002	<0.001
Mode-based	-0.003	0.0004	<0.001
MR-Lasso	**-0.002**	**0.0002**	**<0.001**
cg19693031-eGFR			
MR-Egger (intercept)	0.003	0.001	<0.001
Weighted-median	0.006	0.005	<0.001
Mode-based	0.006	0.003	0.013
MR-Lasso	**0.006**	**0.0004**	**<0.001**
cg16944159-eGFR			
MR-Egger (intercept)	0.000	0.0004	0.884
Weighted-median	-0.0006	0.0004	0.077
Mode-based	-0.0004	0.0007	0.533
MR-Lasso	**-0.0006**	**0.0003**	**0.036**
cg16079347-eGFR			
MR-Egger (intercept)	0.001	0.001	0.455
Weighted-median	0.001	0.001	0.070
Mode-based	0.001	0.002	0.736
MR-Lasso	**0.001**	**0.001**	**0.037**
cg11530914-eGFR			
MR-Egger (intercept)	-0.001	0.001	0.350
Weighted-median	-0.0002	0.0004	0.666
Mode-based	-0.0003	0.001	0.558
MR-Lasso	**-0.0001**	**0.0004**	**0.673**

Bold style font indicates recommended method considering scenarios of MR-assumption.

Abbreviation: eGFR, estimated glomerular filtration rate

Figure 27. Adjusted Mendelian randomization plot of cg00547958-glucose association

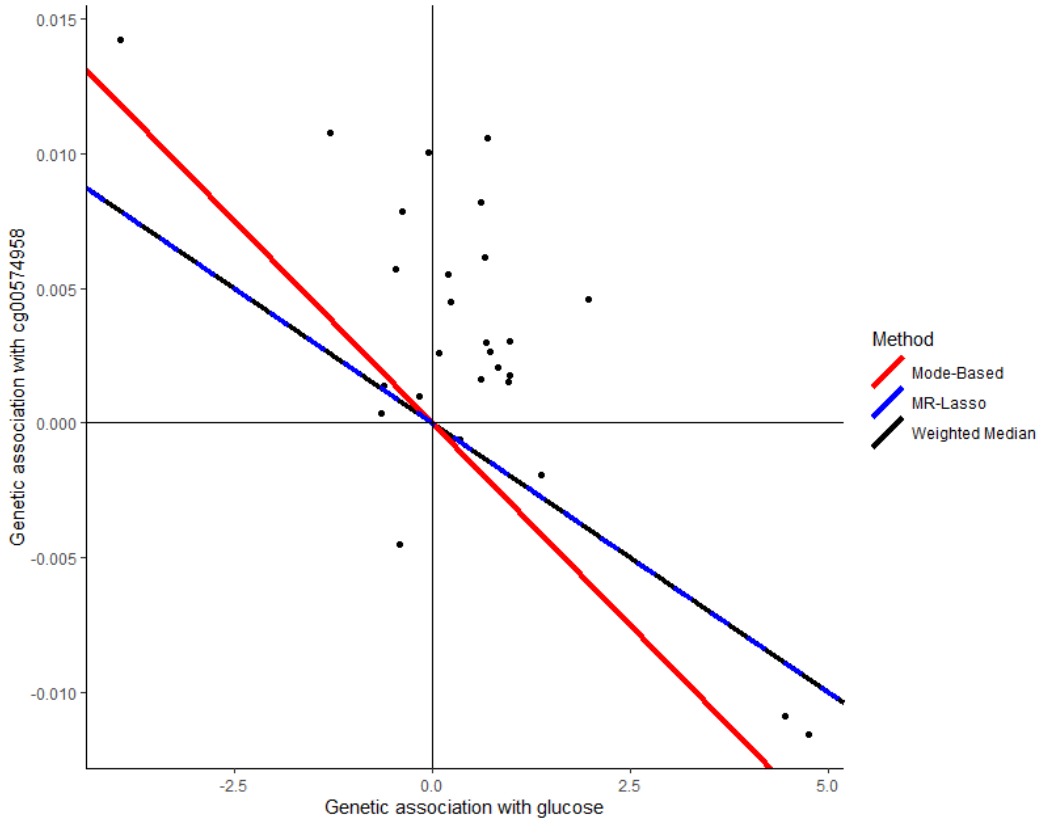


Figure 28. Adjusted Mendelian randomization plot of cg19693031-eGFR association

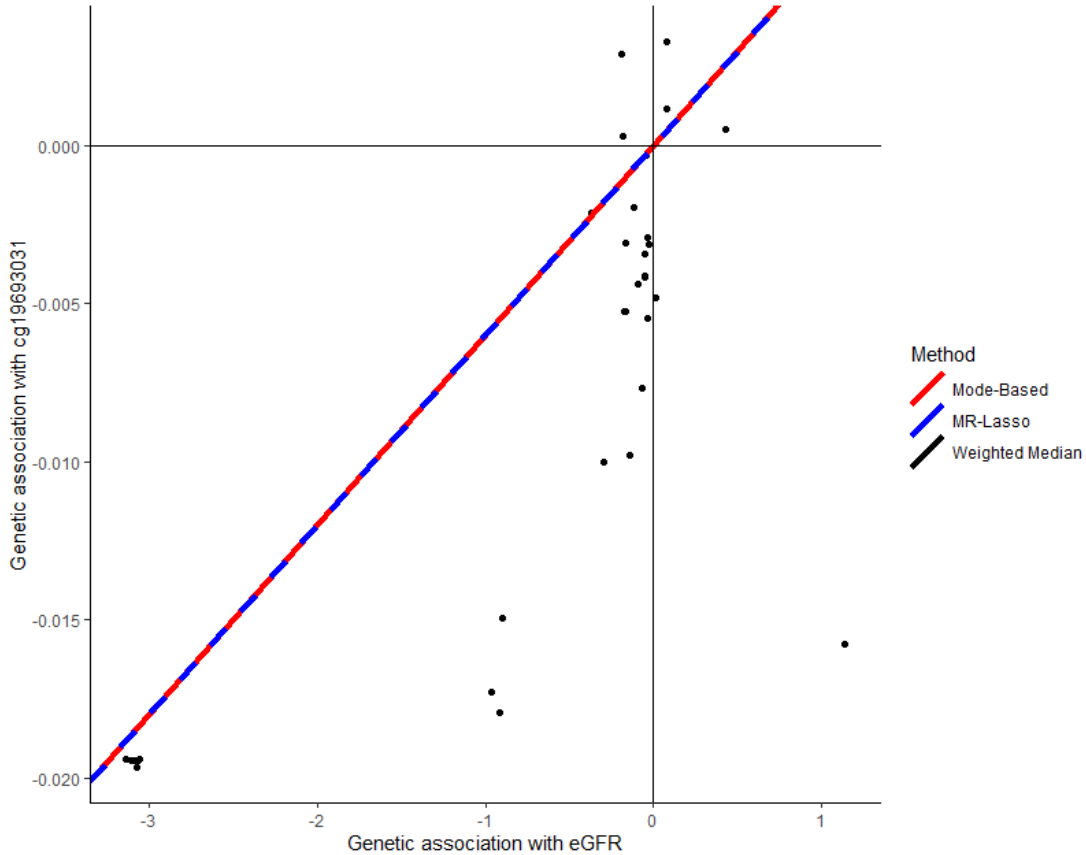


Figure 29. Adjusted Mendelian randomization plot of DKD DMSs-eGFR association

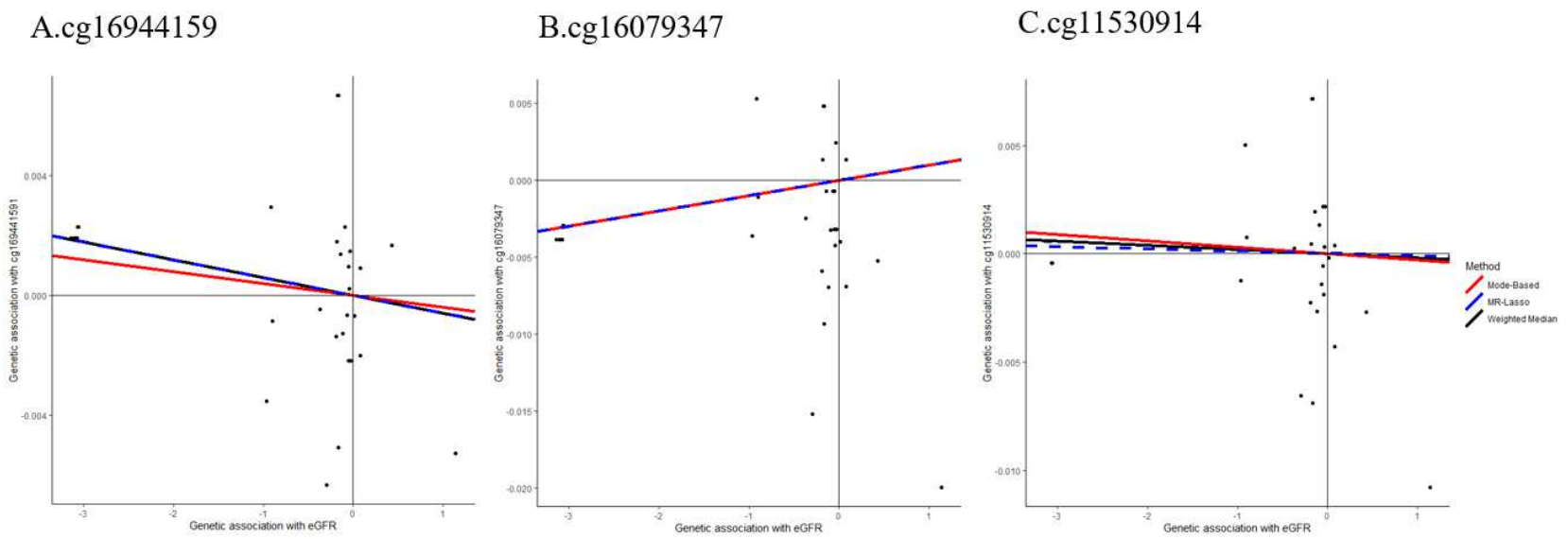


Table 25. Mendelian randomization of Type 2 diabetes/DKD on Type 2 diabetes/DKD associated CpG

Probability of type 2 diabetes			
	Coefficients	**SE**	***P***
cg25139493	0.0011	0.0055	0.842
cg26823705	0.0477	0.0324	0.141
cg26974062	0.0044	0.0180	0.806
cg19693031	-0.0097	0.0207	0.639
cg14530801	-0.0019	0.0027	0.482
cg08867893	-0.0055	0.0031	0.069
cg00574958	0.0083	0.0087	0.345
cg17082373	-0.0039	0.0027	0.141
Probability of DKD			
	Coefficients	**SE**	***P***
cg16944159	0.0032	0.0025	0.206
cg16079347	0.0020	0.0038	0.597
cg11530914	0.0014	0.0020	0.499

Abbreviation: BMI, Body mass index; eGFR, estimated Glomerular Filtration Rate; HbA1c, Glycated hemoglobin; SBP, Systolic blood pressure; SE, Standard error

