Aus dem

Lehrstuhl für Genetische Epidemiologie, IBE

Institut der Ludwig-Maximilians-Universität München

DNA Methylation Patterns Associated with Cardiometabolic Risk Factors

Dissertation

zum Erwerb des Doctor of Philosophy (Ph.D.)

an der Medizinischen Fakultät

der Ludwig-Maximilians-Universität München

vorgelegt von

Liye Lai

aus

Han Dan / China

Jahr

2025

Mit Genehmigung der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

Erstes Gutachten:	Prof. Dr. Annette Peters
Zweites Gutachten:	Prof. Dr. Eva Hoster
Drittes Gutachten:	Prof. Dr. Gunnar Schotta
Viertes Gutachten:	Prof. Dr. Axel Imhof

Dekan: Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung: 27.11.2025

Affidavit

Lai, Liye

Surname, first name

I hereby declare, that the submitted thesis entitled

DNA Methylation Patterns Associated with Cardiometabolic Risk Factors

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the dissertation presented here has not been submitted in the same or similar form to any other institution for the purpose of obtaining an academic degree.

Munich, 17.12.2025

Place, Date

Liye Lai

Signature doctoral candidate

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Dean's Office Medical Faculty
Doctoral Office

Confirmation of congruency between printed and electronic version of the doctoral thesis

Lai, Liye

Surname, first name

I hereby declare that the electronic version of the submitted thesis, entitled

DNA Methylation Patterns Associated with Cardiometabolic Risk Factors

is congruent with the printed version both in content and format.

Munich, 17.12.2025

Place, Date

Liye Lai

Signature doctoral candidate

Table of Contents

List of abbreviations

AHRR	Aryl Hydrocarbon Receptor Repressor
ABCG1	ATP-Binding Cassette Sub-Family G Member 1
BMI	Body Mass Index
BS	Bisulphite
CMDs	Cardiometabolic Diseases
CKD	Chronic Kidney Disease
CPT1A	Carnitine Palmitoyl Transferase 1 A
COPD	Chronic Obstructive Pulmonary Disease
DMPs	Differentially Methylated Positions
EWASs	Epigenome-Wide Association Studies
FPG	Fasting Plasma Glucose
FDR	False Discovery Rate
GPT2	Glutamate Pyruvate Transaminase 2
HbA1c	Hemoglobin A1c
HOMA-B	Homoeostasis Model Assessment of Beta Cell Function
HOMA-IR	Homoeostasis Model Assessment of Insulin Resistance
KORA	Cooperative Health Research in the Region of Augsburg
MAN2A2	Mannosidase Alpha Class 2a Member 2
NGT	Normal Glucose Tolerance

oxBS	Oxidative Bisulphite
OGTT	Oral Glucose Tolerance Test
SNPs	Single Nucleotide Polymorphisms
T2D	Type 2 Diabetes
TXNIP	Thioredoxin-Interacting Protein
5mC	5-methylcytosine
5hmC	5-hydroxymethylcytosine

List of publications

This thesis consists of the following two publications:

1. Lai L, Matías-García PR, Kretschmer A, Gieger C, Wilson R, Linseisen J, Peters A, Waldenberger M. Smoking-Induced DNA Hydroxymethylation Signature Is Less Pronounced than True DNA Methylation: The Population-Based KORA Fit Cohort. Biomolecules. 2024 Jun 5;14(6):662. doi: 10.3390/biom14060662. PMID: 38927065; PMCID: PMC11201877.

2. Lai L, Juntilla DL, Del M, Del C Gomez-Alonso M, Grallert H, Thorand B, Farzeen A, Rathmann W, Winkelmann J, Prokisch H, Gieger C, Herder C, Peters A, Waldenberger M. Longitudinal association between DNA methylation and type 2 diabetes: findings from the KORA F4/FF4 study. Cardiovasc Diabetol. 2025 Jan 18;24(1):19. doi: 10.1186/s12933-024-02558-8. PMID: 39827095; PMCID: PMC11748594.

Contribution to the publications

Contribution to paper I

The first publication, titled ‘Smoking-induced DNA hydroxymethylation signature is less pronounced than true DNA methylation: The Population-Based KORA Fit Cohort’, investigated the relationship between smoking and DNA methylation/hydroxymethylation. We identified smoking-associated hydroxymethylated CpG sites with suggestive links which offer promising avenues for future research. I contributed to the conceptualization, methodology, formal analysis, visualization, writing - original draft, and writing - review & editing.

Contribution to paper II

The second publication, titled ‘Longitudinal Association between DNA Methylation and Type 2 Diabetes: Findings from the KORA F4/FF4 Study’, identified novel differentially methylated loci associated with T2D and changes in diabetes status through a longitudinal approach. I contributed to the conceptualization, methodology, formal analysis, visualization, writing - original draft, and writing - review & editing.

Introductory summary

1 Background

1.1 Cardiometabolic risk factors and DNA methylation patterns

Cardiometabolic diseases (CMDs) comprise a spectrum of interconnected disorders, including metabolic conditions like obesity and T2D, as well as cardiovascular complications such as ischemic heart disease and heart failure. Well-established risk factors such as age, family history, obesity, hypertension, type 2 diabetes (T2D), dyslipidaemia, and smoking play a crucial role in the onset and progression of CMDs. Additionally, non-cardiac conditions such as liver disease and chronic kidney disease (CKD) can worsen the disease severity [1, 2].

Understanding the molecular mechanisms underlying CMDs remains a challenge despite their growing recognition as a major public health concern [3]. Recent studies highlight epigenetic modifications as a potential link between environmental exposures and CMD risk, as they influence gene expression without changing the DNA sequence [4]. Among these modifications, DNA methylation is the most thoroughly investigated, attracting significant attention for its involvement in CMD development through pathways such as inflammation, vascular dysfunction, and insulin resistance [5-8].

Epigenome-wide association studies (EWASs) are used to discover the relationship between DNA methylation and cardiometabolic traits, helping uncover the molecular mechanisms behind CMDs. This understanding can improve CMD diagnostics, facilitate personalized medicine, and support the development of targeted therapies.

1.2 Smoking and DNA methylation/hydroxymethylation

Smoking, a major cardiometabolic risk factor, remains widespread globally and is linked to numerous adverse health effects [9]. DNA methylation is thought to mediate the impact of tobacco exposure by modifying transcriptional activity, with research showing significant methylation changes in smokers [10-12] and in offspring exposed to maternal smoking during pregnancy [13]. However, the bisulfite (BS) transformation approach, frequently employed for identifying DNA methylation, is unable to differentiate between 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), resulting in most studies reporting them together. The impact of smoking on 5hmC, a key intermediate in the demethylation process [14], remains largely unexplored. Due to its presence in enhancers, promoters and transcriptional regulatory elements, 5hmC plays a vital role in gene regulation [15]. Unlike 5mC, which is often associated with gene repression, 5hmC can prevent transcriptional repressor binding, counteracting 5mC's inhibitory effects [16]. The combined use of bisulfite (BS) and oxidative bisulfite (oxBS) treatment enables the separate detection of true 5mC and 5hmC signals [17]. Distinguishing these modifications is crucial for understanding the molecular mechanisms behind smoking-related epigenetic changes.

1.3 Type 2 diabetes and DNA methylation

The global epidemic of obesity and T2D, largely driven by poor dietary habits and physical inactivity, has significantly contributed to the growing burden of CMDs [18]. T2D, marked by chronic hyperglycemia, has been associated with alterations in DNA methylation levels, which may impact transcriptional activity [19, 20]. While cross-sectional studies have identified methylation signatures linked to T2D in various tissues, such as blood and pancreatic islets [21-24], the causal

and temporal relationships remain poorly understood [25]. Methylation changes may either play a causal role in disease development or act as associative biomarkers [26]. Given the dynamic nature of glycemic traits before T2D onset, it is crucial to comprehend how methylation patterns progress to prediabetes and T2D from normal glucose tolerance (NGT) status. Large-scale longitudinal studies are necessary to investigate these associations across multiple time points.

1.4 Aims of this study

This thesis explores DNA methylation-based biomarkers for cardiometabolic-related traits, specifically smoking and T2D, and highlights the potential of epigenetics to revolutionize the understanding and management of cardiometabolic diseases. Specifically, this thesis had two main objectives:

(1) Explore how smoking influences DNA methylation patterns by distinguishing 5mC from 5hmC modifications.

(2) Provide a deeper understanding of methylation dynamics in the progression from normoglycemia to prediabetes and T2D by examining DNA methylation changes over time.

2 Methods

2.1 Study population

2.1.1 KORA Fit

The Cooperative Health Research in the Region of Augsburg (KORA) Fit cohort, a follow-up examination carried out from 2018 to 2019, extended the four foundational cohorts (S1, S2, S3, and S4). Individuals born between 1945 and 1964

who agreed to recontact were requested for a follow-up assessment, with 3,059 individuals participating. Comprehensive details about this study have been previously published [27]. For the analysis, 1,717 participants with DNA methylation data who passed quality control were included. To specifically investigate methylation and hydroxymethylation, a segment of 563 individuals was chosen, consisting of individuals who took part in the initial S4 survey as well as the subsequent KORA Fit assessment.

2.1.2 KORA F4/FF4

The KORA F4 and FF4 studies serve as follow-ups to the KORA S4 study. Comprehensive details regarding the design and measurements of the KORA cohort have been previously published [27]. The examination comprised 3,501 data points (n=2,556) throughout the KORA F4 (n=1,696) and FF4 (n=1,805), all with methylation measurements collected at least once during two time points. Among those individuals, 945 individuals had methylation signatures examined at two visits.

2.2 Exposure assessment

2.2.1 Smoking status measurement

Smoking status was categorized into three categories: current, former, and non-smokers. Individuals with no history of smoking were classified as non-smokers, whereas those who had smoked in the past but were not smoking during the interview period were categorized as former smokers. Participants who reported smoking regularly or occasionally (one cigarette per day or fewer) were categorized as current smokers.

2.2.2 T2D status measurement

Individuals with known T2D were identified through self-report, which was subsequently verified by the primary care physician or through an evaluation of medical records. Participants without a known diagnosis of T2D performed a 75 g oral glucose tolerance test (OGTT) after fasting for at least eight hours. NGT, prediabetes, and the diagnosis of newly identified T2D was established based on the 1999/2006 WHO criteria [28]. Participants with either a new diagnosis of T2D or a prior diagnosis were categorized as having T2D. Fasting plasma glucose (FPG), hemoglobin A1c (HbA1c), HOMA-B (beta-cell function), and HOMA-IR (insulin resistance) were measured as previously described [29].

2.3 Outcome assessment

2.3.1 DNA methylation/hydroxymethylation in KORA Fit study

In the KORA Fit study, after DNA extraction, BS (5mC+5hmC) and oxBS (5mC) treatments were used to differentiate between 5mC and 5hmC. Methylation was quantified using the Illumina EPIC BeadChip and analyzed with GenomeStudio software. Additional quality control and preprocessing were conducted [30], primarily following the CPACOR pipeline. Quality control involved excluding low-quality samples, sex-mismatched probes, and those influenced by single nucleotide polymorphisms (SNPs) or cross-reactivity. For combined 5mC+5hmC and actual 5mC methylation, 1,717 individuals and 734,349 CpG sites were left for the statistical analysis. 5hmC signals at base-pair level precision were determined by removing the oxBS (5mC) signal from the BS (5mC+5hmC) signal for every CpG sites. As to the subsequent hydroxymethylation analysis, the CpG

sites and individuals common to both the total 5mC+5hmC and true 5mC methylation groups were used, leading to 563 individuals and 756,737 CpG sites.

2.3.2 DNA methylation in KORA F4/FF4

In this prospective study, we examined data from the KORA F4 and FF4 cohorts, spanning a period of seven years. Whole blood DNA methylation levels were measured by the Illumina 450K Infinium Methylation BeadChip for the KORA F4 and the Infinium MethylationEPIC BeadChip for the KORA FF4. Quality control procedures were carried out according to the CPACOR preprocessing pipeline, utilizing the minfi2 package [30]. Probe intensities were normalized through quantile normalization for both cohorts. After quality control, 414,872 CpG sites remained in the KORA F4 dataset, while 806,228 CpG sites were retained in the KORA FF4 dataset, with 383,057 CpG sites overlapping between the two. After excluding probes from sex chromosomes, 374,054 CpG candidates were kept.

2.4 Statistical methods

2.4.1 Smoking effects on DNA methylation/hydroxymethylation

EWAS analyses were conducted using multivariate linear regression, with smoking status (current, former, non-smokers) as the predictor and methylation levels as the dependent variable. The analysis was additionally adjusted for several covariates, consisting of age, sex, body mass index (BMI), leukocyte proportions (calculated using the Houseman algorithm), and technical effects captured by principal components. Differentially methylated positions (DMPs) were identified after applying false discovery rate (FDR) multiple correction ($p < 0.05$), while an indicating threshold of $p < 1 \times 10^{-5}$ was used for the hydroxymethylation (5hmC) analyses.

2.4.2 T2D effects on DNA methylation

Linear mixed-effects models by adding random intercepts were employed to investigate the relationship between diabetes status (NGT, prediabetes, T2D) and DNA methylation. These models were adjusted for factors including follow-up time, baseline age, sex, BMI, smoking status, estimated cell types, and technical effects. Additionally, we applied the same model to explore the link between DNA methylation and four glycemic/insulin-related variables (FPG, HbA1c, HOMA-B, and HOMA-IR). Associations were regarded as significant if the *p*_FDR value was <0.05. To assess the variation in the rate of methylation alteration between diabetes categories, we also investigated the interaction impacts between diabetes status and follow-up duration. Further, we investigated CpG sites related to the persistence of prediabetes or T2D, in addition to those linked to the transition from NGT to prediabetes or T2D. Lastly, we evaluated how the CpG sites that were identified correlate with the levels of gene expression.

3 Results

3.1 Smoking-induced DNA hydroxymethylation signature is less pronounced than true DNA methylation

We investigated DNA methylation changes in individuals designated as current, former, and non-smokers. We first investigated the relationship between total 5mC+5hmC methylation signals and smoking status, detecting 38,575 and 82 differentially methylated positions (DMPs) related to current and former smoking, respectively. A significant number of these DMPs have been previously reported, such as those in the aryl hydrocarbon receptor repressor (*AHRR*) gene, alongside some novel findings. Next, we employed sequential BS and oxBS treatments

to differentiate 5hmC from 5mC level. This more detailed analysis revealed 33 DMPs related with current smoking; and 1 DMPs linked to former smoking in the 5mC group, respectively, showing strong consistency in the trend of effects and substantial convergence in loci among the 5mC+5hmC and 5mC methylation categories. Additionally, we identified 8 and 2 DMPs associated with current and former smoking in the 5hmC group, using a suggestive threshold. A notable example is cg16972043, which is labeled as the glutamate pyruvate transaminase 2 (*GPT2*) gene.

3.2 Longitudinal association between DNA methylation and type 2 diabetes

This study leveraged prospective data with longitudinal measurements to investigate the relationship between diabetes status and DNA methylation. We analyzed 3,501 data points from 2,556 individuals employing multivariate linear mixed-effects models with random intercepts, detecting 64 candidate CpG sites linked to T2D. Among these, 49 loci, including the thioredoxin-interacting protein (*TXNIP*) and ATP-binding cassette sub-family G member 1 (*ABCG1*) genes, showed coherent trends in direction in our longitudinal analysis, corroborating previous cross-sectional findings. Notably, we discovered 15 previously uncharacterized CpG sites within 10 distinct genes.

Among the 64 T2D-related CpG candidates, eight exhibited differing annual methylation alteration patterns between the NGT and T2D categories, while seven associated with transition from NGT to prediabetes or T2D, encompassing sites in the mannosidase alpha class 2a member 2 (*MAN2A2*) and carnitine palmitoyltransferase 1A (*CPT1A*) genes. Prospective analysis also identified relationships between methylation and FPG at 128 CpG sites, HbA1c at 41 CpG sites, and HOMA-IR at CpG 57 sites. Furthermore, 104 significant associations were

detected between T2D-related CpG sites and their respective gene expression levels, including 40 distinct CpG candidates and 96 distinct gene transcripts.

4 Discussion

4.1 Smoking-induced DNA hydroxymethylation signature is less pronounced than true DNA methylation

We explored various DNA methylation adjustments in individuals designated as current, former, and non-smokers. This represents, based on available evidence, the pioneering epigenome-wide investigation of smoking's impacts on blood leukocyte DNA methylation, distinguishing between 5mC and 5hmC modifications, particularly applying the Illumina EPIC BeadChip. The *AHRR* gene repeatedly emerged as the highly prominently impacted candidate in smoking-related studies [31, 32], a finding we observed in our cohort as well. In former smokers, while most differentially methylated CpG sites reverted to levels similar to those of non-smokers after quitting smoking, a fraction showed persistent methylation differences even after smoking cessation, though with reduced effect sizes. These specific CpG sites may serve as reliable biomarkers, providing insights into a person's smoking history and indicating long-term wellness impacts [33, 34].

In this study, oxBS treatment enabled the precise quantification of 5mC. All 5mC DMPs significantly related with current smoking were also detected in the traditional 5mC+5hmC category, including well-known loci *AHRR*, *RARA*, and *F2RL3*, confirming their strong association with smoking. Additionally, we observed a high level of agreement in the trends of effects between the 5mC+5hmC and 5mC categories in current smokers, along with many CpG sites exhibiting hypomethylation. The DNA hydroxymethylation pattern associated with smoke exposure

was less evident compared to that of true DNA methylation, likely due to its lower abundance in blood [35, 36]. One notable finding was the identification of the hydroxymethylated CpG site cg16972043, labelled as *GPT2*, which showed a suggestive association with current smoking but did not pass the FDR multiple correction. Recent studies have underscored *GPT2*'s involvement in modulating smoking-induced metabolic changes and harm in respiratory epithelial cells, especially via lipid production [37]. *GPT2* has also been associated with the onset of chronic obstructive pulmonary disease (COPD) in leukocytes, highlighting its significance in pulmonary disorders. The discovery of these innovative smoking-related hydroxymethylated CpG candidates paves the way for further exploration in subsequent studies.

4.2 Longitudinal association between DNA methylation and type 2 diabetes

In our study, we investigated differentially methylated loci associated with T2D and shifts in diabetes status using a prospective method. We identified 64 significant CpG sites that distinguished participants with T2D from those with NGT, spanning 49 unique genomic loci. Notably, *TXNIP* stood out as the most significant gene, aligning with previous research because of its involvement in regulating pancreatic β-cells and its potential as a therapeutic target for diabetes [38, 39]. Given that disorders in glucose metabolism often precede the diagnosis of diabetes, we found that cg19693031 (*TXNIP*) and cg06500161 (*ABCG1*) were concurrently linked to FPG, HbA1c, HOMA-IR, and T2D. This suggests that these loci may function as valuable biomarkers for glycaemic regulation and diabetes risk [40].

DNA methylation is a well-established epigenetic mechanism that is influenced by environmental factors. Various exposures, such as chemical agents and metabolic disorders, can induce global or site-specific methylation alterations, which in turn impact gene expression and transcription factor binding. In our study, we observed that a hypomethylated CpG candidate in *TXNIP* exhibited a faster decrease in participants with T2D in contrast to those with NGT. This resulted in a greater discrepancy in methylation over time, potentially leading to increased *TXNIP* expression [41, 42]. These dynamic methylation patterns underscore their sensitivity to diabetes progression and highlight their potential as targets for therapeutic intervention.

In our following analysis, we discovered seven CpG sites associated with the progression from NGT to prediabetes and T2D, consisting of cg11183227 labelled to *MAN2A2*. These sites hold potential as valuable biomarkers for monitoring disease progression. By incorporating gene expression measurement, 104 associations were discovered between unique 40 CpG candidates and 96 gene transcripts. Of note, methylation at cg06500161 labelled to *ABCG1* exhibited an inverse relationship with its expression level, indicating that hypomethylation at this site may play a role in the progression of T2D and related conditions.

4.3 Strengths and limitations

The present study offers several strengths. We successfully distinguished between true 5mC and 5hmC levels through combined BS and oxBS treatments, particularly when coupled with the Infinium MethylationEPIC BeadChip. Furthermore, our prospective analysis, covering a span of seven years, integrated both DNA methylation signatures and diabetes status, measured via OGTT in participants without a prior diabetes diagnosis. However, there are limitations to our

study. The lack of a validation cohort underscores the necessity for future research to confirm these findings in separate populations. Furthermore, since DNA was sourced from blood, tissue-specific differences in methylation patterns may not have been completely captured.

5 Conclusion

Firstly, by differentiating between 5mC and 5hmC levels in whole blood DNA samples, we uncovered different smoking-related DNA methylation changes. Our findings not only validated previously identified smoking-associated CpG candidates but also identified many novel signatures linked to smoking. While hydroxymethylation was less prominently linked to smoking in whole blood DNA, suggestive CpG candidates warrant further investigation in future studies.

Additionally, our research provided valuable insights into the relationship between DNA methylation and T2D via a prospective method with longitudinal measures. We discovered novel CpG sites linked to T2D and observed differing rates of methylation alterations at candidates across various diabetes status categories. This study also highlighted DNA methylation's potential as a biomarker for tracking diabetes advancement and illustrated its connection to gene expression levels.

In conclusion, this thesis explored DNA methylation-based biomarkers for cardiometabolic traits—smoking and T2D—underscoring the potential of DNA methylation in reshaping the approach to cardiometabolic diseases and advancing personalized healthcare.

References

1. Eroglu, T., F. Capone, and G.G. Schiattarella, The evolving landscape of cardiometabolic diseases. EBioMedicine, 2024. 109: p. 105447.

2. Rutters, F., et al., Lifestyle interventions for cardiometabolic health. Nat Med, 2024. 30(12): p. 3455-3467.

3. Liu, M., et al., Mechanisms of inflammatory microenvironment formation in cardiometabolic diseases: molecular and cellular perspectives. Front Cardiovasc Med, 2024. 11: p. 1529903.

4. Raghubeer, S., The influence of epigenetics and inflammation on cardiometabolic risks. Semin Cell Dev Biol, 2024. 154(Pt C): p. 175-184.

5. Antoun, E., et al., DNA methylation signatures associated with cardiometabolic risk factors in children from India and The Gambia: results from the EMPHASIS study. Clin Epigenetics, 2022. 14(1): p. 6.

6. Barouti, Z., et al., Effects of DNA methylation on cardiometabolic risk factors: a systematic review and meta-analysis. Arch Public Health, 2022. 80(1): p. 150.

7. Colicino, E. and G. Fiorito, DNA methylation-based biomarkers for cardiometabolic-related traits and their importance for risk stratification. Curr Opin Epidemiol Public Health, 2023. 2(2): p. 25-31.

8. Hao, G., et al., The role of DNA methylation in the association between childhood adversity and cardiometabolic disease. Int J Cardiol, 2018. 255: p. 168-174.

9. Sultan, S. and F. Lesloom, Association of cigarette smoking with cardiometabolic risk factors: A cross-sectional study. Tob Induc Dis, 2024. 22.

10. Heikkinen, A., S. Bollepalli, and M. Ollikainen, The potential of DNA methylation as a biomarker for obesity and smoking. J Intern Med, 2022. 292(3): p. 390-408.

11. Joehanes, R., et al., Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet, 2016. 9(5): p. 436-447.
12. Ambatipudi, S., et al., Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics, 2016. 8(5): p. 599-618.
13. Fragou, D., et al., Smoking and DNA methylation: Correlation of methylation with smoking behavior and association with diseases and fetus development following prenatal exposure. Food Chem Toxicol, 2019. 129: p. 312-327.
14. Prasad, R., T.J. Yen, and A. Bellacosa, Active DNA demethylation-The epigenetic gatekeeper of development, immunity, and cancer. Adv Genet (Hoboken), 2021. 2(1): p. e10033.
15. Xu, T. and H. Gao, Hydroxymethylation and tumors: can 5-hydroxymethylation be used as a marker for tumor diagnosis and treatment? Hum Genomics, 2020. 14(1): p. 15.
16. Kranzhöfer, D.K., et al., 5'-Hydroxymethylcytosine Precedes Loss of CpG Methylation in Enhancers and Genes Undergoing Activation in Cardiomyocyte Maturation. PLoS One, 2016. 11(11): p. e0166575.
17. Nestor, C., et al., Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. Biotechniques, 2010. 48(4): p. 317-9.
18. Fontvieille, E., et al., Body mass index and cancer risk among adults with and without cardiometabolic diseases: evidence from the EPIC and UK Biobank prospective cohort studies. BMC Med, 2023. 21(1): p. 418.
19. Kaimala, S., S.A. Ansari, and B.S. Emerald, DNA methylation in the pathogenesis of type 2 diabetes. Vitam Horm, 2023. 122: p. 147-169.
20. Ahmed, S.A.H., et al., The role of DNA methylation in the pathogenesis of type 2 diabetes mellitus. Clin Epigenetics, 2020. 12(1): p. 104.

21. 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.

22. Juvinao-Quintero, D.L., et al., DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. Clin Epigenetics, 2021. 13(1): p. 40.

23. Fraszczyk, E., et al., Epigenome-wide association study of incident type 2 diabetes: a meta-analysis of five prospective European cohorts. Diabetologia, 2022. 65(5): p. 763-776.

24. Hillary, R.F., et al., Blood-based epigenome-wide analyses of 19 common disease states: A longitudinal, population-based linked cohort study of 18,413 Scottish individuals. PLoS Med, 2023. 20(7): p. e1004247.

25. Hong, X., et al., Longitudinal Association of DNA Methylation With Type 2 Diabetes and Glycemic Traits: A 5-Year Cross-Lagged Twin Study. Diabetes, 2022. 71(12): p. 2804-2817.

26. Juvinao-Quintero, D.L., et al., Investigating causality in the association between DNA methylation and type 2 diabetes using bidirectional two-sample Mendelian randomisation. Diabetologia, 2023. 66(7): p. 1247-1259.

27. Holle, R., et al., KORA--a research platform for population based health research. Gesundheitswesen, 2005. 67 Suppl 1: p. S19-25.

28. ElSayed, N.A., et al., 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes-2023. Diabetes Care, 2023. 46(Suppl 1): p. S19-s40.

29. Luo, H., et al., Associations of plasma proteomics with type 2 diabetes and related traits: results from the longitudinal KORA S4/F4/FF4 Study. Diabetologia, 2023. 66(9): p. 1655-1668.

30. Bock, C., Analysing and interpreting DNA methylation data. Nat Rev Genet, 2012. 13(10): p. 705-19.

31. Langsted, A., et al., AHRR hypomethylation as an epigenetic marker of smoking history predicts risk of myocardial infarction in former smokers. Atherosclerosis, 2020. 312: p. 8-15.

32. Wilson, R., et al., The dynamics of smoking-related disturbed methylation: a two time-point study of methylation change in smokers, non-smokers and former smokers. BMC Genomics, 2017. 18(1): p. 805.

33. Morrow, J.D., et al., DNA Methylation Is Predictive of Mortality in Current and Former Smokers. Am J Respir Crit Care Med, 2020. 201(9): p. 1099-1109.

34. Langdon, R.J., et al., Epigenetic modelling of former, current and never smokers. Clin Epigenetics, 2021. 13(1): p. 206.

35. Shi, D.Q., et al., New Insights into 5hmC DNA Modification: Generation, Distribution and Function. Front Genet, 2017. 8: p. 100.

36. Ringh, M.V., et al., Tobacco smoking induces changes in true DNA methylation, hydroxymethylation and gene expression in bronchoalveolar lavage cells. EBioMedicine, 2019. 46: p. 290-304.

37. Yan, F., et al., Roles of glutamic pyruvate transaminase 2 in reprogramming of airway epithelial lipidomic and metabolomic profiles after smoking. Clin Transl Med, 2024. 14(5): p. e1679.

38. Wondafrash, D.Z., et al., Thioredoxin-Interacting Protein as a Novel Potential Therapeutic Target in Diabetes Mellitus and Its Underlying Complications. Diabetes Metab Syndr Obes, 2020. 13: p. 43-51.

39. Kar, A., et al., Thioredoxin Interacting Protein Inhibitors in Diabetes Mellitus: A Critical Review. Curr Drug Res Rev, 2023. 15(3): p. 228-240.

