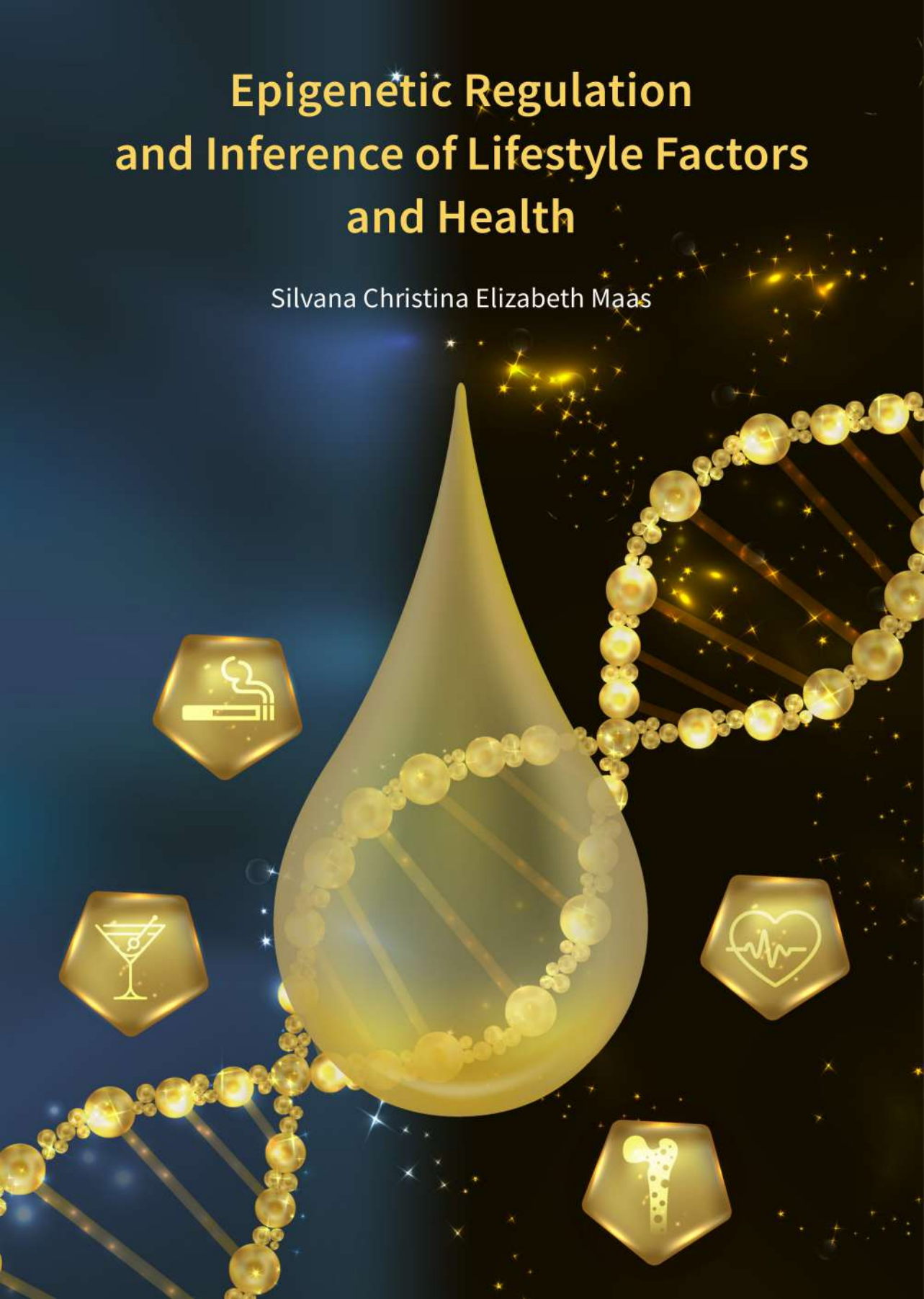
Epigenetic Regulation and Inference of Lifestyle Factors and Health

Silvana Christina Elizabeth Maas



Epigenetic Regulation and Inference of Lifestyle Factors and Health

Silvana Christina Elizabeth Maas

ACKNOWLEDGEMENTS

The research presented in this thesis was performed at the Department of Epidemiology and the Department of Genetic Identification, Erasmus MC, Rotterdam, the Netherlands. The studies described in this thesis were performed within the Rotterdam Study, the Netherlands Twin Register, Cohort on Diabetes and Atherosclerosis Maastricht, Prospective ALS Study Netherlands, Leiden Longevity Study, the Cooperative Health Research in the Region of Augsburg- F4 study, Study of Health in Pomerania- Trend, TwinsUK, LifeLines DEEP, and the Generation R study. We gratefully acknowledge the contributions of the study participants, staff, participating general practitioners, and pharmacists involved in all studies.

Publication of this thesis was kindly supported by the Department of Epidemiology of the Erasmus Medical Center and by the Erasmus University Rotterdam.

ISBN: 978-94-6361-612-6

Layout, cover design and printing by Optima Grafische Communicatie

Epigenetic Regulation and Inference of Lifestyle Factors and Health

Epigenetische regulatie en inferentie van leefstijlfactoren en gezondheid

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. A.L. Bredenoord

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 9 februari 2022 om 13.00 uur

door

Silvana Christina Elizabeth Maas

geboren te Breda

Erasmus University Rotterdam

Erasmus

PROMOTIECOMMISSIE

Promotoren:	Prof.dr. M.H. Kayser
	Prof.dr. M.A. Ikram
Overige leden:	Prof.dr. J.B.J. van Meurs
	Prof.dr. B.T. Heijmans
	Dr. A. Dehghan
Copromotoren:	Dr. M. Ghanbari
	Dr. A. Vidaki

Paranimfen:	Carolina Patricia Ochoa Rosales
	Irma Karabegović

To Marcelo and Mateo-Millan

MANUSCRIPTS THAT FORM THE BASIS FOR THIS THESIS

Chapter 2. Lifestyle factor inference from DNA methylation

Silvana C.E. Maas, Athina Vidaki, Alexander Teumer, Ricardo Costeira, Rory Wilson, Jenny van Dongen, Marian Beekman, Uwe Völker, Hans J. Grabe, Sonja Kunze BIOS Consortium, Karl-Heinz Ladwig, Joyce B.J. van Meurs, André G. Uitterlinden, Trudy Voortman, Dorret I. Boomsma, P. Eline Slagboom, Diana van Heemst, Carla J.H. van der Kallen, Leonard H. van den Berg, Melanie Waldenberger, Henry Völzke, Annette Peters, Jordana T. Bell, M. Arfan Ikram, Mohsen Ghanbari*, Manfred Kayser*. Validating biomarkers and models for epigenetic inference of alcohol consumption from blood. *Clinical Epigenetics* 2021;13(1):198. Doi: 10.1186/s13148-021-01186-3.

Silvana C.E. Maas, Athina Vidaki, Rory Wilson, Alexander Teumer, Fan Liu, Joyce B.J. van Meurs, André G. Uitterlinden, Dorret I. Boomsma, Eco J.C. de Geus, Gonneke Willemsen, Jenny van Dongen, Carla J.H. van der Kallen, P. Eline Slagboom, Marian Beekman, Diana van Heemst, Leonard H. van den Berg, BIOS Consortium, Liesbeth Duijts, Vincent W.V. Jaddoe, Karl-Heinz Ladwig, Sonja Kunze, Annette Peters, M. Arfan Ikram, Hans J. Grabe, Janine F. Felix, Melanie Waldenberger, Oscar H. Franco, Mohsen Ghanbari*, Manfred Kayser*. Validated inference of smoking habits from blood with a finite DNA methylation marker set. *European Journal of Epidemiology* 2019;34(11):1055-74. Doi: 10.1007/s10654-019-00555-w.

Chapter 3. Lifestyle factor induced changes in DNA methylation and cardiovascular outcomes

Silvana C.E. Maas, Michelle M.J. Mens, Brigitte Kühnel, Joyce B.J. van Meurs, André G. Uitterlinden, Annette Peters, Holger Prokisch, Christian Herder, Harald Grallert, Sonja Kunze, Melanie Waldenberger, Maryam Kavousi, Manfred Kayser, Mohsen Ghanbari. Smoking-related changes in DNA methylation and gene expression are associated with cardiometabolic traits. *Clinical Epigenetics* 2020;12(1):157. Doi: 10.1186/s13148-020-00951-0.

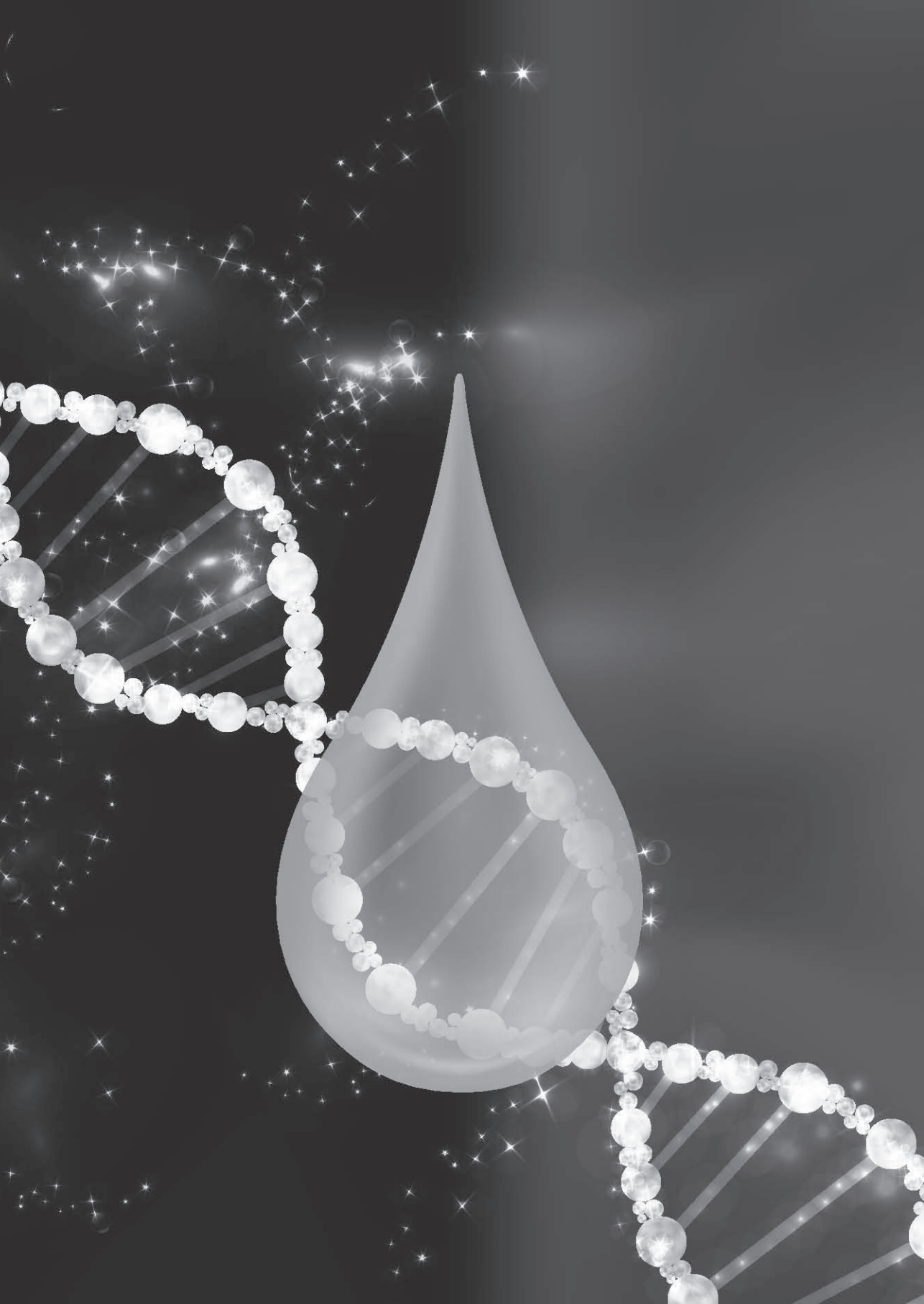
Chapter 4. MicroRNA expression and health outcomes

Michelle M.J. Mens, **Silvana C.E. Maas**, Jaco Klap, Gerrit Jan Weverling, Paul Klatser, Just P.J. Brakenhoff, Joyce B.J. van Meurs, André G. Uitterlinden, M. Arfan Ikram, Maryam Kavousi, Mohsen Ghanbari. Multi-Omics Analysis Reveals MicroRNAs Associated With Cardiometabolic Traits. *Frontiers in Genetics* 2020;11:110. Doi: 10.3389/fgene.2020.00110.

Irma Karabegović, **Silvana Maas**, Carolina Medina-Gomez, Maša Zrimšek, Sjur Reppe, Kaare M. Gautvik, André G. Uitterlinden, Fernando Rivadeneira, Mohsen Ghanbari. Genetic Polymorphism of miR-196a-2 is Associated with Bone Mineral Density (BMD). *International Journal of Molecular Sciences* 2017;18(12):2529. Doi: 10.3390/ijms18122529.

*Denotes equal contribution

TABLE OF CONTENTS



Chapter 1

General introduction

INTRODUCTION

Lifestyle factors are modifiable behaviors, including someone's diet, smoking habits, alcohol consumption, physical activity, and others. It has been shown that these lifestyle factors are associated with disease risk [1]. For example, high salt intake, the lack of physical activity (sedentary behavior), being a smoker, and/or heavy alcohol consumption are associated with a wide range of health outcomes, including most non-communicable diseases (NCDs) [1, 2]. NCDs, also known as chronic diseases, are the leading cause of death worldwide [3]. The risk factors leading to NCDs can be categorized as genetics, physiological, environmental, and lifestyle factors. Substantial advances have been made in the diagnoses and treatment of these diseases; nevertheless, their prevalence continues to increase worldwide. Hence, it would be important to not only focus on patients that already have a disease but also on disease prevention. This highlights the importance of more in depth research investigating the underlying mechanisms of disease risk factors, and subsequently to disease onset. It is impossible, for now, to change someone's genetic information and also the environmental exposures are not always easy adaptable. Therefore, the impact of lifestyle behavior should have a more prominent role in NCD-related research. Most often, lifestyle information is studied using information obtained via subjective measurements like self-reported questionnaires or interviews. To better understand the lifestyle-related effects on disease, new and/or improved objective measurements are required to overcome any possible discrepancies. Additionally, it is not yet clear via which molecular mechanism these lifestyle factors affect disease onset. It has been shown that the interplay between environmental and lifestyle factors together with genetics direct the dynamic epigenome [4]. For that reason, epigenetics has been proposed as a possible mechanism linking lifestyle to disease risk, possibly via altering gene expression (**Figure 1**). The integration of genetic, epigenetic, and gene expression information while studying disease risk might be the solution to disentangle these complex interplays (**Figure 1**). Moreover, epigenetic variation determined by lifestyle factors may provide a suitable resource to develop biomarker for predicting such lifestyle factors from human biological material, which eventually could become useful in various areas of medical and non-medical research and applications, even for investigative purposes in forensics [5].

1.1 Genomics and Gene expression

1.1.1 Genomics

The human genome consists of the complete set of nuclear deoxyribonucleic acid (DNA), the inheritable blueprint for life that together holds the information for cell function activity [6]. The total DNA consist of ~3.2 billion base pairs with a >99.5% similarity between any two individuals. With the completion of the 1000 Genomes Project, more

than 88 million variants in the DNA sequence were discovered. On average, there is a difference of 4.1 to 5 million sites between any typical genome and the reference genome, making any person unique [7]. Monozygotic twins are the only exception to this as they share the same DNA sequence. Changes in the DNA sequence can occur in the form of single nucleotide polymorphisms (SNPs) or structural variations, including deletions, duplications, copy-number variants, insertions, inversions, and translocations [8]. Depending on the location of the DNA variation, a variant can affect gene function in the form of protein truncation, peptide-sequence alteration, and by altering regulatory regions, such as promoters, enhancers, or transcription factor binding sites [7].

Alterations in the DNA sequence can lead to the susceptibility and onset of disease; therefore, they are of great interest in disease studies [6, 9, 10]. Due to advances in SNP microarray technology development, it became possible to identify genotype information for hundreds of thousands SNPs with a single array. These genotype data can subsequently be used in polygenic disease studies, in which SNPs in several genes play a role. A commonly used approach to investigate these diseases is via genome-wide association studies (GWAS), in which the obtained genotypes across the genome from participants with the disease of interest are compared to those of non-diseased participants in a hypothesis-free approach to identify disease-related genetic variants [11].

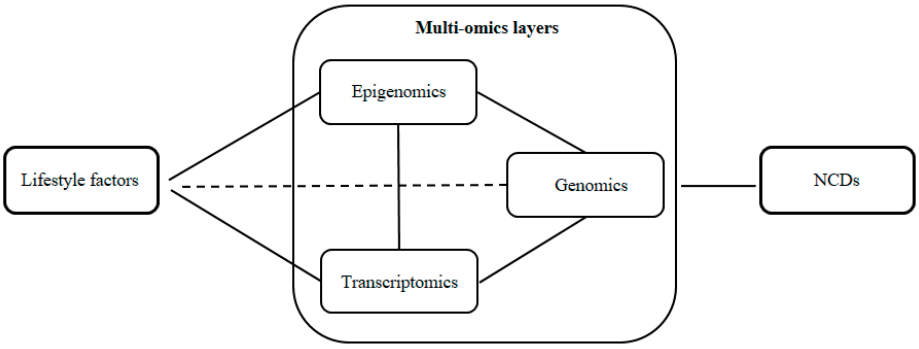


Figure 1. The conceptual relationship between lifestyle factors, multi-omics layers, and health outcomes. Lifestyle factors are associated with multi-omics layers, including epigenetic markers, gene expression levels, and to some extent the genetic sequence. Lifestyle factors are also associated with several diseases, including most non-communicable diseases (NCDs), possibly via alterations in the multi-omics layers.

1.1.2 Gene expression

The DNA sequence contains genes that are used as a template during the transcription process, in which a complementary single-stranded ribonucleic acid (RNA) strand is formed (**Figure 2**) [12, 13]. The complete set of RNA transcripts (transcriptome) reflects the gene expression pattern that varies across tissues [14, 15]. The nuclear DNA sequence contains around 20,000 protein-coding genes that are transcribed into messenger RNA

(mRNA) and subsequently, translated into a protein (**Figure 2**) [12, 13]. As proteins direct the activities of cells and functions of the body, changes in gene expression might have severe consequences for disease risk. The complete transcriptome is investigated for its disease association using a hypotheses-free approach with transcriptome-wide association analysis (TWAS). In TWAS, the transcriptome between participants with a disease is compared to non-diseased participants to identify disease-related genes. Genetic mutations in the coding regions could affect the gene translation into proteins (**Figure 2**), while a mutation in the non-coding regions could affect the gene transcription process [14]. Therefore, when a SNP is associated with a disease, the expression of the host gene that contains the SNP could be further investigated for its association with the disease of interest.

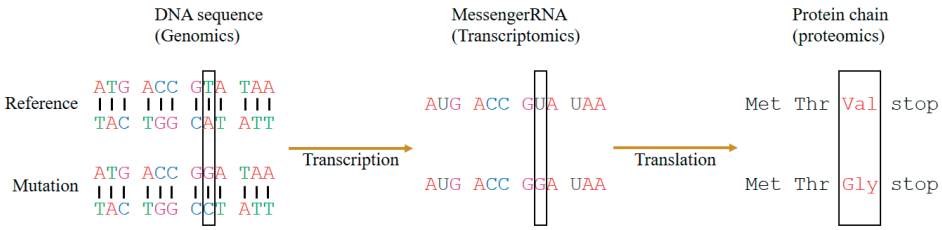


Figure 2. From DNA sequence to protein. Nuclear DNA contains protein-coding genes that are transcribed into messenger RNA (mRNA) and subsequently translated into a protein. During transcription, a variation in the DNA sequence (highlighted with blocks) results in an altered mRNA sequence. The translation of the altered mRNA sequence can subsequently lead to a different amino acid in the protein chain. For example, the reference mRNA sequence "GUA" codes for the amino acid "Val", while the mutated sequence "GGA" codes for "Gly".

1.2 The epigenome and epigenetic mechanisms

1.2.1 The epigenome

Besides due to the effect of DNA variation, gene expression can also be altered via epigenetic mechanisms on both the transcriptional and translational levels. Epigenetic mechanisms change the gene transcription and translation without alterations in the DNA sequence [16]. Epigenetics functions in a tissue- or cell-specific manner by controlling stable repression of genes not required in specific cell types [17, 18]. In this line, epigenetics is an important regulator in mammalian development mechanisms, including X-chromosome inactivation, differentiation of pluripotent stem cells, mediating of allele-skewed gene expression, such as differential allele expression, and allele-specific gene expression, such as imprinting [17-21]. The epigenetic markers studied in this thesis include the microRNAs and DNA methylation markers.

1.2.2 MicroRNA biogenesis and gene expression regulation

The transcripts that do not possess any protein-coding capacity, the non-coding (nc) RNAs, include epigenetic markers that play an important role in the post-transcriptional as well as translational coordination of gene expression [22-25]. So far, the best-characterized ncRNA is the class of microRNAs (miRNAs), which have gained widespread attention as important modulators of different biological processes. MiRNAs can be encoded from the introns of protein-coding genes, long non-coding transcripts, and the chromosomal regions between two genes: the intergenic regions [26-28]. MiRNAs are transcribed in the nucleus into primary miRNA (pri-miRNA) and subsequently cleaved into a hairpin precursor miRNA (pre-miRNA) that is then exported into the cytoplasm [28-31]. Here, the loop structure is cleaved of the pre-miRNA, resulting in a ~22bp double-stranded miRNA. After dissociation of the miRNA, the passenger strand is degraded and the guide strand (the mature miRNA) is fused into the RNA-induced silencing complex (RISC), together with Dicer, TRBP, PACT, and Argonaute (Ago) proteins [22, 28]. This RISC complex interacts with the 3'-untranslated region (3'UTR) of target mRNA. The "seed" sequence is the core of the mature miRNA (nucleotides 2 to 7-8, from the 5' end). Perfect complementarity between this region and the 3'UTR sequence of the target mRNA will result in site-specific cleavage of the mRNA [22, 28, 32], while imperfect complementarity will repress translational and/or mRNA destabilization (**Figure 3**) [28, 33-35]. Alterations in DNA methylation patterns or variation in miRNA-related sequences can lead to changes in the miRNA's biogenesis, secondary structure, free-binding energy, and expression, leading to phenotypic changes [36-45]. MiRNAs are selectively sorted into extracellular vesicles, including exosomes and microvesicles, providing them a stable form for cell secretion to nearby or distant targets (**Figure 3**) [46]. Thus, miRNAs are also found in extra cellular fluids, including plasma, making them easily accessible and a possible target for disease biomarkers [47, 48]. The expression of these circulating miRNAs could be obtained in an individual manner or a transcriptome-wide approach using arrays for disease testing [49].

A more novel approach to identify potentially important miRNAs is via the use of publicly available GWAS statistics data [50]. In this context, the genomic position of human miRNAs is obtained using publicly available databases, including FANTOM5 [40], miRBase database [51], and ProMiR II [52], and SNPs located in miRNA-related sequences via the dbSNP database [53]. By means of publicly available GWAS meta-analysis summary statistics, it is possible to identify the association for these miRNA-related SNPs with several disease traits.

1.2.3 DNA methylation and gene expression

Another epigenetic regulator is DNA methylation, the most studied epigenetic mechanism, in which DNA methyltransferases (DNMTs) transfer a methyl group (-CH_3) from

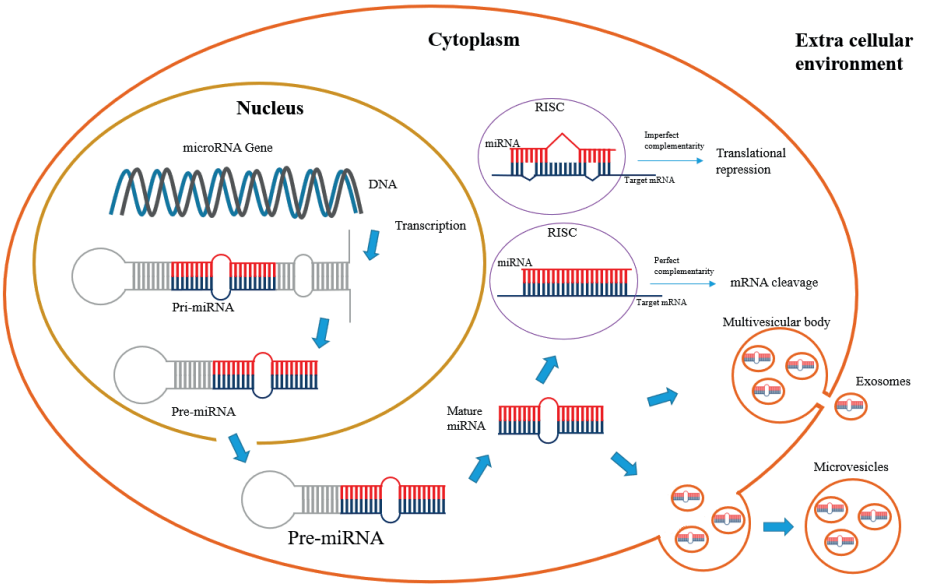


Figure 3. MiRNA biogenesis and target mRNA interaction. MiRNAs are transcribed from a miRNA gene in the DNA sequence into primary miRNA (pri-miRNA). The pri-miRNA is cleaved into a hairpin precursor miRNA (pre-miRNA) and exported into the cytoplasm, where the loop structure is cleaved of. After dissociation of the miRNA, the mature miRNA is fused into RISC, which interacts with the 3'UTR of the target mRNA. Perfect complementarity between the miRNA "seed" sequence and the mRNA 3'UTR sequence results in mRNA cleavage, while imperfect complementarity will repress translational and/or mRNA destabilization. MiRNAs are secreted from cells sorted into extracellular vesicles to nearby or distant targets.

S-Adenosyl methionine to the 5' position of carbon in cytosines (C) neighboring guanine (G) (**Figure 4**). These sites are referred to as CpG sites (CpGs); 5'-Cytosine-phosphate-Guanine-3' [54]. Such a methyl group can affect DNA accessibility for the RNA polymerase during the transcription process, resulting in pre-transcriptional alteration of gene expression [55]. CpGs associated with changes in gene expression are referred to as expression quantitative trait methylation (eQTM). Methylation of CpGs located in the promoter, enhancer, or the transcription start site usually silences gene expression; conversely, hypomethylation in these regions is generally associated with increased transcription (**Figure 4**) [17]. Therefore, when testing their association with gene expression, most studies focus on CpGs located in these regions. Due to the importance of DNA methylation in regulating crucial aspects of the genome's function, it is extensively studied for its association with complex traits. Experimental samples contain several cells in which a single CpG site can be either methylated or not. The proportion of a CpG methylated in a sample can be measured using DNA methylation arrays and is used as a quantitative trait [56]. Changes in DNA methylation levels at independent CpGs are tested for their association with disease, using candidate CpG approaches or large-scale epigenome-wide association studies (EWAS) [57].

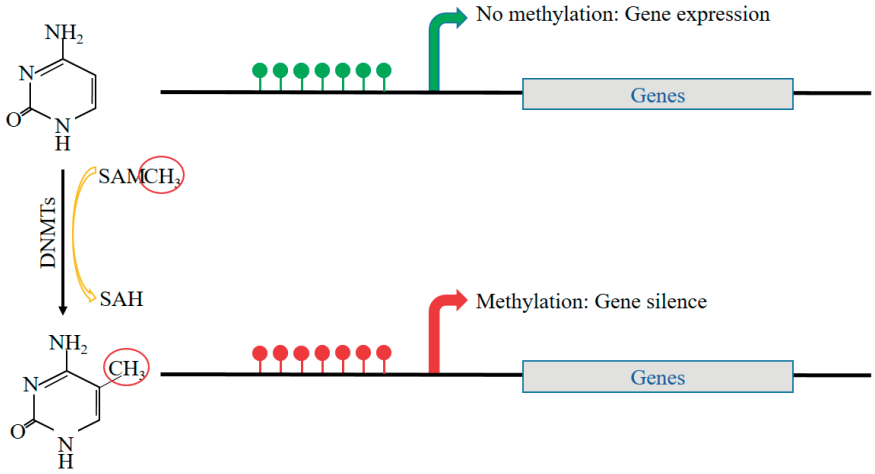


Figure 4. DNA methylation and gene expression. DNA methylation refers to the addition of a methyl group (-CH_3) to CpG sites. DNA methylation is performed by DNA methyltransferases (DNMTs) that transfer the methyl group from S-Adenosyl methionine (SAM). DNA accessibility for the RNA polymerase can be affected by the presence of such methyl group resulting in altered gene expression.

1.3 DNA methylation-based lifestyle inference

1.3.1 Objective measures for lifestyle factors

The dynamic epigenome is directed by alterations in the genome and by exposure to different environmental factors [4], such as a person's lifestyle habits like tobacco smoking habits and alcohol consumption, but also the exposure from one's surrounding. Smoking is the most studied modifiable lifestyle factor. Tobacco smoking, second-hand exposure, and chewing combined is the second-leading risk factor globally for attributable deaths, accounting for 15.4% (95% CI = 14.6%–16.2%) of all deaths in 2019 [1]. Moreover, alcohol consumption is another important modifiable lifestyle factor that is a major risk factor for disease development. Alcohol consumption was estimated to be the leading risk factor for those aged 25–49 years and the eight-leading risk factor in all men, accounting for 2.07 million (95% CI = 1.79–2.37) deaths, and the 14th-leading risk factor in women, accounting for 374 thousand (95% CI = 298–461) deaths globally in 2019 [1]. Lifestyle information, including smoking and alcohol consumption, is often studied as the main exposure for disease risk or as a confounding factor for adjustment. However, most studies collect lifestyle data using self-reported questionnaires, which are prone to error, usually underreporting negatively viewed lifestyle factors such as smoking and alcohol consumption [58]. Therefore, an accurate biomarker that could infer this information would be helpful for health care providers and researchers to complement, or even replace, self-reported questionnaires.

Blood-based toxicological tests exist for both smoking and alcohol consumption; however, they come with several limitations, including short half-time, low prediction

accuracies and/or solely assessing recent and excessive exposure [59-62]. A more recent proposed mechanism for lifestyle inference, including smoking and alcohol consumption, is DNA methylation. Several large cohort studies already implemented DNA methylation measurements in their data collection procedures due to its possible association with health outcomes. Therefore, compared to standard blood-based toxicological tests, implementing a DNA methylation-based prediction model would not result in extra costs and would, thus, have a large benefit for cohort studies.

1.3.2 DNA methylation-based inference of smoking habits

Several EWAS studies have already identified hundreds of CpGs associated with smoking habits [63-81]. The largest EWAS to date was conducted by Joehanes *et al.* [82], implementing a large-scale meta-analysis employing 15,907 participants embedded in 16 cohorts. In total, 2,623 CpGs were differentially methylated between smokers and never smokers ($P< 1\times10^{-7}$). Also, a few studies have already explored the possibility to infer smoking habits using DNA methylation [64, 83-89]. These currently available DNA methylation-based prediction models come with several limitations, including small sample size, limited validation, exclusion of categories in the models, and/or utilizing large numbers of CpGs. In particular, Philibert *et al.* [84] obtained an AUC of 0.99 using only the methylation levels of cg05575921 (*AHRR*), employing 35 non-smokers and 26 smokers. Similarly, Endo *et al.* [88] obtained an AUC of 0.96 using a DNAm rate cut-off point (58.96%) for cg23576855 (*AHRR*) for current smoking using 19 never smokers, 7 former smokers, and 7 current smokers. Although both studies provide high AUCs, the use of very small sample size questions the reliability of the obtained prediction accuracies, which needs to be established from much larger data.

Smoking prediction models using subsets of categories were developed that will be applicable in certain settings. For instance, Shenker *et al.* [83] focused on distinguishing former smokers from never smokers obtaining an AUC of 0.82 (95% CI = 0.64–0.99) in the test set (N= 81) and 0.83 (95% CI = 0.70–0.96) in the validation set (N=180). Similarly, Zhang *et al.* [85] used cotinine levels, DNA methylation levels of cg05575921 *(AHRR*), a methylation score, and their combination with cotinine. Discriminating current smokers from never smokers obtained for all methods in the validation set (N=500) AUCs ≥0.96, while distinguishing former from never smokers resulted in an AUC of 0.54 for cotinine, 0.78 for cg05575921, and 0.83 for the methylation score. Although models using subsets of the categories could be helpful in certain settings, it is important to note that for a prediction model to be applicable to the general population it would need to include all smoking categories.

A few studies have obtained high accuracies while employing a large number of CpGs. For example, Elliot *et al.* [64] conducted a DNA methylation-based smoking score using 183 CpGs previously associated with smoking status [66], that can identify smokers with

100% sensitivity and 97% specificity in Europeans (N=95). Similarly, Mc Cartney *et al.* [89] used 233 CpGs to distinguish current from never smokers, obtaining an AUC of 0.98 (95% CI = 0.97–1.0). A large marker set might result in a higher chance of missing values in one or more of the markers due to the strict quality controls implemented in cohort studies. For example, Elliot *et al.* [64] used only 183 CpGs out of the 186 CpGs previously identified by Zeilinger *et al*. [66] as the three additional CpGs did not pass quality control measures. Also, studies often include a limited sample size, which might result in model overfitting when too many predictive markers are included [90].

1.3.3 DNA methylation-based inference of alcohol consumption

Several studies have conducted EWASs on alcohol consumption, alcohol use disorder, and alcohol withdrawal [91-101]. The largest EWAS to date was done by Liu *et al.* [102], employing 9,643 participants of European ancestry identifying 363 CpGs ($P< 1\times10^{-7}$) associated with alcohol consumption in grams/day. This study also developed alcohol prediction models for four alcohol consumption categories: heavy drinker, at-risk drinkers, light drinkers, and non-drinkers. Out of 361 CpGs ($P< 5\times10^{-6}$, N= 6,926), 144 CpGs and three subsets (78, 23, and 5 CpGs) obtained high prediction accuracies inferring subsets of the four alcohol categories. A major limitation of the study was the lack of independent validation of the obtained prediction models. Unique marker-weights were used during the external replication phase rather than the weights obtained in the model building phase. Another limitation of the study is the use of subsets of the four alcohol categories. In particular, in the model distinguishing heavy drinkers from non-drinkers, only data from heavy and non-drinkers were used while excluding the data from participants categorized as light and at-risk drinkers.

Similar to the smoking models, Philibert *et al.* [101] and Endo *et al.* [88] developed alcohol consumption prediction models using a small sample size including 343 and 33 participants, respectively. Also, McCartney e*t al.* [89] developed a model to distinguish light-to-moderate drinkers (N=745) from heavy drinkers (N=150) using 450 CpGs. This emphasizes the need for the development of a validated and precise prediction model including all possible categories for smoking habits and alcohol consumption based on a finite set of DNA methylation markers.

1.3.4 Lifestyle inference as an application in forensic investigations

Although not yet established, molecular biomarkers for lifestyle factors that allow inferring such factors from human biological materials could be helpful in forensic casework, particularly for investigative purposes to find unknown perpetrators of crime who in principle cannot be identified with forensic DNA profiling [5]. Forensic DNA profiling is comparative in nature and uses a set of highly polymorphic autosomal short tandem repeats (STRs) [103-105]. In forensic cases, these STR profiles are obtained from bio-

logical samples collected at crimes scenes and compared to the STR profiles of suspects or profiles of known offenders stored in national forensic DNA databases. However, in cases where no match is found, because the perpetrator is not amongst the known case suspects and is not in the forensic DNA database, comparative DNA profiling fails in identifying a perpetrator. In such case, lifestyle information of an unknown perpetrator inferred from a biological trace left behind at a crime scene could help narrowing down the suspect pool by further detailing the currently considered DNA-predicted externally visible characteristic information on age, bio-geographic ancestry, and appearance [5, 106, 107], thereby providing a very different application of predictive epigenetics with benefit to society outside the field of medicine and public health.

1.4 Risk factors for non-communicable diseases

1.4.1 Smoking, epigenetics, and cardio-metabolic traits

Smoking is a major risk factor for disease development, including for non-communicable diseases (NCDs) [2]. Smoking attributes to one in six of all deaths resulting from NCDs [2]. Cardiovascular disease (CVD) accounts for the largest number of deaths from the NCDs with an estimated 17.8 million (95% CI = 17.5–18.0) deaths [3]. Smoking is a major risk factor for development of CVD and the cardio-metabolic traits, major biological risk factors of CVD [108, 109]. The cardio-metabolic traits include insulin resistance, impaired glucose tolerance, hypertension, intra-abdominal adiposity, and dyslipidaemia; defined as increased low-density lipoprotein (LDL), decreased high-density lipoprotein (HDL), and/or increased triglyceride concentrations [110, 111]. Substantial advances have been made in the diagnoses and treatment of these biological risk factors; nevertheless, their numbers continue to increase worldwide. Also, the exact molecular mechanisms linking smoking to the cardio-metabolic traits and CVD is still unclear. The understanding of this mechanism would provide a better insight into the disease etiology. This highlights the importance of more in-depth research investigating the underlying mechanisms of smoking that lead to these biological risk factors, and subsequently, to disease onset.

Over the recent years, great progress has been made in identifying the independent genetic markers associated with the cardio-metabolic traits in large consortiums [112-122]. Using hypothesis-free GWAS, changes in the DNA sequence have been found explaining a fraction of the variance in coronary artery disease and the cardio-metabolic traits [112-122]. Not all identified genetic variants affect protein sequences but can possibly affect gene regulation via regulatory mechanisms. In this line, changes in gene expression, DNA methylation levels, and miRNAs are also associated with the cardio-metabolic traits and smoking habits [82, 123-137]. Besides the great progress in identifying markers in the independent omics-fields, far less studies have investigated the integration of different omics layers. Multi-omics studies can limit passive correlations and provide a more comprehensive view of disease biology.

1.4.2 Missing variance explained for bone mineral density variation

Another highly prevalent NCD is osteoporosis, which is the most common bone disease affecting one in three women and one in five men above 50 years of age, with more than 200 million patients with osteoporotic hip fractures worldwide [138, 139]. Osteoporosis is characterized by reduced bone mass, disruption of bone micro-architectural, deterioration of bone tissue, with a subsequent increase in bone fragility, resulting in an increased risk of fractures [140, 141]. The decline in bone mass and prevalence of osteoporosis increase with age, especially in postmenopausal women [142]. Bone mineral density (BMD) measurements are used as a diagnostic proxy to assess osteoporosis risk in the clinical field [143]. BMD is the amount of bone mass per unit volume (volumetric density, g/cm^3) or per unit area (areal density, g/cm^2). Osteoporosis is diagnosed if the BMD measured by dual X-ray absorptiometry is more than 2.5 standard deviations below the age sex-matched mean [144, 145].

Osteoporosis is a highly heritable and complex polygenic disease. Twin and family studies reported high heritability (H^2 = 0.5–0.8) for both osteoporosis and BMD [146, 147]. Using GWAS, changes in the DNA sequence have been identified in association with BMD, including femoral neck, lumbar spine, and forearm BMD [148-150], as well as sex-specific associations of genetic variants with BMD [151, 152]. So far, 518 loci are identified in association with BMD, explaining 20% of its variance [153]. Further research is needed to investigate the additional of the explaining variance, possibly via epigenetic markers.

AIM OF THIS THESIS AND OUTLINE

The overall aim of this thesis is to investigate epigenetic mechanisms as possible biomarkers for disease risk, as a possible mediator between lifestyle factors and disease risk, and for inferring lifestyle factors from human materials (**Figure 5**).

Chapter 2 of this thesis investigates the possibility of developing lifestyle prediction models using DNA methylation markers (blue line in **Figure 5**). Specifically, **Chapter 2.1** focuses on validating previously published methods and markers that were used to develop a DNA methylation-based prediction model for alcohol consumption. **Chapter 2.2** aims to develop a DNA methylation-based prediction model for smoking habits using a finite set of CpGs. Specifically, I develop five prediction models able to establishes someone's smoking status, including 1) current *vs.* non-smokers, 2) current *vs.* former *vs.* never-smokers, 3) pack-years in current smokers, 4) cessation time in former smokers, and finally, a model that can predict 5) lifetime smoking habits including pack-years in current smokers *vs.* cessation time in former smokers *vs.* never smokers.

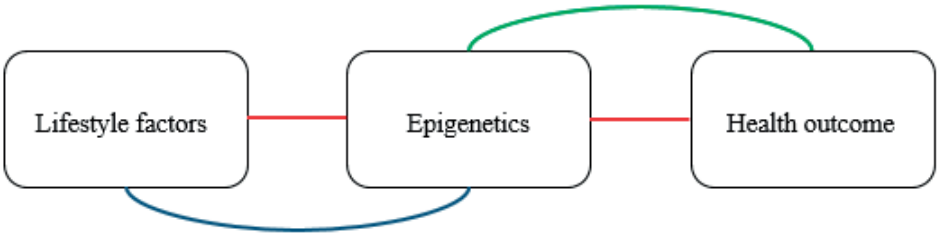


Figure 5. Overview of the included study aims in this thesis. In this thesis, I investigate DNA methylation markers for lifestyle inference (blue line). Also, I study the impact of smoking-related changes in DNA methylation and gene expression on cardio-metabolic traits (red lines). Moreover, I examine the role of epigenetic markers in relation to health outcomes (green line).

Chapter 3 aims to investigate the possibility of epigenetic alterations as a mechanism linking smoking status to cardio-metabolic traits (red lines in **Figure 5**). Specifically, I test the association between smoking-related changes in DNA methylation and gene expression and the association between these changes and cardio-metabolic traits.

Chapter 4 focuses on identifying miRNAs that are associated with health outcomes (green line in **Figure 5**). In **Chapter 4.1**, I use a multi-omics approach including previously published large-scale GWAS data, DNA methylation, and miRNA expression data to identify miRNAs associated with cardio-metabolic traits. In **Chapter 4.2**, I study the association between genetic variants in miRNA-related sequences and BMD using previously published large-scale GWAS data. Then, I investigate the potential target genes and pathways that may mediate the function of identified miRNAs in bone tissue and BMD.

Finally, in **Chapter 5**, I give an overview of the main findings of this thesis, discuss methodological issues, and examine the implications of the results.

STUDY POPULATION

Chapter 2.1 of this thesis used data from five cohort studies embedded in the Biobank-based Integrative Omics Study (BIOS) consortium [154]; the Rotterdam Study [155], a population-based prospective cohort study. The Rotterdam Study was initiated in 1990 and includes middle-aged and elderly participants living in the Ommoord district in Rotterdam, the Netherlands. In total, 14,926 participants were enrolled until 2008 during three separated recruitment periods; Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) [156], consisting of a selection of 547 subjects from a larger population-based cohort including participants of Caucasian descent of 40 years of age and older with an moderately increased risk of developing cardio-metabolic diseases; the Netherlands Twin Register (NTR) [157], established in 1987 including 52% of all Dutch twin-pairs born between 1987 and 2017 translating to an enrolment of around 120,000 twins and a roughly equal number of their relatives with a total of 255,729 registered participants

as of 2019; Leiden Longevity Study (LLS) [158], includes long-lived siblings of European descent together with their offspring and their offspring's partners including a total of 944 long-lived siblings from 421 families, together with 1,671 of their offspring and 744 partners; and Prospective ALS Study Netherlands (PAN) [159], a population-based study including patients of 15 year and older that diagnosed with suspected, possible, probable or definite ALS according to the El Escorial criteria and control samples with a total of 3,200 participants as of 2016. In addition, we used data from the Cooperative Health Research in the Region of Augsburg (KORA)-F4 study [160]. The KORA-F4 study (examined 2006-2008) is a seven-year follow-up study of the KORA-S4 survey (examined 1999-2001) conducted in 3,080 participants living in the region of Augsburg, Southern Germany. The external validation of this chapter used data from the Study of Health in Pomerania (SHIP)- Trend [161], the second cohort from SHIP including 4,420 participants aged 20 to 81 years at baseline in 2008; TwinsUK [162], established in 1992 to recruit monozygotic and dizygotic same-sex twins. All subjects are Caucasian females and ascertained to be free from severe disease at sample collected. In total, more than 13,000 twin participants between 16 to 98 years old are included from all regions across the United Kingdom; we included additional, non-overlapping participants from the Rotterdam study [155].

Chapter 2.2 used data from six cohort studies embedded in the Biobank-based Integrative Omics Study (BIOS) consortium [154]; the Rotterdam Study [155]; CODAM [156]; NTR [157]; LLS [158]; PAN [159]; and LifeLines DEEP [163], a sub-cohort of 1,461 participants from the LifeLines cohort [164]. The external model validation was conducted in data from the KORA-F4 study [160], SHIP-Trend [161], and the Generation R study [165], a population-based prospective birth-cohort study from fetal life onwards, conducted in Rotterdam. In total, 9,778 mothers living in Rotterdam with a delivery date from April 2002 until January 2006 were enrolled in the study.

Chapter 3 used data from the Rotterdam Study [155] in the discovery phase and data from the KORA-F4 study [160] in the replication phase.

Chapter 4.1 used GWAS summary statistics data from large consortiums, including from the Genetic Investigation of Anthropometric Traits (GIANT) consortium [112, 113], the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) [114-118], the Diabetes Genetics Replication And Meta-analysis (DIAGRAM) consortium [119], the Global Lipids Genetics Consortium (GLGC) [120], the Coronary ARtery DIsease Genome wide Replication and Meta-analysis plus The Coronary Artery Disease Genetics consortium (CARDIoGRM plusC4D) [121], and the International Consortium for Blood Pressure (ICBP) [122]. In addition, data from the Rotterdam study [155] were used.

Chapter 4.2 used GWAS summary statistics data from the GEnetic Factors for OSteoporosis (GEFOS) Consortium [166], data from the Rotterdam study [155], and data from 84 non-related postmenopausal ethnic Norwegian women (50–86 years) consecutively recruited at the Lovisenberg Deacon Hospital, the Out-patient Clinic, Oslo [167].