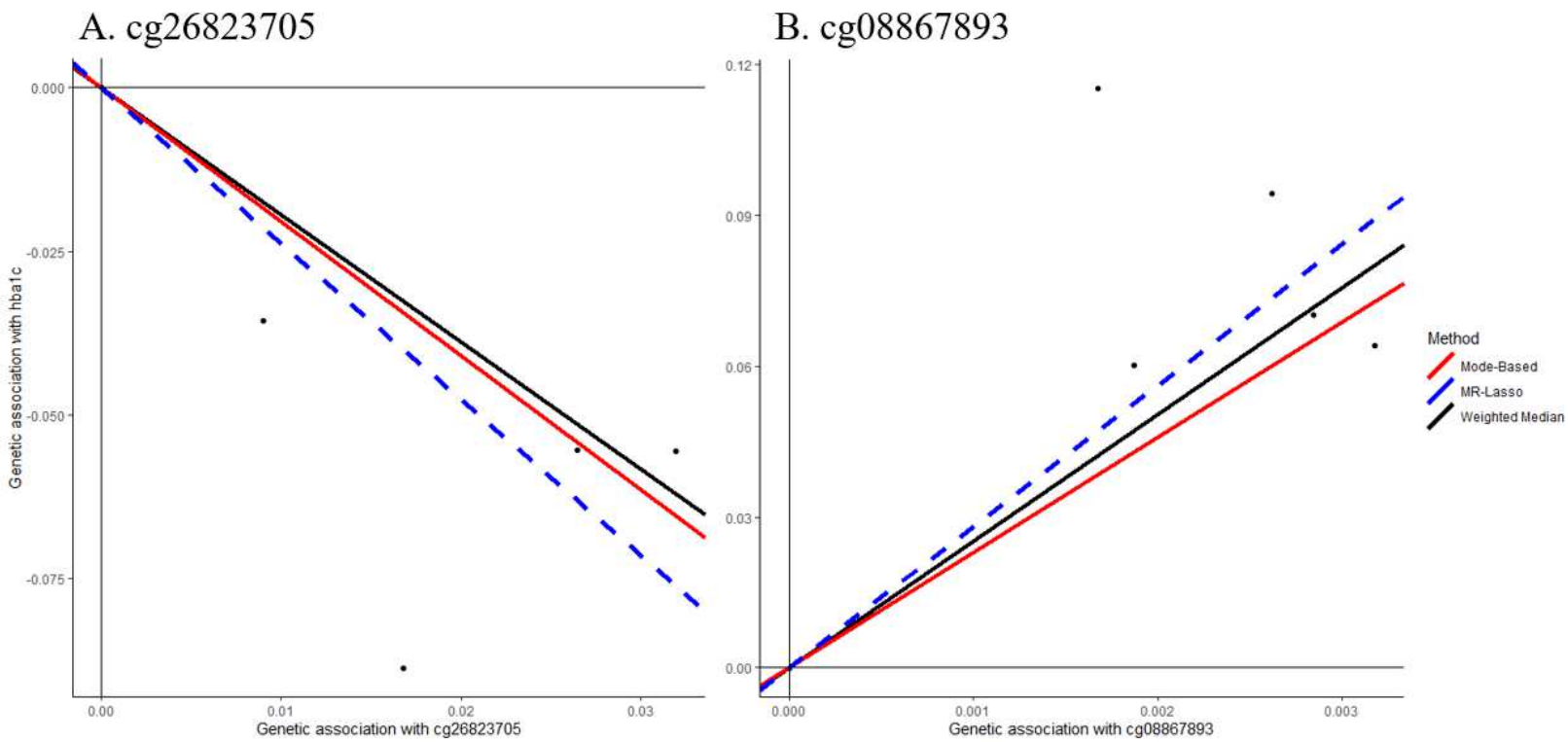
Table 26. Mendelian randomization of the DNAm on metabolic traits

Metabolic Traits	Coefficient (SE)	P	Coefficient (SE)	P
	cg25139493		**cg14530801**	
BMI	-0.08 (0.06)	0.194	0.35 (0.17)	0.044
Fasting glucose	-0.18 (0.35)	0.601	-1.97 (1.38)	0.154
HbA1c	0.00 (0.02)	0.936	0.02 (0.07)	0.744
eGFR	-0.01 (4.89)	0.999	5.68 (18.99)	0.765
	cg26823705		**cg08867893**	
BMI	-0.01 (0.02)	0.606	0.03 (0.06)	0.612
Fasting glucose	0.04 (0.17)	0.797	-0.11 (0.45)	0.8
HbA1c	-0.04 (0.01)	**<0.001**	0.11 (0.02)	**<0.001**
eGFR	-0.36 (2.12)	0.865	0.92 (5.36)	0.863
	cg26974062		**cg00574958**	
BMI	0.01 (0.02)	0.775	0.09 (0.06)	0.123
Fasting glucose	-0.25 (0.17)	0.129	-0.44 (0.38)	0.246
HbA1c	0.00 (0.01)	0.693	-0.01 (0.02)	0.773
eGFR	-0.82 (2.90)	0.777	-1.98 (5.27)	0.706
	cg19693031		**cg17082373**	
BMI	0.02 (0.01)	0.203	-0.22 (0.22)	0.313
Fasting glucose	-0.02 (0.09)	0.803	-2.29 (-2.29)	0.097
HbA1c	0.00 (0.00)	0.925	-0.31 (0.08)	**<0.001**
eGFR	1.17 (1.40)	0.401	-13.79 (20.38)	0.499
	cg16944159		**cg16079347**	
BMI	-0.01 (0.02)	0.617	0.02 (0.07)	0.73
Fasting glucose	-0.50 (0.40)	0.210	-0.58 (0.80)	0.47
HbA1c	-0.01 (0.01)	0.296	-0.03 (0.03)	0.30
eGFR	-0.35 (0.19)	0.07	0.12 (0.42)	0.78
	cg11530914			
BMI	0.01 (0.03)	0.81		
Fasting glucose	-0.76 (0.65)	0.240		
HbA1c	-0.03 (0.02)	0.192		
eGFR	-0.73 (0.47)	0.123		

Table 27. Causal association of DNAm on HbA1c using adjusted Mendelian randomization of significant results from the individual level MR

	Estimate	**SE**	***P***
HbA1c- cg26823705			
MR-Egger (intercept)	3.96E-06	1.30E-06	0.002
Weighted-median	-1.97	2.38	0.408
Mode-based	-2.11	0.85	0.013
MR-Lasso	-2.55	0.66	**<0.001**
HbA1c-cg08867893			
MR-Egger (intercept)	-7.36E-06	3.29E-06	0.025
Weighted-median	25.23	19.98	0.207
Mode-based	22.91	11.08	0.039
MR-Lasso	28.05	8.19	**0.001**

Figure 30. Adjusted Mendelian randomization plot of HbA1c-DNAm association



Supplementary Table 8. Number of SNPs used for IV in MR analysis

	BMI	Glucose	HbA1c	eGFR	Type 2 diabetes	DKD
1. Number of SNP previously reported	5834	126	180	1,117	2280	16
2. Available in our data	2477	115	175	838	1207	12
3. Replicated with metabolic traits/type 2 diabetes and DKD in our data	218	32	33	33	192	2
4. Not associated with type 2 diabetes/metabolic traits in our data	218	27	18	33	64	1

Supplementary Table 9. Information of SNPs used for IV in Mendelian randomization analysis

BMI						***Fasting glucose***					
Chromosome	SNP	RA	MAF	Coefficient	*P*	Chromosome	SNP	RA	MAF	Coefficient	*P*
1	rs3934834	T	0.09	0.29	0.01	1	rs2375278	A	0.26	1.32	<0.01
1	rs12044597	G	0.47	-0.15	0.03	2	rs780094	C	0.46	0.71	0.05
1	rs9660180	A	0.47	-0.15	0.03	2	rs12712928	C	0.38	0.67	0.01
1	rs4243830	C	0.14	0.2	0.03	2	rs560887	T	0.03	-2.16	0.05
1	rs12711521	C	0.36	0.14	0.05	2	rs12053049	C	0.35	0.78	0.04
1	rs491603	C	0.2	-0.17	0.04	3	rs8192675	C	0.21	-1.44	<0.01
1	rs2296172	G	0.14	0.19	0.05	3	rs11924032	A	0.18	-1.4	<0.01
1	rs1993709	A	0.01	-0.65	0.03	3	rs11705729	T	0.31	1.36	<0.01
1	rs4970658	A	0.04	0.35	0.04	6	rs7756992	A	0.46	-1.71	<0.01
1	rs10747472	G	0.19	-0.17	0.05	6	rs3765467	A	0.21	-1.6	<0.01
1	rs1749405	A	0.15	0.21	0.02	7	rs17168486	T	0.41	0.72	<0.01
1	rs34720381	T	0.07	0.46	<0.01	7	rs2191349	G	0.33	-1.67	<0.01
1	rs633715	C	0.25	0.23	<0.01	7	rs17140821	A	0.22	0.98	0.02
1	rs516636	A	0.25	0.23	<0.01	7	rs2971672	C	0.49	0.54	0.04
1	rs574367	T	0.25	0.24	<0.01	7	rs799165	A	0.14	1.36	<0.01
1	rs527248	G	0.24	0.24	<0.01	8	rs13266634	T	0.4	-1.29	<0.01
1	rs539515	C	0.25	0.24	<0.01	8	rs11558471	G	0.42	-1.5	<0.01
1	rs543874	G	0.25	0.24	<0.01	8	rs2980853	A	0.42	-0.91	0.01

1	rs545608	C	0.28	0.26	<0.01	9	rs57884925	G	0.41	1.1	0.05
1	rs591120	C	0.42	0.16	0.02	9	rs2383208	G	0.41	-1.27	<0.01
1	rs1998710	C	0.39	-0.14	0.04	10	rs7081476	C	0.02	2.76	0.02
1	rs340874	C	0.37	-0.16	0.02	11	rs2237896	A	0.4	-1.25	<0.01
1	rs6689335	C	0.28	0.15	0.04	11	rs11603334	A	0.06	-1.9	<0.01
1	rs12027371	T	0.28	0.15	0.04	12	rs79105258	A	0.18	-1.2	0.04
2	rs4854307	C	0.38	0.24	<0.01	12	rs10774625	A	0	23.36	0.03
2	rs10172678	T	0.33	-0.14	0.04	13	rs282587	G	0.02	2.69	0.03
2	rs4410328	T	0.25	-0.21	<0.01	20	rs1203936	G	0.16	-1.71	0.02
2	rs11892371	T	0.25	-0.21	<0.01	21	rs460976	A	0.32	0.78	0.04
2	rs4581940	C	0.25	-0.21	<0.01	***HbA1c***					
2	rs73966422	G	0.05	0.37	0.01	Chromosome	SNP	RA	MAF	Coefficient	*P*
2	rs222826	T	0.05	-0.37	0.02	2	rs12712928	C	0.38	0.04	0.02
2	rs3769885	A	0.29	0.19	0.01	7	rs6975024	C	0.19	0.05	0.02
2	rs355914	C	0.48	0.23	<0.01	8	rs13266634	T	0.4	-0.04	0.02
2	rs355810	G	0.48	-0.23	<0.01	8	rs11558471	G	0.42	-0.04	0.01
2	rs12477385	T	0.35	-0.28	<0.01	9	rs579459	C	0.26	0.04	0.03
2	rs61051952	G	0.37	-0.3	<0.01	11	rs174594	C	0.34	-0.04	0.05
2	rs1019612	C	0.24	-0.17	0.03	11	rs1387153	T	0.42	0.04	0.03
2	rs11679338	C	0.24	-0.18	0.02	11	rs10830963	G	0.43	0.03	0.04
2	rs1263629	G	0.23	0.17	0.03	12	rs2110073	T	0.04	0.12	<0.01
2	rs9973932	G	0.18	0.22	0.01	12	rs147538848	A	0.13	0.05	0.04

2	rs17203016	G	0.17	0.18	0.04	16	rs837763	C	0.32	-0.05	0.01
2	rs6435622	C	0.39	0.15	0.03	16	rs2968478	T	0.35	-0.04	0.04
3	rs2574704	T	0.36	-0.14	0.04	17	rs2748427	G	0.23	0.05	0.01
3	rs8192472	C	0.46	-0.17	0.02	17	rs2748424	G	0.14	0.08	<0.01
3	rs754635	C	0.31	-0.16	0.03	19	rs11667918	T	0.38	-0.04	0.01
3	rs10460960	G	0.31	-0.15	0.04	19	rs11086054	T	0.39	-0.04	0.03
3	rs1046953	T	0.23	0.25	<0.01	19	rs57601949	T	0.32	0.07	<0.01
3	rs10510757	G	0.33	-0.17	0.02	22	rs855791	G	0.48	-0.04	0.02
3	rs1916801	A	0.39	-0.18	0.01	***eGFR***					
3	rs815715	C	0.38	-0.16	0.02	Chromosome	SNP	RA	MAF	Coefficient	*P*
3	rs1289736	T	0.08	0.26	0.03	1	rs12563200	T	0.2	1.39	<0.01
3	rs4273371	C	0.19	-0.23	<0.01	1	rs4525087	A	0.39	0.79	0.04
3	rs7632505	G	0.18	0.22	0.01	1	rs78614739	T	0.02	4	<0.01
3	rs9857883	A	0.37	0.2	<0.01	1	rs10874312	G	0.47	-1.27	<0.01
3	rs1730028	T	0.37	0.19	<0.01	1	rs1887252	G	0.47	-1.23	<0.01
3	rs2682406	T	0.37	0.19	<0.01	2	rs807624	G	0.17	-1.36	<0.01
3	rs11924032	A	0.18	0.19	0.03	2	rs807603	T	0.17	-1.24	0.02
3	rs6809651	A	0.02	-0.58	<0.01	2	rs807601	G	0.17	-1.23	0.02
3	rs4234589	G	0.02	-0.57	<0.01	2	rs780094	C	0.46	-0.81	0.04
3	rs10513801	G	0.02	-0.57	<0.01	2	rs780093	C	0.46	-0.9	0.02
3	rs1516725	T	0.02	-0.57	<0.01	2	rs3788967	A	0.13	-1.1	0.05