40. Davidsen, L., et al., Efficacy and safety of continuous glucose monitoring on glycaemic control in patients with chronic pancreatitis and insulin-treated diabetes: A randomised, open-label, crossover trial. Diabetes Obes Metab, 2025.
41. Yamazaki, M., et al., DNA methylation level of the gene encoding thioredoxin-interacting protein in peripheral blood cells is associated with metabolic syndrome in the Japanese general population. Endocr J, 2022. 69(3): p. 319-326.
42. Maeda, K., et al., Association between DNA methylation levels of thioredoxin-interacting protein (TXNIP) and changes in glycemic traits: a longitudinal population-based study. Endocr J, 2024. 71(6): p. 593-601.

Publication I

Title:	Smoking-Induced DNA Hydroxymethyla-tion Signature Is Less Pronounced than True DNA Methylation: The Population-Based KORA Fit Cohort
Authors:	Lai L, Matías-García PR, Kretschmer A, Gieger C, Wilson R, Linseisen J, Peters A, Waldenberger M
Journal:	Biomolecules
Status:	Published
Volume:	14(6)
Page:	662
Year:	2024
doi:	10.3390/biom14060662

biomolecules

MDPI

Article

Smoking-Induced DNA Hydroxymethylation Signature Is Less Pronounced than True DNA Methylation: The Population-Based KORA Fit Cohort

Liye Lai [1,2,3,*], Pamela R. Matías-García [1,3], Anja Kretschmer [3], Christian Gieger [1,3], Rory Wilson [1,3], Jakob Linseisen [4], Annette Peters [1,2,3,5] and Melanie Waldenberger [1,3,5,*]

1 Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), 85764 Neuherberg, Germany; pamela.matiasgarcia@helmholtz-munich.de (P.R.M.-G.); christian.gieger@helmholtz-munich.de (C.G.); wilson.rory@gmail.com (R.W.); annette.peters@helmholtz-munich.de (A.P.)
2 Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Pettenkofer School of Public Health, Faculty of Medicine, Ludwig Maximilians University, 81377 Munich, Germany
3 Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), 85764 Neuherberg, Germany; anja.kretschmer@helmholtz-munich.de
4 Epidemiology, Faculty of Medicine, University Hospital of Augsburg, University of Augsburg, 86156 Augsburg, Germany; jakob.linseisen@med.uni-augsburg.de
5 German Centre for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, 81377 Munich, Germany
* Correspondence: liye.lai@helmholtz-munich.de (L.L.); melanie.waldenberger@helmholtz-munich.de (M.W.); Tel.: +49-89-3187-1270 (M.W.)

Abstract: Despite extensive research on 5-methylcytosine (5mC) in relation to smoking, there has been limited exploration into the interaction between smoking and 5-hydroxymethylcytosine (5hmC). In this study, total DNA methylation (5mC+5hmC), true DNA methylation (5mC) and hydroxymethylation (5hmC) levels were profiled utilizing conventional bisulphite (BS) and oxidative bisulphite (oxBS) treatment, measured with the Illumina Infinium Methylation EPIC BeadChip. An epigenome-wide association study (EWAS) of 5mC+5hmC methylation revealed a total of 38,575 differentially methylated positions (DMPs) and 2023 differentially methylated regions (DMRs) associated with current smoking, along with 82 DMPs and 76 DMRs associated with former smoking (FDR-adjusted $p < 0.05$). Additionally, a focused examination of 5mC identified 33 DMPs linked to current smoking and 1 DMP associated with former smoking (FDR-adjusted $p < 0.05$). In the 5hmC category, eight DMPs related to current smoking and two DMPs tied to former smoking were identified, each meeting a suggestive threshold ($p < 1 \times 10^{-5}$). The substantial number of recognized DMPs, including 5mC+5hmC (7069/38,575, 2/82), 5mC (0/33, 1/1), and 5hmC (2/8, 0/2), have not been previously reported. Our findings corroborated previously established methylation positions and revealed novel candidates linked to tobacco smoking. Moreover, the identification of hydroxymethylated CpG sites with suggestive links provides avenues for future research.

Keywords: smoking; DNA methylation; hydroxymethylation; differentially methylated positions (DMPs); differentially methylated regions (DMRs); Illumina Infinium Methylation EPIC BeadChip

Citation: Lai, L.; Matías-García, P.R.; Kretschmer, A.; Gieger, C.; Wilson, R.; Linseisen, J.; Peters, A.; Waldenberger, M. Smoking-Induced DNA Hydroxymethylation Signature Is Less Pronounced than True DNA Methylation: The Population-Based KORA Fit Cohort. *Biomolecules* **2024**, *14*, 662. https://doi.org/10.3390/biom14060662

Academic Editor: Gyeong Hoon Kang

Received: 1 May 2024
Revised: 31 May 2024
Accepted: 3 June 2024
Published: 5 June 2024

1. Introduction

Although tobacco smoking is widely recognized as a harmful behaviour with significant impacts on human health, smoking or exposure to smoke continues to be prevalent worldwide. Tobacco smoking is a risk factor for and is a frequent cause of many adverse health consequences, such as chronic obstructive pulmonary disease (COPD) [1], cardiovascular diseases [2], asthma [3] and various forms of cancer, in particular lung cancer [4,5]. Moreover, smoking status appears to contribute to a poor prognosis in COVID-19 patients [6]. While the precise pathogenic mechanisms remain under investigation, it

is widely acknowledged that the induction of oxidative stress through the generation of excessive reactive oxygen species (ROS) by harmful chemicals is a key molecular event that predisposes individuals to inflammation, senescence and smoking-related illnesses [7,8].

Epigenetic mechanisms, specifically alterations in DNA methylation, have been suggested to moderate the impact of tobacco smoking, leading to changes in transcriptional activity and contributing to smoking-related diseases [9]. With the update of DNA methylation arrays, the impact of smoking on DNA 5-methylcytosine (5mC) methylation has been thoroughly investigated in blood cells from adults, revealing significant disparities between smokers and non-smokers [10,11], which can be even more conspicuous in specific tissues like vascular endothelial cells [12], and vulnerable groups like cancer patients [4]. The impact of tobacco smoking on DNA methylation is also prominent in the blood of newborns whose mothers smoked during pregnancy [13]. Previous studies also demonstrated that the link between cigarette smoking and methylation is dynamic, showing ongoing fluctuations in methylation levels even decades after smoking cessation. However, only a few studies have delved into the effect of smoking on DNA 5-hydroxymethylcytosine (5hmC) methylation, an intermediate oxidized form of 5mC involved in the active demethylation process. During active demethylation process, the ten-eleven translocation (TET) enzymes play a crucial role by oxidizing 5mC into 5hmC, further converting 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). Subsequently, the thymine DNA glycosylase (TDG)-dependent base excision repair (BER) transforms 5fC and 5caC into an unmethylated cytosine [14,15]. Due to their low abundance in the genome, 5fC and 5caC demonstrate limited stability [16]. In contrast to 5fC and 5caC, 5hmC is relatively stable and presents tissue specificity [17]. Given its enrichment in promoters, enhancers and transcriptional regulatory elements, 5hmC is intimately associated with the regulation of gene expression [18].

Recent studies have highlighted that smoking-induced oxidative stress can initiate the DNA demethylation pathway [19]. Additionally, 5hmC has emerged as an informative biomarker in mammalian development and diseases [20,21]. However, the traditional bisulphite (BS) conversion method, commonly used for detecting DNA methylation, cannot distinguish between 5mC and 5hmC [22]. As a result, most of the existing literature on DNA methylation reports 5mC and 5hmC signals jointly. Moreover, the Infinium HumanMethylation450 BeadChip has been predominantly utilized to identify smoking-associated differentially methylated positions (DMPs). In this study, the oxidative bisulphite (oxBS) treatment was employed to measure true 5mC and 5hmC signals separately (Figure 1A). We hypothesized that smoking-induced differential DNA methylation could potentially influence not only 5mC but also 5hmC patterns in leucocytes from blood samples. Initially, we examined total 5mC+5hmC methylation levels in 1717 participants classified as current, former and non-smokers from the Cooperative Health Research in the Region of Augsburg (KORA) Fit population-based cohort (Figure 1B). We employed the latest HumanMethylation EPIC BeadChip, providing expanded CpG site coverage compared to prior arrays (over 850,000 CpG sites). Subsequently, we evaluated 5mC and 5hmC methylation levels separately in a subset of 563 individuals.

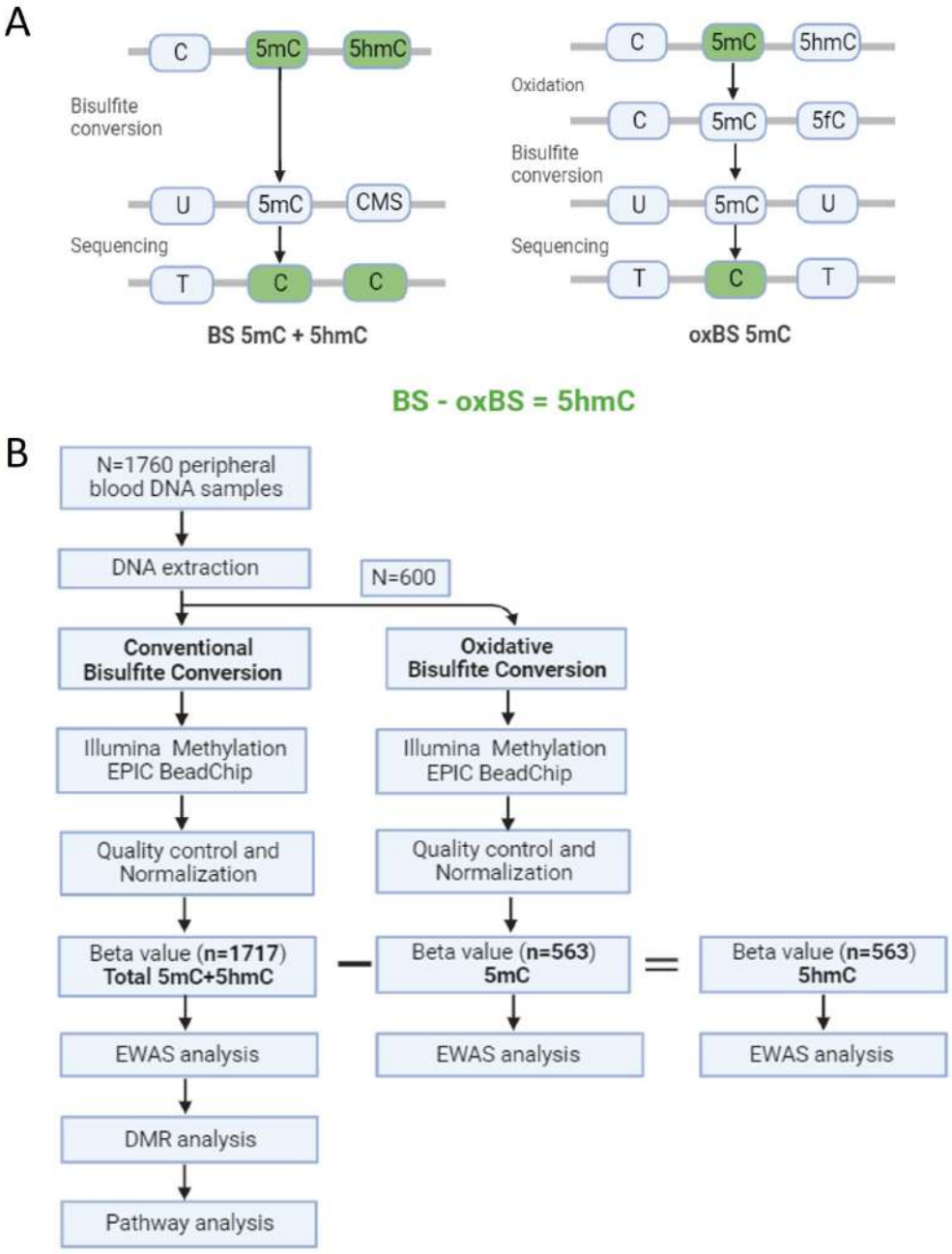


Figure 1. (**A**) Schematic overview depicting bisulphite conversion (BS) and oxidative BS. (**B**) Illustration of the study design.

2. Materials and Methods

2.1. Study Population

The analysis was based on data from the KORA Fit study, a follow-up study conducted between early 2018 and mid-2019, building upon the 4 cross-sectional baseline surveys (KORA S1, S2, S3 and S4 cohorts). All living participants of the KORA cohorts born between 1945 and 1964 who consented to be recontacted were invited for a new examination (n = 3059 or 64.4% of all eligible participants). Exhaustive information about this study has been described previously [23]. In total, 1760 participants with available data on DNA methylation were included in the analysis. Specifically, for the investigation into true methylation and hydroxymethylation, a subgroup comprising 600 participants from the KORA Fit study was considered. This subgroup included individuals who participated in both the S4 baseline survey and the KORA Fit examination. Individuals who self-declared

as either regular or occasional smokers (defined as 1 cigarette per day or less) at the time of the interview were classified as current smokers. Those who had never smoked were categorized as non-smokers, while individuals who had previously smoked but were not currently smoking at the time of the interview were classified as former smokers.

2.2. DNA Extraction and DNA Methylation Quantification

DNA extraction followed standard procedures. For the total 5mC+5hmC methylation processing, genomic DNA (750 ng) from 1160 individuals underwent BS conversion using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA, USA). Meanwhile, genomic DNA (1500 ng) from 600 individuals was split (750 ng each), and separate aliquots of each DNA sample were processed in parallel. One aliquot underwent BS treatment to generate total methylation (5mC+5hmC) signals, while the other aliquot underwent oxidation and then BS treatment to generate true methylation (5mC) signals, both using the TrueMethyl oxBS Module (Tecan Genomics, Redwood City, CA, USA). During BS treatment, 5mC and 5hmC are preserved as cytosines, whereas unmethylated cytosines are deaminated to uracil. Consequently, DNA methylation measured by the BS treatment reflects an amalgamation of 5mC and 5hmC. Upon oxidation, 5mC remains as 5mC, while 5hmC is converted into 5fC. The 5fC is susceptible to BS treatment, and it is deaminated into uracil (equivalent to an unmethylated cytosine), while 5mC is preserved as a cytosine upon BS treatment. Thus, oxBS conversion enables the specific measurement of nucleotide-level 5mC [24,25]. Subsequent methylation analysis for all samples was conducted on an Illumina (San Diego, CA, USA) iScan platform using the Infinium Methylation EPIC BeadChip v1, following standard protocols provided by Illumina. Initial quality control procedures of assay performance and generation of methylation data export files were carried out using GenomeStudio software version 2011.1 with Methylation Module version 1.9.0.

2.3. Preprocessing and Normalization

Raw intensities were imported, and further quality control and preprocessing were performed in R software (R v4.3.3), with the minfi package v1.48.0, primarily following the CPACOR pipeline [26]. Total methylation (5mC+5hmC) and true methylation (5mC) were processed separately. Samples with defective chips and over 20% missing values, along with sex-mismatching samples, were removed. Probes with detection *p*-values great than 0.01 in more than 5% of samples were set to missing. Furthermore, sex chromosomes and cross-reactive and SNP-related probes were removed. Subsequently, quantile normalization (QN) was independently performed on the signal intensities, which were categorized into the 6 probe types: type II red, type II green, type I green unmethylated, type I green methylated, type I red unmethylated, type I red methylated. β-values were then calculated by initiating with the BS signal, representing the total methylation (5mC+5hmC) signal at each CpG site. Total methylation β-values were computed as the ratio of the methylated signal over the sum of the methylated and unmethylated signals [27]. For the analysis of total 5mC+5hmC methylation, 1717 samples and 734,349 probes were retained for the final analysis. Similarly, 5mC β-values were calculated using the oxBS signal. Lastly, the level of 5hmC at a single-nucleotide resolution was estimated by subtracting the oxBS measure (5mC) from the BS measure (5mC+5hmC) at each probe. Specifically, for the hydroxymethylation, only probes and samples that were common between the 5mC+5hmC and 5mC datasets were kept, resulting in 563 samples and 756,737 probes. Additionally, subtracting 5mC from 5mC+5hmC is known to introduce negative β-values, so any negative β-values were set to a value close to zero (1×10^{-7}).

2.4. Differential Methylation Analysis

An Epigenome-wide association study (EWAS) was carried out using a multivariate linear regression model, where smoking status (current, former, non-smokers) served as the exposure variable, and untransformed methylation β-values (ranging from 0 to 1) were used as the outcome. Recognizing that methylation levels in blood can be significantly

influenced by leukocyte composition, the houseman algorithm was employed to estimate white blood cell type proportions [28]. Additionally, principal components (PCs) of all non-negative control probes were calculated to account for technical effects. All epigenome-wide analyses were adjusted for the age at blood collection, sex, BMI, six estimated cell type proportions (monocytes, granulocytes, natural killer cells, B cells, CD4T cells and CD8T cells) and the first 5 principal components (PCs). To assess the epigenome-wide distribution of p values compared to the expected null distribution of p values, we calculated the inflation factor λ and generated quantile–quantile (QQ) plots. The inflation factor was defined as the ratio of the median of the observed log10-transformed p values to the median of the expected log10-transformed p values. We also applied bacon correction to mitigate bias and inflation of the test statistic. A probe was considered significantly differentially methylated with a false discovery rate (FDR)-adjusted (Benjamini–Hochberg) p value less than 0.05. Given the anticipated lower range of 5hmC methylation values, a less stringent suggestive threshold of $p < 1 \times 10^{-5}$ was employed when identifying 5hmC-associated differential methylation. EWAS Catalog (a database of epigenome-wide association studies) [29] was used to compare and select the novel smoking-associated CpG candidates. DMRs represent genomic regions with consistently different DNA methylations across multiple adjacent CpG sites. In addition to the single-site DMP analysis, we applied the comb-p function using the Enmix package (version 1.38.01), which provides quality control, analysis and visualization tools for Illumina DNA methylation BeadChip, to detect DMRs among current, former and non-smokers. In this analysis, regions were defined as sets of all probes containing ≥3 DMPs within 1000 base pairs of another probe and having false discovery rate (FDR)-adjusted p values less than 0.05.

2.5. Gene Enrichment Analyses

To gain insights into potential smoking-relevant biological processes, gene pathway analysis was performed in the context of differentially methylated CpG sites. This analysis utilized the GOmeth function from the missMethyl package (version 1.38.0), which accounts for the number of CpG sites per gene on the 450K/EPIC array and multi-gene-annotated CpGs. Independent pathways with an FDR $p < 0.05$ were considered significantly associated with smoking. Gene annotation was performed using the HumanMethylation EPIC probe annotation file.

3. Results

3.1. Characteristics of the Study Population

A total of 1717 participants were included in our study for further analyses after quality control, consisting of 217 current smokers, 719 former smokers and 781 non-smokers. The cohort characteristics are described in Table 1. Current smokers were younger and exhibited a lower prevalence of hypertension compared to non-smokers. Former smokers had a larger proportion of males and a higher BMI level. Both current and former smokers displayed an increased daily alcohol consumption, lower HDL cholesterol levels and higher triglycerides levels. All groups were comparable in terms of physical activity, diabetes status, HOMA-IR and HOMA-Beta levels.

Table 1. Characteristics of the study population.

Characteristics	All Participants	Current Smokers	Former Smokers	Non-Smokers
	1717	217	719	781
Age (years)	63 (59, 68)	61 (57, 65) ***	64 (59, 68)	63 (59, 68)
Male (%)	814 (46.3%)	105 (47.3%)	393 (53.5%) ###	316 (39.4%)
BMI (kg/m^2)	27.4 (24.5, 30.8)	26.2 (23.7, 30)	27.6 (24.8, 31.3) #	27.3 (24.5, 30.3)
Physical activity	1268 (72.1%)	159 (71.6%)	535 (72.8%)	574 (71.6%)
Alcohol intake (g/day)	6.6 (0, 22.9)	8.6 (0, 30) *	8.6 (0.2, 23.8) ##	5.7 (0, 20)

Table 1. *Cont.*

Characteristics	All Participants	Current Smokers	Former Smokers	Non-Smokers
Hypertension	855 (48.7%)	82 (36.9%) *	395 (53.8%) #	378 (47.2%)
Diabetes mellitus	135 (7.7%)	14 (6.3%)	65 (8.9%)	56 (7%)
HDL-cholesterol (mg/dL)	61.7 (51.1, 75)	58.5 (49, 69.9) ***	61.2 (50, 75) #	62.8 (53, 77.2)
LDL-cholesterol (mg/dL)	122.8 (99.1, 146.5)	124.7 (99.9, 147.4)	119.6 (95.6, 144) ##	126.2 (103, 147.8)
Total cholesterol (mg/dL)	212.4 (185.1, 238.3)	211.9 (184.4, 234.7)	208.9 (181.8, 236.1) ##	215.8 (189.6, 241.9)
Triglycerides (mg/dL)	106 (77.7, 145.6)	109.3 (85.4, 153.5) *	107.7 (77.9, 149.2) #	103 (76.2, 139)
Fasting glucose (mg/dL)	98 (92, 107)	96 (91, 104)	100 (93, 109) ###	97 (92, 105)
HOMA-IR	2.3 (1.5, 3.5)	2.1 (1.4, 3)	2.3 (1.5, 3.6)	2.3 (1.5, 3.4)
HOMA-Beta	97.8 (71.2, 132)	93.1 (68.7, 124.2)	97.1 (68.9, 132.3)	101 (73.9, 132.7)
HbA1c (%)	5.5 (5.3, 5.8)	5.6 (5.3, 5.8) *	5.5 (5.3, 5.8)	5.5 (5.2, 5.8)

Basic characterization of individuals in our cohort. Continuous variables are presented as median (25th, 75th), while categorical variables are expressed as *n* (%). Statistical analyses employed the Kruskal–Wallis Test for continuous variables and the Chi-square test for categorical variables. Significance levels for comparisons between current and non-smokers are denoted as * $p < 0.05$, *** $p < 0.001$. For comparisons between former and non-smokers, significance levels are indicated as # $p < 0.05$, ## $p < 0.01$, ### $p < 0.001$.

3.2. Distribution of Methylation β-Values

The methylation β-values, ranging from 0 to 1, were computed as the ratio of the methylated signal to the sum of the methylated and unmethylated signals. The distribution of methylation β-values are described in Figure 2. The distribution of β-values for total 5mC+5hmC and 5mC methylation were notably similar, with the median values of 0.75 (interquartile range (IQR) = 0.03) and 0.56 (IQR = 0.03), respectively. Both distributions follow an obvious binomial pattern, drastically compressed within the low (0–0.2) and high (0.8–1.0) ranges. However, the values for 5hmC were notably low, with a median value of 0.03 (IQR = 0.02).

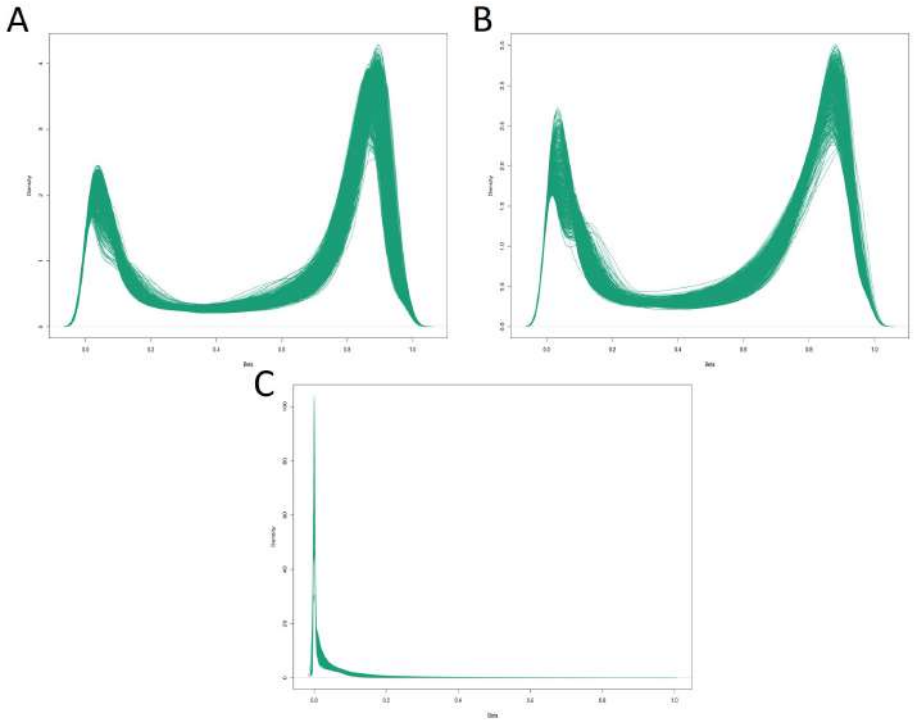


Figure 2. Density plots illustrating the distribution of methylation β-values. The *x*-axis represents the β-values ranging from 0 to 1, while the *y*-axis depicts the corresponding density. (**A**) Density plot for total 5mC+5hmC methylation β-values. (**B**) Density plot for true 5mC methylation β-values. (**C**) Density plot for 5hmC hydroxymethylation β-values.

3.3. Site-Specific Changes in Total 5mC+5hmC Associated with Smoking

The EWAS was conducted to determine epigenome-wide differences in total 5mC+5hmC methylation among current, former and non-smokers. Additionally, we employed bacon correction to mitigate bias and inflation of the test statistic, resulting in a correction of the inflation factor to 1.38 (Supplementary Material S1: Figure S1A,B), which is consistent with many CpG sites being impacted by tobacco smoking. The analysis of 5mC+5hmC methylation data revealed 38,575 DMPs associated with current smoking and 82 DMPs associated with former smoking (FDR-adjusted $p < 0.05$). A summary of the top 10 most significant 5mC+5hmC DMPs associated with both current and former smoking is shown in Table 2, and the complete list of significant 5mC+5hmC DMPs can be found in Supplementary Material S2: Tables S1 and S2.

Table 2. Summary of top 10 most significant 5mC+5hmC DMPs from current and former smokers.

Probe	Delta Beta	*p* Value	FDR	CHR	Gene	MAPINFO	EPIC
Current DMPs	data	data					
cg05575921	−22.72%	2.13×10^{-245}	1.56×10^{-239}	5	*AHRR*	373378	
cg21566642	−16.26%	1.89×10^{-162}	6.94×10^{-157}	2		233284661	
cg01940273	−9.67%	5.22×10^{-147}	1.27×10^{-141}	2		233284934	
cg03636183	−9.88%	5.45×10^{-140}	1.00×10^{-134}	19	*F2RL3*	17000585	
cg21161138	−6.88%	1.91×10^{-111}	2.80×10^{-106}	5	*AHRR*	399360	
cg17739917	−10.21%	4.62×10^{-110}	5.65×10^{-105}	17	*RARA*	38477572	*
cg14391737	−10.12%	5.50×10^{-82}	5.77×10^{-77}	11	*PRSS23*	86513429	*
cg26703534	−4.88%	1.90×10^{-78}	1.75×10^{-73}	5	*AHRR*	377358	
cg17087741	−6.13%	4.22×10^{-77}	3.44×10^{-72}	2		233283010	
cg21911711	−5.65%	1.44×10^{-71}	1.06×10^{-66}	19	*F2RL3*	16998668	*
Former DMPs							
cg14391737	−4.56%	2.23×10^{-40}	1.63×10^{-34}	11	*PRSS23*	86513429	*
cg21566642	−4.62%	1.74×10^{-36}	6.40×10^{-31}	2		233284661	
cg05575921	−4.06%	1.20×10^{-25}	2.95×10^{-20}	5	*AHRR*	373378	
cg06644428	−2.20%	3.45×10^{-23}	6.34×10^{-18}	2		233284112	
cg01940273	−2.24%	1.74×10^{-22}	2.56×10^{-17}	2		233284934	
cg16841366	−2.62%	2.90×10^{-16}	3.56×10^{-11}	2		233286192	*
cg11660018	−1.65%	4.39×10^{-16}	4.61×10^{-11}	11	*PRSS23*	86510915	
cg00475490	−1.53%	1.04×10^{-15}	9.56×10^{-11}	11	*PRSS23*	86517110	*
cg03636183	−1.88%	5.66×10^{-15}	1.35×10^{-9}	19	*F2RL3*	17000585	
cg17739917	−2.20%	1.85×10^{-14}	1.35×10^{-9}	17	*RARA*	38477572	*
cg14391737	−4.56%	2.23×10^{-40}	1.63×10^{-34}	11	*PRSS23*	86513429	*

Probe: Unique identifier from the Illumina CG database; Delta Beta: Mean methylation difference between smokers and non-smokers; FDR: Benjamini–Hochberg corrected *p* value (FDR); CHR: Chromosome; Gene: Target gene name from the UCSC database; MAPINFO: Chromosomal coordinates of the CpG (Build 37); EPIC: * indicates CpG sites that are exclusively present in the Infinium Methylation EPIC BeadChip.