REFERENCES

1. Collaborators GBDRF. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396(10258):1223-49.
2. Beaglehole R, Bonita R, Horton R, Adams C, Alleyne G, Asaria P, et al. Priority actions for the non-communicable disease crisis. Lancet. 2011;377(9775):1438-47.
3. Collaborators GBDCoD. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736-88.
4. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007;128(4):669-81.
5. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol. 2017;18(1):238.
6. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.
7. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
8. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85-97.
9. Shastry BS. SNPs in disease gene mapping, medicinal drug development and evolution. Journal of Human Genetics. 2007;52(11):871-80.
10. Ball EV, Stenson PD, Abeysinghe SS, Krawczak M, Cooper DN, Chuzhanova NA. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat. 2005;26(3):205-13.
11. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005;6(2):95-108.
12. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152(6):1237-51.
13. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-3.
14. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
15. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489(7414):101-8.
16. Bird A. Perceptions of epigenetics. Nature. 2007;447(7143):396-8.
17. Wolffe AP, Matzke MA. Epigenetics: regulation through repression. Science. 1999;286(5439):481-6.
18. Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol. 2010;28(10):1057-68.
19. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. Science. 2001;293(5532):1089-93.
20. Kacem S, Feil R. Chromatin mechanisms in genomic imprinting. Mamm Genome. 2009;20(9-10):544-56.
21. Reik W, Lewis A. Co-evolution of X-chromosome inactivation and imprinting in mammals. Nat Rev Genet. 2005;6(5):403-10.
22. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215-33.
23. Ebert MS, Sharp PA. Roles for microRNAs in conferring robustness to biological processes. Cell. 2012;149(3):515-24.
24. Lee JT. Epigenetic regulation by long noncoding RNAs. Science. 2012;338(6113):1435-9.

25. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145-66.
26. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. Identification of mammalian microRNA host genes and transcription units. Genome Res. 2004;14(10A):1902-10.
27. Saini HK, Griffiths-Jones S, Enright AJ. Genomic analysis of human microRNA transcripts. Proc Natl Acad Sci U S A. 2007;104(45):17719-24.
28. Shukla GC, Singh J, Barik S. MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. Mol Cell Pharmacol. 2011;3(3):83-92.
29. Bohnsack MT, Czaplinski K, Gorlich D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. Rna. 2004;10(2):185-91.
30. Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U. Nuclear export of microRNA precursors. Science. 2004;303(5654):95-8.
31. Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, et al. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell. 2006;125(5):887-901.
32. Hutvágner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. Science. 2002;297(5589):2056-60.
33. Doench JG, Petersen CP, Sharp PA. siRNAs can function as miRNAs. Genes Dev. 2003;17(4):438-42.
34. Zeng Y, Yi R, Cullen BR. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. Proc Natl Acad Sci U S A. 2003;100(17):9779-84.
35. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281-97.
36. Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G, et al. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum Mutat. 2012;33(1):254-63.
37. Bushati N, Cohen SM. microRNA functions. Annu Rev Cell Dev Biol. 2007;23:175-205.
38. Ardekani AM, Naeini MM. The Role of MicroRNAs in Human Diseases. Avicenna J Med Biotechnol. 2010;2(4):161-79.
39. Lu J, Clark AG. Impact of microRNA regulation on variation in human gene expression. Genome Res. 2012;22(7):1243-54.
40. de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol. 2017;35(9):872-8.
41. Dole NS, Delany AM. MicroRNA variants as genetic determinants of bone mass. Bone. 2016;84:57-68.
42. Haas U, Sczakiel G, Laufer SD. MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3'-UTR via altered RNA structure. RNA Biol. 2012;9(6):924-37.
43. Mahen EM, Watson PY, Cottrell JW, Fedor MJ. mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS Biol. 2010;8(2):e1000307.
44. Huan T, Mendelson M, Joehanes R, Yao C, Liu C, Song C, et al. Epigenome-wide association study of DNA methylation and microRNA expression highlights novel pathways for human complex traits. Epigenetics. 2020;15(1-2):183-98.
45. Han L, Witmer PD, Casey E, Valle D, Sukumar S. DNA methylation regulates MicroRNA expression. Cancer Biol Ther. 2007;6(8):1284-8.
46. Groot M, Lee H. Sorting Mechanisms for MicroRNAs into Extracellular Vesicles and Their Associated Diseases. Cells. 2020;9(4).
47. Weber JA, Baxter DH, Zhang S, Huang DY, How Huang K, Jen Lee M, et al. The MicroRNA Spectrum in 12 Body Fluids. Clinical Chemistry. 2010;56(11):1733-41.
48. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-

based markers for cancer detection. Proc Natl Acad Sci U S A. 2008;105(30):10513-8.
49. de Planell-Saguer M, Rodicio MC. Detection methods for microRNAs in clinic practice. Clin Biochem. 2013;46(10-11):869-78.
50. Ghanbari M, Ikram MA, de Looper HWJ, Hofman A, Erkeland SJ, Franco OH, et al. Genome-wide identification of microRNA-related variants associated with risk of Alzheimer's disease. Sci Rep. 2016;6:28387.
51. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68-73.
52. Nam JW, Kim J, Kim SK, Zhang BT. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Res. 2006;34(Web Server issue):W455-8.
53. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11.
54. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38(1):23-38.
55. Wu H, Zhang Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell. 2014;156(1-2):45-68.
56. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288-95.
57. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529-41.
58. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7.
59. Peterson K. Biomarkers for alcohol use and abuse--a summary. Alcohol Res Health. 2004;28(1):30-7.
60. Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, Koren G. Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. Ther Drug Monit. 2009;31(1):14-30.
61. Benowitz NL. Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiol Rev. 1996;18(2):188-204.
62. Andresen-Streichert H, Müller A, Glahn A, Skopp G, Sterneck M. Alcohol Biomarkers in Clinical and Forensic Contexts. Dtsch Arztebl Int. 2018;115(18):309-15.
63. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011;88(4):450-7.
64. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6(1):4.
65. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22(5):843-51.
66. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One. 2013;8(5):e63812.
67. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environ Health Perspect. 2014;122(7):673-8.
68. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382-96.

69. Allione A, Marcon F, Fiorito G, Guarrera S, Siniscalchi E, Zijno A, et al. Novel epigenetic changes unveiled by monozygotic twins discordant for smoking habits. PLoS One. 2015;10(6):e0128265.
70. Besingi W, Johansson A. Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet. 2014;23(9):2290-7.
71. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.
72. Sayols-Baixeras S, Lluís-Ganella C, Subirana I, Salas LA, Vilahur N, Corella D, et al. Corrigendum. Identification of a new locus and validation of previously reported loci showing differential methylation associated with smoking. The REGICOR study. Epigenetics. 2016;11(2):174.
73. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8(5):599-618.
74. Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2012;120(10):1425-31.
75. Zhu X, Li J, Deng S, Yu K, Liu X, Deng Q, et al. Genome-Wide Analysis of DNA Methylation and Cigarette Smoking in a Chinese Population. Environ Health Perspect. 2016;124(7):966-73.
76. Prince C, Hammerton G, Taylor AE, Anderson EL, Timpson NJ, Davey Smith G, et al. Investigating the impact of cigarette smoking behaviours on DNA methylation patterns in adolescence. Hum Mol Genet. 2019;28(1):155-65.
77. de Vries M, van der Plaat DA, Nedeljkovic I, Verkaik-Schakel RN, Kooistra W, Amin N, et al. From blood to lung tissue: effect of cigarette smoke on DNA methylation and lung function. Respir Res. 2018;19(1):212.
78. Su D, Wang X, Campbell MR, Porter DK, Pittman GS, Bennett BD, et al. Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes. PLoS One. 2016;11(12):e0166486.
79. Wilson R, Wahl S, Pfeiffer L, Ward-Caviness CK, Kunze S, Kretschmer A, 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):805.
80. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet. 2012;21(13):3073-82.
81. Sun YV, Smith AK, Conneely KN, Chang Q, Li W, Lazarus A, et al. Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum Genet. 2013;132(9):1027-37.
82. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;9(5):436-47.
83. Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, et al. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology. 2013;24(5):712-6.
84. Philibert R, Hollenbeck N, Andersen E, Osborn T, Gerrard M, Gibbons FX, et al. A quantitative epigenetic approach for the assessment of cigarette consumption. Front Psychol. 2015;6:656.
85. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res. 2016;146:395-403.

86. Kondratyev N, Golov A, Alfimova M, Lezheiko T, Golimbet V. Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation. Clin Epigenetics. 2018;10(1):130.
87. Sugden K, Hannon EJ, Arseneault L, Belsky DW, Broadbent JM, Corcoran DL, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Transl Psychiatry. 2019;9(1):92.
88. Endo K, Li J, Nakanishi M, Asada T, Ikesue M, Goto Y, et al. Establishment of the MethyLight Assay for Assessing Aging, Cigarette Smoking, and Alcohol Consumption. Biomed Res Int. 2015;2015:451981.
89. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136.
90. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
91. Zhang R, Miao Q, Wang C, Zhao R, Li W, Haile CN, et al. Genome-wide DNA methylation analysis in alcohol dependence. Addict Biol. 2013;18(2):392-403.
92. Zhao R, Zhang R, Li W, Liao Y, Tang J, Miao Q, et al. Genome-wide DNA methylation patterns in discordant sib pairs with alcohol dependence. Asia Pac Psychiatry. 2013;5(1):39-50.
93. Philibert RA, Penaluna B, White T, Shires S, Gunter T, Liesveld J, et al. A pilot examination of the genome-wide DNA methylation signatures of subjects entering and exiting short-term alcohol dependence treatment programs. Epigenetics. 2014;9(9):1212-9.
94. Philibert RA, Plume JM, Gibbons FX, Brody GH, Beach SR. The impact of recent alcohol use on genome wide DNA methylation signatures. Front Genet. 2012;3:54.
95. Weng JT, Wu LS, Lee CS, Hsu PW, Cheng AT. Integrative epigenetic profiling analysis identifies DNA methylation changes associated with chronic alcohol consumption. Comput Biol Med. 2015;64:299-306.
96. Xu K, Montalvo-Ortiz JL, Zhang X, Southwick SM, Krystal JH, Pietrzak RH, et al. Epigenome-Wide DNA Methylation Association Analysis Identified Novel Loci in Peripheral Cells for Alcohol Consumption Among European American Male Veterans. Alcohol Clin Exp Res. 2019;43(10):2111-21.
97. Lu M, Xueying Q, Hexiang P, Wenjing G, Hägg S, Weihua C, et al. Genome-wide associations between alcohol consumption and blood DNA methylation: evidence from twin study. Epigenomics. 2021;13(12):939-51.
98. Lohoff FW, Roy A, Jung J, Longley M, Rosoff DB, Luo A, et al. Epigenome-wide association study and multi-tissue replication of individuals with alcohol use disorder: evidence for abnormal glucocorticoid signaling pathway gene regulation. Mol Psychiatry. 2020.
99. Witt SH, Frank J, Frischknecht U, Treutlein J, Streit F, Foo JC, et al. Acute alcohol withdrawal and recovery in men lead to profound changes in DNA methylation profiles: a longitudinal clinical study. Addiction. 2020;115(11):2034-44.
100. Hagerty SL, Bidwell LC, Harlaar N, Hutchison KE. An Exploratory Association Study of Alcohol Use Disorder and DNA Methylation. Alcohol Clin Exp Res. 2016;40(8):1633-40.
101. Philibert R, Dogan M, Noel A, Miller S, Krukow B, Papworth E, et al. Genome-wide and digital polymerase chain reaction epigenetic assessments of alcohol consumption. Am J Med Genet B Neuropsychiatr Genet. 2018;177(5):479-88.
102. Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23(2):422-33.

103. Jeffreys AJ, Wilson V, Thein SL. Hypervariable 'minisatellite' regions in human DNA. Nature. 1985;314(6006):67-73.
104. Jeffreys AJ, Wilson V, Thein SL. Individual-specific 'fingerprints' of human DNA. Nature. 1985;316(6023):76-9.
105. Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. Nat Rev Genet. 2004;5(10):739-51.
106. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Sci Int Genet. 2017;28:225-36.
107. Xavier C, de la Puente M, Mosquera-Miguel A, Freire-Aradas A, Kalamara V, Vidaki A, et al. Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA. Forensic Sci Int Genet. 2020;48:102336.
108. Collaborators GBDRF. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1345-422.
109. Sun K, Liu J, Ning G. Active smoking and risk of metabolic syndrome: a meta-analysis of prospective studies. PLoS One. 2012;7(10):e47791.
110. Mottillo S, Filion KB, Genest J, Joseph L, Pilote L, Poirier P, et al. The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis. J Am Coll Cardiol. 2010;56(14):1113-32.
111. Wilson PW, D'Agostino RB, Parise H, Sullivan L, Meigs JB. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. Circulation. 2005;112(20):3066-72.
112. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.
113. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Magi R, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature. 2015;518(7538):187-96.
114. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012;44(6):659-69.
115. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet. 2010;42(2):142-8.
116. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes. 2011;60(10):2624-34.
117. Wheeler E, Leong A, Liu CT, Hivert MF, Strawbridge RJ, Podmore C, et al. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med. 2017;14(9):e1002383.
118. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42(2):105-16.
119. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66(11):2888-902.
120. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with

lipid levels. Nat Genet. 2013;45(11):1274-83.
121. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47(10):1121-30.
122. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9.
123. Kaur G, Begum R, Thota S, Batra S. A systematic review of smoking-related epigenetic alterations. Arch Toxicol. 2019;93(10):2715-40.
124. Huan T, Joehanes R, Schurmann C, Schramm K, Pilling LC, Peters MJ, et al. A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. Hum Mol Genet. 2016;25(21):4611-23.
125. Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. Clin Epigenetics. 2017;9:15.
126. Richard MA, Huan T, Ligthart S, Gondalia R, Jhun MA, Brody JA, et al. DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. Am J Hum Genet. 2017;101(6):888-902.
127. Huan T, Esko T, Peters MJ, Pilling LC, Schramm K, Schurmann C, et al. A meta-analysis of gene expression signatures of blood pressure and hypertension. PLoS Genet. 2015;11(3):e1005035.
128. Liu J, Carnero-Montoro E, van Dongen J, Lent S, Nedeljkovic I, Ligthart S, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. Nat Commun. 2019;10(1):2581.
129. Chen BH, Hivert MF, Peters MJ, Pilling LC, Hogan JD, Pham LM, et al. Peripheral Blood Transcriptomic Signatures of Fasting Glucose and Insulin Concentrations. Diabetes. 2016;65(12):3794-804.
130. Dhana K, Braun KVE, Nano J, Voortman T, Demerath EW, Guan W, et al. An Epigenome-Wide Association Study of Obesity-Related Traits. Am J Epidemiol. 2018;187(8):1662-9.
131. Nikpay M, Beehler K, Valsesia A, Hager J, Harper ME, Dent R, et al. Genome-wide identification of circulating-miRNA expression quantitative trait loci reveals the role of several miRNAs in the regulation of cardiometabolic phenotypes. Cardiovasc Res. 2019;115(11):1629-45.
132. Stauffer BL, Russell G, Nunley K, Miyamoto SD, Sucharov CC. miRNA expression in pediatric failing human heart. J Mol Cell Cardiol. 2013;57:43-6.
133. Rotllan N, Price N, Pati P, Goedeke L, Fernández-Hernando C. microRNAs in lipoprotein metabolism and cardiometabolic disorders. Atherosclerosis. 2016;246:352-60.
134. Wang ZH, Sun XY, Li CL, Sun YM, Li J, Wang LF, et al. miRNA-21 Expression in the Serum of Elderly Patients with Acute Myocardial Infarction. Med Sci Monit. 2017;23:5728-34.
135. Ovchinnikova ES, Schmitter D, Vegter EL, Ter Maaten JM, Valente MA, Liu LC, et al. Signature of circulating microRNAs in patients with acute heart failure. Eur J Heart Fail. 2016;18(4):414-23.
136. Zhou SS, Jin JP, Wang JQ, Zhang ZG, Freedman JH, Zheng Y, et al. miRNAS in cardiovascular diseases: potential biomarkers, therapeutic targets and challenges. Acta Pharmacol Sin. 2018;39(7):1073-84.
137. Kaur G, Begum R, Thota S, Batra S. A systematic review of smoking-related epigenetic alterations. Archives of Toxicology. 2019;93(10):2715-40.
138. Cooper C, Campion G, Melton LJ, 3rd. Hip fractures in the elderly: a world-wide projection. Osteoporos Int. 1992;2(6):285-9.

139. Rivadeneira F, Makitie O. Osteoporosis and Bone Mass Disorders: From Gene Pathways to Treatments. Trends Endocrinol Metab. 2016;27(5):262-81.
140. Hernlund E, Svedbom A, Ivergard M, Compston J, Cooper C, Stenmark J, et al. Osteoporosis in the European Union: medical management, epidemiology and economic burden. A report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA). Arch Osteoporos. 2013;8:136.
141. Nih Consensus Development Panel on Osteoporosis Prevention D, Therapy. Osteoporosis prevention, diagnosis, and therapy. Jama. 2001;285(6):785-95.
142. Cummings SR, Melton LJ. Epidemiology and outcomes of osteoporotic fractures. Lancet. 2002;359(9319):1761-7.
143. Blake GM, Fogelman I. The role of DXA bone density scans in the diagnosis and treatment of osteoporosis. Postgrad Med J. 2007;83(982):509-17.
144. WHO. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Geneva: WHO, 1994.
145. Sözen T, Özışık L, Başaran N. An overview and management of osteoporosis. Eur J Rheumatol. 2017;4(1):46-56.
146. Ralston SH, Uitterlinden AG. Genetics of osteoporosis. Endocr Rev. 2010;31(5):629-62.
147. Al-Barghouthi BM, Farber CR. Dissecting the Genetics of Osteoporosis using Systems Approaches. Trends Genet. 2019;35(1):55-67.
148. Rivadeneira F, Styrkarsdottir U, Estrada K, Halldorsson BV, Hsu YH, Richards JB, et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. Nat Genet. 2009;41(11):1199-206.
149. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, Ntzani EE, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat Genet. 2012;44(5):491-501.
150. Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature. 2015;526(7571):112-7.
151. Karasik D, Ferrari SL. Contribution of gender-specific genetic factors to osteoporosis risk. Ann Hum Genet. 2008;72(Pt 5):696-714.
152. Naganathan V, Macgregor A, Snieder H, Nguyen T, Spector T, Sambrook P. Gender differences in the genetic factors responsible for variation in bone density and ultrasound. J Bone Miner Res. 2002;17(4):725-33.
153. Morris JA, Kemp JP, Youlten SE, Laurent L, Logan JG, Chai RC, et al. An atlas of genetic influences on osteoporosis in humans and mice. Nat Genet. 2019;51(2):258-66.
154. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131-8.
155. Ikram MA, Brusselle G, Ghanbari M, Goedegebure A, Ikram MK, Kavousi M, et al. Objectives, design and main findings until 2020 from the Rotterdam Study. Eur J Epidemiol. 2020;35(5):483-517.
156. van Greevenbroek MM, Jacobs M, van der Kallen CJ, Vermeulen VM, Jansen EH, Schalkwijk CG, et al. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). Eur J Clin Invest. 2011;41(4):372-9.
157. Ligthart L, van Beijsterveldt CEM, Kevenaar ST, de Zeeuw E, van Bergen E, Bruins S, et al. The Netherlands Twin Register: Longitudinal Research Based on Twin and

Twin-Family Designs. Twin Res Hum Genet. 2019:1-14.

158. Schoenmaker M, de Craen AJ, de Meijer PH, Beekman M, Blauw GJ, Slagboom PE, et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. Eur J Hum Genet. 2006;14(1):79-84.

159. Huisman MH, de Jong SW, van Doormaal PT, Weinreich SS, Schelhaas HJ, van der Kooi AJ, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. J Neurol Neurosurg Psychiatry. 2011;82(10):1165-70.

160. Holle R, Happich M, Lowel H, Wichmann HE, Group MKS. KORA--a research platform for population based health research. Gesundheitswesen. 2005;67 Suppl 1:S19-25.

161. Volzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, et al. Cohort profile: the study of health in Pomerania. Int J Epidemiol. 2011;40(2):294-307.

162. Moayyeri A, Hammond CJ, Valdes AM, Spector TD. Cohort Profile: TwinsUK and healthy ageing twin study. Int J Epidemiol. 2013;42(1):76-85.

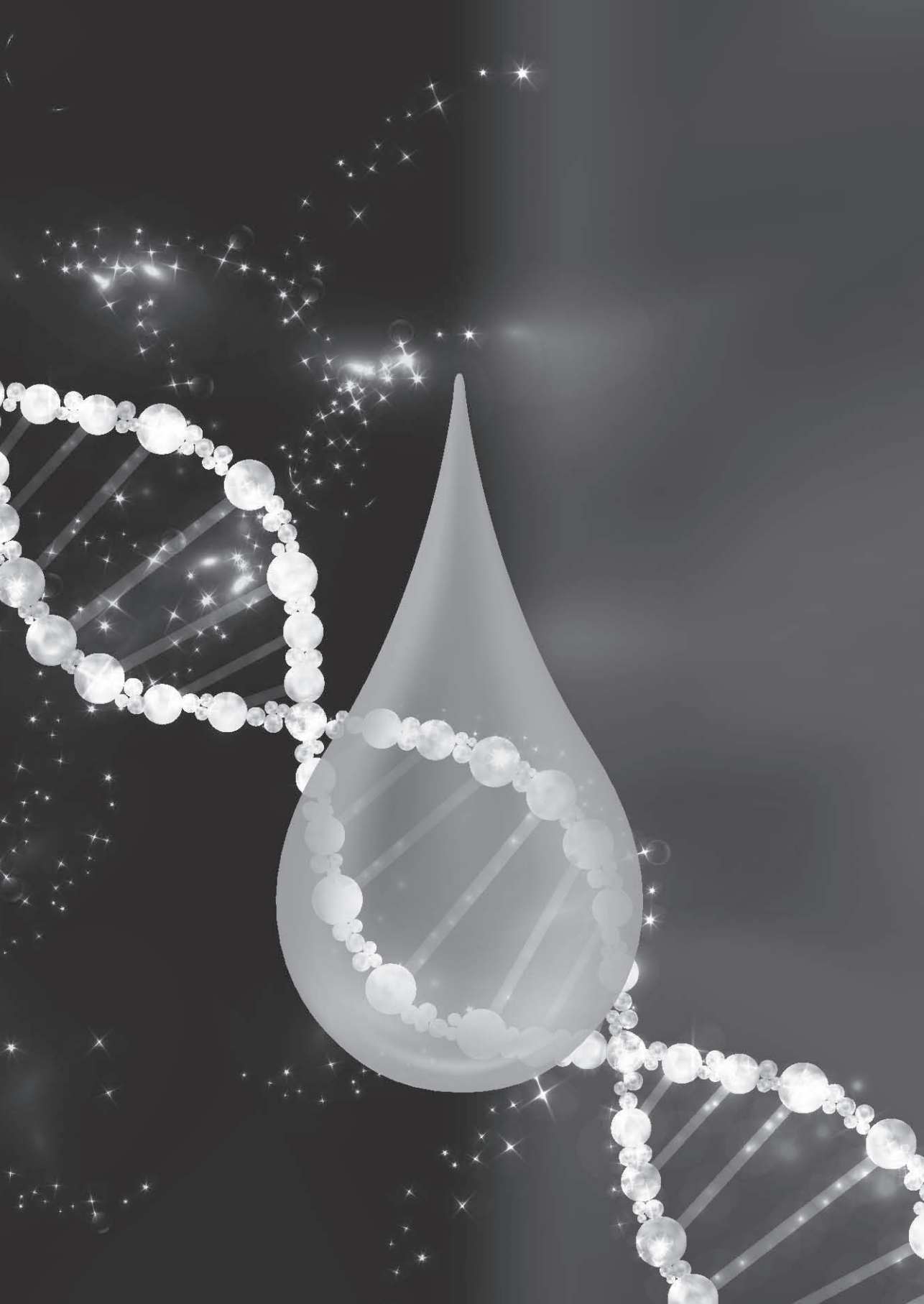
163. Tigchelaar EF, Zhernakova A, Dekens JA, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015;5(8):e006772.

164. Scholtens S, Smidt N, Swertz MA, Bakker SJ, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol. 2015;44(4):1172-80.

165. Kooijman MN, Kruithof CJ, van Duijn CM, Duijts L, Franco OH, van IMH, et al. The Generation R Study: design and cohort update 2017. Eur J Epidemiol. 2016;31(12):1243-64.

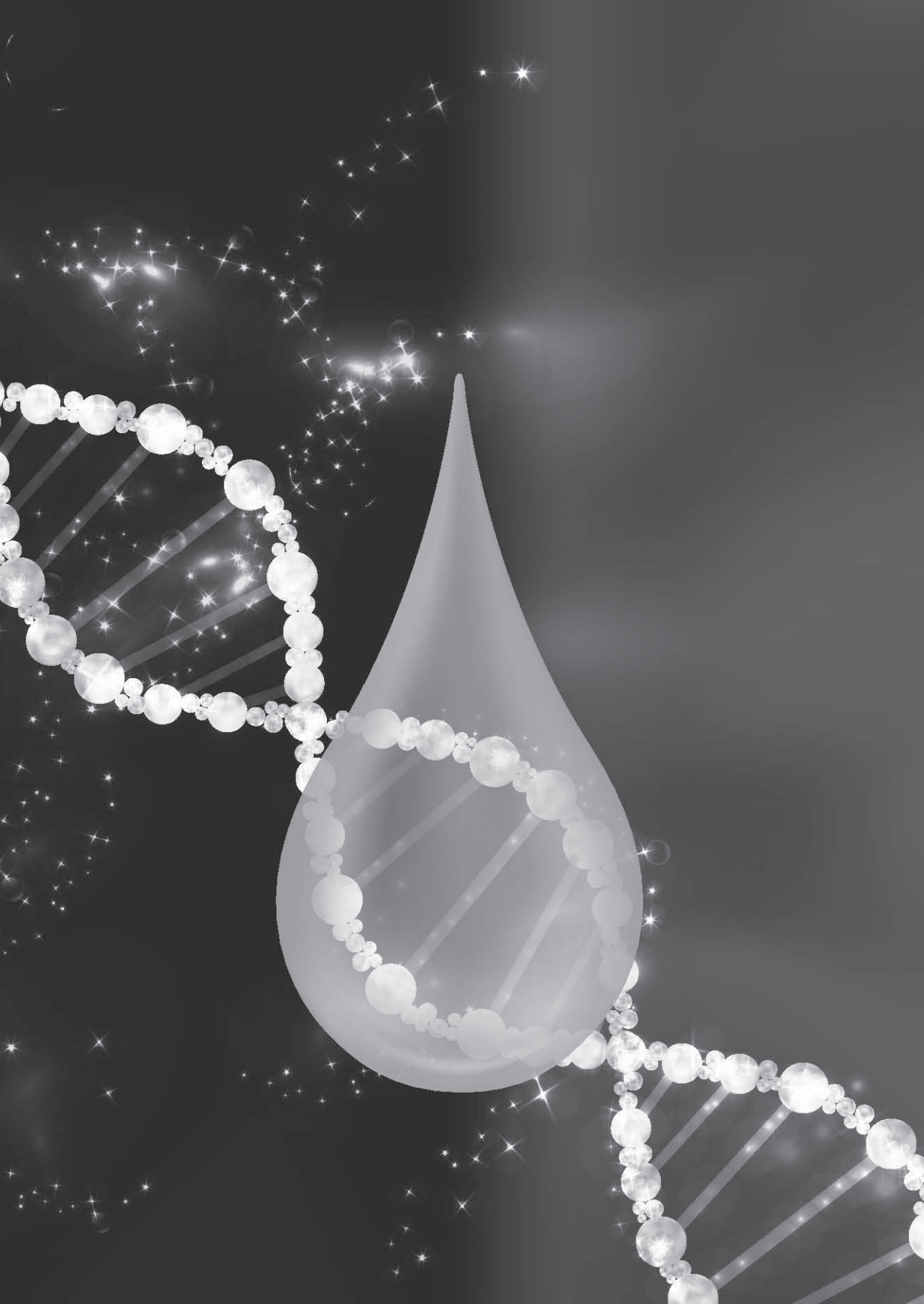
166. Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature. 2015;526(7571):112-7.

167. Reppe S, Refvem H, Gautvik VT, Olstad OK, Høvring PI, Reinholt FP, et al. Eight genes are highly associated with BMD variation in postmenopausal Caucasian women. Bone. 2010;46(3):604-12.



Chapter 2

Lifestyle factor inference from DNA methylation



Chapter 2.1

Validating biomarkers and models for epigenetic inference of alcohol consumption from blood

Silvana C.E. Maas, Athina Vidaki, Alexander Teumer, Ricardo Costeira, Rory Wilson, Jenny van Dongen, Marian Beekman, Uwe Völker, Hans J. Grabe, Sonja Kunze, Karl-Heinz Ladwig, Joyce B.J. van Meurs, André G. Uitterlinden, Trudy Voortman, Dorret I. Boomsma, P. Eline Slagboom, Diana van Heemst, Carla J.H. van der Kallen, Leonard H. van den Berg, Melanie Waldenberger, Henry Völzke, Annette Peters, Jordana T. Bell, M. Arfan Ikram, Mohsen Ghanbari*, Manfred Kayser*

Clinical Epigenetics 2021;13(1):198

***Denotes equal contribution**

ABSTRACT

Background: Information on long-term alcohol consumption is relevant for medical and public health research, disease therapy, and other areas. Recently, DNA methylation-based inference of alcohol consumption from blood was reported with high accuracy, but these results were based on employing the same dataset for model training and testing, which can lead to accuracy overestimation. Moreover, only subsets of alcohol consumption categories were used, which makes it impossible to extrapolate such models to the general population. By using data from eight population-based European cohorts (N=4677), we internally and externally validated the previously reported biomarkers and models for epigenetic inference of alcohol consumption from blood and developed new models comprising all data from all categories.

Results: By employing data from six European cohorts (N=2883), we empirically tested the reproducibility of the previously suggested biomarkers and prediction models via ten-fold internal cross-validation. In contrast to previous findings, all seven models based on 144-CpGs yielded lower mean AUCs compared to the models with less CpGs. For instance, the 144-CpG heavy *versus* non-drinkers model gave an AUC of 0.78±0.06, while the 5 and 23 CpG models achieved 0.83±0.05, respectively. The transportability of the models was empirically tested via external validation in three independent European cohorts (N=1794), revealing high AUC variance between datasets within models. For instance, the 144-CpG heavy *versus* non-drinkers model yielded AUCs ranging from 0.60 to 0.84 between datasets. The newly developed models that considered data from all categories showed low AUCs but gave low AUC variation in the external validation. For instance, the 144-CpG heavy and at-risk *versus* light and non-drinkers model achieved AUCs of 0.67±0.02 in the internal cross-validation and 0.61-0.66 in the external validation datasets.

Conclusions: The outcomes of our internal and external validation demonstrate that the previously reported prediction models suffer from both overfitting and accuracy overestimation. Our results show that the previously proposed biomarkers are not yet sufficient for accurate and robust inference of alcohol consumption from blood. Overall, our findings imply that DNA methylation prediction biomarkers and models need to be improved considerably before epigenetic inference of alcohol consumption from blood can be considered for practical applications.

INTRODUCTION

Alcohol consumption is a modifiable lifestyle factor associated with morbidity and mortality worldwide [1]. It was estimated to be the seventh-leading risk factor for disability-adjusted life-years (DALYs) and deaths in 2016, accounting for 5.2% (95% CI 4.4–6.0) of deaths globally [1]. Various diseases are caused or strongly influenced by excessive alcohol consumption, often in a dose-dependent manner, such as different forms of cancer, various liver diseases, cardiovascular disease, epilepsy, and unipolar depressive disorder [2].

Recent alcohol consumption is detectable by breathalyzers or direct measurement of the alcohol concentration in blood and urine; however, such measurements only provide information on few hours since the last alcohol consumption. For example, ethanol can be detected in urine within ten to twelve hours after the last drink, but not later [3]. Blood-based toxicological tests for alcohol consumption are also available, which are based on direct or indirect biomarkers. A direct biomarker is the result from ethanol metabolism or its reaction with other substances in the body, including ethyl glucuronide (EtG), ethyl sulfate (EtS), and phospholipid phosphatidylethanol (PEth). Indirect biomarkers are derived from cellular processes that undergo changes as a response to alcohol consumption, including carbohydrate-deficient transferrin (CDT), mean corpuscular volume (MCV), aspartate-aminotransferase (AST), alanine aminotransferase (ALT), and gamma-glutamyl transferase activity (GGT) [4, 5]. It is important to note that these direct and indirect biomarkers are specifically useful to determine the extreme categories, including excessive alcohol consumption or abstinence, and for recent alcohol consumption [5]. For example, CDT can distinguish excessive alcohol consumption of on average >50–80 gram ethanol per day over a period of 2 weeks [4]. In contrast, there are no reliable biomarkers available that can determine overall alcohol consumption habits like to distinguish heavy and at-risk drinkers from light and non-drinkers, or drinkers from non-drinkers and that are informative for alcohol consumption for longer periods of time. Therefore, due to the limited progress in previous alcohol biomarker research, information on long-term alcohol consumption is typically still collected using self-reports, although they are known to be unreliable [6]. Accurate and reliable biomarkers that reflects habitual alcohol consumption over months and years are needed to better diagnose and treat alcohol-related diseases and for objective exposure assessment in studies on alcohol consumption and health [7].

DNA methylation has been proposed as a biomarker for the detection of lifestyle factors in general [8] and several studies have already shown that alcohol consumption is associated with changes in DNA methylation levels in particular [9-12]. A few studies have also explored the possibility of epigenetic inference of alcohol consumption from blood [12-15]. A large benefit from epigenetic-based inference is the increasing

availability of DNA methylation information in study participants, as DNA methylation is extensively studied for its association with diseases. The most extensive study investigating the epigenetic association and inference of alcohol consumption was done by Liu *et al.* [12]. In this study, an epigenome-wide association study (EWAS) meta-analysis on alcohol consumption was conducted in 9643 individuals of European ancestry from blood-derived DNA [12]. The authors identified 363 CpGs significantly associated ($P<1\times10^{-7}$) with alcohol consumption levels used as a continuous variable (grams/day). A meta-analysis was performed for prediction marker discovery in a subset of 6926 participants of European ancestry, which identified 361 CpGs ($P<5\times10^{-6}$). The study also reports impressively high prediction accuracies, expressed as area under the curve (AUC) estimates, for DNA methylation-based prediction models for categorical alcohol consumption based on sets of 5, 23, 78, or 144 CpG markers plus age, sex, and BMI. These models include pairwise combinations of four alcohol consumption categories with the highest AUC obtained for the models with the extreme categories. For instance, the reported 144-CpG model showed discrimination of heavy drinkers *versus* (*vs.*) non-drinkers with an AUC of 0.91-1.0 (an AUC of 1.0 means completely accurate inference) in the discovery dataset and all four replication cohorts as well as 0.86-1.0 for heavy drinkers *vs.* light drinkers [12]. The authors demonstrated increase in AUC with increased number of CpG predictors included in the models.

The high prediction accuracies reported by Liu *et al.* [12] were questioned based on methodological grounds by Hattab *et al.* [16]. Liu *et al.* were particularly criticized for not having used the coefficients from the discovery dataset to determine prediction accuracies in the replication datasets, but instead, they re-estimated these coefficients in each replication cohort using the same dataset for model training and testing. Hattab *et al.* [16] concluded that the prediction accuracies published by Liu *et al.* represent overestimates. However, Hattab *et al.* based their conclusions entirely on simulated data instead of empirical data. In a subsequent study, Yousefi *et al.* [17] found only half of the alcohol consumption variance explained by the DNA methylation markers in their independent data, compared to the explained variance values reported by Liu *et al.* [12]. In addition, Yousefi *et al.* [17] generated DNA methylation-derived scores using the coefficients made available by Liu *et al.*; based on these coefficients, they obtained much lower AUCs for the same models as reported by Liu *et al*. For instance, for adults at midlife, the reported AUCs were between 0.48 to 0.57 for distinguishing heavy drinkers from non-drinkers and AUCs between 0.55 to 0.57 for heavy drinkers *vs*. light drinkers. Although the Yousefi *et al.* study used empirical data, an important limitation of the study is their relatively small sample size, comprising only of 14 heavy drinkers, 67 at-risk drinkers, 748 light drinkers, and 54 non-drinkers.

Another source for the Liu *et al.* [12] AUCs putatively reflecting overestimations is their use of category subsets, and therewith participant subsets, in their prediction modeling

approach. For instance, for estimating AUC for heavy drinkers *vs.* non-drinkers, Liu *et al.* only used data from heavy and non-drinkers thereby excluding the data from light drinkers and at-risk drinkers. Such use of partial data in prediction modeling is expected to result in overestimated prediction outcomes compared to a model that would include all available categories. Moreover, models that exclude participants based on their non-considered categories cannot be applied to the general populations where people with the excluded categories exist but can never be inferred correctly because their category was excluded from the model.

In the current study, we firstly aimed at replicating the association between alcohol consumption and the 363 CpGs previously identified by Liu *et al.* [12], using data from 2042 independent participants from five cohorts [18-22]. Then, by using a total of 4677 individuals from eight European cohorts [18-25], we aimed to thoroughly validate the DNA methylation biomarker sets and prediction models for the epigenetic inference of alcohol consumption from blood previously used by Liu *et al*. [12]. In addition, we trained and validated two new models including all alcohol consumption categories.

RESULTS

Study populations and data sets

For replicating the association between alcohol consumption and the 363 CpGs previously reported by Liu *et al.* [12], we used data from 2042 individuals of five European cohorts as part of the Biobank-based Integrative Omics Study (BIOS) consortium [18-22, 26].

For prediction model building and internal validation, we employed a total dataset of 2883 Europeans, including the 2042 individuals from the BIOS consortium [26] together with 841 participants from The Cooperative Health Research in the Region of Augsburg (KORA) study (F4) [23]. Only participants with complete alcohol consumption data and DNA methylation data of all 144 predictive CpGs were included. Notably, there is no overlap between these data and those used by Liu *et al*. [12] in their prediction marker discovery EWAS. This makes our model building dataset completely independent from that of Liu *et al*. The KORA data included here were previously used by Liu *et al.* for prediction replication analysis; thus, its use for model building here provides no data dependency problem.

For external validation, we applied data from three European cohorts not applied for model training and internal validation: i.e., participants from the Rotterdam Study (sub-cohort RS-III-1) [18] (N=648) not included in the BIOS consortium, from the Study of Health in Pomerania (SHIP)-Trend cohort (N=433) [24], and two datasets from the TwinsUK Study, TwinsUK (N=713) and TwinsUK2 (N=442) [25]. The TwinsUK2 (N=442) dataset

comprises a subset of the TwinsUK (N=713) participants but with re-processed DNA methylation dataset and a different alcohol consumption collection method (**Additional file 1**: Supplementary Methods). Of note, the TwinsUK and RS-III-1 data were previously used by Liu *et al.* [12] in their prediction marker discovery EWAS that identified the 361 associated CpGs ($P<5\times10^{-6}$). Testing the inference ability of these 361 alcohol associated CpGs by Liu *et al*. was solely conducted in the Framingham Heart Study data [27], which identified the 5, 23, 78, and 144 CpG marker sets used for prediction modelling by Liu *et al.* and therefore here as well. However, since these data were used in the initial marker discovery EWAS, we cannot exclude an overestimation effect in our prediction accuracy estimates obtained from these two cohorts (see below).

An overview of the datasets included in each analysis step of our study is provided in **Figure 1**, their characteristics are summarized in **Table 1** and described in detail in **Additional file 1**: Supplementary Methods.

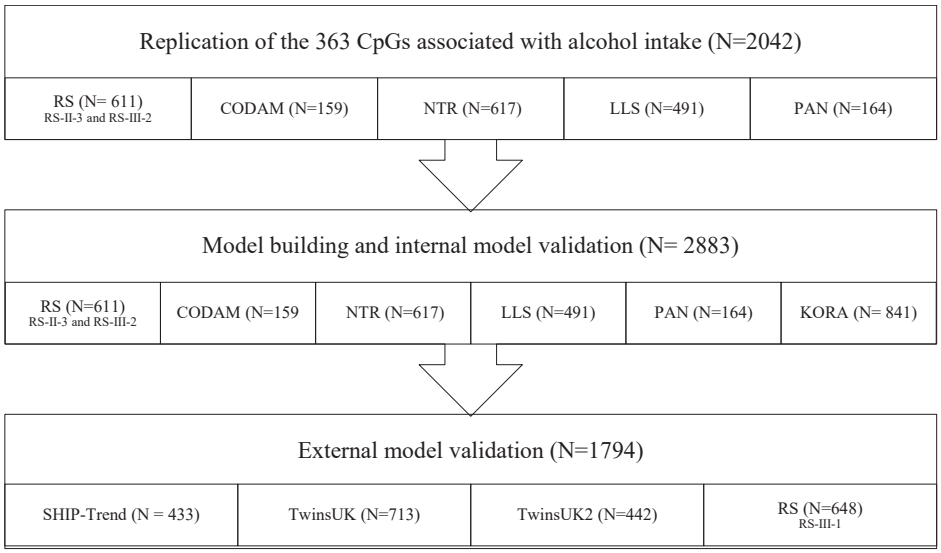


Figure 1. Use of study populations in each analysis. The 363 alcohol-associated CpGs previously identified by Liu et al. were replicated using data from 2042 participants of five cohorts studies embedded within the BIOS consortium. An additional 841 participants from the KORA F4 study were combined with these 2042 participants and together comprises our model building dataset. The model building dataset was used to train the prediction models and to test the reproducibility of the prediction models via internal cross-validation. The transportability of the models was tested in the external validation phase based on 1794 participants from three cohorts that were independent from the data used for model building and internal validation. Abbreviations: CODAM, Cohort on Diabetes and Atherosclerosis Maastricht; KORA, Cooperative Health Research in the Region of Augsburg study; LLS, Leiden Longevity Study; NTR, Netherlands Twin Register; PAN, Prospective ALS Study Netherlands; RS, Rotterdam Study; SHIP-Trend, Study of Health in Pomerania-Trend; TwinsUK- The TwinsUK Study; TwinsUK2- Subset of the TwinsUK Study.

Table 1: Dataset characteristics used in model building, internal and external validation.

	N	Age (years), mean (SD)	Men (%)	BMI mean (SD)	Alcohol gr/ day, Median (min, max)	Non-drinkers (%)	Light drinkers (%)	At-risk drinkers (%)	Heavy drinkers (%)
Model building and internal validation dataset									
RS-II-3/III-2	611	67 (6)	275 (45)	27.8 (4)	8.6 (1, 57)	0 (0)	545 (89)	52 (9)	14 (2)
CODAM	159	66 (7)	86 (54)	28.9 (4)	7.9 (0, 72)	12 (8)	117 (74)	23 (14)	7 (4)
NTR	617	39 (14)	188 (31)	24.6 (4)	5.1 (0, 69)	195 (32)	348 (56)	44 (7)	30 (5)
LLS	491	58 (6)	231 (47)	25.3 (3)	13.0 (0, 90)	36 (7)	309 (63)	98 (20)	49 (10)
PAN	164	62 (9)	100 (61)	26.0 (4)	11.0 (0, 77)	1 (1)	127 (77)	20 (12)	16 (10)
KORA F4	841	61 (9)	415 (49)	28.0 (5)	7.6 (0, 150)	251 (30)	354 (42)	133 (16)	103 (12)
Total dataset	*2883*	*57 (14)*	*1295 (45)*	*26.7 (4)*	*8.0 (0, 150)*	*495 (17)*	*1800 (62)*	*370 (13)*	*218 (8)*
External validation datasets									
SHIP-Trend	433	51 (14)	205 (47)	27.2 (4.1)	3.6 (0, 82)	47 (11)	346 (80)	28 (6)	12 (3)
TwinsUK	713	58 (10)	0 (0)	26.7 (5)	2.3 (0, 101)	187 (26)	423 (59)	67 (9)	36 (5)
TwinsUK2	442	59 (9)	0 (0)	26.6 (5)	5.3 (0, 94)	36 (8)	311 (70)	46 (10)	49 (11)
RS-III-1	648	59.6 (8)	298 (46)	27.7 (5)	6.4 (0, 57)	64 (10)	495 (76)	79 (12)	10 (2)

The total model building dataset was also used for internal ten-fold cross-validation. Abbreviations: BMI- body mass index; CODAM- Cohort on Diabetes and Atherosclerosis Maastricht; KORA F4- The Cooperative Health Research in the Region of Augsburg study; LLS- Leiden Longevity Study; NTR- Netherlands Twin Register; PAN- Prospective ALS Study Netherlands; RS- Rotterdam Study; SD- standard deviation; SHIP- Study of Health in Pomerania-Trend cohort; TwinsUK- The TwinsUK Study; TwinsUK2- Subset of the TwinsUK Study. The alcohol categories were defined as; non-drinkers were defined as participants with no alcohol consumption; light drinkers with an alcohol consumption of 0< g per day ≤28 in men and 0< g per day ≤14 in women; and heavy drinkers with an alcohol consumption of ≥42 g per day in men and ≥28 g per day in women.

Replication of alcohol consumption associations

We aimed at replicating the association between alcohol consumption and the 363 CpGs previously identified by Liu *et al*. ($P<1\times10^{-7}$) [12], using data from the BIOS consortium (N=2042), which does not overlap with the Liu *et al.* data. This analysis revealed successful replication of 106 (29%) of these 363 CpGs after applying the Bonferroni-corrected significance threshold of $P<1.4\times10^{-4}$ *(0.05/363)* and 283 (78%) CpGs based on the uncorrected nominal significance threshold of $P<0.05$. All but one (cg06603309) of the 106 CpGs replicated after Bonferroni correction showed an inverse relationship with alcohol consumption, i.e., lower DNA methylation levels were associated with higher alcohol consumption, in line with the findings from the initial discovery EWAS by Liu *et al*. [12].

The top CpG in the Liu *et al*. discovery EWAS was cg02583484, annotated to the heterogeneous nuclear ribonucleoprotein A1 gene ($P=1.50\times10^{-19}$, $\beta=-0.0004$), which was replicated in our independent dataset with a P-value of 1.16×10^{-13} and β= -0.0055. In our dataset, this CpG had a methylation range with a minimum DNA methylation beta-value of 0.1457 and a maximum of 0.4189. However, this marker was not among the 144 CpGs used by Liu *et al.* for inference and thus was not used by us for prediction validation (see below). Out of the 144 predictive CpGs, we replicated 29 CpGs ($P<1.4\times10^{-4}$), which were

all included in the 144-CpG model, 19 in the 78-CpG model, 6 in the 23-CpG model, and 3 in the 5-CpG model. A summary of the results is presented in **Additional file 2**: **Table S1**.