3	rs73052033	C	0.02	-0.6	<0.01	2	rs57989581	A	0.08	1.58	0.01
3	rs7647305	T	0.04	-0.47	<0.01	2	rs16856823	T	0.2	1.34	<0.01
3	rs9816226	A	0.02	-0.78	<0.01	2	rs3770636	G	0.2	1.27	<0.01
3	rs7644678	G	0.15	-0.29	<0.01	2	rs35284526	A	0.09	2	<0.01
4	rs10938353	A	0.12	0.23	0.03	2	rs12471433	A	0.46	0.89	0.02
4	rs13130484	T	0.28	0.23	<0.01	2	rs34468415	G	0.08	1.94	<0.01
4	rs12641981	T	0.28	0.23	<0.01	2	rs35955110	T	0.46	0.99	0.01
4	rs12507026	T	0.28	0.23	<0.01	2	rs17581525	G	0.08	2.23	<0.01
4	rs10938397	G	0.28	0.23	<0.01	3	rs7651407	T	0.41	1.04	<0.01
4	rs4398538	C	0.28	-0.24	<0.01	3	rs67616801	G	0.49	0.84	0.03
4	rs2391540	T	0.17	0.18	0.05	4	rs17319721	A	0.1	-1.29	0.04
5	rs782971	A	0.34	-0.14	0.04	4	rs2725261	C	0.39	0.95	0.01
5	rs1503526	T	0.17	0.23	0.01	4	rs17263971	C	0.14	1.06	0.05
5	rs1895407	T	0.17	0.23	0.01	5	rs2010352	A	0.46	0.79	0.04
5	rs1895408	A	0.17	0.23	0.01	5	rs11747973	C	0.23	-0.91	0.04
5	rs4700608	T	0.17	0.23	0.01	6	rs9368805	T	0.17	-1.42	<0.01
5	rs6881648	A	0.43	0.13	0.05	6	rs7766720	C	0.16	-1.08	0.03
5	rs9942416	G	0.47	-0.31	0.03	7	rs6948759	T	0.22	0.94	0.04
5	rs13182155	T	0.17	-1.03	0.03	7	rs41301394	T	0.42	0.85	0.03
6	rs2206271	A	0.38	0.19	<0.01	7	rs868822	G	0.44	-0.81	0.04
6	rs2206277	T	0.3	0.2	<0.01	7	rs2365286	G	0.44	1.22	<0.01
6	rs987237	G	0.19	0.19	0.02	8	rs2442604	T	0.49	-1.07	<0.01

6	rs72892910	T	0.19	0.18	0.03	8	rs3758086	A	0.2	-1.12	0.02
6	rs4715210	T	0.2	0.19	0.02	8	rs36071802	C	0.25	-1.1	0.01
6	rs72887147	T	0.2	0.19	0.02	8	rs7007761	T	0.2	-1.22	<0.01
6	rs4141973	T	0.24	0.16	0.04	8	rs4871905	C	0.25	-1.29	<0.01
6	rs1321512	G	0.26	0.18	0.02	8	rs34861762	T	0.2	-1.29	<0.01
6	rs12199003	T	0.26	0.17	0.03	8	rs10109414	T	0.21	-1.13	0.01
6	rs947612	A	0.24	-0.16	0.04	8	rs1705694	A	0.19	-1.42	<0.01
6	rs7769594	T	0.08	0.28	0.02	8	rs1913641	G	0.41	1	<0.01
6	rs1179907	G	0.08	0.27	0.03	10	rs4918943	A	0.33	0.89	0.03
6	rs2781668	T	0.24	0.18	0.02	10	rs11191686	A	0.3	-1.09	<0.01
6	rs2246012	C	0.38	0.18	0.01	10	rs1536225	G	0.29	-0.84	0.05
6	rs227458	T	0	5.04	<0.01	11	rs3925584	C	0.36	1.19	<0.01
7	rs1167821	T	0.03	-0.43	0.04	13	rs7326821	G	0.03	-2.42	0.03
7	rs1167827	G	0.05	-0.39	0.02	15	rs476633	G	0.3	0.87	0.04
7	rs7777084	A	0.39	0.14	0.04	15	rs2928148	G	0.3	0.87	0.04
7	rs2237579	G	0.39	0.14	0.04	15	rs28522606	C	0.33	-0.79	0.05
7	rs34696181	C	0.39	0.14	0.04	15	rs12443279	G	0.5	1.05	<0.01
7	rs10487505	G	0.3	-0.17	0.02	15	rs28607641	T	0.45	-0.82	0.03
8	rs2081493	C	0.22	0.23	<0.01	15	rs7177266	T	0.45	-0.78	0.04
8	rs10091344	G	0.5	0.18	<0.01	16	rs9932625	A	0.36	-0.91	0.02
8	rs4739570	A	0.5	0.24	<0.01	16	rs12935539	C	0.32	-1.04	<0.01
8	rs10954968	G	0.48	-0.27	<0.01	16	rs12927956	T	0.32	-1.18	<0.01

8	rs7844647	C	0.48	-0.27	<0.01
8	rs6994670	G	0.26	0.15	0.05
8	rs2941428	A	0.44	0.15	0.03
8	rs11997238	G	0.16	0.21	0.02
9	rs1323489	A	0.09	0.27	0.02
9	rs10757826	A	0.4	0.15	0.03
9	rs11142387	C	0.39	0.2	<0.01
9	rs2777777	T	0.09	0.28	0.01
9	rs7869969	A	0.5	-0.15	0.02
9	rs10985968	C	0.49	-0.17	0.02
9	rs10760279	G	0.37	0.17	0.03
9	rs2235056	C	0.47	0.15	0.04
10	rs11189058	G	0.3	0.18	0.02
10	rs7075281	A	0.32	-0.16	0.03
10	rs12765002	T	0.33	0.14	0.05
10	rs11191514	T	0.25	0.2	<0.01
10	rs11191560	C	0.25	0.19	0.02
10	rs11191580	C	0.25	0.2	0.01
10	rs718948	C	0.3	-0.24	<0.01
10	rs2172131	T	0.08	0.26	0.04
10	rs4880341	C	0.09	0.24	0.05
11	rs2237892	T	0.38	0.15	0.03
16	rs7203398	C	0.21	-1.34	<0.01
16	rs62053077	G	0.31	1.15	<0.01
17	rs11657044	T	0.41	-1.19	<0.01
17	rs9895661	C	0.48	-1.13	<0.01
17	rs887258	C	0.29	-1.05	0.02
18	rs1785418	A	0.22	-0.92	0.05

Type 2 diabetes

Chromosome	SNP	RA	MAF	OR	*P*
2	rs1009358	C	0.32	0.87	0.01
2	rs2249105	G	0.34	0.88	0.01
2	rs7572857	A	0.08	0.83	0.04
2	rs2052261	G	0.34	0.89	0.02
2	rs7565310	G	0.45	0.76	0.02
2	rs35999103	T	0.30	0.90	0.04
2	rs7578326	G	0.16	1.22	0.04
2	rs10211205	T	0.45	0.90	0.03
3	rs6780569	A	0.18	0.85	0.01
3	rs7612463	A	0.19	0.85	0.01
3	rs1845900	G	0.20	0.79	<0.01
3	rs1496653	G	0.20	0.84	<0.01
3	rs35352848	C	0.19	0.83	<0.01
3	rs17013314	G	0.16	1.19	<0.01

11	rs11041833	A	0.44	-0.15	0.03	3	rs450495	T	0.09	0.77	<0.01
11	rs223058	A	0.02	-0.5	0.02	3	rs2134223	A	0.49	1.12	0.02
11	rs223051	C	0.02	-0.48	0.03	3	rs12629668	C	0.32	0.90	0.04
11	rs12788343	C	0.42	-0.18	0.01	4	rs6446601	C	0.35	1.14	0.01
12	rs11047132	G	0.24	0.16	0.04	4	rs28690107	C	0.38	0.76	<0.01
12	rs7975187	G	0.35	0.16	0.02	4	rs7685296	T	0.44	0.90	0.03
12	rs704061	C	0.44	0.15	0.03	4	rs7660590	T	0.45	0.89	0.01
12	rs111260184	A	0.48	0.16	0.02	4	rs11099942	T	0.07	0.81	0.04
12	rs12321904	T	0.26	0.16	0.04	7	rs56805921	C	0.38	1.19	0.04
13	rs9536382	A	0.43	0.14	0.04	7	rs2908334	C	0.31	0.89	0.03
13	rs12429545	A	0.25	0.16	0.03	7	rs849135	A	0.00	2.39	0.04
13	rs4477562	T	0.25	0.16	0.04	7	rs6944685	T	0.02	1.49	0.01
13	rs9568867	A	0.25	0.16	0.04	7	rs6951280	G	0.34	1.14	0.01
13	rs1927790	C	0.41	0.13	0.05	7	rs791595	A	0.11	1.16	0.04
13	rs1536053	T	0.04	-0.41	0.02	7	rs4731420	C	0.11	1.16	0.04
14	rs2300835	A	0.09	0.25	0.03	8	rs13262861	A	0.13	0.85	0.03
14	rs75766425	C	0.16	0.23	0.01	8	rs515071	A	0.18	0.86	0.02
14	rs12894211	T	0.12	-0.25	0.02	9	rs7847880	T	0.24	0.78	0.02
14	rs8009329	A	0.19	-0.18	0.03	9	rs1873747	T	0.07	0.79	0.02
14	rs3783890	C	0.3	-0.16	0.03	9	rs11787792	G	0.05	0.65	<0.01
14	rs7147503	T	0.46	0.17	0.01	10	rs1631619	A	0.43	1.11	0.03
14	rs12431682	T	0.46	0.18	0.01	10	rs177045	A	0.37	0.88	0.01

14	rs8008758	T	0.1	-0.23	0.04	10	rs12416116	A	0.27	0.88	0.02
15	rs4432245	C	0.47	-0.15	0.03	10	rs4933236	T	0.28	0.72	<0.01
15	rs12899905	T	0.04	0.47	<0.01	10	rs34872471	C	0.03	1.52	<0.01
15	rs12593036	G	0.24	-0.19	0.02	10	rs7901695	C	0.03	1.53	<0.01
15	rs8036171	A	0.28	0.17	0.02	10	rs4506565	T	0.03	1.51	<0.01
15	rs11633626	C	0.28	0.17	0.02	10	rs7903146	T	0.03	1.45	<0.01
15	rs7181498	T	0.28	0.17	0.02	10	rs10886863	T	0.34	0.88	0.01
15	rs11638950	C	0.32	-0.18	0.01	11	rs10734252	G	0.39	1.14	0.01
16	rs1053874	G	0.45	-0.14	0.05	11	rs5215	C	0.39	1.13	0.01
16	rs2531995	T	0.35	0.15	0.04	11	rs5219	T	0.39	1.13	0.01
16	rs62034325	G	0.11	0.26	0.02	11	rs2074314	C	0.39	1.13	0.01
16	rs2008514	A	0.13	0.21	0.03	11	rs7115753	G	0.21	0.80	0.04
16	rs8049439	C	0.23	0.17	0.03	12	rs2261181	T	0.12	1.15	0.04
16	rs12325113	C	0.13	0.21	0.04	12	rs2258238	T	0.12	1.16	0.04
16	rs4788102	A	0.13	0.21	0.03	12	rs343092	T	0.34	1.10	0.05
16	rs7498665	G	0.13	0.22	0.03	12	rs7955901	T	0.30	1.13	0.02
16	rs7359397	T	0.13	0.21	0.03	12	rs7138300	T	0.30	1.13	0.02
16	rs3888190	A	0.13	0.21	0.03	12	rs1169299	T	0.25	1.13	0.03
16	rs8061590	G	0.13	0.21	0.03	13	rs79490558	A	0.06	0.66	0.03
16	rs3814883	T	0.45	0.15	0.03	13	rs7985179	A	0.15	0.82	<0.01
16	rs1421085	C	0.12	0.23	0.03	16	rs8046545	G	0.13	1.20	0.01
16	rs11642015	T	0.12	0.21	0.04	17	rs989128	A	0.04	0.45	<0.01