The results supported many previously reported gene loci, including CpG sites annotated to aryl hydrocarbon receptor repressor (*AHRR*), retinoic acid receptor alpha (*RARA*), F2R-like thrombin or trypsin receptor 3 (*F2RL3*) and serine protease 23 (*PRSS23*). Notably, cg05575921 (annotated to *AHRR*), which has consistently emerged as the most significant DMP in previous smoking studies, demonstrated remarkable significance ($p = 1.56 \times 10^{-239}$) and exhibited the largest effect size in our analysis (−22.72% difference in methylation). Out of the 38,575 DMPs, 59.32% (22,884/38,575) were exclusive to EPIC BeadChip and did not present on the previous 450k BeadChip. Moreover, 18.33% (7069/38,575) of the DMPs were novel candidates, not previously reported in the EWAS Catalog (Supplementary Material S2: Table S3). A predominant fraction of DMPs, comprising 77.71% (29,977/38,575), exhibited hypomethylation due to current smoking, with a mean methylation difference of 1.07% (SD = 0.53%). Conversely, 22.29% (8598/38,575) of the DMPs displayed hypermethylation, showing a mean percentage difference of 1.03% (SD = 0.53%). The Manhattan

plot (Figure 3A) and the Volcano plot (Supplementary Material S1: Figure S2A) illustrated EWAS results for 5mC+5hmC methylation related to current smoking.

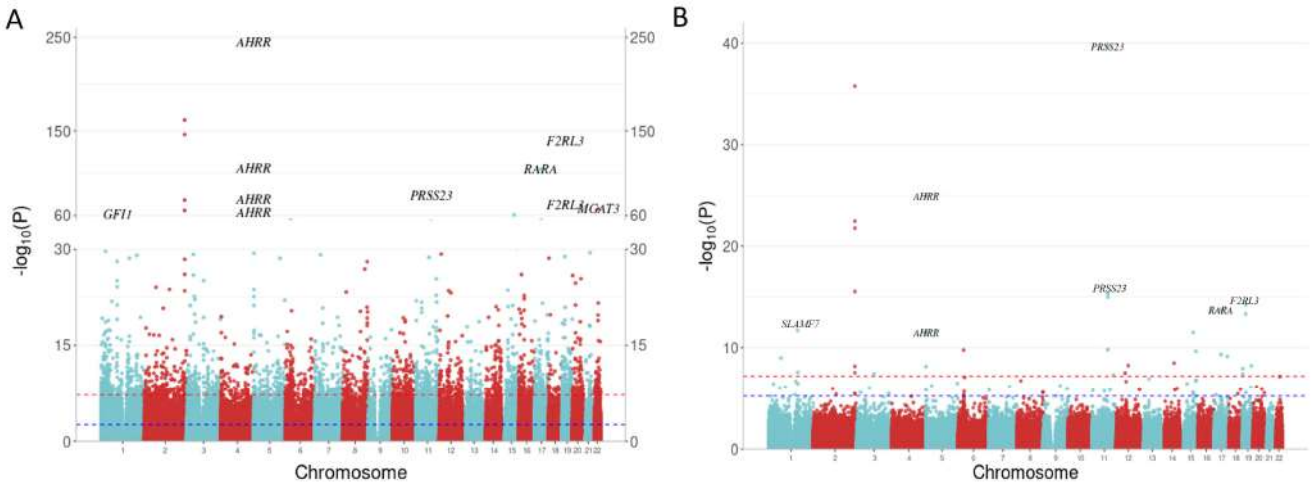


Figure 3. Manhattan plots illustrating smoking EWAS results for 5mC+5hmC methylation. The *x*-axis indicates the chromosome location, and the *y*-axis represents the −log10 (*p*-value). The Bonferroni threshold of 6.81×10^{-8} is marked by a red dashed line, while the Benjamini–Hochberg (FDR) threshold ($p < 0.05$) is indicated by a blue dashed line. The ggbreak package (version 0.1.2) was used to effectively utilize plotting space and handle large *y*-axis values for currents smokers. (**A**) Manhattan plot for current vs. non-smokers; (**B**) Manhattan plot for former vs. non-smokers.

In former smokers, only 82 CpG sites remained differentially methylated, although with reduced effect sizes compared to the observed effects in current smokers. Genomic inflation was not strongly evident (λ = 1.13). All annotated genes associated with former smoking, including *PRSS23*, *AHRR*, *F2RL3* and *RARA*, overlapped with genes associated with current smoking. In contrast to current smokers, the most significant CpG site in former smokers was cg14391737, annotated to *PRSS23* ($p = 1.63 \times 10^{-34}$, effect size: −4.56%), surpassing cg05575921, annotated to *AHRR* ($p = 2.95 \times 10^{-20}$, effect size: −4.06%). Of the 82 identified DMPs, 51.22% (42/82) were exclusive to the EPIC BeadChip and 2.44% (2/82) DMPs were novel candidates (Supplementary Material S2: Table S4). For 90.24% (74/82) of DMPs displaying decreased methylation in response to former smoking, the mean methylation percentage difference was 1.37% (SD = 0.78%). For 9.76% (8/82) of DMPs showing increased methylation in response to former smoking, the mean percentage difference was 1.55% (SD = 0.67%). The Manhattan plot (Figure 3B) and the Volcano plot (Supplementary Material S1: Figure S2B) illustrate EWAS results for 5mC+5hmC methylation related to former smoking.

3.4. Site-Specific True Methylation Changes Associated with Smoking

True DNA methylation (5mC) was measured by oxBS treatment. A total of 33 DMPs were associated with current smoking and 1 5mC DMP was identified between former vs. non-smokers. There was no evidence of inflation (λ = 0.996 for current smokers, λ = 1.009 for former smokers). The count of 5mC DMPs for both current and former smoking was prominently lower than of 5mC+5hmC DMPs. Remarkably, all 33 of the 5mC DMPs, linked to current smoking, were encompassed within the 5mC+5hmC results (Figure 4), and the overall pattern of the 5mC+5hmC and 5mC methylation changes exhibited similarity. For example, the cg05575921, annotated to *AHRR*, consistently retained its position as the most strongly associated with current smoking ($p = 1.27 \times 10^{-77}$) and showed a slightly stronger effect size difference (−24.01%) in the 5mC methylation dataset. In line with 5mC+5hmC, 72.73% (24/33) of the DMPs exhibited hypomethylation in the 5mC dataset, demonstrating a mean difference in methylation of −7.75% (SD = 4.46%). Additionally, 27.27% (9/33) of the DMPs displayed hypermethylation with a mean difference

in methylation of −7.09% (SD = 1.66%). For former smokers, only cg24476099, annotated to megakaryoblastic leukemia 1 (*MKL1*), reached statistical significance with an effect size of −4.34%, and it is specific to the EPIC BeadChip. The most significant 5mC DMPs are shown in Table 3, and the complete list can be found in Supplementary Material S2: Tables S5 and S6. The Manhattan plot (Figure 5A,B) and Volcano plot (Supplementary Material S1: Figure S4A,B) illustrate EWAS results for 5mC methylation related to current and former smoking.

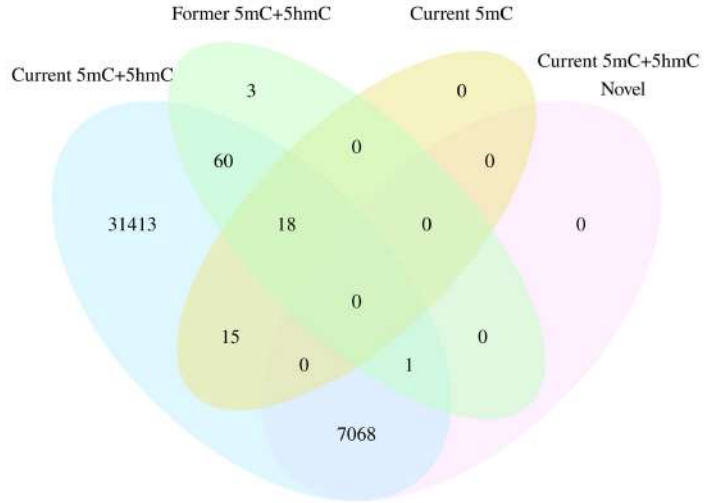


Figure 4. Venn plot illustrating the overlap number of DMPs in different methylation dataset. The blue and cyan colours represent the number of significant DMPs, related with current and former smoking respectively, in the context of 5mC+5hmC methylation. The yellow colour represents the number of significant DMPs related with current smoking in the context of 5mC methylation. The pink colour represents the number of novel DMPs related with current in the context of 5mC+5hmC methylation.

Table 3. Summary of significant true 5mC and 5hmC DMPs from current and former smokers.

Probe	Delta Beta	*p* Value	FDR	CHR	Gene	MAPINFO	EPIC
5mC Current	data	data					
cg05575921	−24.01%	1.68×10^{-77}	1.27×10^{-71}	5	*AHRR*	373378	
cg21566642	−14.63%	2.26×10^{-34}	8.58×10^{-29}	2		233284661	
cg01940273	−9.32%	2.02×10^{-26}	5.10×10^{-21}	2		233284934	
cg03636183	−8.41%	7.61×10^{-25}	1.43×10^{-19}	19	*F2RL3*	17000585	
cg14391737	−11.13%	6.90×10^{-17}	1.04×10^{-11}	11	*PRSS23*	86513429	
5mC Former							
cg24476099	−4.34%	3.95×10^{-8}	0.03	22	*MKL1*	40925033	*
5hmC Current	−4.62%	1.74×10^{-36}	6.40×10^{-31}	2		233284661	
cg16972043	4.14%	1.36×10^{-7}	0.103	16	*GPT2*	46932066	*
cg01483713	1.97%	1.89×10^{-6}	0.718	4		6252582	*
cg15297506	1.22%	4.42×10^{-6}	0.784	10	*SH3PXD2A*	105453418	*
cg04131101	3.50%	4.90×10^{-6}	0.784	11		94427846	
cg22377040	1.68%	5.40×10^{-6}	0.784	6	*TRIM31*	30071412	
5hmC Former	−1.53%	1.04×10^{-15}	9.56×10^{-11}	11	*PRSS23*	86517110	*
cg24012880	3.61%	4.45×10^{-7}	0.337	11	*TSPAN18*	44880910	
cg10148425	2.58%	6.77×10^{-6}	0.985	19		184224630	*

Probe: Unique identifier from the Illumina CG database; Delta Beta: Mean methylation difference between smokers and non-smokers; FDR: Benjamini–Hochberg corrected *p* value (FDR); CHR: Chromosome; Gene: Target gene name from the UCSC database; MAPINFO: Chromosomal coordinates of the CpG (Build 37); EPIC: * indicates CpG sites that are exclusively present in the Infinium Methylation EPIC BeadChip.

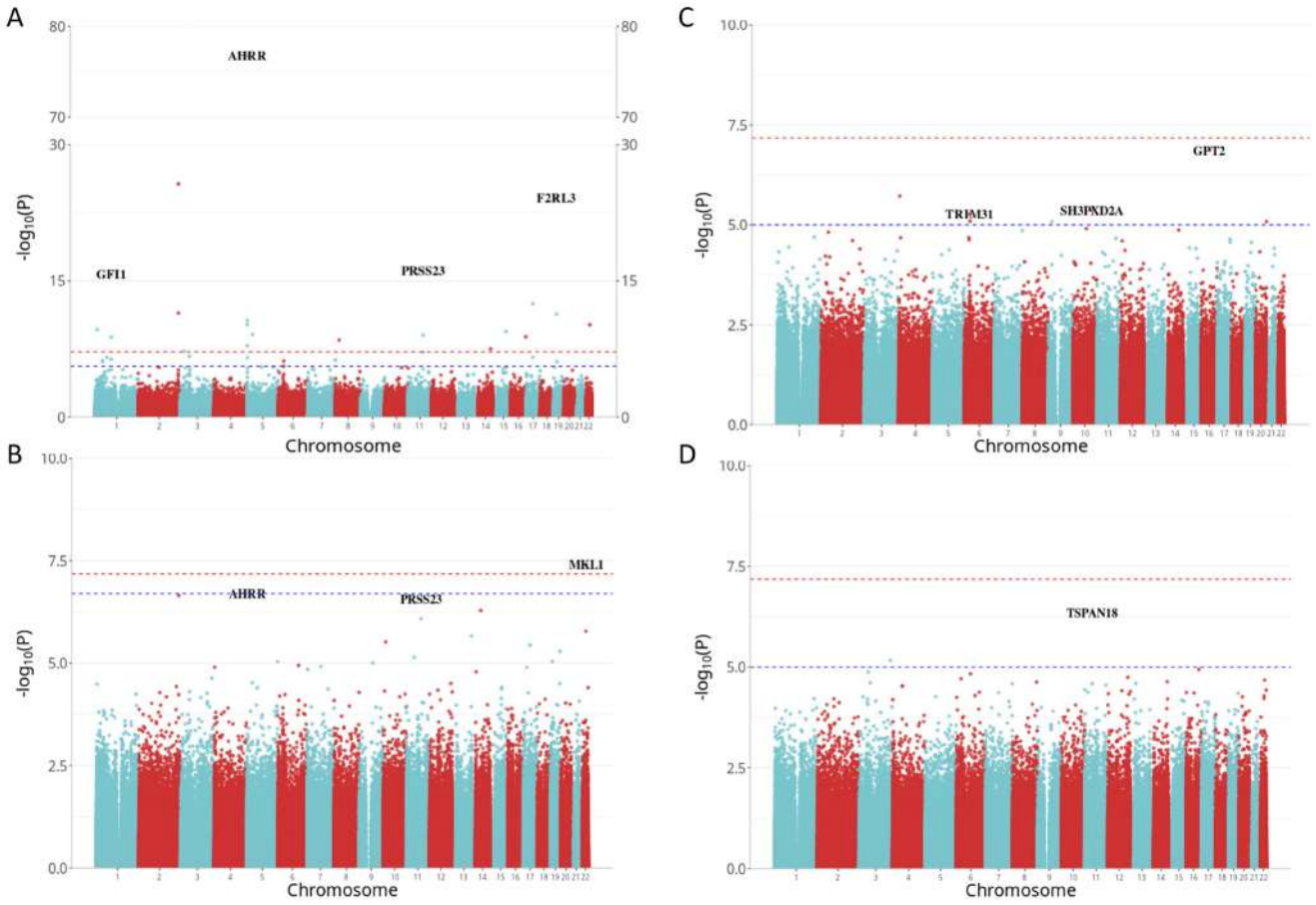


Figure 5. Manhattan plots illustrating smoking EWAS results for both 5mC and 5hmC methylation. The *x*-axis represents the chromosome location, while the *y*-axis represents the −log10(*p* value). The Bonferroni threshold of 6.61×10^{-8} is marked by a red dashed line, while the Benjamini–Hochberg (FDR) threshold ($p < 0.05$) is indicated by a blue dashed line. The ggbreak packagewas used to effectively utilize plotting space and handle large *y*-axis values for currents smokers. (**A**) Manhattan plot for current vs. non-smokers in 5mC dataset; (**B**) Manhattan plot for former vs. non-smokers in 5mC dataset; (**C**) Manhattan plot for current vs. non-smokers in 5hmC dataset; (**D**) Manhattan plot for former vs. non-smokers in 5hmC dataset.

3.5. Site-Specific Hydroxymethylation Changes Associated with Smoking

The total 5mC+5hmC methylation levels were determined using BS treatment, while true DNA methylation (5mC) was measured by oxBS treatment. The quantification of 5hmC involved subtracting 5mC β-values from the combined 5mC+5hmC β-values. 5hmC methylation values were observed at a lower level, so a suggestive threshold of $p < 1 \times 10^{-5}$ was set, revealing eight and two significant 5hmC DMPs between current vs. non-smokers and former vs. non-smokers, respectively. No strong evidence of inflation was detected (λ = 1.132 for current smokers, λ = 1.018 for former smokers). The cg16972043, annotated to the glutamate pyruvate transaminase 2 (*GPT2*) gene, emerged as the most strongly associated ($p = 1.26 \times 10^{-7}$) with current smoking and displayed the largest effect size difference (4.14%) in the 5hmC methylation dataset. Conversely, the cg24012880, annotated to the tetraspanin 18 (*TSPAN18*) gene, demonstrated the strongest association ($p = 4.45 \times 10^{-7}$) with former smoking, displaying an effect size difference of 3.61%. In contrast with methylation changes observed in 5mC+5hmC and 5mC datasets, almost all the top 5hmC DMPs were hypermethylated, demonstrating a mean methylation difference of 2.32% (SD = 1.11%) in current smokers and 0.99% (SD = 0.04%) in former smokers. The most significant 5hmC DMPs are shown in Table 3, and the complete list can be found in Supplementary Material S2: Tables S7 and S8. The Manhattan plot (Figure 5C,D) and

the Volcano plot (Supplementary Material S1: Figure S4C,D) illustrated EWAS results for 5hmC methylation associated with current and former smoking.

3.6. Region-Specific Changes Associated with Smoking

In the total 5mC+5hmC dataset, there were 2023 distinct DMRs linked to current smoking, encompassing 9367 measured CpG sites annotated across 1553 genes. The most prominent DMR uncovered in individuals who currently smoke was situated in a region on chromosome 1, annotated to the growth factor independent 1 transcriptional repressor (*GFI1*) gene, spanning nine CpG sites. The DMR displaying the second strongest association comprised seven CpG sites and was annotated to *AHRR*. A substantial overlap of genes (1542/1553, 99.29%) was observed between the genes identified in the DMP and DMR analyses, which included notable genes like *GFI1*, *AHRR* and HIVEP Zinc Finger 3 (*HIVEP3*). Notably, DMR analyses produced 11 additional genes not identified in DMP analyses, such as Retinoic Acid Receptor Responder 2 (*RARRES2*), Ring Finger Protein 40 (*RNF40*) and Solute Carrier Family 1 Member 5 (*SLC1A5*). During the DMR analysis comparing former smokers and non-smokers, a total of 76 distinct DMRs were identified, containing 390 measured CpG sites and annotated to 61 different genes. Only a minimal overlap of 9.83% (6/61) was observed with previously identified DMPs, specifically Alanyl Aminopeptidase Membrane (*ANPEP*) and *PRSS23*. Additionally, 55 annotated genes such as Proline Rich Transmembrane Protein 1 (*PRRT1*) were exclusively detected in the DMR results. In the true 5mC dataset, there were 14 distinct DMRs linked to current smoking, encompassing 85 measured CpG sites annotated across 12 genes such as *HIVEP3*, *GFI1* and Valyl-TRNA Synthetase 1 (*VARS*). Additionally, there were five distinct DMRs linked to former smoking, encompassing 25 CpG sites annotated across four genes. In the 5hmC dataset, we did not find any DMRs related to current or former smoking. The top 10 most significant DMRs linked to both current and former smoking are presented in Table 4. The complete list of DMRs can be found in Supplementary Material S2: Tables S9–S12; Manhattan plots illustrating DMR results for the 5mC+5hmC and true 5mC methylation datasets related to current and former smoking can be found in Supplementary Materials S1: Figures S3 and S6.

Table 4. Summary of top 10 most significant total 5mC+5hmC DMRs from current and former smokers.

Gene	CHR	Start	End	*p* Value	FDR	Nprobe
Current smokers						
	2	233283010	233286291	5.02×10^{-212}	3.97×10^{-208}	12
GFI1	1	92945668	92947962	5.74×10^{-130}	3.03×10^{-126}	9
AHRR	5	399360	400833	1.16×10^{-63}	2.29×10^{-60}	7
C5orf62	5	150161299	150162069	7.24×10^{-53}	8.20×10^{-50}	3
SLC1A5	19	47287778	47289612	3.52×10^{-51}	3.72×10^{-48}	12
	19	1265877	1266000	1.66×10^{-48}	1.65×10^{-45}	3
	14	106329158	106331863	2.67×10^{-46}	2.49×10^{-43}	19
HIVEP3	1	42384002	42385942	5.62×10^{-46}	4.69×10^{-43}	15
ITGAL	16	30485296	30485967	1.09×10^{-44}	8.68×10^{-42}	7
	6	30719807	30720485	4.34×10^{-42}	2.86×10^{-39}	6
Former smokers						
	2	233283010	233286291	1.53×10^{-61}	2.38×10^{-59}	12
PRRT1	6	32118204	32118458	4.68×10^{-22}	1.81×10^{-20}	13
NBL1	1	19971709	19972778	2.37×10^{-17}	7.37×10^{-16}	9
	19	1265877	1266000	2.98×10^{-16}	7.71×10^{-15}	3
ANPEP	15	90345999	90346095	8.64×10^{-16}	1.91×10^{-14}	3
	1	161708999	161710014	2.05×10^{-13}	3.17×10^{-12}	3
PRSS23	11	86510915	86511218	8.38×10^{-13}	1.18×10^{-11}	5
PPT2	6	32120955	32121556	1.70×10^{-12}	2.19×10^{-11}	20
VARS	6	31762353	31762902	3.91×10^{-12}	3.56×10^{-11}	15
GNA12	7	2847477	2847576	1.47×10^{-11}	1.26×10^{-10}	3
	2	233283010	233286291	1.53×10^{-61}	2.38×10^{-59}	12

Gene: UCSC gene name; CHR: Chromosome; Start: Start CHR position of this region; End: End CHR position of this region; FDR: Benjamini–Hochberg corrected *p* value; Nprobe: number of CpG probes in this region.

3.7. Gene Enrichment Analysis

The genes associated with DMPs that passed the significant threshold (FDR-adjusted $p < 0.05$) were identified. Exploratory downstream enrichment analyses were performed on those genes using the missMethyl package with the KEGG dataset. In the total 5mC+5hmC methylation dataset, DMPs associated with current smoking exhibited enrichment in 27 pathways, whereas DMPs associated with former smoking showed enrichment in 1 pathway. However, we did not find any significant pathway from the true 5mC and 5hmC datasets. These findings suggest a potential link between cigarette smoking and alterations in various molecular pathways, including mechanisms of cardiovascular diseases and cancers. The top 10 ranked biological pathways based on DMPs related to current and former smoking from total 5mC+5hmC are illustrated in Figure 6. The complete lists of pathways, from the total 5mC+5hmC, true 5mC and 5hmC methylation datasets, can be found in Supplementary Material S2: Tables S13–S18.

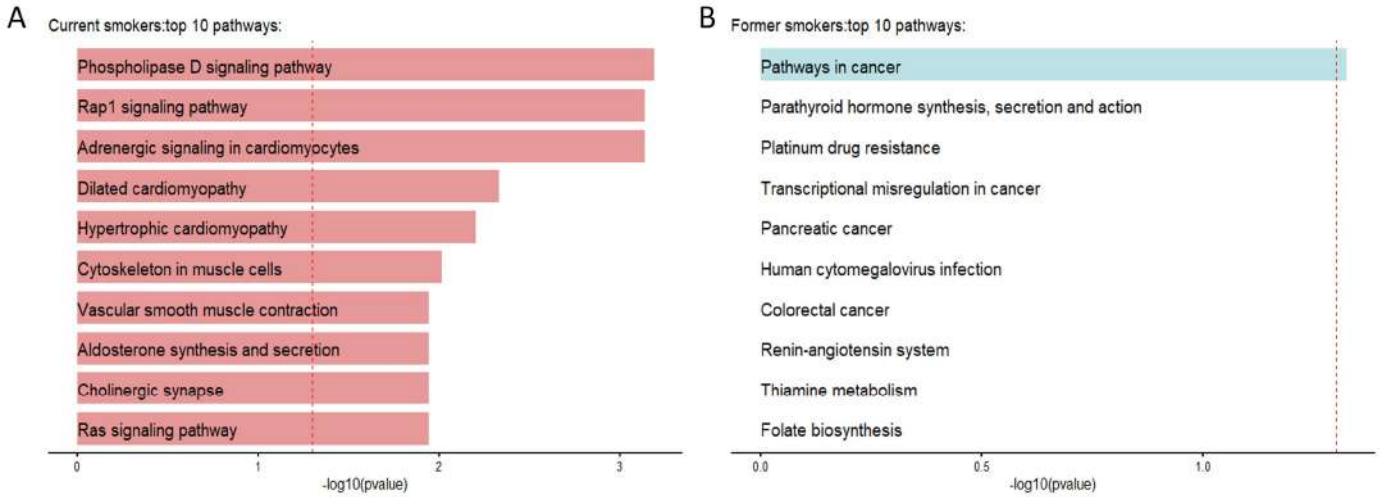


Figure 6. Enrichment analysis results of total 5mC+5hmC methylation. The *x*-axis represents the −log10(*p*-value), and the red dashed line represents the significant threshold (FDR-adjusted $p < 0.05$). (**A**) The top 10 most significant pathways derived from 5mC+5hmC methylation between current and non-smokers. (**B**) The top 10 most significant pathways derived from 5mC+5hmC methylation between former and non-smokers.

4. Discussion

We have investigated different DNA methylation modifications among individuals categorized as current, former and non-smokers. This is, to the best of our knowledge, the first epigenome-wide methylation study of smoking's effects on blood leucocyte samples, analysing true 5mC and 5hmC as distinct DNA methylation modifications, especially in conjunction with the Illumina EPIC BeadChip. Initially, we explored the association between smoking status and total 5mC+5hmC methylation levels, identifying 38,575 and 82 DMPs associated with current and former smoking, many of which are novel candidates. Subsequently, employing tandem BS and oxBS treatment, we differentiated 5hmC from 5mC at the single-nucleotide level. Within this refined analysis, we discovered 33 and 1 DMPs associated with current and former smoking in the 5mC category, respectively. Additionally, eight and two DMPs linked to current and former smoking were identified in the 5hmC category, respectively. We observed a high concordance in the direction of effects and a large overlap in the identified loci between 5mC+5hmC and 5mC groups.

Robust associations have been established between smoking exposure and alterations in blood DNA methylation, supported by the identification of numerous specific loci [11,30]. For example, the most extensive meta-analysis of smoking-associated epigenome-wide DNA methylation was conducted using the 450K array to analyse 15,907 blood-derived DNA samples from individuals across 16 cohorts. A total of 2623 CpG sites, annotated

to 1405 genes, demonstrated associations with current smoking [10]. In this study, we replicated many previously reported sites, including those annotated to *AHRR*, *RARA*, *F2RL3*, *PRSS23* and *GFI1* [31], and identified a substantial number of the novel smoking-associated candidates by using the latest EPIC BeadChip. The *AHRR* gene consistently appeared as the most significantly affected genomic locus in studies investigating the impact of smoking [32,33], a pattern also evident in our cohort. Specifically, 41 DMPs associated with current smoking were annotated to *AHRR* in the 5mC+5hmC dataset, and 11 in the 5mC dataset. All these findings substantiate the robustness and reliability of our study results.

The global initiatives for smoking cessation, coupled with legislative measures, have led to a decline in the number of cigarette smokers and a concomitant rise in the population of former smokers. Decades after cessation, cigarette smoking continues to pose a long-term risk for diseases, and DNA methylation also leaves a persistent signature after smoking exposure [34]. In our analysis, despite the majority of differently methylated CpG sites returning to the methylation levels like non-smokers following smoking cessation, a subset of CpG sites exhibited sustained different methylation even after quitting smoking, albeit with diminished effect sizes in former smokers. The impact of smoking on these specific CpG sites holds the potential to function as robust biomarkers, offering insights into an individual's historical smoking behaviour and reflecting enduring health consequences [35,36].

Clusters of neighbouring probes associated with a phenotype, known as DMRs, may enhance the ability to detect associations between DNA methylation and diseases or phenotypes of interest [37]. For instance, in newborns exposed to maternal gestational diabetes mellitus (GDM) in utero compared to control subjects, only two DMRs were identified without significant DMPs [38]. Therefore, we evaluated methylation differences not only on the individual CpG level but also the regional level using a dimension reduction approach (comb-p). Our analysis revealed 2023 DMRs in current smokers and 76 DMRs in former smokers in the context of 5mC+5hmC. The DMRs associated with smoking exhibited a substantial overlap with the DMP results in both current and former smokers. Notably, CpG sites within these regions were annotated to previously reported genes, including *GFI1*. In addition, a few annotated genes were exclusively identified in the DMRs results; some examples include *RARRES2*, *RNF40* and *SLC1A5*, associated with current smoking, and *PRRT1*, linked to former smoking. Our findings highlight the importance of regional analysis as an additional approach to validate known or identify novel smoking-related genes. Cigarette smoking is linked to increased cancer incidence and poorer cancer-related clinical outcomes. The results of the enrichment analyses also suggest that the discerned smoking-related effects on DNA methylation are likely to carry implications for the risk of various pathologies, including cardiovascular diseases and cancers.