A total of 77 genes were annotated to the 106 replicated CpGs after Bonferroni correction. Gene ontology enrichment analysis via http://geneontology.org/page/go-enrichmentanalysis showed that these 77 genes were enriched in two biological processes. The 'negative regulation of cellular macromolecule biosynthetic process' included enrichment of 16 genes (3.49-fold, FDR = 4.91×10^{-2}) and 25 genes were enriched in the 'cellular response to chemical stimulus' (2.63-fold, FDR = 3.80×10^{-2}).

Internal validation of alcohol consumption prediction models

To test the reproducibility of the seven prediction models reported by Liu *et al.* [12], we performed internal validation in our model building dataset via ten-fold cross-validation. The CpGs included per marker set and their average DNA methylation β-value per alcohol consumption category are presented in **Additional file 3**: **Table S2**. The mean AUC±SD obtained by the ten logistic regression models are denoted as 'Internal Validation' in **Figure 2**, **Additional file 4**: **Figures S1-S5** and **Additional file 5**: **Tables S3-S9**. The highest mean AUC of 0.83±0.05 was obtained for both the 5 and 23-CpG model for heavy drinkers *vs.* non-drinkers (**Figure 2A** and **Supplemental Table S3**). For the other six models, we obtained for all marker sets an average AUC ≤ 0.75 in three models and ≤ 0.70 in the other three models (**Figure 2**). Among all predictive marker sets, the lowest AUC of 0.61±0.04 was obtained for the 144-CpG model for light drinkers *vs.* non-drinkers (**Supplemental Table S9** and **Supplemental Figure S5**). In all seven prediction models, we obtained lower mean AUCs based on 144-CpGs compared to the models with lower numbers of CpG predictors. For example, the 144-CpG model for heavy drinkers *vs.* non-drinkers yielded an AUC of 0.78±0.06 compared to 0.83±0.05 obtained in the 5 and 23-CpG models (**Figure 2A** and **Supplemental Table S3**). Similar results were obtained in the other models, and for some of the 78-CpG models, as shown in **Additional file 4**: **Figures S1-S5** and **Additional file 5**: **Tables S3-S9**. Notably, these findings contrasts with that of Liu *et al.*, who reported increased prediction accuracies with increased numbers of CpG predictors [12].

External validation of alcohol consumption prediction models

Aiming to test the transportability of the prediction models trained in our complete model building dataset (N=2883), we performed external validation using data from three European cohorts (N=1794) not considered for model building and internal validation: the Rotterdam study (RS-III-1), SHIP-Trend, and two datasets from the TwinsUK study. The obtained AUCs are denoted as 'External Validation' in **Figure 2**, **Additional file 4: Figures S1-S5** and **Additional file 5: Tables S3-S9**. The AUCs obtained from external validation varied strongly per model between the external validation datasets

and differed with those obtained in the internal cross-validation. For example, the 144-CpG model for the heavy *vs.* non-drinkers yielded an AUC of 0.80 in RS and 0.84 in SHIP-Trend, while in TwinsUK and TwinsUK2 they were considerably lower with 0.68 and 0.60, respectively, and the mean AUC in the internal cross-validation was 0.78±0.06 (**Figure 2A** and **Additional file 5: Tables S3**). Similarly, the 23-CpG heavy *vs.* non-drinker model yielded AUCs of 0.81, 0.87, 0.65, 0.61, and 0.83±0.05, respectively. The high variance

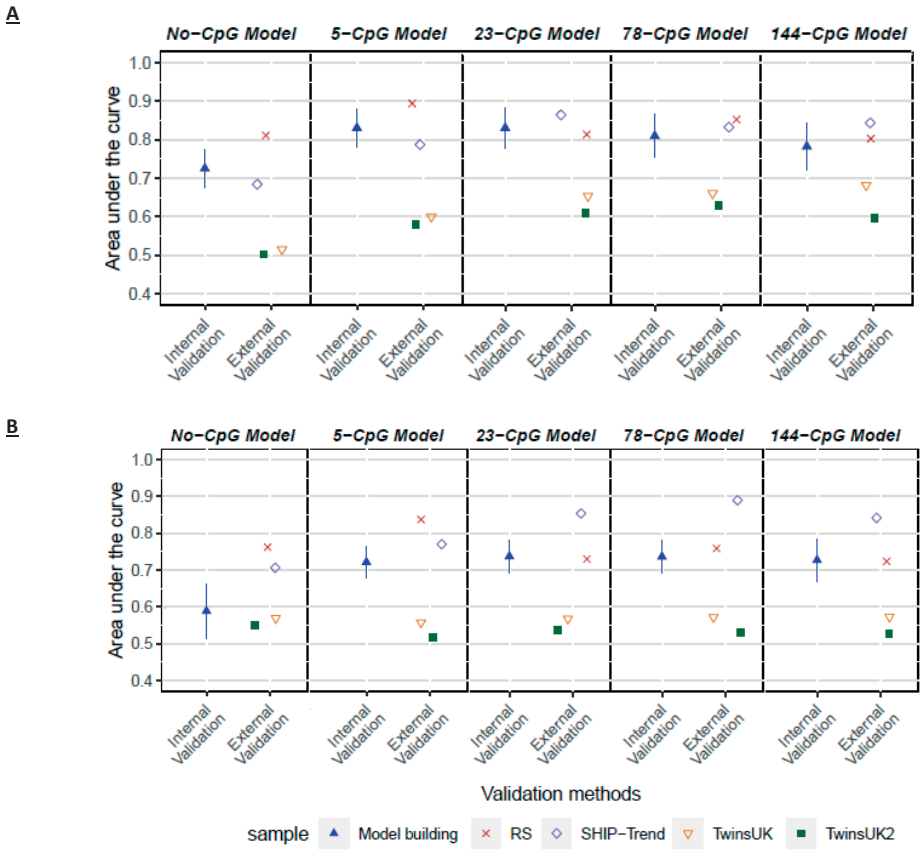


Figure 2. Epigenetic inference of alcohol consumption from blood based on Liu *et al.* biomarkers and models. Prediction accuracy for alcohol consumption expressed as Area Under the Curve (AUC) for (A) heavy drinkers *vs.* non-drinkers and (B) heavy drinkers *vs.* light drinkers using the CpG marker sets from Liu *et al.* [12]. Data from participants who do not fit the inferred categories were excluded from the respective prediction models following the approach used by Liu *et al*. 'Internal Validation': Mean AUC and SD from internal validation using ten-fold cross-validation in our model building dataset. 'External Validation': AUCs from external validation by applying our models trained in the model building dataset to independent data from three external validation cohorts (Rotterdam Study, N= 648; SHIP-Trend, N= 433; and TwinsUK, N= 713 and N= 442). Based on interview or self-reported information, non-drinkers were defined as participants with no alcohol consumption; light drinkers with an alcohol consumption of 0< g per day ≤28 in men and 0< g per day ≤14 in women; and heavy drinkers with an alcohol consumption of ≥42 g per day in men and ≥28 g per day in women. Abbreviations: RS- The Rotterdam Study; SHIP- Study of Health in Pomerania-Trend cohort; TwinsUK- The TwinsUK Study; TwinsUK2- Subset of the TwinsUK Study.

between obtained AUCs in the different external validation datasets was also observed in several other models as shown in **Additional file 4: Figures S1-S5** and **Additional file 5: Tables S3-S9**. The high AUC variance we observed in the external validation between datasets indicate non-robust performance of these prediction models, when applied to independent datasets.

New models for epigenetic inference of alcohol consumption using all categories

Finally, we developed two new models for epigenetic inference of alcohol consumption from blood by considering all data from all individuals of all four alcohol consumption categories in our prediction models, thereby refraining from excluding categories from prediction modeling as was done by Liu *et al*. [12]. To this end, we used all individuals from the model building dataset to build and internally validate via ten-fold cross-validation the models, as well as all individuals from our external validation datasets to externally validate the models. This was done for two different models. Model 1 comprised all heavy and at-risk drinkers combined *vs.* all light and non-drinkers combined. Model 2 included all heavy, at-risk, and light drinkers combined (i.e., all drinkers no matter the level of alcohol consumption) *vs.* all non-drinkers. The average AUCs ±SDs from internal cross-validation in the model building dataset were denoted as 'Internal Validation' and the four AUCs from the four external validation datasets as 'External Validation' (**Figure 3** and **Additional file 5: Tables S10** and **S11**).

Regarding model 1 for inferring heavy and at-risk *vs.* light and non-drinkers, the (mean) AUCs from internal cross-validation and from external validations ranged between 0.67-0.68 and 0.60-0.70, respectively across all marker sets (**Figure 3A** and **Additional file 5: Table S10**). Regarding model 2 for inferring all drinkers (heavy plus risk plus light) *vs.* non-drinkers, the AUCs from the two validation approaches based on the 5-CpG and the 23-CpG models were between 0.54-0.55 and 0.54-0.61, respectively. For the 78-CpG and the 144-CpG models, similarly low AUCs were seen in the internal validation, between 0.55-0.56, with slightly higher AUCs in the external validation datasets, between 0.57-0.63 (**Figure 3B** and **Additional file 5: Table S11**). Thus, compared to the Liu *et al.* models based on an approach that leaves out data, the new models based on all data achieved generally lower AUCs, while the AUC variance in the external validation was much less pronounced between the datasets than observed for the Liu *et al*. models (**Figure 2**, **Additional file 4: Figures S1-S5**, and **Additional file 5: Table S3-S9**).

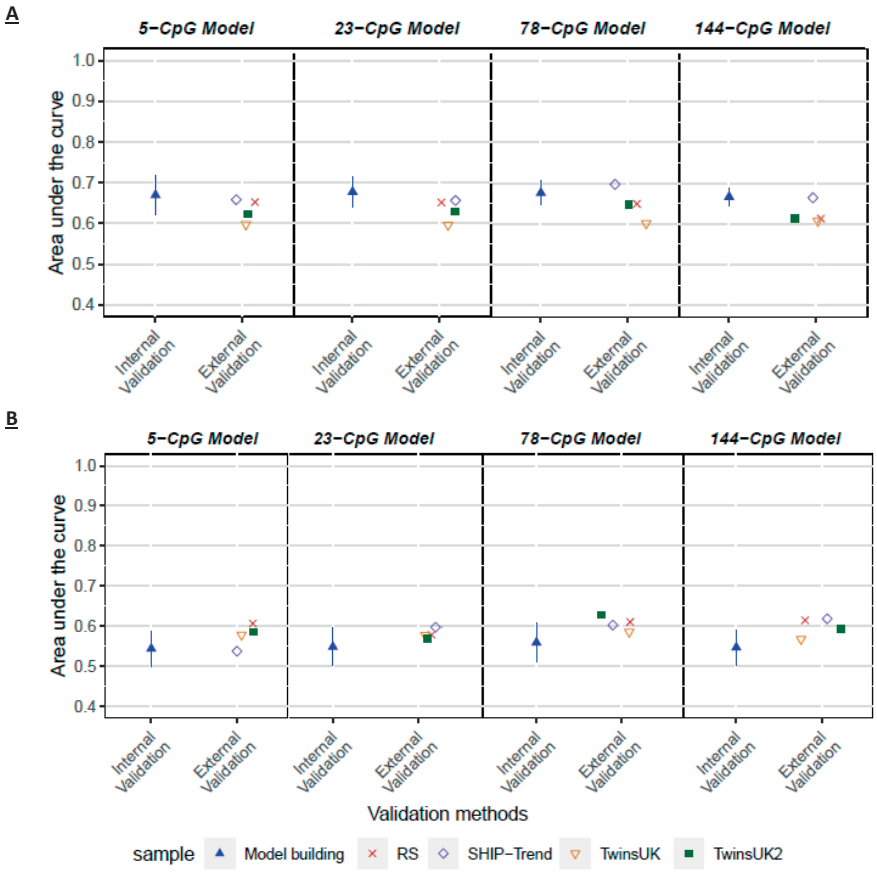


Figure 3. Epigenetic inference of alcohol consumption from blood based on newly developed models including all categories. Prediction accuracy for alcohol consumption expressed as Area Under the Curve (AUC) for (A) heavy and at-risk drinkers *vs.* light and non-drinkers and (B) heavy, at-risk and light drinkers *vs.* non-drinkers. In these models, all available participants from all categories were included, in contrast to Fig. 2. 'Internal Validation': Mean AUC and SD from internal validation using ten-fold cross-validation in our model building data set. 'External Validation': AUCs from external validation by applying our model trained in the model building dataset to independent data from three external validation cohorts (Rotterdam Study, N= 648; SHIP-Trend, N= 433; and TwinsUK, N= 713 and N= 442). For phenotype definition, see legend of Fig 2. Abbreviations: RS- The Rotterdam Study; SHIP- Study of Health in Pomerania-Trend cohort; TwinsUK- The TwinsUK Study; TwinsUK2- Subset of the TwinsUK Study.

DISCUSSION

In this study, we firstly performed replication analysis in an independent dataset of the EWAS results on alcohol consumption previously reported by Liu *et al.* [12], which delivered Bonferroni-corrected significant replication of close to one-third of the previously identified CpGs. Our smaller sample size of 2042 compared to 9642 in the Liu *et al.* study might be the reason why we only replicated one-third of the previously identified CpGs.

However, using the nominal significance threshold (*P<0.05*), we replicated the association of close to 80% of these CpGs.

Secondly, by using data from eight population-based cohorts, we performed in-depth validation of the biomarkers and models reported by Liu *et al*. [12] to infer alcohol consumption from blood. Reproducibility assesses the degree to which the model fits the real patterns rather than random noise in the data [28]. To test for reproducibility of the models, we performed internal model validation by implementing a ten-fold cross-validation scheme. The heavy *vs.* non-drinkers model obtained the highest average AUCs of the seven models in the cross-validation. Interestingly, the 144 and 78-CpG models obtained a lower average AUC than the 5 and 23-CpG models. In addition, in all models we observed a higher AUC for the models including less CpGs compared to the 144-CpG model and to some extent also for the 78-CpG models. In contrast, Liu *et al.* reported increased prediction accuracies for models with increased number of CpG predictors [12]. Our findings provide evidence that these 144-CpG models are over-fitted and thus, likely not reproducible. This increased risk for overfitting by including an increasing number of CpGs was also suggested by Hattab *et al.* [16]. Overfitting of a model is more likely to be observed when the ratio of the number of variables to the number of samples is small [29]. In this context, Harrell *et al.* [30] suggested that for generalizable binary models, no more than one predictor per ten participants in the smallest outcome category should be examined when fitting a regression model. As some of the findings of the analysis come from partially fitting to the noise on top of the true signal, noise features may be assigned nonzero coefficients due to chance associations with response to the training set [31]. Overall, the AUCs we achieved via internal cross-validation for the different models and marker sets were considerably lower than those reported by Liu *et al.* [12]. Also, the results obtained in our internal validation were much lower compared to the results we obtained when applying the same methods as Liu *et al.*, e.g. training and testing the model in the same dataset, in our model building dataset (see **Additional file 6** for results). This confirms previous conclusions [16] that the prediction accuracies reported by Liu *et al.* represent overestimates.

The transportability of the prediction models was tested by applying the models (trained in the model building dataset) to four validation datasets from three cohorts. Three models yielded an AUC ≤0.75 in the internal validation and in all four external validation datasets across all marker sets. For the other four models, a large variability in AUCs was obtained between the different datasets. Overall, in these four models, we obtained similar to higher AUCs for the Rotterdam Study and SHIP-Trend compared to the internal validation, while both TwinsUK datasets provided lower AUCs than the internal validation. It is important to note that the datasets from the Rotterdam Study and the TwinsUK (N=713) were both included in the EWAS for predictive marker discovery by Liu *et al.* [12]. The use of the same participants here and by Liu *et al.* could have led to an

overestimation of the prediction accuracies. Surprisingly, the AUCs we obtained in the TwinsUK (N=713) in the current study are in most models much lower than the AUCs we obtain in SHIP-Trend and the Rotterdam study. The results obtained in the Rotterdam Study were overall more similar to those obtained by SHIP-Trend. These results suggest that the use of the same participant, here and by Liu *et al.,* did not positively impact the prediction accuracies obtained in our study. The subset of the TwinsUK (N=442) includes re-processed DNA methylation data and a different FFQ-based approach for alcohol consumption information. Nevertheless, also in this dataset we obtain lower AUCs compared to the Rotterdam Study and SHIP-Trend, with very similar result as for the total TwinsUK (N=713) dataset. Notably, the AUCs from external validation were generally lower than the AUCs reported by Liu *et al.* [12] and as the similarly high AUCs we obtained from our model building dataset, when applying the same methods as Liu *et al*. (see **Additional file 6** for results), providing further evidence that the prediction accuracies reported by Liu *et al.* represent overestimates. This is in line with our conclusion from internal validation and as suggested by Hattab *et al.* [16].

Yousefi *et al.* [17] estimated DNA methylation-derived scores using the coefficients made available by Liu *et al.* [12] in participants of the Accessible Resource for Integrated Epigenomic Studies (ARIES) parental generation at midlife cohort (N = 1049, mean age = 50.2 ± 5.4 SD) as discovery dataset. A limitation of the study by Yousefi *et al.* [17] was the relatively small sample size in the higher alcohol consumption categories, with only 14 heavy drinkers and 67 at-risk drinkers. As a result, the lower AUCs obtained by Yousefi *et al.* [17] compared to Liu *et al.* [12] could possibly be due to the small sample size rather than an accurate representation of the true model prediction accuracies. In the current study, however, we have implemented 2883 participants, including 495 non-drinkers, 1800 light drinkers, 370 at-risk drinkers, and 218 heavy drinkers, with an age range of 19-87 years (mean age 57.4 ± 13.8 SD). By including more participants, especially in the categories with higher alcohol consumption, we overcome this possible sample size limitation and thus provide a more reliable representation of the models' prediction accuracies. Yousefi *et al.* [17] obtained low AUCs from 0.48 to 0.57 to distinguish heavy drinkers *vs.* non-drinkers and 0.55 to 0.57 for heavy drinkers *vs.* light drinkers in adults at midlife. In our external validation, we obtained AUCs from 0.80 to 0.89 in the Rotterdam Study, 0.68 to 0.87 in SHIP-Trend, 0.52 to 0.68 in TwinsUK (N=713), and 0.50-0.63 in TwinsUK2 (N=442) for distinguishing heavy *vs.* non-drinkers and 0.72 to 0.84 in the Rotterdam Study, 0.71 to 0.89 in SHIP-trend, 0.56 to 0.57 in TwinsUK (N=713), and 0.52-0.55 in TwinsUK2 (N=442) for heavy drinkers *vs.* light drinkers. The results in the Rotterdam Study and SHIP-Trend are overall higher than those obtained by Yousefi *et al.*, while the results obtained in the TwinsUK are very similar to those obtained by Yousefi *et al.* In addition, the high variability in the obtained AUCs in our study and the close to random inference obtained by Yousefi *et al.* [17] study suggest that the tested CpGs are

not as suitable as previously suggested for achieving transportable and accurate alcohol consumption prediction models.

The exclusion from prediction modelling of data from participants who did not fit the inferred categories, as done by Liu *et al.* [12], means that such models cannot be applied to the general population, where individuals with the excluded categories exist and can never be inferred correctly because their category was not considered in the prediction model. Therefore, for a prediction models to be applicable in cohort studies used in epidemiology research, or any practical applications in the clinic and beyond, should be designed in data that realistically reflects the general population. For that reason, we have developed two additional models in which data from all individuals of all alcohol categories were included and validated them internally via cross-validation as well as externally in independent datasets. The first model for heavy and at-risk drinkers *vs.* light and non-drinkers provided cross-validated AUCs between 0.67 and 0.68 across all four CpG marker sets. These results are close to the lower 95% CI of the 450-CpG based model previously developed by McCartney *et al*. [15], which had an AUC of 0.73 (95% CI= 0.69–0.78) to distinguish light-to-moderate drinkers from heavy drinkers. Four CpGs overlap between this 450-CpG model and the 23-CpG model: cg00252472, cg06690548, cg11613559, and cg12825509. In addition, two more CpGs overlap with the 144-CpG model; cg11376147 and cg18032812. In the external validation, we obtained AUCs in the range of 0.60-0.70 across all marker sets and all external validation cohorts. In the second model, which distinguishes heavy, at-risk, and light drinkers *vs.* non-drinkers, we obtained AUCs at 0.54-0.63 in both internal and external validation. Thus, when applying appropriate prediction methodology by not excluding participant data and performing external validation, the CpG marker sets reported by Liu *et al.* [12] yield much lower prediction accuracies as compared to the AUCs previously published and obtained here based on the previous approach.

Our study has strengths and limitations that should be considered when interpreting the results. The main strengths of our study are the use of a large dataset from several cohorts with similar numbers for the different categories as Liu *et al.* [12], and the use of four datasets for external model validation. Moreover, our findings agree with a previous validation study based on a different methodology [17], while our larger dataset improved the limitations of the limited data used in the previous validation study. The main limitation of our study, as well as in the previous studies, is that the alcohol consumption information is based on interviews or self-reported questionnaires, which are generally considered unreliable in terms of underestimating actual alcohol consumption. Regarding the putative inaccuracy of interviews and self-reported alcohol consumption used here as phenotypes, we cannot know how error-prone these reports are. In particular, it is possible that heavy drinkers might not be able to or might be hesitant or unwilling to accurately recall or report their high alcohol consumption. Also, there is variability

in the questionnaires regarding the reference time window. For example, KORA participants were asked about alcohol consumption in the past few days, which may or may not be representative of the participants' long-term alcohol consumption habits. Also, non-drinkers may include lifetime non-drinkers but also sober alcoholics; however, it is not yet clear how this could affect the obtained DNA methylation patterns. Because all available studies, including the EWAS that identified CpGs associated with alcohol consumption, used interviews or self-reported alcohol consumption information, this is a general limitation that cannot be easily solved, as methods to empirically measure alcohol concentrations are not suitable for estimating long-term alcohol consumption. Another source of uncertainty may lie in the calculation for alcohol consumption in grams/day, which presents a slight variation in the formula used between the different cohorts. The variation in alcohol consumption data collection between the cohorts might also play a role in the variance we obtain in the prediction AUCs.

Another shortcoming of our study was the inclusion of only participants from European ancestry. As DNA methylation patterns might differ between populations [32], the absence of non-European participants during marker discovery and model building might prohibit accurate model transportability to non-European populations. Hence, future studies would benefit from a trans-ethnic prediction marker discovery, model building, and validation.

Overall, our extensive validation testing of the different CpG sets reported by Liu *et al*. [12] for inferring alcohol consumption from blood demonstrates that using appropriate prediction methodology regarding both separating datasets for model building and model testing by performing internal cross-validation and external validation, and including all alcohol consumption categories and individuals in the prediction modelling, yields much lower prediction accuracies and with a high variance between validation cohorts for the Liu *et al*. models as were previously published. This allows us to conclude that the currently available DNA methylation predictors for alcohol consumption need to be improved considerably before epigenetic inference of alcohol consumption from blood can be considered for practical applications in the clinic and beyond. Our study implies that, currently, we are far away from epigenetic inference of alcohol consumption from blood in research and practical applications, despite EWASs having already delivered hundreds of associated CpGs. Thus, further EWASs on alcohol consumption are necessary to increase the number of associated CpGs, including replication studies for the identified CpGs. Established CpGs replicated in several independent studies could provide better predictive markers than CpGs identified in one large meta-analysis. These CpGs will need to be carefully tested for their value to improve the low accuracy in inferring alcohol consumption from blood achieved with the currently available marker sets.

METHODS

Study populations

This study was embedded within the Biobank-based Integrative Omics Study (BIOS) consortium [26], by including participants from the Rotterdam Study (sub-cohorts RS-II-3 and RS-III-2) (N=611) [18], Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) (N=159) [19], the Netherlands Twin Register (NTR) (N=617) [20], the Leiden Longevity Study (LLS) (N=491) [21], and the Prospective ALS Study Netherlands (PAN) (N=164) [22]. Additionally, we included 841 participants from The Cooperative Health Research in the Region of Augsburg (KORA) study (F4) [23]. External validation was conducted in independent samples (i.e., not used for model building and internal validation) from the Study of Health in Pomerania (SHIP)-Trend cohort (N = 433) [24], two datasets from the TwinsUK Study with overlapping participants; TwinsUK (N=713) and TwinsUK2 (N=442) [25], and participants from the Rotterdam Study sub-cohort RS-III-1 (N=648) that are not included in the BIOS consortium. Alcohol consumption information was obtained via interviews or self-reported questionnaires. Cohort specific data collection and dataset characteristics are summarized in **Table 1** and described in detail in the Supplementary Methods (**Additional file 1**).

Microarray-based DNA methylation quantification

DNA was extracted from whole peripheral blood and analyzed with the Illumina Infinium Human Methylation 450 K BeadChip (Illumina Inc, San Diego, CA, USA) or the Infinium MethylationEPIC BeadChip (Illumina Inc, San Diego, CA, USA) to obtain the DNA methylation measurements. Details on cohort-specific methods are provided in Supplementary Methods (**Additional file 1**). The methylation proportion of a CpG site was reported as the methylation β-value in the range of 0 to 1.

Candidate-CpG association study for alcohol consumption and gene annotation

Using the data from the BIOS Consortium (N=2042), we tested the association of the 363 CpGs previously found to be significantly associated ($P<1\times10^{-7}$) with alcohol consumption [12]. Alcohol consumption levels (grams/day) were right-skewed and contained non-drinkers; therefore, the log-transformed alcohol consumption (log (g per day+1)) was used as the independent variable. The β-values of the 363 CpGs were included as the dependent variable and the analysis was adjusted for age, sex, BMI, batch effects (plate, plate location, and cohort ID), and Houseman-imputed white blood cell counts (WBC) for CD4T cells, CD8T cells, natural killer cells, B-cells, granulocytes, and monocytes [33]. The Bonferroni multiple-test corrected 5% significance level of $P<1.4\times10^{-4}$ *(0.05/363)* was applied. All analyses were performed using the statistical package R, version 3.4.3.

We obtained the genes annotated to the replicated CpGs using the annotation file provided by Illumina and performed Gene ontology (http://geneontology.org/page/go-enrichmentanalysis) enrichment analysis for these genes.

Validation of the previously published prediction models

The BIOS and KORA DNA methylation data were combined as the model building dataset (N=2883) using the "ComBat" function [34] (R-package "sva" [35]) to adjust for the known batches via an empirical Bayesian framework adjusting for age and sex. Then, possible confounders were regressed out using linear regression models, obtaining the residuals for each CpG adjusted for age, sex, BMI, batch effects (plate, plate location, and cohort ID), and Houseman-imputed WBC (CpG = age+ sex+ BMI+ batch effects+ WBC).

The self-reported phenotypic data on alcohol consumption were categorized according to their alcohol consumption levels for which we used the same cut-off categories as described by Liu *et al*. [12], to allow for direct comparison of the models' performance; non-drinkers: participants with no alcohol consumption; light drinkers: participants with alcohol consumption of 0< g per day ≤28 in men and 0< g per day ≤14 in women; at-risk drinkers: participants with alcohol consumption of 28< g per day <42 in men and 14< g per day <28 in women; heavy drinkers: participants with alcohol consumption of ≥42 g per day in men and ≥28 g per day in women.

The alcohol categories used in each model were inferred using the same seven prediction models as previously applied by Liu *et al.* with heavy drinkers *vs.* all other categories separately, i.e., heavy drinkers *vs.* (1) non-drinkers, (2) light drinkers, (3) pooled individuals of light or non-drinkers, (4) at-risk drinkers, as well as two-category combinations between the other categories including (5) at-risk drinkers *vs.* non-drinkers, (6) at-risk drinkers *vs.* light drinkers, and (7) light drinkers *vs.* non-drinkers. In all models, the former category was the 'cases' (coded as "1") and the latter was the 'control' group (coded as "0"). Selecting a subset of categories in prediction modeling and AUC estimation, as was done by Liu *et al.,* may limit the possibility to extrapolate the result to the general population. Hence, we replicated this approach solely for outcome compatibility reasons. All seven models were trained for the null model, which only includes age, sex, and BMI, and subsequently the null model combined with the residuals of the four CpG sets (5, 23, 78, or 144 CpGs). The CpGs included per model and the average DNA methylation β-values per category are presented in **Additional file 3: Table S2**.

Internal and external validation of the previously published prediction models

We tested the reproducibility (internal validation) and transportability (external validation) of the prediction models conducted by Liu *et al.* [12, 28]. First, we adopted a ten-fold cross-validation scheme [36] in which the whole model building dataset (N=2883) was

randomly distributed into ten non-overlapping subsets. The logistic regression model was trained in a combination of nine subsets (90% of the data), which was then applied to the remaining subset (10% of the data) to infer the participants' alcohol status. This method results in ten different training (90%) and testing (10%) sets. We trained the seven models in the training sets (90%) using binomial regression analysis with the alcohol categories (coded as 1/0) as the dependent variable and age, sex, and BMI without (the null model) or with a set of (the residuals of the) CpGs as the independent variables (Alcohol category = age + sex + BMI (+$ResCpGs_{5,\ 23,\ 78,\ 144}$)). For this purpose, the "glm" function with "binomial" as family and "logit" as link were used. The models were then applied to the test set (10%) using the "predict" function. The prediction performance of the models was assessed using "roc" (R-package "pROC") that calculates the AUC per model. This method resulted in ten logistic regression models and consequently, ten AUCs from which average values were estimated and standard deviation were obtained.

Secondly, we externally validated the models that were trained in the complete model building dataset (N=2883) by testing them in four external validation datasets, using the "predict" function. The "roc" function (R-package "pROC") was again used to calculate the AUC per model. The independent cohorts used our previously described pre-processing procedure by regressing out the potential covariates. The TwinsUK study used a linear mixed model to additionally adjust for twin family structure and zygosity using random effects. Also, sex was not included in the pre-processing steps because solely women were included in the TwinsUK analysis. Notably, according to the above-described scenarios, both internal and external validations followed the same approach previously applied by Liu *et al.* [12] in that individuals not fitting the inferred categories were excluded from the prediction analysis.

Prediction modeling without excluding categories and data

Finally, we trained as well as internally and externally validated two new prediction models comprising all individuals in the prediction modeling, i.e., 1) heavy and at-risk drinkers *vs.* light and non-drinkers and 2) heavy, at-risk and light drinkers (i.e. all drinkers no matter how much) *vs.* non-drinkers. These two models were internally validated via ten-fold cross-validation and externally validated in four datasets. As age, sex, and BMI are already accounted for in the residuals we solely included the four CpG marker sets in these models. The coefficients for these models are presented in **Additional file 7**: **Supplementary Table S21**.

SUPPLEMENTARY MATERIAL

Figure legend: Figures S1- S5

Prediction accuracy for alcohol consumption expressed as Area Under the Curve (AUC) using the CpG marker sets from Liu et al. 'Internal Validation': Mean AUC and SD from internal validation using ten-fold cross-validation in our model building dataset (6 cohorts, N= 2883). 'External Validation': AUCs from external validation by applying our model trained in the model building dataset to data from three external validation cohorts (Rotterdam Study, N= 648; SHIP-Trend, N= 433; and TwinsUK, N= 713 and N=442). Based on interview or self-reported information, non-drinkers were defined as participants with no alcohol consumption; light drinkers with an alcohol consumption of 0< g per day ≤28 in men and 0< g per day ≤14 in women; and heavy drinkers with an alcohol consumption of ≥42 g per day in men and ≥28 g per day in women. Abbreviations: RS- The Rotterdam Study; TwinsUK- The TwinsUK Study; TwinsUK2- Subset of the TwinsUK Study; SHIP- Study of Health in Pomerania-Trend cohort

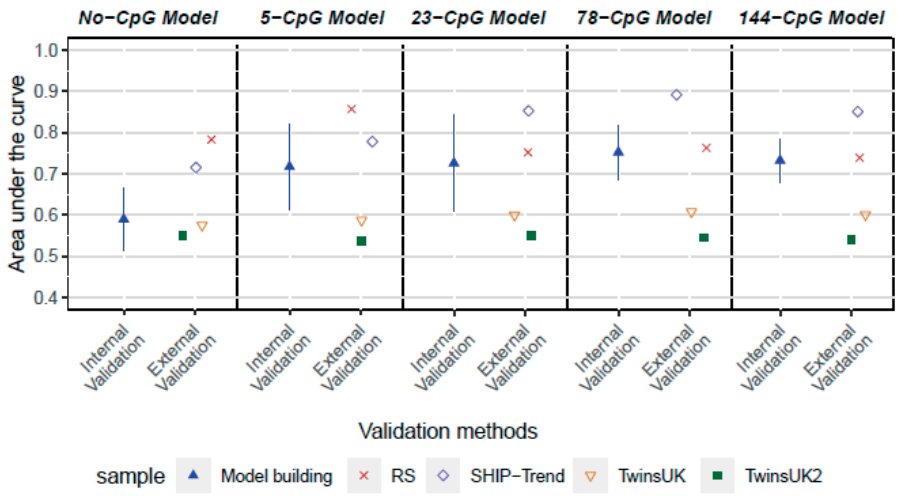


Figure S1. Accuracy of epigenetic inference of heavy drinkers *vs.* light and non-drinkers using different marker sets.

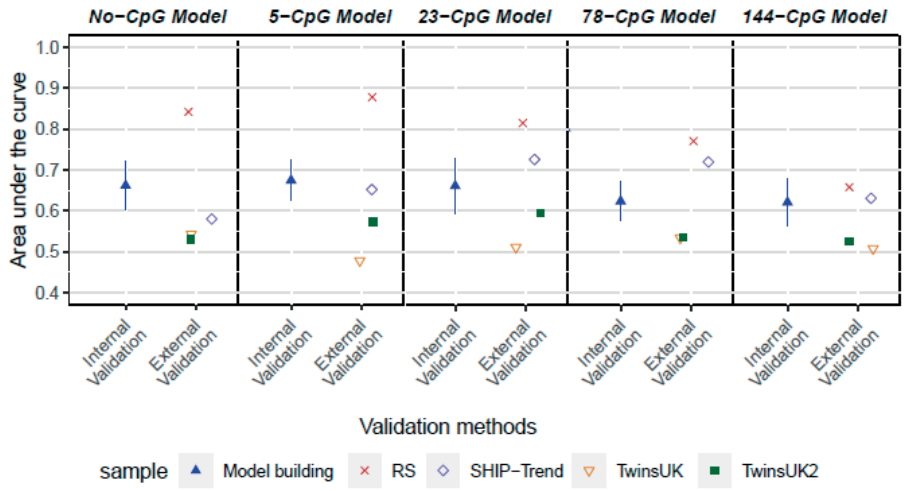


Figure S2. Accuracy of epigenetic inference of heavy drinkers *vs.* at-risk drinkers using different marker sets.

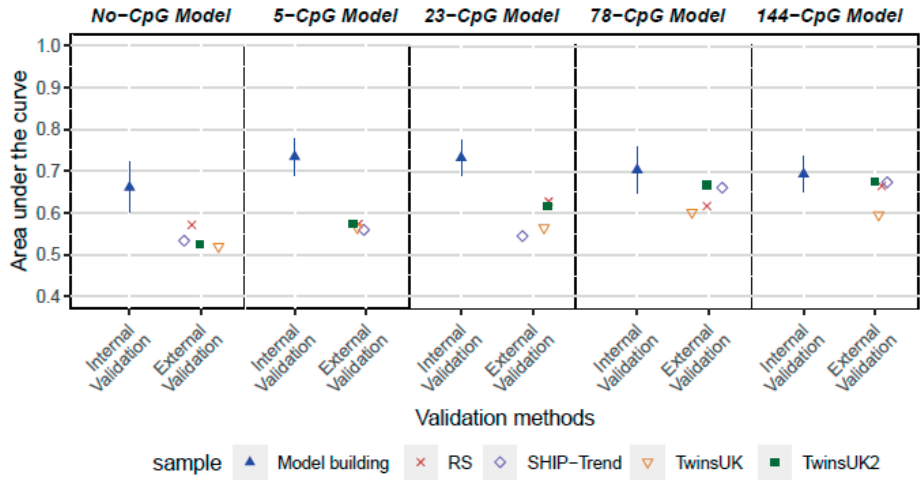


Figure S3. Accuracy of epigenetic inference of at-risk drinkers *vs.* non-drinkers using different marker sets.

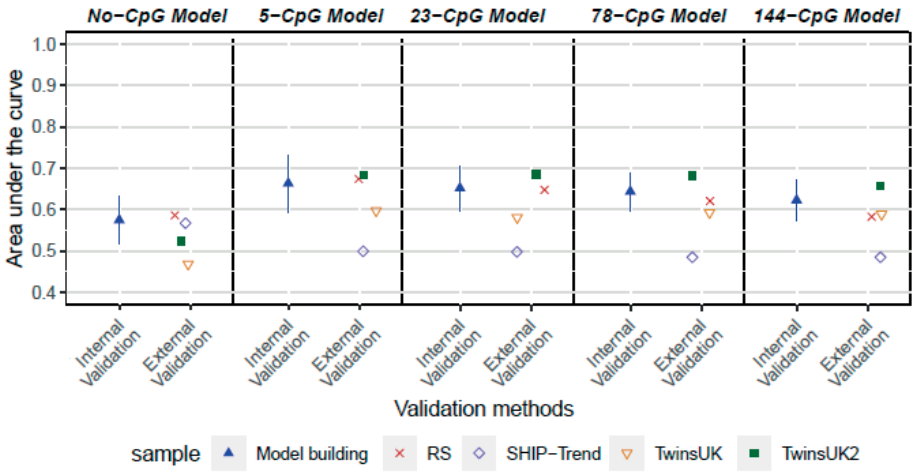


Figure S4. Accuracy of epigenetic inference of at-risk drinkers *vs.* light drinkers using different marker sets.

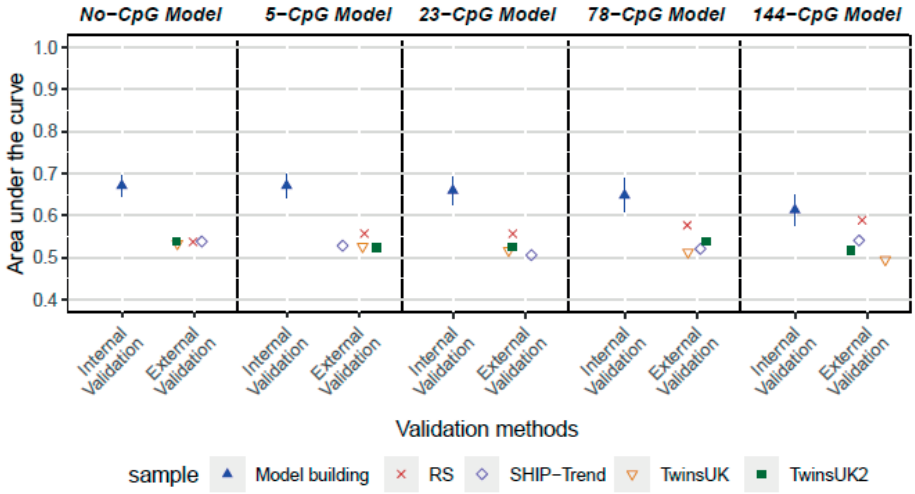


Figure S5. Accuracy of epigenetic inference of light drinkers *vs.* non-drinkers using different marker sets.

Table legend: Table S3- Table S11

Prediction accuracy for alcohol consumption expressed as Area Under the Curve (AUC) using the CpG marker sets from Liu *et al.* 'Internal Validation': Obtained AUCs using ten-fold cross-validation in the model building data set; 'External Validation': AUCs from external validation by applying our model trained in the model building dataset to data from three external validation cohorts (Rotterdam Study, N= 648; SHIP-Trend, N= 433; and TwinsUK, N= 713 and N=442). Based on interview or self-reported information, non-drinkers were defined as participants with no alcohol consumption; light drinkers with an alcohol consumption of 0< g per day ≤28 in men and 0< g per day ≤14 in women; and heavy drinkers with an alcohol consumption of ≥42 g per day in men and ≥28 g per day in women. Abbreviations: RS- The Rotterdam Study; TwinsUK- The TwinsUK Study; TwinsUK2- Subset of the TwinsUK Study; SHIP- Study of Health in Pomerania-Trend cohort; ABS- Null model including only age, body mass index, and sex.

Table S3 Accuracy of epigenetic inference of heavy drinkers *vs.* non-drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.78±0.06	0.80	0.84	0.68	0.60
78-CpGs	0.81±0.06	0.85	0.83	0.66	0.63
23-CpGs	0.83±0.05	0.81	0.87	0.65	0.61
5-CpGs	0.83±0.05	0.89	0.79	0.60	0.58
ABS	0.73±0.05	0.81	0.68	0.52	0.50

Table S4 Accuracy of epigenetic inference of heavy drinkers *vs.* light drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.73±0.06	0.72	0.84	0.57	0.53
78-CpGs	0.74±0.04	0.76	0.89	0.57	0.53
23-CpGs	0.74±0.04	0.73	0.85	0.57	0.54
5-CpGs	0.72±0.04	0.84	0.77	0.56	0.52
ABS	0.59±0.07	0.76	0.71	0.57	0.55

Table S5 Accuracy of epigenetic inference of heavy drinkers *vs.* light and non-drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.73±0.05	0.74	0.85	0.60	0.54
78-CpGs	0.75±0.07	0.76	0.89	0.61	0.55
23-CpGs	0.73±0.12	0.75	0.85	0.60	0.55
5-CpGs	0.72±0.10	0.86	0.78	0.59	0.54
ABS	0.59±0.08	0.78	0.72	0.58	0.55

Table S6 Accuracy of epigenetic inference of heavy drinkers *vs.* at-risk drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.62±0.06	0.66	0.63	0.51	0.53
78-CpGs	0.62±0.05	0.77	0.72	0.53	0.54
23-CpGs	0.66±0.07	0.82	0.73	0.51	0.60
5-CpGs	0.67±0.05	0.88	0.65	0.48	0.57
ABS	0.66±0.06	0.84	0.58	0.54	0.53

Table S7 Accuracy of epigenetic inference of at-risk drinkers *vs.* non-drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.69±0.04	0.67	0.67	0.60	0.68
78-CpGs	0.70±0.06	0.62	0.66	0.60	0.67
23-CpGs	0.73±0.04	0.63	0.55	0.57	0.62
5-CpGs	0.73±0.04	0.57	0.56	0.56	0.57
ABS	0.66±0.06	0.57	0.53	0.52	0.53

Table S8 Accuracy of epigenetic inference of at-risk drinkers *vs.* light drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.62±0.05	0.58	0.49	0.59	0.66
78-CpGs	0.64±0.05	0.62	0.49	0.59	0.68
23-CpGs	0.65±0.05	0.65	0.50	0.58	0.69
5-CpGs	0.66±0.07	0.67	0.50	0.60	0.68
ABS	0.58±0.06	0.59	0.57	0.47	0.52

Table S9 Accuracy of epigenetic inference of light drinkers *vs.* non-drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.61±0.04	0.59	0.54	0.50	0.52
78-CpGs	0.65±0.04	0.58	0.52	0.51	0.54
23-CpGs	0.66±0.03	0.56	0.51	0.52	0.53
5-CpGs	0.67±0.03	0.56	0.53	0.53	0.52
ABS	0.67±0.03	0.54	0.54	0.53	0.54

Table S10 Accuracy of epigenetic inference of heavy and at-risk drinkers *vs.* light and non-drinkers using different marker sets.

	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.67±0.02	0.61	0.66	0.61	0.61
78-CpGs	0.68±0.03	0.65	0.70	0.60	0.65
23-CpGs	0.68±0.04	0.65	0.66	0.60	0.63
5-CpGs	0.67±0.05	0.65	0.66	0.60	0.62

Table S11 Accuracy of epigenetic inference of heavy, at-risk and light drinkers *vs.* non-drinker using different marker sets.

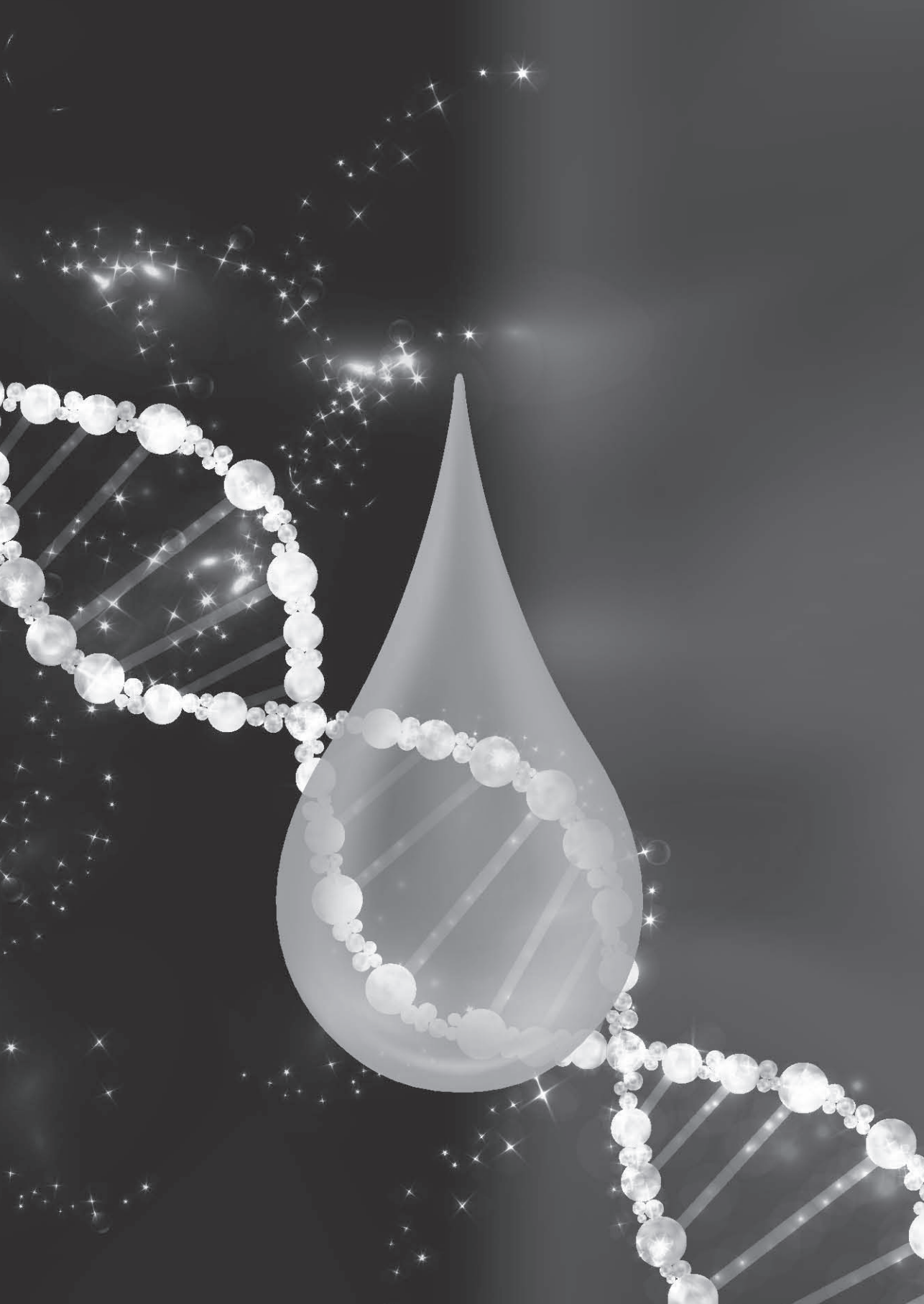
	Internal	*External validation*			
Marker set	*validation*	*RS*	*SHIP-Trend*	*TwinsUK*	*TwinsUK2*
144-CpGs	0.55±0.04	0.61	0.62	0.57	0.59
78-CpGs	0.56±0.05	0.61	0.60	0.59	0.63
23-CpGs	0.55±0.05	0.58	0.60	0.58	0.57
5-CpGs	0.54±0.04	0.61	0.54	0.58	0.59

Additional supplemental material for this chapter can be found in the online version of the paper via https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-021-01186-3.