16	rs1558902	A	0.12	0.22	0.03
16	rs55872725	T	0.12	0.23	0.02
16	rs7193144	C	0.12	0.22	0.03
16	rs62033400	G	0.12	0.21	0.04
16	rs17817449	G	0.12	0.22	0.03
16	rs8043757	T	0.12	0.22	0.03
16	rs8050136	A	0.12	0.23	0.03
16	rs3751812	T	0.12	0.22	0.03
16	rs9936385	C	0.12	0.22	0.03
16	rs11075990	G	0.12	0.22	0.03
16	rs9926289	A	0.12	0.21	0.04
16	rs9939609	A	0.12	0.22	0.03
16	rs17817712	G	0.12	0.22	0.04
16	rs7202116	G	0.12	0.22	0.04
16	rs7185735	G	0.12	0.22	0.03
16	rs9932690	C	0.49	-0.16	0.03
17	rs2242449	T	0.41	0.16	0.03
17	rs3930349	A	0.31	-0.17	0.02
17	rs676387	A	0.39	-0.21	<0.01
17	rs668799	T	0.39	-0.2	<0.01
17	rs4144743	C	0.49	0.16	0.02
19	rs58542926	T	0.07	1.29	0.05
19	rs150268548	A	0.01	2.27	0.02
19	rs3794991	T	0.09	1.40	<0.01
19	rs57504626	T	0.09	1.33	0.02
20	rs11699802	T	0.50	0.84	0.05
22	rs71313049	C	0.10	1.40	0.02

DKD

Chromosome	SNP	RA	MAF	OR	*P*
15	rs16977473	G	0.09	2.39	0.02

17	rs208015	T	0.17	0.19	0.04
17	rs113866544	C	0.17	0.19	0.04
17	rs6504108	C	0.23	0.22	<0.01
17	rs7218014	T	0.35	-0.16	0.02
17	rs12602912	C	0.35	-0.16	0.02
17	rs4790981	A	0.35	-0.17	0.02
17	rs9891146	C	0.29	-0.2	0.01
17	rs4969049	C	0.23	0.17	0.03
17	rs9910745	T	0.32	-0.22	<0.01
17	rs12939549	G	0.32	-0.22	<0.01
17	rs12940622	A	0.32	-0.21	<0.01
17	rs11150745	G	0.3	-0.18	0.01
17	rs3935190	G	0.35	0.18	0.01
18	rs512121	T	0.28	0.16	0.04
18	rs72880429	A	0.02	-0.56	0.02
18	rs4327120	C	0.24	-0.26	<0.01
18	rs9304204	A	0.24	-0.22	<0.01
19	rs17513613	C	0.09	0.25	0.04
19	rs2238691	A	0.18	-0.27	<0.01
19	rs11672660	T	0.18	-0.27	<0.01
19	rs1800437	C	0.18	-0.28	<0.01
19	rs10423928	A	0.18	-0.29	<0.01

19	rs2287019	T	0.18	-0.29	<0.01
19	rs8105198	G	0.12	0.23	0.04
20	rs6089584	G	0.31	0.14	0.05
21	rs394608	T	0.48	-0.15	0.03

Risk allele which showed negative association with trait flipped to the other allele when building GRS.

Abbreviations: MAF, Minor Allele Frequency; RA, Risk Allele

Supplementary Table 10. DNAm changes associated with metabolic traits in Type 2 diabetes/DKD associated CpGs

	Non-diabetic Controls							
	KoGES (N=356)		HTS (N=463)		SNUH (N=197)		Meta-analysis	
	Coefficients (M-values)	*P* *	Coefficients (M-values)	*P* *	Coefficients (M-values)	*P* *	Coefficients (M-values)	*P* *
Type 2 diabetes								
BMI (Kg/m^2)								
cg25139493	-0.001	0.78	0.019	0.292	0.006	0.574	0.001	0.816
cg26823705	0.005	0.642	0.005	0.581	0.004	0.61	0.005	0.439
cg26974062	0.001	0.821	-0.015	0.415	0.012	0.416	0.001	0.916
cg19693031	-0.007	0.317	0.014	0.371	0.022	0.123	0.001	0.888
cg14530801	0.014	0.130	-0.006	0.711	0.033	0.227	0.011	0.159
cg08867893	0.008	0.216	0.032	0.094	-0.003	0.836	0.009	0.141
cg00574958	-0.018	**0.003**	-0.011	0.314	-0.019	0.067	-0.017	**3.87E-04**
cg17082373	0.0005	0.92	0.006	0.697	-0.018	0.136	-0.001	0.784
Fasting Glucose (mmol/L)								
cg25139493	-0.016	0.184	-0.225	0.026	-0.056	0.397	-0.02	0.087
cg26823705	0.040	0.081	0.047	0.372	0.087	0.091	0.067	0.067
cg26974062	-0.045	0.645	-0.026	0.795	-0.052	0.565	-0.04	0.532
cg19693031	-0.083	**2.79E-07**	-0.021	0.820	-0.075	0.387	-0.081	**1.51E-07**

cg14530801	-2.3E-05	0.984	-0.024	0.806	-0.086	0.603	-0.002	0.89
cg08867893	-0.015	0.331	-0.01	0.924	-0.098	0.27	-0.018	0.246
cg00574958	-0.051	**2.84.E-04**	0.056	0.350	0.022	0.736	-0.041	**1.46.E-03**
cg17082373	-0.012	0.242	0.017	0.841	-0.001	0.988	-0.011	0.26
HbA1c (mmol/mol)								
cg25139493	-0.003	0.143	-0.022	0.238	-0.025	0.016	-0.003	0.053
cg26823705	-0.008	0.233	-0.008	0.486	-0.008	0.344	-0.008	0.233
cg26974062	-0.029	**0.002**	-0.025	0.272	-0.04	0.00461	-0.037	**0.003**
cg19693031	-0.011	**8.70.E-06**	-0.018	0.372	-0.014	0.326	-0.011	**2.90.E-06**
cg14530801	-0.002	0.498	-0.028	0.170	-0.011	0.677	-0.003	0.351
cg08867893	-0.003	0.149	0.007	0.751	0.019	0.176	-0.003	0.241
cg00574958	-0.006	**0.003**	0.000	0.984	0.001	0.926	-0.006	**4.50E-03**
cg17082373	-0.001	0.409	-0.02	0.229	0.002	0.839	-0.001	0.366
eGFR (ml/min)								
cg25139493	-0.0003	0.759	0.004	0.281	0.002	0.388	0.0002	0.772
cg26823705	-0.0005	0.584	-0.003	0.119	0.001	0.502	-0.0004	0.711
cg26974062	-0.0087	0.168	-0.004	0.251	-0.003	0.28	-0.0033	0.122
cg19693031	-0.0005	0.726	-0.005	0.141	-0.001	0.684	-0.001	0.347
cg14530801	0.0014	0.402	-0.001	0.853	-0.001	0.868	0.0008	0.557
cg08867893	0.0004	0.735	-0.002	0.637	-0.002	0.513	-0.0001	0.939
cg00574958	0.0009	0.437	-1.E-04	0.948	0.000	0.879	0.0006	0.507

cg17082373	0.0008	0.304	0.002	0.584	0.002	0.245	0.0011	0.139
DKD								
BMI (Kg/m^2)								
cg16944159	-0.002	0.566	0.005	0.594	-0.015	0.101	-0.003	0.375
cg16079347	0.001	0.671	0.001	0.831	0.003	0.681	0.002	0.535
cg11530914	0.004	0.335	0.019	**0.006**	0.004	0.672	0.007	0.030
Fasting Glucose (mmol/L)								
cg16944159	0.006	0.578	0.004	0.937	0.013	0.825	0.005	0.54
cg16079347	-0.005	0.489	-0.035	0.314	0.048	0.247	-0.005	0.486
cg11530914	-0.006	0.448	-0.008	0.828	-0.023	0.675	-0.007	0.389
HbA1c (mmol/mol)								
cg16944159	0.002	0.291	-0.002	0.813	0.002	0.811	0.016	0.296
cg16079347	0.000	0.973	-0.007	0.307	0.006	0.384	0.0003	0.977
cg11530914	-0.001	0.539	0.004	0.962	0.003	0.756	-0.007	0.574
eGFR (ml/min)								
cg16944159	-0.0003	0.716	0.0025	0.205	0.0021	0.213	0.0004	0.572
cg16079347	-0.0005	0.372	0.0002	0.885	0.0007	0.537	-0.0002	0.643
cg11530914	4.E-05	0.945	0.0002	0.897	-0.0006	0.731	-0.0001	0.997

Supplementary Table 11. Association between age and genetic risk score

Genetic risk scores (GRS)	Coefficient (SE)	P-value
Age		
GRS of BMI	-0.0001 (0.0002)	0.680
GRS of Fasting Glucose	0.0001 (0.0004)	0.753
GRS of HbA1c	-7.32×10^{-6} (0.0006)	0.901
GRS of eGFR	-0.0002 (0.0004)	0.554
GRS of T2D	-9.41×10^{-6} (0.0003)	0.769
GRS of DKD	-0.0001 (0.0002)	0.547
Sex	**Coefficient (SE)**	**P-value**
GRS of BMI	-0.0077 (0.0006)	0.187
GRS of Fasting Glucose	0.0012 (0.0098)	0.205
GRS of HbA1c	-0.0006 (0.0013)	0.669
GRS of eGFR	9.25×10^{-6} (0.0009)	0.992
GRS of T2D	0.0007 (0.0007)	0.306
GRS of DKD	-0.0023 (0.0053)	0.669

Abbreviation: SE, Standard error

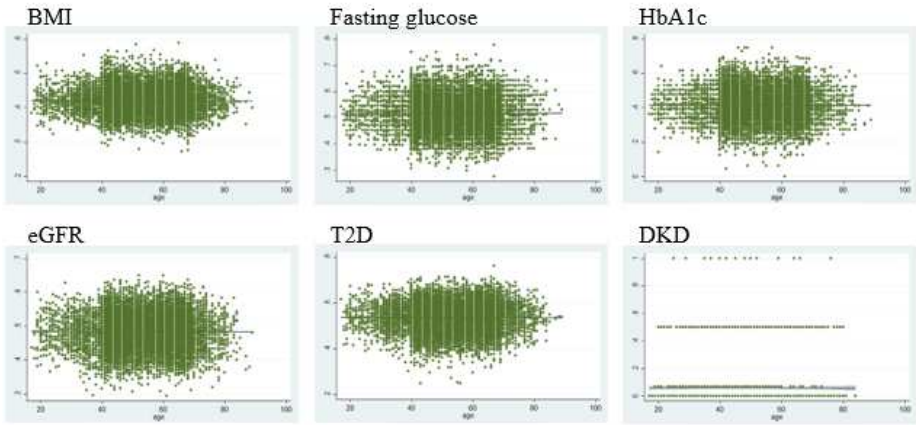
Supplementary Table 12. The entire information of Mendelian randomization of metabolic traits on type 2 diabetes/DKD associated CpGs

	non-diabetic controls		
	Coefficients (β-values)	**SE**	***P***
Type 2 diabetes			
BMI (Kg/m^2)			
cg25139493	0.0006	0.0012	0.651
cg26823705	0.0018	0.0085	0.834
cg26974062	-0.0032	0.0046	0.490
cg19693031	-0.0100	0.0056	0.075
cg14530801	-0.0005	0.0008	0.497
cg08867893	0.0004	0.0008	0.564
cg00574958	-0.00001	0.0018	0.994
cg17082373	-0.0010	0.0006	0.122
Fasting Glucose (mmol/L)			
cg25139493	-0.0013	0.0046	0.778
cg26823705	0.0254	0.0307	0.409
cg26974062	-0.0129	0.0176	0.462
cg19693031	-0.0319	0.0170	0.061
cg14530801	0.0004	0.0023	0.872
cg08867893	0.0013	0.0030	0.664
cg00574958	-0.0204	0.0077	**0.008**
cg17082373	0.0031	0.0026	0.230
HbA1c (mmol/mol)			
cg25139493	0.0001	0.0004	0.857
cg26823705	-0.00003	0.0032	0.992
cg26974062	-0.0001	0.0018	0.963
cg19693031	0.0012	0.0017	0.467
cg14530801	-0.00004	0.0003	0.889
cg08867893	0.0003	0.0003	0.275
cg00574958	-0.0007	0.0008	0.386
cg17082373	-0.0001	0.0002	0.519
eGFR (10ml/min)			
cg25139493	-0.0180	0.0092	0.051
cg26823705	-0.0819	0.0606	0.177
cg26974062	-0.0568	0.0332	0.087
cg19693031	-0.0702	0.0350	**0.045**
cg14530801	-0.0049	0.0037	0.183
cg08867893	-0.0085	0.0063	0.176
cg00574958	-0.0272	0.0146	0.063
cg17082373	-0.0093	0.0068	0.176
DKD			
BMI (kg/m^2)			