In the present study, oxBS conversion allowed the specific measurement of nucleotide-level 5mC, which holds promise as a biomarker for various diseases [39] and accurate measurement of the true 5mC signal is crucial to prevent false positive findings. In our study, all significant 5mC DMPs associated with current smoking were also found in the conventional 5mC+5hmC dataset, such as *AHRR*, *RARA* and *F2RL3*, proving that these CpG sites are strongly related to smoking. Furthermore, we noted a substantial concordance in the direction of effects between 5mC+5hmC and 5mC groups in current smokers, with a majority of loci displaying hypomethylation. For example, *AHRR* hypomethylation, serving as an epigenetic marker of smoking history, was reported to predict the risk of myocardial infarction, particularly in former smokers [33]. The CpG site cg24476099, annotated to *MLK1*, emerged as the sole novel significant 5mC linked to former smoking in this study. It is noteworthy that prior research has identified other CpG sites annotated to *MLK1*, demonstrating associations with smoking, incident COPD and prevalent type 2 diabetes [40].

Different methylation modifications possess distinct properties, including varying affinities to transcription factors. Unlike 5mC, often linked to gene repression, 5hmC can

inhibit the binding to transcriptional repressors and thereby display the repressive impact of 5mC [41,42]. Hence, the differentiation between 5mC and 5hmC is essential to comprehending the underlying molecular alterations associated with smoking. Most tissues contain approximately 4% 5mC, whereas 5hmC content varies and is typically below 1% in various tissue types [43]. The abundance of 5hmC is remarkably higher in adult neurons and during embryogenesis [44]. Previous research has identified 67 5hmC DMPs between healthy smokers and non-smokers using lung bronchoalveolar lavage cells, providing evidence of 5hmC being involved in the effects of smoking. These findings also suggested that smoking-related differences may involve DNA demethylation of 5mC with a 5hmC intermediate, as inferred from the observed contrasting hypomethylated 5mC and hypermethylated 5hmC data [45]. Our study aligns with this interpretation, further supporting the notion that smoking-induced oxidative stress can trigger DNA demethylation through the sequential oxidation procedure. As expected, given its low abundance in blood, the DNA hydroxymethylation signature linked to smoke exposure exhibited a lesser prominence compared to true DNA methylation, even under a less stringent threshold. The CpG sites cg16972043 (annotated to *GPT2*) and cg24012880 (annotated to *TSPAN18*) emerged as the most significant and novel hydroxymethylated CpG sites associated with current and former smoking, respectively. *GPT2* serves as a crucial link between glycolysis and glutaminases and exhibits significant upregulation in aggressive breast cancers [46]. Recent research has unveiled *GPT2*'s role in regulating smoking-induced metabolism and damage in airway epithelial cells through its impact on lipid synthesis [47]. Furthermore, both *GPT2* and *TSPAN18* have been implicated in incident COPD in leukocytes [40], underscoring their relevance in respiratory conditions. The identification of these novel smoking-associated hydroxymethylated CpG sites holds promise for guiding future research endeavours. The present study has several strengths. Our multivariate linear regression model was meticulously adjusted for many potential confounders, including estimated cell fractions. To enhance the precision of our findings, we differentiated between true 5mC and 5hmC signals using the tandem BS and oxBS treatment, effectively minimizing the likelihood of identifying false positives, especially in combination with Infinium Methylation EPIC BeadChip. Additionally, the study's robustness was further fortified by the assessment of DMRs in addition to individual CpG sites. However, our study does have limitations. Passive smoking was not considered, and additional continuous smoking variables like pack years were unavailable, limiting the comprehensive analysis of smoking effects. The absence of a replication cohort emphasizes the need for future studies to validate our findings in independent populations. Additionally, the use of DNA derived from blood may not fully capture tissue-specific variations in methylation patterns; exploring specific tissues could offer more nuanced information on the impact of smoking on both true DNA methylation and hydroxymethylation.

5. Conclusions

Our results confirmed previously reported smoking-associated CpG sites with the Illumina Infinium Methylation EPIC BeadChip, but also revealed many novel smoking-associated signatures. By distinguishing 5mC and 5hmC data from peripheral blood DNA samples, our study identified distinct smoking-associated DNA methylation modifications. Hydroxymethylation was not strongly associated with smoking in peripheral blood DNA samples, but suggestive hydroxymethylated CpG sites might inform future research.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biom14060662/s1, Figure S1: QQ plots for total 5mC+5hmC methylation; Figure S2: Volcano plots of smoking association effect sizes for total 5mC+5hmC methylation; Figure S3: Manhattan plots of DMR results for total 5mC+5hmC methylation, Figure S4: Volcano plots of smoking association effect sizes for 5mC and 5hmC methylation, Figure S5: QQ plots for 5mC and 5hmC methylation; Figure S6: Manhattan plots of DMR results for 5mC methylation; Figure S7: Gene enrichment analysis plots of true 5mC and 5hmC methylation. Tables S1–S2: the significant DMPs related to current and former smoking from total 5mC+5hmC methylation dataset;

Tables S3–S4: the novel DMPs related to current and former smoking from total 5mC+5hmC methylation dataset; Tables S5–S6: the significant DMPs related to current and former smoking from 5mC methylation dataset; Tables S7–S8: the significant DMPs related to current and former smoking from 5hmC methylation dataset. Tables S9–S12: the significant DMRs related to current and former smoking from total 5mC+5hmC and true 5mC methylation datasets; Tables S13–S18: the pathways related to current and former smoking from total 5mC+5hmC, true 5mC and 5hmC methylation datasets.

Author Contributions: L.L. and M.W. contributed to the design of the study. R.W. and L.L. conducted the data processing and analyses. L.L. and M.W. interpreted the data. L.L. wrote the manuscript. A.K., C.G., J.L., A.P. and M.W. contributed to population-based cohorts. P.R.M.-G., A.K., C.G., J.L., A.P. and M.W. provided suggestions and revisions to manuscript drafts. All authors have read and agreed to the published version of the manuscript.

Funding: The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, supported by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Additionally, KORA has received support within the Munich Center of Health Sciences (MC-Health) at Lud-wig-Maximilians-Universität as part of LMUinnovativ. L.-L. was supported by a scholarship under the State Scholarship Fund by the China Scholarship Council (File No. 202106010104).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Bavarian Medical Association (KORA-Fit EC No 17040).

Informed Consent Statement: All research participants provided signed informed consent before participating in any research activities.

Data Availability Statement: Data are contained within the article and Supplementary Files. The KORA data are available upon request from the KORA Project Application Self-Service Tool (https://www.helmholtz-munich.de/en/epi/cohort/kora, accessed on 10 April 2024).

Acknowledgments: We extend our gratitude to all study participants and research staff of the KORA cohort for their invaluable contributions to the data collection and pre-processing.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wei, X.; Guo, K.; Shang, X.; Wang, S.; Yang, C.; Li, J.; Li, Y.; Yang, K.; Li, X.; Zhang, X. Effects of different interventions on smoking cessation in chronic obstructive pulmonary disease patients: A systematic review and network meta-analysis. *Int. J. Nurs. Stud.* **2022**, *136*, 104362. [CrossRef]
2. Kondo, T.; Nakano, Y.; Adachi, S.; Murohara, T. Effects of tobacco smoking on cardiovascular disease. *Circ. J.* **2019**, *83*, 1980–1985. [CrossRef]
3. Thomson, N.C.; Polosa, R.; Sin, D.D. Cigarette smoking and asthma. *J. Allergy Clin. Immunol. Pract.* **2022**, *10*, 2783–2797. [CrossRef]
4. Domingo-Relloso, A.; Joehanes, R.; Rodriguez-Hernandez, Z.; Lahousse, L.; Haack, K.; Fallin, M.D.; Herreros-Martinez, M.; Umans, J.G.; Best, L.G.; Huan, T.; et al. Smoking, blood DNA methylation sites and lung cancer risk. *Environ. Pollut.* **2023**, *334*, 122153. [CrossRef]
5. Skvortsova, K.; Stirzaker, C.; Taberlay, P. The DNA methylation landscape in cancer. *Essays Biochem.* **2019**, *63*, 797–811. [CrossRef]
6. Gallus, S.; Scala, M.; Possenti, I.; Jarach, C.M.; Clancy, L.; Fernandez, E.; Gorini, G.; Carreras, G.; Malevolti, M.C.; Commar, A.; et al. The role of smoking in COVID-19 progression: A comprehensive meta-analysis. *Eur. Respir. Rev.* **2023**, *32*, 220191. [CrossRef]
7. Seo, Y.S.; Park, J.M.; Kim, J.H.; Lee, M.Y. Cigarette smoke-induced reactive oxygen species formation: A concise review. *Antioxidants* **2023**, *12*, 1732. [CrossRef]
8. Caliri, A.W.; Tommasi, S.; Besaratinia, A. Relationships among smoking, oxidative stress, inflammation, macromolecular damage, and cancer. *Mutat. Res. Rev. Mutat. Res.* **2021**, *787*, 108365. [CrossRef]
9. Heikkinen, A.; Bollepalli, S.; Ollikainen, M. The potential of DNA methylation as a biomarker for obesity and smoking. *J. Intern. Med.* **2022**, *292*, 390–408. [CrossRef]
10. Joehanes, R.; Just, A.C.; Marioni, R.E.; Pilling, L.C.; Reynolds, L.M.; Mandaviya, P.R.; Guan, W.; Xu, T.; Elks, C.E.; Aslibekyan, S.; et al. Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* **2016**, *9*, 436–447. [CrossRef]
11. Ambatipudi, S.; Cuenin, C.; Hernandez-Vargas, H.; Ghantous, A.; Le Calvez-Kelm, F.; Kaaks, R.; Barrdahl, M.; Boeing, H.; Aleksandrova, K.; Trichopoulou, A.; et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* **2016**, *8*, 599–618. [CrossRef]
12. Higashi, Y. Smoking cessation and vascular endothelial function. *Hypertens. Res.* **2023**, *46*, 2670–2678. [CrossRef]

13. Fragou, D.; Pakkidi, E.; Aschner, M.; Samanidou, V.; Kovatsi, L. Smoking and DNA methylation: Correlation of methylation with smoking behavior and association with diseases and fetus development following prenatal exposure. *Food Chem. Toxicol.* **2019**, *129*, 312–327. [CrossRef]
14. Yano, N.; Fedulov, A.V. Targeted DNA demethylation: Vectors, effectors and perspectives. *Biomedicines* **2023**, *11*, 1334. [CrossRef]
15. Prasad, R.; Yen, T.J.; Bellacosa, A. Active DNA demethylation-The epigenetic gatekeeper of development, immunity, and cancer. *Adv. Genet.* **2021**, *2*, e10033. [CrossRef]
16. Klungland, A.; Robertson, A.B. Oxidized C5-methyl cytosine bases in DNA: 5-Hydroxymethylcytosine; 5-formylcytosine; and 5-carboxycytosine. *Free Radic. Biol. Med.* **2017**, *107*, 62–68. [CrossRef]
17. Xu, T.; Gao, H. Hydroxymethylation and tumors: Can 5-hydroxymethylation be used as a marker for tumor diagnosis and treatment? *Hum. Genom.* **2020**, *14*, 15. [CrossRef]
18. Kranzhöfer, D.K.; Gilsbach, R.; Grüning, B.A.; Backofen, R.; Nührenberg, T.G.; Hein, L. 5′-Hydroxymethylcytosine precedes loss of CpGmethylation in enhancers and genes undergoing activation in cardiomyocyte maturation. *PLoS ONE* **2016**, *11*, e0166575. [CrossRef]
19. Zhou, X.; Zhuang, Z.; Wang, W.; He, L.; Wu, H.; Cao, Y.; Pan, F.; Zhao, J.; Hu, Z.; Sekhar, C.; et al. OGG1 is essential in oxidative stress induced DNA demethylation. *Cell. Signal.* **2016**, *28*, 1163–1171. [CrossRef]
20. Lu, M.J.; Lu, Y. 5-Hydroxymethylcytosine (5hmC) at or near cancer mutation hot spots as potential targets for early cancer detection. *BMC Res. Notes* **2022**, *15*, 143. [CrossRef]
21. Wang, Z.; Du, M.; Yuan, Q.; Guo, Y.; Hutchinson, J.N.; Su, L.; Zheng, Y.; Wang, J.; Mucci, L.A.; Lin, X.; et al. Epigenomic analysis of 5-hydroxymethylcytosine (5hmC) reveals novel DNA methylation markers for lung cancers. *Neoplasia* **2020**, *22*, 154–161. [CrossRef]
22. Nestor, C.; Ruzov, A.; Meehan, R.; Dunican, D. Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. *Biotechniques* **2010**, *48*, 317–319. [CrossRef]
23. Holle, R.; Happich, M.; Löwel, H.; Wichmann, H.E. KORA—A research platform for population based health research. *Gesundheitswesen* **2005**, *67* (Suppl. S1), S19–S25. [CrossRef]
24. Booth, M.J.; Ost, T.W.; Beraldi, D.; Bell, N.M.; Branco, M.R.; Reik, W.; Balasubramanian, S. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat. Protoc.* **2013**, *8*, 1841–1851. [CrossRef]
25. De Borre, M.; Branco, M.R. Oxidative bisulfite sequencing: An experimental and computational protocol. *Methods Mol. Biol.* **2021**, *2198*, 333–348. [CrossRef]
26. Hattori, N.; Liu, Y.Y.; Ushijima, T. DNA methylation analysis. *Methods Mol. Biol.* **2023**, *2691*, 165–183. [CrossRef]
27. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **2012**, *13*, 705–719. [CrossRef]
28. Houseman, E.A.; Accomando, W.P.; Koestler, D.C.; Christensen, B.C.; Marsit, C.J.; Nelson, H.H.; Wiencke, J.K.; Kelsey, K.T. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform.* **2012**, *13*, 86. [CrossRef]
29. Battram, T.; Yousefi, P.; Crawford, G.; Prince, C.; Babaei, M.S.; Sharp, G.; Hatcher, C.; Vega-Salas, M.J.; Khodabakhsh, S.; Whitehurst, O.; et al. The EWAS Catalog: A database of epigenome-wide association studies. *Wellcome Open Res.* **2022**, *7*, 41. [CrossRef]
30. Christiansen, C.; Castillo-Fernandez, J.E.; Domingo-Relloso, A.; Zhao, W.; El-Sayed Moustafa, J.S.; Tsai, P.C.; Maddock, J.; Haack, K.; Cole, S.A.; Kardia, S.L.R.; et al. Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. *Clin. Epigenetics* **2021**, *13*, 36. [CrossRef]
31. Silva, C.P.; Kamens, H.M. Cigarette smoke-induced alterations in blood: A review of research on DNA methylation and gene expression. *Exp. Clin. Psychopharmacol.* **2021**, *29*, 116–135. [CrossRef]
32. Nomura, S.; Morita, H. Dysregulation of *DNA* methylation in the aryl-hydrocarbon receptor repressor (*AHRR*) gene. *Circ. J.* **2022**, *86*, 993–994. [CrossRef]
33. Langsted, A.; Bojesen, S.E.; Stroes, E.S.G.; Nordestgaard, B.G. *AHRR* hypomethylation as an epigenetic marker of smoking history predicts risk of myocardial infarction in former smokers. *Atherosclerosis* **2020**, *312*, 8–15. [CrossRef]
34. Wilson, R.; Wahl, S.; Pfeiffer, L.; Ward-Caviness, C.K.; Kunze, S.; Kretschmer, A.; Reischl, E.; Peters, A.; Gieger, C.; Waldenberger, M. The dynamics of smoking-related disturbed methylation: A two time-point study of methylation change in smokers, non-smokers and former smokers. *BMC Genom.* **2017**, *18*, 805. [CrossRef]
35. Morrow, J.D.; Make, B.; Regan, E.; Han, M.; Hersh, C.P.; Tal-Singer, R.; Quackenbush, J.; Choi, A.M.K.; Silverman, E.K.; DeMeo, D.L. DNA methylation is predictive of mortality in current and former smokers. *Am. J. Respir. Crit. Care Med.* **2020**, *201*, 1099–1109. [CrossRef]
36. Langdon, R.J.; Yousefi, P.; Relton, C.L.; Suderman, M.J. Epigenetic modelling of former, current and never smokers. *Clin. Epigenetics* **2021**, *13*, 206. [CrossRef]
37. Yan, Q.; Forno, E.; Celedón, J.C.; Chen, W. A region-based method for causal mediation analysis of DNA methylation data. *Epigenetics* **2022**, *17*, 286–296. [CrossRef]
38. Howe, C.G.; Cox, B.; Fore, R.; Jungius, J.; Kvist, T.; Lent, S.; Miles, H.E.; Salas, L.A.; Rifas-Shiman, S.; Starling, A.P.; et al. Maternal gestational diabetes mellitus and newborn DNA methylation: Findings from the pregnancy and childhood epigenetics consortium. *Diabetes Care* **2020**, *43*, 98–105. [CrossRef]
39. Zeng, Y.; Chen, T. DNA methylation reprogramming during mammalian development. *Genes* **2019**, *10*, 257. [CrossRef]

40. Robert, F.H.; Daniel, L.M.; Elena, B.; Danni, A.G.; Yi-Peng, C.; Aleksandra, D.C.; Hannah, M.S.; Lee, M.; Nicola, W.; Archie, C.; et al. Blood-based epigenome-wide analyses on the prevalence and incidence of nineteen common disease states. *medRxiv* **2023**. [CrossRef]
41. Szyf, M. The elusive role of 5′-hydroxymethylcytosine. *Epigenomics* **2016**, *8*, 1539–1551. [CrossRef]
42. Taylor, S.E.; Li, Y.H.; Smeriglio, P.; Rath, M.; Wong, W.H.; Bhutani, N. Stable 5-hydroxymethylcytosine (5hmC) acquisition marks gene activation during chondrogenic differentiation. *J. Bone Miner. Res.* **2016**, *31*, 524–534. [CrossRef]
43. Zhang, Z.; Lee, M.K.; Perreard, L.; Kelsey, K.T.; Christensen, B.C.; Salas, L.A. Navigating the hydroxymethylome: Experimental biases and quality control tools for the tandem bisulfite and oxidative bisulfite Illumina microarrays. *Epigenomics* **2022**, *14*, 139–152. [CrossRef]
44. Shi, D.Q.; Ali, I.; Tang, J.; Yang, W.C. New Insights into 5hmC DNA modification: Generation, distribution and function. *Front. Genet.* **2017**, *8*, 100. [CrossRef]
45. Ringh, M.V.; Hagemann-Jensen, M.; Needhamsen, M.; Kular, L.; Breeze, C.E.; Sjöholm, L.K.; Slavec, L.; Kullberg, S.; Wahlström, J.; Grunewald, J.; et al. Tobacco smoking induces changes in true DNA methylation, hydroxymethylation and gene expression in bronchoalveolar lavage cells. *eBioMedicine* **2019**, *46*, 290–304. [CrossRef]
46. Mitra, D.; Vega-Rubin-de-Celis, S.; Royla, N.; Bernhardt, S.; Wilhelm, H.; Tarade, N.; Poschet, G.; Buettner, M.; Binenbaum, I.; Borgoni, S.; et al. Abrogating *GPT2* in triple-negative breast cancer inhibits tumor growth and promotes autophagy. *Int. J. Cancer* **2021**, *148*, 1993–2009. [CrossRef]
47. Yan, F.; Zhang, L.; Duan, L.; Li, L.; Liu, X.; Liu, Y.; Qiao, T.; Zeng, Y.; Fang, H.; Wu, D.; et al. Roles of glutamic pyruvate transaminase 2 in reprogramming of airway epithelial lipidomic and metabolomic profiles after smoking. *Clin. Transl. Med.* **2024**, *14*, e1679. [CrossRef]

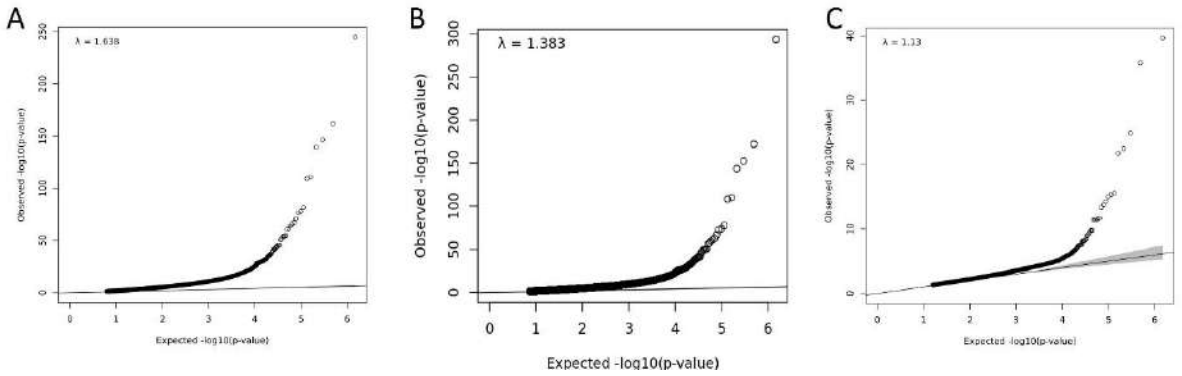


Figure S1. QQ plots for total 5mC+5hmC methylation. The x-axis represents the expected –log10 (P-value) and the y-axis represents the observed –log10 (P-value). **(A)** QQ plot for current vs non-smokers; **(B)** QQ plot for current vs non-smokers after bacon correction; **(C)** QQ plot for former vs non-smokers.

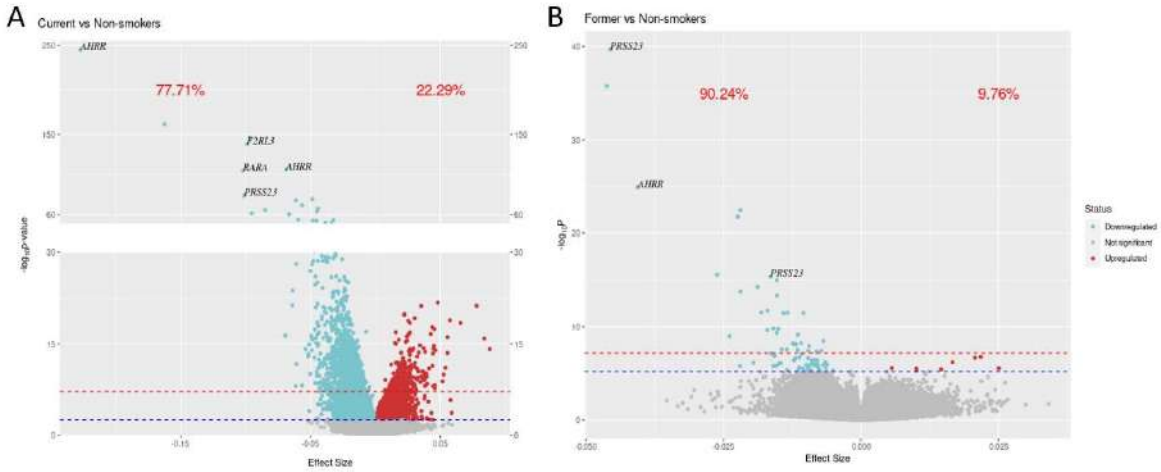


Figure S2. Volcano plots of smoking association effect sizes for total 5mC+5hmC methylation. The x-axis represents the effect size (the methylation value difference between groups), and the y-axis represents the –log10 (*P*-value). The Bonferroni threshold of 6.81×10^{-8} is marked by a red dashed line, while the Benjamini-Hochberg (FDR) threshold ($P<0.05$) is indicated by a blue dashed line. The ggbreak package was used to effectively utilize plotting space and handle large y-axis for currents smokers. **(A)** Volcano plot for current vs non-smokers; **(B)** Volcano plot for former vs non-smokers.

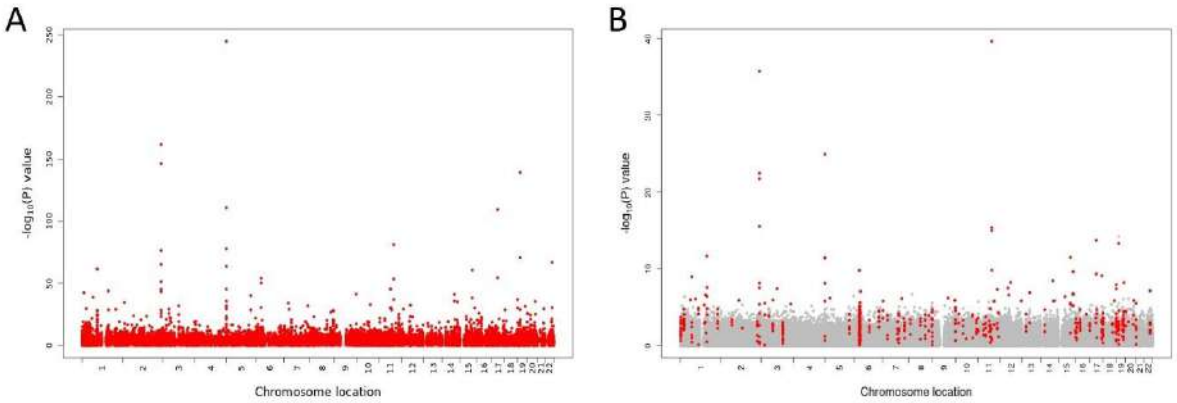


Figure S3. Manhattan plots of DMR results for total 5mC+5hmC methylation. The x-axis represents the chromosome location, and the y-axis represents the –log10 (*P*-value). **(A)** Manhattan plot for current vs non-smokers; **(B)** Manhattan plot for former vs non-smokers.

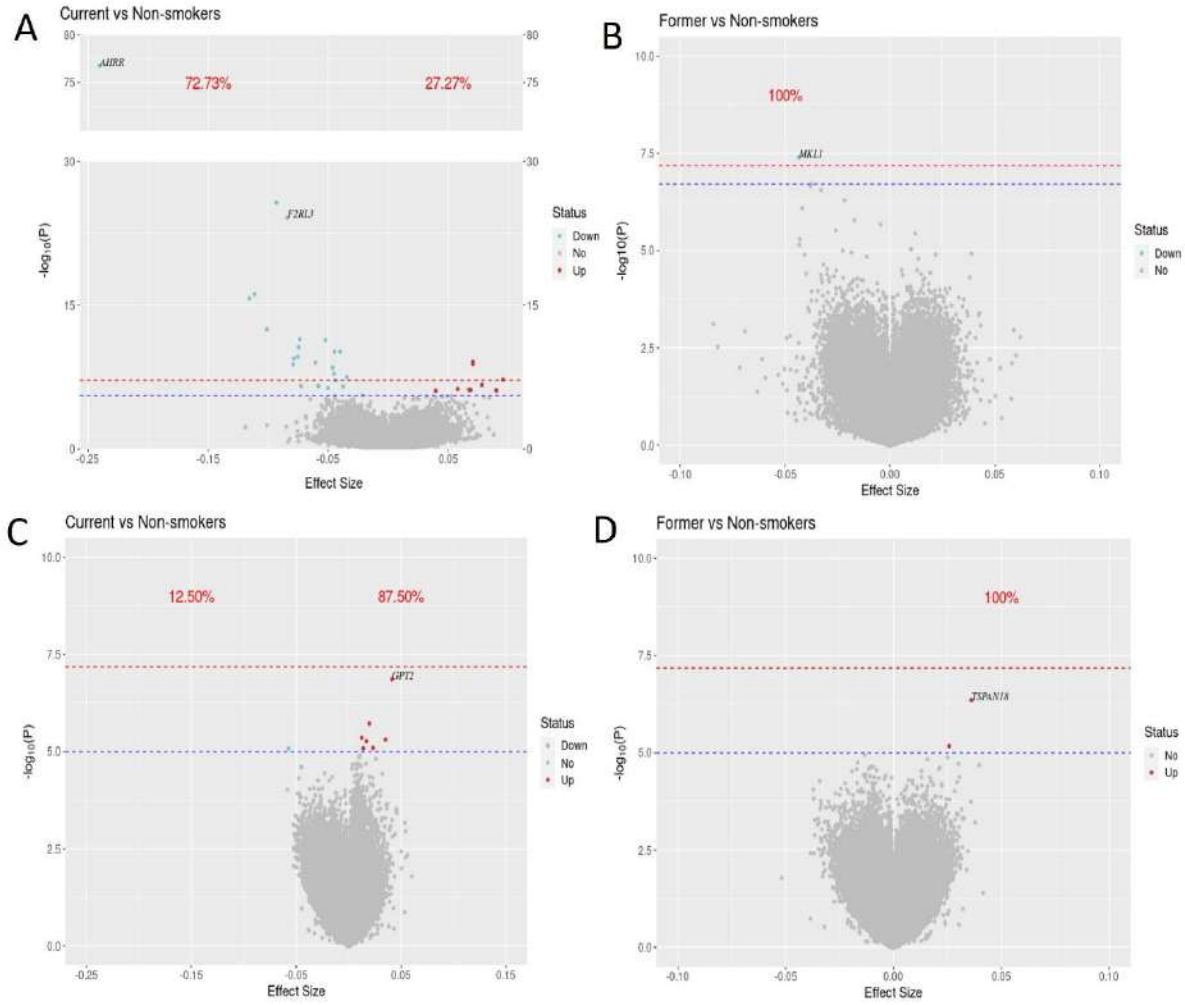


Figure S4. Volcano plots of smoking association effect sizes for 5mC and 5hmC methylation. The x-axis represents the effect size (the methylation value difference between groups), and the y-axis represents the –log10 (*P*-value). The Bonferroni threshold of 6.61×10^{-8} is marked by a red dashed line, while the Benjamini-Hochberg (FDR) threshold ($P<0.05$) is indicated by a blue dashed line. The ggbreak package was used to effectively utilize plotting space and handle large y-axis for

currents smokers. **(A)** Volcano plot for current vs non-smokers from 5mC dataset; **(B)** Volcano plot for former vs non-smokers from 5mC dataset; **(C)** Volcano plot for current vs non-smokers from 5hmC dataset; **(D)** Volcano plot for former vs non-smokers from 5hmC dataset.