REFERENCES

1. Collaborators GBDRF. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1345-422.
2. Shield KD, Parry C, Rehm J. Chronic diseases and conditions related to alcohol use. Alcohol Res. 2013;35(2):155-73.
3. Helander A, Eriksson CJ, State WISo, Trait Markers ofAlcohol U, Dependence I. Laboratory tests for acute alcohol consumption: results of the WHO/ISBRA Study on State and Trait Markers of Alcohol Use and Dependence. Alcohol Clin Exp Res. 2002;26(7):1070-7.
4. Helander A. Biological markers in alcoholism. J Neural Transm Suppl. 2003(66):15-32.
5. Andresen-Streichert H, Müller A, Glahn A, Skopp G, Sterneck M. Alcohol Biomarkers in Clinical and Forensic Contexts. Dtsch Arztebl Int. 2018;115(18):309-15.
6. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7.
7. Ladd-Acosta C. Epigenetic Signatures as Biomarkers of Exposure. Curr Environ Health Rep. 2015;2(2):117-25.
8. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biology. 2017;18(1):238.
9. Philibert RA, Plume JM, Gibbons FX, Brody GH, Beach SR. The impact of recent alcohol use on genome wide DNA methylation signatures. Front Genet. 2012;3:54.
10. Wilson LE, Xu Z, Harlid S, White AJ, Troester MA, Sandler DP, et al. Alcohol and DNA Methylation: An Epigenome-Wide Association Study in Blood and Normal Breast Tissue. Am J Epidemiol. 2019;188(6):1055-65.
11. Xu K, Montalvo-Ortiz JL, Zhang X, Southwick SM, Krystal JH, Pietrzak RH, et al. Epigenome-Wide DNA Methylation Association Analysis Identified Novel Loci in Peripheral Cells for Alcohol Consumption Among European American Male Veterans. Alcohol Clin Exp Res. 2019;43(10):2111-21.
12. Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23(2):422-33.
13. Endo K, Li J, Nakanishi M, Asada T, Ikesue M, Goto Y, et al. Establishment of the MethyLight Assay for Assessing Aging, Cigarette Smoking, and Alcohol Consumption. Biomed Res Int. 2015;2015:451981.
14. Philibert R, Miller S, Noel A, Dawes K, Papworth E, Black DW, et al. A Four Marker Digital PCR Toolkit for Detecting Heavy Alcohol Consumption and the Effectiveness of Its Treatment. J Insur Med. 2019;48(1):90-102.
15. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136.
16. Hattab MW, Clark SL, van den Oord E. Overestimation of the classification accuracy of a biomarker for assessing heavy alcohol use. Mol Psychiatry. 2018;23(11):2114-5.
17. Yousefi PD, Richmond R, Langdon R, Ness A, Liu C, Levy D, et al. Validation and characterisation of a DNA methylation alcohol biomarker across the life course. Clin Epigenetics. 2019;11(1):163.
18. Ikram MA, Brusselle G, Ghanbari M, Goedegebure A, Ikram MK, Kavousi M, et al. Objectives, design and main findings until 2020 from the Rotterdam Study. Eur J Epidemiol. 2020;35(5):483-517.
19. van Greevenbroek MM, Jacobs M, van der Kallen CJ, Vermeulen VM, Jansen EH, Schalkwijk CG, et al. The cross-sectional association between insulin resistance

and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). Eur J Clin Invest. 2011;41(4):372-9.
20. Willemsen G, Vink JM, Abdellaoui A, den Braber A, van Beek JH, Draisma HH, et al. The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. Twin Res Hum Genet. 2013;16(1):271-81.
21. Schoenmaker M, de Craen AJ, de Meijer PH, Beekman M, Blauw GJ, Slagboom PE, et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. Eur J Hum Genet. 2006;14(1):79-84.
22. Huisman MH, de Jong SW, van Doormaal PT, Weinreich SS, Schelhaas HJ, van der Kooi AJ, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. J Neurol Neurosurg Psychiatry. 2011;82(10):1165-70.
23. Holle R, Happich M, Lowel H, Wichmann HE, Group MKS. KORA--a research platform for population based health research. Gesundheitswesen. 2005;67 Suppl 1:S19-25.
24. Völzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, et al. Cohort profile: the study of health in Pomerania. Int J Epidemiol. 2011;40(2):294-307.
25. Moayyeri A, Hammond CJ, Valdes AM, Spector TD. Cohort Profile: TwinsUK and healthy ageing twin study. Int J Epidemiol. 2013;42(1):76-85.
26. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131-8.
27. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol. 1979;110(3):281-90.
28. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130(6):515-24.
29. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
30. Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med. 1984;3(2):143-52.
31. Frank E. Harrell J. Regression Modeling Strategies; With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Second ed. New York: Springer; 2001.
32. Kader F, Ghai M. DNA methylation-based variation between human populations. Molecular Genetics and Genomics. 2017;292(1):5-35.
33. Houseman EA, Kelsey KT, Wiencke JK, Marsit CJ. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. BMC Bioinformatics. 2015;16:95.
34. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
35. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.
36. Hastie T TR, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second ed. New York: Springer; 2009.



Chapter 2.2

Validated inference of smoking habits from blood with a finite DNA methylation marker set

Silvana C.E. Maas, Athina Vidaki, Rory Wilson, Alexander Teumer, Fan Liu, Joyce B.J. van Meurs, André G. Uitterlinden, Dorret I. Boomsma, Eco J.C. de Geus, Gonneke Willemsen, Jenny van Dongen, Carla J.H. van der Kallen, P. Eline Slagboom, Marian Beekman, Diana van Heemst, Leonard H. van den Berg, BIOS Consortium, Liesbeth Duijts, Vincent W.V. Jaddoe, Karl-Heinz Ladwig, Sonja Kunze, Annette Peters, M. Arfan Ikram, Hans J. Grabe, Janine F. Felix, Melanie Waldenberger, Oscar H. Franco, Mohsen Ghanbari*, Manfred Kayser*

European Journal of Epidemiology 2019;34(11):1055-74

***Denotes equal contribution**

ABSTRACT

Inferring a person's smoking habit and history from blood is relevant for complementing or replacing self-reports in epidemiological and public health research, and for forensic applications. However, a finite DNA methylation marker set and a validated statistical model based on a large dataset are not yet available. Employing 14 epigenome-wide association studies for marker discovery, and using data from six population-based cohorts (N=3,764) for model building, we identified 13 CpGs most suitable for inferring smoking versus non-smoking status from blood with a cumulative Area Under the Curve (AUC) of 0.901. Internal five-fold cross-validation yielded an average AUC of 0.897±0.137, while external model validation in an independent population-based cohort (N=1,608) achieved an AUC of 0.911. These 13 CpGs also provided accurate inference of current (average $AUC_{crossvalidation}$ 0.925±0.021, $AUC_{externalvalidation}$0.914), former (0.766±0.023, 0.699) and never smoking (0.830±0.019, 0.781) status, allowed inferring pack-years in current smokers (10 pack-years 0.800±0.068, 0.796; 15 pack-years 0.767±0.102, 0.752) and inferring smoking cessation time in former smokers (5 years 0.774±0.024, 0.760; 10 years 0.766±0.033, 0.764; 15 years 0.767±0.020, 0.754). Model application to children revealed highly accurate inference of the true non-smoking status (6 years of age: accuracy 0.994, N=355; 10 years: 0.994, N=309), suggesting prenatal and passive smoking exposure having no impact on model applications in adults. The finite set of DNA methylation markers allow reliable and accurate inference of smoking habit, with comparable accuracy as plasma cotinine use, and smoking history from blood, which we envision becoming useful in epidemiology and public health research, and in medical and forensic applications.

INTRODUCTION

Several studies suggest that tobacco smoking impacts the human epigenome, particularly by changing DNA methylation patterns [1, 2]. DNA methylation is catalyzed by DNA methyltransferases (DNMT's); the carcinogens in cigarette smoke cause double-strand DNA breaks and the DNA repair sites recruit DNMT1 [3], which methylates cytosines in CpGs adjacent to the repaired nucleotides [4]. Nicotine was shown to down-regulate DNMT1, and mRNA and protein expression [5]. Furthermore, cigarette smoke condensate increases expression of Sp1, a transcription factor that binds to GC-rich motifs in gene promoters, preventing *de novo* methylation [6-9]. In recent years, various epigenome-wide association studies (EWASs) have provided a long list of CpGs significantly associated with tobacco smoking habits in blood [10]. Although there are strong smoking associations across the epigenome, some studies suggest that after smoking cessation, DNA methylation patterns can return back to those found in never smokers [11, 12].

Smoking is a well-known risk factor for the development of several diseases [13, 14]. Therefore, studies that investigate smoking and its effect on mortality and morbidity rely on accurate assessments of smoking exposure. These studies use mainly self-reported smoking questionnaires to collect this information, which could result in underestimation and misrepresent the degree of the true smoking exposure [15]. In particular, it is possible that specific groups of participants, for instance pregnant women, are more reluctant to confide that they smoke [16]. Hence, the ability to reliably and accurately infer a person's smoking habit from blood is relevant in epidemiology and public health research as well as in medical practice, because such an approach could complement, or even replace, self-reported smoking questionnaires.

Moreover, inference of a person's smoking habit from blood traces found at crime scenes would allow the broadening of DNA investigative intelligence beyond the currently considered parameters of appearance, bio-geographic ancestry and age, thus helping to better find unknown perpetrators of crime who are not identifiable via standard forensic DNA profiling [17]. Blood-based toxicological tests for measurement of tobacco exposure exist; however, they assess current and acute, rather than habitual, smoking [18]. In addition, biomarkers used include nicotine itself or its metabolite cotinine, and their accurate detection of current smokers is affected by their short half-lives (2-3h versus 15-19h for nicotine and cotinine, respectively) and individual variation in metabolic rates [19]. Therefore, when using the cotinine-based approach false-negatives can be easily obtained, and also false-positives may occur in former smokers that use nicotine replacement therapy [20]. Given these constrains of current toxicology blood measures, and considering the recent progress in understanding the impact of smoking on epigenetic variation, we envision DNA methylation from blood as a promising approach for long-term habitual smoking behaviour.

Although progress has been made in understanding the epigenetic impact of smoking [1], only a limited number of studies have explored the inference of smoking habits from blood with DNA methylation markers, albeit with various limitations such as small sample size, limited validation, restricting to smokers and non-smokers and not considering former smokers in the model building, and/or utilizing large numbers of CpGs [21-27]. Reliable studies on the validated inference of a person's smoking habits and history from blood with a finite set of DNA methylation markers and based on statistical models with large underlying data are not available as of yet. A finite number of DNA methylation markers achieving maximal prediction accuracy would be especially beneficial for those practical applications where - due to limited DNA quality and quantity, a common problem in forensics - it is impossible to apply standard DNA methylation microarray technology [17].

With this study, we aimed to identify a robust, finite set of DNA methylation markers in blood and, based on this finite biomarker set, develop accurate, reliable and validated statistical models for inferring a person's tobacco smoking habits and history from blood, which we envision becoming useful in future epidemiology and public health research as well as medical and forensic applications.

METHODS

Study population

This study was embedded within the Biobank-based Integrative Omics Study (BIOS) Consortium [28], which consists of six Dutch cohorts (N=3,118), including the Rotterdam Study (RS) (N=584) [29], Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) (N=156) [30], The Netherlands Twin Register (NTR) (N=894) [31], Leiden Longevity Study (LLS) (N= 625) [32], Prospective ALS Study Netherlands (PAN) (N=167) [33] and LifeLines (LL) (N=692) [34]. Additionally, we included another 646 unrelated participants from the Rotterdam Study (RS-III-1) not included in BIOS. We externally validated our model in the Kooperative Gesundheitsforschung in der Region Augsburg (KORA) study (F4, N=1,608) [35], as well as in the Study of Health in Pomerania (SHIP)-Trend (N=244) [36] cohort. Characteristics of all cohorts used can be found in Online Resource 1: **Table S1**. We additionally tested our model in samples from children included in the Generation R Study [37], in particular, we used data from children participating at birth (N=1,111), at the age of six years (N=355), and at the age of ten years (N=309), of which 197 overlapped between all three time points, providing longitudinal data (Online Resource 1: **Table S2**). The smoking status information was obtained using questionnaires. The study characteristics are described in detail in Online Resource 2: **Supplemental methods**.

DNA methylation quantification

DNA was extracted from whole peripheral blood in all studies using standard procedures. All studies used the Illumina Infinium Human Methylation450K BeadChip (Illumina Inc, San Diego, CA, USA) for epigenome-wide DNA methylation measurements, except the SHIP-Trend study, which used the more recent Infinium MethylationEPIC BeadChip (Illumina Inc, San Diego, CA, USA). DNA methylation data pre-processing for cohorts included in the BIOS consortium were conducted together via the pipeline created by Tobi *et al.*[38, 39]. The DNA methylation data pre-processing in the external validation cohorts and the Generation R Study were done independently. The methylation proportion of a CpG site was reported as a methylation β-value in the range of 0 (representing completely non-methylated sites) to 1 (representing completely methylated sites). Further study-specific methods can be found in Online Resource 2: **Supplemental methods**.

Ascertainment of smoking-associated CpGs

EWASs using the Illumina Infinium Human Methylation 27K or 450K BeadChip investigating smoking-induced changes in DNA methylation patterns were reviewed [2, 21, 40-50]. We excluded studies [11] that used cohorts included in our model- building dataset, to avoid over-estimation of our model. Envisioning future laboratory tool development, we only selected robust CpGs hat were (1) highlighted in two or more studies, (2) with at least 10% difference in mean or median (depending on availability per EWAS) β- values between current smokers and never-smokers (or non-smokers when non-smoking data was available) in at least one of the studies, and (3) with the same direction in β- value difference between current smokers and never/non-smokers in all studies investigated.

Statistical modeling for current smoking habits

Of the total participants considered for model building (N_{total}=5,178), we excluded those with (1) missing data for smoking habits (1,206 participants), (2) missing β- values for the predictive CpGs (82 participants), or (3) extreme outliers for one or more CpGs (mean +/- 4 SD) (126 participants). In the end, we included 3,764 participants in the final model building set, who were then categorized based on their smoking habits as (1) current smokers or (2) former and never smokers combined. The association between the candidate CpGs and smoking habits (smokers vs non- smokers) was replicated in our model building dataset using binomial regression analysis adjusted for age and sex using the “glm” function with “binomial” as family and “logit” as link. To identify the most informative set of DNA methylation predictors from the candidate CpGs, the association between the complete set of predictive CpGs and smoking habits was assessed in a binary logistic regression analysis, using the “glm” function with “binomial” as family and “logit” as link. Backward elimination procedures were used for the marker selection process. We excluded the CpGs one by one based on their absolute z-statistic

per regression (calculated by dividing the regression coefficient by its standard error) assessed using the "VarImp" function (r-package "caret"). The predictive CpG with the lowest absolute z-statistic in the regression was removed. The model was applied to the dataset with the "predict" function (type= "response") and the confusion matrix (r-package "caret") was conducted using a probability threshold of 0.5. The prediction performance of the model was additionally assessed using "prediction" and "performance" (r-package "ROCR"), the Area Under the Curve (AUC) per model was calculated (r-package "ROCR") and a cumulative AUC profile was conducted for each model to obtain a cumulative AUC profile. We selected the best-fit prediction model using a combination of the backward elimination approach and the Chi-squared test. In particular, we compared the model including all CpGs ($model_{FULL}$) with the model excluding one CpGs, ($model_{FULL\text{-}1CpG}$), this model $_{FULL\text{-}1CpG}$ was then compared with the model excluding another CpG ($model_{FULL\text{-}2CpGs}$), following the same order as conducted via the backward approach, and so on until we noticed a statistically significant difference between two models in the backward approach. Subsequently, we tested the inclusion of age, sex and cell counts to the final model.

Former smokers as additional category

Participants included in the model building dataset (N =3,764) without additional smoking data, including the age someone stopped smoking (former smokers) or the age someone started smoking or the number of cigarettes someone smokes per day (current smokers), were excluded, resulting in a dataset including 2,939 participants. The association between the previously selected predictive CpGs and the three smoking categories was assessed in a multinomial regression analysis, using the "multinom" function (r-package "nnet"). We predicted the smoking categories using the "predict" function (type= "class") and the confusion matrix (r-package "caret") was conducted. The AUC per category was conducted using the "predict" function (type= "probs") and "roc" function (r-package "pROC").

Smoking cessation time inference in former smokers

In the former smokers (N=1,332), smoking cessation time was calculated as one's age minus the age one stopped smoking. The participants were split into two categories for three models. For model 1, ≥5 years cessation time were coded as "1" and <5 years smoking cessation were coded as "0", for model 2, ≥10 years cessation time were coded as "1" and <10 years smoking cessation were coded as "0", and for model 3, ≥15 years cessation time were coded as "1" and < 15 years smoking cessation were coded as "0". The predictions were conducted using the same method as described for the current *versus* non- smokers model. Probability thresholds were set to 0.8733, 0.7650 and 0.6397 respectively.

Pack-year inference in current smokers

For the current smokers (N=364) the pack-years were calculated as the number of cigarettes smoked per day divided by 20, multiplied by the total years of smoking. The participants were categorized into two categories for two models. For model 1, ≥15 pack-years were coded as "1" and <15 pack-years coded as "0", for model 2, ≥10 pack- years were coded as "1" and <10 pack-years coded as "0". The predictions were conducted using the same method as described for the current *vs* non-smokers model.

Pack-years (current-), smoking cessation time (former-) and never smokers

We combined the pack-year inference in current smokers with the cessation time in former and never smokers, resulting into five categories in two models (N=2,939) for inferring life-time smoking information. For model 1, the current smokers ≥ 15pack-years were coded as "5", with < 15 pack-years were coded as "4", the former smokers ≤ 10 years smoking cessation were coded as "3", with > 10 years smoking cessation were coded as "2" and never smokers were coded as "1". In the second model the same categories were used except for the pack-years which were now divided in ≥ 10 pack-years (coded as "5") and < 10 pack-years (coded as "4"). The predictions were conducted using the same method as described for the current *vs* former *vs* never smokers model.

Internal validation of the developed prediction models

For internal validation of the developed predictive models, we adopted a fivefold cross-validation scheme [51], in which the whole dataset is first randomly distributed into five equal and non-overlapping subsets. Four of the subsets (80% of the data) are combined to form a dataset used to train the logistic regression model which is then tested by inferring the smoking habits in the remaining dataset (20% of the data). This resulted in five different training (80%) and testing (20%) sets. The model was trained in the five training sets and applied to corresponding testing sets, resulting in five logistic regression models. Subsequently, we used the bootstrap method (r- packages "boot" and "parallel") as additional internal validation to correct for potential overestimation of the prediction, since we use the same data for model building and predictions. We generated 1,000 bootstrap samples, with replacement from the dataset for which we estimated the model and applied each fitted model to the original sample, resulting in 1,000 AUC estimates. Thereafter, we recalculated the prediction accuracy by applying the fitted model to the bootstrap sample itself. The performance in the bootstrap sample represents an estimation of the apparent performance, and the performance in the original sample represents test performance. The difference between the average of the two conducted AUCs is a stable estimate of the optimism. We corrected for prediction overestimation by subtracting the optimism from the apparent AUC, to obtain an improved estimate of the prediction AUC [52, 53].

External validation of the developed prediction models

We externally validated our prediction models in two independent cohorts from German- European origin. The full models were validated in the KORA F4 study (N=1,608). Additionally, we externally validated our models in the SHIP-Trend study (N=244). In this cohort, the EPIC methylation array was used which does not include all CpGs of the 450K array. We therefore first generated the prediction models based on the overlapping CpGs in the model building dataset and subsequently externally validated them in the SHIP-Trend dataset.

Comparing performance of CpG-based model with cotinine level cut-off

We compared the outcomes of the CpG model to infer current *vs* non-smokers with the outcomes using a cotinine level cut-off of 50 ng/mL [54, 55] and applied smoking information from self-reports as reference. We employed a subset of our model building dataset (N=488 participants included in NTR [56]) in which both DNA methylation levels and cotinine levels were available. First, participants were categorized as smokers when their plasma cotinine levels were > 50 ng/mL, or as non- smokers with cotinine levels ≤ 50 ng/mL, threshold according to previous studies including the used cotinine data [54, 55]. Second, the current *vs* non- smokers CpG model was applied to this subset, obtaining the inferred smoking status for the participants. Third, we compared the obtained smoking status for both models with the information obtained from the self-reported questionnaires and computed the sensitivity and specificity per model.

Application of the developed prediction model in newborns and young children

Studies have shown the impact of prenatal smoking exposure on the DNA methylation pattern of the offspring [57] and the ability of predicting maternal smoking status using these patterns [58]. In this context, we wanted to test the effect of prenatal exposure on model application in adults. Hence, when an adult does not smoke, but was exposed to prenatal smoking, do we predict this person indeed as a true non-smoker? To test for this putative impact of exposure to prenatal smoking on epigenetic inference of smoking habits using our model, we tested our model in umbilical cord blood of newborns (N=1,111), and in whole blood of children at the ages of six (N=355) and 10 years (N=309). We used five different analyses to evaluate the effects of active smoking of the mothers and passive smoking of the mothers (i.e. smoking of others in the mother's home and work environment) during pregnancy on smoking habit inference using our model. In our first analysis, we did not take the smoking habits of the pregnant mothers or others in the pregnant mother's home and work environment into account and all children were coded as non-smokers. The proportion of accurately predicted cases was calculated using a probability threshold of 0.5. In each of the following analyses, we coded

the children "1" if their parents met the smoking habit criteria, otherwise they were coded as "0". So, in the second analysis, only sustained maternal smoking throughout pregnancy was considered. Therefore, the children of mothers that smoked during the whole pregnancy were coded as "1". In the third analysis, we additionally included the children of mothers who stopped smoking when they realized that they were pregnant by coding these children as "1". In the fourth analysis, we additionally included smoking of the father and/or others in the mother's household / at work (> 1h per day) during pregnancy (i.e. passive smoking). In the fifth analysis, we assessed the sole effect of passive smoking i.e., where the mother did not smoke but the father or someone else in the house or at work (> 1h per day) smoked during the pregnancy of the mother. For 197 children, DNA methylation levels were measured at all three time points, i.e. birth, 6 years of age and 10 years of age; hence, we repeated the previous models again in these children to allow a direct comparison of the findings at these three time points in the same individuals.

RESULTS

Ascertaining candidate DNA methylation markers for inferring smoking habits from blood

We inspected 14 published EWASs on tobacco smoking habits (N_{total}= 7,015) [2, 21, 40-50] to identify smoking-associated CpGs as candidate DNA methylation markers for prediction modeling of smoking habits. CpGs were selected as candidate prediction markers if they met three criteria as mentioned in the method section. This procedure highlighted 20 top smoking-associated CpGs as candidate markers used for further analyses (**Table 1**). The differences in β-values between smokers and never- /non-smokers reported previously for these 20 top smoking-associated CpGs are illustrated in **Figure 1**.

Building CpG-based models for inferring smoking habit and history from blood

Following the replication of the association between the CpGs and smoking habits (smokers *vs* non- smokers) after adjusting for age and sex (Online source **Table 3**), we assessed the predictive effect of the selected 20 candidate markers in the model building dataset (N=3,764). Starting with a model including all 20 CpGs, the CpG with the lowest z-value per model was sequentially removed, and the AUC was calculated for each model to obtain a cumulative AUC profile (**Table 1**, **Figure 2**).

Table 1. Top 20 smoking-associated CpGs from 14 previous EWASs considered here for marker sub-selection and their contributions to smoking inference from blood.

CpG ID	Chr:position♦♦	Gene ID♦♦♦	Location of CpG	Cumulative AUC
cg05575921♦	5:373,378	*AHRR*	Gene body	0.8801
cg13039251♦	5:32,018,601	*PDZD2*	Gene body	0.8888
cg03636183♦	19:17,000,585	*F2RL3*	Gene body	0.8883
cg12803068♦	7:45,002,919	*MYO1G*	Gene body	0.8889
cg22132788♦	7:45,002,486	*MYO1G*	Gene body	0.8934
cg06126421♦	6:30,720,080	NA	-	0.8929
cg21566642♦	2:233,284,661	NA	-	0.8957
cg23576855♦	5:373,299	*AHRR*	Gene body	0.8967
cg15693572♦	3:22,412,385	NA	-	0.8982
cg05951221♦	2:233,284,402	NA	-	0.8989
cg01940273♦	2:233,284,934	NA	-	0.8998
cg12876356♦	1:92,946,825	*GFI1*	Gene body	0.9005
cg09935388♦	1:92,947,588	*GFI1*	Gene body	0.9010
cg19572487	17:38,476,024	*RARA*	5'UTR	0.9012
cg19859270	3:98,251,294	*GPR15*	Gene body (1st Exon)	0.9015
cg18146737	1:92,946,700	*GFI1*	Gene body	0.9015
cg21161138	5:399,360	*AHRR*	Gene body	0.9015
cg23480021	3:22,412,746	NA	-	0.9016
cg21188533	3:53,700,263	*CACNA1D*	Gene body	0.9015
cg03274391	3:22,413,232	NA	-	0.9015

♦ CpGs included in our final 13 CpG model
♦♦ Genome coordinates provided by Illumina (GRCh37/hg19)
♦♦♦ According to the Illumina Infinium HumanMethylation450K annotation file
NA not annotated to any gene according to the Illumina InfiniumHumanMethylation450K annotation file
AUC Area under the Curve

To identify the minimal number of CpGs required to achieve maximum prediction accuracy, we additionally used Chi-squared tests. Applying this backward approach, the first significant difference between two models was noted when we compared the model with and without cg09935388 (**Table 1**, **Figure 2**). The combined marker elimination approach resulted in a finite set of DNA methylation markers comprising 13 CpGs (**Table 1**, **Figure 2**). The AUC for the identified 13-CpG model was 0.901 for distinguishing between smokers versus non-smokers (for other prediction accuracy measures, see **Table 2**). The remaining 7 CpGs raised the cumulative AUC only on the 4th decimal i.e. from 0.9010 to 0.9016 (**Table 1**, **Figure 2**). Hence, this finite set of 13 CpGs was used for subsequent prediction modeling. Using the 13-CpG model, we inferred the smoking status of the participants included in our model building dataset; the inferred probabilities are presented in a histogram in **Figure 3**, where each probability bin is overlaid with the percentage of accurately inferred smoking habits in that probability range.

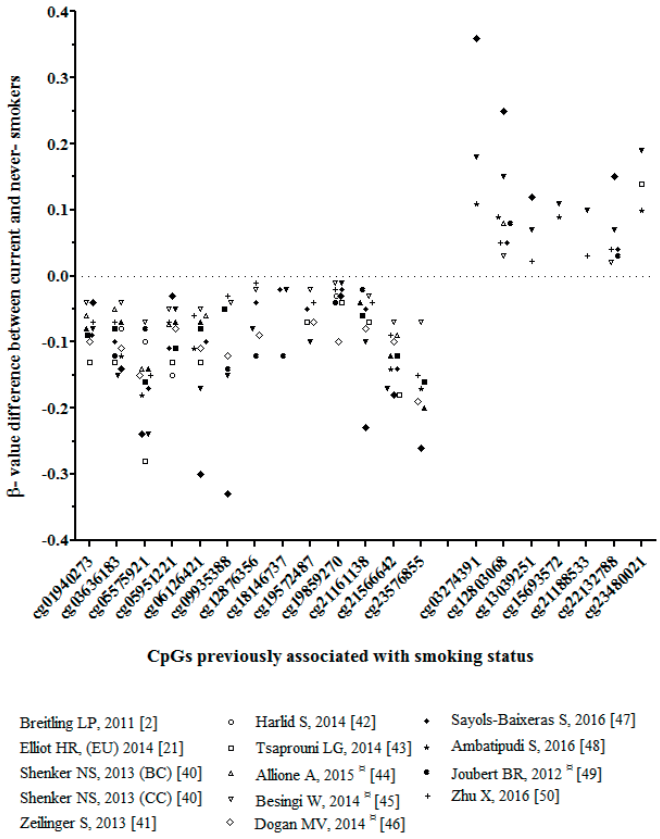


Figure 1. DNA methylation β- value differences between smokers and never-smokers for the top 20 smoking-associated CpGs. Previously reported differences in β-values in mean or median (depending on availability per EWAS) between smokers and never-smokers (or non-smokers ¤ when non-smoking data was available) for the selected 20 top-associated CpGs obtained from the 14 reviewed EWASs on smoking habits that did not include samples used here for model building.

Table 2. Outcomes of the two-category-model (smokers vs. non-smokers) for inferring smoking habits from blood based on CpGs.

	13-CpG model			10-CpG model♦		
	Model building data set (N=3,764)		External validation	Model building data set (N=3,764)		External validation
	model building	five-fold cross-validation	KORA (N=1,608)	model building	five-fold cross-validation	SHIP-Trend (N= 244)
Accuracy$ (95% CI) ± SD	0.923 (0.914, 0.931)	0.921±0.008	0.926 (0.912, 0.938)	0.917 (0.908, 0.926)	0.917±0.011	0.873 (0.825, 0.912)
Specificity	0.976	0.976±0.005	0.983	0.975	0.975±0.006	0.995
Sensitivity	0.585	0.577±0.044	0.580	0.548	0.551±0.042	0.412
AUC	0.901	0.897±0.137	0.911	0.896	0.893±0.012	0.888

♦ three CpGs (cg06126421, cg22132788 and cg05951221) are not included in the EPIC methylation microarray dataset from SHIP-Trend, this model is included here to demonstrate a second external validation in SHIP next to KORA with the full 13-CpG model. $ proportion accurately inferred smoking habits, 95% confidence interval (CI). Cross-validation analysis results are presented as mean ± standard deviation. *AUC* Area under the Curve

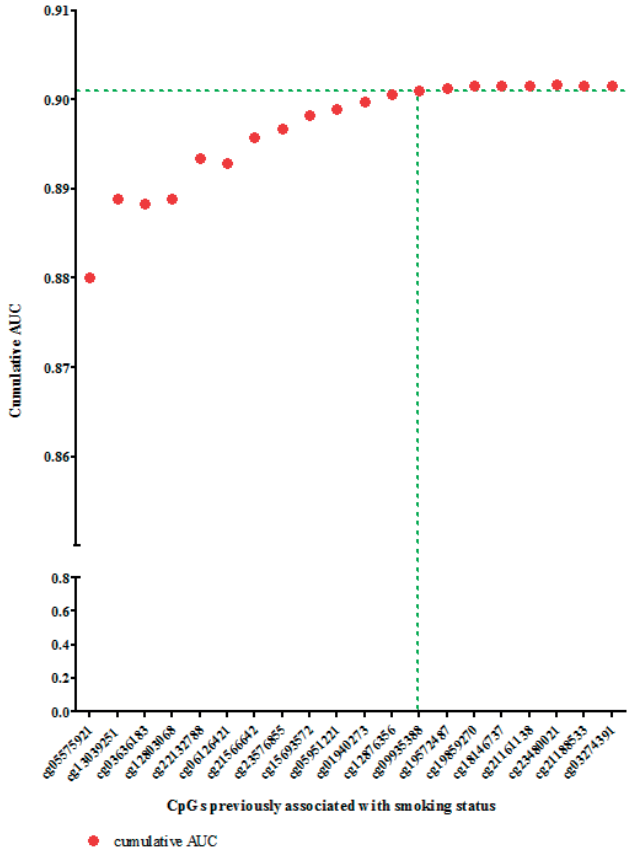


Figure 2. Cumulative AUC profile for smoking habit inference from blood based on the top 20 CpGs.
The 20 CpGs were selected from previous EWASs on smoking habits (see Figure 1) and were tested in the model-building set (N=3,764). Presented is the cumulative contribution of each of the selected 20 CpGs to the model-based smoking habit inference, shown as the AUC plotted against the number of CpGs included in the binary logistic regression model. In the model selection process, first all CpGs were included, and using backward elimination procedures, those with the lowest z-statistic per model were removed one by one. After 13 CpGs, the AUC plateaus; therefore, and by considering the results from Chi^2 testing, these 13 CpGs were used for further analyses.

Adjusting the prediction model for age resulted in a minor AUC increase from 0.901 to 0.907, adjusting for sex from 0.901 to 0.903 and including both age and sex in the model increased the AUC slightly from 0.901 to 0.911 (Online Resource 1: **Table S4**). Additionally, we tested the influence of cell counts on the model accuracy. In the subset of participants for which cell count measures were available (N=3,402), our 13-CpG model without cell counts achieved an AUC of 0.906. Including the cell count measurements for monocytes, granulocytes and lymphocytes in our 13-CpG model, the AUC was almost identical at 0.907 (Online Resource 1: **Table S5**). Since age, sex and cell counts only had a minor impact on the prediction accuracy, these three non-epigenetic factors were not considered in the final model used in the subsequent analyses. Next, we considered

former smokers as an additional, separate category in the prediction model building based on the finite set of 13 CpGs, resulting in a three-category prediction model. To this end, we considered a subset of 2,939 participants for which the relevant smoking habit information was available. We obtained for the current smokers (N=364) an AUC of 0.928, for the former smokers (N=1,332) 0.772, and for the never smokers (N=1,243) 0.835 (for other accuracy measures, see **Table 3**).

Table 3. Outcomes of the three-category-model for inferring smoking habits from blood based on CpGs.

Model building data set (N=2,939): model building 13-CpG model			
	Never (N=1,243)	**Former (N=1,332)**	**Current (N=364)**
Specificity	0.746	0.770	0.997
Sensitivity	0.780	0.652	0.668
AUC	0.835	0.772	0.928
Model building data set (N=2,939): five-fold cross-validation 13-CpG model			
	Never (N=1,243)	**Former (N=1,332)**	**Current (N=364)**
Specificity	0.739±0.017	0.766±0.053	0.975±0.008
Sensitivity	0.769±0.060	0.643±0.039	0.669±0.056
AUC	0.830±0.019	0.766±0.023	0.925±0.021
External replication in KORA (N=1,608): 13-CpG model			
	Never (N=675)	**Former (N=707)**	**Current (N=226)**
Specificity	0.539	0.870	0.980
Sensitivity	0.916	0.392	0.615
AUC	0.781	0.699	0.914
Model building data set (N=2,939): model building 10-CpG model♦			
	Never (N=1,243)	**Former (N=1,332)**	**Current (N=364)**
Specificity	0.749	0.737	0.974
Sensitivity	0.751	0.648	0.626
AUC	0.825	0.753	0.922
Model building data set (N=2,939): five-fold cross-validation 10-CpG model♦			
	Never (N=1,243)	**Former (N=1,332)**	**Current (N=364)**
Specificity	0.745±0.013	0.735±0.042	0.975±0.010
Sensitivity	0.747±0.050	0.645±0.026	0.627±0.025
AUC	0.823±0.018	0.748±0.023	0.919±0.019
External replication in SHIP-Trend (N=244): 10-CpG model♦			
	Never (N=101)	**Former (N=92)**	**Current (N=51)**
Specificity	0.490	0.822	0.990
Sensitivity	0.891	0.315	0.451
AUC	0.778	0.654	0.882

Cross-validation analysis results are presented as mean ± standard deviation

AUC Area under the Curve

♦ three CpGs (cg06126421, cg22132788 and cg05951221) are not included in the EPIC methylation microarray dataset from SHIP-Trend

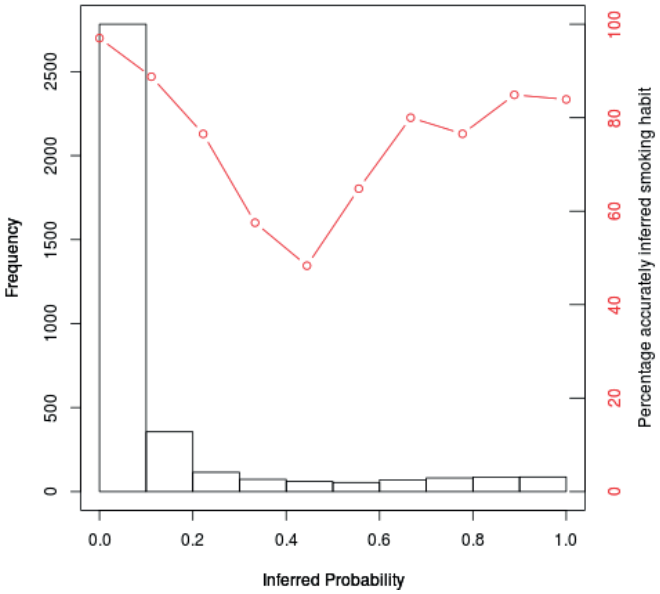


Figure 3. Inferred probability of being a smoker versus the percentage of correctly inferred smoking habits. Histogram of predicted probabilities in our model building dataset (N=3,764), probabilities determined using the 13 CpGs included in the final prediction model. The y-axis presents the number of individuals for whom the predicted probability of being a smoker was within the given probability range (x-axis). The red dots present the percentage of individuals in each probability bin that were accurately inferred using a > 0.5 probability threshold for being a smoker.

Additionally, we calculated smoking cessation time for the former smokers (N=1,332), and used the 13-CpGs to infer smoking cessation for ≥5 years (N=1160) *vs* <5 years (N=172), which resulted in an AUC of 0.793, for ≥ 10 *vs* <10 years smoking cessation time (N=1028 and N= 304, respectively) an AUC of 0.778 was obtained and for ≥15 *vs* <15 years smoking cessation time (N=887 and N= 445, respectively) an AUC of 0.779 was obtained (**Table 4**). Furthermore, for the current smokers (N=364) we calculated the pack-years (see methods) and used the 13 CpG markers to infer pack-years for ≥ 15 pack-years (N=210) *vs* <15 pack-years (N=154), which resulted in an AUC of 0.815. For ≥10 *vs* <10 pack-years (N=246 and N= 118, respectively) an AUC of 0.846 was obtained (**Table 5**).

Finally, we combined the pack-years in current smokers, smoking cessation in former smokers with the never smokers (N=2,939) into one model for life-time smoking information inferring. We obtained for the current smokers with ≥15 pack-years (N=210) an AUC of 0.949, <15 pack-years (N=154) an AUC of 0.869, in former smokers with ≤10 years smoking cessation (N=311) an AUC of 0.793, with >10 years smoking cessation (N=1021) an AUC of 0.739 and the never smokers (N=1,243) an AUC of 0.835 (**Table 6**). We obtained for the current smokers with ≥ 10 pack-years (N=246) an AUC of 0.948, < 10 pack-years (N=118) an AUC of 0.863, former smokers with ≤ 10 years smoking cessation (N=311) an AUC of 0.794, with > 10 years smoking cessation (N=1021) an AUC of 0.739, and the never smokers (N=1,243) an AUC of 0.835 (**Table 6**).

Table 4. Outcomes of the two-category model for inferring smoking history in former smokers from blood based on 13 CpGs.

	Former <5y vs Former ≥5y			Former <10y vs Former ≥10y			Former <15y vs Former ≥15y		
	Model building data set (N=1,332)		External validation	Model building data set (N=1,332)		External validation	Model building data set (N=1,332)		External validation
	model building	five-fold cross-validation	KORA (N=652)	model building	five-fold cross-validation	KORA (N=652)	model building	five-fold cross-validation	KORA (N=652)
Accuracy[$] (95% CI) ± SD	0.725 (0.700, 0.749)	0.715±0.020	0.830 (0.799, 0.858)	0.730 (0.705, 0.753)	0.721±0.029	0.799 (0.766, 0.829)	0.732 (0.707, 0.756)	0.718±0.016	0.759
Specificity	0.715	0.691±0.090	0.494	0.694	0.682±0.063	0.471	0.663	0.644±0.033	0.449
Sensitivity	0.727	0.718±0.026	0.879	0.740	0.733±0.026	0.900	0.767	0.756±0.015	0.902
AUC	0.793	0.774±0.024	0.760	0.778	0.766±0.033	0.764	0.779	0.767±0.020	0.754

[$] proportion accurately inferred smoking habits, 95% confidence interval (CI)
Cross-validation analysis results are presented as mean ± standard deviation
AUC Area under the Curve

Table 5. Outcomes of model applications to infer smoking history in in current smokers (N=364) from blood based on CpGs.

	More or less than 10 pack-years					
	13-CpG model			10-CpG model♦		
	Model Building N=364	five-fold Cross-validation	KORA F4 N=224	Model Building N=364	five-fold Cross-validation	SHIP-Trend N= 41
Accuracy[$]	0.824	0.783±0.05	0.813	0.808	0.770±0.035	0.805
(95% CI)	(0.781, 0.862)		(0.755, 0.861)	(0.76, 0.847)		(0.651, 0.912)
Specificity	0.644	0.577±0.131	0.343	0.602	0.548±0.14	0.778
Sensitivity	0.911	0.882±0.045	0.899	0.907	0.879±0.046	0.813
AUC	0.846	0.800±0.068	0.796	0.834	0.809±0.039	0.837
	More or less than 15 pack-years					
	13-CpG model			10-CpG model♦		
	Model Building N=364	five-fold Cross-validation	KORA F4 N=224	Model Building N=364	five-fold Cross-validation	SHIP-Trend N= 41
Accuracy[$]	0.733	0.719±0.093	0.786	0.728	0.709±0.059	0.659
(95% CI)	(0.685, 0.778)		(0.726, 0.838)	(0.679, 0.773)		(0.494, 0.799)
Specificity	0.617	0.600±0.204	0.455	0.597	0.575±0.143	0.533
Sensitivity	0.819	0.805±0.042	0.894	0.824	0.808±0.035	0.731
AUC	0.815	0.767±0.102	0.752	0.786	0.757±0.077	0.779

♦ Three CpGs (cg06126421, cg22132788 and cg05951221) are not included in the EPIC methylation microarray dataset from SHIP-Trend.
[$]proportion accurately inferred smoking habits; 95% CI, confidence interval; AUC, Area under the Curve,
Cross-validation analysis results are presented as mean ± standard deviation.

Table 6. Outcomes of the five-category-model for inferring smoking habits and history from blood based on 13 CpGs.

Never vs Former >10 years vs Former =< 10 years vs <15 Pack-years vs >=15 Pack-years					
			Model building data set (N=2,939)		
	Never (N=1,243)	**F > 10y (N=1,021)**	**F ≤10y (N=311)**	**<15PY (N=154)**	**≥15PY (N=210)**
Specificity	0.712	0.777	0.979	0.987	0.967
Sensitivity	0.817	0.554	0.206	0.299	0.724
AUC	0.835	0.739	0.793	0.869	0.949
			Model building data set five-fold cross-validation		
Specificity	0.711±0.022	0.775±0.036	0.977±0.009	0.984±0.009	0.963±0.014
Sensitivity	0.809±0.047	0.545±0.040	0.199±0.042	0.274±0.128	0.695±0.064
AUC	0.832±0.014	0.731±0.026	0.779±0.018	0.855±0.046	0.947±0.016
			External replication in KORA (N=1,551)		
	Never (N=675)	**F > 10y (N=488)**	**F≤10y (N=164)**	**<15 PY (N=55)**	**≥15PY (N=169)**
Specificity	0.534	0.830	0.994	0.994	0.979
Sensitivity	0.927	0.299	0.122	0.018	0.728
AUC	0.788	0.650	0.791	0.710	0.955
Never vs Former >10 years vs Former =< 10 years vs <10 Pack-years vs >=10 Pack-years					
			Model building data set (N=2,939)		
	Never (N=1,243)	**F > 10y (N=1,021)**	**F ≤10y (N=311)**	**<10 PY (N=118)**	**≤10PY (N=246)**
Specificity	0.714	0.776	0.981	0.994	0.963
Sensitivity	0.817	0.554	0.193	0.220	0.772
AUC	0.835	0.739	0.794	0.863	0.948
			Model building data set five-fold cross-validation		
Specificity	0.709±0.023	0.774±0.034	0.980±0.006	0.992±0.003	0.960±0.008
Sensitivity	0.808±0.045	0.542±0.042	0.194±0.043	0.206±0.066	0.758±0.067
AUC	0.831±0.014	0.730±0.027	0.780±0.018	0.847±0.047	0.946±0.023
			External replication in KORA (N=1,551)		
	Never (N=675)	**F > 10y (N=488)**	**F ≤10y (N=164)**	**<10 PY (N=35)**	**≥10PY (N=189)**
Specificity	0.535	0.827	0.994	0.998	0.977
Sensitivity	0.926	0.299	0.110	0.000	0.683
AUC	0.788	0.651	0.791	0.694	0.943

Cross-validation analysis results are presented as mean ± standard deviation
AUC Area under the Curve, *F* Former smokers, *PY* pack-years

Validating CpG-based models for inferring smoking habit and history from blood

We validated the newly developed prediction models based on the 13 selected CpGs via both internal and external validation procedures. Internal validation was carried out in the model building set using fivefold cross-validation and bootstrapping. For the two-category model (smokers *vs* non-smokers), the optimism from bootstrap internal validation was 0.0032, resulting in a bootstrap-adjusted AUC of 0.898 (0.901-0.0032), see **Table 2** for other accuracy measures and cross- validation results. For the three-category model (smokers *vs* former smokers *vs* never smokers) the bootstrap conducted optimisms are 0.0032 for current smokers, 0.0063 for former smokers and 0.0036 for never smokers resulting in bootstrap adjusted AUCs of 0.925 (0.928-0.0032) for current smokers, 0.766 (0.772-0.0063) for former smokers and 0.831 (0.835-0.0036) for never smokers (**Table 3**).