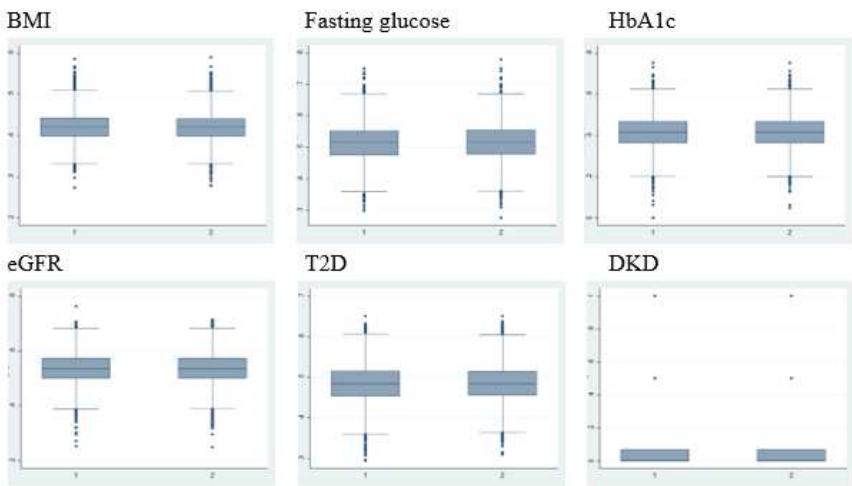
cg16944159	0.0062	0.0129	0.632
cg16079347	-0.0074	0.0214	0.730
cg11530914	-0.0045	0.0108	0.675
Fasting glucose (mmol/L)			
cg16944159	0.0073	0.0086	0.398
cg16079347	0.0186	0.0166	0.263
cg11530914	0.0132	0.0095	0.162
HbA1c (mmol/mol)			
cg16944159	-0.0010	0.0010	0.347
cg16079347	0.0012	0.0021	0.573
cg11530914	0.0017	0.0013	0.170
eGFR (ml/min)			
cg16944159	-0.0108	0.0029	**0.001**
cg16079347	-0.0188	0.0060	**0.002**
cg11530914	-0.0093	0.0034	**0.006**

Bold style font indicates P under nominal significance (P <0.05)
Abbreviation: BMI, Body mass index; eGFR, estimated Glomerular Filtration Rate; HbA1c, Glycated hemoglobin; SBP, Systolic blood pressure; SE, Standard error

Supplementary Figure 2. Scatter plot between age and GRS of phenotypes



Supplementary Figure 3. Box plot between age and GRS of phenotypes



V. Overall summary and Conclusion

Type 2 diabetes is a complex disease, which is characterized by hyper-glycemia due to impaired beta-cell function and insulin resistance. The prevalence of type 2 diabetes is estimated to be 9.3% of the adult population worldwide in 2019 and is projected to rise to 10.9% by 2030 [8]. In Korea, 12.4% of adults have type 2 diabetes in 2018 [9]. Despite the high prevalence of type 2 diabetes, the mechanisms underlying the development of diabetes are not fully established due to its complexity. Therefore, understanding pathophysiology of type 2 diabetes became the important global health agenda to develop improved prevention method and treatments for diabetes and associated complications.

DNA methylation (DNAm), the one of epigenetic modification, is often considered as the result of environmental stimuli [1]. Epigenetic modification is heritable changes to the gene expression that do not related to changes in DNA sequences [4]. Epigenetic changes by biological, lifestyle and environmental factors are stably remembered by cells across the lifespan and generations.

To date, the numerous family-based study and genome-wide association studies (GWAS) of type 2 diabetes demonstrated the genetic influence on type 2 diabetes. Currently, more than 560 genetic loci have been identified to be associated with type 2 diabetes through studies involving more than 1,400,000 participants [15]. Still, identified variants do not fully explain the expected heritability [16]. Also increasing evidence showed that type 2 diabetes arises from an interaction between genetic and non-genetic factors.

Epigenetics is complex intermediate which is both influenced by genetic and environmental factors. Environmental stimuli changes DNAm level which finally lead to gene expression change. Furthermore, some DNAm sites are affect by genetics, which is called methylation quantitative loci (mQTL) [37, 38, 40]. Therefore, the role of epigenetics emerged to find ‘missing heritability’. Epigenetic changes are considered to be a link between genetics and non-genetic

factors.

There are several studies about the DNAm difference between ethnic groups [24-30]. As DNAm patterns reflect social and environmental factors, the genetic and cultural differences between diverse ethnic groups might lead to different DNAm profiles. Therefore, discovering East-Asian specific DNA methylation loci would contribute to understanding the risk factor of type 2 diabetes in the East-Asian population.

This study focused on DNAm changes of type 2 diabetes in the East Asian population. Understanding DNA methylation profiles characterized in type 2 diabetes patients and the cause of those methylation changes would help to provide better understanding type 2 diabetes pathophysiology and develop prevention and early diagnosis tool.

There were three of research question: 1) Is there East Asian specific methylated sites in type 2 diabetes patients compared to the non-diabetic population in East Asians? 2) Is there DKD patients specific methylated sites compared to type 2 diabetes patients through metabolic memory phenomena? 3) Is metabolic traits characteristic of type 2 diabetes patients affect DNA methylation changes associated with type 2 diabetes /DKD?

To answer research questions, study objectives were set: 1) Identify DMSs associated with type 2 diabetes through MWAS and test prediction of type 2 diabetes incidence using the aggregated score of DMSs. 2) Identify DMSs associated with DKD through MWAS and investigate overlapped markers between DMSs of type 2 diabetes and DKD. 3) Explore the association and the causation of these DMSs with type 2 diabetes-related metabolic traits in non-diabetic controls.

Firstly, an overall review of epigenetics and DNA methylation was conducted in **Chapter I**. Differentially methylated markers of type 2 diabetes through earlier MWAS in European, African, and Indian Asian were reviewed. Imputation of DNA methylation markers from Illumina 450K platform to Illumina EPIC

platform was conducted using methyLImp based on correlation matrix calculation method.

In **Chapter II**, five new DMSs associated with type 2 diabetes were identified with methylome-wide statistical significance (9 x 10-8) in the East Asian population. We performed an initial MWAS in 232 participants in a case-control study design from SNUH cohort with type 2 diabetes and 197 non-diabetic controls. In addition, seven type 2 diabetes DMSs reported predominantly in a European population were replicated in this study. DNA methylation profile of type 2 diabetes' DMSs helped to predict future incidence risk of type 2 diabetes. In mQTL and Mendelian randomization (MR) analysis, discovered DMSs are not causally related to type 2 diabetes.

In **Chapter III**, three novel DMSs associated with DKD with methylome-wide statistical significance were identified in the East-Asian population. Type 2 diabetes group from Chapter II was subdivided to 87 DKD cases and 80 non-DKD controls. COMMD1, TMOD1, and FHOD1 where DKD associated DMSs are located are linked to kidney function. There was no overlap between type 2 diabetes-associated DMSs and DKD associated DMSs Also, two DMSs of DKD are largely influenced by genetics and have a causal effect on DKD in MR analysis.

For the last **Chapter IV**, to explore the evidence that DNA methylation affects type 2 diabetes via metabolic memory phenomenon, the risk of metabolic syndrome is investigated through MWAS and MR analysis. To investigate the association of the identified DMSs with diabetes-related metabolic traits, two additional population-based cohorts were used: 1) the Korean Genome and Epidemiology Study (KoGES) consisting of 356 non-diabetic controls [52] and 2) the Healthy Twin Study, Korea (HTS), consisting of 463 non-diabetic controls [53]. Fasting glucose showed a causal relationship with hypo-methylation of cg00574958 which lead to insulin resistance via CPT1A expression. Three DMSs of type 2 diabetes affect the HbA1c level.

Overall summary of discovered DMSs in this study is presented in **Table 1**.

The clinical characteristics of diabetes in East Asians showing early pancreatic β-cell dysfunction in insulin resistance are not commonly observed in Europeans [155, 156]. Although it is not clear whether these differences can be explained by distinct DNA methylation changes, β-cell function is well known to be affected by epigenetic regulations. Thus, it would be interesting to find the difference of DNAm profiling between Europeans and East Asian population.

Discovered DMSs in ABCG1, CPT1A, PHOSPHO1, SOCS3, SREBF1, and TXNIP have been well-replicated in European population (6-10,12,13,49). This finding suggests that there are common DNAm marks that are consistently associated with T2D across different ethnic groups. However, we found 5 novel DMSs out of eight in Korean suggesting there are different characteristic in DNAm profiles in Korean distinguished from other populations.

Findings from this study support the hypothesis that glycemic variation influences DNAm in type 2 diabetes. We found that hypomethylation of cg00574958 is a marker of hyperglycemia that could proceed with the development of type 2 diabetes. On the other hand, it has been reported that methylation at this DMS was causally associated with type 2 diabetes in an MR study using methylation quantitative loci. The overlap between the DMSs of type 2 diabetes and metabolic traits and the findings from the MR analysis suggest that the metabolic effects may be engraved as epigenetic marks, which further modulate the risk of type 2 diabetes, at least in part.

Three DMSs of type 2 diabetes (cg26823705, cg08867893, and cg17082373) affect the increase of HbA1c level. Unfortunately, the genetic mechanisms of three genes (NBPF20, ZNF365, and TRIM37) which induce HbA1c change are unknown. In our knowledge, the expression of these genes is associated with cancer or tumor proliferation [150-152]. There is a lot of evidence that a higher HbA1c level is associated with higher incidence and mortality risk for all cancers [153, 154]. The Discovery of relevance between DNA methylation of three genes

and HbA1c could explain the association of HbA1c level and cancer risk.

Epidemiology is defined as the study of the distribution and determinants of health-related states and events in specified populations. Recently, traditional epidemiology has expanding the field using new technologies. Epigenetic epidemiology is defined as "the integration of epigenetic analysis into population-based epidemiological research with the goal of identifying both the causes and phenotypic consequences of the epigenomic variation" [157]. This is one of the newly emerged field of epidemiology to give beneficial knowledge into disease and public health. Epigenetic variation could be the cause of the disease, mediator of the effect of exposure and biomarker of exposure and disease.

In this study, we discovered that the DNA methylation profile of the East Asian population is somewhat distinguished from that of Europeans, which is suggesting that pathophysiology between two populations are different. This discovery could contribute to develop more suitable therapy on East-Asian. The future incidence of type 2 diabetes in high-risk individuals could be predicted by using methylation profiles.

It is worth noting that DMSs of type 2 diabetes was not the cause of the disease. We discovered DNAm profiles as a biomarker of type 2 diabetes/DKD, which could be utilize to develop type 2 diabetes risks assessment and prognosis prediction in especially East-Asian specific manner. Some DMSs are identified to be a result of glycemic variation, which would help to detection high risk individuals of type 2 diabetes before its onset. Others are identified to be a cause of HbA1c variation. Uncontrolled HbA1c has higher risk of complication incidence in type 2 diabetes patients. By using DNAm profiles, it would be possible to give an earlier intensive intervention to type 2 diabetes patients who have high risk of diabetic complication. Individuals with high risk of type 2 diabetes/DKD DNAm profiles can be identified and be targeted for risk-reducing intervention strategies.

In conclusion, this study would help a better understanding of the pathophysiology

of type 2 diabetes/DKD and enhance the prediction of type 2 diabetes. Also, this study would contribute to developing a better intervention in type 2 diabetes for not to progress to DKD.