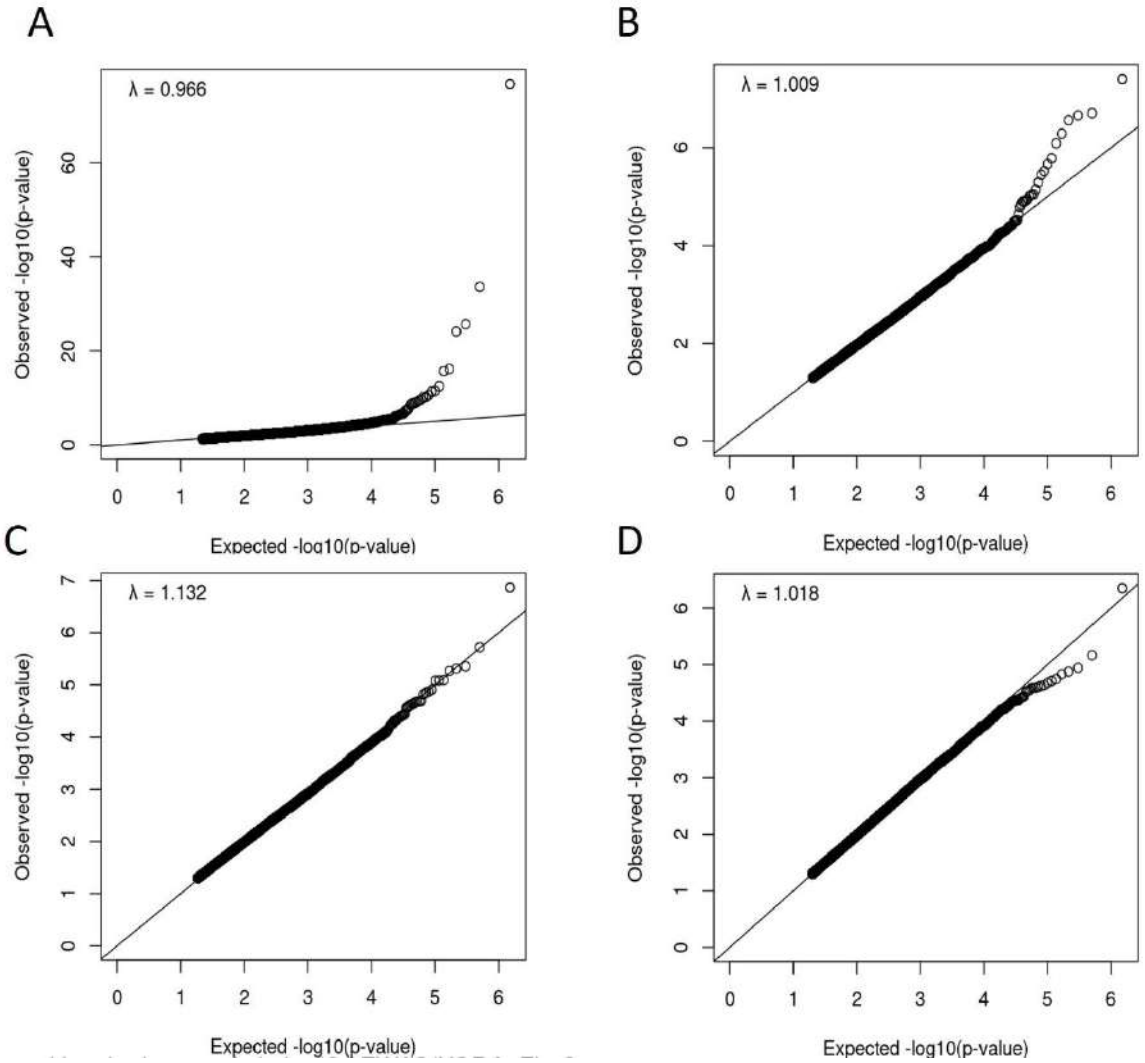


Figure S5. QQs plots of for 5mC and 5hmC methylation. The x-axis represents the expected –log10 (P-value) and the y-axis represents the observed –log10 (*P*-value). **(A)** QQ plot for current vs non-smokers from 5mC dataset; **(B)** QQ plot for former vs non-smokers from 5mC dataset; **(C)** QQ plot for current vs non-smokers from 5hmC dataset; **(D)** QQ plot for former vs non-smokers from 5hmC dataset.

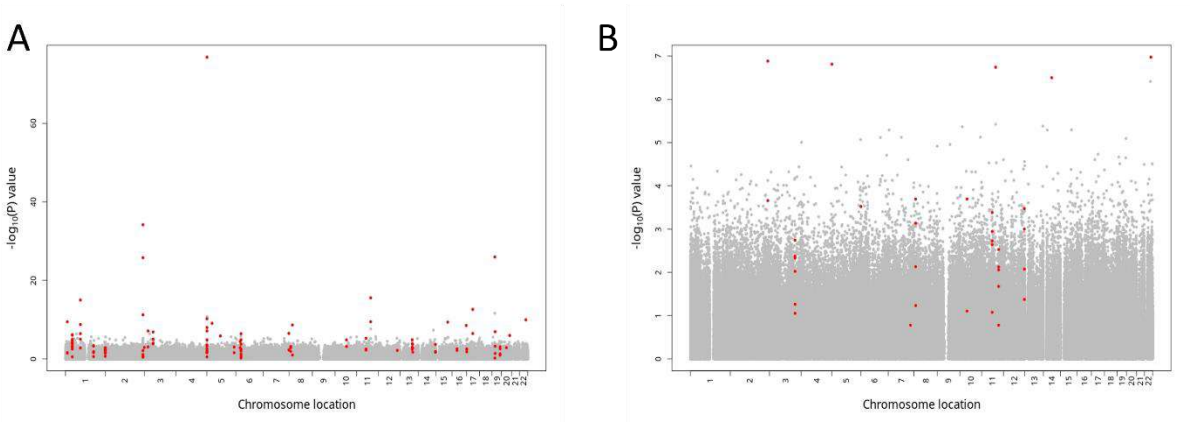


Figure S6. Manhattan plots of DMR results for 5mC methylation. The x-axis represents the chromosome location, and the y-axis represents the –log10 (*P* value). **(A)** Manhattan plot for current vs non-smokers from 5mC dataset; **(B)** Manhattan plot for former vs non-smokers from 5mC dataset.

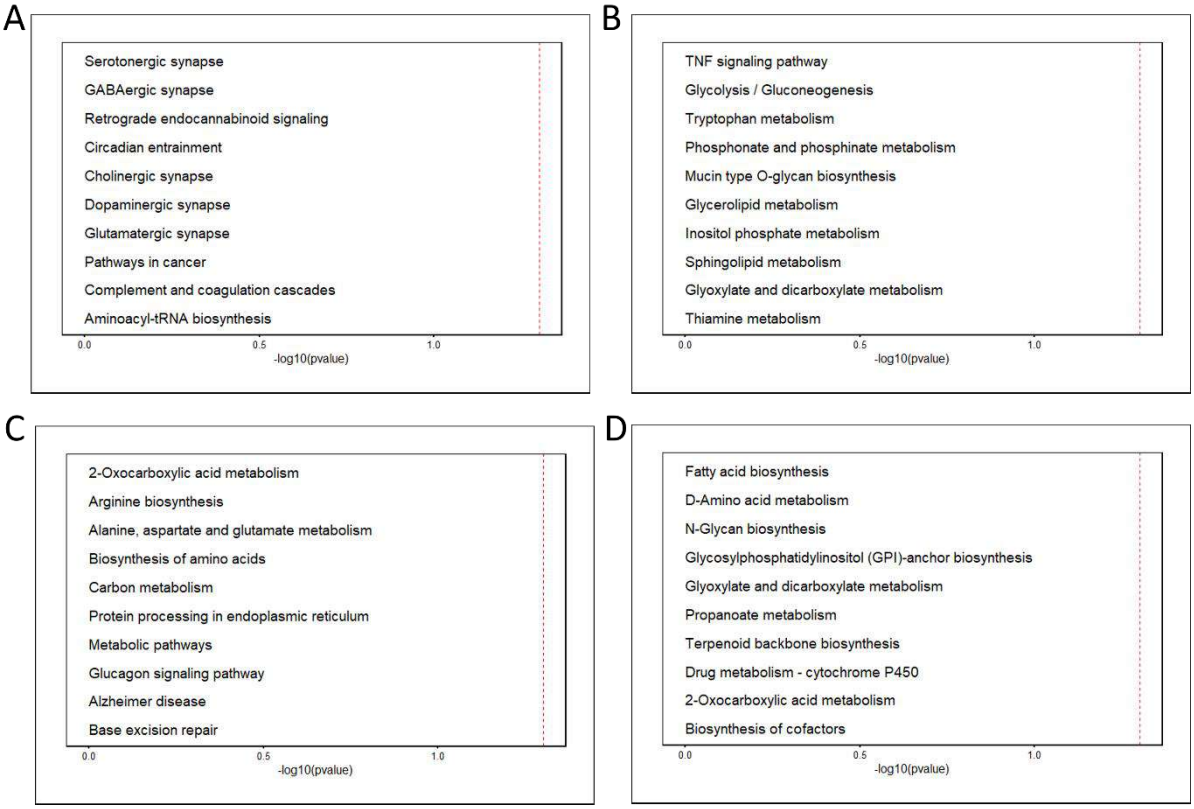


Figure S7. Gene enrichment analysis plots of true 5mC and 5hmC methylation. The x axis represents the -log10(*P*-value), and the red dashed line represents the significant threshold (FDR-adjusted *P*<0.05). (A) illustrate the top 10 pathways derived from true 5mC methylation between current vs non-smokers. (B) illustrate the top 10 pathways derived from true 5mC methylation between former vs non-smokers. (C) illustrate the top 10 pathways derived from 5hmC methylation between current vs non-smokers. (D) illustrate the top 10 pathways derived from 5hmC methylation between former vs non-smokers.

Publication II

Title:	Longitudinal association between DNA methylation and type 2 diabetes: findings from the KORA F4/FF4 study
Authors:	Lai L, Juntilla DL, Del M, Del C Gomez-Alonso M, Grallert H, Thorand B, Farzeen A, Rathmann W, Winkelmann J, Prokisch H, Gieger C, Herder C, Peters A, Waldenberger M
Journal:	Cardiovasc Diabetol
Status:	Published
Volume:	24(1)
Page:	19
Year:	2025
doi:	10.1186/s12933-024-02558-8

Cardiovascular Diabetology

RESEARCH

Open Access

Longitudinal association between DNA methylation and type 2 diabetes: findings from the KORA F4/FF4 study

Liye Lai[1,2,3*], Dave Laurence Juntilla[1,2,3], Monica Del[1], Monica Del C Gomez-Alonso[1,2], Harald Grallert[1,2,4], Barbara Thorand[2,3,4], Aiman Farzeen[1,2,6], Wolfgang Rathmann[7], Juliane Winkelmann[5,6,8,9], Holger Prokisch[5,6], Christian Gieger[1,2,4], Christian Herder[4,10,11], Annette Peters[2,3,4,12†] and Melanie Waldenberger[1,2,12*†]

Abstract

Background Type 2 diabetes (T2D) has been linked to changes in DNA methylation levels, which can, in turn, alter transcriptional activity. However, most studies for epigenome-wide associations between T2D and DNA methylation comes from cross-sectional design. Few large-scale investigations have explored these associations longitudinally over multiple time-points.

Methods In this longitudinal study, we examined data from the Cooperative Health Research in the Region of Augsburg (KORA) F4 and FF4 studies, conducted approximately seven years apart. Leucocyte DNA methylation was assessed using the Illumina EPIC and 450K arrays. Linear mixed-effects models were employed to identify significant associations between methylation sites and diabetes status, as well as with fasting plasma glucose (FPG), hemoglobin A1c (HbA1c), homoeostasis model assessment of beta cell function (HOMA-B), and homoeostasis model assessment of insulin resistance (HOMA-IR). Interaction effects between diabetes status and follow-up time were also examined. Additionally, we explored CpG sites associated with persistent prediabetes or T2D, as well as the progression from normal glucose tolerance (NGT) to prediabetes or T2D. Finally, we assessed the associations between the identified CpG sites and their corresponding gene expression levels.

Results A total of 3,501 observations from 2,556 participants, with methylation measured at least once across two visits, were included in the analyses. We identified 64 sites associated with T2D including 15 novel sites as well as known associations like those with the thioredoxin-interacting protein (*TXNIP*) and ATP-binding cassette sub-family G member 1 (*ABCG1*) genes. Of these, eight CpG sites exhibited different rates of annual methylation change between the NGT and T2D groups, and seven CpG sites were linked to the progression from NGT to prediabetes or T2D, including those annotated to mannosidase alpha class 2a member 2 (*MAN2A2*) and carnitine palmitoyl transferase 1 A (*CPT1A*). Longitudinal analysis revealed significant associations between methylation and FPG at 128 sites, HbA1c at

†Annette Peters and Melanie Waldenberger: Shared last authors.

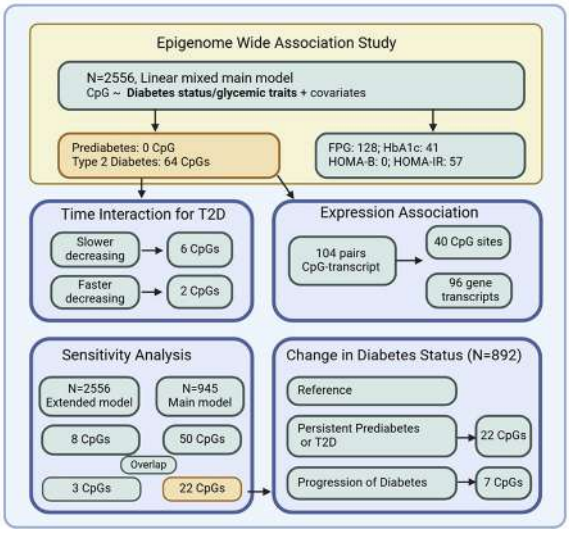
*Correspondence:
Liye Lai
liye.lai@helmholtz-munich.de
Melanie Waldenberger
melanie.waldenberger@helmholtz-munich.de

Full list of author information is available at the end of the article

41 sites, and HOMA-IR at 57 sites. Additionally, we identified 104 CpG-transcript pairs in whole blood, comprising 40 unique CpG sites and 96 unique gene transcripts.

Conclusions Our study identified novel differentially methylated loci linked to T2D as well as to changes in diabetes status through a longitudinal approach. We report CpG sites with different rates of annual methylation change and demonstrate that DNA methylation associated with T2D is linked to following transcriptional differences. These findings provide new insights into the molecular mechanisms of diabetes development.

Graphical abstract



Keywords DNA methylation, Type 2 diabetes, Glycemic traits, Diabetes progression, Gene expression

Background

Type 2 diabetes (T2D) is a major public health concern, characterized by chronic hyperglycemia. The prevalence of T2D is rising rapidly worldwide, projected to affect 783 million adults by 2045 [1]. Individuals with T2D are at risk of developing severe and life-threatening complications, leading to increased medical needs and reduced quality of life. Despite extensive research on T2D pathophysiology, the underlying mechanisms are not yet fully elucidated. Epigenetic modifications, especially DNA methylation—where methyl groups are added to DNA molecules affecting gene expression without altering the DNA sequence—are emerging as crucial links between genetic, environmental, and lifestyle factors in T2D development and progression [2–5]. Identification of novel biomarkers linked to T2D and early glucose disturbances can enhance our understanding of the disease's etiology and improve prevention and prediction strategies [6, 7].

Advances in methylation technology have facilitated the simultaneous measurement of numerous cytosine-phosphate-guanine (CpG) dinucleotide sites, leading to the identification of various CpG sites associated with prevalent T2D and glycemic traits in cross-sectional epigenome-wide association studies (EWAS) [8–11]. Recent comprehensive analyses, including a systematic review of 32 studies, have summarized evidence linking DNA methylation patterns to T2D pathophysiology, utilizing samples from blood, pancreatic islet, adipose tissue, liver, spermatozoa and skeletal muscle [12]. Additionally, a study involving over 18,000 Scottish individuals examined the relationship between blood DNA methylation and the prevalence and incidence of multiple diseases, including T2D [13]. Furthermore, genome-wide DNA methylation changes in early life, particularly among offspring exposed to gestational diabetes, have been proposed as a potential mechanism that increase the risk of obesity, glucose intolerance, and T2D [14–16].

Previous studies have been cross-sectional, limiting insights into temporality. Methylation changes may either be part of the causal pathway to disease or serve as non-causal biomarkers [17, 18]. Considering the fluctuating nature of glucose and insulin metabolism prior to T2D development, it is essential to understand the evolution of methylation patterns in the progression from normal glucose tolerance (NGT) to prediabetes and T2D. For instance, maternal glycemia during pregnancy has been linked to longitudinal variations in blood DNA methylation at the fibronectin type III and spry domain containing 1 like (*FSD1L*) loci from birth to age five [19]. In

addition, a cross-lagged analysis of twin samples in China demonstrated bidirectional associations between DNA methylation and T2D or glycemic traits, with significant paths from T2D influencing subsequent DNA methylation and vice versa [20]. In summary, few studies have examined longitudinal changes in methylation across multiple time points and existing longitudinal research often focuses on specific individuals or ancestries with small sample sizes. In our study, we aimed to investigate the association between DNA methylation and diabetes status, as well as four related traits—fasting plasma glucose (FPG), hemoglobin A1c (HbA1c), homoeostasis model assessment of insulin resistance (HOMA-IR) and homoeostasis model assessment of beta-cell function (HOMA-B)—within a longitudinal, population-based cohort comprising 2,556 individuals, utilizing up to two repeated measurements of DNA methylation as well as glucose- and insulin-related traits.

Illustration of the selection criteria for study participants and CpG sites included in the analysis.

Methods

Study population

This study used data from the Cooperative Health Research in the Region of Augsburg (KORA) F4 (2006–2008) and FF4 (2013–2014) studies, both follow-up studies of the KORA S4 study (1999–2001). Detailed information on the KORA cohort design, measurement, and data collection has been previously described [21]. In total, 3,501 observations from 2,556 participants in KORA F4 (1,696) and FF4 (1,805), with methylation data at least once across two visits, were included in the analysis. Of these participants, 945 participants (36.97%) had methylation patterns measured at both time points. Detailed information about the inclusion of study participants can be found in Additional file 1: Text S1.

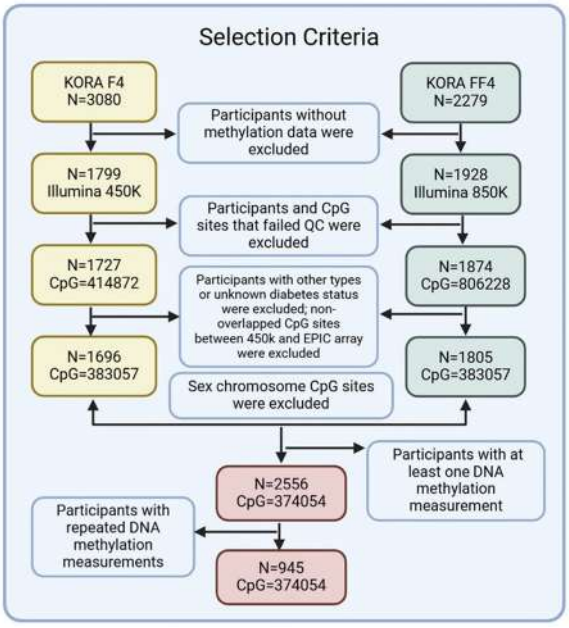


Fig. 1 Illustration of the selection criteria for study participants and CpG sites included in the analysis

Measures of epigenome-wide DNA methylation and gene expression

In the KORA F4 study, genome-wide DNA methylation in whole blood was analysed using the Illumina 450K Infinium Methylation BeadChip (Illumina Inc., San Diego, CA, USA). For the KORA FF4 study, the Infinium MethylationEPIC BeadChip (Illumina Inc., San Diego, CA, USA) was used. DNA methylation was quantified on a scale of 0 to 1, with 1 signifying 100% methylation. We followed the general outline of the CPACOR preprocessing for quality control by using minfi2 package [22]. A total of 374,054 CpG sites were left for the analysis and detailed information about the quality control step and inclusion of CpG sites can be found in Fig. 1 and Additional file 1: Text S2 and Text S3. The proportions of white blood cell types (CD8T, CD4T, natural killer (NK) cells, B lymphocytes, monocytes and granulocytes) were estimated using the Reinius reference-based houseman algorithm implemented in the minfi package [23]. The algorithm is based on methylation values obtained from purified cell types in whole blood. These proportions were then utilized as covariates in the model to mitigate cell type confounding. The KORA F4 and FF4 datasets each included 470 and 448 non-negative control probes from the methylation arrays, respectively, with 430 probes overlapping. To address technical effects during the experiment, we conducted principal component analysis (PCA) on the overlapping probes. The resulting principal components (PCs) are believed to capture technical variability, and the first five control probe PCs, which accounted for 70% of the variance, were included as covariates in the model to eliminate technical biases. The generation and processing of the RNA-seq data of KORA FF4 are described in Additional file 1: Text S4. After quality control, the RNA-seq data were available for 1,543 individuals, with 10,671 gene counts retained for subsequent analysis.

Measures of diabetes status

Previously known T2D was identified by self-report, validated by the responsible physician or medical chart review, or by self-reported current use of glucose-lowering medication. After an overnight fast of at least eight hours, participants without known diabetes underwent a standard 75 g oral glucose tolerance test (OGTT). NGT, prediabetes and newly diagnosed T2D were defined according to the 1999/2006 World health organization (WHO) criteria [24]. The specific cutoff values for

the definition of T2D can be found in Additional file 1: Text S5. For this study, individuals with newly diagnosed T2D or previously known T2D were categorized as having T2D. Since this study involves longitudinal data, an individual's diabetes status may change between time points. Abbreviations separated by a dash indicate diabetes status at baseline and at follow-up. For example, " prediabetes-T2D" refers to individuals who had prediabetes at baseline and had T2D at follow-up. FPG, HbA1c, HOMA-IR, and HOMA-B were assessed as described earlier [25].

Statistical analysis

Epigenome wide association studies

We applied linear mixed-effects models with random participant-specific intercepts to examine the associations between DNA methylation (measured beta values ranging from 0 to 1) and diabetes status (NGT vs. prediabetes and T2D). The association between DNA methylation and diabetes status were identified by the epigenome wide association studies, adjusting for follow-up time (0 for baseline and the time difference to follow-up), age at baseline (years), sex (male, female), body mass index (BMI, kg/m^2), smoking status (never, former, current), estimated cell types (monocytes, B Cells, CD4 T cells, CD8 T cells, and NK cells) and technical effects. An interaction term between sex and T2D was incorporated into the EWAS model to assess the differences in methylation levels between male and female individuals. We used the false discovery rate (FDR) (Benjamini–Hochberg method) to account for multiple testing. An association was considered statistically significant at a p_FDR value < 0.05. The same linear mixed effect model was applied to explore the association between DNA methylation and four continuous outcomes (FPG, HbA1c, HOMA-B and HOMA-IR), which were log-transformed to increase the conformity to normal distributions of residuals. Differentially methylated regions (DMRs) are genomic areas characterized by consistently differing DNA methylation levels across multiple adjacent CpG sites. Alongside the single-site position analysis, we utilized the comb-p function from the Enmix package (version 1.38.01) to identify diabetes-related DMRs. These were defined as groups of probes containing three or more positions within 1,000 base pairs of one another, with FDR-adjusted p-values of less than 0.05. To determine whether the identified diabetes-related CpG sites are also associated with other diseases or exhibit methylation changes in tissues beyond whole blood samples, we checked each significant CpG site in the EWAS Catalog [26].

Time interaction analysis

For CpG sites significantly associated with T2D in the main model, we examined their interaction effects between diabetes status and follow-up time. This interaction effect represents the difference in the rate of methylation change per year between individuals with and without T2D.

Sensitivity analysis

We conducted two sensitivity analyses to evaluate the robustness of our findings. First, we expanded our analysis by including additional confounding variables: parental history of diabetes (positive: at least one parent with diabetes; negative: both parents without diabetes; unknown), use of glucose-lowering medication (yes or no), HDL-cholesterol levels, triglyceride levels, and hypertension (yes or no). The detailed criteria used to assess or define these cofounders have been previously explained [27]. Second, we included only participants with repeated measures of both DNA methylation and glucose- and insulin-related traits, allowing for within-person comparisons over time (945 participants with 1,890 observations).

Association between DNA methylation and changing diabetes status

To investigate the association between DNA methylation and changing diabetes status over time, we categorized 945 participants individuals into 3 groups according to the diabetes status both at baseline and at follow-up: (i) 169 individuals who had either prediabetes or T2D at both time-points (prediabetes-prediabetes:67, T2D-T2D:102), (ii) 200 individuals who progressed from NGT to prediabetes or T2D, or from prediabetes to T2D (prediabetes-T2D:57, NGT-T2D:22, NGT-prediabetes:121), and (iii) 523 individuals who had NGT at both time-points (NGT-NGT: 523). We further excluded 53 individuals whose conditions improved over time, including those with T2D at baseline who had prediabetes or NGT at follow-up, and those with prediabetes at baseline who had NGT at follow-up (T2D-prediabetes:6, T2D-NGT:1, prediabetes-NGT:46) and finally 892 individuals left for the analysis. We focused on the previously identified overlapping significant CpG sites from the analysis of all individuals with methylation measured at least once across two visits (N = 2,556), as well as the subset with repeated DNA methylation measurements (N = 945).

Association between DNA methylation and gene expression

To investigate the relationship between the identified T2D-related CpG sites and gene expression, and to improve annotation, we analysed associations with gene expression probes within a 500 kb window surrounding the significant CpG sites. The MatrixEQTL (version 2.3)

package was used to identify significant CpG-transcript associations. Linear models were adjusted for age, sex, measured white blood cell proportions (neutrophils, monocytes, basophils, and eosinophils) and technical variation with FDR correction for multiple testing.

Pathway analysis

To gain insights into potential biological processes relevant to diabetes or glycemic regulation, we performed gene pathway analysis using the GOmeth function from the missMethyl package (version 1.38.0). Pathways with an $p\_FDR < 0.05$ were considered significant association.

Results

Characteristics of the study population

The analysis included 3,501 observations from 2,556 participants in the KORA F4 (1,696) and FF4 (1,805) studies. Table 1 presents the characteristics of all participants, while Additional file 1: Table S1 shows the characteristics of the 945 individuals with methylation measured at both time points. For all participants, the mean age was 61.0 years in F4 and 58.0 years in FF4. Among the 945 participants with repeated methylation measurements, the mean age was 57.0 years in F4 and 64.0 years in FF4. Due to differences in average age between the two cohorts, we included baseline age as a covariate in our linear mixed effects model to control for age-related variability. The mean BMI was 27.5 kg/m^2 in F4 and 27.0 kg/m^2 in FF4. Male participants comprised 48.8% of the F4 cohort and 48.1% of the FF4 cohort. Additionally, 14.5% of participants in F4 and 13.2% in FF4 had T2D, while 22.4% and 27.8%, respectively, had a parental history of diabetes.