For the smoking cessation time inference in former smoker, (1) for ≥5 *vs* <5 years smoking cessation the bootstrap optimism was 0.0170 resulting in a bootstrap-adjusted AUC of 0.776 (0.793-0.0170); (2) for ≥10 *vs* <10 years smoking cessation the bootstrap resulted in an optimism of 0.0112, giving a bootstrap-adjusted AUC of 0.767 (0.778 - 0.0112); (3) ≥15 *vs* <15 years smoking cessation the bootstrap resulted in an optimism of 0.0096, giving a bootstrap-adjusted AUC of 0.769 (0.779 - 0.0096) (**Table 4**). For the two pack-year models, (1) the bootstrap optimism for ≥15 *vs* < 15 pack- was 0.029 resulting in a bootstrap-adjusted AUC of 0.786 (0.815- 0.029); and (2) for ≥10 *vs* < 10 pack-years the bootstrap resulted in an optimism of 0.026, giving a bootstrap-adjusted AUC of 0.820 (0.846- 0.026) (**Table 5**). Finally, for the life-time smoking information inferring, we obtained for ≥15 pack-years a bootstrap optimism of 0.0034 resulting in a bootstrap-adjusted AUC of 0.946 (0.949- 0.0034), for <15 pack-years a bootstrap-adjusted AUC of 0.860 (0.869- 0.0091), for ≤10 smoking cessation a bootstrap-adjusted AUC of 0.782 (0.793- 0.0106), >10 years smoking cessation a bootstrap optimism of 0.0075 resulting in a bootstrap-adjusted AUC of 0.732 (0.739- 0.0075) and for never smokers a bootstrap-adjusted AUC of 0.831 (0.835- 0.0037) (**Table 6**). For the second five-category model, very similar results were obtained (**Table 6**).

External validation was performed in independent samples of two population-based studies, KORA and SHIP-Trend. In KORA (F4, N=1,608), an AUC of 0.911 was achieved for the full 13-CpG two-category model (**Table 2**). In SHIP-Trend (N=244), an AUC of 0.888 was obtained for the two-category model based on a subset of ten CpGs, since the EPIC-array applied for SHIP-Trend is missing three of the 13 CpGs (cg06126421, cg22132788 and cg05951221). This 10-CpG model in the model building set gave a cross-validated average AUC of 0.893±0.012 (**Table 2**). External validation of the three-category model in the KORA study (F4, N=1,608) achieved an AUC of 0.914 for the current smokers (N=226), 0.699 for the former smokers (N=707), and 0.781 for the never smokers (N=675) (**Table 3**).

The three-category model validation in SHIP-Trend for the 10-CpG model resulted in an AUC of 0.882 for current smokers (N=51), 0.654 for former smokers (N=92), and 0.778 for never smokers (N=101) (**Table 3**). For comparison, in the model building set, this three category 10-CpG model gave a cross-validated average AUC of 0.919±0.019 for current smokers, 0.748±0.023 for former smokers, and 0.823±0.018 for never smokers (**Table 3**). External validation of smoking cessation time inference in former smokers in the KORA study (N=652) resulted in an AUC of 0.760 for ≥5 *vs* <5 years of smoking cessation time, an AUC of 0.764 for ≥10 *vs* <10 years of smoking cessation time, and of 0.754 for ≥15 *vs* <15 years of smoking cessation time (**Table 4**). Furthermore, we externally validated the prediction of pack-years in the current smokers of the KORA study (F4, N=224) and obtained an AUC of 0.752 for inferring ≥15 *vs*<15 pack-years and an AUC of 0.796 for ≥10 *vs* <10 pack-years (**Table 5**). The pack-year validation in the current smokers of SHIP-Trend (N=41) for the 10-CpG model resulted in an AUC of 0.779 for ≥15 *vs* < 15 pack-years (AUC of 0.757±0.077 in the model building set) and an AUC of 0.837 for ≥10 *vs*<10 pack-years (AUC of 0.809±0.039 in the model building) (**Table 5**). The external validation of the five-category models in the KORA study resulted for the current smokers with ≥15 pack-years in an AUC of 0.955, for <15 pack-years an AUC of 0.710, for ≤10 years smoking cessation an AUC of 0.791, >10 years smoking cessation an AUC of 0.650 and for never smokers an AUC of 0.788. For the second five-category model, we obtained in the KORA study an AUC of 0.943 for ≥10 pack-years, of 0.694 for <15 pack-years, an AUC of 0.791 for ≤10 years smoking cessation, of 0.651 ≥10 years smoking cessation and an AUC of 0.788 for never smokers (**Table 6**).

Comparing CpG-based with cotinine-based inference of smoking habit

In a subset of 488 participants for which we had CpG, cotinine and smoking information available, we compared our validated CpG-based prediction model for current *vs* non-smokers with the use of a cotinine cut-off to determine current smoking, using smoking information from self-reported questionnaires as reference. Using our CpG-model, we accurately inferred 87 of the 140 smokers and 344 of the 348 non-smokers (sensitivity of 0.621 and specificity of 0.989) compared to 105 of the 140 smokers and 342 of the 348 non-smokers using the cotinine level cut-off of 50 ng/mL (sensitivity of 0.750 and specificity of 0.983). Out of the 87 accurately inferred smokers with our CpG model, 75 (86%) were also accurately selected as smokers based on cotinine, and out of the 105 participants correctly selected with cotinine as smokers, 75 (71%) were accurately inferred as smokers with our CpG model. For the non-smokers, out of the 344 accurately inferred with our CpG model, 340 (99%) were also selected with cotinine as non-smokers, and 340 (99%) out of the 342 accurately selected non-smokers with cotinine, were accurately inferred as non-smokers with our CpG model. Finally, when comparing all three methods (questionnaires/cotinine levels/DNA methylation prediction), 340 participants

were highlighted as non-smokers and 75 as smokers with all three methods, 12 were selected as smokers based on questionnaires and DNA methylation inference, 30 as smokers with both questionnaires and cotinine, 2 were determined as smokers with both cotinine and DNA methylation inference, whereas 23 were determined as smokers with questionnaires only, 2 as smokers with DNA methylation inference only, and 4 as smokers with cotinine only.

Investigating prenatal smoking exposure effects on CpG-based inference of smoking habit

Next, we investigated the putative effect of prenatal smoking exposure and passive smoking on the epigenetic inference of smoking habits achievable with our validated model. When applying our model to the DNA methylation data at time of birth collected from cord blood, the proportion of children accurately inferred as non-smokers was surprisingly low at 0.114 (N=1,111) (Online Resource 1: **Table S6**). We then classified children whose mothers smoked throughout pregnancy as "smokers", and obtained an AUC of 0.773, with a high sensitivity of 0.981 and a low specificity of 0.131. The AUC decreased to 0.664 when additionally considering mothers who stopped smoking when they became aware of their pregnancy (generally in the first trimester), and decreased further to 0.591 when additionally considering passive smoking of the mother during pregnancy; assessing the latter solely, an AUC of 0.460 was obtained, reflecting random prediction.

Additionally, we applied our model to data of children from the Generation R Study obtained from blood collected at the ages of 6 (N=355) and 10 (N=309) years. In contrast to the results for newborns obtained from cord blood, we found that the proportion of six- and ten-year-old children accurately inferred as non-smokers with our model was very high at 0.994 for both age groups (**Table 7**). This suggests no impact of prenatal smoking exposure nor passive smoking exposure during early childhood on the model performance. Subsequently, we applied our model to those 197 children for which epigenetic data were available from serial samples collected at birth, 6, and 10 years of age. The proportion of children that with our model accurately inferred as non-smokers at birth was 0.112, whereas it was 0.994 at six and 0.995 at 10 years of age, which was highly similar to the results obtained from the total datasets available for these three time points. The β-values per CpG for the model building set and the three time points in Generation R are shown in Online Resource 3: **Figures S1-15**.

DISCUSSION

In this study, we introduce a robust, finite set of DNA methylation markers and carefully validated statistical models based on reasonably large population-based data, which together allow accurate and reliable inference of a person's tobacco smoking habit and history from blood DNA.

Previous studies have identified numerous CpGs associated with tobacco smoking in blood, and showed that DNA methylation patterns of specific genes are modified by smoking habits [2, 21, 40-50]; here we took advantage of these EWASs as a marker discovery resource. From the 20 top smoking-associated CpGs consistently highlighted in previous EWASs and by using new population-based cohort data not overlapping with these previous EWASs, we identified a robust, finite set of 13 CpG markers as being most suitable for inferring a person's smoking habit from blood DNA. Eight of these 13 CpGs are annotated to five known genes i.e., *AHRR* (2 CpGs), *GFI1* (2), *MYO1G* (2), *F2RL3* (1) and *PDZD2* (1), while the remaining 5 CpGs are not annotated to any coding regions. The highest AUC (0.880) for a given CpG among the 13 biomarkers in the model was achieved for cg05575921, which, together with one other CpG in the model (cg23576855), is located in the *AHRR* gene. The *AHRR* gene was shown to interact with the aryl hydrocarbon receptor (AHR), the induction point for the xenobiotic pathway, which includes several P450 enzymes, and is responsible for the degradation of environmental toxins [59-61]. Notably, *AHRR* provides the strongest epigenetic response to tobacco smoking known today [59, 62].

While a few previous studies have investigated DNA methylation markers for inferring smoking habits from blood, they all suffered from one or more limitations, including small sample size, limited model validation, exclusion of the former smoker category from the prediction model building, using a large number of CpGs and others [21-26]. For example, Philibert *et al.* [23] reported on the performance of five CpGs yielding AUCs 0.86-0.99 but only using 61 subjects. Notably, all five CpGs were among the 20 markers investigated in our study and are also included in our final 13-CpG model. For cg05575921, Philibert *et al.* estimated an AUC of 0.99 [23]; when testing this DNA methylation marker in our model building set of 3,764 samples, a considerably lower AUC of 0.8801 was achieved. In another study, Elliot *et al*. [21] reported a methylation score based on 183 CpGs to distinguish between current, former and never smokers, with a sensitivity of 100% and a specificity of 97% using 96 subjects only. When generating the methylation score using the methods described by Elliot *et al*., and applying it to our model building set (N=3,764), we obtained a specificity of 0.864 and sensitivity of 0.747 with an AUC of 0.806, considerably lower than reported by Elliot *et al*. These two examples illustrate that previously reported prediction accuracies obtained from studies using small sample size likely reflect overestimation caused by small sample size.

Given the relatively larger sample size for model building and internal validation, and for external validation with independent samples as utilized here, our results demonstrate that the new 13-CpG model introduced here provides more robust and reliable accuracy outcomes than previously reported models.

Previous studies have shown that DNA methylation patterns can be altered by age, sex and various lifestyle factors other than tobacco smoking [63, 64]. Additionally, recent papers suggest that the change in DNA methylation measurements due to smoking are mainly caused by the smoking induced changes in cell types [65-68]. We therefore tested the impact of age, sex, and cell counts on the model performance and found that these covariates only provide a slight increase in the prediction accuracy our model provides. Notably, a model that does not consider sex, age, and cell counts is beneficial for those applications where (some of) this information is not easily available, such as in forensics.

A recent study reported that the DNA methylation of most CpGs returns to never smoker levels within 5 years of smoking cessation, while some do not go back completely [11]. Also, previous work demonstrated that there is an association between smoking cessation time and smoking pack-years with DNA methylation scores [65, 69]. We therefore tested to what degree the 13 selected CpGs can distinguish former smokers from current smokers and never-smokers, and how well they allow inferring smoking history such as smoking cessation time and pack-years. Our results demonstrate that our 3-category model allows as first the inference of the former smoking category (smoking cessation between 0.1 and 58.86 years) together with current smokers and never smokers and also a more in depth inference possibility for cessation time categories as of more *vs* less than 5, 10 and 15 years of smoking cessation, although not as accurately as current and never smokers, as may be expected. The 13 CpGs also allowed accurate prediction of the pack-years in current smokers with a high AUCs for distinguishing between more or less than 10 pack-years, and for distinguishing between more or less than 15 pack-years. Finally, we show, to the best of our knowledge, for the first time an inference model able of inferring life-time smoking information in one model including the never smokers, cessation time in former smokers and pack-years in current smokers. Thus, the finite set of 13 DNA methylation markers and models we introduce here not only allow inferring information on current smoking or non-smoking status, but additionally provide information on former smoking and cessation time, smoking intensity in current smokers, and can additionally, as the first model to date, also provide complete life-time smoking information as of five different smoking categories.

Cotinine is the primary metabolite of nicotine and is therefore used as a reliable measurement for current smoking [19]. However, due to the short half-live of cotinine (between 15-19h), a false-negative prediction of current smoking can be easily obtained when there is a long time between the last cigarette and blood drawn [19]. In addition, former smokers that use nicotine replacement therapy to reduce the motivation to

smoke and for nicotine withdrawal symptoms, might result in false-positive predictions since cotinine, nicotine's metabolite, will still be traceable [20, 70]. Finally, due to protein instability over time, cotinine levels would only be accurately measurable in fresh blood samples, which are not always available such as in forensic investigations. Zhang *et al.* [24] showed that both DNA methylation and cotinine can accurately distinguish current from never smokers, but also emphasized that only DNA methylation is able to provide more in depth life-time smoking information. In line with this, we show in the current study that using both cotinine (sensitivity 0.750, specificity 0.983) and DNA methylation (sensitivity 0.621, specificity 0.989) we can infer current smokers with high accuracy. However, the sensitivity of our CpG model is slightly lower than the use of the cotinine cut-off in this subset. Nonetheless, with the upcoming availability of DNA methylation data in large cohort studies, the availability of a reliable smoking inference model, giving extending life-time smoking information inference, would be more widely accessible than information on cotinine levels.

Maternal smoking during pregnancy has been shown to influence fetal DNA methylation patterns [57, 71], which in principle could affect epigenetic inference of smoking habits in adults. Additionally, it is shown that maternal smoking status can be predicted from DNA methylation retrieved from newborns [72, 73]. Therefore, we employed data from the Generation R study to test the influence of prenatal smoking exposure on the inference of smoking status in adolescence. Hence, we tested our prediction model using epigenetic data from cord blood collected at time of birth, and peripheral blood collected at 6 and 10 years of age [37]. Our results showed that at the age of 6 years, 353 of the 355 children were correctly inferred as non-smokers (accuracy of 0.994), and at the age of 10 years 307 of the 309 children (accuracy of 0.994) were correctly inferred as non-smokers. This might indicate that prenatal smoking exposure and passive smoking exposure does not affect DNA methylation levels to such an extent that they are detected with our inference model. At time of birth, our model incorrectly inferred 984 (88.57%) of the 1,111 children as smokers (accuracy of 0.114). To test whether the newborns were inferred wrongly as smokers due to prenatal smoking exposure, we further classified the newborns as smokers when their mothers smoked throughout pregnancy (N=161). This resulted in a high AUC (0.773), with high sensitivity (0.981) but low specificity (0.131). Retrieving this low specificity while correcting for prenatal smoking exposure may indicate that the incorrect smoking inference of newborns achieved with our model can only in part be explained by smoking exposure during pregnancy. Other explanations may be developmental effects, and perhaps the tissue difference between whole blood and cord blood and therefore the difference in cell composition, given that the applied model was developed using whole blood [74]. Previous studies have shown specific changes in DNA methylation during early childhood that were explained by developmental effects [71, 75]. In any case, given that envisioned applications of epigenetic inference of smoking

habit in medical and forensic practice, as well as in most epidemiological and public health research, are typically performed in adults, our findings in children of advanced age imply that our model will indeed deliver smoking habit information of the adult individual tested, independent of prenatal smoking exposure or other effects.

The main strengths of our study are (1) the use of robust DNA methylation markers highlighted in multiple epigenome-wide association studies, (2) the use of independent population-based studies for marker discovery, model building and external model validation, and (3) the employment of thousands of samples for model building and validation. We therefore expect that the high prediction accuracy (AUC of 0.911) obtained from the full 13-CpG model in the KORA samples used for external validation reflects a realistic characterization of the performance of our model. This is also supported in part by the SHIP-Trend outcomes (AUC of 0.888) of the partial 10-CpG model. As the Illumina 450K array on which our marker selection was initially based is no longer available, the SHIP-Trend results using 10-CpG subset from the current Infinium MethylationEPIC BeadChip indicate that this sub-model would be applicable to new studies moving forward.

This study, however, does not come without limitations. Our model is based on smoking habit data retrieved from self-reported questionnaires, which are generally considered unreliable in terms of underestimating actual smoking levels [15]. Regarding the putative inaccuracy of self-reported smoking habits used here as phenotypes, we cannot know how error-prone these reports are. In particular, it is possible that specific groups of volunteers, for instance pregnant women such as those involved in the Generation R Study, are more reluctant to confide that they smoke [16]. However, we did not use the Generation R Study data for model building or validation purposes. Moreover, we included cotinine data to confirm the self-reported smoking habits for subset of participants (N = 488). Overall, we expect that smoking phenotype inaccuracy did not strongly impact the performance outcomes of our models. Lastly, all but one of the studies included in the model building and model validation are population-based studies, which therefore can include participants with various diseases. Though, due to the large sample sizes used for model building and validation, we expect that disease status does not strongly impact our model performance. Another limitation for the pack-year model is the formula used to calculate the pack-years. For this estimation, the number of cigarettes the participant currently smokes is used, which might have changed over the life span, and if so, this phenotypic variation is not considered. Additionally, the start-age is used to calculate the number of years someone smoked or has been smoking, which might be prone to recall bias especially for elderly people.

We envision that future works may provide targeted laboratory tools for analysing the 13 CpGs included in our final model in different types of blood samples and possible translation to different tissues, as is recently already shown to be promising for our top hit CpG (cg0557592) in saliva [76]. This would enhance the spectrum of practical

applications of epigenetic smoking habit inference. Given the finite set of DNA methylation markers introduced here, it is impractical to apply genome-wide DNA methylation microarrays just for the purpose of analyzing 13 CpGs. Moreover, there can be blood samples where microarrays do not produce reliable DNA methylation data, such as when the amount of DNA is low and/or the DNA is degraded such as DNA obtained from crime scene traces [17]. Hence, the future development of a fast and cheap laboratory tool that allows the reliable targeted analysis of the 13 CpGs highlighted here by employing a technology that can handle low quality and/or quantity DNA would be valuable. Foreseeing the future development of such a lab tool, we only included CpGs with at least a β-value difference ≥10% in mean or median (depending on availability per EWAS) in at least one published EWAS, to ensure detectability of the DNA methylation differences with targeted analysis technologies currently available [77, 78]. We view the positive results on epigenetic inference of smoking habits from blood presented here as a promising starting point for inferring more lifestyle factors using DNA methylation markers within the concept of epigenetic fingerprinting [17]. This requires continuous progress in identifying candidate DNA methylation predictors of lifestyle factors via dedicated EWASs, the subsequent use of these biomarkers in prediction modeling and validation studies to generate reliable and accurate models such as that reported here for tobacco smoking, and the development of robust and sensitive lab tools that allow the successful analysis of the DNA samples of interest, including those of limited quality and quantity.

SUPPLEMENTARY MATERIAL

Table S1. Characteristics of the dataset used for model building, internal (N=3,764) and external validation (N=1,852).

Study		No. of individuals	Females (%)	Average age (SD)
Model building and internal validation dataset (Dutch Europeans)				
Two-category model	*Total*	3,764	2,148 (57.07)	53.65 (15.45)
	Smokers	511	304 (59.49)	49.66 (15.36)
	Non-smokers♦	3,253	1,844 (56.69)	54.27 (15.38)
Three-category model	*Total*	2,939	1,670 (56.82)	55.76 (15.17)
	Smokers	364	218 (59.89)	52.53 (14.51)
	Former smokers	1,332	646 (48.50)	61.06 (10.75)
	Never smokers	1,243	807 (64.92)	51.02 (17.40)
External model validation dataset (German Europeans)				
KORA F4	*Total*	1,608	831 (51.68)	60.93 (8.83)
Two-category model				
	Smokers	226	106 (46.90)	57.02 (6.79)
	Non-smokers♦	1,382	725 (52.46)	61.57 (8.96)
Three-category model				
	Smokers	226	106 (46.90)	57.02 (6.79)
	Former smokers	707	277 (39.18)	61.02 (8.96)
	Never smokers	675	448 (66.37)	62.14 (8.93)
SHIP-Trend	*Total*	244	127 (52.0)	51.3(13.8)
Two-category model				
	Smokers	51	29 (56.9)	45.7 (11.8)
	Non-smokers♦	193	95 (50.8)	52.8 (13.9)
Three-category model				
	Smokers	51	29 (56.9)	45.7 (11.8)
	Former smokers	92	31 (33.7)	53.3 (13.3)
	Never smokers	101	67 (66.3)	52.3 (14.5)

♦Non- smokers: Former smokers and never smokers combined
SD standard deviation.
SHIP –Trend Study of health in Pomerania,
KORA (F4) Kooperative Gesundheitsforschung in der Region Augsburg F4 Study
Model building and validation set contains participant data from 6 cohorts; Cohort on Diabetes and Atherosclerosis Maastricht (*CODAM*), LifeLines (*LL*), Leiden Longevity Study (*LLS*), Netherlands Twin Register (*NTR)*, Prospective ALS Study Netherlands (*PAN*), Rotterdam Study (*RS*).

Table S2. Characteristics of Generation R Study data used for investigating the effect of prenatal smoking exposure.

Sample set	No. of individuals	Females (%)	Maternal smoking during pregnancy			Paternal / other smoking during pregnancy		
			Whole pregnancy	Until first trimester	Never	Yes	No	♦Missing
Birth	1,111	545 (49.1)	161	106	844	465	598	48
6 years old	355	185 (52.3)	46	39	270	130	194	31
10 years old	309	150 (48.5)	35	37	237	112	170	27
Serial samples♦♦	197	95 (48.2)	24	26	147	73	119	5

♦ Participants without information on paternal / passive smoking exposure.
♦♦ Serial samples of 197 children measured at all three time-points

Table S3. Association and prediction results of the top 20 CpGs in the model building dataset (N=3,764).

	Independent association with smoking habits♦♦		Association in the full model with smoking habits♦♦♦	
Marker	**Coefficient**	**p-value**	**Coefficient**	**p-value**
cg05575921 ♦	-17.443	< 2.22e-16	-18.270	< 2.22e-16
cg13039251 ♦	9.241	< 2.00e-16	5.492	2.81e-05
cg03636183 ♦	-13.909	< 2.22e-16	6.755	2.86e-05
cg12803068 ♦	6.479	< 2.00e-16	-9.766	6.07e-10
cg22132788 ♦	9.800	< 2.22e-16	13.670	5.27e-10
cg06126421 ♦	-9.415	< 2.22e-16	4.658	3.02e-05
cg21566642 ♦	-20.218	< 2.22e-16	-3.677	0.033
cg23576855 ♦	-4.390	< 2.22e-16	-1.598	0.001
cg15693572 ♦	4.189	< 2.22e-16	2.368	0.055
cg05951221 ♦	-13.816	< 2.22e-16	4.751	0.012
cg01940273 ♦	-16.834	< 2.22e-16	-5.432	0.017
cg12876356 ♦	-4.741	< 2.22e-16	5.444	0.015
cg09935388 ♦	-7.205	< 2.22e-16	-3.213	0.023
cg19572487	-10.216	< 2.00e-16	-0.666	0.518
cg19859270	-47.049	< 2.00e-16	-2.716	0.456
cg18146737	-4.602	< 2.22e-16	-0.980	0.606
cg21161138	-17.857	< 2.22e-16	-0.937	0.598
cg23480021	4.741	< 2.22e-16	-0.362	0.874
cg21188533	3.638	< 2.22e-16	0.106	0.870
cg03274391	4.715	< 2.22e-16	0.140	0.916

♦ CpGs included in our final model
♦♦ The association between the selected CpG sites and smoking habits (smokers vs non- smokers) is tested in our dataset using binominal regression adjusted for age and sex (e.g. smoking ~ CpG_1 + age + sex)
♦♦♦ The statistical summary from the full model; testing the association between smoking habits (smokers vs non- smokers) and all 20 CpGs included in our model building procedure using binominal regression (smoking ~ $CpG_{1\text{-}20}$)

Table S4. Prediction results when including age and sex in the model.

	13 CpGs	13 CpGs + Age	13 CpGs + Sex	13 CpGs + Age + Sex
Accuracy[$]	0.923	0.925	0.925	0.923
(95% CI)	(0.914, 0.931)	(0.916, 0.933)	(0.916, 0.933)	(0.915, 0.932)
Specificity	0.976	0.975	0.976	0.976
Sensitivity	0.585	0.603	0.595	0.589
AUC	0.901	0.907	0.903	0.911

[$] Proportion accurately inferred smoking habits
95% CI: confidence interval; AUC: Area under the Curve.

Table S5. Prediction results when including cell count in the model .

	13 CpGs	13 CpGs + Cell count
Accuracy[$]	0.925	0.925
(95% CI)	(0.915, 0.933)	(0.916, 0.934)
Specificity	0.975	0.975
Sensitivity	0.616	0.618
AUC	0.906	0.907

The table shows prediction results by including cell count in the model in 3,402 participants (477 current smokers and 2,925 non- smokers).
[$] Proportion accurately inferred smoking habits
95%CI: confidence interval; AUC: Area under the Curve.

Table S6. Model application to children from the Generation R study at time of birth using cord blood.

		Birth (N=1,111)	Birth (N=197)
		Whole dataset	**Serial samples**
Child non-smoking (all "0")	Accuracy[$]	0.114	0.112
Sustained prenatal smoking of mother throughout pregnancy	N	0: 950	0: 173
		1: 161	1: 24
	Specificity	0.131	0.121
	Sensitivity	0.981	0.958
	AUC	0.773	0.751
Sustained prenatal smoking of mother throughout pregnancy *or* mother stopped smoking when aware of pregnancy	N	0: 844	0: 147
		1: 267	1: 50
	Specificity	0.133	0.129
	Sensitivity	0.944	0.940
	AUC	0.664	0.571
Active *or* passive smoking of mother during pregnancy[♦]	N	0: 576	0: 108
		1: 535	1: 89
	Specificity	0.135	0.148
	Sensitivity	0.908	0.932
	AUC	0.591	0.562
Only passive smoking of mother during pregnancy[♦♦]	N	0: 843	0: 158
		1: 268	1: 39
	Specificity	0.110	0.120
	Sensitivity	0.873	0.923
	AUC	0.460	0.512

[♦] Active smoking: sustained smoking of mother throughout pregnancy or until mother became aware of pregnancy, generally in the first trimester, passive smoking: smoking of others in the pregnant mother's household or at her place of work
[♦♦] Passive smoking: mother never smoked during pregnancy but others smoked in the pregnant mother's household or at her place of work
[$] Proportion of children correctly predicted as non-smokers
AUC: Area under the Curve

Additional supplemental material for this chapter can be found in the online version of the paper via https://link.springer.com/article/10.1007%2Fs10654-019-00555-w.

REFERENCES

1. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. Front Genet. 2013;4:132.
2. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011;88(4):450-7.
3. Mortusewicz O, Schermelleh L, Walter J, Cardoso MC, Leonhardt H. Recruitment of DNA methyltransferase I to DNA repair sites. Proc Natl Acad Sci U S A. 2005;102(25):8905-9.
4. Cuozzo C, Porcellini A, Angrisano T, Morano A, Lee B, Di Pardo A, et al. DNA damage, homology-directed repair, and DNA methylation. PLoS Genet. 2007;3(7):e110.
5. Satta R, Maloku E, Zhubi A, Pibiri F, Hajos M, Costa E, et al. Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. Proc Natl Acad Sci U S A. 2008;105(42):16356-61.
6. Mercer BA, Wallace AM, Brinckerhoff CE, D'Armiento JM. Identification of a cigarette smoke-responsive region in the distal MMP-1 promoter. Am J Respir Cell Mol Biol. 2009;40(1):4-12.
7. Kadonaga JT, Carner KR, Masiarz FR, Tjian R. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell. 1987;51(6):1079-90.
8. Di YP, Zhao J, Harper R. Cigarette smoke induces MUC5AC protein expression through the activation of Sp1. J Biol Chem. 2012;287(33):27948-58.
9. Han L, Lin IG, Hsieh CL. Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. Mol Cell Biol. 2001;21(10):3416-24.
10. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics. 2015;7:113.
11. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;9(5):436-47.
12. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet. 2012;21(13):3073-82.
13. Ligthart S, Steenaard RV, Peters MJ, van Meurs JB, Sijbrands EJ, Uitterlinden AG, et al. Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. Diabetologia. 2016.
14. Steenaard RV, Ligthart S, Stolk L, Peters MJ, van Meurs JB, Uitterlinden AG, et al. Tobacco smoking is associated with methylation of genes related to coronary artery disease. Clin Epigenetics. 2015;7(1):54.
15. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res. 2009;11(1):12-24.
16. Shipton D, Tappin DM, Vadiveloo T, Crossley JA, Aitken DA, Chalmers J. Reliability of self reported smoking status by pregnant women for estimating smoking prevalence: a retrospective, cross sectional study. Bmj. 2009;339:b4347.
17. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol. 2017;18(1):238.
18. Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, Koren G. Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus

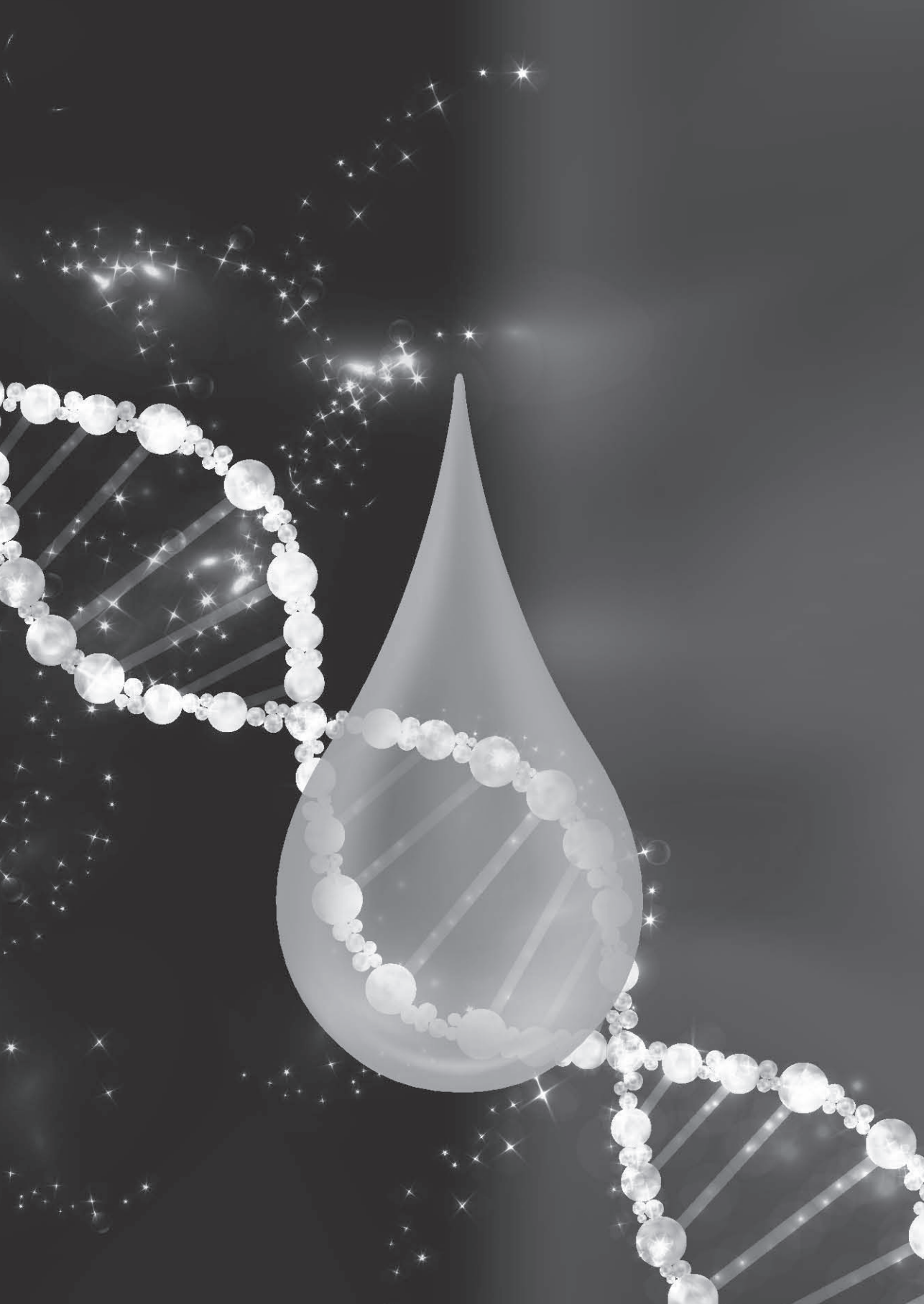
on developmental toxicology. Ther Drug Monit. 2009;31(1):14-30.
19. Benowitz NL. Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiol Rev. 1996;18(2):188-204.
20. Benowitz NL, Hukkanen J, Jacob P, 3rd. Nicotine chemistry, metabolism, kinetics and biomarkers. Handb Exp Pharmacol. 2009(192):29-60.
21. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6(1):4.
22. Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, et al. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology. 2013;24(5):712-6.
23. Philibert R, Hollenbeck N, Andersen E, Osborn T, Gerrard M, Gibbons FX, et al. A quantitative epigenetic approach for the assessment of cigarette consumption. Front Psychol. 2015;6:656.
24. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res. 2016;146:395-403.
25. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136.
26. Kondratyev N, Golov A, Alfimova M, Lezheiko T, Golimbet V. Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation. Clin Epigenetics. 2018;10(1):130.
27. Sugden K, Hannon EJ, Arseneault L, Belsky DW, Broadbent JM, Corcoran DL, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Transl Psychiatry. 2019;9(1):92.
28. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131-8.
29. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. Eur J Epidemiol. 2017;32(9):807-50.
30. van Greevenbroek MM, Jacobs M, van der Kallen CJ, Vermeulen VM, Jansen EH, Schalkwijk CG, et al. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). Eur J Clin Invest. 2011;41(4):372-9.
31. Willemsen G, Vink JM, Abdellaoui A, den Braber A, van Beek JH, Draisma HH, et al. The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. Twin Res Hum Genet. 2013;16(1):271-81.
32. Schoenmaker M, de Craen AJ, de Meijer PH, Beekman M, Blauw GJ, Slagboom PE, et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. Eur J Hum Genet. 2006;14(1):79-84.
33. Huisman MH, de Jong SW, van Doormaal PT, Weinreich SS, Schelhaas HJ, van der Kooi AJ, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. J Neurol Neurosurg Psychiatry. 2011;82(10):1165-70.
34. Tigchelaar EF, Zhernakova A, Dekens JA, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015;5(8):e006772.
35. Holle R, Happich M, Lowel H, Wichmann HE, Group MKS. KORA--a research platform for population based health research. Gesundheitswesen. 2005;67 Suppl 1:S19-25.

36. Jurgens C, Volzke H, Tost F. [Study of health in Pomerania (SHIP-Trend): : Important aspects for healthcare research in ophthalmology] Study of Health in Pomerania (SHIP-Trend) : Wichtige Aspekte fur die ophthalmologische Versorgungsforschung. Ophthalmologe. 2014;111(5):443-7.
37. Kooijman MN, Kruithof CJ, van Duijn CM, Duijts L, Franco OH, van IMH, et al. The Generation R Study: design and cohort update 2017. Eur J Epidemiol. 2016;31(12):1243-64.
38. Tobi EW, Slieker RC, Stein AD, Suchiman HE, Slagboom PE, van Zwet EW, et al. Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome. Int J Epidemiol. 2015;44(4):1211-23.
39. van Iterson M, Tobi EW, Slieker RC, den Hollander W, Luijk R, Slagboom PE, et al. MethylAid: visual and interactive quality control of large Illumina 450k datasets. Bioinformatics. 2014;30(23):3435-7.
40. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22(5):843-51.
41. Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One. 2013;8(5):e63812.
42. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environ Health Perspect. 2014;122(7):673-8.
43. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382-96.
44. Allione A, Marcon F, Fiorito G, Guarrera S, Siniscalchi E, Zijno A, et al. Novel epigenetic changes unveiled by monozygotic twins discordant for smoking habits. PLoS One. 2015;10(6):e0128265.
45. Besingi W, Johansson A. Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet. 2014;23(9):2290-7.
46. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.
47. Sayols-Baixeras S, Lluis-Ganella C, Subirana I, Salas LA, Vilahur N, Corella D, et al. Corrigendum. Identification of a new locus and validation of previously reported loci showing differential methylation associated with smoking. The REGICOR study. Epigenetics. 2016;11(2):174.
48. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8(5):599-618.
49. Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2012;120(10):1425-31.
50. Zhu X, Li J, Deng S, Yu K, Liu X, Deng Q, et al. Genome-Wide Analysis of DNA Methylation and Cigarette Smoking in a Chinese Population. Environ Health Perspect. 2016;124(7):966-73.
51. Hastie T, Tibshirani, Robert, Friedman, Jerome. The Elements of Statistical Learning; Data Mining, Inference, and Prediction. 2 ed: Springer; 2009.

52. Bradley Efron RJT. An Introduction to the Bootstrap: Chapman and Hall/CRC; 1994.
53. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-81.
54. Ware JJ, Chen X, Vink J, Loukola A, Minica C, Pool R, et al. Genome-Wide Meta-Analysis of Cotinine Levels in Cigarette Smokers Identifies Locus at 4q13.2. Sci Rep. 2016;6:20092-.
55. Gupta R, van Dongen J, Fu Y, Abdellaoui A, Tyndale RF, Velagapudi V, et al. Epigenome-wide association study of serum cotinine in current smokers reveals novel genetically driven loci. Clinical Epigenetics. 2019;11(1):1.
56. Bot M, Vink JM, Willemsen G, Smit JH, Neuteboom J, Kluft C, et al. Exposure to secondhand smoke and depression and anxiety: a report from two studies in the Netherlands. J Psychosom Res. 2013;75(5):431-6.
57. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. Am J Hum Genet. 2016;98(4):680-96.
58. Richmond RC, Suderman M, Langdon R, Relton CL, Davey Smith G. DNA methylation as a marker for prenatal smoke exposure in adults. Int J Epidemiol. 2018;47(4):1120-30.
59. Philibert RA, Beach SR, Lei MK, Brody GH. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. Clin Epigenetics. 2013;5(1):19.
60. Esser C. Biology and function of the aryl hydrocarbon receptor: report of an international and interdisciplinary conference. Arch Toxicol. 2012;86(8):1323-9.
61. Nguyen TA, Hoivik D, Lee JE, Safe S. Interactions of nuclear receptor coactivator/corepressor proteins with the aryl hydrocarbon receptor complex. Arch Biochem Biophys. 1999;367(2):250-7.
62. Bojesen SE, Timpson N, Relton C, Davey Smith G, Nordestgaard BG. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. Thorax. 2017;72(7):646-53.
63. Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K. Association of DNA methylation with age, gender, and smoking in an Arab population. Clin Epigenetics. 2015;7:6.
64. Lim U, Song MA. Dietary and lifestyle factors of DNA methylation. Methods Mol Biol. 2012;863:359-76.
65. Su D, Wang X, Campbell MR, Porter DK, Pittman GS, Bennett BD, et al. Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes. PLoS One. 2016;11(12):e0166486.
66. Bauer M, Fink B, Thurmann L, Eszlinger M, Herberth G, Lehmann I. Tobacco smoking differently influences cell types of the innate and adaptive immune system-indications from CpG site methylation. Clin Epigenetics. 2015;7:83.
67. Bauer M, Linsel G, Fink B, Offenberg K, Hahn AM, Sack U, et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin Epigenetics. 2015;7:81.
68. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9(5):757-68.
69. Zhang Y, Schottker B, Florath I, Stock C, Butterbach K, Holleczek B, et al. Smoking-Associated DNA Methylation Biomarkers and Their Predictive Value for All-Cause and Cardiovascular Mortality. Environ Health Perspect. 2016;124(1):67-74.
70. Stead LF, Perera R, Bullen C, Mant D, Hartmann-Boyce J, Cahill K, et al. Nicotine

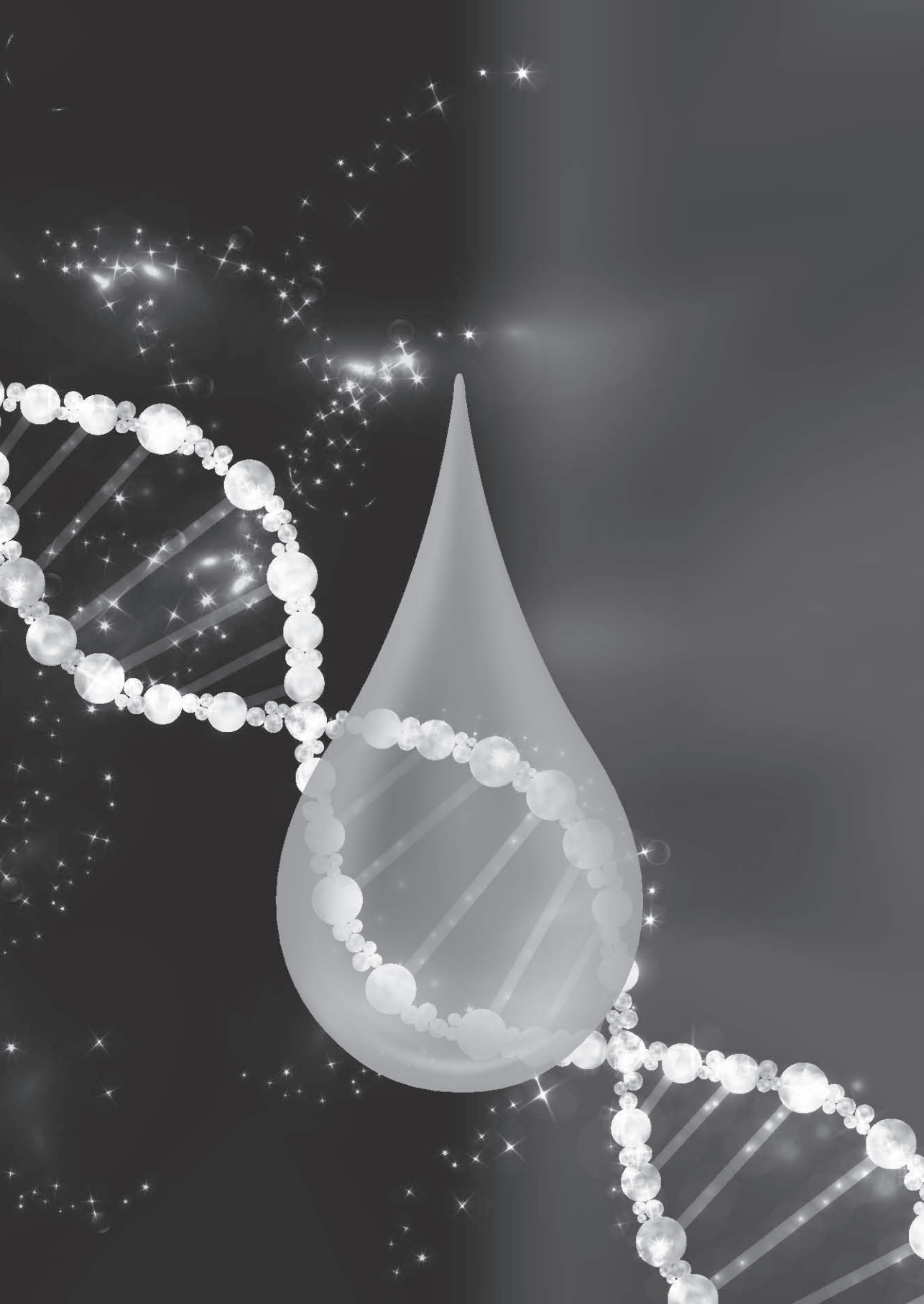
replacement therapy for smoking cessation. Cochrane Database Syst Rev. 2012;11:CD000146.

71. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet. 2015;24(8):2201-17.
72. Ladd-Acosta C, Shu C, Lee BK, Gidaya N, Singer A, Schieve LA, et al. Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. Environ Res. 2016;144(Pt A):139-48.
73. Reese SE, Zhao S, Wu MC, Joubert BR, Parr CL, Haberg SE, et al. DNA Methylation Score as a Biomarker in Newborns for Sustained Maternal Smoking during Pregnancy. Environ Health Perspect. 2017;125(4):760-6.
74. Bergens MA, Pittman GS, Thompson IJB, Campbell MR, Wang X, Hoyo C, et al. Smoking-associated AHRR demethylation in cord blood DNA: impact of CD235a+ nucleated red blood cells. Clin Epigenetics. 2019;11(1):87.
75. Xu CJ, Bonder MJ, Soderhall C, Bustamante M, Baiz N, Gehring U, et al. The emerging landscape of dynamic DNA methylation in early childhood. BMC Genomics. 2017;18(1):25.
76. Dawes K, Andersen A, Vercande K, Papworth E, Philibert W, Beach SRH, et al. Saliva DNA Methylation Detects Nascent Smoking in Adolescents. J Child Adolesc Psychopharmacol. 2019.
77. Vidaki A, Johansson C, Giangasparo F, Denise Syndercombe C. Differentially methylated embryonal Fyn-associated substrate (EFS) gene as a blood-specific epigenetic marker and its potential application in forensic casework. Forensic Sci Int Genet. 2017;29:165-73.
78. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Sci Int Genet. 2017;28:225-36.



Chapter 3

Lifestyle factor induced changes in DNA methylation and cardiovascular outcomes



Chapter 3.1

Smoking induced epigenetic changes and its association with cardiometabolic traits

Silvana C.E. Maas, Michelle M.J. Mens, Brigitte Kühnel, Joyce B.J. van Meurs, André G. Uitterlinden, Annette Peters, Holger Prokisch, Christian Herder, Harald Grallert, Sonja Kunze, Melanie Waldenberger, Maryam Kavousi, Manfred Kayser, Mohsen Ghanbari

Clinical Epigenetics 2020;12(1):157

ABSTRACT

Background: Tobacco smoking is a well-known modifiable risk factor for many chronic diseases, including cardiovascular disease (CVD). One of the proposed underlying mechanism linking smoking to disease is via epigenetic modifications, which could affect the expression of disease-associated genes. Here, we conducted a three-way association study to identify the relationship between smoking-related changes in DNA methylation and gene expression and their associations with cardio-metabolic traits.