Table 28. Overall summary of discovered DMSs

CpG	Reported information				Discovered information				
	Reported MWAS traits	Nearby gene	Reported Population	Reported mQTL (Reference)	Discovered mQTL[a]	MR 1: Effect of CpG methylation on Type 2 diabetes/DKD	MR 2: Effect of Type 2 diabetes/DKD on CpG methylation	MR 3: Effect of metabolic traits on CpG methylation [b]	MR 4: Effect of CpG methylation on metabolic traits
						OR (P)	Coefficients (P)	Trait: Coefficients (P)	Trait: Coefficients (P)
Type 2 diabetes									
cg25139493	-	*BMP8A*	-	-	-	1.08 (0.103)	0.11 (0.84)	-	-
cg26823705	HbA1c	*NBPF20*	-	-	-	1.00 (0.982)	4.77 (0.14)	-	-
cg26974062	T2D	*TXNIP*	SAS	-	-	0.99 (0.582)	0.44 (0.81)	-	-
cg19693031	Glucose, HbA1c, lipid metabolites, T2D	*TXNIP*	AMR, SAS, EUR	rs17387733, rs11451644, rs3768036 and 27 others (mQTL db)	-	0.99 (0.682)	-0.97 (0.64)	eGFR: -7.02 (0.045) [c]	-
cg14530801	-	*STX18*	-	-	-	0.74 (0.121)	-0.19 (0.48)	-	
cg08867893	-	*ZNF365*	-	-	-	1.00 (0.983)	-0.55 (0.07)	-	HbA1c:0.11 (<0.001)
cg00574958	T2D	*CPT1A*	AMR, EUR	rs964184, rs3741298 (BIOS QTL browser)	-	0.90 (0.052)	0.83 (0.35)	Fasting glucose: -2.04 (0.008)	-

cg17082373	-	*TRIM37*	-		-	0.76 (0.173)	-0.39 (0.14)	-	-HbA1c: 0.31 (<0.001)
DKD									
cg16944159	-	*COMMD1*	-	-	rs79841299	1.15 (<0.001)	0.32 (0.21)	eGFR: -1.08 (0.001) [c]	-
cg16079347	*Sex*	*TMOD1*	-	-	chr9:111425306:A:G	1.20 (0.263)	0.20 (0.60)	eGFR: -1.88 (0.002) [c]	-
cg11530914	-	*FHOD1*	-	-	-	1.19 (0.001)	0.14 (0.50)	eGFR: -0.93 (0.006) [c]	-

Population: AMR (Ad mixed American), SAS (South Asian), EUR (European)

Abbreviation: MR, Mendelian randomization; mQTL, methylation quantitative trait loci

a. cis-QTL SNPs (≤500 kb) with genome-wide significance

b Only significant association were shown.

c. Result nullified in adjusted MR analysis

- mark denotes that there are no significance results

VI. References

1. Griffith, J.S. and H.R. Mahler, *DNA ticketing theory of memory.* Nature, 1969. **223**(5206): p. 580-2.
2. Paul, D.S., et al., *Increased DNA methylation variability in type 1 diabetes across three immune effector cell types.* Nat Commun, 2016. **7**: p. 13555.
3. Wong, C.C., et al., *A longitudinal study of epigenetic variation in twins.* Epigenetics, 2010. **5**(6): p. 516-26.
4. Painter, R.C., T.J. Roseboom, and O.P. Bleker, *Prenatal exposure to the Dutch famine and disease in later life: an overview.* Reprod Toxicol, 2005. **20**(3): p. 345-52.
5. Stirzaker, C., et al., *Mining cancer methylomes: prospects and challenges.* Trends in Genetics, 2014. **30**(2): p. 75-84.
6. Gunderson, K.L., et al., *Decoding randomly ordered DNA arrays.* Genome Res, 2004. **14**(5): p. 870-7.
7. Miao, F., et al., *Evaluating the role of epigenetic histone modifications in the metabolic memory of type 1 diabetes.* Diabetes, 2014. **63**(5): p. 1748-1762.
8. Atlanta, G.C.f.D.C.a.P., U.S. Dept of Health and Human Services,, *Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020.* 2020.
9. KOSIS, *당뇨병 유병률 추이 (URL: https://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_11702_N102).* 2020-02-11.
10. KOSIS, *2018 사망원인 통계 (URL: http://kostat.go.kr/portal/korea/kor_nw/1/6/2/index.board?bmode=read&bSeq=&aSeq=377606&pageNo=1&rowNum=10&navCount=10&currPg=&searchInfo=&sTarget=title&sTxt=).* 2019-09-24.
11. Centers for Disease Control and Prevention, N.C.f.h.s., *Leading cause of death.* 2018.
12. Kirkman, M.S., et al., *Diabetes in Older Adults.* Diabetes Care, 2012. **35**(12): p. 2650.
13. Hu, F.B., et al., *Diet, lifestyle, and the risk of type 2 diabetes mellitus in women.* N Engl J Med, 2001. **345**(11): p. 790-7.
14. Almgren, P., et al., *Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study.* Diabetologia, 2011. **54**(11): p. 2811-2819.
15. Vujkovic, M., et al., *Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis.* Nat Genet, 2020. **52**(7): p. 680-691.
16. Xue, A., et al., *Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes.* Nature communications, 2018. **9**(1): p. 1-14.
17. Hall, E., et al., *DNA methylation of the glucagon-like peptide 1 receptor (GLP1R) in human pancreatic islets.* BMC medical genetics, 2013. **14**(1): p. 76.
18. Ling, C., et al., *Epigenetic regulation of PPARGC1A in human type 2 diabetic islets and effect on insulin secretion.* Diabetologia, 2008. **51**(4): p. 615-622.
19. Volkov, P., et al., *Whole-Genome Bisulfite Sequencing of Human Pancreatic Islets Reveals Novel Differentially Methylated Regions in Type 2 Diabetes Pathogenesis.* Diabetes, 2017. **66**(4): p. 1074-1085.
20. Kulkarni, S.S., et al., *Mitochondrial regulators of fatty acid metabolism reflect metabolic dysfunction in type 2 diabetes mellitus.* Metabolism, 2012. **61**(2): p. 175-85.
21. Nilsson, E., et al., *Altered DNA methylation and differential expression of genes*

influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. Diabetes, 2014. **63**(9): p. 2962-2976.
22. Abderrahmani, A., et al., *Increased hepatic PDGF-AA signaling mediates liver insulin resistance in obesity-associated type 2 diabetes.* Diabetes, 2018. **67**(7): p. 1310-1321.
23. Kirchner, H., et al., *Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients.* Mol Metab, 2016. **5**(3): p. 171-183.
24. Cardona, A., et al., *Epigenome-wide association study of incident Type 2 diabetes in a British population: EPIC-Norfolk study.* Diabetes, 2019. **68**(12): p. 2315-2326.
25. Chambers, J.C., et al., *Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study.* The lancet Diabetes endocrinology, 2015. **3**(7): p. 526-534.
26. Kulkarni, H., et al., *Novel epigenetic determinants of type 2 diabetes in Mexican-American families.* Human molecular genetics, 2015. **24**(18): p. 5330-5344.
27. Soriano-Tarraga, C., et al., *Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia.* Hum Mol Genet, 2016. **25**(3): p. 609-19.
28. Florath, I., et al., *Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults.* Diabetologia, 2016. **59**(1): p. 130-138.
29. Al Muftah, W.A., et al., *Epigenetic associations of type 2 diabetes and BMI in an Arab population.* Clinical epigenetics, 2016. **8**(1): p. 13.
30. Meeks, K.A., et al., *Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study.* International journal of epidemiology, 2019. **48**(1): p. 58-70.
31. Teschendorff, A.E., et al., *An epigenetic signature in peripheral blood predicts active ovarian cancer.* PloS one, 2009. **4**(12).
32. Woodson, K., et al., *Hypomethylation of p53 in peripheral blood DNA is associated with the development of lung cancer.* Cancer Epidemiology and Prevention Biomarkers, 2001. **10**(1): p. 69-74.
33. Houseman, E.A., et al., *DNA methylation in whole blood: uses and challenges.* Current environmental health reports, 2015. **2**(2): p. 145-154.
34. Alicic, R.Z., M.T. Rooney, and K.R. Tuttle, *Diabetic kidney disease: challenges, progress, and possibilities.* Clinical Journal of the American Society of Nephrology, 2017. **12**(12): p. 2032-2045.
35. Dos Santos Nunes, M.K., et al., *Analysis of the DNA methylation profiles of miR-9-3, miR-34a, and miR-137 promoters in patients with diabetic retinopathy and nephropathy.* J Diabetes Complications, 2018. **32**(6): p. 593-601.
36. Aldemir, O., F. Turgut, and C. Gokce, *The association between methylation levels of targeted genes and albuminuria in patients with early diabetic kidney disease.* Ren Fail, 2017. **39**(1): p. 597-601.
37. Hannon, E., et al., *Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci.* Nat Neurosci, 2016. **19**(1): p. 48-54.
38. Wagner, J.R., et al., *The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts.* Genome Biol, 2014. **15**(2): p. R37.
39. Gibbs, J.R., et al., *Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain.* PLoS Genet, 2010. **6**(5): p. e1000952.
40. Olsson, A.H., et al., *Genome-wide associations between genetic and epigenetic variation influence mRNA expression and insulin secretion in human pancreatic islets.* PLoS Genet, 2014. **10**(11): p. e1004735.
41. Grundberg, E., et al., *Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements.*

Am J Hum Genet, 2013. **93**(5): p. 876-90.
42. Bell, J.T., et al., *Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.* PLoS Genet, 2012. **8**(4): p. e1002629.
43. Elliott, H.R., et al., *Role of DNA Methylation in Type 2 Diabetes Etiology: Using Genotype as a Causal Anchor.* Diabetes, 2017. **66**(6): p. 1713.
44. Angermueller, C., et al., *DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning.* Genome Biology, 2017. **18**(1): p. 67.
45. Di Lena, P., et al., *Missing value estimation methods for DNA methylation data.* Bioinformatics, 2019. **35**(19): p. 3786-3793.
46. Lövkvist, C., et al., *DNA methylation in human epigenomes depends on local topology of CpG sites.* Nucleic Acids Res, 2016. **44**(11): p. 5123-32.
47. Zhang, L., et al., *DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells.* Biophys J, 2017. **113**(7): p. 1395-1404.
48. Tian, Q., et al., *MRCNN: a deep learning model for regression of genome-wide DNA methylation.* BMC Genomics, 2019. **20**(2): p. 192.
49. Yu, F., et al., *A novel computational strategy for DNA methylation imputation using mixture regression model (MRM).* BMC Bioinformatics, 2020. **21**(1): p. 552.
50. Li, G., et al., *CUE: CpG impUtation ensemble for DNA methylation levels across the human methylation450 (HM450) and EPIC (HM850) BeadChip platforms.* Epigenetics, 2020: p. 1-11.
51. Kwak, S.H., et al., *Nonsynonymous Variants in PAX4 and GLP1R Are Associated With Type 2 Diabetes in an East Asian Population.* Diabetes, 2018. **67**(9): p. 1892-1902.
52. Kim, Y., B.-G. Han, and G.E.S.g. Ko, *Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium.* International journal of epidemiology, 2017. **46**(2): p. e20-e20.
53. Sung, J., et al., *Healthy Twin: a twin-family study of Korea—protocols and current status.* Twin Research and Human Genetics, 2006. **9**(6): p. 844-848.
54. American Diabetes Association, *11. Microvascular Complications and Foot Care: Standards of Medical Care in Diabetes-2020.* Diabetes Care, 2020. **43**: p. S135.
55. Assenov, Y., et al., *Comprehensive analysis of DNA methylation data with RnBeads.* Nature methods, 2014. **11**(11): p. 1138.
56. Teschendorff, A.E., et al., *A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.* Bioinformatics, 2013. **29**(2): p. 189-196.
57. Rahmani, E., et al., *GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data.* Bioinformatics, 2017. **33**(12): p. 1870-1872.
58. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments.* Bioinformatics, 2012. **28**(6): p. 882-883.
59. Smyth Gordon, K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**(1): p. 1-25.
60. Mansell, G., et al., *Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array.* BMC genomics, 2019. **20**(1): p. 366.
61. Li, M., et al., *EWAS Atlas: a curated knowledgebase of epigenome-wide association studies.* Nucleic Acids Res, 2019. **47**(D1): p. D983-d988.
62. Shen, J., et al., *Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips.* Epigenetics, 2013. **8**(1): p. 34-43.
63. Elliott, H.R., et al., *Differences in smoking associated DNA methylation patterns in South Asians and Europeans.* Clinical Epigenetics, 2014. **6**(1): p. 4.
64. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and*