Longitudinal association between DNA methylation and diabetes status

An EWAS was conducted to identify differences in DNA methylation among individuals with NGT, prediabetes and T2D using linear mixed effect models with individual-specific random intercepts in a longitudinal study. Among the 374,054 CpG sites examined, none showed a significant association with prediabetes, while 64 sites (annotated to 47 unique genes) exhibited significant associations with T2D, with 21 sites being hypomethylated and 43 sites being hypermethylated compared to individuals with NGT. Diabetes-by-sex interaction analysis revealed no significant differences between men and women. The Miami plot (Fig. 2) illustrates the distribution of CpG sites associated with T2D. Table 2 provides a summary of the 15 most significant CpG sites, while Additional file 2: Table S1 lists all significant CpG sites linked to T2D. Notably, cg19693031, annotated to thioredoxin-interacting protein (*TXNIP*), emerged as the most significant CpG site (p value: 9.51×10^{-27}) and demonstrated the most significant effect size in our analysis (−2.92%). The results confirm 49 previously reported cross-sectionally associated gene loci, including those annotated to *TXNIP*, ATP-binding cassette sub-family G member 1 (*ABCG1*), carnitine palmitoyl transferase 1 A

Table 1 Characteristics of the study population

Characteristics	KORA F4			KORA FF4				
	All $N=1696$	NGT $N=1113$	Prediabetes $N=338$	T2D $N=245$	All $N=1805$	NGT $N=1262$	Prediabetes $N=304$	T2D $N=239$
Age (years)	61 (14)	58 (14)	65 (14)	67 (10)	58 (18)	54.5 (16)	63 (16)	68 (13.5)
Male (%)	828 (48.8%)	499 (44.8%)	184 (54.4%)	145 (59.2%)	868 (48.1%)	554 (43.9%)	172 (56.6%)	142 (59.4%)
BMI (kg/m^2)	27.5 (5.8)	26.2(5.2)	29.3 (5.7)	30.7(6.7)	27.0 (6.2)	26.0 (5.4)	29.2 (5.2)	30.4 (7.2)
Smoking								
Never smoker	710 (41.9%)	460 (41.3%)	156 (46.2%)	94 (38.4%)	746 (41.3%)	522 (41.4%)	118 (38.8%)	106 (44.4%)
Former smoker	737 (43.5%)	462 (41.5%)	156 (46.2%)	119 (48.6%)	766 (42.4%)	517 (41.0%)	138 (45.4%)	111 (46.4%)
Current smoker	247 (14.6%)	189 (17.0%)	26 (7.7%)	32 (13.1%)	293 (16.2%)	223 (17.7%)	48 (15.8%)	22 (9.2%)
Hypertension (%)	772 (45.5%)	377 (33.9%)	198 (58.6%)	197 (80.4%)	646 (35.8%)	317 (25.1%)	159 (52.3%)	170 (71.1%)
Fasting glucose	5.4 (0.9)	5.2 (0.6)	5.8 (0.9)	6.9 (1.9)	5.4 (0.9)	5.2 (0.6)	6.1 (0.8)	7.2 (2.0)
HOMA-IR	2.2 (1.8)	1.9 (1.3)	3.1 (2.5)	5.1 (4.0)	2.1 (1.9)	1.8 (1.4)	3.5 (2.2)	4.8 (4.2)
HOMA-beta	102.0 (65.7)	101.0 (62.7)	110.0 (79.0)	93.5 (97.4)	94.8 (65.5)	93.1 (61.0)	110.0 (87.7)	102. (70.3)
HbA1c	37.0 (6.0)	36.0 (5.0)	38.5 (5.0)	46.0 (12.0)	36.0 (6.0)	34.0 (5.0)	38.0 (5.0)	45.0 (10.8)
HDL-cholesterol	1.4 (0.5)	1.5 (0.5)	1.3 (0.5)	1.2 (0.4)	1.6 (0.7)	1.7 (0.7)	1.5 (0.6)	1.4 (0.5)
Triglycerides	1.3 (0.9)	1.1 (0.8)	1.5 (1.0)	1.7 (1.2)	1.2 (0.8)	1.1 (0.7)	1.5 (1.0)	1.6 (1.2)
Medication	128.0 (7.6%)	0 (0%)	0 (0%)	128 (52.2%)	133 (7.4%)	0 (0%)	0 (0%)	133 (55.6%)
Parental history								
Yes	380 (22.4%)	239 (21.5%)	71 (21.0%)	70 (28.6%)	501 (27.8%)	314 (24.9%)	94 (30.9%)	93 (38.9%)
No	773 (45.6%)	582 (52.3%)	135 (39.9%)	56 (22.9%)	1131 (62.7%)	844 (66.9%)	177 (58.2%)	110 (46.0%)
Unknown	254 (15.0%)	159 (14.3%)	53 (15.7%)	42 (17.1%)	173 (9.6%)	104 (8.2%)	33 (10.9%)	36 (15.1%)

Data are median (IQR) for continuous variables and n (%) for categorical variables. The unit for both fasting glucose and HbA1c is mmol/mol. The unit for both HDL-cholesterol and triglycerides is mmol/l. Medication means the glucose-lowering medication

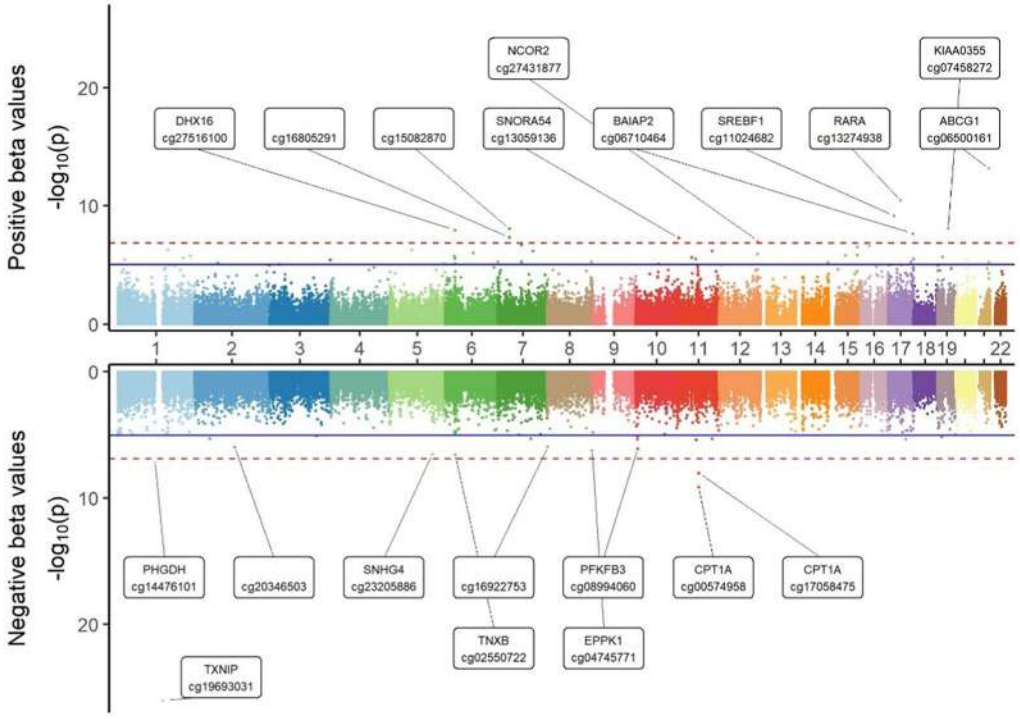


Fig. 2 Miami plot illustrating EWAS results associated with T2D

Table 2 Summary of top 15 significant CpG sites associated with T2D

Probe	Delta beta (%)	*p* value	*p*_FDR	CHR	Gene	MAPINFO	Gene_group
cg19693031	− 2.92	9.51E−27	3.55E−21	1	*TXNIP*	145,441,552	3′UTR
cg06500161	1.22	6.69E−14	1.25E−08	21	*ABCG1*	43,656,587	Body
cg13274938	0.91	3.30E−11	4.12E−06	17	*RARA*	38,493,822	Body
cg11024682	0.95	6.78E−10	5.43E−05	17	*SREBF1*	17,730,094	Body
cg00574958	− 0.73	7.26E−10	5.43E−05	11	*CPT1A*	68,607,622	5′UTR
cg07458272	1.02	7.75E−09	4.45E−04	19	*KIAA0355*	34,744,396	TSS1500
cg15082870	0.91	8.44E−09	4.45E−04	7	#	36,022,841	#
cg17058475	− 1.06	9.53E−09	4.45E−04	11	*CPT1A*	68,607,737	5′UTR
cg27516100	0.83	1.11E−08	4.64E−04	6	*DHX16*	30,624,520	Body
cg06710464	0.94	2.24E−08	8.38E−04	17	*BAIAP2*	79,047,695	Body
cg16805291	1.15	4.70E−08	1.58E−03	7	#	36,022,575	#
cg13059136	1.06	5.08E−08	1.58E−03	11	*SNORA54*	2,986,541	TSS1500
cg14476101	− 1.46	6.48E−08	1.86E−03	1	*PHGDH*	120,255,992	Body
cg27431877	0.60	8.83E−08	2.35E−03	12	*NCOR2*	124,911,924	Body
cg01676795	1.18	2.02E−07	5.04E−03	7	*POR*	75,586,348	Body

Probe: Unique identifier from the Illumina CG database; Delta Beta: Mean methylation difference between T2D and NGT; *p*_FDR: Benjamini-Hochberg corrected *p* value (FDR); CHR: Chromosome; Gene: Target gene name from the UCSC database (# indicates no annotated gene); MAPINFO: Chromosomal coordinates of the CpG (Build 37); Gene_Group: Gene region feature category describing the CpG position from UCSC

(*CPT1A*), and sterol regulatory element-binding transcription factor 1 (*SREBF1*). Importantly, the effect direction of these associations in this longitudinal study was consistent with those of the cross-sectional results for all 49 known sites listed in the EWAS catalogue [26]. Additionally, 15 CpG sites annotated to 10 unique genes were identified as novel associations, including cg02550722 annotated to tenascin XB (*TNXB*), cg04745771 annotated to epiplakin 1 (*EPPK1*), cg23661483 annotated to ilvb acetolactate synthase like (*ILVBL*), cg13947735 annotated to UDP-glcnac: betagal beta-1,3-n-acetylglucosaminyltransferase like 1 (*B3GNTL1*), cg15418499 annotated to interleukin-18 (*IL18*), cg14172849 annotated to X-ray repair cross complementing 3 (*XRCC3*), cg20661985 annotated to open reading frame 3 encoded at human chromosome 20 (*C20orf3*). The DMR analysis identified 44 significant regions associated with 36 unique genes. This analysis confirmed 7 genes previously identified in

the single position analysis and uncovered 29 novel genes linked to T2D, such as valyl-tRNA synthetase (*VARS*), or solute carrier family 1 member 5 (*SLC1A5*). Detailed information related to the DMR analysis is available in Additional file 2: Table S2. The identified T2D-related CpG sites are also linked to other diseases, including metabolic syndrome and cardiovascular diseases, and show methylation changes in specific tissues, such as the liver. For detailed information, please refer to Additional file 2: Table S3.

Miami plot illustrating EWAS results associated with T2D. The x axis indicates the chromosome location, and the y-axis represents the −log10 (*p*-value). The Bonferroni threshold of 1.34×10^{-7} is marked by a red dashed line, while the Benjamini–Hochberg (FDR) threshold (p_FDR < 0.05) is indicated by a blue solid line. The upper side represents the positive estimates, and the lower side represents the negative estimates.

Longitudinal association between DNA methylation and glycemic traits

The same EWAS model was employed to evaluate the longitudinal association between DNA methylation and four glycemic traits: FPG, HbA1c, HOMA-B, and HOMA-IR. Out of the 374,054 CpG sites examined, 128 were associated with FPG, 41 with HbA1c, none with HOMA-B, and 57 with HOMA-IR. Notably, two CpG sites, cg19693031 (*TXNIP*) and cg06500161 (*ABCG1*), were associated with FPG, HbA1c, HOMA-IR, and T2D. The glycemic trait analysis identified an additional 161 unique CpG sites distinct from those associated with T2D, bringing the total number of unique CpG sites linked to both T2D and glycemic traits to 225. Volcano plots (Fig. 3) illustrate the direction of association of the significant CpG sites related to glycemic traits. Additional file 2: Tables S4-6 provide detailed information on all significant CpG sites linked to glycemic traits.

Volcano plots illustrating the results for glycemic traits. The x axis indicates the effect size, and the y-axis represents the −log10 (*p*-value). The Bonferroni threshold of $p = 1.34 \times 10^{-7}$ is marked by a red dashed line, while the Benjamini–Hochberg (FDR) threshold (p_FDR < 0.05) is indicated by a blue dashed line. (A) Volcano plot for FPG. (B) Volcano plot for HbA1c. (C) Volcano plot for HOMA-B. (D) Volcano plot for HOMA-IR.

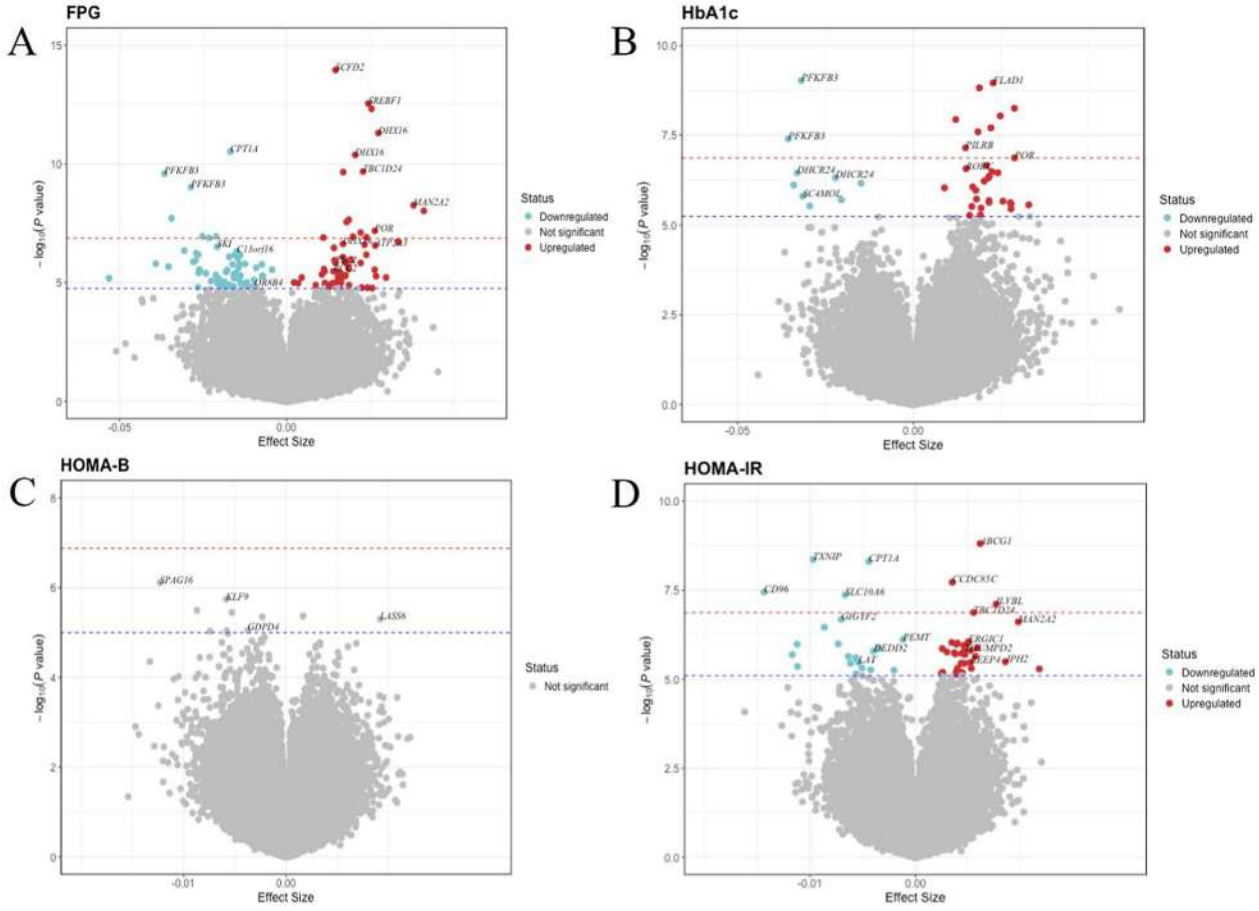


Fig. 3 Volcano plots illustrating the results for glycemic traits

Interaction between diabetes status and follow-up time

We focused on the 64 CpG sites that showed significant associations with T2D in the main model and added an interaction term between T2D and follow-up time to the model. This estimate indicates the difference of the methylation change rates between individuals with T2D and NGT. Eight CpG sites were considered significant (p_FDR value < 0.05). All 8 CpG sites showed a decrease in methylation levels over time. Two CpG sites, cg20346503 and cg19693031 (annotated to *TXNIP*), exhibited a steeper decline in methylation for individuals with T2D compared to those with NGT, with methylation rates of -1.22% and – 1.01% for NGT, versus – 1.31% and – 1.15% for T2D, respectively. In contrast, six CpG sites (cg10442325, cg15418499 annotated to *IL18*, cg20507228, annotated to *MAN2A2*, cg04334723 annotated to calreticulin (*CALR*), cg20661985 and cg00574958 annotated to *CPT1A*) exhibited a slower decrease in methylation change over time for individuals with T2D compared to those with NGT. For instance, the slope for *CPT1A* was – 0.17% for NGT versus – 0.10% for T2D. Furthermore, our analysis demonstrated that there are no interaction effects among male and female participants. Table 3 and Additional file 2: Table S7 provide summary information about the CpG sites which showed interaction effects with follow-up time. Figure 3 and Additional file 1: Fig.S1 illustrate the rate of methylation change over time for the NGT and T2D groups (Fig. 4).

Line plots illustrating the rate of methylation change over time for the NGT and T2D groups. The red and blue line represents the individuals with NGT and T2D, respectively. (A) cg19693031 (*TXNIP*); (B) cg00574958 (*CPT1A*); (C) cg15418499 (*IL18*); (D) cg20507228 (*MAN2A2*).

Sensitivity analysis

In our sensitivity analysis, we further adjusted for medication use, parental history of diabetes, HDL-cholesterol, triglycerides, and hypertension as the extended model. Among the 374,054 CpG sites examined, 8 sites were associated with T2D. Of these, 3 CpG sites remained significant and consistent with our main analysis results. These include cg19693031 annotated to *TXNIP* (effect size: -1.83%, p value: 1.31×10^{-7}), cg06500161 annotated to *ABCG1* (effect size: 0.20%, p value: 1.41×10^{-7}), and cg13274938 annotated to retinoic acid receptor alpha (*RARA*) (effect size: 0.92%, p value: 9.93×10^{-7}).

We also conducted a sensitivity analysis on a subset of 945 individuals with repeated methylation measurements. Among the 374,054 CpG sites examined, 50 CpG sites were associated with T2D and the associations for 22 of these sites, including *TXNIP*, *ABCG1* and *RARA*, remained robust. The correlation coefficients of estimates and p values between the full cohort (N = 2,556) and the repeated methylation measurement subset (N = 945) was strong (r = 0.78) and moderate (r = 0.45), respectively. The Venn diagram (Fig. 5) illustrates the overlap of CpG sites across different datasets, while the Manhattan plots (Additional file 1: Fig. S2) and Additional file 2: Tables S8-9 present results from the extended model and the subset analysis.

Venn diagram illustrating the overlap of CpG sites (with annotated gene names) in the sensitivity analysis. The light cyan colour represents the number of significant CpG sites associated with T2D in the main analysis with all individuals. The greyish-yellow colour represents the number of significant CpG sites associated with T2D in the extended models with all individuals. The light pink colour represents the number of significant CpG sites associated with T2D from individuals with repeated methylation measurements at two time points.

Association between DNA methylation and changing diabetes status over time

The analysis focused on the 22 CpG sites that were associated with T2D in both the full cohort (N = 2,556) and the subset cohort (N = 945). Among these 22 CpG sites, all showed significant associations with persistent

Table 3 Summary of 8 significant CpG sites with different methylation change rates over time for individuals with T2D compared to those with NGT

Probe	Estimate1 (%)	Estimate2 (%)	Estimate3 (%)	*p*value	*p*_FDR	Gene	Gene_group
cg10442325	– 0.86	– 0.71	0.14	3.81E-05	0.002	#	#
cg15418499	– 0.98	– 0.81	0.17	8.19E-04	0.023	*IL18*	5′UTR
cg20507228	– 1.16	– 0.96	0.19	1.10E-03	0.023	*MAN2A2*	Body
cg04334723	– 0.79	– 0.68	0.10	2.15E-03	0.031	*CALR*	Body
cg20346503	– 1.22	– 1.31	– 0.09	2.48E-03	0.031	#	#
cg19693031	– 1.01	– 1.15	– 0.14	3.40E-03	0.031	*TXNIP*	3′UTR
cg20661985	– 1.39	– 1.25	0.13	3.46E-03	0.031	*C20orf3*	Body
cg00574958	– 0.17	– 0.10	0.07	6.02E-03	0.048	*CPT1A*	5′UTR

Probe: Unique identifier from the Illumina CG database; Estimate1: the estimate of follow-up time indicating the methylation change rate per year for individuals with NGT; Estimate2: the methylation change rate per year for individuals with T2D by adding Estimate1 and Estimate3; Esimate3: the estimate of the interaction term between diabetes and follow-up time indicating the difference of methylation change rates between NGT and T2D; p_FDR: Benjamini-Hochberg corrected p value; Gene: Target gene name from the UCSC database. Gene_Group: Gene region feature category describing the CpG position from UCSC

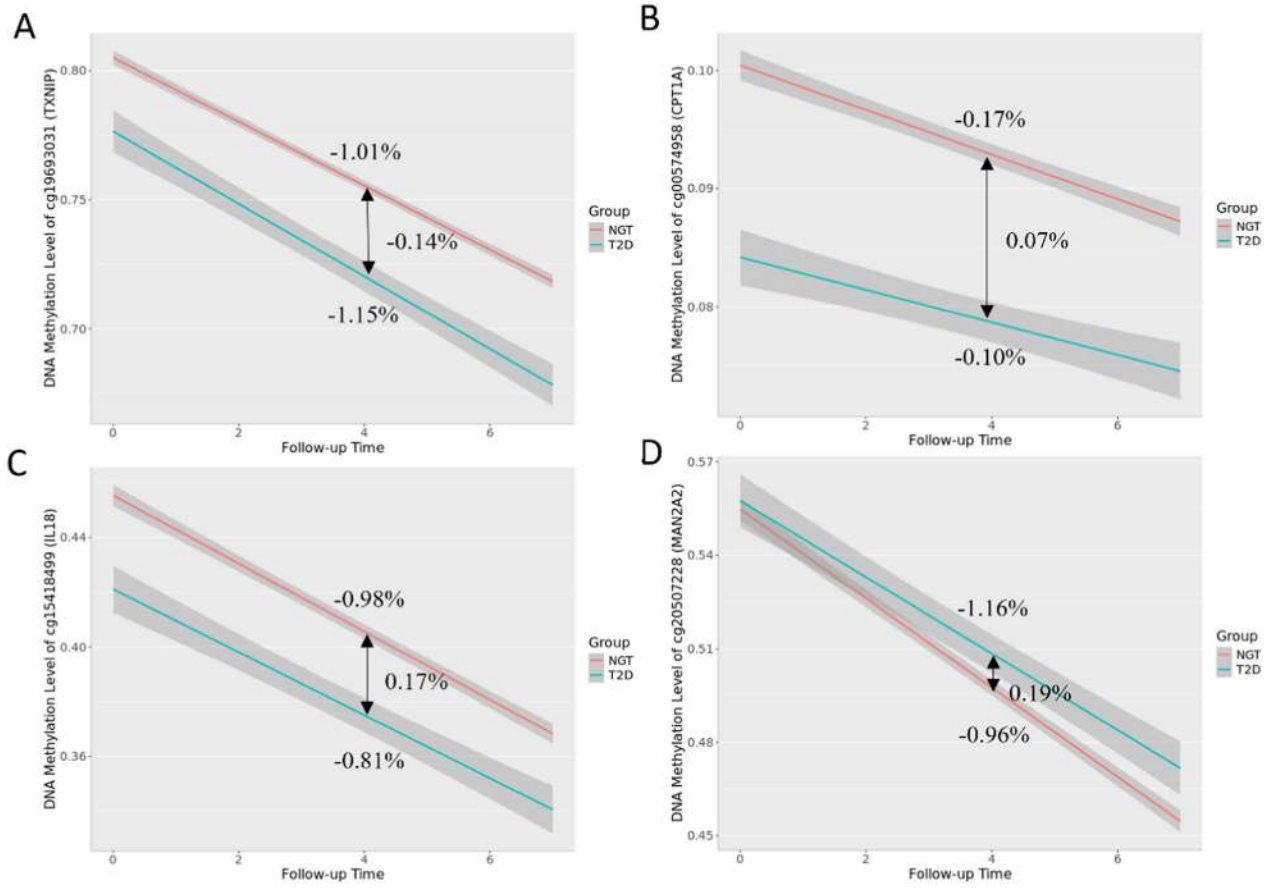


Fig. 4 Line plots illustrating the rate of methylation change over time for the NGT and T2D groups

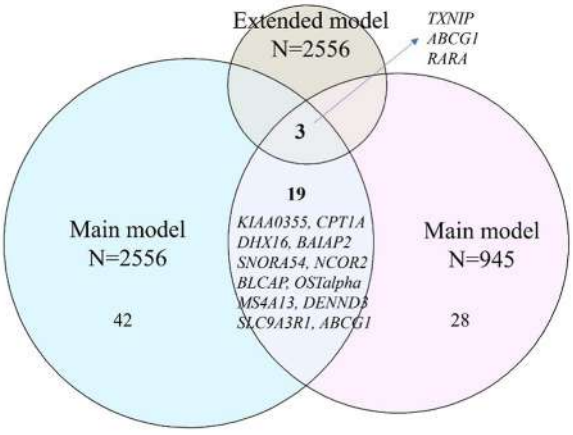


Fig. 5 Venn diagram illustrating the overlap of CpG sites (with annotated gene names) in the sensitivity analysis

prediabetes or T2D at both timepoints, while 7 showed significant associations with progression of diabetes status either from NGT to prediabetes or T2D or from prediabetes to T2D. Notably, these 7 CpG sites, including cg23436042, cg11183227 annotated to *MAN2A2*, cg06500161 annotated to *ABCG1*, cg08788930 annotated to DENN domain-containing protein 3 (*DENND3*), cg11311053 annotated to nuclear receptor corepressor 2 (*NCOR2*), cg06710464 annotated to BAR/IMD domain containing adaptor protein 2 (*BAIAP2*), and cg17058475 annotated to *CPT1A*, demonstrated associations with both persistent and progressed diabetes status. Volcano plots (Fig. 6) illustrate the direction of associations of these significant CpG sites, while the Venn plot (Additional file 1: Fig.S3) shows the overlap of CpG sites across different groups. Additional file 2: Tables S10-11 provide summaries of the significant CpG sites linked to persistent and progressed diabetes status, respectively.

Volcano plots illustrating the association between DNA methylation and changing diabetes status over time. The x axis indicates the effect size, and the y-axis represents the $-\log10$ (p-value). The Bonferroni threshold of 2.27×10^{-3} is marked by a red dashed line, while the Benjamini–Hochberg (FDR) threshold ($p\_FDR < 0.05$) is indicated by a blue dashed line. (A) Volcano plot for the persistent prediabetes or T2D. (B) Volcano plot for the progression of diabetes.