Results: We selected 2,549 CpG sites and 443 gene expression probes associated with current *versus* never smokers, from the largest epigenome-wide association study and transcriptome-wide association study to date. We examined three-way associations, including CpG *versus* gene expression, cardio-metabolic trait *versus* CpG, and cardio-metabolic trait *versus* gene expression, in the Rotterdam study. Subsequently, we replicated our findings in The Cooperative Health Research in the Region of Augsburg (KORA) study. After correction for multiple testing, we identified both *cis*- and *trans*-expression quantitative trait methylation (eQTM) associations in blood. Specifically, we found 1,224 smoking-related CpGs associated with at least one of the 443 gene expression probes, and 200 smoking-related gene expression probes to be associated with at least one of the 2,549 CpGs. Out of these, 109 CpGs and 27 genes were associated with at least one cardio-metabolic trait in the Rotterdam Study. We were able to replicate the associations with cardio-metabolic traits of 26 CpGs and 19 genes in the KORA study. Furthermore, we identified a three-way association of triglycerides with two CpGs and two genes (*GZMA*; *CLDND1*), and BMI with six CpGs and two genes (*PID1*; *LRRN3*). Finally, our results revealed the mediation effect of cg03636183 (*F2RL3*), cg06096336 (*PSMD1*), cg13708645 (*KDM2B*), and cg17287155 (*AHRR*) within the association between smoking and *LRRN3* expression.

Conclusions: Our study indicates that smoking-related changes in DNA methylation and gene expression are associated with cardio-metabolic risk factors. These findings may provide additional insights into the molecular mechanisms linking smoking to the development of CVD.

INTRODUCTION

Tobacco smoking is a major modifiable risk factor for premature death and non-communicable diseases worldwide (1). With almost 18 million deaths in 2017, cardiovascular diseases (CVD) account for the largest number of deaths of non-communicable diseases (2). Smoking is also associated with cardio-metabolic traits, such as dyslipidemia, hypertension, insulin resistance, and obesity, which are major risk factors leading to CVD (3, 4). Furthermore, persistent smoking has an excessive impact on DNA methylation (5-7) and gene expression (8-10), which their alterations are also linked to cardio-metabolic traits and risk of CVD (11-16).

Extensive studies have shown the independent association of smoking with DNA methylation, gene expression levels, and disease risk. In this context, smoking is associated with alteration in DNA methylation levels of several genes related to type 2 diabetes (17) and coronary artery disease (18). Additionally, smoking-related CpGs have a strong association with all-cause and cardiovascular mortality (19). Nevertheless, much less research has investigated smoking-related changes in DNA methylation and gene expression concurrently and in relation to health outcomes. A recent study identified a link between smoking-related DNA methylation and gene expression changes with metabolic health (20). Their results indicate possible molecular pathways in which smoking affects disease development.

In this study, we hypothesized that smoking-related modifications in DNA methylation and gene expression are associated with each other and, additionally, with cardio-metabolic traits. Hence, we first determined three-way associations, including CpGs *vs.* gene expression, cardio-metabolic traits *versus* CpGs, cardio-metabolic traits *versus* gene expression. To this end, we selected CpGs and gene expression probes associated with current *versus* never smokers using the largest published epigenome-wide association study (EWAS) (6) and transcriptome-wide association study (TWAS) (8) to date. Next, we used data from the Rotterdam Study to test the expression quantitative trait methylation (eQTM) association between the selected CpGs and gene expression probes. Subsequently, we tested the association for these CpGs and genes with different cardio-metabolic traits, including lipids, glycemic indices, blood pressure, and obesity-related traits. Moreover, we performed mediation analysis to test the mediating effect of (1) DNA methylation in the association between smoking and cardio-metabolic traits, (2) gene expression in the association between smoking and cardio-metabolic traits, and (3) DNA methylation in the association between smoking and expression levels of smoking-related genes. To test the validity of our findings, we further replicated our results in an independent cohort, The Cooperative Health Research in the Region of Augsburg (KORA) study.

RESULTS

An overview of our study design is illustrated in **Figure 1**. The discovery dataset consisted of 1,412 participants with DNA methylation data from the two sub-cohorts of the Rotterdam study; RS-II and RS-III. Of these, 716 participants from RS-III had also gene expression data (21). The replication dataset comprises 1,727 participants with DNA methylation data, of whom 687 also had gene expression data, from the KORA study (F4) (22). Both the discovery and replication cohorts consisted of both males and females (53.3%) and current, former and never smokers. In the current study, the former and never smokers are combined in the non-smoker category (83.6%). General characteristics of the study population are listed in **Table 1**.

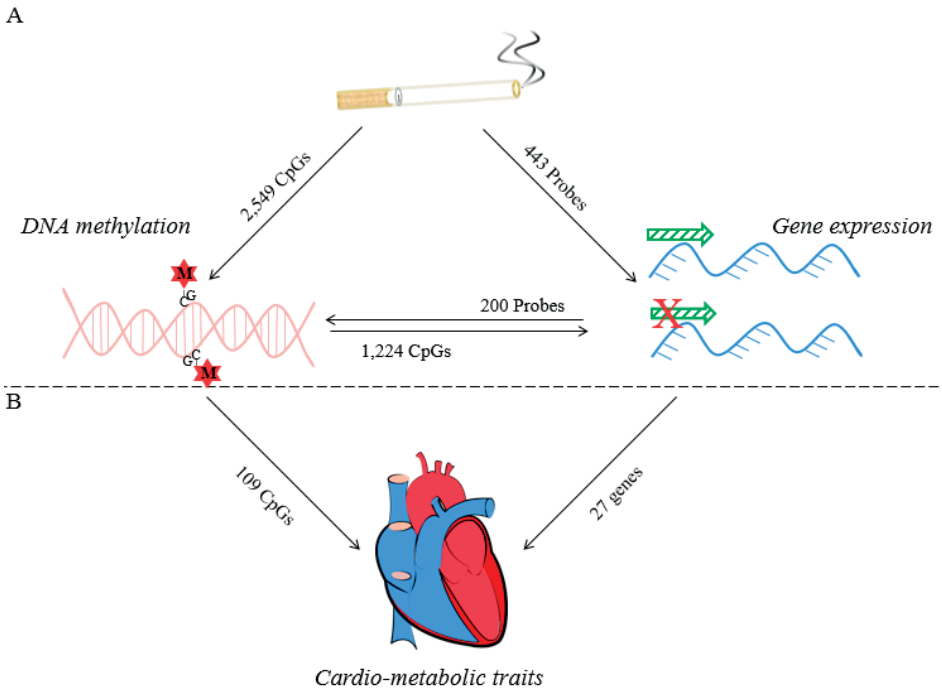


Figure 1. Schematic overview of the study design. In the current paper, previously identified CpGs by the largest available EWAS (6) and genes by the largest available TWAS (8) associated with current *versus* never smokers were used to test the link between smoking and cardio-metabolic traits. To this end, we first examined the association between smoking and alterations in gene expression (A). Second, we checked the association between the smoking-related CpGs and the smoking-related gene expression probes (A). Third, the smoking-related CpGs and gene expression probes that were in eQTM with each other were tested for their association with cardio-metabolic traits (B).

Table 1. Population Characteristics

	Discovery dataset		Replication dataset	
	Gene expression data set	DNA methylation data set	Gene expression data set	DNA methylation data set
N	716	1,412	687	1,727
Female	389 (54.3%)	791 (56.0%)	339 (49.3%)	882 (51.0%)
Age (years)	59.8 (±8.1)	63.6 (±8.1)	69.1 (±4.4)	61.0 (±8.8)
BMI (kg/m^2)	27.6 (±4.6)	27.7 (±4.4)	28.9 (±4.5)	28.1 (±4.8)
WHR	0.9 (±0.1)	0.9 (±0.1)	0.9 (±0.1)	0.9 (±0.1)
Current smokers	193 (27.0%)	266 (18.0%)	53 (7.7%)	250 (14.5%)
Triglycerides (mmol/L)	1.5 (±0.9)	1.5 (±0.8)	1.5 (±0.9)	1.5 (±1.1)
HDL-cholesterol (mmol/L)	1.4 (±0.4)	1.5 (±0.4)	1.4 (±0.4)	1.5 (±0.4)
LDL-cholesterol (mmol/L)	3.9 (±1.0)	3.8 (±1.0)	3.7 (±0.9)	3.6 (±0.9)
Total cholesterol (mmol/L)	5.6 (±1.1)	5.5 (±1.0)	5.8 (±1.0)	5.7 (±1.0)
Lipid lowering medication (yes)	190 (26.5%)	404 (28.6%)	172 (25.0%)	283 (16.4%)
Systolic blood pressure (mm Hg)	134.2 (±19.8)	139.5 (±21.5)	128.7 (±19.4)	124.8 (±18.7)
Diastolic blood pressure (mm Hg)	82.8 (±11.4)	83.6 (±11.5)	74.7 (±10.0)	76.1 (±10.0)
Anti-hypertensive medication (yes)	215 (30.0%)	517 (36.6%)	383 (55.7%)	650 (37.6%)
Glucose (mmol/L)	5.6 (±1.0)	5.6 (±1.1)	NA	NA
Insulin (pmol/L)	96.0 (±63.0)	89.3 (±56.6)	88.2 (±122.0)	81.3 (±91.0)
Anti-diabetic medication (yes)	39 (5.4%)	95 (6.7%)	76 (11.1%)	134 (7.8%)

Values are presented as mean ±(SD) or N (%)

BMI, body mass index; *WHR*, waist to hip ratio; *HDL,* High-density lipoproteins;

LDL, Low-density lipoprotein

The participants included in the gene expression data are a subset of the total DNA methylation dataset

NA, not applicable; the associations with glucose levels in model 2 from the discovery did not pass the significance threshold

Correlation between smoking-related changes in DNA methylation and gene expression

We selected 2,623 CpGs previously reported as being significantly (P<1×10^{-7}) differentially methylated between smokers and never smokers (6). Of these, 2,549 CpGs passed the quality control in the Rotterdam Study. Furthermore, we selected 502 gene expression probes that were differently expressed between smokers and never smokers (*FDR<0.05*), and replicated in an independent dataset as part of the same study (8). Of these, 443 gene expression probes passed quality control in the Rotterdam Study. Then, we investigated the eQTM associations to test the possible impact of smoking-related DNA methylation changes on the smoking-related genes, or vice versa. To this end we computed the residuals for both the CpGs and gene expression probes. Then, we tested the association between all the smoking-related CpGs with all the smoking-related gene expression probes. Here, we investigated *cis*-eQTMs in which the CpG regulates

transcription of a neighboring gene (≤ 250Kb from each side of the transcription start site). Also, we studied the *trans*-eQTM association in which a CpG regulates distant genes located >250Kb of the transcription start site (23). Notably, out of the 2,549 smoking-related CpGs, 1224 were associated with at least one of the gene expression probes at the significance threshold of $P< 4.4\times10^{-8}$ (0.05/443×2,549). Of the 443 tested gene expression probes, 200 probes were significantly associated with at least one of the 2,549 CpGs, after correcting for multiple testing (**Additional file 1: Table S1**). The R code to generate the residuals for the CpGs and gene expression probes, and for the eQTM analysis are included in **Additional File 2**.

To examine the possible enrichment due to the smoking effect, we further tested if the number of significant eQTM associations is higher while using smoking-related CpGs and genes, compared to the number of eQTM associations while using randomly selected CpGs and genes. When testing the association between the 2,549 smoking-related CpGs with 443 randomly selected gene expression probes, we found that only 325 CpGs are associated with at least one of these gene expression probes and 186 gene expression probes with at least one smoking-related CpG. Using the chi-square test of independence to compare the use of smoking-related gene expression probes *versus* randomly selected gene expression probes, we obtained for the CpGs (1,224 *vs.* 325, respectively) $P< 1.0\times10^{-5}$ and for the genes expression probes (200 *vs.* 186, respectively) a *P*-value of 0.38. Similarly, when testing the association between 2,549 randomly selected CpGs with the 443 smoking-related gene expression probes, we found only 465 CpGs associated with at least one smoking-related gene expression probe, and 19 gene expression probes with at least one smoking-related CpG. Using the chi-square test of independence, comparing the use of smoking-related CpGs *versus* randomly selected CpGs, we found a significant difference ($P<1.0\times10^{-5}$) for both the CpGs (1,224 *vs.* 465, respectively) and the gene expression probes (200 *vs.*19, respectively). These results indicate enrichment of smoking-related genes in smoking-related DNA methylation sites and vice versa.

The replication in the KORA study confirmed the association of 134 smoking-related CpGs with at least one gene expression probe and 50 smoking-related gene expression probes with at least one smoking-related CpG, after correcting for multiple testing, at the significance threshold of $P< 2.04\times10^{-7}$ (0.05/200×1,224).

Association of smoking-related changes in DNA methylation and gene expression with cardio-metabolic traits

We tested the association of the 1,224 CpGs and the 200 gene expression probes with cardio-metabolic traits, including high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides (TG) and serum cholesterol, fasting glucose and insulin levels, systolic blood pressure (SBP) and diastolic blood pressure (DBP), waist to hip ratio

(WHR) and body mass index (BMI) in the Rotterdam Study. After adjusting for age, sex, blood cell count, and technical covariates (model 1), we found significant associations between 202 out of the 1,224 smoking-related CpGs and any cardio-metabolic trait at $P<4.08\times10^{-5}$ (0.05/1224) (n= 1,412 participants) (**Additional file 3: Table S2**). Among these, we observed associations with HDL (126 CpGs), TG (84 CpGs), glucose (2 CpGs), insulin (10 CpGs), DBP (1 CpG), WHR (21 CpGs), and BMI (16 CpGs). After further adjustment for BMI and relevant medication in the model 2, associations with 109 CpGs remained significant, including HDL (58 CpGs), TG (35 CpGs), DBP (1 CpG), WHR (6 CpG), and BMI (16 CpG same as model 1) (**Additional file 3: Table S3**). The R code to test the association between cardio-metabolic traits and the CpGs are included in **Additional File 4**. We pursued replication in the KORA study for the CpGs reaching significance in the model 2 and found that 26 CpGs surpassed the nominal significance ($P<0.05$, n= 1,727 participants), including 8 CpGs for HDL, 8 CpGs for TG, 4 CpGs for WHR, and 7 CpGs for BMI (**Table 2**, **Additional file 3: Table S3**). The direction of associations with cardio-metabolic traits was consistent in all 26 replicated CpGs. Based on the stringent Bonferroni-adjusted P-value threshold, the replication signals were significant at 2 CpGs for TG ($P< 0.05/35=1.43\times10^{-3}$), 3 CpGs for WHR ($P< 0.05/6=8.33\times10^{-3}$), and 4 CpGs with BMI ($P< 0.05/16=3.13\times10^{-3}$) (**Table 2** and **Figure 2**).

Furthermore, out of the 200 smoking-related gene expression probes 39 (35 genes) were significantly associated with at least one cardio-metabolic trait at $P< 2.5\times10^{-4}$ (0.05/200) in the Rotterdam Study (n=716 participants) (**Additional file 2: Table S4**). In the Illumina HumanHT-12 Expression BeadChip array, some of the annotated genes have more than one probe. Therefore, we adjusted the analysis for the number of probes we tested and provided both the probe ID and annotated gene in the tables. Of the 39 probes, we found associations with HDL (15 genes), LDL (1 gene), TG (18 genes), cholesterol (1 gene), glucose (3 genes), insulin (13 genes), WHR (6 genes), and BMI (14 genes). After further adjustments in model 2, the associations of 29 probes (27 genes) remained significant, including HDL (5 genes), LDL (1 gene), TG (14 genes), cholesterol (1 gene), and insulin (2 genes), for the association with BMI nothing changed (14 genes) (**Additional file 3: Table S5**). The R code to test the association between cardio-metabolic traits and the gene expression probes is included in **Additional File 5**. Replication in the KORA study for the gene expression probes that reached significance in model 2 showed 21 probes (19 genes) that passed the nominal significance ($P<0.05$, n= 687 participants). These include 2 genes for HDL, 13 genes for TG, 1 gene for insulin, and 10 genes for BMI (**Table 3**, **Additional file 3: Table S5**). The direction of the association between gene expression and cardio-metabolic traits was consistent for all these genes. Based on the stringent Bonferroni-adjusted P-value in which we adjusted for the number of probes, the replication signal was significant at 2 genes for HDL ($P< 0.05/5=0.01$), 11 genes for TG ($P< 0.05/15=3.33\times10^{-3}$), 1 gene for insulin ($P< 0.05/2=0.03$) and 4 genes

Table 2. CpG sites associated with cardio-metabolic traits in DNA methylation analysis.

CpG	Chr:position*	Gene ID**	Trait	Model 1		Model 2		Replication	
				Effect	P-value	Effect	P-value	Effect	P-value
cg04716530	16:30485684	*ITGAL*	HDL	0.01700	2.02E-07	0.01550	6.34E-06	0.00014	9.82E-04
cg07826859	7:45020086	*MYO1G*	HDL	0.01440	6.79E-06	0.01390	3.61E-05	0.00018	1.61E-03
cg26724967	16:3115223	*IL32*	HDL	0.01290	2.89E-06	0.01220	2.34E-05	0.00015	1.63E-03
cg16391678	16:30485597	*ITGAL*	HDL	0.01520	1.03E-06	0.01400	1.60E-05	0.00013	5.14E-03
cg16519923	16:30485810	*ITGAL*	HDL	0.01980	8.23E-08	0.01790	3.44E-06	0.00012	9.75E-03
cg10310310	7:157367150	*PTPRN2;MIR153-2*	HDL	0.01130	4.49E-06	0.01060	4.05E-05	0.00011	1.96E-02
cg24323726	3:111314186	*ZBED2;CD96*	HDL	0.01300	4.84E-07	0.01230	5.32E-06	0.00009	3.41E-02
cg07929642	16:89390685	*ANKRD11*	HDL	0.01650	1.87E-07	0.01550	3.13E-06	0.00009	4.25E-02
cg21566642	**2:233284661**	**-**	**TG**	**-0.01990**	**3.89E-05**	**-0.02150**	**1.91E-05**	**-0.01967**	**1.14E-05**
cg04716530	**16:30485684**	***ITGAL***	**TG**	**-0.01370**	**1.23E-06**	**-0.01220**	**3.58E-05**	**-0.00380**	**5.34E-04**
cg27409015	2:158114424	*GALNT5*	TG	0.01660	8.39E-07	0.01490	2.16E-05	0.00683	1.50E-03
cg06635952	2:70025869	*ANXA4*	TG	0.01300	9.06E-07	0.01280	3.36E-06	0.00502	5.95E-03
cg11095027	11:1297066	*TOLLIP*	TG	0.00996	2.72E-05	0.01040	2.64E-05	0.00375	7.82E-03
cg26219092	8:134388022	-	TG	0.01050	2.37E-06	0.00991	2.05E-05	0.00285	1.73E-02
cg10919522	14:74227441	*C14orf43*	TG	-0.01410	7.90E-09	-0.01260	7.97E-07	-0.00491	1.91E-02
cg22635096	21:46550644	*ADARB1*	TG	0.01370	6.61E-08	0.01300	8.65E-07	0.00392	3.55E-02
cg00310412	**15:74724918**	***SEMA7A***	**WHR**	**-0.05000**	**3.96E-05**	**-0.07150**	**1.95E-07**	**-0.06952**	**4.54E-05**
cg04424621	**6:27101941**	***HIST1H2BJ***	**WHR**	**-0.06490**	**1.06E-05**	**-0.07360**	**1.03E-05**	**-0.07191**	**5.73E-05**
cg04583842	**16:88103117**	***BANP***	**WHR**	**0.12600**	**5.29E-08**	**0.12500**	**1.56E-06**	**0.06802**	**3.71E-03**
cg13755776	11:3602845	-	WHR	-0.08530	1.03E-06	-0.08200	3.33E-05	-0.04521	3.71E-02
cg17287155	**5:393347**	***AHRR***	**BMI**	**0.00117**	**8.69E-06**	NA	NA	**0.00053**	**3.11E-06**
cg26361535	**8:144576604**	***ZC3H3***	**BMI**	**0.00155**	**1.85E-07**	NA	NA	**0.00102**	**2.38E-05**
cg06096336	**2:231989800**	***PSMD1;HTR2B***	**BMI**	**0.00168**	**9.51E-07**	NA	NA	**0.00111**	**1.66E-04**
cg13708645	**12:121974305**	***KDM2B***	**BMI**	**0.00152**	**6.72E-07**	NA	NA	**0.00089**	**1.01E-03**
cg25649826	17:20938740	*USP22*	BMI	0.00086	1.63E-05	NA	NA	0.00041	3.15E-03
cg24539517	10:121161258	*GRK5*	BMI	0.00149	2.95E-05	NA	NA	0.00078	4.33E-03
cg03636183	19:17000585	*F2RL3*	BMI	0.00160	3.04E-05	NA	NA	0.00063	3.40E-02

The table shows 26 CpGs that are associated to at least one cardio-metabolic trait and in eQTM with at least one smoking related gene-expression probe.

Only CpGs significantly associated in both models and nominally significant (*P< 0.05*) in the replication are presented in this table.

HDL, high-density-lipoprotein; *TG*, triglycerides; *WHR*, waist to hip ratio; *BMI*, body mass index; NA, Not applicable (because of adjusting for BMI).

Model 1: Adjusted for age, sex, cell count and technical covariates. Model 2: Model 1 + BMI and relevant medication.

We did not correct for additional covariates when testing the association for BMI

P-value threshold for discovery *P*< (0.05/1,224=) 4.08×10^{-5}

P-value threshold for replication: HDL; *P*< (0.05/58=)8.62×10^{-4}, TG; *P*< (0.05/35=)1.43×10^{-3}, WHR; *P*< (0.05/6=)8.33×10^{-3}, BMI; *P*< (0.05/16=)3.13×10^{-3}

CpGs that are presented bold passed the replication P-value threshold in 1,727 participants of the KORA study

* Genome coordinates provided by Illumina (GRCh37/hg19), ** According to the Illumina Infinium HumanMethylation450K annotation file

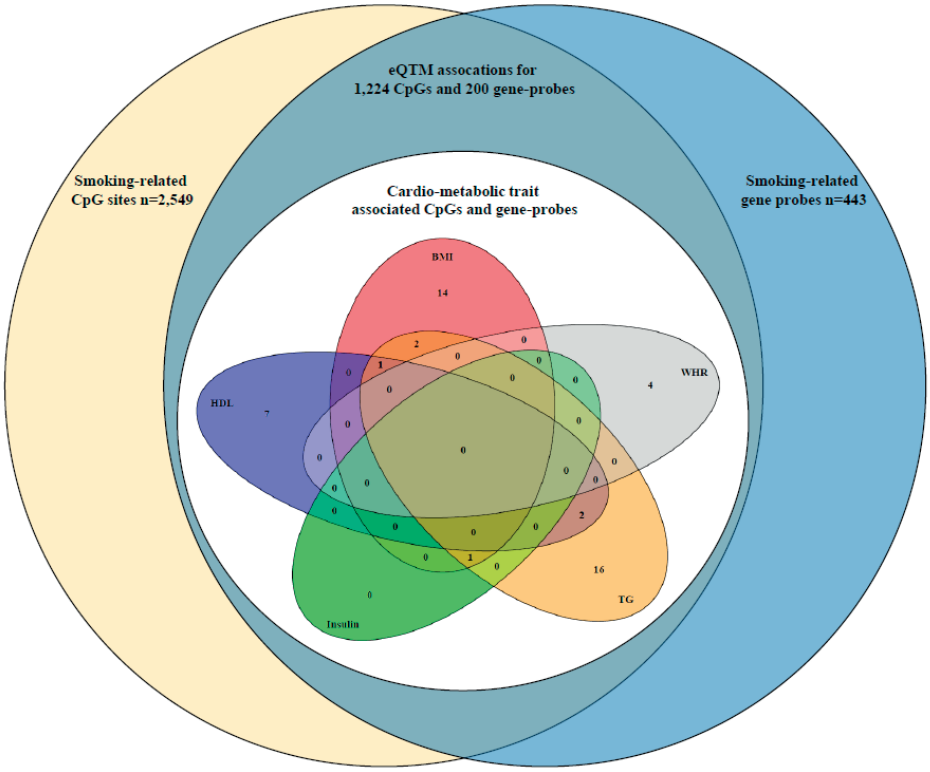


Figure 2. The overlap of smoking-related CpGs and genes in association with cardio-metabolic traits.
In the current study, 2,549 smoking-related CpGs and 443 smoking-related gene expression probes were included. Of these, 1,224 CpGs and 200 gene expression probes showed eQTM association. We found the association for 26 CpGs and 19 genes (21 expression probes) with at least one cardio-metabolic traits, which surpassing the nominal significance ($P < 0.05$) in the KORA replication study.

for BMI ($P< 0.05/16=3.13\times10^{-3}$). Several of these genes were associated in model 2 with more than one cardio-metabolic trait and were replicated at least at the nominal significance ($P< 0.05$). For example, *KLRB1* (ILMN_2079655), *ITM2C* (ILMN_2366041), and *CD3D* (ILMN_2261416) were associated with both TG and BMI, and *OCIAD2* (ILMN_1700306) was associated with both HDL and TG, and *EFHD2* (ILMN_1761463) was associated with HDL, TG, and BMI (**Table 3** and **Figure 2**).

Next, we explored whether there is an overlap in the results obtained with DNA methylation and gene expression data, which possibly explain the link between smoking and cardio-metabolic traits. **Table 4** shows the overlap of the replicated association of cardio-metabolic traits with gene expression, which were both also associated with the smoking-related CpGs, indicating a three-way association (**Figure 1**). **Additional file 3: Table S6** displays the three-way association as obtained in our discovery dataset. For example, we found in the Rotterdam Study overlapping association of serum HDL levels with four CpGs (cg01305745, cg06177555, cg07990556, and cg16448702) and expression

Table 3. Gene expression probes associated with cardio-metabolic traits.

				Model 1		Model 2		Replication	
Probe ID	**Gene ID***	**Chr.***	**Trait**	**Effect**	**P-value**	**Effect**	**P-value**	**Effect**	**P-value**
ILMN_1700306	***OCIAD2***	**4**	**HDL**	**-0.4114**	**8.39E-07**	**-0.3573**	**5.15E-05**	**-0.0040**	**9.20E-05**
ILMN_1761463	***EFHD2***	**1**	**HDL**	**0.4668**	**1.60E-07**	**0.3576**	**0.00013**	**0.0022**	**1.38E-03**
ILMN_2261416	***CD3D***	**11**	**TG**	**0.9089**	**2.22E-15**	**0.8328**	**3.84E-12**	**0.3233**	**2.58E-15**
ILMN_2079655	***KLRB1***	**12**	**TG**	**1.1666**	**6.48E-14**	**1.0518**	**1.01E-10**	**0.3946**	**4.09E-15**
ILMN_1779324	***GZMA***	**5**	**TG**	**1.0562**	**1.99E-09**	**1.0033**	**6.01E-08**	**0.2734**	**2.06E-13**
ILMN_1761463	***EFHD2***	**1**	**TG**	**-0.4238**	**2.59E-09**	**-0.3434**	**3.54E-06**	**-0.1298**	**9.99E-13**
ILMN_1700306	***OCIAD2***	**4**	**TG**	**0.3413**	**3.32E-07**	**0.2998**	**1.93E-05**	**0.1691**	**7.64E-10**
ILMN_1808939	***RPS6***	**9**	**TG**	**0.6145**	**5.58E-11**	**0.5687**	**7.54E-09**	**0.3001**	**1.17E-09**
ILMN_1812191	***C12orf57***	**12**	**TG**	**0.4392**	**1.37E-05**	**0.3897**	**0.00023**	**0.1772**	**1.07E-07**
ILMN_1776181	***BIRC3***	**11**	**TG**	**0.6940**	**2.86E-10**	**0.6176**	**8.34E-08**	**0.1473**	**4.89E-07**
ILMN_1813836	***DARS***	**2**	**TG**	**0.2657**	**9.37E-07**	**0.2878**	**4.67E-07**	**0.0894**	**3.18E-06**
ILMN_1669927	***ICOS***	**2**	**TG**	**0.2859**	**6.88E-06**	**0.2692**	**5.75E-05**	**0.0925**	**2.95E-04**
ILMN_2198878	***INPP4B***	**4**	**TG**	**0.2950**	**2.99E-07**	**0.2853**	**2.53E-06**	**0.0668**	**5.93E-04**
ILMN_2366041	*ITM2C*	2	TG	-0.5648	6.81E-09	-0.4159	3.35E-05	-0.0818	6.33E-03
ILMN_1680453	*ITM2C*	2	TG	-0.5880	5.50E-08	-0.4295	0.000124	-0.0818	7.05E-03
ILMN_2352563	*CLDND1*	3	TG	0.3877	4.61E-05	0.4003	6.52E-05	0.0656	3.53E-02
ILMN_2079655	***KLRB1***	**12**	**Insulin**	**0.7694**	**7.01E-09**	**0.6393**	5.21E-05	**0.2120**	**5.59E-05**
ILMN_1766657	***STOM***	**9**	**BMI**	**0.0549**	**3.85E-07**	NA	NA	**0.0196**	**2.56E-08**
ILMN_1671891	***PID1***	**2**	**BMI**	**-0.0425**	**7.23E-09**	NA	NA	**-0.0137**	**3.27E-06**
ILMN_2366041	***ITM2C***	**2**	**BMI**	**-0.0577**	**1.86E-09**	NA	NA	**-0.0123**	**9.53E-05**
ILMN_1773650	***LRRN3***	**7**	**BMI**	**-0.0669**	**3.77E-05**	NA	NA	**-0.0162**	**1.70E-03**
ILMN_1661599	*DDIT4*	10	BMI	-0.0658	2.75E-07	NA	NA	-0.0096	4.06E-03
ILMN_2048591	*LRRN3*	7	BMI	-0.0604	1.46E-05	NA	NA	-0.0086	6.95E-03
ILMN_2377669	*CD247*	1	BMI	-0.0370	5.95E-05	NA	NA	-0.0058	8.01E-03
ILMN_2109197	*EPB41L3*	18	BMI	-0.0322	0.000112	NA	NA	-0.0072	1.12E-02
ILMN_2261416	*CD3D*	11	BMI	0.0458	6.47E-05	NA	NA	0.0107	1.37E-02
ILMN_2079655	*KLRB1*	12	BMI	0.0669	1.63E-05	NA	NA	0.0122	2.29E-02
ILMN_1761463	*EFHD2*	1	BMI	-0.0339	1.46E-06	NA	NA	-0.0038	4.68E-02

The table shows 21 probes annotated to 19 genes that are significantly associated with cardio-metabolic traits and in eQTM with at least one smoking related CpG.

Only probes significantly associated in both models and nominally significant ($P< 0.05$) in the replication are presented in this table. *HDL*, high-density-lipoprotein; *TG*, triglycerides; *BMI*, body mass index; *NA*, Not applicable (because of adjusting for BMI). We did not correct for additional covariates when testing the association for BMI.

Model 1: Adjusted for age, sex, cell count, RNA quality score and technical covariates. Model 2: Model 1 + BMI and relevant medication.

P-value threshold $P< (0.05/200=) 2.25\times10^{-4}$; P-value threshold for replication: HDL; $P< (0.05/5=)0.01$, TG; $P< (0.05/15=)3.33\times10^{-3}$, Insul; $P< (0.05/2=)0.03$, BMI; $P< (0.05/16=)3.13\times10^{-3}$.

Genes that are presented bold passed the replication p-value threshold in 687 participants of the KORA study.

*According to the by Illumina provided annotation file

levels of three genes *(EFHD2*, *PRF1,* and *OSBPL5*). Likewise, we found the association of TG levels with 18 CpGs and six genes (*ICOS*, *GZMA*, *C12orf57*, *CD3D, CLDND1,* and *EFHD2*). Finally, we found BMI to be associated with 16 CpGs and five genes (*LRRN3, EFHD2, PID1, STOM,* and *CD3D*) (**Additional file 3: Table S6**). Of these, we were able to replicate the three-way association of TG with DNA methylation levels of cg04716530 and expression levels of *GZMA*, and DNA methylation levels of cg21566642 and expression levels of *CLDND1* in the KORA study. Furthermore, we found BMI to be associated with DNA methylation levels of 6 CpGs and expression of two genes (*LRRN3* and *PID1*) (**Table 4**).

In the three-way association (**Table 4**), we also identified CpGs associated with expression levels of genes far approximate from their annotated gene/loci. We did a lookup for the identified CpGs for eQTM association using data from the BIOS-BBMRI database (http://www.genenetwork.nl/ biosqtlbrowser/). Here we found *cis*- eQTMs between cg17287155 and expression of *EXOC3* and between cg03636183 and expression of *F2RL3*. In the Rotterdam Study, both *EXOC3* and *F2RL3* gene expression probes did not pass the QC. Hence, we could not test the influence of these genes in the identified eQTM associations in a three-way analysis.

Table 4. The DNA methylation sites associated with gene expression.

Gene Expression*				DNA methylation **			eQTM***	
ProbeID	**Effect**	**P-value**	**Trait**	**CpG**	**Effect**	**P-value**	**Coeff**	**P-value**
ILMN_1779324 (*GZMA*)	1.0033	6.01E-08	TG	**cg04716530**	-0.0122	3.58E-05	-11.7641	6.91E-12
ILMN_2352563 (*CLDND1*)	0.4003	6.52E-05	TG	**cg21566642**	-0.0215	1.91E-05	-5.1957	3.54E-19
ILMN_1671891 (*PID1*)	-0.0425	7.23E-09	BMI	cg03636183	0.0016	3.04E-05	-3.9797	1.28E-11
ILMN_1773650 (*LRRN3*)	-0.0669	3.77E-05	BMI	cg03636183	0.0016	3.04E-05	-16.4622	**3.46E-41**
				cg06096336	0.0017	9.51E-07	-15.0031	2.91E-24
				cg13708645	0.0015	6.72E-07	-9.5025	2.31E-09
				cg17287155	0.0012	8.69E-06	-26.2306	**3.09E-54**
				cg25649826	0.0009	1.63E-05	-14.4989	3.43E-08
				cg26361535	0.0016	1.85E-07	-12.549	2.30E-11
ILMN_2048591 (*LRRN3*)	-0.0604	1.46E-05	BMI	cg03636183	0.0016	3.04E-05	**-14.4435**	**2.67E-43**
				cg06096336	0.0017	9.51E-07	-11.5428	1.40E-19
				cg13708645	0.0015	6.72E-07	-8.2627	1.36E-09
				cg17287155	0.0012	8.69E-06	**-21.4408**	**1.19E-48**

The table shows an overview of the overlap of the hits with nominal significant (*P< 0.05*) replication in KORA in all three association analyses, including the association between 1) DNA methylation and cardio-metabolic traits, 2) gene expression and cardio-metabolic traits, and 3) the eQTM results for the gene and CpG that are associated with the same cardio-metabolic trait. P-value thresholds in the discovery for DNA methylation $P< (0.05/1{,}224=)4.08\times10^{-5}$, gene expression $P< (0.05/200=)2.25\times10^{-4}$ and for eQTM $P< (0.05/(443\times2{,}549))=4.4\times10^{-8}$. P-value thresholds in the replication for TG; gene expression $P< (0.05/15=)3.33\times10^{-3}$, DNA methylation $P< (0.05/35=)1.43\times10^{-3}$, and BMI; gene expression $P< (0.05/16=)3.13\times10^{-3}$, DNA methylation $P< (0.05/16=)3.13\times10^{-3}$, and eQTM $P< (0.05/(1{,}224\times200)=)2.04\times10^{-7}$ Results that are presented bold passed the replication P-value threshold in the KORA study * Expression probe ~ cardio-metabolic trait + age, sex, cell count, RNA quality score, technical covariates, BMI and relevant medication ** CpGs ~ cardio-metabolic trait + age, sex, cell count, technical covariates, BMI and relevant medication *** Expression probe ~ CpGs + age, sex *TG*, triglycerides; *BMI*, body mass index

Mediation analysis for smoking-related CpGs and genes associated with cardio-metabolic traits

As shown in **Figure 3**, we used mediation analysis to investigate the effect of DNA methylation and gene expression, independently, in the association between smoking and cardio-metabolic traits. Also, we tested the mediating effect of DNA methylation in the association between smoking and gene expression. In total, we conducted three different models; first, gene expression as a mediator in the observed association between smoking and cardio-metabolic traits (A1 and A2 in **Figure 3**); second, DNA methylation as a mediator in the observed association between smoking and gene expression (B1 and B2 in **Figure 3**); and third, DNA methylation as the mediator in the association between smoking and cardio-metabolic traits (C1 and C2 in **Figure 3**). We conducted the average causal mediation effect (ACME), average direct effect (ADE), and the proportion mediated (Prop. med.), which are illustrated in **Table 5** (and **Additional file 3: Table S7**). The ADE reflects the effect of smoking on the tested outcome that does not depend on the mediator. The R code for the mediation analyses is included in **Additional File 6** and an example input file is provided in **Additional file 7**.

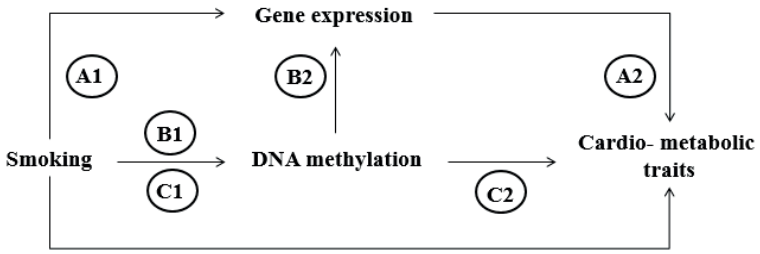


Figure 3. Schematic overview of the mediation analyses. We used mediation analysis to test the mediation effect of gene expression in the association between smoking and cardio-metabolic traits (A1 and A2). Furthermore, we tested the mediation effect of DNA methylation in the associations between smoking and gene expression (B1 and B2) and the mediation effect of DNA methylation in the association between smoking and cardio-metabolic traits (C1 and C2).

The mediation effect of the three-way associations as obtained in our discovery dataset (**Additional file 1: Table S6**) are provided in **Additional file 3: Table S7**. Out of the 124 mediation analysis conducted, there was significant mediation effect in 69 of them in the Rotterdam Study (**Additional file 3: Table S7**). Of these, we were able to replicate the mediating effect of cg01305745 (*VKORC1*) and cg16448702 (*INPP5D*) in the association between smoking and *PRF1* expression (ILMN_1740633).

Also, we identified the mediating effect of cg16448702 (*INPP5D*) in the association between smoking and *OSBPL5* (ILMN_1802151). Furthermore, we replicated the mediation effect of 9 CpGs in the association between smoking and *LRRN3* expression (ILMN_1773650 and ILMN_2048591) (**Additional file 3: Table S7**). Finally, of the replicated three-way associations as shown in **Table 4**, we were able to replicate the media-

tion effect of cg03636183 (*F2RL3*), cg06096336 (*PSMD1; HTR2B*), cg13708645 (*KDM2B*), and cg17287155 (*AHRR*) in the association between smoking and *LRRN3* expression (**Table 5**). We conducted the ρ at which ACME is 0, to test the models' sensitivity. Here, we obtained ρ's in the range between -0.1 and -0.5, and 0.1 and 0.4. A value of ρ close to 0 indicates that the assumption we made is sensitive to violations (24).

Table 5. Mediation effect of DNA methylation and gene expression in the association between smoking and cardio-metabolic traits.

Mediator	Outcome	ACME (95%CI)	ADE (95%CI)	Total Effect (95%CI)	Prob. Med. (95%CI)	ρ at which ACME is 0*
cg03636183 *(F2RL3)*	ILMN_1773650 *(LRRN3)*	0.6835 (0.4731/0.8869)	1.9603 (1.5832/2.3822)	2.6438 (2.2907/3.0048)	0.2585 (0.1767/0.3432)	-0.3
cg06096336 *(PSMD1;HTR2B)*	ILMN_1773650 *(LRRN3)*	0.1237 (0.0263/0.2396)	2.5202 (2.1796/2.849)	2.6438 (2.2907/3.0048)	0.0468 (0.0102/0.0886)	-0.4
cg13708645 *(KDM2B)*	ILMN_1773650 *(LRRN3)*	0.0768 (0.025/0.1408)	2.5671 (2.2153/2.9309)	2.6438 (2.2907/3.0048)	0.0290 (0.0092/0.0533)	-0.1
cg17287155 *(AHRR)*	ILMN_1773650 *(LRRN3)*	0.6357 (0.4798/0.8094)	2.0081 (1.6771/2.333)	2.6438 (2.2907/3.0048)	0.2405 (0.1835/0.3036)	-0.5
cg06096336 *(PSMD1;HTR2B)*	ILMN_2048591 *(LRRN3)*	0.0992 (0.0198/0.1915)	2.2838 (1.9578/2.5975)	2.3830 (2.0445/2.721)	0.0416 (0.0085/0.0779)	-0.3
cg17287155 *(AHRR)*	ILMN_2048591 *(LRRN3)*	0.5004 (0.3828/0.6542)	1.8826 (1.5691/2.2123)	2.3830 (2.0445/2.721)	0.2100 (0.1603/0.2724)	-0.4

The table shows the results of mediation analysis, in which current smoking is always used as exposure and are adjusted for age and sex.
ACME; Average Causal Mediation Effect, *ADE*; Average Direct Effect, *Prop. Med*; Proportion mediated
* ρ at which ACME is 0 indicates how sensitive our model is to the non-unmeasured confounding assumption.

DISCUSSION

The associations of smoking, gene expression, and DNA methylation with cardio-metabolic traits have been studied independently and reviewed in great detail (11, 25-28); however, the overlap between epigenetics and transcriptomics in the association between smoking and cardio-metabolic traits has been studied much less. This study investigated the relationship between previously identified smoking-related changes in DNA methylation (6) and gene expression (8), followed by their associations with cardio-metabolic traits within two population-based cohort studies. In this line, we first showed several significant *cis*- and *trans*-eQTM associations between smoking-related CpGs and gene expression probes. Furthermore, we replicated 26 smoking-related CpGs and 19 smoking-related genes (21 probes) associated with cardio-metabolic traits. Moreover, we showed three-way association of TG with two CpGs and two genes (*GZMA* and *CLDND1*), and BMI with six CpGs and two genes (*PID1* and *LRRN*). Finally, our study

demonstrated a mediating effect of 4 CpGs (cg03636183, cg06096336, cg13708645, and cg17287155) in the association between smoking and the BMI-related gene *LRRN3*.

Our results showed a three-way association between TG with the decrease in DNA methylation levels of cg21566642 and the increase in expression levels of *CLDND1*. In this line, smoking was associated with an increase in the expression of *CLDND1* (8) and a decrease in cg21566642 DNA methylation levels (6); and here, we showed the inverse relation between *CLDND1* expression and methylation levels at cg21566642. The expression of *CLDND1*, a tight junction protein, is shown to be highly increased in human Colon cancer samples and cell lines, and also positively correlated with tumor growth and disease progression (29). The inverse association between DNA methylation levels at cg21566642 and smoking was previously shown in blood samples with cross-tissue replications in adipose tissue and skin tissue (20). Additionally, cg21566642 is inversely associated with CVD risk (30), all-cause mortality (31), and with left ventricular mass (LVM) index in young adults (32). LVM index is an important cardiac remodeling trait that is an intermediate phenotype for heart failure. In line with this, an increased LVM index is associated with high levels of TG (33, 34) and with an increased risk of depressed left ventricular ejection fraction, coronary heart disease, congestive heart failure, and stroke (35, 36).

In the three-way association for BMI, we found that smoking is associated with lower BMI, indicating that current smokers are less likely to be obese than never smoker, which has been reported in several previous studies as well (37-39). Our results further showed that cg03636183 (*F2RL3*) was positively associated with BMI and negatively associated with the expression of *PID1* and *LRRN3*. Smoking was inversely associated with cg03636183 (6) and positively with *PID1* and *LRRN3* expression (8). Here we found an inverse relation between cg03636183 and expression levels of *PID1* and *LRRN3*. Due to the quality control implemented within the Rotterdam Study gene expression profiling data, gene expression data on *F2RL3* was not available. Therefore, we could not test if the association of cg03636183 with *PID1* and *LRRN3* expression levels was independent or via a downstream effect of *F2RL3* expression. Nonetheless, the inverse correlation between DNA methylation levels at cg03636183 and expression of *F2RL3* was previously shown (20). This might indicate that the identified eQTM associations are, at least partly, via *F2RL3* expression. *F2RL3* encodes the protease-activated receptor-4 (PAR-4), a protein expressed in various tissues that introduce platelet activation, intimal hyperplasia, and inflammation (40). Furthermore, the expression of *F2RL3* was associated with metabolic disease risk phenotypes, including a negative association with visceral fat mass and a positive association with total fat mass and android-to-gynoid fat ratio (20). Additionally, the inverse association between DNA methylation levels at cg03636183 and smoking has been shown in blood samples with cross-tissue replications in adipose and skin tissues (20). The inverse relation between DNA methylation levels at cg03636183 and

TG (41), all-cause mortality (31), lung cancer incidence and mortality (42), as well as total mortality and cardiovascular mortality (43) was also previously identified. Also, a smoking-related decrease in cg03636183 methylation levels appears to increase serum levels of IL-18 (44). IL-18 promotes the synthesis of IL-6, which stimulates the production of serum CRP (45, 46). The increase of IL-18 and IL-6 leads to a higher risk ratio for CHD development (47). Moreover, the increase in serum CRP concentrations results in increased risk ratios for CHD, ischaemic stroke, vascular mortality, and non-vascular mortality (48).

Two of the CpGs, cg26361535 (*ZC3H3*) and cg25649826 (*USP22*), for which we found a three-way association with BMI and *LRRN3*, have been reported to be positively associated with BMI (49). Both CpGs are cross-tissue replicated in adipose tissue and in isolated adipocytes for obese cases *versus* normal-weight controls. The association with cg26361535 was in the same direction and for cg25649826 in the opposite direction as obtained in our results (49). Additionally, both CpGs were positively associated in blood with weight, WHR, glucose, insulin, TG, and CRP, and negatively with HDL. Furthermore, cg26361535 was positively associated with SBP and DBP (49) and all-cause mortality (31).