compare ROC curves. BMC bioinformatics, 2011. **12**(1): p. 77.
65. Carpenter, J. and J. Bithell, *Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians.* Stat Med, 2000. **19**(9): p. 1141-64.
66. Emdin, C.A., A.V. Khera, and S. Kathiresan, *Mendelian Randomization.* JAMA, 2017. **318**(19): p. 1925-1926.
67. Bonder, M.J., et al., *Disease variants alter transcription factor levels and methylation of their binding sites.* Nature genetics, 2017. **49**(1): p. 131.
68. Zhernakova, D.V., et al., *Identification of context-dependent expression quantitative trait loci in whole blood.* Nat Genet, 2017. **49**(1): p. 139-145.
69. Gaunt, T.R., et al., *Systematic identification of genetic influences on methylation across the human life course.* Genome biology, 2016. **17**(1): p. 61.
70. *Package : PLINK [v1.90b3.31], Authors : Shaun Purcell, Christopher Chang, URL : www.cog-genomics.org/plink/1.9/.*
71. Zhou, Y., S.R. Browning, and B.L. Browning, *IBDkin: fast estimation of kinship coefficients from identity by descent segments.* Bioinformatics, 2020. **36**(16): p. 4519-4520.
72. Ziyatdinov, A., et al., *solarius: an R interface to SOLAR for variance component analysis in pedigrees.* Bioinformatics, 2016. **32**(12): p. 1901-2.
73. Walaszczyk, E., et al., *DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA1c levels: a systematic review and replication in a case-control sample of the Lifelines study.* Diabetologia, 2018. **61**(2): p. 354-368.
74. Soranzo, N., et al., *Common variants at 10 genomic loci influence hemoglobin A_1 (C) levels via glycemic and nonglycemic pathways.* Diabetes, 2010. **59**(12): p. 3229-39.
75. Wojcik, G.L., et al., *Genetic analyses of diverse populations improves discovery for complex traits.* Nature, 2019. **570**(7762): p. 514-518.
76. Kanai, M., et al., *Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases.* Nat Genet, 2018. **50**(3): p. 390-400.
77. Wheeler, E., et al., *Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis.* PLoS Med, 2017. **14**(9): p. e1002383.
78. Kulminski, A.M., et al., *Strong impact of natural-selection-free heterogeneity in genetics of age-related phenotypes.* Aging (Albany NY), 2018. **10**(3): p. 492-514.
79. Hwang, J.Y., et al., *Genome-wide association meta-analysis identifies novel variants associated with fasting plasma glucose in East Asians.* Diabetes, 2015. **64**(1): p. 291-8.
80. Wessel, J., et al., *Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility.* Nat Commun, 2015. **6**: p. 5897.
81. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet, 2010. **42**(2): p. 105-16.
82. Kraja, A.T., et al., *A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium.* Diabetes, 2011. **60**(4): p. 1329-39.
83. Gallois, A., et al., *A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context.* Nat Commun, 2019. **10**(1): p. 4788.
84. Krumsiek, J., et al., *Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information.* PLoS Genet, 2012. **8**(10): p. e1003005.
85. Johansen, C.T., et al., *Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia.* Nat Genet, 2010. **42**(8): p. 684-7.
86. Kilpeläinen, T.O., et al., *Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity.* Nat Commun, 2019. **10**(1): p. 376.
87. Kulminski, A.M., et al., *Quantitative and Qualitative Role of Antagonistic*

Heterogeneity in Genetics of Blood Lipids. J Gerontol A Biol Sci Med Sci, 2020. **75**(10): p. 1811-1819.
88. Krause, C., et al., *Critical evaluation of the DNA-methylation markers ABCG1 and SREBF1 for Type 2 diabetes stratification.* Epigenomics, 2019. **11**(8): p. 885-897.
89. Mahajan, A., et al., *Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps.* Nat Genet, 2018. **50**(11): p. 1505-1513.
90. Whittle, A.J., et al., *BMP8B increases brown adipose tissue thermogenesis through both central and peripheral actions.* Cell, 2012. **149**(4): p. 871-885.
91. Zhong, S., et al., *Spatial and temporal expression of bmp8a and its role in regulation of lipid metabolism in zebrafish Danio rerio.* Gene Reports, 2018. **10**: p. 33-41.
92. Nilsson, E., et al., *Transcriptional and epigenetic changes influencing skeletal muscle metabolism in women with polycystic ovary syndrome.* The Journal of clinical endocrinology and metabolism, 2018. **103**(12): p. 4465-4477.
93. Pulit, S.L., et al., *Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry.* Hum Mol Genet, 2019. **28**(1): p. 166-174.
94. Karlberg, N., et al., *Insulin resistance syndrome in subjects with mutated RING finger protein TRIM37.* Diabetes, 2005. **54**(12): p. 3577-81.
95. Aslibekyan, S., et al., *Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference.* Obesity (Silver Spring), 2015. **23**(7): p. 1493-501.
96. Dick, K.J., et al., *DNA methylation and body-mass index: a genome-wide analysis.* The Lancet, 2014. **383**(9933): p. 1990-1998.
97. Reed, Z.E., et al., *The association of DNA methylation with body mass index: distinguishing between predictors and biomarkers.* Clinical Epigenetics, 2020. **12**(1): p. 50.
98. Kader, F. and M. Ghai, *DNA methylation-based variation between human populations.* Molecular Genetics and Genomics, 2017. **292**(1): p. 5-35.
99. Yuan, V., et al., *Accurate ethnicity prediction from placental DNA methylation data.* Epigenetics & Chromatin, 2019. **12**(1): p. 51.
100. Galanter, J.M., et al., *Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures.* Elife, 2017. **6**.
101. Lokk, K., et al., *DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns.* Genome Biol, 2014. **15**(4): p. r54.
102. Widschwendter, M., et al., *Epigenotyping in peripheral blood cell DNA and breast cancer risk: a proof of principle study.* PLoS One, 2008. **3**(7): p. e2656.
103. Christensen, B.C., et al., *Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context.* PLoS Genet, 2009. **5**(8): p. e1000602.
104. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC Bioinformatics, 2012. **13**: p. 86.
105. Reinius, L.E., et al., *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.* PLoS One, 2012. **7**(7): p. e41361.
106. Fraga, M.F. and M. Esteller, *DNA methylation: a profile of methods and applications.* Biotechniques, 2002. **33**(3): p. 632, 634, 636-49.
107. Modification of Diet in Renal Disease Study Group, *The Modification of Diet in Renal Disease Study: design, methods, and results from the feasibility study.* Am J Kidney Dis, 1992. **20**(1): p. 18-33.
108. Devlin, B., K. Roeder, and L. Wasserman, *Genomic control, a new approach to genetic-based association studies.* Theoretical population biology, 2001. **60**(3): p. 155-166.
109. Yavorska, O.O. and S. Burgess, *MendelianRandomization: an R package for*

performing Mendelian randomization analyses using summarized data. International journal of epidemiology, 2017. **46**(6): p. 1734-1739.
110. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* International Journal of Epidemiology, 2015. **44**(2): p. 512-525.
111. Bowden, J., et al., *Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.* Genet Epidemiol, 2016. **40**(4): p. 304-14.
112. Hartwig, F.P., G. Davey Smith, and J. Bowden, *Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.* Int J Epidemiol, 2017. **46**(6): p. 1985-1998.
113. Verbanck, M., et al., *Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases.* Nat Genet, 2018. **50**(5): p. 693-698.
114. Rees, J.M.B., et al., *Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates.* PLoS One, 2019. **14**(9): p. e0222362.
115. Slob, E.A.W. and S. Burgess, *A comparison of robust Mendelian randomization methods using summary data.* Genet Epidemiol, 2020. **44**(4): p. 313-329.
116. Kanwar, Y.S., et al., *A glimpse of various pathogenetic mechanisms of diabetic nephropathy.* Annu Rev Pathol, 2011. **6**: p. 395-423.
117. Kim, M.K., *Treatment of diabetic kidney disease: current and future targets.* Korean J Intern Med, 2017. **32**(4): p. 622-630.
118. *Sustained effect of intensive treatment of type 1 diabetes mellitus on development and progression of diabetic nephropathy: the Epidemiology of Diabetes Interventions and Complications (EDIC) study.* Jama, 2003. **290**(16): p. 2159-67.
119. Nathan, D.M., et al., *Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes.* N Engl J Med, 2005. **353**(25): p. 2643-53.
120. de Boer, I.H., et al., *Intensive diabetes therapy and glomerular filtration rate in type 1 diabetes.* N Engl J Med, 2011. **365**(25): p. 2366-76.
121. The Writing Team for the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group, *Effect of intensive therapy on the microvascular complications of type 1 diabetes mellitus.* JAMA: the journal of the American Medical Association, 2002. **287**(19): p. 2563.
122. Ke, Y., et al., *COMMD1 downregulates the epithelial sodium channel through Nedd4-2.* Am J Physiol Renal Physiol, 2010. **298**(6): p. F1445-56.
123. Wang, W., et al., *Quantitative proteomics reveals TMOD1-related proteins associated with water balance regulation.* PLoS One, 2019. **14**(7): p. e0219932.
124. Bechtel, W., et al., *Methylation determines fibroblast activation and fibrogenesis in the kidney.* Nat Med, 2010. **16**(5): p. 544-50.
125. Hayashi, K., et al., *KLF4-dependent epigenetic remodeling modulates podocyte phenotypes and attenuates proteinuria.* J Clin Invest, 2014. **124**(6): p. 2523-37.
126. Marumo, T., et al., *Diabetes Induces Aberrant DNA Methylation in the Proximal Tubules of the Kidney.* J Am Soc Nephrol, 2015. **26**(10): p. 2388-97.
127. Sapienza, C., et al., *DNA methylation profiling identifies epigenetic differences between diabetes patients with ESRD and diabetes patients without nephropathy.* Epigenetics, 2011. **6**(1): p. 20-8.
128. Swan, E.J., A.P. Maxwell, and A.J. McKnight, *Distinct methylation patterns in genes that affect mitochondrial function are associated with kidney disease in blood-derived DNA from individuals with Type 1 diabetes.* Diabet Med, 2015. **32**(8): p. 1110-5.
129. Gautier, J.F., et al., *Kidney Dysfunction in Adult Offspring Exposed In Utero to Type 1 Diabetes Is Associated with Alterations in Genome-Wide DNA Methylation.* PLoS One, 2015. **10**(8): p. e0134654.
130. Chen, Z., et al., *Epigenomic profiling reveals an association between persistence of*