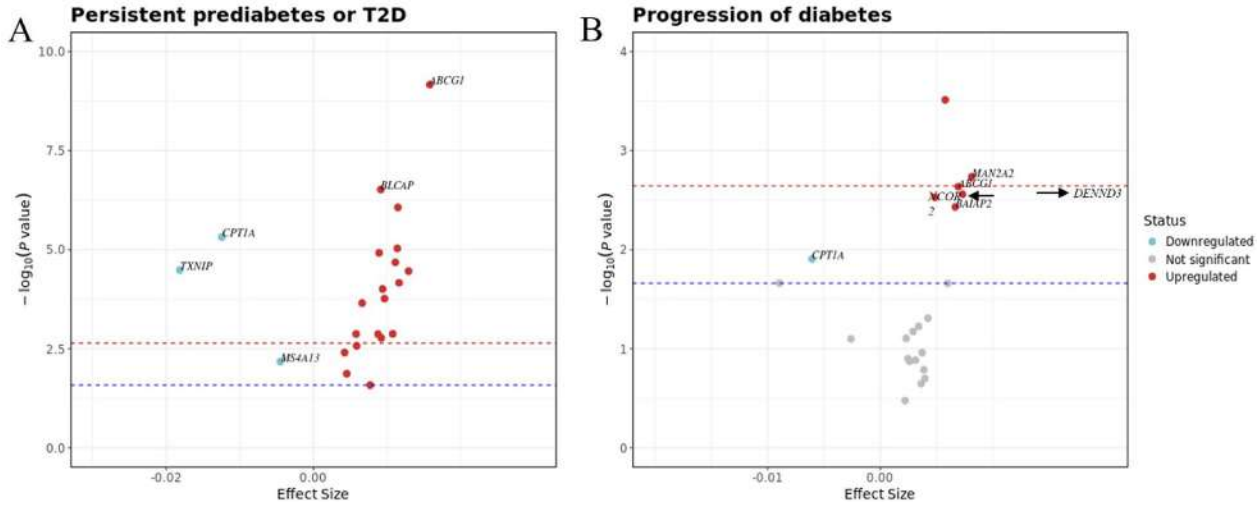


Fig. 6 Volcano plots illustrating the association between DNA methylation and changing diabetes status over time

Table 4 Top 10 associated CpG-transcript pairs

CpG	Gene	*p* value	FDR	Beta
cg06500161	*ABCG1*	1.17E−46	5.80E−44	−4.76
cg06710464	*BAIAP2*	2.84E−41	7.04E−39	−3.30
cg24704287	*LPHN1*	2.11E−29	3.49E−27	1.02
cg27243685	*ABCG1*	3.97E−28	4.91E−26	−4.76
cg11024682	*SREBF1*	4.32E−26	4.28E−24	−2.31
cg01676795	*POR*	3.23E−22	2.67E−20	−1.18
cg06710464	*BAIAP2-AS1*	4.83E−22	3.41E−20	−2.12
cg00851028	*TARBP1*	2.69E−15	1.66E−13	−1.36
cg26340740	*MPEG1*	4.90E−12	2.69E−10	−1.21
cg10691109	*COG5*	1.10E−11	5.48E−10	−0.95

Statistically significant associations between metabolic measure-associated CpG sites and expression of cis-transcripts in whole blood (FDR-adjusted significance threshold $p < 0.05$). Gene: transcript ID; beta: coefficient between methylation and gene transcripts.

Association between DNA methylation and gene expression

Focusing on the 64 significant T2D-related CpG sites, we identified 104 CpG-transcript pairs in whole blood, involving 40 unique CpG sites and 96 unique gene transcripts. Of these, 48 pairs showed positive associations with an average effect size of 0.58, while 56 pairs showed negative associations with an average effect size of -1.02. For example, cg06500161 in *ABCG1* and cg06710464 in *BAIAP2* were negatively associated with their corresponding gene transcripts, while cg24704287 in latrophilin 1 (*LPHN1)* was positively associated with its corresponding gene transcript. Table 4 shows the top 10 significant associations; Additional file 2: Tables S12 summarizes the CpG-transcript associations.

Pathway analysis

In the pathway analysis of the 225 CpG sites associated with T2D and glycemic traits, no significant pathways were identified. The list of non-significant pathways is provided in Additional file 1: Fig.S5 and Additional file 2: Table S13.

Discussion

This study employed longitudinal data with repeated measurements to explore the association between DNA methylation and diabetes status, as well as glycemic traits. We analysed 3,501 observations from 2,556 participants using linear mixed-effects models and identified 64 CpG sites associated with T2D. Notably, DNA methylation at 49 of these loci, including *TXNIP*, *ABCG1*, *CPT1A*, and *SREBF1*, exhibited consistent directional associations in our longitudinal analysis compared to previously reported cross-sectional studies [13, 28]. Importantly, our study revealed 15 novel CpG sites within 10 unique genes. Furthermore, we observed a distinct rate of methylation change for 8 CpG sites between the NGT and T2D groups, including those annotated to *IL18*, *MAN2A2*, *CALR*, *C20orf3* and *CPT1A*, which exhibited either faster or slower decreasing trends. Additionally, 7 CpG sites annotated to *MAN2A2*, *ABCG1*, *DENND3*, *NCOR2*, *BAIAP2* and *CPT1A* were linked to changes in diabetes status. Moreover, we identified 104 associations between identified significant T2D-related CpG sites and their corresponding gene expression levels.

The 64 significant sites that differ between individuals with T2D and NGT in our longitudinal study are annotated to 49 unique genomic loci. *TXNIP* (1 site) has consistently emerged as the most significant gene associated with T2D in previous EWAS studies [29] due to its role in regulating pancreatic β-cells production and survival [30] and has arisen as a novel potential therapeutic target in diabetes mellitus and its complications [31]. *RARA* (1 site), the gene encoding retinoic acid receptor alpha, is a well-known gene linked to cigarette smoking [32]. *FoxK2* (1 site), a major target of insulin signalling, plays

a critical role in apoptosis, metabolism, and mitochondrial function [33] and could regulate aerobic glycolysis [34]. Dyslipidaemia and diabetes are closely related, and epigenome-wide approaches have identified differential methylation of genes known to have a key role in lipid metabolism and lipid traits, particularly *CPT1A*, *ABCG1*, *SREBF1* [35–38]. *ABCG1* (2 sites) is crucial for cholesterol efflux [39], and cg06500161 within *ABCG1* has been reported to mediate the association between statins and risk of T2D [40]. *CPT1A* (2 sites) is associated with an increased risk of gestational diabetes mellitus (GDM) [41]. And multi-tissue epigenetic analysis has revealed distinct associations between the *CPT1A* locus and insulin resistance [42]. Risk group stratification based on cg11024682 (*SREBF1*) was reported to be valuable for personalized T2D risk prediction [43, 44]. Our study found that after controlling for lipid levels in extended models, the associations at the *ABCG1* loci remained robust. In contrast, the associations for *CPT1A* and *SREBF1* were not maintained, suggesting that these associations might be driven by alterations in lipid metabolism.

Our study identified 15 novel CpG sites annotated to 10 unique genes, including *TNXB*, *EPPK1*, *ILVBL*, *B3GNTL1*, *IL18*, *XRCC3*, *C20orf3*. Hypomethylation of *TNXB* gene and differential expression of *EPPK1* protein in the placenta has been reported to be associated with GDM [45, 46]. In a mouse model of diabetes, *ILVBL* has been reported to be involved in the formation of increased dimethylglyoxal, which induces oxidative stress and disrupts the blood-brain barrier, potentially leading to neurological complications in diabetes [47]. *B3GNTL1* was identified as part of a trans-omics biomarker for diabetic kidney disease in diabetic patients [48]. *XRCC3*, a DNA repair gene, has been significantly associated with T2D and diabetic nephropathy in a Turkish population [49]. *C20orf3*, an adipocyte plasma membrane-associated protein, was found to be down-regulated in omental adipose tissues from individuals with GDM [50]. Previous studies have shown that blood methylation patterns in adipose tissue change after bariatric surgery, particularly in genes related to immune system, suggesting that blood DNA methylation reflects the inflammatory state of adipose tissue post-surgery [51]. In our study, we also found that the identified T2D-related CpG sites are also showed methylation changes in specific tissues, such as the liver, by comparing them to the EWAS catalog.

Prolonged disturbances in glucose metabolism are often observed before diabetes diagnosis. Diagnostic tools like FPG and HbA1c are critical for identifying diabetes, underscoring the significance of investigating their effects on DNA methylation. A systematic review and meta-analysis revealed that high HOMA-IR values were positively associated with an increase in risk of T2D [52]. Previous studies have explored the association between DNA methylation changes and hyperglycaemia exposure using the longitudinal D.E.S.I.R. cohort over a six-year period but did not find significant results [53]. Notably, in our study, two CpG sites, cg19693031 (*TXNIP*) and cg06500161 (*ABCG1*), were simultaneously associated with FPG, HbA1c, HOMA-IR, and T2D. These findings highlight the link between glycemic parameters, insulin resistance and DNA methylation, suggesting that alterations at specific CpG sites could serve as biomarkers for glycaemic control and diabetes risk prediction.

DNA methylation is the most studied epigenetic regulator related to environmental exposures. Various environmental triggers, including chemical exposures and complex disease conditions, can lead to global or site-specific DNA methylation changes. This regulation allows for immediate environmental adaptations, potentially affecting transcription factor binding and gene expression. Importantly, we observed that the rate of methylation change varied across diabetes groups. Eight CpG sites, annotated to six unique genes—*IL18*, *MAN2A2*, *CALR*, *TXNIP*, *C20orf03*, and *CPT1A*—all showed decreasing methylation values over time. Low blood *TXNIP* DNA methylation has been linked to increased glucose levels and an increased risk of T2D. In our study, a hypomethylated CpG site annotated to *TXNIP* showed a faster rate of methylation decline in individuals with T2D compared to NGT individuals, resulting in a larger methylation difference between groups, potentially leading to a higher *TXNIP* gene expression over time. Conversely, *IL18*, an inflammation-induced cytokine that is secreted by immune cells and adipocytes [54], was identified as one of the novel sites in our research, showed a slower decrease in methylation values in individuals with T2D compared to NGT. Inflammation-driven processes in the innate immune system can lead to apoptosis, tissue fibrosis, and organ dysfunction, contributing to insulin resistance, impaired insulin secretion, and renal failure [55]. The changing methylation signatures at these 7 CpG loci over time confirm their responsiveness to variations of diabetes status and suggesting their potential as therapeutic targets for future interventions.

In our follow-up study, we considered the evolving nature of diabetes status and identified seven methylation sites linked to the progression from NGT to prediabetes and T2D: cg23436042, cg11183227 (*MAN2A2*), cg06500161 (*ABCG1*), cg08788930 (*DENND3*), cg11311053 (*NCOR2*), cg06710464 (*BAIAP2*), and cg17058475 (*CPT1A*). *MAN2A2* (2 sites), involved in carbohydrate formation, was linked to fasting insulin in an integrative cross-omics analysis [56]. *DENND3* is a positive regulator of starvation-induced autophagy [57]. *NCOR2* has been identified as a potential target gene for T2D screening in the context of cell-free DNA (cfDNA)

methylation changes [58]. It has also been recognized as a potential druggable target for T2D based on an interactome-transcriptome analysis of peripheral blood mononuclear cells (PBMC) in a case-control study of Chinese T2D patients and age- and sex-matched healthy people [59]. *BAIAP2*, the tenth significant site in our study (effect size: 0.94%, p value: 2.24×10^{-8}), encodes the insulin-responsive protein of 53kDa (*IRSp53*). In our EWAS analysis, we did not identify any CpG sites linked to prediabetes. However, within the progression analysis involving individuals transitioning from NGT to prediabetes or T2D, we observed that 2 out of 7 CpG sites—*MAN2A1* and *ABCG1*—exhibited suggestive significance or nominal significance to prediabetes. This suggests that prediabetes may indeed influence the progression of diabetes from NGT to prediabetes. Our findings reveal that DNA methylation is associated with the progression of diabetes status and the identified CpG sites could serve as valuable biomarkers for tracking disease evolution and guiding personalized treatments. Further investigation with larger sample sizes may be necessary to better understand the epigenetic changes associated with prediabetes.

DNA methylation is a recognized regulator of gene expression. By integrating gene expression data, we identified 104 associations between 40 CpG sites and 96 unique gene transcripts in whole blood. Notably, among the seven CpG sites liked to the diabetes progression, five showed a negative correlation with gene expression levels, including cg23436042, cg11183227 (*MAN2A2*), cg06500161 (*ABCG1*), cg06710464 (*BAIAP2*), and cg17058475 (*CPT1A*), while cg08788930 (*DENND3*) and cg11311053 (*NCOR2*) did not. For instance, methylation at cg06500161 in the *ABCG1* gene was negatively associated with its expression levels, providing evidence for a potential link between hypomethylation at this site and upregulated gene expression, which may contribute to T2D and related diseases. Although methylation at cg19693031, which is annotated to *TXNIP*, was negatively associated with T2D, our analysis in blood did not identify any associations involving the *TXNIP* gene transcript. Prior research has demonstrated that hyperglycemia-induced overexpression of *TXNIP* can lead to pancreatic β-cell apoptosis, cardiomyopathy, and metabolic disorders [46]. However, the EWAS results indicated no significant association between DNA methylation and HOMA-beta function; likely due to the nature of the blood samples used. *TXNIP* gene expression has been found to be upregulated in skeletal muscle samples from individuals with diabetes and prediabetes [55], supporting our hypothesis. As a metabolically active tissue, blood plays a crucial role in the inflammatory and vascular effects associated with adiposity, thus making it relevant to our investigation. Moreover, the advantages of utilizing blood samples include their accessibility, cost-effectiveness, and potential for early diagnosis and treatment, which enhances their practicality for clinical applications.

Our study has notable strengths. Firstly, we have comprehensive CpG site coverage through EPIC and 450k arrays, in contrast to candidate locus studies which typically utilize pyrosequencing methods. Secondly, we conducted a longitudinal analysis spanning seven years, incorporating both DNA methylation profiles and diabetes status assessed, through OGTT in those without a clinical diabetes diagnosis. Lastly, we employed different statistical models to control for potential confounders, thereby enhancing the robustness and reliability of our findings. Our study also has limitations. We did not account for other types of diabetes such as type 1 diabetes and gestational diabetes, which may exhibit different methylation patterns and disease mechanisms. Furthermore, utilizing DNA derived from blood may not completely reflect tissue-specific variations in methylation patterns. Additionally, the lack of a replication cohort from diverse ancestries, focusing solely on individuals of European ancestry, highlights the necessity for future studies to validate our findings across different populations.

Conclusion

Our study provides new insights into the associations between DNA methylation and T2D through a longitudinal approach involving repeated measurements. We identified novel CpG sites associated with T2D and revealed varying rates of methylation changes at specific loci across different diabetes status groups. Moreover, we underscored the potential of DNA methylation as a biomarker for diabetes progression and demonstrated the relationship between DNA methylation and the gene expression levels.

Abbreviations

ABCG1	ATP-binding cassette sub-family G member 1
BAIAP2	BAR/IMD domain containing adaptor protein 2
B3GNTL1	UDP-GlcNAc Beta-1,3-N-Acetylglucosaminyltransferase like 1
BMI	Body mass index
CALR	Calreticulin
C20orf3	Open reading frame 3 encoded at human chromosome 20
CPT1A	Carnitine palmitoyl transferase 1 A
CpG	Cytosine-phosphate-guanine
cfDNA	cell-free DNA
DENND3	DENN domain-containing protein 3
DMRs	Differentially methylated regions
EPPK1	Epiplakin 1
EWAS	Epigenome-wide association studies
FDR	False discovery rate
FPG	Fasting plasma glucose
FoxK2	Forkhead Box K2
FSD1L	Fibronectin type III and spry domain containing 1 like
GDM	Gestational diabetes mellitus
HbA1c	Hemoglobin A1c
HOMA-B	Homoeostasis model assessment of beta cell function

HOMA-IR	Homoeostasis model assessment of insulin resistance
IFG	Impaired fasting glucose
IGT	Impaired glucose tolerance
IL18	Interleukin 18
ILVBL	Ilvb acetolactate synthase like
IRSp53	Insulin-responsive protein of 53 kDa
KORA	Cooperative Health Research in the Region of Augsburg
LINE-1	Long interspersed nucleotide element-1
MAN2A2	Mannosidase alpha class 2a member 2
NCOR2	Nuclear receptor corepressor 2
NGT	Normal glucose tolerance
OGTT	Oral glucose tolerance test
PCA	Principal component analysis
PCs	Principal components
PBMC	Peripheral blood mononuclear cells
RARA	Retinoic acid receptor alpha
SNPs	Single nucleotide polymorphisms
SREBF1	Sterol regulatory element-binding transcription factor 1
SLC1A5	Solute carrier family 1 member 5
T2D	Type 2 diabetes
TNXB	Tenascin XB
TXNIP	Thioredoxin-interacting protein
VARS	Valyl-tRNA synthetase
WHO	World health organization
XRCC3	X-ray repair cross complementing 3

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12933-024-02558-8.

Additional file 1. Text S1: Selection criteria of individuals in KORA F4 and FF4; **Text S2**: CPACOR Preprocessing Pipeline; **Text S3**: Selection criteria of CpG sites in KORA F4 and FF4; **Text S4**: Quality control for KORA FF4 gene expression data; **Text S5**: WHO criteria of type 2 diabetes. **Fig.S1**: Line plot illustrates the rate of methylation change over time across different groups; **Fig.S2**: Manhattan plots of sensitivity analysis; **Fig.S3**: Venn plot illustrating the overlap of CpG sites from different analysis; **Fig.S4**: The top 10 non-significant pathways associated with T2D and glycemic traits.

Additional file 2. Table S1: Significant CpG sites associated with T2D; **Table S2**: Significant differentially methylated regions associated with T2D; **Table S3**: Related diseases or tissues; **Table S4**: Significant CpG sites associated with FPG; **Table S5**: Significant CpG sites associated with HbA1c; **Table S6**: Significant CpG sites associated with HOMA-IR; **Table S7**: Significant CpG sites associated with interaction between T2D and follow-up time; **Table S8**: Significant CpG sites associated with T2D from extended model; **Table S9**: Significant CpG sites associated with T2D from repeated methylation measurements; **Table S10**: Significant CpG sites associated with persistent diabetes; **Table S11**: Significant CpG sites associated with progression of diabetes status from NGT to diabetes; **Table S12**: Associated CpG-transcripts pairs; **Table S13**: Pathways associated to T2D.

Acknowledgements

We would like to express our gratitude to all study participants and the research staff of the KORA cohort for their invaluable contributions to data collection and preprocessing.

Author contributions

LL and MW contributed to the design of the study. LL conducted the data processing and analyses. LL and MW interpreted the data. LL wrote the manuscript. DLJ, AF and MCGA analysed the gene expression data. HG, BT, WR, JW, HP, CG, CH, AP, and MW contributed to population-based cohorts. HG, BT, WR, JW, HP, CG, CH, AP, and MW provided suggestions and revisions to manuscript drafts. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.
The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, supported by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Additionally, KORA has received support within the Munich Center of Health Sciences (MC-Health) at Ludwig-Maximilians-Universität as part of LMUinnovativ. The German Diabetes Center (DDZ) is funded by the German Federal Ministry of Health (Berlin, Germany) and the Ministry of Culture and Science of the state North Rhine-Westphalia (Düsseldorf, Germany) and receives additional funding from the German Federal Ministry of Education and Research (BMBF) through the German Center for Diabetes Research (DZD e.V.). L.L. was supported by a scholarship under the State Scholarship Fund by the China Scholarship Council (File No. 202106010104). This project is also funded by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de).

Data availability

The dataset(s) supporting the conclusions of this article is (are) included within the article (and its additional files). The KORA data are available upon request from KORA Project Application Self-Service Tool (https://www.helmholtz-munich.de/en/epi/cohort/kora); data requests can be submitted online and are subject to approval by the KORA Board.

Declarations

Ethics approval and consent to participate

Ethical approval for the KORA cohort was granted by the ethics committee of the Bavarian Medical Association and all procedures were conducted in accordance with the principles of the Declaration of Helsinki. All research participants provided signed informed consent before participating in any research activities. The KORA data protection procedures were approved by the responsible data protection officer of the Helmholtz Zentrum München.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany
[2]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany
[3]Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Pettenkofer School of Public Health, Faculty of Medicine, Ludwig Maximilians University, Munich, Germany
[4]German Center for Diabetes Research (DZD), Neuherberg, Germany
[5]Institute of Human Genetics, School of Medicine, Technical University Munich, Munich, Germany
[6]Institute of Neurogenomics, Computational Health Center, Helmholtz Zentrum München, Neuherberg, Germany
[7]Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[8]Cluster for Systems Neurology (SyNergy), Munich, Germany
[9]Chair of Neurogenetics, Technische Universität München, Munich, Germany
[10]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[11]Department of Endocrinology and Diabetology, Medical Faculty, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[12]German Centre for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Munich, Germany

Received: 16 October 2024 / Accepted: 23 December 2024
Published online: 18 January 2025

References

1. Magliano DJ, Boyko EJ. I.D.F. D.A.t.e.s. committee, IDF Diabetes Atlas, in Idf diabetes atlas. 2021, International Diabetes Federation © International Diabetes Federation, Brussels. 2021.
2. Kaimala S, Ansari SA, Emerald BS. DNA methylation in the pathogenesis of type 2 diabetes. Vitam Horm. 2023;122:147–69.
3. Ling C, Rönn T. Epigenetics in human obesity and type 2 diabetes. Cell Metab. 2019;29(5):1028–44.
4. Ahmed SAH, et al. The role of DNA methylation in the pathogenesis of type 2 diabetes mellitus. Clin Epigenetics. 2020;12(1):104.
5. 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):130–8.
6. Willmer T, et al. Blood-based DNA methylation biomarkers for type 2 diabetes: potential for clinical applications. Front Endocrinol (Lausanne). 2018;9:744.
7. Raciti GA et al. DNA methylation and type 2 diabetes: novel biomarkers for risk assessment? Int J Mol Sci, 2021. 22(21).
8. Kriebel J, et al. Association between DNA methylation in whole blood and measures of glucose metabolism: KORA F4 study. PLoS ONE. 2016;11(3):e0152314.
9. Juvinao-Quintero DL, et al. DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. Clin Epigenet. 2021;13(1):40.
10. Fraszczyk E, et al. Epigenome-wide association study of incident type 2 diabetes: a meta-analysis of five prospective European cohorts. Diabetologia. 2022;65(5):763–76.
11. Baca P, et al. DNA methylation and gene expression analysis in adipose tissue to identify new loci associated with T2D development in obesity. Nutr Diabetes. 2022;12(1):50.
12. Nadiger N, et al. DNA methylation and type 2 diabetes: a systematic review. Clin Epigenetics. 2024;16(1):67.
13. Hillary RF, et al. Blood-based epigenome-wide analyses of 19 common disease states: a longitudinal, population-based linked cohort study of 18413 Scottish individuals. PLoS Med. 2023;20(7):e1004247.
14. Howe CG, et al. Maternal gestational diabetes mellitus and newborn DNA methylation: findings from the pregnancy and childhood epigenetics consortium. Diabetes Care. 2020;43(1):98–105.
15. Finer S, et al. Maternal gestational diabetes is associated with genome-wide DNA methylation variation in placenta and cord blood of exposed offspring. Hum Mol Genet. 2015;24(11):3021–9.
16. Tobi EW, et al. Maternal glycemic dysregulation during pregnancy and neonatal blood DNA methylation: meta-analyses of epigenome-wide association studies. Diabetes Care. 2022;45(3):614–23.
17. Elliott HR, et al. Role of DNA methylation in type 2 diabetes etiology: using genotype as a causal anchor. Diabetes. 2017;66(6):1713–22.
18. Juvinao-Quintero DL, et al. Investigating causality in the association between DNA methylation and type 2 diabetes using bidirectional two-sample mendelian randomisation. Diabetologia. 2023;66(7):1247–59.
19. Taschereau A, et al. Maternal glycemia in pregnancy is longitudinally associated with blood DNAm variation at the FSD1L gene from birth to 5 years of age. Clin Epigenetics. 2023;15(1):107.
20. Hong X, et al. Longitudinal association of DNA methylation with type 2 diabetes and glycemic traits: a 5-year cross-lagged twin study. Diabetes. 2022;71(12):2804–17.
21. Holle R, et al. KORA–a research platform for population based health research. Gesundheitswesen. 2005;67(Suppl 1):S19–25.
22. Lehne B, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 2015;16(1):37.
23. Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
24. ElSayed NA, et al. 2. Classification and diagnosis of diabetes: standards of care in diabetes-2023. Diabetes Care. 2023;46(Suppl 1):S19–40.
25. Huth C, et al. Protein markers and risk of type 2 diabetes and prediabetes: a targeted proteomics approach in the KORA F4/FF4 study. Eur J Epidemiol. 2019;34(4):409–22.
26. Battram T, et al. The EWAS catalog: a database of epigenome-wide association studies. Wellcome Open Res. 2022;7:41.
27. Luo H, et al. Associations of plasma proteomics with type 2 diabetes and related traits: results from the longitudinal KORA S4/F4/FF4 study. Diabetologia. 2023;66(9):1655–68.
28. Walaszczyk E, et al. DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA(1c) levels: a systematic review and replication in a case-control sample of the lifelines study. Diabetologia. 2018;61(2):354–68.
29. Wondafrash DZ, et al. Thioredoxin-interacting protein as a novel potential therapeutic target in diabetes mellitus and its underlying complications. Diabetes Metab Syndr Obes. 2020;13:43–51.
30. Kar A, et al. Thioredoxin interacting protein inhibitors in diabetes mellitus: a critical review. Curr Drug Res Rev. 2023;15(3):228–40.
31. Basnet R, et al. Overview on thioredoxin-interacting protein (TXNIP): a potential target for diabetes intervention. Curr Drug Targets. 2022;23(7):761–7.
32. Lai L et al. Smoking-induced DNA hydroxymethylation signature is less pronounced than true DNA methylation: the population-based KORA fit cohort. Biomolecules, 2024. 14(6).
33. Sakaguchi M, et al. FoxK1 and FoxK2 in insulin regulation of cellular and mitochondrial metabolism. Nat Commun. 2019;10(1):1582.
34. Sukonina V, et al. FOXK1 and FOXK2 regulate aerobic glycolysis. Nature. 2019;566(7743):279–83.
35. Jhun MA, et al. A multi-ethnic epigenome-wide association study of leukocyte DNA methylation and blood lipids. Nat Commun. 2021;12(1):3987.
36. Braun KV, et al. The role of DNA methylation in dyslipidaemia: a systematic review. Prog Lipid Res. 2016;64:178–91.
37. Jones AC, et al. Lipid phenotypes and DNA methylation: a review of the literature. Curr Atheroscler Rep. 2021;23(11):71.
38. Mazaheri-Tehrani S, et al. A systematic review and metaanalysis of observational studies on the effects of epigenetic factors on serum triglycerides. Arch Endocrinol Metab. 2022;66(3):407–19.
39. Zeng GG, et al. A new perspective on the current and future development potential of ABCG1. Curr Probl Cardiol. 2024;49(1 Pt C):102161.
40. Ochoa-Rosales C, et al. Epigenetic link between statin therapy and type 2 diabetes. Diabetes Care. 2020;43(4):875–84.
41. Ren Q, et al. Association of CPT1A gene polymorphism with the risk of gestational diabetes mellitus: a case-control study. J Assist Reprod Genet. 2021;38(7):1861–9.
42. Sarnowski C, et al. Multi-tissue epigenetic analysis identifies distinct associations underlying insulin resistance and Alzheimer's disease at CPT1A locus. Clin Epigenetics. 2023;15(1):173.
43. Krause C, et al. Critical evaluation of the DNA-methylation markers ABCG1 and SREBF1 for type 2 diabetes stratification. Epigenomics. 2019;11(8):885–97.
44. Ling C. Epigenetic regulation of insulin action and secretion—role in the pathogenesis of type 2 diabetes. J Intern Med. 2020;288(2):158–67.
45. Wang WJ, et al. Genome-wide placental gene methylations in gestational diabetes mellitus, fetal growth and metabolic health biomarkers in cord blood. Front Endocrinol (Lausanne). 2022;13:875180.
46. Ge L, et al. The new landscape of differentially expression proteins in placenta tissues of gestational diabetes based on iTRAQ proteomics. Placenta. 2023;131:36–48.
47. Rhein S, et al. The reactive pyruvate metabolite dimethylglyoxal mediates neurological consequences of diabetes. Nat Commun. 2024;15(1):5745.
48. Wu IW, et al. Discovering a trans-omics biomarker signature that predisposes high risk diabetic patients to diabetic kidney disease. NPJ Digit Med. 2022;5(1):166.
49. Yesil-Devecioglu T, et al. Role of DNA repair genes XRCC3 and XRCC1 in predisposition to type 2 diabetes mellitus and diabetic nephropathy. Endocrinol Diabetes Nutr. 2019;66(2):90–8.
50. Ma Y, et al. Identification of a novel function of adipocyte plasma membrane-associated protein (APMAP) in gestational diabetes mellitus by proteomic analysis of omental adipose tissue. J Proteome Res. 2016;15(2):628–37.
51. Müller L, et al. Blood methylation pattern reflects epigenetic remodelling in adipose tissue after bariatric surgery. EBioMedicine. 2024;106:105242.
52. González-González JG, et al. HOMA-IR as a predictor of health outcomes in patients with metabolic risk factors: a systematic review and meta-analysis. High Blood Press Cardiovasc Prev. 2022;29(6):547–64.
53. Khamis A, et al. Epigenetic changes associated with hyperglycaemia exposure in the longitudinal D.E.S.I.R. cohort. Diabetes Metab. 2022;48(4):101347.
54. Skurk T, et al. The proatherogenic cytokine interleukin-18 is secreted by human adipocytes. Eur J Endocrinol. 2005;152(6):863–8.
55. Wada J, Makino H. Innate immunity in diabetes and diabetic nephropathy. Nat Rev Nephrol. 2016;12(1):13–26.
56. Liu J, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. Nat Commun. 2019;10(1):2581.
57. Xu J, McPherson PS. DENND3: a signaling/trafficking interface in autophagy. Cell Cycle. 2015;14(17):2717–8.