Finally, we identified a three-way association between BMI, an increase in methylation levels at cg17287155 (*AHRR*), and *LRRN3* expression. Smoking is negatively associated with DNA methylation levels at cg17287155 (6) and, as we replicated here, positively associated with BMI (50). Notably, in the eQTM look-up we found a *cis*– eQTM for cg17287155 with the expression of *EXOC3,* instead of with its annotated gene (*AHRR*). *AHRR* is a well-studied gene in relation to smoking (5) and is a key regulator of the Xenobiotic metabolism pathway responsible for detoxification of polyaromatic hydrocarbons (PAHs) in tobacco smoke (51, 52). Nevertheless, *EXOC3* overexpression increases insulin-induced glucose uptake in adipocytes (53), indicating a possible link for *EXOC3* with CVD related risk factors. Further research is needed to verify the eQTM-associations for cg17287155 with *EXOC3* and its impact on the eQTM-associations identified in the current study.

The identified associations and mediating effects in our study indicate a possible regulatory effect of DNA methylation on the expression levels of genes far from the neighboring methylation site, which so-called *trans*-regulatory effect of methylated CpG sites on gene expression (54). So far, most previous studies have limited their research to the correlation between gene expression and DNA methylation at CpGs located in the nearby regions and in the gene body, or the *cis*-regulatory effect. In this line, a recent study has shown the *trans*-regulatory effect of DNA methylation in the associations with gene expression and chronic obstructive pulmonary disease (54). Therefore, future research is needed with a broader methodological approach, including examining pos-

sible *trans*-regulatory effects to gain more insight into the epigenetic regulatory effects in disease studies.

This study has strengths as well as limitations that should be considered when interpreting the results. The main strengths of this study include the availability of DNA methylation data in a large sample of adults from the general population overlapping with transcriptomic and clinical data. Another strength is the use of the largest available EWAS (6) and TWAS (8) to date for selecting the CpGs and genes of interest associated with smoking. A limitation of the current study could be that data on smoking habits are retrieved from questionnaires, which might be underestimating actual smoking levels possibly leading to information bias (55-57). This self-reporting bias can arise due to several reasons, such as recall bias in which a participant might not remember the true exposure or social desirability bias in which participants deliberately underestimate due to the socially stigmatized nature (57). However, we expect the underestimation to be primarily quantitative and should not significantly impact the current *versus* non-smoker categorization we used in this study. Also, the questionnaires used for smoking data-collection did not include information regarding passive smoking, which is a risk factor for CVD (58). As a result, we were not able to adjust for the passive-smoking effect in our analysis. As these participants are included in the non-smoker group, this might have underestimated the true effect.

Furthermore, due to the nature of the current study we have included the same participants in all mediation analyses and have used the mediator and exposure measurements on the same time-point; therefore, we cannot rule out reverse causality. Another limitation is that DNA methylation and gene expression levels were only measured at baseline; hence, we have no access to pre-measurement covariates. Consequently, we could not further adjust our models without risking the adjustment of a mediator, which could explain the ρ values close to 0 we obtained in a subset of our models in the sensitivity analysis. However, we did include additional adjustments (e.g. BMI and relevant medication) in the association analysis between cardio-metabolic traits with DNA methylation and gene expression, indicating the robustness of the identified three-way associations. Also, due to the stringent quality control in the Rotterdam Study, we were not able to test the impact of the *cis*- eQTM genes in the identified eQTMs. Finally, the use of whole-blood for the quantification of DNA methylation and transcriptomics associated with smoking and cardio-metabolic traits could be a limitation, since DNA methylation and gene expression are tissue-specific. Nonetheless, these data from other tissues are currently not available in the majority of population-based studies including the two participating cohorts in this study.

CONCLUSION

In this study, we tested the association of smoking-related changes in DNA methylation and gene expression with cardio-metabolic traits. We found a three-way association of TG and BMI with CVD-relevant CpG sites and genes. Our results may provide further insight into the possible molecular cascades linking smoking to metabolic risk factors leading to CVD. Further research is warranted to conduct experimental research on the molecular mechanisms of the impact of smoking on cardiovascular disease and its risk factors through changes in DNA methylation and gene expression levels.

METHODS

Study population

The discovery data set comprised a total of 1,412 participants included in the Rotterdam Study; the design from the Rotterdam Study has been described elsewhere (21). Briefly, in 1990 all residents of Ommoord, a district in Rotterdam, aged 55 years and older, were invited for participation (RS-I). In 2000, the cohort was extended with participants who had reached the age of 55 years or who had moved into the district (RS-II). An additional group was invited in 2006, from the age of 45 years and older (RS-III). Participants have been re-examined every 3–4 years. In the current study, we used data from the third visit from RS-II (RS-II-3) and the first and second visit of RS-III (RS-III-1 and RS-III-2). In total, DNA methylation measurements of 1,412 participants from RS-III-1, RS-II-3, and RS-III-2 were included in our analysis. Additionally, gene expression data was available for 716 participants included in RS-III-1. Smoking information was collected via self-reported questionnaires, additional data collection details are described in **Additional file 8**.

The replication data comprised a total of 1,717 participants included in The Cooperative Health Research in the Region of Augsburg (KORA) study. The KORA study is a series of independent population-based epidemiological surveys and follow-up studies of participants living in the region of Augsburg, Southern Germany. The KORA F4 study, a 7-year follow-up study of the KORA S4 survey (examined 1999-2001), was conducted between 2006 and 2008. The standardized examinations applied in the survey have been described in detail elsewhere (21). A total of 3,080 subjects with ages ranging from 32 to 81 years participated in the examination. In a random subgroup of 1,802 KORA F4 subjects, the genome-wide DNA methylation patterns were analyzed as described in **Additional file 3**. Smoking information was collected via self-reported questionnaires, additional data collection details are described in **Additional file 8**.

DNA methylation data

DNA methylation in the Rotterdam Study and KORA study was extracted from whole peripheral blood and DNA methylation measurements were obtained using the Illumina Infinium Human Methylation 450K BeadChip (Illumina Inc, San Diego, CA, USA). The DNA methylation pre-processing procedures are described in **Additional file 3**. The methylation proportion of a CpG site was reported as a methylation β-value in the range of 0 to 1. Genome coordinates provided by Illumina (GRCh37/hg19) were used to identify independent loci.

In the current study, CpGs of interest were selected using a recent EWAS (6) investigating the association between tobacco smoking and changes in DNA methylation values in the epigenome. In total, 2,623 CpG sites were identified as being significantly ($P<1\times10^{-7}$) differentially methylated between smokers and never smokers. In the Rotterdam Study, 2,549 out of the 2,623 CpGs passed the quality control and are included in this study (**Additional file 3: Table S8**).

RNA expression data

In the Rotterdam Study, RNA was isolated from whole blood and gene expression profiling was performed using the IlluminaHumanHT-12v4 Expression Beadchips (Illumina, San Diego, CA, USA). The expression dataset is available at Gene Expression Omnibus (GEO) public repository under the accession GSE33828: 881 samples are available for analysis. In KORA F4, total RNA was extracted from whole blood and the Illumina Human HT-12 v3 Expression BeadChip (Illumina, San Diego, CA, USA) was used for gene expression profiling (59). A more detailed description is implemented in **Additional file 8**.

In the current study, genes of interest were selected using a previous TWAS testing the association between gene expression and current *versus* never-smoking status (8). In this TWAS, the meta-analysis was performed on all transcripts with matching gene Entrez IDs. Employing a significance threshold of FDR<0.05, 886 significant gene Entrez IDs were identified, of which 387 replicated in an independent dataset. Employing the annotation file provided by the Illumina (HumanHT-12_V4), we found 502 gene expression probes to be annotated to these gene Entrez IDs out of which 443 were present in the Rotterdam Study and were included in the current study (**Additional file 8: Table S9**).

Correlation between DNA methylation and gene expression

Since DNA methylation and gene expression may affect each other (i.e. eQTMs), we tested the association between 2,549 CpGs and 443 gene expression probes linked to smoking in participants who had both methylation and gene expression data available in the Rotterdam Study (N= 716). We regressed out age, sex, blood cell counts (fixed effect), and technical covariates (random effect) on the normalized beta-values of the

CpGs and separately on the mRNA expression levels using a linear mixed model analysis. The association between the residuals of DNA methylation (independent variable) and gene expression (dependent variable) was examined using a linear regression model. The robust Bonferroni-corrected P-value threshold for a significant association was $P< 4.4\times10^{-8}$ (0.05/(443×2549)).

Additionally, we randomly selected 443 gene expression probes from the IlluminaHumanHT12v4 Expression Beadchips, and 2,549 CpGs from the Illumina Human 450K array, that were available in the Rotterdam Study. Using the same methods mentioned above, we tested the association between the 2,549 smoking-related CpGs with the 443 randomly selected gene expression probes, and the association between 2,549 randomly selected CpGs with the 443 smoking- associated gene expression probes. The chi-square test of independence was used to test possible enrichment for the smoking effect.

Association of DNA methylation and gene expression with cardio-metabolic traits

We studied the relationship of cardio-metabolic traits with (1) smoking-CpGs associated with at least one smoking-gene probe, and (2) smoking-gene probes associated with at least one smoking-CpG. We included the following cardio-metabolic related phenotypes: HDL, LDL, TG, serum cholesterol, fasting glucose and insulin levels, SBP, DBP, WHR, and BMI.

First, we tested the association between the smoking-related CpGs (dependent variable) with the cardio-metabolic traits (exposure variable) using linear mixed effects models (LME4 package in R). The selected covariates in model 1 with fixed effects were age, sex, and cell counts for granulocytes, lymphocytes and monocytes. Array number and position number on array were added in the model as covariates with random effect to correct for batch effect. In model 2, we additionally adjusted for BMI and relevant medication, including for lipid exposures (lipid-lowering medication), for glycemic traits (glucose-lowering medication), for SBP and DBP (lipid-lowering medication and anti-hypertensives, diuretics, beta-blockers, calcium channel blockers, and RAAS modifying agents).

Second, we tested the association between gene expression (dependent variable) and the cardio-metabolic traits (exposure variable) using linear mixed-effects models (LME4 package in R), adjusting for age, sex, blood cell counts (granulocytes, lymphocytes, and monocytes), RNA quality score and batch effect. In model 2, we additionally adjusted for BMI and relevant medication (as described for DNA methylation).

Third, we combined our EWAS and TWAS results and showed the obtained three-way association; CpG *versus* gene expression; cardio-metabolic trait *versus* CpG; cardio-metabolic trait versus gene expression. For the CpG *versus* gene expression, we did a lookup for the identified CpGs to identify possible *cis*- eQTM associations using data from five

Dutch biobanks (BIOS-BBMRI database) in a total of 3,841 whole blood samples (http://www.genenetwork.nl/ biosqtlbrowser/).

Mediation analysis

CpGs and gene expression probes associated with each other and associated with the same cardio-metabolic trait were reviewed in three mediation analyses (**Figure 3**); (1) the mediation of gene expression in the association between smoking status and the cardio-metabolic trait, (2) the mediation of DNA methylation in the association between smoking status and gene expression changes, and (3) the mediation of DNA methylation in the association between smoking status and the cardio-metabolic trait. In all three analyses, we included the same participants, current *versus* non-smokers as exposure and all models are corrected for age and sex. In the first analysis, we used the gene expression as potential mediator and the cardio-metabolic trait as outcome. In the second analysis, we used DNA methylation as possible mediator and the gene expression as outcome. In the third analysis, we used DNA methylation as possible mediator and the cardio-metabolic trait as outcome. We used the "mediate" function in the mediation package in R (60), using the bootstrap method including 1000 simulations and confidence intervals using the BCa method (61). The proportion mediated describes the average magnitude of indirect association between smoking status and the gene expression or cardio-metabolic trait attributed through changes in DNA methylation or gene expression relative to the average total association, and it is calculated by dividing the average causal mediation effect by the average total effect (62). Asymptotic 95% confidence intervals (CI) were obtained from nonparametric bootstrapping with 1000 iterations. These mediation analyses assumed no additional unmeasured confounding; however, if unobserved variables confound the models, the unmeasured confounding assumption is violated. Therefore, we used the sensitivity analysis included in the mediation package using the "medsens" function conducted by varying the values of ρ and determine the ρ at which ACME is 0 per model. Obtaining a value of ρ close to 0 indicates that the assumption is sensitive to violations, meaning that having a confounder with a higher correlation than the value of ρ, the assumption of no additional unmeasured confounding likely does not hold (24).

Replication in the KORA study

The identified associations in the Rotterdam Study were replicated using the same models in the KORA study. The adjustment for blood cell counts (monocytes, granulocytes, and lymphocytes) was based on Houseman estimates rather than laboratory measurements (63). Furthermore, principal components were used to adjust for technical covariates rather than plate number and position on array.

Statistical analysis

All analyses were performed using the statistical package R. The eQTM analysis and the associations of the cardio-metabolic traits with smoking-related CpGs and genes were conducted in R (version 3.2.0) under a Linux operating system, using the “LME4” package (version 1.1-16) and the “parallel” package (version 3.2.0). The mediation analyses were conducted in R studio Desktop (version 3.2.0) under Windows operating system using the “mediation” package (version 4.4.6.). Data collection and related statistical methods are provided in **Additional file 8**.

SUPPLEMENTARY MATERIAL

Supplemental material for this chapter can be found in the online version of the paper via https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-020-00951-0.

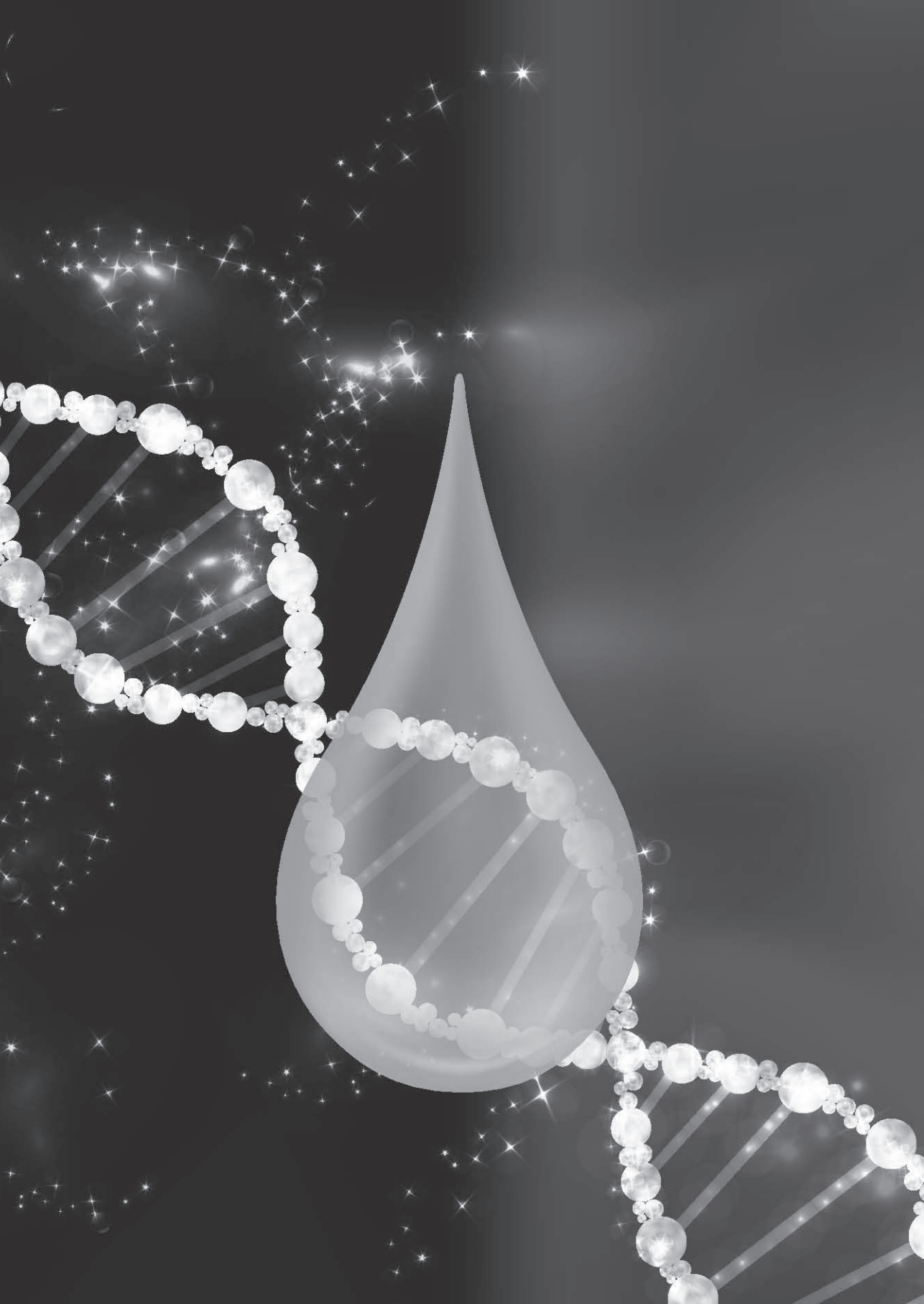
REFERENCES

1. WHO. WHO global report on mortality attributale to tobacco. 2012:392.
2. Collaborators GCoD. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736-88.
3. Sun K, Liu J, Ning G. Active smoking and risk of metabolic syndrome: a meta-analysis of prospective studies. PLoS One. 2012;7(10):e47791.
4. Mottillo S, Filion KB, Genest J, Joseph L, Pilote L, Poirier P, et al. The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis. J Am Coll Cardiol. 2010;56(14):1113-32.
5. Kaur G, Begum R, Thota S, Batra S. A systematic review of smoking-related epigenetic alterations. Arch Toxicol. 2019;93(10):2715-40.
6. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;9(5):436-47.
7. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One. 2013;8(5):e63812.
8. Huan T, Joehanes R, Schurmann C, Schramm K, Pilling LC, Peters MJ, et al. A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. Hum Mol Genet. 2016;25(21):4611-23.
9. Charlesworth JC, Curran JE, Johnson MP, Goring HH, Dyer TD, Diego VP, et al. Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. BMC Med Genomics. 2010;3:29.
10. Vink JM, Jansen R, Brooks A, Willemsen G, van Grootheest G, de Geus E, et al. Differential gene expression patterns between smokers and non-smokers: cause or consequence? Addict Biol. 2017;22(2):550-60.
11. Dhana K, Braun KVE, Nano J, Voortman T, Demerath EW, Guan W, et al. An Epigenome-Wide Association Study of Obesity-Related Traits. Am J Epidemiol. 2018;187(8):1662-9.
12. Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. Clin Epigenetics. 2017;9:15.
13. Richard MA, Huan T, Ligthart S, Gondalia R, Jhun MA, Brody JA, et al. DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. Am J Hum Genet. 2017;101(6):888-902.
14. Liu J, Carnero-Montoro E, van Dongen J, Lent S, Nedeljkovic I, Ligthart S, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. Nat Commun. 2019;10(1):2581.
15. Chen BH, Hivert MF, Peters MJ, Pilling LC, Hogan JD, Pham LM, et al. Peripheral Blood Transcriptomic Signatures of Fasting Glucose and Insulin Concentrations. Diabetes. 2016;65(12):3794-804.
16. Huan T, Esko T, Peters MJ, Pilling LC, Schramm K, Schurmann C, et al. A meta-analysis of gene expression signatures of blood pressure and hypertension. PLoS Genet. 2015;11(3):e1005035.
17. Ligthart S, Steenaard RV, Peters MJ, van Meurs JB, Sijbrands EJ, Uitterlinden AG, et al. Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. Diabetologia. 2016.
18. Steenaard RV, Ligthart S, Stolk L, Peters MJ, van Meurs JB, Uitterlinden AG, et al. Tobacco smoking is associated with methylation of genes related to coronary artery disease. Clin Epigenetics. 2015;7(1):54.

19. Zhang Y, Schöttker B, Florath I, Stock C, Butterbach K, Holleczek B, et al. Smoking-Associated DNA Methylation Biomarkers and Their Predictive Value for All-Cause and Cardiovascular Mortality. Environ Health Perspect. 2016;124(1):67-74.
20. Tsai PC, Glastonbury CA, Eliot MN, Bollepalli S, Yet I, Castillo-Fernandez JE, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. Clin Epigenetics. 2018;10(1):126.
21. Ikram MA, Brusselle G, Ghanbari M, Goedegebure A, Ikram MK, Kavousi M, et al. Objectives, design and main findings until 2020 from the Rotterdam Study. Eur J Epidemiol. 2020;35(5):483-517.
22. Holle R, Happich M, Löwel H, Wichmann HE, Group MKS. KORA--a research platform for population based health research. Gesundheitswesen. 2005;67 Suppl 1:S19-25.
23. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131-8.
24. Muthén BO, Muthén LK, Asparouhov T. Regression and mediation analysis using Mplus: Muthén & Muthén Los Angeles, CA; 2017.
25. Joehanes R, Ying S, Huan T, Johnson AD, Raghavachari N, Wang R, et al. Gene expression signatures of coronary heart disease. Arterioscler Thromb Vasc Biol. 2013;33(6):1418-26.
26. Huan T, Zhang B, Wang Z, Joehanes R, Zhu J, Johnson AD, et al. A systems biology framework identifies molecular underpinnings of coronary heart disease. Arterioscler Thromb Vasc Biol. 2013;33(6):1427-34.
27. Burns DM. Epidemiology of smoking-induced cardiovascular disease. Prog Cardiovasc Dis. 2003;46(1):11-29.
28. Muka T, Nano J, Voortman T, Braun KVE, Ligthart S, Stranges S, et al. The role of global and regional DNA methylation and histone modifications in glycemic traits and type 2 diabetes: A systematic review. Nutr Metab Cardiovasc Dis. 2016;26(7):553-66.
29. Dhawan P, Singh AB, Deane NG, No Y, Shiou SR, Schmidt C, et al. Claudin-1 regulates cellular transformation and metastatic behavior in colon cancer. J Clin Invest. 2005;115(7):1765-76.
30. Fernandez-Sanles A, Sayols-Baixeras S, Curcio S, Subirana I, Marrugat J, Elosua R. DNA Methylation and Age-Independent Cardiovascular Risk, an Epigenome-Wide Approach: The REGICOR Study (REgistre GIroni del COR). Arterioscler Thromb Vasc Biol. 2018;38(3):645-52.
31. Svane AM, Soerensen M, Lund J, Tan Q, Jylhava J, Wang Y, et al. DNA Methylation and All-Cause Mortality in Middle-Aged and Elderly Danish Twins. Genes (Basel). 2018;9(2).
32. Sabogal C, Su S, Tingen M, Kapuku G, Wang X. Cigarette smoking related DNA methylation in peripheral leukocytes and cardiovascular risk in young adults. Int J Cardiol. 2020;306:203-5.
33. Jorgensen PG, Jensen MT, Biering-Sorensen T, Mogelvang R, Galatius S, Fritz-Hansen T, et al. Cholesterol remnants and triglycerides are associated with decreased myocardial function in patients with type 2 diabetes. Cardiovasc Diabetol. 2016;15(1):137.
34. de las Fuentes L, Waggoner AD, Brown AL, Davila-Roman VG. Plasma triglyceride level is an independent predictor of altered left ventricular relaxation. J Am Soc Echocardiogr. 2005;18(12):1285-91.
35. Drazner MH, Rame JE, Marino EK, Gottdiener JS, Kitzman DW, Gardin JM, et al. Increased left ventricular mass is a risk factor for the development of a depressed left ventricular ejection fraction within five

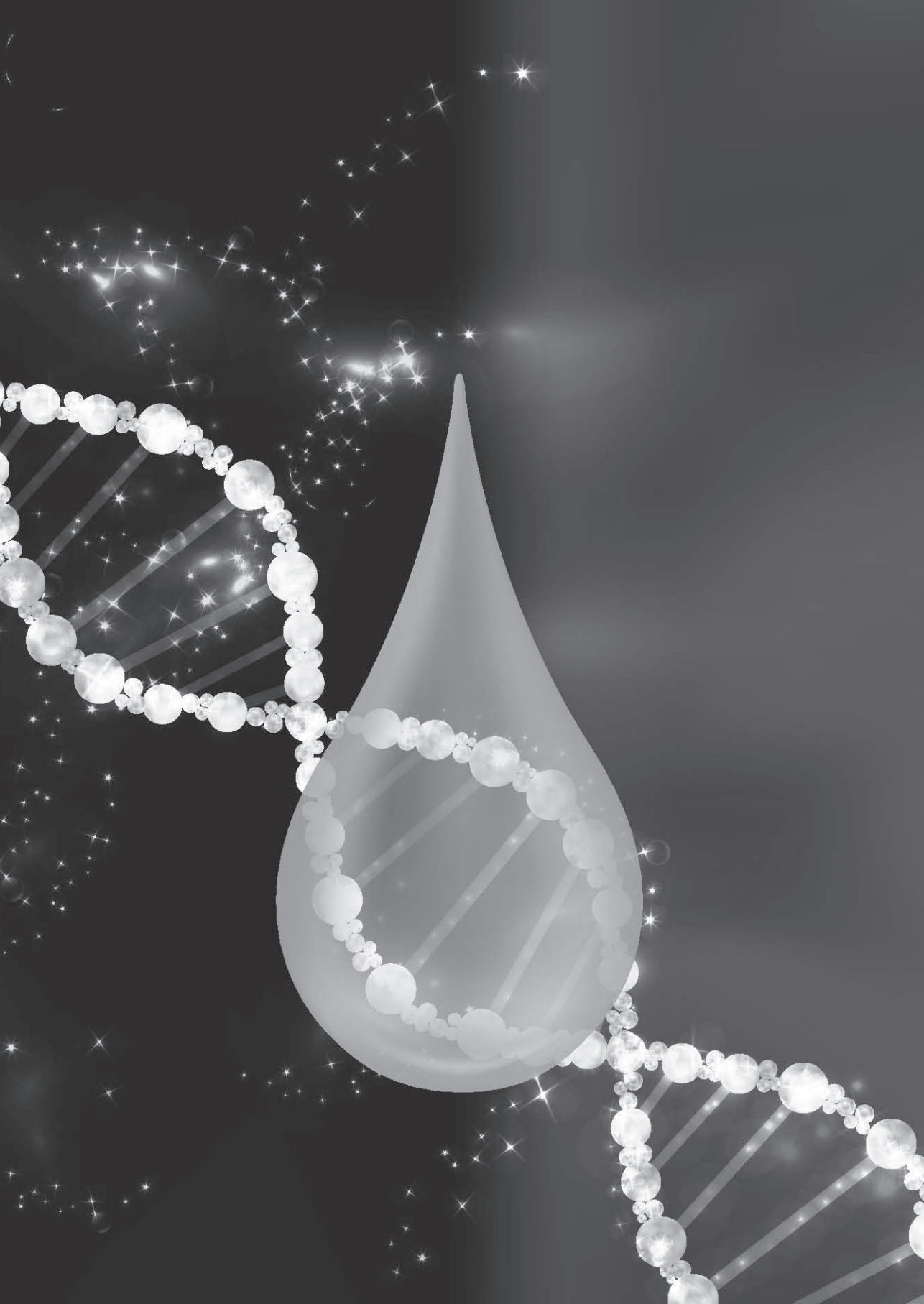
years: the Cardiovascular Health Study. J Am Coll Cardiol. 2004;43(12):2207-15.
36. Gardin JM, McClelland R, Kitzman D, Lima JA, Bommer W, Klopfenstein HS, et al. M-mode echocardiographic predictors of six- to seven-year incidence of coronary heart disease, stroke, congestive heart failure, and mortality in an elderly cohort (the Cardiovascular Health Study). Am J Cardiol. 2001;87(9):1051-7.
37. Molarius A, Seidell JC, Kuulasmaa K, Dobson AJ, Sans S. Smoking and relative body weight: an international perspective from the WHO MONICA Project. J Epidemiol Community Health. 1997;51(3):252-60.
38. Shimokata H, Muller DC, Andres R. Studies in the distribution of body fat. III. Effects of cigarette smoking. Jama. 1989;261(8):1169-73.
39. Dare S, Mackay DF, Pell JP. Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the UK general population. PLoS One. 2015;10(4):e0123579.
40. Leger AJ, Covic L, Kuliopulos A. Protease-activated receptors in cardiovascular diseases. Circulation. 2006;114(10):1070-7.
41. Dekkers KF, van Iterson M, Slieker RC, Moed MH, Bonder MJ, van Galen M, et al. Blood lipids influence DNA methylation in circulating cells. Genome Biol. 2016;17(1):138.
42. Zhang Y, Schottker B, Ordonez-Mena J, Holleczek B, Yang R, Burwinkel B, et al. F2RL3 methylation, lung cancer incidence and mortality. Int J Cancer. 2015;137(7):1739-48.
43. Zhang Y, Yang R, Burwinkel B, Breitling LP, Holleczek B, Schottker B, et al. F2RL3 methylation in blood DNA is a strong predictor of mortality. Int J Epidemiol. 2014;43(4):1215-25.
44. Jhun MA, Smith JA, Ware EB, Kardia SLR, Mosley TH, Jr., Turner ST, et al. Modeling the Causal Role of DNA Methylation in the Association Between Cigarette Smoking and Inflammation in African Americans: A 2-Step Epigenetic Mendelian Randomization Study. Am J Epidemiol. 2017;186(10):1149-58.
45. Gerdes N, Sukhova GK, Libby P, Reynolds RS, Young JL, Schonbeck U. Expression of interleukin (IL)-18 and functional IL-18 receptor on human vascular endothelial cells, smooth muscle cells, and macrophages: implications for atherogenesis. J Exp Med. 2002;195(2):245-57.
46. Jones SA, Novick D, Horiuchi S, Yamamoto N, Szalai AJ, Fuller GM. C-reactive protein: a physiological activator of interleukin 6 receptor shedding. J Exp Med. 1999;189(3):599-604.
47. Kaptoge S, Seshasai SR, Gao P, Freitag DF, Butterworth AS, Borglykke A, et al. Inflammatory cytokines and risk of coronary heart disease: new prospective study and updated meta-analysis. Eur Heart J. 2014;35(9):578-89.
48. Emerging Risk Factors C, Kaptoge S, Di Angelantonio E, Lowe G, Pepys MB, Thompson SG, et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet. 2010;375(9709):132-40.
49. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541(7635):81-6.
50. Aslibekyan S, Demerath EW, Mendelson M, Zhi D, Guan W, Liang L, et al. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. Obesity (Silver Spring). 2015;23(7):1493-501.
51. Larigot L, Juricek L, Dairou J, Coumoul X. AhR signaling pathways and regulatory functions. Biochim Open. 2018;7:1-9.
52. Vu AT, Taylor KM, Holman MR, Ding YS, Hearn B, Watson CH. Polycyclic Aromatic Hydrocarbons in the Mainstream Smoke of

Popular U.S. Cigarettes. Chem Res Toxicol. 2015;28(8):1616-26.
53. Ewart MA, Clarke M, Kane S, Chamberlain LH, Gould GW. Evidence for a role of the exocyst in insulin-stimulated Glut4 trafficking in 3T3-L1 adipocytes. J Biol Chem. 2005;280(5):3812-6.
54. Yoo S, Takikawa S, Geraghty P, Argmann C, Campbell J, Lin L, et al. Integrative analysis of DNA methylation and gene expression data identifies EPAS1 as a key regulator of COPD. PLoS Genet. 2015;11(1):e1004898.
55. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res. 2009;11(1):12-24.
56. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. Nephron Clin Pract. 2010;115(2):c94-9.
57. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7.
58. Khoramdad M, Vahedian-Azimi A, Karimi L, Rahimi-Bashar F, Amini H, Sahebkar A. Association between passive smoking and cardiovascular disease: A systematic review and meta-analysis. IUBMB Life. 2020;72(4):677-86.
59. Schurmann C, Heim K, Schillert A, Blankenberg S, Carstensen M, Dörr M, et al. Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. PLoS One. 2012;7(12):e50938.
60. Dustin Tingley TY, Kentaro Hirose, Luke Keele, Kosuke Imai. mediation: R Package for Causal Mediation Analysis. Journal of Statistical Software. 2014;59 (2014)(5).
61. DiCiccio TJ, Efron B. Bootstrap confidence intervals. 1996;11(3):189-228.
62. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods. 2010;15(4):309-34.
63. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.



Chapter 4

MicroRNA expression and health outcomes



Chapter 4.1

Multi-omics analysis reveals microRNAs associated with cardiometabolic disorders

Michelle M. J. Mens, **Silvana C. E. Maas**, Jaco Klap, Gerrit Jan Weverling, Paul Klatser, Just P. J. Brakenhoff, Joyce B. J. van Meurs, André G. Uitterlinden, M. Arfan Ikram, Maryam Kavousi and Mohsen Ghanbari

Frontiers in Genetics 2020;11:110

ABSTRACT

MicroRNAs (miRNAs) are non-coding RNA molecules that regulate gene expression. Extensive research has explored the role of miRNAs in the risk for type 2 diabetes (T2D) and coronary heart disease (CHD) using single-omics data, but much less by leveraging population-based omics data. Here we aimed to conduct a multi-omics analysis to identify miRNAs associated with cardiometabolic risk factors and diseases. First, we used publicly available summary statistics from large-scale genome-wide association studies to find genetic variants in miRNA-related sequences associated with various cardiometabolic traits, including lipid and obesity-related traits, glycemic indices, blood pressure, and disease prevalence of T2D and CHD. Then, we used DNA methylation and miRNA expression data from participants of the Rotterdam Study to further investigate the link between associated miRNAs and cardiometabolic traits. After correcting for multiple testing, 180 genetic variants annotated to 67 independent miRNAs were associated with the studied traits. Alterations in DNA methylation levels of CpG sites annotated to 38 of these miRNAs were associated with the same trait(s). Moreover, we found that plasma expression levels of 8 of the 67 identified miRNAs were also associated with the same trait. Integrating the results of different omics data showed miR-10b-5p, miR-148a-3p, miR-125b-5p, and miR-100-5p to be strongly linked to lipid traits. Collectively, our multi-omics analysis revealed multiple miRNAs that could be considered as potential biomarkers for early diagnosis and progression of cardiometabolic diseases.

INTRODUCTION

Type 2 diabetes mellitus (T2D) is a complex metabolic disease that is characterized by insulin resistance and impairment of insulin secretion, which leads to hyperglycemia. The presence of T2D leads to a two- to four-fold increase risk of developing coronary heart disease (CHD) (1), which is among the leading causes of morbidity and mortality worldwide (2). Many risk factors are identified as mediators of these diseases, including hypertension, dyslipidemia, central adiposity and elevated blood glucose, which are together known as cardiometabolic traits (3). Despite substantial advances in diagnosis and widely prescribed drugs for these diseases, their rate continue to increase worldwide, emphasizing the need for deeper insights into underlying mechanisms and innovative therapeutic strategies. Cardiometabolic traits and diseases have underlying genetic components and many loci have been discovered through large-scale genome- and epigenome-wide association studies (4, 5). However, most of the identified genetic variants do not affect protein sequences, but are thought to affect gene regulation. One of the potential regulatory mechanisms involved might be microRNAs (miRNAs).

MiRNAs represent a class of small non-coding RNAs, which function as post-transcriptional regulators of gene expression via targeting the 3' untranslated region of target transcripts (6). Over the past years, miRNAs have emerged as key regulators of biological processes underlying T2D and CHD. In this context, aberrant expression and function of miRNAs, such as miR-33, miR-208, miR-133, and miR-124, have been shown to be associated with lipid metabolism, insulin secretion, myocardial infarction and T2D (7-9). Most of the disease-associated miRNAs have been discovered in cells originated from tissue of interest in small number of samples or animal studies. But advances in high-throughput technologies make it possible to study miRNAs in a population-based manner. In particular cell-derived vesicles, known as exosomes, release miRNAs in the blood stream that are very stable and can be used as biomarkers for disease (10).

Similar to other regulatory RNA molecules, the function and expression of miRNAs can be affected by genetic variants. Single-nucleotide polymorphisms (SNPs) can occur at various stages of the miRNA biogenesis including precursor- and mature miRNA sequences (11) as well as within regulatory elements, such as promoter regions (12). Also, DNA methylation can control transcription, which have been reported to be associated with the expression level of miRNAs (13). In this context, epigenome-wide association studies (EWAS) have shown that altered DNA methylation within miRNA promoters is associated with miRNAs expression levels and therewith modify disease risk (14). However, previous studies are mainly based on single omics data or small sample size (15, 16). As each type of omics data provides associations that can be useful for detecting development or progression of disease, integrating different omics layers can limit passive correlations and provide a more comprehensive view of the disease biology.

In this study, we applied a multi-omics approach to identify miRNAs associated with cardiometabolic traits. First, we identified genetic variants in miRNA sequences and their potential regulatory regions associated with different cardiometabolic risk factors and diseases using genetic association data from the available genome-wide association studies (GWAS). We then integrated population-based DNA methylation and miRNA expression data from the Rotterdam Study to link omics layers, strengthening the association of the identified miRNAs with cardiometabolic traits. We envision that the identified miRNAs could be considered as potential biomarkers for early diagnosis of cardiometabolic diseases.

METHODS

A graphical overview of the multi-omics approach used in this study is illustrated in **Figure 1**.

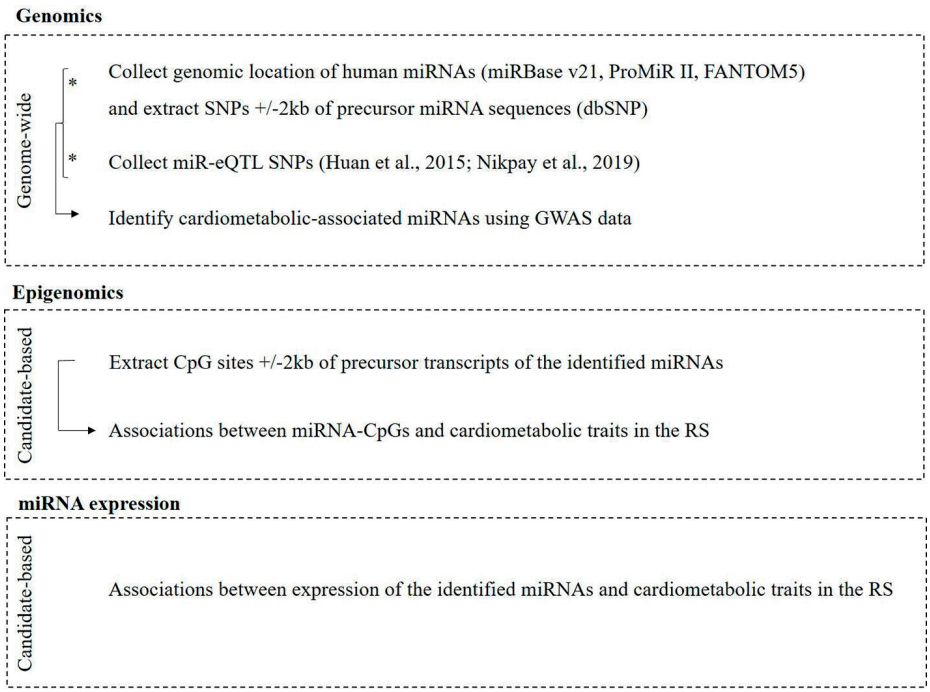


Figure 1. Overview of the multi-omics layers used in this study.

Retrieval of SNPs in miRNA-related regions

The primary transcripts of miRNAs for the processing to mature miRNAs are approximately 3-4kb in length (17). We collected the genomic position of all human miRNAs employing the miRBase database (v21) (18), ProMiR II (19) and FANTOM5 (12). Using dbSNP database (20), we extracted 18,545 SNPs located in +/-2kb of the precursor miRNA sequences (pre-miRNA) of 1,554 known miRNAs. Of these, 2,420 SNPs are located in pre- and mature sequences of miRNAs. Genetic variants have been found to alter miRNA expression and are known as miRNA expression quantitative trait loci (miR-eQTLs). To this end, we included 5,528 miR-eQTLs that change the expression of 221 mature miRNA using data from the Framingham Heart Study (FHS) (21) and from the Ottawa Hospital Bariatric Centre (22).The FHS focused on *cis*-miR-eQTLs, of which the majority was located 300-500kb away from their target miRNA. Nikpay et al. (2019) investigated both *cis*-miR-eQTLs and *trans*-miR-eQTLs, however, they reported likewise FHS that most cis-miR-eQTLs were distal regulators of the miRNAs. There were 83 miR-eQTLs overlapping with the SNPs in +/-2kb of the precursor miRNA sequence. Altogether, 23,990 unique SNPs were included in our analysis.

The genomic location of miRNAs can be discriminated among intergenic and intragenic. Roughly half of the known miRNAs are found to be transcribed from intergenic regions of the genome, suggesting that these miRNAs are transcribed under independent control of regulatory elements (23). The intragenic miRNAs are embedded within sequences of protein-coding genes, including intronic and exonic regions. If the intragenic miRNA and its host gene share the same promoter, the miRNA is likely to be co-expressed with the host gene (24). Here, the genomic location of the identified miRNAs was obtained using miRIAD (25).

Genome-Wide Association Studies of Cardiometabolic Traits

Cardiometabolic risk factors and diseases in this study were classified into four specific trait groups based on their shared pathophysiology and underlying pathways. These include (i) Anthropometric traits: body mass index (BMI), waist to hip ratio (WHR) and waist circumference (WC); (ii) Glycemic traits: fasting glucose (FG), glucose 2 hours (G2H), fasting insulin (FI), proinsulin (Pro-Ins), hemoglobin A1c (HbA1c), homeostatic model assessment of insulin resistance (HOMA-IR), β-cell function (HOMA-β), and type 2 diabetes mellitus (T2D); (iii) Lipid traits: low-density lipoprotein (LDL), high-density lipoprotein (HDL), total serum cholesterol (TC), and triglycerides (TG); and (iv) Cardiovascular traits: coronary artery disease (CAD), diastolic (DBP), and systolic blood pressure (SBP). To test the association of miRNA-related SNPs with cardiometabolic traits, we used publicly available GWAS summary statistics. A description of GWAS meta-analysis data and corresponding consortia used in this study is provided in **Supplementary Table S1**. To obtain the number of independent SNPs, we used the linkage disequilibrium (LD) based

SNP pruning in PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/), in which we excluded the SNPs with $R^2 > 0.7$. Bonferroni correction was used to adjust for multiple testing based on the number of independent SNPs available in the GWAS data (HapMap or 1000G project imputed data).

Prioritization of miRNA-related SNPs associated with cardiometabolic traits

For miRNA-related SNPs significantly associated with cardiometabolic traits, we performed *in silico* analysis to prioritize the SNPs that are more likely to be functional in their corresponding loci based on the following criteria: (i) association between the miRNA-related SNP and the cardiometabolic trait, (ii) association between the miRNA-related SNP and the expression level of miRNA/miRNA hosting gene, and (iii) expression of the miRNA in tissues relevant to cardiometabolic traits. In this regard, regional association plots were generated (using LocusZoom web tool, Version 1.1) to visualize the physical position and evaluate the association of the cardiometabolic traits with the miRNA-related SNP and its proxy SNPs ($R^2 > 0.8$) in the corresponding locus: (i) To explore whether the SNP is associated with the expression of related miRNA or miRNA hosting genes in relevant tissues (e.g., adipose tissue, liver, pancreas, muscle and blood), we used eQTL data from GTEx Portal (), (ii) We used two online databases; miRmine and Human miRNA tissue atlas (26, 27) to test where a miRNA is expressed in tissues relevant to cardiometabolic traits (e.g., adipose tissue, liver, pancreas, muscle, and blood), (iii) The Vienna RNAfold algorithm was used to check miRNA secondary structure and free energy changes with wild-type and mutant alleles of SNPs located in miRNA sequences (28).

Determination of methylation quantitative trait loci (me-QTLs)

To determine if the identified SNPs have an effect on the methylation levels of CpG sites (me-QTLs), we used data of a recent me-QTL study performed in five cohorts, including the RS, with a total of 3,841 individuals (29). We incorporated both *cis*-me-QTLs and *trans*-me-QTLs. Where *cis*-me-QTLs were defined as the effect of SNPs on the methylation levels of a CpG sites no further than 250kb apart, *trans*-me-QTLs were defined as the effect of distal SNPs on the CpG methylation levels. Details on the me-QTL mapping are described elsewhere (29). We tested if the cardiometabolic-associated SNPs found in the current study were identified as me-QTLs.

DNA methylation analysis in the Rotterdam Study

The Rotterdam Study (RS) is a large prospective population-based cohort study conducted among middle-aged and elderly people in the suburb Ommoord in Rotterdam, the Netherlands. In 1989, 7,983 inhabitants aged 55 and older were recruited in the first cohort (RS-I) (78% of 10,215 invitees). In 2000, the RS was extended with a second cohort

of 3,011 participants that moved to Ommoord or turned 55 years old (RS-II). In 2006, the third cohort (RS-III) was initiated in which inhabitants aged 45-54 years were invited and included 3,932 participants. A detailed description of RS can be found elsewhere (30). In the current study, we used DNA methylation data from a random subset (n = 717) of the third visit of RS-II (RS-II-3) and second visit of RS-III (RS-III-2) and a random subset (n = 721) of the first visit of RS-III (RS-III-1). There was no overlap in participants. The RS has been approved by the institutional review board (Medical Ethics Committee) of the Erasmus Medical Center and by the review board of The Netherlands Ministry of Health, Welfare and Sports. All participants gave written consent before participation in the study. Participant characteristics are presented in **Supplementary Table S2**.

DNA was extracted from whole peripheral blood using standardized salting out methods, of which 500ng was bisulfite treated using the Zymo EZ-96 DNA methylation kit (Zymo Research, Irvine, CA, USA). Bisulfite converted DNA was hybridized to the Illumina Human 450K array (Illumina, San Diego, CA, USA), according to manufacturer's protocol. Data preprocessing was performed using an R programming pipeline based on the pipeline developed by Touleimat and Tost (31). The genome coordinates provided by Illumina (GRCh37/hg19) were used to identify independent loci. We extracted 12,939 unique CpGs located in +/-2kb of the pre-miRNA sequences using the Illumina450K array annotation file as provided by Illumina (32). Among these, 12,617 CpGs were located in the regulatory region of 1,269 miRNAs and 450 CpGs were located in the pre- and mature sequence of 391 miRNAs. We tested the association of these CpGs with different cardiometabolic traits using linear mixed models. Data collection on these traits in the RS is described in **Supplementary Methods**. The models were adjusted for age, gender, current smoking, blood cell counts (monocytes, granulocytes, lymphocytes) as fixed effects and technical covariates as random effects. Models were further adjusted for covariates per group as follows: (i) for Anthropometric traits we adjusted WC and WHR for BMI, (ii) for Glycemic traits we adjusted for BMI and diabetic medication, (iii) for Lipid traits we adjusted for BMI and lipid medication, and (iv) for Cardiovascular traits we adjusted for BMI, blood pressure lowering medication and lipid medication. A candidate-based approach was used to sought overlap between identified miRNAs. A nominal p-value of <0.05 was found to be significant.