DNA methylation and metabolic memory in the DCCT/EDIC type 1 diabetes cohort. Proc Natl Acad Sci U S A, 2016. **113**(21): p. E3002-11.
131. Ko, Y.A., et al., *Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development.* Genome Biol, 2013. **14**(10): p. R108.
132. Chu, A.Y., et al., *Epigenome-wide association studies identify DNA methylation associated with kidney function.* Nat Commun, 2017. **8**(1): p. 1286.
133. Brennan, E., et al., *The genetics of diabetic nephropathy.* Genes (Basel), 2013. **4**(4): p. 596-619.
134. Palmer, N.D. and B.I. Freedman, *Insights into the genetic architecture of diabetic nephropathy.* Curr Diab Rep, 2012. **12**(4): p. 423-31.
135. Badal, S.S. and F.R. Danesh, *New insights into molecular mechanisms of diabetic kidney disease.* Am J Kidney Dis, 2014. **63**(2 Suppl 2): p. S63-83.
136. Cho, Y.S., et al., *An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes.* Nat Commun, 2016. **7**: p. 13637.
137. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
138. Viechtbauer, W., *Conducting meta-analyses in R with the metafor package.* Journal of statistical software, 2010. **36**(3): p. 1-48.
139. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic acids research, 2019. **47**(D1): p. D1005-D1012.
140. Geach, T., *Methylation a consequence not a cause.* Nature Reviews Endocrinology, 2017. **13**(3): p. 127-127.
141. Kim, M., *DNA methylation: a cause and consequence of type 2 diabetes.* Genomics & informatics, 2019. **17**(4): p. e38-e38.
142. Zhang, N., *Epigenetic modulation of DNA methylation by nutrition and its mechanisms in animals.* Anim Nutr, 2015. **1**(3): p. 144-151.
143. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.
144. Alam, F., et al., *DNA Methylation: An Epigenetic Insight into Type 2 Diabetes Mellitus.* Curr Pharm Des, 2016. **22**(28): p. 4398-419.
145. Liu, J., et al., *An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis.* Nat Commun, 2019. **10**(1): p. 2581.
146. Briant, L.J.B., et al., *CPT1a-Dependent Long-Chain Fatty Acid Oxidation Contributes to Maintaining Glucagon Secretion from Pancreatic Islets.* Cell Rep, 2018. **23**(11): p. 3300-3311.
147. Pugliese, G., et al., *Diabetic kidney disease: new clinical and therapeutic issues. Joint position statement of the Italian Diabetes Society and the Italian Society of Nephrology on "The natural history of diabetic kidney disease and treatment of hyperglycemia in patients with type 2 diabetes and impaired renal function"*. Journal of Nephrology, 2020. **33**(1): p. 9-35.
148. Wing, M.R., et al., *DNA methylation profile associated with rapid decline in kidney function: findings from the CRIC study.* Nephrol Dial Transplant, 2014. **29**(4): p. 864-72.
149. Graham, S.E., et al., *Sex-specific and pleiotropic effects underlying kidney function identified from GWAS meta-analysis.* Nature Communications, 2019. **10**(1): p. 1847.
150. Wang, C., et al., *Downregulation of ZNF365 by methylation predicts poor prognosis in patients with colorectal cancer by decreasing phospho-p53 (Ser15) expression.* Oncol Lett, 2020. **20**(4): p. 85.
151. Brigant, B., et al., *TRIMming down to TRIM37: Relevance to Inflammation, Cardiovascular Disorders, and Cancer in MULIBREY Nanism.* Int J Mol Sci, 2018. **20**(1).

152. Tao, Y., et al., *TRIM37 promotes tumor cell proliferation and drug resistance in pediatric osteosarcoma*. Oncol Lett, 2017. **14**(6): p. 6365-6372.
153. de Beer, J.C. and L. Liebenberg, *Does cancer risk increase with HbA1c, independent of diabetes?* Br J Cancer, 2014. **110**(9): p. 2361-8.
154. Hope, C., et al., *Relationship between HbA1c and cancer in people with or without diabetes: a systematic review.* Diabet Med, 2016. **33**(8): p. 1013-25.
155. Kwak, S.H. and K.S. Park, *Pathophysiology of Type 2 Diabetes in Koreans.* Endocrinol Metab (Seoul), 2018. **33**(1): p. 9-16.
156. Ohn, J.H., et al., *10-year trajectory of β-cell function and insulin sensitivity in the development of type 2 diabetes: a community-based prospective cohort study.* Lancet Diabetes Endocrinol, 2016. **4**(1): p. 27-34.
157. Neidhart, M., *Chapter 5 - DNA Methylation and Epidemiology*, in *DNA Methylation and Complex Human Disease*, M. Neidhart, Editor. 2016, Academic Press: Oxford. p. 67-79.

VII. 국문초록

제2형 당뇨 관련 후성유전학 지표 발굴 연구

서울대학교 보건대학원

보건학과 유전체 & 건강 빅데이터 전공

김 하 경

제2형 당뇨는 췌장의 베타 세포 (beta cell)의 기능 장애와 인슐린 저항성으로 인한 고혈당증을 특징으로 하는 만성질환이다. 한국의 제2형 당뇨 유병률은 성인의 12.4%에 달하고 있다. 제2형 당뇨는 유전적 요인과 환경적 요인, 그리고 두 요인의 상호작용으로 인해 발생한다고 알려져 있지만, 추정유전율이 25~69%에 달하는 것에 비해 전장유전체분석 (Genome-wide association study, GWAS)를 통해 확인한 변이 (genetic variants)는 추정 유전율의 일부만을 설명할 뿐이다. 이러한 관점에서 유전적 요인과 환경적 요인의 상호작용을 밝히기 위해 후성유전학적 연구가 활발히 진행되고 있다.

후성유전(Epigenetics)는 DNA 염기서열의 변화 없이 유전자 발현에 영향을 주는 현상을 의미하며, 이는 생애에서 노출되는 환경적인 요인으로 인해 영향을 받는다고 알려져 있다. 후성유전의 대표적인 기전은 DNA 메틸화 (DNAm)와 히스톤 변형 (Histone modification)이 있다. 최근 연구에서는 DNA 메틸화를 포함한 후성유전학적 변화가 제2형 당뇨 및 미세 혈관 합병증의 발병 위험에 영향을 미친다는 것이

밝혀졌다. 뿐만 아니라, 고혈당에의 노출이 DNA 메틸화의 변화를 일으키고 나아가 당뇨병성 신장질환의 위험을 증가시킨다는 '대사성 기억 (metabolic memory)' 현상이 제기됨에 따라 제2형 당뇨에서의 DNA 메틸화 변화에 대한 연구의 중요성이 증가하고 있다. 따라서 본 연구는 한국인에서 제2형 당뇨와 당뇨병성 신장질환 (Diabetic kidney disease, DKD) 특이적인 DNA 메틸화 지표를 발굴하고 해당 지표들과 대사성 특성(metabolic trait)과의 연관성을 탐구하고자 수행되었다.

먼저, 전장메틸화영역 연관분석 (Methylome-wide association study)을 통해 제2형 당뇨 유병여부에 따른 DNA 메틸화 차이를 관찰하고 차이가 특이적인 지표를 발견하는 분석을 수행하였다. 해당 지표들을 이용해 DNA 메틸화 점수를 계산하였고, 이 점수에 따라 10년간 추적 관찰한 대상자들에게서 제2형 당뇨병의 발생이 차등적으로 관찰되는지에 대한 분석을 수행해 제2형 당뇨 특이적 DNA 메틸화 분석결과의 신뢰성을 확보하였다. 또한, 같은 분석방법을 사용해 당뇨병성 신장질환 특이적인 DNA 메틸화 지표를 발견하는 분석이 이어서 수행되었다. 마지막으로, 발견된 제2형 당뇨/당뇨병성 신장질환의 DNA 메틸화 마커와 대사성 특성의 연관, 인과관계를 분석해 제2형 당뇨에서의 DNA 메틸화의 기전을 밝히고자 하였다.

말초 혈액 백혈구 샘플에서 DNA를 추출한 후, KoGES 코호트의 356 명과 HTS 코호트의 일부 대상자는 Illumina 사의 Infinium HumanMethylation 450 BeadChip으로 어세이 되어 유전체 내 약 450,000 개 이상의 DNA 메틸화 위치에 대한 DNA 메틸화 수준의 데이터를 얻었으며, 일부 HTS 코호트 대상자와 서울대병원 당뇨 코호트의 대상자들은 Illumina 사의 Infinium Methylation Epic Beadchip으로

어세이되어 유전체 내 총 850,000개 이상의 DNA 메틸화 수준의 데이터를 얻을 수 있었다. 이 중 Infinium HumanMethylation 450 BeadChip으로 어세이 된 자료들은 DNA 메틸화 마커 대치 (imputation) 방법 중 하나인 'methylImp'를 사용하여 주변 마커들과의 상관관계를 이용한 원리로 대치되어 총 850,000개 이상의 마커를 얻을 수 있었다.

본 연구에서는 197명의 제2형 당뇨병 대상자와 232명의 비 당뇨병 대조군을 사용한 사례-대조 연구 설계 (서울대병원 당뇨 코호트)에서 전장메틸화영역연관분석을 수행하였다. 제2형 당뇨병 그룹은 87명의 당뇨병성 신장질환 대상자와 80명의 비당뇨병성 신장질환군 (제2형 당뇨 유병 기간 10년 이상의 환자로 환장)으로 세분화되었다. 추가적으로 한국인 가족-쌍둥이 (HTS) 코호트 및 한국인유전체역학조사사업 (KoGES) 코호트의 2 개의 인구 기반 코호트에서 추가로 819 명의 대상자를 사용하여 발견된 DNA 메틸화 지표와 대사성 특성의 연관성을 조사하였다. 대사성 특성과 DNA 메틸화 지표의 인과 관계를 조사하기 위해 멘델리안 무작위 분석방법 (Mendelian randomization, MR)이 적용되었다.

본 연구에서는 제2형 당뇨의 8 개의 차등메틸화 영역 (Differentially methylated site, DMS) (각각 BMP8A, NBPF20, STX18, ZNF365, CPT1A, TRIM37 및 TXNIP에서 2 개)을 DNA 메틸화연구에서의 유의성 임계값($P < 9.0 \times 10^{-8}$)에서 확인했으며 이 중 TXNIP과 CPT1A에 위치한 3 개의 차등메틸화 영역은 기존 연구에서 확인된 영역이었다. 8개의 DNA 메틸화 마커로 구성된 DNAm 점수를 개인별로 계산하였으며, DNAm 점수의 1분위와 10분위의 상대적 위험이 2.86 (95% 신뢰구간 1.10-7.44)으로 나타났다. 전장메틸화연관분석이 수행된 것과 독립적인 전향

코호트 (KoGES)에서 제2형 당뇨병 발병 위험을 예측할 수 있어, 발견된 제2형 당뇨 특이적인 DNA 메틸화 마커의 신뢰성을 확보할 수 있었다. 멘델리안 무작위분석 방법을 이용하여 제2형 당뇨와 관련된 차등메틸화영역은 당뇨병 발병에 인과적인 효과가 없다는 것을 확인하였다. 이 과정에서 이용된 내생변수로 DNA 메틸화 정량 유전자좌 (methylation quantitative loci, mQTL)를 사용하였는데, 이 중 선행된 전장유전체 분석에서 보고된 제2형 당뇨와 대사특성 관련 단일염기다형성 (Single nucleotide polymorphism, SNP)이 발견되었다. 이는 해당 DNA 메틸화 정량 유전자좌와 관련이 있는 차등메틸화영역과 제2형 당뇨의 연관성에 교란효과가 있을 수 있음을 시사한다. CPT1A 및 TXNIP의 DMR은 공복 혈당, HbA1c 및 체질량 지수를 포함한 정량적 대사 특성과 관련이 있었으며, 특히 CPT1A의 DNA 메틸화는 공복 혈당에 인과적으로 조절되는 것으로 보였다. 또한 167명의 당뇨병성 신장 대상자 및 비 당뇨병성신증 대조군의 전장메틸화연관분석 연구에서 당뇨병성 신장질환과 관련된 3 개의 차등메틸화 영역(COMMD1, TMOD1 및 FHOD1)을 확인할 수 있었다. 당뇨병성 신장질환의 차등메틸화영역은 제2형 당뇨병에서 발견된 차등메틸화영역과 통계적으로 의미 있는 수준에서 공통적인 부분을 발견할 수 없었다. 당뇨병성 신장질환의 차등메틸화 영역은 유전변이에 의해 많은 부분이 설명 되고 있었으며, 멘델리안 무작위분석방법을 통해 당뇨병성 신장질환에 인과적인 영향을 미치고 있음을 확인할 수 있었다. 추정된 사구체 여과율 (eGFR)은 당뇨병성 신장질환에서 발견된 3개의 차등메틸화 영역과 인과 관계가 관찰되었다.

결론적으로, 우리는 동아시아 인구에서 제2형 당뇨병과 관련된 8 개의

차등메틸화영역을 메틸화연구에서의 유의성 임계값 ($P < 9.0 \times 10^{-8}$)에서 발견하였으며 이 중 5개는 본 연구에서 새롭게 밝혀진 DNA 메틸화 지표이다. 또한 당뇨병성 신장질환과 관련된 3개의 차등메틸화영역을 확인하였으며, 이는 제2형 당뇨의 영역과는 공통적인 부분이 발견되지 않았다. 본 연구 결과는 당뇨병성 신장질환의 후성유전학적 기전이 제2형 당뇨의 기전과는 다를 것임을 시사한다.

Keywords: 후성유전학, DNA 메틸화, 후성유전체연관분석, 전장메틸화영역연관분석, 제2형 당뇨, 당뇨병성신장질환

Student Number: 2018-35969