58. Emantoko D, Putra S, et al. Epigenetics of diabetes: a bioinformatic approach. Clin Chim Acta. 2024;557:117856.
59. Li JW, et al. Interactome-transcriptome analysis discovers signatures complementary to GWAS loci of type 2 diabetes. Sci Rep. 2016;6:35228.

Publisher's note

Text S1 Selection criteria of individuals in KORA F4 and FF4

The KORA F4 study included 3,080 participants, while the KORA FF4 study involved 2,279 participants. Methylation measurements were available for 1,799 participants in KORA F4 and 1,928 in KORA FF4, using the Illumina 450K Infinium Methylation BeadChip and Infinium MethylationEPIC BeadChip, respectively. Samples with greater than 5% missing values (based on the autosomes only) were removed, as well as whose predicted sex differed from the sex recorded at the time of the interview. After quality control, 1727 individuals remained in KORA F4 and 1874 in KORA FF4.

For this study, individuals with newly diagnosed T2D measured by oral glucose tolerance test (OGTT) or previously known T2D were categorized as having T2D. We further excluded observations due to either other types of diabetes or unknown diabetes status at KORA F4 or FF4. The longitudinal analyses of diabetes status and glycemic and insulin-related traits were restricted to 2,556 participants with at least one DNA methylation measurement at either F4 or FF4. In total, 3,501 observations from 2,556 participants in KORA F4 (1696) and FF4 (1,805) were included in the analysis. Of these participants, 945 (36.97%) had methylation data at both time points.

Text S2 CPACOR Preprocessing Pipeline

1. DNA methylation measurement: In the KORA F4 study, genome-wide DNA methylation in whole blood was analysed using the Illumina 450K Infinium Methylation BeadChip (Illumina Inc., San Diego, CA, USA). For the KORA FF4 study, the Infinium MethylationEPIC BeadChip (Illumina Inc., San Diego, CA, USA) was used according to standard protocols provided by Illumina. GenomeStudio software version 2011.1 with Methylation Module version 1.9.0 was used for initial quality control of assay performance and for generation of methylation data export files.

2. Reading in the data: Raw IDAT files were read into R (v4.3.0) using the command read.metharray from the Bioconductor package minfi (v1.46.0) and background corrected using the command bgcorrect.illumina.

3. Sex prediction: When we used the command getSex (minfi v1.46.0) on the raw data. In KORA F4, there was no individuals with a predicted sex different from the sex given at the time of the interview (cut-off -1.5). In KORA FF4, there were two individuals with predicted sex different to the sex given at the time of the interview and these individuals were removed.

4. Quality control on raw intensities: We used the command getQC (minfi v1.46.0) on the raw data. In KORA F4, 1 individual failed the QC (cut-off 9) and was removed. In KORA FF4, individuals were removed whose median intensity was less than 50% of the experiment-wide mean, or less than 2000 arbitrary units (33 individuals).

5. Detection p-value filter: Probes whose detection p-values were greater than 0.01 were set to missing.

6. Sample call rate filter: Samples with greater than 5% missing values (testing the autosomes only) were removed. In KORA F4, this led to the exclusion of 72 individuals. In KORA FF4, 9 individuals were excluded among which 4 individuals were overlapped with those failing raw intensity quality control.

7. CpG call rate filter: In KORA F4, CpG sites with greater than 5% missing values on the autosomes were removed (N= 14541). In KORA FF4, probes with greater than 5% missing values on the autosomes were also removed (N=5786).

8. CpG probe exclusion: In KORA F4, we use the manifest HM450.hg19.manifest.pop.tsv.gz (Population-specific masking HM450 file from

https://zwdzwd.github.io/InfiniumAnnotation) and set MASK_general_EUR to TRUE to obtain a reliable list of probes to be excluded. This is based on PMID: 27924034. This yields 59186 CpG sites to exclude. In KORA FF4, 1) Cross-reactive probes: There are publications providing lists for probes that hybridize to multiple possible regions (PMID: 27717381, PMID: 27330998). A total of 44493 unique probes were removed. 2) SNPs within the probe-binding region: The R package minfi v1.28.3 provides a list of SNPs within the probe-binding regions for each CpG. Probes for CpG sites known to be SNPs with minor allele frequency >0.05 (as given by minfi), or probes that had SNPs in the single base extension with minor allele frequency >0.05 were removed (11370 and 5597, respectively).

9. Quantile normalization: Quantile normalization was performed separately on the signal intensities divided into the 6 probe types: type II red, type II green, type I green unmethylated, type I green methylated, type I red unmethylated, type I red methylated (PMID: 25853392). The quantile normalized intensities were then used to generate methylation beta values, a measure from 0 to 1 indicating what percent of the cells were methylated at this locus. This step was performed separately for the autosomes, and for the sex chromosomes. For the sex chromosomes this step was performed separately for men and women. QN was performed using the R package limma v3.56.2 (PMID: 25605792).

10. Blood disorders: In KORA FF4, seven individuals have strong blood disorders. 1 had already been removed due to failing quality control, and the remaining six were removed from the dataset.

11. Cell type heterogeneity: White blood cell type proportions were estimated using the Houseman algorithm (PMID: 22568884) as implemented using the command estimateCellCounts (minfi v1.46.0) on the raw intensities and the default parameters. estimate were performed using the default types: "CD8T", "CD4T", "NK", "Bcell", "Mono",” Gran”.

12. Technical covariates: We calculated the principal components (PCs) of all the non-negative control probes, as per the CPACOR pipeline. Up to 30 control probe PCs can be used as covariates in the regression models to adjust for technical affects. Alternatively, some combination of plate, chip and chip position can be used.

13. Probe count summary: In KORA F4, the original 450K array has 485577 probes, of which 65 are SNP probes for quality control and were removed. Then the array contains 485512 probes (473864 on the autosomes, 11232 on the X chromosome, 416 on the Y chromosome). 59186 were probes to be excluded based on the population-specific masking HM450 file, and 14541 failed the detection p-value filter, a total of 73727. However, some probes overlapped both categories: a total of 70640 were removed. This leaves a total of 414872 probes: 404837 from the autosomes, 9792 from the X chromosome, 243 from the Y chromosome. In KORA FF4, the original EPIC array had 866895 probes, of which 59 are SNP probes for quality control. A "Product Quality Notice" (Tracking Number: PQN0223) issued by Illumina on April 19, 2017, indicated that 977 probes were removed due to underperformance, hence the total of 865859. 40 samples from batch 1 had defective chips and were missing 598 CpG sites. For these individuals the missing CpG sites were simply replaced with missing values in the data. Then the array contains 865859 probes (846232 on the autosomes, 19090 on the X chromosome, 537 on the Y chromosome). 44493 were cross-reactive probes, 11370 and 5597 had SNPs in the CG position and single base extensions respectively, and 5786 failed the detection p-value filter, a total of 67246. However, many probes overlapped multiple categories: a total of 59631 were removed. This leaves a total of 806228 probes: 788106 from the autosomes, 17743 X chromosome, 379 Y chromosome.

14. Sample count summary: In KORA F4, 1799 individuals were measured in one batch using the Illumina HumanMethylation 450 BeadChip. A total of 72 were removed due to quality

control: these all failed the detection rate threshold, and 1 additionally failed the median intensity step. This leaves 1727 individuals passing quality control. In KORA FF4, 1928 individuals were measured in two rounds. 2 were removed due to sex mismatch, 33 removed due to failing quality control on the raw intensities and 9 failed the detection p-value filter (4 overlap with intensity filter), leaving 1888 individuals passing quality control. In the first round, there were N=488 KORA FF4 samples. In the second round, there were N=1440 KORA FF4 samples. They were both measured using the Illumina EPIC BeadChip. Seven individuals had a noted strong blood disorder or unusual cell counts, one of whom had already been removed from the dataset. The further 6 individuals were removed. After all these steps, 8 individuals withdrew consent for their data to be used, leaving 1874 individuals.

Text S3 Selection criteria of CpG sites in KORA F4 and FF4

Probes with more than 5% missing values on the autosomes were excluded. Additionally, probes containing single nucleotide polymorphisms (SNPs) within the probe-binding regions were removed. Probes were also filtered out if the detection P-value exceeded 0.01, or if they were found to hybridize to multiple genomic regions. Probe intensities were normalized using the quantile normalization procedure for both KORA F4 and FF4. After quality control, 414,872 CpG sites remained in KORA F4 and 806,228 in KORA FF4, with 383,057 overlapping CpG sites. Following the exclusion of sex chromosome CpG sites, 374,054 CpG sites were left in the final analysis.

Text S4 Quality control for KORA FF4 gene expression data

After RNA isolation using PAXgene Blood RNA Kit, RNA integrity number (RIN) was measured using the Agilent 2100 Bioanalyzer system. RNA samples with RIN values of approximately 6 or more were selected for mRNA sequencing (poly-A selected). The libraries were prepared using the Illumina stranded mRNA prep ligation kit (Illumina), following the

kit's instructions. After a final QC, the libraries were sequenced in a paired-end mode (2x100 bases) in the Novaseq6000 sequencer (Illumina) with a depth of ⩾ 40 Million reads per sample. After demultiplexing, FASTQ files from each sample are processed using standard tools. Alignment to UCSC Genome Browser hg19 human reference genome using STAR v2.4.2a (PMID: 23104886). Unaligned reads are discarded. Sequencing QC was done using RNASeQC v1.1.8.1 (PMID: 22539670). Properly aligned reads are then processed with HTSeq-count v0.6.1 (PMID: 25260700) to generate read counts which can be interpreted as quantified gene expression. The reads are then normalized for exon length and total sequencing yield to generate Fragments Per Kilobase of transcript per Million mapped reads (FPKM), and this is done through dividing the fragments per gene by the product of length of the gene in kilobase and million reads sequenced.

After sequencing QC, samples QC was done. Samples with < 30 million reads were discarded. Exonic, intronic, intragenic, intergenic and rRNA rates calculated by RNAseQC were examined for outliers but no such outliers were found, and no samples were excluded based on these. Only the genes with FPKM of ⩾ 1 in at least 5% of the samples were selected. Number of the selected genes in each sample were calculated. Samples having less than 5750 genes were excluded. Sex mismatches in the phenotype tables and those discerned from looking at the expression of XIST and UTY genes were also excluded.

Text S5 WHO criteria of type 2 diabetes

Normal glucose tolerance (fasting glucose <6.1 mmol/l and 2h glucose <7.8 mmol/l); prediabetes defined as (1) impaired fasting glucose (IFG; fasting glucose ⩾6.1 mmol/l but <7.0 mmol/l, and 2h-glucose <7.8 mmol/l), (2) impaired glucose tolerance (IGT; fasting glucose <6.1 mmol/l and 2h glucose ⩾7.8 mmol/l but <11.1 mmol/l) or (3) combination of (1) and (2);

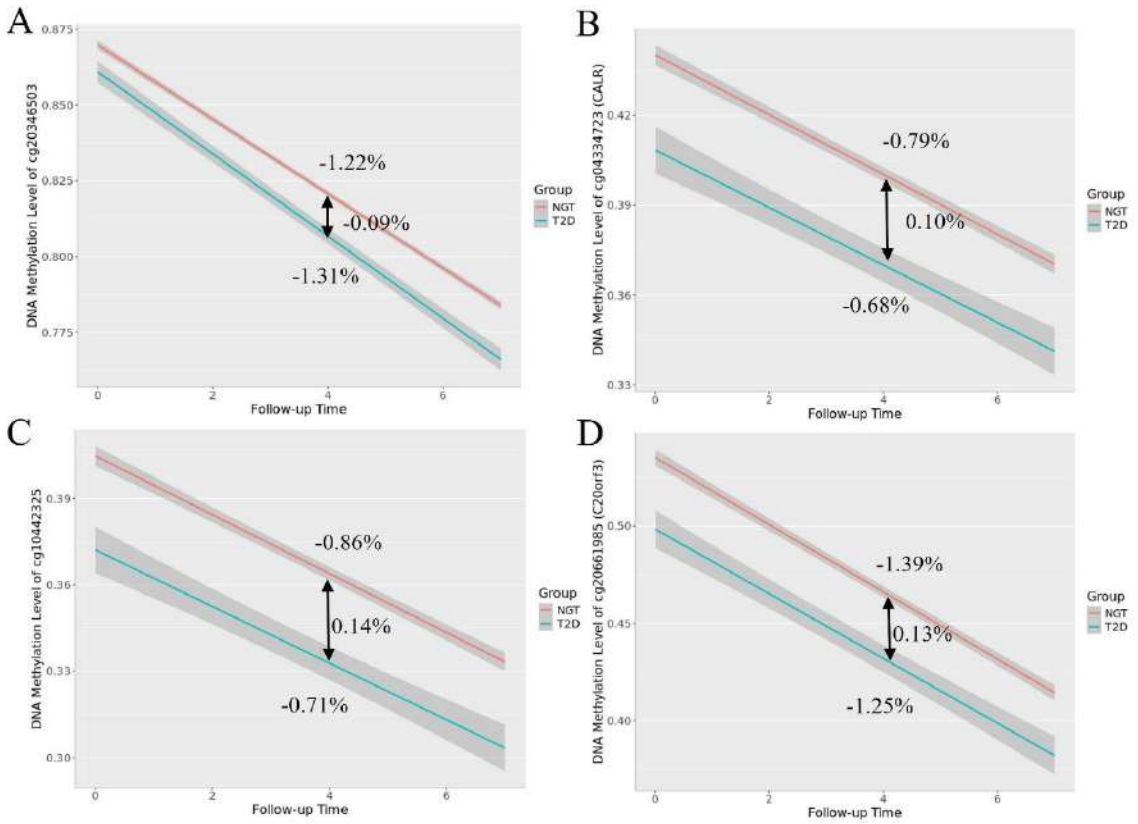
and newly diagnosed T2D (fasting glucose ⩾7.0 mmol/l or 2h-glucose ⩾11.1 mmol/l) were defined according to the 1999/2006 WHO criteria.

Table S1 Characteristics of population with repeated methylation measurements

Characteristics	KORA F4				KORA FF4			
	All N=945	NGT N=666	Prediabetes N=170	T2D N=109	All N=945	NGT N=570	Prediabetes N=194	T2D N=181
Age (years)	57 (12)	56 (11)	60 (12)	64 (12)	64 (12)	62 (12)	65 (11)	69 (12)
Male (%)	459 (48.6%)	297 (44.6%)	97 (57.1%)	65 (59.6%)	459 (48.6%)	239 (41.9%)	110 (56.7%)	110 (60.8%)
BMI (kg/m2)	27.1 (5.9)	26.1 (5.1)	29.8 (6.0)	30.4 (6.5)	27.4 (6.2)	26.2 (5.6)	29.0 (5.50)	29.9 (7.18)
Smoking								
Never smoker	389 (41.2%)	278 (41.7%)	71 (41.8%)	40 (36.7%)	389 (41.2%)	237 (41.6%)	77 (39.7%)	75 (41.4%)
Former smoker	415 (43.9%)	274 (41.1%)	84 (49.4%)	57 (52.3%)	436 (46.1%)	256 (44.9%)	91 (46.9%)	89 (49.2%)
Current smoker	141 (14.9%)	114 (17.1%)	15 (8.8%)	12 (11.0%)	120 (12.7%)	77 (13.5%)	26 (13.4%)	17 (9.39%)
Hypertension	367 (38.8%)	193 (29.0%)	91 (53.5%)	83 (76.2%)	447 (47.3%)	201 (35.3%)	114 (58.8%)	132 (72.9%)
Fasting glucose	5.3 (0.8)	5.2 (0.6)	5.9 (1.0)	7 (2.3)	5.6 (1.0)	5.3 (0.6)	6.1 (0.7)	7.3 (2)
HOMA-IR	2.04 (1.7)	1.8 (1.2)	3.11 (2.5)	4.61 (3.5)	2.3 (2)	2.0 (1.3)	3.6 (2.2)	4.8 (4.3)
HOMA-B	99.4 (64.1)	98.8 (58.4)	115.0 (78.5)	84.2 (76.1)	96.0 (67.7)	95 (61.9)	110. (84.2)	88.4 (71.7)
HbA1c	37.0 (7)	36 (5)	38 (4.8)	46 (11)	37.0 (6)	35 (5)	38 (4)	45 (10)
HDL-cholesterol	1.4 (0.5)	1.5 (0.5)	1.3 (0.4)	1.2 (0.4)	1.7 (0.7)	1.8 (0.7)	1.5 (0.5)	1.4 (0.5)
Triglycerides	1.3 (0.9)	1.1 (0.8)	1.6 (1.1)	1.3 (1.2)	1.3 (0.8)	1.1 (0.6)	1.4 (1.0)	1.6 (1.3)
Medication	46 (4.9%)	0 (0%)	0 (0%)	46 (42.2%)	104 (11.0%)	0 (0%)	0 (0%)	104 (57.5%)
Parental history								
Yes	247 (26.1%)	161 (24.2%)	48 (28.2%)	38 (34.9%)	268 (28.4%)	140 (24.6%)	57 (29.4%)	71 (39.2%)
No	476 (50.4%)	365 (54.8%)	78 (45.9%)	33 (30.3%)	569 (60.2%)	373 (65.4%)	115 (59.3%)	81 (44.8%)
Unknown	254 (13.3%)	90 (13.5%)	25 (14.7%)	11 (10.1%)	108 (11.4%)	57 (10%)	22 (11.3%)	29 (16.0%)

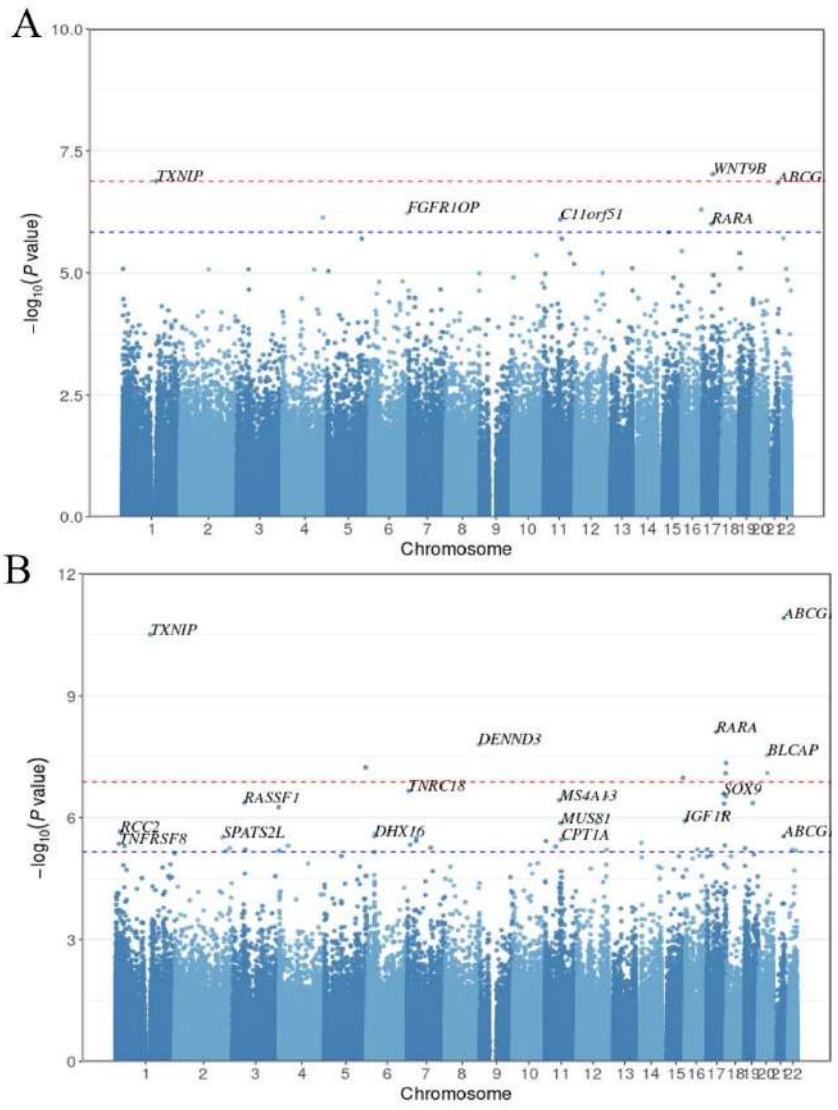
Data are median (IQR) for continuous variables and n (%) for categorical variables. The unit for both fasting glucose and HbA1c is mmol/mol. The unit for both HDL-cholesterol and triglycerides is mmol/l. Medication means the glucose-lowering medication.

Fig. S1



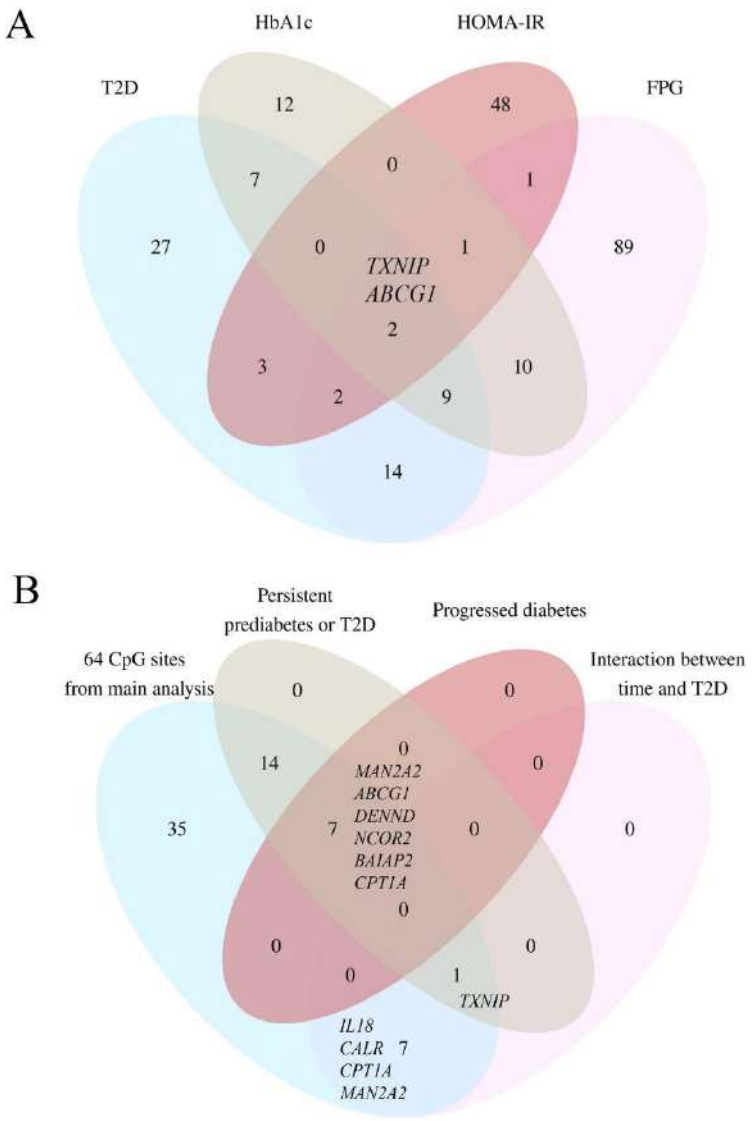
Line plots illustrate the rate of methylation change over time across different groups. The red and blue line represents the individuals with NGT and T2D, respectively. (A) cg20346503; (B) cg04334723 (*CALR*); (C) cg10442325; (D) cg20661985 (*C20orf3*).

Fig. S2



Manhattan plots of sensitivity analysis. The x axis indicates the chromosome location, and the y-axis represents the −log10 (*p*-value). The Bonferroni threshold of 1.34×10^{-7} is marked by a blue solid line, while the Benjamini–Hochberg (FDR) threshold (p_FDR < 0.05) is indicated by a red dashed line. (A) Manhattan plot of EWAS results from extended model. (B) Manhattan plots of EWAS results from individuals with two-time points methylation data.

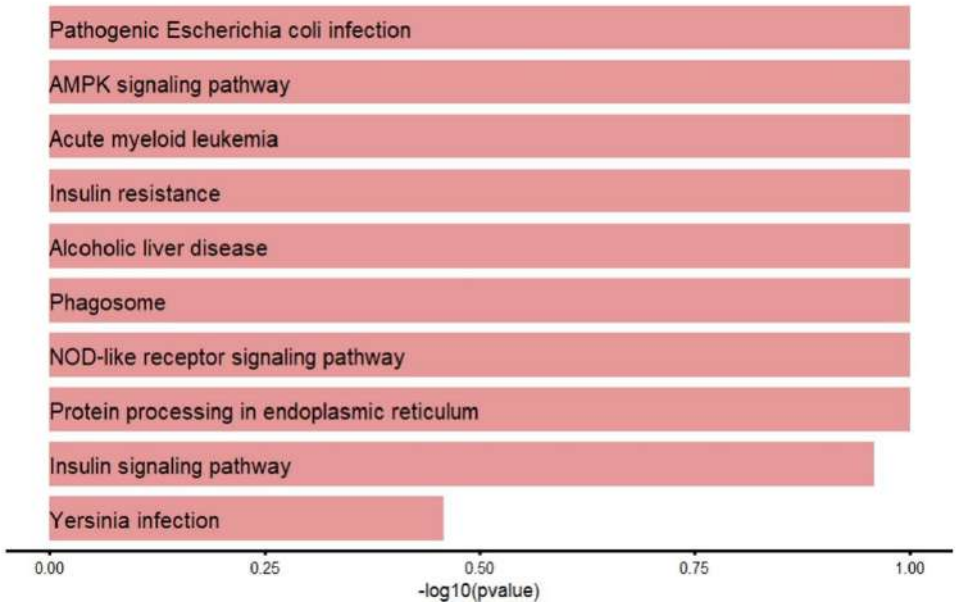
Fig. S3



166

167 Venn plot illustrates the overlap of CpG sites from different analysis.

Fig. S4



168

169 The top 10 non-significant pathways associated with T2D and glycemic traits. The x-axis represents the

170 −log10(p-value), and the red dashed line represents the significant threshold (p_FDR < 0.05).

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisors, Prof. Dr. Annette Peters, Director of the Institute of Epidemiology and Dr. Melanie Waldenberger, Head of Research Group "Complex Diseases" at the Research Unit of Molecular Epidemiology (AME) of Helmholtz München, for their invaluable guidance, support and encouragement throughout my PhD journey. Their invaluable expertise, enthusiasm and patience were instrumental in the successful completion of these research projects. I am especially thankful for their insightful advice, engaging discussions, and invaluable feedback.

I would like to thank my research team for their hard work and dedication. I am particularly grateful to Dr. Pamela R. Matías-García, and Dr. Thomas Delerue for their thoughtful insights, productive discussions, support, and guidance during my research.

I would like to express my gratitude to Prof. Dr. Eva Hoster, Prof. Dr. Christian Herder, and Dr. Rory Wilson, who provided invaluable feedback and support on my research.

I am deeply indebted to Chinese Scholarship Council for their financial support throughout my research.

To my parents and husband- words cannot express how grateful I am for your unconditional love and support throughout my educational journey.

Finally, I would like to thank all those who have contributed to my PhD projects in one way or another.