Determination of miR-eQTMs

To identify association between the methylation level of CpGs and the expression of miRNAs (miR-eQTMs), we used miR-eQTM data from a recent study (13). The latter study analyzed associations of expression levels of 283 miRNAs with methylation of CpGs from 3,565 individuals, in which they identified 227 miR-eQTMs at FDR < 0.01. We tested if any of the cardiometabolic-associated CpGs in the current study was among the identified miR-eQTM (13).

MiRNA expression profiling in the Rotterdam Study

We performed miRNA expression analysis in 2,000 RS participants, including a random subset (n = 1,000) of the fourth visit of RS-I (RS-I-4) and a random subset (n = 1,000) of the second visit of RS-II (RS-II-2). Plasma miRNA levels were determined using the HTG EdgeSeq miRNA Whole Transcriptome Assay (WTA), which measures the expression of 2,083 mature human miRNAs (HTG Molecular Diagnostics, Tuscon, AZ, USA) and using the Illumina NextSeq 500 sequencer (Illumina, San Diego, CA, USA). The WTA characterizes miRNA expression patterns, and measures the expression of 13 housekeeping genes, that allows flexibility in data normalization and analysis. Quantification of miRNA expression was based on counts per million (CPM). Log2 transformation of CPM was used as standardization and adjustment for total reads within each sample. MiRNAs with Log2 CPM < 1.0 were indicated as not expressed in the samples. The lower limit of quantification (LLOQ) was used to select well-expressed miRNAs. The LLOQ level was based on a monotonic decreasing spline curve fit between the means and standard deviations of all miRNAs. In our definition well-expressed miRNA levels in plasma were those with >50% values above LLOQ. Out of the 2,083 measured miRNAs, 591 miRNAs were expressed at good levels in plasma.

The miRNAs significantly associated with cardiometabolic traits, in the genetic association studies, were tested for the association of their plasma expression levels with the same cardiometabolic trait(s). Linear models were used to test the association between available continuous traits in the RS (incl. BMI, WC, WHR, FG, HDL, TC, SBD, and DBP) and miRNA expression. Additionally, we used binomial models to test the association between disease prevalence (incl. T2D and CHD) and miRNA expression. We used the cardiometabolic traits as dependent variable and plasma miRNAs level as explanatory variable, adjusting for age, gender and current smoking. Models were further adjusted for covariates per group as follows: (i) for Anthropometric traits we adjusted WC and WHR for BMI, (ii) for Glycemic traits we adjusted for BMI and diabetic medication, (iii) for Lipid traits we adjusted for BMI and lipid medication, and (iv) for Cardiovascular traits we adjusted for BMI, blood pressure lowering medication and lipid medication. A candidate-based approach was used to sought overlap between identified miRNAs. A nominal p-value of <0.05 was found to be significant.

In addition, we extracted strongly validated target genes, defined as being validated by western blot and/or luciferase reporter assay, of the identified miRNAs from the miRTarBase database (33). Next, we extracted SNPs in these target genes and tested their associations with cardiometabolic traits using summary statistics of previously mentioned GWAS data.

RESULTS

Association of miRNA-SNPs with cardiometabolic traits and diseases

Out of 23,990 miRNA-related SNPs, 2,358 independent SNPs were present in the GWAS data based on HapMap and 8,652 independent SNPs were present in the 1000G project. Bonferroni correction was used to set the significance threshold, at p-value $<2.12\times10^{-5}$ (0.05/2,358) for GWAS with HapMap imputed data and p-value $< 5.78\times10^{-6}$ (0.05/8,652) for GWAS with 1000 Genomes project imputed data. Of these, 180 SNPs annotated to 67 miRNAs passed the significance threshold to be associated with at least one cardiometabolic trait (**Table 1**). Out of the 180 identified SNPs, 89 SNPs were located in +/-2kb of 57 primary miRNA transcripts (**Supplementary Table S3**) and 92 SNPs were among the previously reported miR-eQTLs of 15 mature miRNAs (**Supplementary Table S4**). Manhattan plots illustrated in **Figure 2** present the miRNA-annotated genetic variants associated with lipid traits and the prevalence of T2D and CHD. **Table 2** shows the top miRNA-related SNPs associated with cardiometabolic traits, which were annotated to 20 miRNAs.

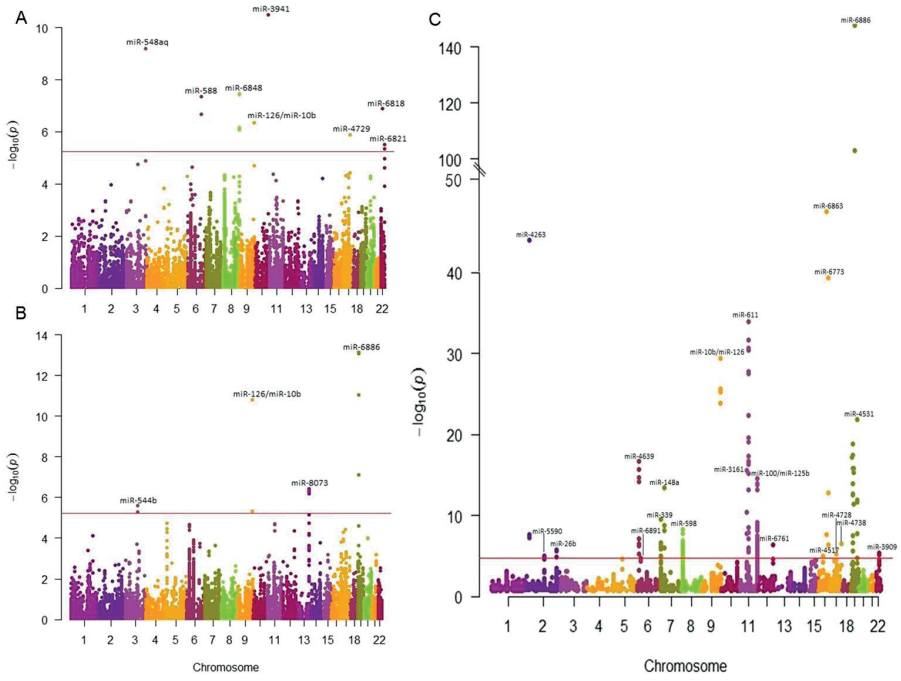


Figure 2. Manhattan plots showing the association of miRNA-SNPs with T2D, CAD, and lipid traits. The association miRNA-related SNPs and cardiometabolic traits were examined using the publicly available GWAS data. We reported the most significantly associated miRNA of each SNP loci. The horizontal red line indicates the study significance threshold. (A) Manhattan plot showing the association of miRNA-SNPs with T2D in which 12 SNPs in 8 miRNAs passed the significant threshold. (B) Manhattan plot showing the association of miRNA-SNPs with CAD in which 13 SNPs in 9 miRNAs passed the significance threshold. (C) Manhattan plot showing the association of miRNA-SNPs with lipid traits in which 107 SNPs in 36 miRNAs passed the significant threshold. When SNPs were present in more traits, the most associated SNP was plotted.

Table 1. Description of genome-wide association studies (GWAS) of cardiometabolic traits and associated miRNA single-nucleotide polymorphisms (SNPs).

Phenotype	Consortium	SNPs in +/- 2kb miR*	SNPs in miR-seq*	SNPs in miR-QTL*	Associated miR loci†
Anthropometric traits					
Body-mass index	GIANT(34)	9	0	9	7
Waist to hip ratio	GIANT(35)	2	1	1	4
Waist circumference	GIANT(35)	10	0	1	8
Glycemic traits					
Glucose fasting	MAGIC(36)	3	0	1	4
Glucose after 2h	MAGIC(37)	0	0	0	0
Insulin fasting	MAGIC(36)	1	0	2	2
Proinsulin	MAGIC(38)	3	0	4	3
HbA1c	MAGIC(39)	1	0	15	4
HOMA-IR	MAGIC(40)	0	0	0	0
HOMA-β	MAGIC(40)	0	0	0	0
Type 2 diabetes	DIAGRAM(41)	12	0	1	8
Lipid traits					
LDL	GLGC(42)	22	1	20	11
HDL	GLGC(42)	12	1	23	9
Total cholesterol	GLGC(42)	26	1	40	13
Triglyceride	GLGC(42)	8	1	27	7
Cardiovascular traits					
CAD	CARDIoGRM plusC4D(43)	10	0	2	4
DBP	ICBP(44)	3	0	-	2
SBP	ICBP(44)	2	0	-	2

Shown are SNPs located within +/-2kb of primary miRNA transcripts, pre- and mature miRNA sequences, miRNA-eQTL SNPs).

* Number of SNPs that passed the significance threshold (p-value $<2.12x10^{-5}$ for SNPs imputed with HapMap and p-value $<5.78x10^{-6}$ for SNPs imputed with 1000G)

† Number of independent loci.

In order to prioritize miRNA-related SNPs based on potential functionality in relation to the associated cardiometabolic traits, we created regional association plots to visualize the LD of miRNA SNP with the top SNP in the corresponding locus (**Figure 3**). We found three top SNPs in their loci, including rs7117842 associated with TC ($p = 2.48\times10^{-15}$, $\beta = 0.029$) and located ~512kb upstream of miR-100-5p/miR-125b-5p (**Figure 3A**), rs1997243 associated with TC ($p = 2.72\times10^{-10}$, $\beta = 0.033$) and located ~21kB upstream of miR-339-3p (**Figure 3B**), and rs7607369 associated with BMI ($p = 1.10\times10^{-7}$, $\beta = -0.016$) and located ~11.7kb upstream of miR-26b-5p (**Figure 3C**). These three SNPs were previously identified as miR-eQTLs that change the expression levels of related miRNAs in blood (21). In addition, rs4722551 located ~2kb upstream of miR-148a shows the strongest association with LDL ($p = 3.95\times10^{-14}$, $\beta = 0.039$) on the Chr7p15.2 locus (**Figure 3D**).

Table 2. The top 20 miRNAs with single-nucleotide polymorphisms (SNPs) in related regions association with cardiometabolic traits.

miRNA	SNPID	Chr.	Position	Alleles (A/R)	Annotated gene	Associated trait	Effect	P value
miR-6886†	rs17248720	19	11198187	C/T	*LDLR*	LDL	0.226	$2.40x10^{-148}$
miR-6863†	rs13306673	16	56900931	C/T	*SLC12A3*	HDL	0.098	$2.76x10^{-48}$
miR-4263†	rs2305929	2	28113911	G/A	*BRE*	TG	0.064	$1.13x10^{-44}$
miR-6773†	rs8057119	16	68268836	T/C	*ESRP2*	HDL	0.072	$5.21x10^{-40}$
miR-611†	rs174538	11	61560081	G/A	*THEM258*	LDL	0.050	$1.07x10^{-34}$
miR-1908-5p‡	rs174548	11	61571348	C/G	*FADS1*	LDL	0.047	$2.29x10^{-31}$
miR-10b-5p/126-5p‡	rs532436	9	136149830	A/G	*ABO*	LDL	0.079	$4.02x10^{-30}$
miR-4721†	rs4788099	16	28763228	G/A	*TUMF*	BMI	0.031	$1.09x10^{-24}$
miR-4531†	rs6509170	19	45159636	C/A	*LOC107985305*	LDL	0.127	$1.54x10^{-22}$
miR-199a-1†	rs11085748	19	10927540	T/C	*DNM2*	LDL	0.055	$1.46x10^{-19}$
miR-4999†	rs7254882	19	8359822	C/T	*MIR4999*	HDL	0.033	$6.66x10^{-18}$
miR-4639†	rs3757354	6	16127407	C/T	*MYLIP*	LDL	0.038	$2.09x10^{-17}$
miR-640†	rs1000237	19	19518316	T/A	*GATAD2A*	TG	0.033	$1.61x10^{-16}$
miR-3161†	rs79837139	11	48000780	C/T	*PTPRJ*	HDL	0.062	$2.99x10^{-16}$
miR-100-5p/125b-5p‡	rs7117842	11	122663796	C/T	*UBASH3B*	TC	0.029	$2.48x10^{-15}$
miR-148a†	rs4722551	7	25991826	C/T	*MIR148A*	LDL	0.039	$3.95x10^{-14}$
miR-139†	rs11605042	11	72700619	A/G	*ARAP1*	Pro-Ins	-0.053	$5.24x10^{-13}$
miR-3941†	rs71486610	10	124134803	C/G	*PLEKHA1*	T2D	-0.081	$3.30x10^{-11}$
miR-6745†	rs901750	11	47209472	A/G	*PACSIN3*	HDL	0.024	$3.95x10^{-11}$
miR-196a-2-3p*	rs11614913	12	53991815	C/T	*MIR196A2*	WHR	0.029	$6.90x10^{-11}$

* SNP located in pre- and mature miRNA sequence
† SNP located within +/- 2kb of primary miRNA transcript
‡ miR-eQTL SNPs

Moreover, rs174561 has previously been reported by (22) to change the expression of miR-1908-5p. We found this SNP, located in the coding sequence of miR-1908-5p, to be associated with lipid traits (LDL, HDL, TC, and TG), and rs11614913, located in the coding sequence of miR-196a2-3p, to be associated with WHR. These two variants have previously been reported to be associated with lipid traits and WHR and have been suggested to change the miRNA structure and expression (45). We also found a suggestive association between rs58834075, located in the pre-miR-656 sequence (T > C, Chr14:101066756) and T2D ($p = 6.30\times10^{-5}$, $\beta = -0.170$). The miRNA secondary structure and free energy changes of both wild-type and mutant alleles of these three SNPs (rs174561, rs11614913 and rs58834075) are illustrated in **Supplementary Figure S1**.

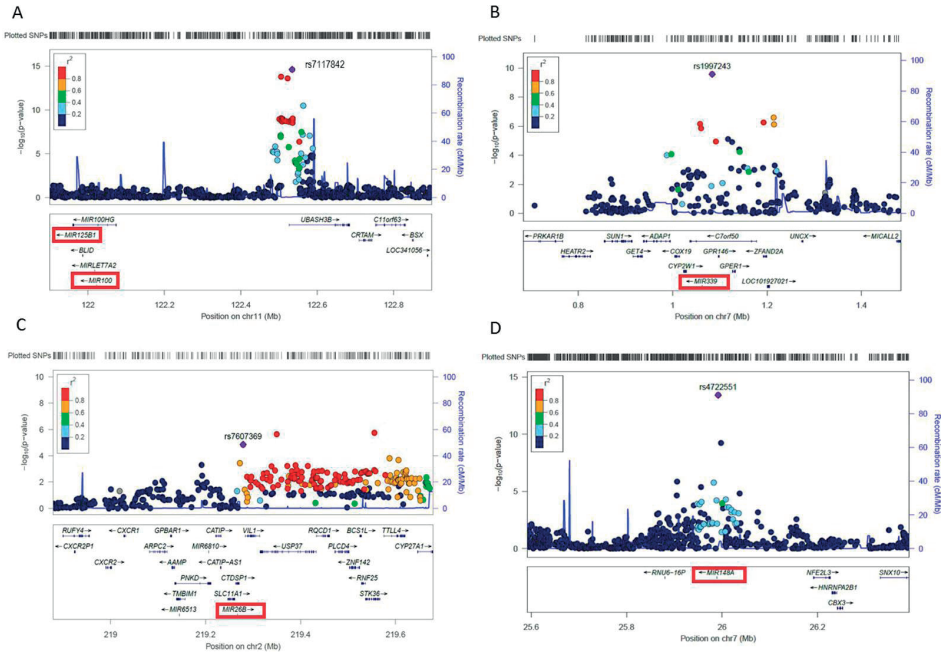


Figure 3. Regional plots showing the association of four top miRNA-SNPs with cardiometabolic traits. The most significant SNP in the region, according to P-value, is represented by a purple diamond, and the degree of linkage disequilibrium of other SNPs in the region to the lead SNP is representative by the color scale shown in the legend. Genes are illustrated below. The associated miRNA is illustrated with a red box. (A) Regional plot showing the association of rs7117842 located ~512kb upstream of miR-100-5p/125b-5p with TC, LDL and HDL on the Chr11q24.1 locus. (B) Regional plot showing the association of rs1997243 located ~21kb upstream of the primary transcript of miR-339-3p with TC and HDL on the Chr7p22.3 locus. (C) Regional plot showing the association of rs7607369 located ~11.7kb upstream of the primary transcript of miR-26b-5p with BMI and TG on the Chr2q35 locus. (D) Regional plot showing the association of rs472551 located ~2kb upstream of the primary transcript of miR-148a with LDL, TG, and TC on the Chr7p15.2 locus.

Identification of methylation quantitative trait loci (me-QTLs)

We identified 29 *cis*-me-QTL effects for 47 independent CpGs at FDR < 0.05 (49 SNP-CpG pairs). Among these, we found 14 *cis*-me-QTLs that were associated with both the expression level of 8 miRNAs and the methylation level of 26 CpGs (**Supplementary Table S5**). In total there were 7 *cis*-me-QTLs (for 8 CpGs) that were associated with a cardiometabolic trait in the current study (**Table 3**). Furthermore, 4 *trans*-me-QTL effects for 21 independent CpGs were found at FDR < 0.05 (27 SNP-CpG pairs) (**Supplementary Table S5**). Two out of the four *trans*-me-QTL were miR-eQTL SNPs (rs174548 for miR-1908-5p and rs1997243 for miR-339-3p). None of the associated CpGs *in trans* were found in the current study to be associated with cardiometabolic traits.

Table 3. Identified me-QTLs with cardiometabolic-associated CpGs.

miRNA	SNPID	CpG	Cis†/ Trans	miR-eQTL SNP*	SNP associated with cardiometabolic trait	CpG associated with cardiometabolic trait
miR-611	rs174538	cg16150798	Cis	-	FG, LDL, HDL, TG, TC	WC
miR-588	rs9388486	cg20229609	Cis	-	T2D	SBP, DBP
miR-1908-5p	rs174548	cg03921599	Cis	√	FG, HbA1c, LDL, HDL, TG, TC	LDL, TC
miR-199a-1	rs3786719	cg02907064	Cis	-	LDL, TC	LDL
miR-6745	rs901750	cg00724111	Cis	-	HDL	FI, SBP, DBP
miR-8073	rs3809346	cg22382805	Cis	-	CAD	FI
miR-653, miR-489	rs2528521	cg06934092	Cis	-	BMI	FG, FI, TC
miR-8073	rs3809346	cg19700260	Cis	-	CAD	DBP

Shown are 7 me-QTLs associated with methylation levels of 8 cardiometabolic-associated CpG sites.
* miR-eQTL SNP is associated to change the expression of miRNA level
† SNP and CpG are located not further away than 250kb

Testing DNA methylation and expression of miRNAs associated with cardiometabolic traits

To access the relationship between miRNAs and cardiometabolic traits in other omics layers, we performed a candidate-based test to check whether the 67 identified miRNAs, with SNPs associated with cardiometabolic traits, show also association between DNA methylation and miRNA expression with cardiometabolic traits. Using DNA methylation data from 1,438 RS participants, we found 278 CpG sites annotated to 64 out of the 67 miRNAs, to be associated with any cardiometabolic trait (**Supplementary Table S6**). By integrating our DNA methylation results with the GWAS data, we observed an overlap of 38 miRNAs (79 CpGs) that had both a SNP and a CpG associated with the same trait (**Supplementary Table S7**). The CpG site showing the most significant association was cg15616915 which is located in the regulatory region of miR-26b and is positively associated with TG ($p = 1.59\times10^{-4}$, $\beta = 0.009$). We found 16 cardiometabolic-associated CpGs that are annotated to more than one miRNA. For example, cg03722243 associated with BMI ($p = 1.55\times10^{-3}$, $\beta = 0.001$) is annotated to miR-489 and miR-653, which are clustered on chromosome 7. In addition, cg15334028 associated with WC, HDL, LDL, and TG is annotated to three miRNAs (miR-638, miR-6793, and miR-4748) on chromosome 19.

We identified two CpGs that are associated with the expression level of miRNAs (miR-eQTM) at FDR < 0.01. The most significant cis-miR-eQTM, cg26363555 has been reported to be negatively associated with both miR-125b-5p (~2kb downstream) and miR-100-5p (~50kb upstream) expression levels (13). The CpG cg26363555 was positively associated with FG ($\beta = 0.012$) and DBP ($\beta = 2.00\times10^{-4}$) and negatively associated with HDL ($\beta = -0.004$) in the RS. In addition, cg03891346 has been reported to be negatively associated with the expression level of miR-100-5p (~53kb downstream) (13). This CpG, which is also annotated to MIR125B1, was positively associated with WC ($\beta = 5.00\times10^{-4}$) in the RS.

Next, we tested whether the 67 identified miRNAs show differential expression in plasma in relation to the associated cardiometabolic trait(s). Of the 67 miRNAs, we could only test the association of 28 mature miRNAs that were well-expressed in plasma and of which the phenotype of interest was available in the RS. Of these, plasma levels of 22 miRNAs were nominally associated with at least one cardiometabolic traits (**Supplementary Table S8**). Furthermore, out of the 67 miRNAs, we found 12 differently expressed mature miRNAs to be associated with the same trait (**Table 4**). Plasma levels of miR-126-3p, miR-126-5p, miR-10b-5p, miR-148a-3p, miR-199a-1-3p, miR-199a-1-5p, miR-125b-5p, and miR-100-5p were positively associated with serum TC levels. In contrast, miR-6886 was negatively associated with serum TC levels. A negative association between miR-126-5p and miR-126-3p and CHD was found. Furthermore, we observed a negative association between miR-4681 levels and WC. An overview of the number of associated miRNAs using different omics data is illustrated in **Figure 4**.

Table 4. Plasma expression levels of miRNAs associated with cardiometabolic traits.

miRNA	Effect	P value	Associated trait
miR-126-3p	0.379	$1.09x10^{-14}$	TC†
miR-10b-5p	0.352	$3.30x10^{-11}$	TC†
miR-126-5p	0.258	$3.75x10^{-11}$	TC†
miR-148a-3p	0.189	$8.01x10^{-06}$	TC†
miR-199a-1-3p	0.171	$3.38x10^{-05}$	TC†
miR-125b-5p	0.159	$2.43x10^{-03}$	TC†
miR-100-5p	0.141	$3.15x10^{-03}$	TC†
miR-6886-3p	-0.083	$9.49x10^{-03}$	TC†
miR-126-5p	-0.365	$1.24x10^{-02}$	CHD‡
miR-4681	-0.440	$2.13x10^{-02}$	WC*
miR-199a-1-5p	0.074	$3.38x10^{-02}$	TC†
miR-126-3p	-0.385	$3.54x10^{-02}$	CHD‡

Model 1: adjusted for: age, gender, current smoking
* Model 1 + BMI
† Model 1 + BMI, lipid medication
‡ Model 1 + BMI, blood pressure lowering medication, lipid medication

Furthermore, out of 22 miRNAs that were associated with at least one cardiometabolic trait, we found validated target genes for 14 miRNAs. We tested the association between these target genes and cardiometabolic traits using summary statistics GWAS data. After correcting for multiple testing, based on the number of tested SNPs in the target genes of a miRNA, we found 24 unique target genes for 9 of the 14 miRNAs to be associated with cardiometabolic traits (**Supplementary Table S9**).

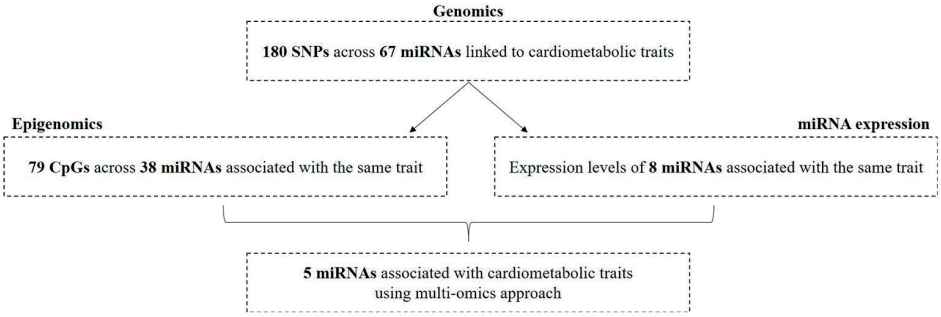


Figure 4. Overview of miRNAs associated with cardiometabolic traits by integrating three omics layers.

Finally, we sought overlapping miRNAs that were associated with the same cardiometabolic trait in the three different omics analyses (**Supplementary Table S10**). Since not all related phenotypes were available within the RS and not all miRNAs were expressed, we tested 64 miRNAs that had DNA methylation sites and 22 mature miRNAs that were available for miRNA expression analyses using the RS. We found five miRNAs, including miR-10b-5p, miR-148a-3p, miR-100-5p, miR-125b-5p, and miR-6886 that had at least one CpG and of which the expression was also associated with the same cardiometabolic trait. After prioritization based on the suggested criteria for potential functionality, miR-10b-5p, miR-148a-3p, miR-125b-5p, and miR-100-5p were highlighted as the most likely miRNAs involved in the pathogenesis of risk factors for T2D and CHD (**Table 5**).

DISCUSSION

In this study, we integrated different population-based omics data (including genetics, epigenetics and miRNA expression) to identify miRNAs associated with cardiometabolic traits. Genetic variants related to 67 miRNAs were associated with the studied traits. Alterations in DNA methylation of CpG sites annotated to 38 of these miRNAs and plasma expression levels of 8 of them were also associated with the same trait. In principle, the association between a miRNA and trait of interest in more than two layers of omics may strengthen its potential to play a role in the disease underlying mechanisms. In this context, we sought to identify overlap between miRNAs that were associated with the same cardiometabolic trait across different approaches. This integration analysis revealed the correlation between four miRNAs (miR-10b-5p, miR-148a-3p, miR-125b-5p, and miR-100-5p) and lipid traits.

MiR-10b-5p is a highly conserved miRNA across multiple species and is located inside the homeobox D cluster on chromosome 2. A recent study showed a mediating role for miR-10b between obesity and primary breast cancer (46). Moreover, previous research in mice found a negative regulatory role of miR-10b on cholesterol efflux via targeting

Table 5. Cardiometabolic-associated miRNAs after integrative multi-omics data and in silico prioritization analysis.

miRNA	Traits	SNPID	P value	Proxy SNPs (Non-syn.)	GWAS	miRNA CpG	P value	Trait	miR-eQTM	Trait	miRNA Exp.	Trait
miR-10b-5p*	LDL†, TC, CAD, T2D	rs532436(‡)	$4.02x10^{-30}$	9 (0)	√	cg25820279	$4.37x10^{-2}$	TC	-	-	$3.30x10^{-11}$	TC
miR-148a-3p*	LDL†, TC, TG	rs4722551	$3.95x10^{-14}$	1 (0)	√	cg18188200	$1.94x10^{-3}$	TC	-	-	$8.01x10^{-6}$	TC
miR-125b-5p*	TC†, LDL, HDL	rs7117842(‡)	$2.48x10^{-15}$	52 (0)	√	cg06749053	$2.39x10^{-2}$	LDL	cg26363555	HDL	$2.43x10^{-3}$	TC
miR-100-5p*	TC†, LDL, HDL	rs7117842(‡)	$2.48x10^{-15}$	52 (0)	√	cg14724899	$4.02x10^{-2}$	HDL	-	-	$3.15x10^{-3}$	TC
miR-6886-3p	LDL†, TC, CAD	rs17248720	$2.40x10^{-148}$	49 (0)	-	cg19751789	$8.03x10^{-4}$	TC	-	-	$9.49x10^{-3}$	TC

* MiRNAs which were associated with multi-omics approach and fulfill most criteria to play likely a functional role in the progression or development of cardiometabolic traits.† Trait to which the SNP is associated. ‡ miR-eQTL SNP

Shown here are 5 miRNAs which (partly) fulfill criteria for being potentially functional in their loci. Proxy: Number of proxy SNPs (R^2>0.8) in strong linkage disequilibrium (LD) with the given variant. Non-synonymous: Number of non-synonymous variants that are in high LD with the given variant. GWAS: (√ SNP represented to be top SNP with the strongest association with the related trait on the certain genomic position). miR-eQTM: CpG is associated with expression level of miRNA.

the ATP binding cassette transporter gene (ABCA1) (47). MiR-10b has been also shown to be involved in the progression of atherosclerosis, which is a major cause of cardiovascular disease (48). We found a genetic variant (rs532436;A > G) annotated to the Alpha 1-3-N-acetylgalactosaminyltransferase (*ABO*) gene to be positively associated with LDL, TC, CAD, and T2D. The *ABO* gene has been linked to cholesterol absorption and cardiovascular disease (49). Rs532436, located on chromosome 9, has been reported as *trans*-miR-eQTL for miR-10b-5p (22). In this study, we further showed that a CpG site (cg25820279) annotated to Homeobox D3 (*HOXD3*), is located in the regulatory region of miR-10b and is associated with total cholesterol levels in serum. In addition, the expression level of miR-10b-5p in plasma showed a positive association with total cholesterol levels, which further support the crucial role of miR-10b-5p in lipid metabolism.

MiR-148a-3p has been shown to control the LDL uptake and cholesterol efflux through affecting the expression of low-density lipoprotein receptor (LDLR) (50). Moreover, *in vivo* studies in mouse models have confirmed that miR-148a-3p is upregulated in adipogenesis and highly expressed in liver tissue (51). We found rs4722551, located ~2kb upstream of miR-148a, associated with LDL, TC and TG. It has been suggested previously that a large part of regulatory elements such as promoter regions are located within +/-2kb of pre-miRNAs (17). Rs4722551 has previously been reported to be positively associated with serum lipid levels *via cis*-miR-eQTL in liver tissue (52)). Our findings may shed light on the mechanism that associates the rs4722551 risk allele (T >C) with an increased miR-148a-3p expression, which is subsequently associated with higher serum cholesterol levels. Furthermore, our results showed a CpG site (cg18188200) in the regulatory region of miR-148a to be associated with LDL, TC, and TG and demonstrated that the plasma expression level of miR-148a-3p is also associated with total serum cholesterol levels. These data are in line with the findings from previous studies reporting a functional role for miR-148a-3p in lipid metabolism confirmed by various *in vivo* and *in vitro* validation experiments (50, 52).

We found strong associations of rs7117842, located ~512kb upstream of miR-100-5p/125b-5p, with TC, LDL, and HDL, suggesting these two miRNAs to play a role in lipid metabolism. The SNP has been previously shown to be negatively associated with the expression levels of miR-100-5p and miR-125b-5p in blood (21). In our analysis, plasma expression levels of miR-100-5p and miR-125b-5p are positively associated with TC. These findings could be interpreted in a way that carrying the risk allele of rs7117842 (T > C) is associated with decreased expression of miR-100-5p/125b-5p, which is associated with a reduced increase of total serum cholesterol levels. In addition, cg26363555, located in the promoter region of miR-125b-5p, was previously reported to act as miR-eQTM by changing the expression levels of both miR-100-5p and miR-125b-5p (13). We found cg26363555 associated with HDL in the RS. In addition, cg03891346 annotated to MIR125B1 was reported to be associated with the expression level of miR-100-5p (13).

Our DNA methylation analysis results showed the association between cg03891346 and waist circumference in the RS. Our findings are partly in line with previous research investigating the role of miR-125b-5p on adipogenesis where it is observed that miR-125b-5p downregulates the anti-adipogenic gene *MMP11* in human, indicating that miR-125b-5p *via MMP11* positively regulate adipogenesis (53) . Conversely, the same study demonstrated a direct effect of reduction in fat accumulation through overexpression of miR-125b-5p (53). In addition to the role of miR-125b-5p on lipid metabolism in human, its regulatory role has been investigated in other organisms including zebrafish and mice. Over-expression of miR-125b in zebrafish is linked to lipid metabolism in brain, heart and liver tissue (54). This study observed that overexpression of miR-125b inhibits osteoblastic differentiation and promotes fat synthesis. Moreover, the expression of miR-125b is activated by estrogen via ERα *in vitro* and *in vivo* in mice, in which they demonstrated that miR-125b can limit fat accumulation in liver tissue (55). These contradictory findings may implicate that miR-125b-5p plays an important role in lipid metabolism via a complex molecular cascade. However, the role of miR-100-5p in regard to lipid metabolism and cardiovascular disease yet to be further investigated. Since miR-100-5p and miR-125b-5p are located in the same locus on chromosome 11, it could be possible that miR-125b-5p is the driving miRNA in relation to the associated lipid traits. Future research is warranted to confirm the regulatory role of miR-100-5p in lipid metabolism.

The main strengths of this study include the use of robust data from the large-scale GWAS studies and multi-omics implementation of a large sample size in the Rotterdam Study, which indicates with more confidence that miRNAs are involved in the pathophysiology of cardiometabolic diseases. Our study, however, does not come without limitations. First, our study design is based on associations rather than causations, therefore this approach does not prove that the identified miRNAs play a causal role in the studied traits. To test for causal inferences between miRNAs and disease risk, future studies should test mediating effects and incorporate functional follow-up experiments. Furthermore, our study design was based on a cross-sectional approach, which means that individuals included in this study were not free of CHD or T2D. In regard to test whether the identified miRNAs are associated with the risk of developing disease, future longitudinal studies are warranted. Another limited factor is that we were unable to link all identified miRNAs with epigenetic and expression analyses in the RS, since not all phenotypic data were available for each trait of interest nor were all miRNAs well-expressed in plasma. In addition, different sub cohorts of the RS were used for DNA methylation and miRNA expression analysis due to the availability of data. DNA methylation and miRNA signatures are dynamic over time and could have yield in confounding results. The challenge of this multi-omics approach includes the intra-individual variation and thereby lack of generalizability between datasets. However, the sub cohorts of RS-II and

RS-III are extensions of the RS-I cohort. Previous epigenetic (DNA methylation) studies using the RS data showed that the results are replicated after additional adjustment for sub cohort (56-58). This may indicate that the intra-individual differences between variables in these RS sub cohorts have not significantly affected by exposing to different environmental factors. Yet in an optimal setting one should apply the multi-omics analysis in the same individuals and the same timeframe. Furthermore, we used whole blood to determine DNA methylation and plasma to check expression levels of miRNAs, which are not the most relevant tissue for cardiometabolic traits. This could have resulted in overlooking some of the miRNAs, but the found associations are comparable because both analyses were performed in the same tissue. In an optimal setting one should examine the observed associations using next-generation sequencing covering all miRNAs in target tissues (e.g., adipose tissue, heart, pancreas and liver). Such infrastructure is not yet available in large epidemiologic studies with validated clinical data. However, for the use of miRNAs as targets for early diagnosis or progression of T2D and CHD, blood might be a very good test tissue since it is a non-invasive method for biomarker measurements in clinical diagnosis. In addition, regarding potential missed cardiometabolic-associated SNPs, our study could have benefited from denser genotyping methods including 1000 Genomes project or the Haplotype Reference Consortium (HRC).

CONCLUSION

In this study, we systematically examined the association of miRNAs with cardiometabolic risk factors and diseases using population-based genetic, DNA methylation and miRNA expression data. By integrating these omics data we found several cardiometabolic- associated miRNAs, such as miR-10b-5p, miR-148a-3p, miR-125b-5p, and miR-100-5p involved in lipid metabolism, that can be viewed as potential biomarkers for early diagnosis or progression of T2D and CHD. Future experimental studies are warranted to elucidate pathways underlying the link between these miRNAs and cardiometabolic risk factors such as dyslipidemia, central adiposity and elevated blood glucose levels.

SUPPLEMENTARY MATERIAL

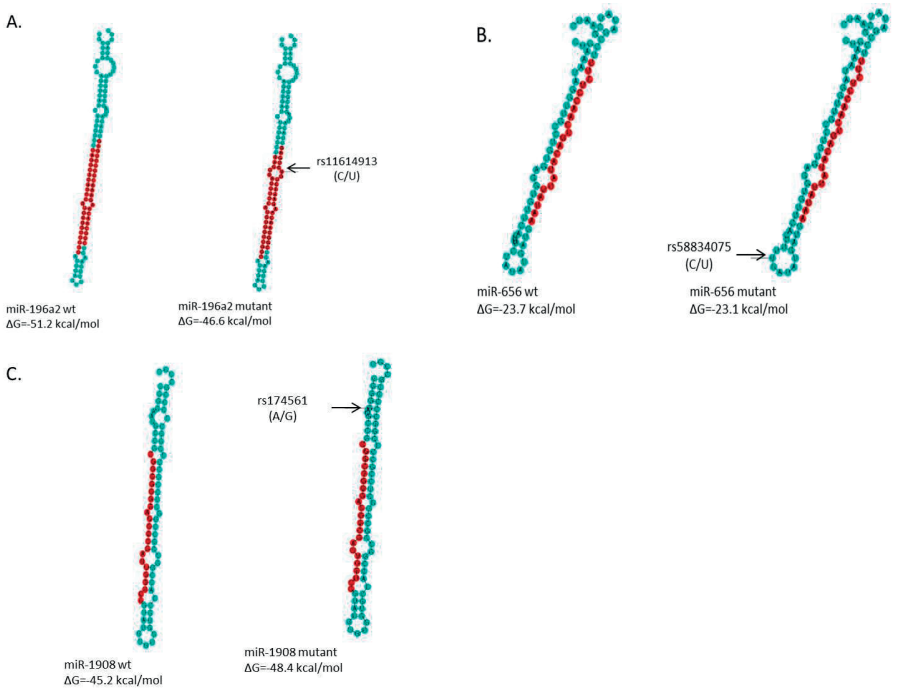
Supplementary Table 1. Description of GWAS meta-analysis and consortia used in this study Phenotype.

	Consortia	Sample size	All candidate SNPs in GWAS
Anthropometric traits			
Body-mass index	GIANT	241,258	2,549
Waist to hip ratio	GIANT	241,258	2,531
Waist circumference	GIANT	241,258	2,533
Glycemic traits			
Glucose fasting	MAGIC	133,010	2,639
Glucose after 2h	MAGIC	45,854	48
Insulin fasting	MAGIC	108,557	2,637
Pro-insulin	MAGIC	10,701	2,506
HbA1c	MAGIC	123,665	887
HOMA-IR	MAGIC	46,186	64
HOMA-β	MAGIC	46,186	37
Type 2 diabetes	DIAGRAM	26,676 cases/ 132,532 controls	10,690
Lipid traits			
Low-density lipoprotein	GLGC	173,000	2,377
High-density lipoprotein	GLGC	187,000	2,385
Total cholesterol	GLGC	187,000	2,385
Triglyceride	GLGC	178,000	2,376
Cardiovascular traits			
Coronary artery disease	CARDIoGRMplusC4D	60,801 cases/ 123,504 controls	10,706
Diastolic blood pressure	ICBP	71,255	2,345
Systolic blood pressure	ICBP	71,255	2,345

Supplementary Table 2. Participant characteristics of the Rotterdam Study for DNA methylation analysis and miRNA expression profiling DNA methylation.

	DNA methylation	DNA methylation	miRNA expression	P value*
	(RS-II-3 & RS-III-2)	(RS-III-1)	(RS-I-4 & RS-II-2)	
N	717	721	1999	
Female	413 (57.6%)	391 (54.2%)	1141 (57.1%)	<0.001
Age (years)	67.5 (5.93)	59.8 (8.16)	71.6 (7.58)	<0.001
BMI (kg/m2)	27.7 (4.12)	27.6 (4.63)	27.7 (4.11)	0.905
Waist circumference	94.4 (12.00)	93.75 (12.92)	93.6 (11.98)	0.142
WHR	0.9 (0.09)	0.9 (0.08)	0.9 (0.09)	<0.001
Current smoking (yes)	76 (10.6%)	193 (26.8%)	288 (14.4%)	<0.001
Triglycerides (mmol/L)	1.5 (0.79)	1.5 (0.88)	NA	0.298
HDL-cholesterol (mmol/L)	1.5 (0.44)	1.4 (0.41)	1.4 (0.39)	<0.001
LDL-cholesterol (mmol/L)	3.7 (0.94)	3.9 (1.00)	NA	<0.001
Total cholesterol (mmol/L)	5.5 (1.02)	5.6 (1.07)	5.6 (0.99)	0.004
Lipid lowering medication (yes)	225 (31.4%)	191 (26.5%)	450 (22.5%)	<0.001
Systolic blood pressure	144.8 (21.91)	134.2 (19.76)	148.2 (20.82)	<0.001
Diastolic blood pressure	84.4 (11.66)	82.8 (11.38)	79.6 (10.84)	<0.001
Coronary heart disease	28 (3.9%)	46 (6.4%)	214 (10.7%)	<0.001
Anti-hypertensive medication (yes)	310 (43.2%)	217 (30.1%)	880 (44.0%)	<0.001
Glucose (mmol/L)	5.7 (1.11)	5.6 (1.04)	5.8 (1.09)	0.001
Insulin (pmol/L)	82.6 (48.26)	96.0 (63.04)	NA	<0.001
Prevalence type 2 diabetes	96 (13.4%)	74 (10.3%)	278 (13.9%)	0.04
Anti-diabetic medication	59 (8.2%)	39 (5.4%)	132 (6.6%)	0.0985

Values are presented as mean ±(SD) or N (%). *Differences between groups were addressed using ANOVA in the case variables were available among three groups. Student's T-tests in the case variables were available in two groups. NA: Not Available



Supplementary Figure I. Predicted secondary structure of miRNA wild type and variant. Location of the SNP is demonstrated by an arrow. The red part shows the mature sequence and the blue part shows the rest of the pre-miR. The corresponding minimum free energy (MFE) is illustrated with the thermodynamic ensemble ΔG. A, Secondary structure of miR- 196a2-3p wildtype and variant (rs11614913) located in mature miRNA sequence. MFE changes by -4.6kcal/mol. B, Secondary structure of miR-656 wildtype and variant (rs58834075) located in precursor miRNA sequence. MFE changes by -0.6kcal/mol. C, Secondary structure of miR-1908-5p wildtype and variant (rs174561) located in precursor miRNA sequence. MFE changes by +3.2kcal/mol.

Additional supplemental material for this chapter can be found in the online version of the paper via https://www.frontiersin.org/articles/10.3389/fgene.2020.00110/full.

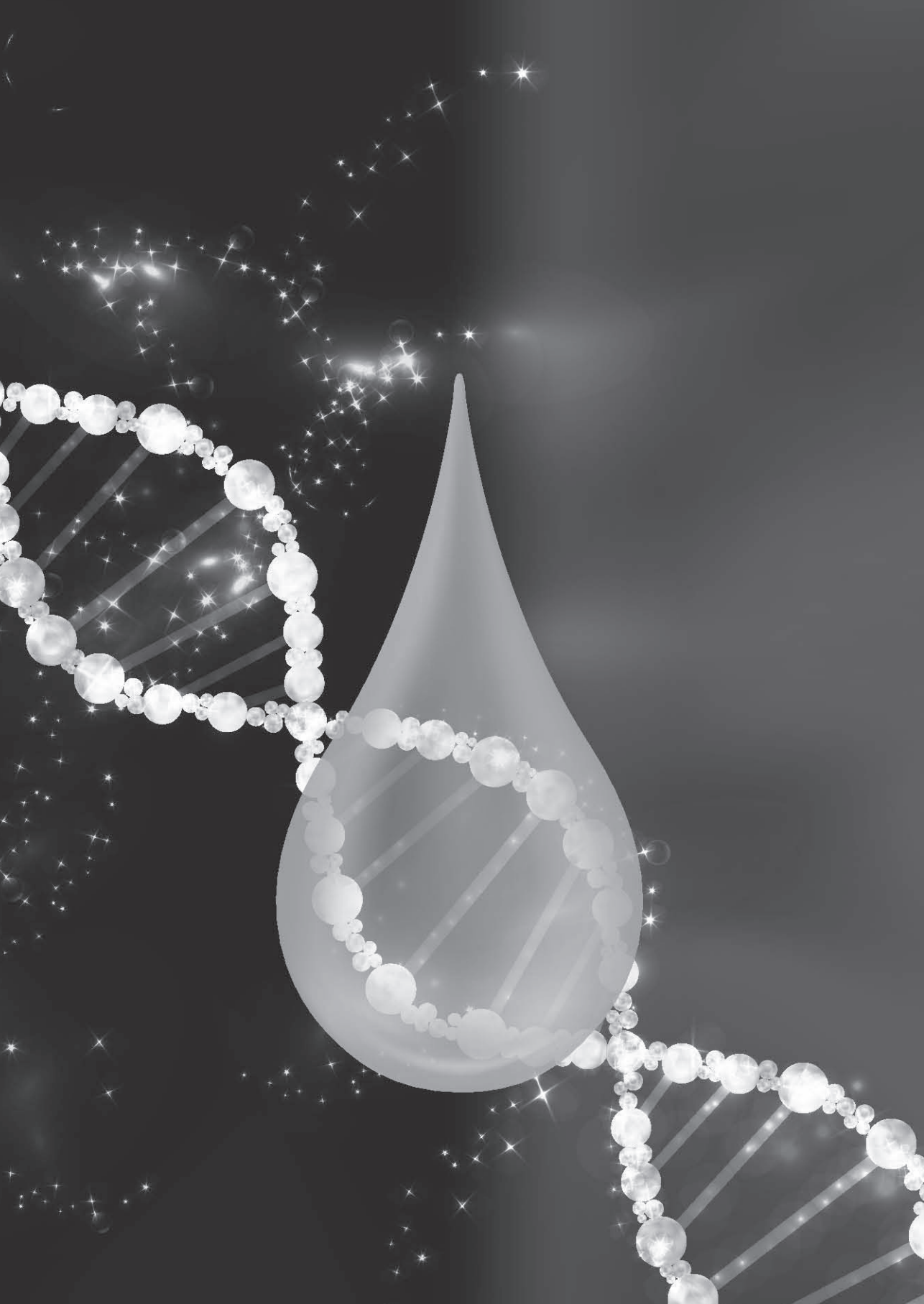
REFERENCES

1. Kannel WB, McGee DL. Diabetes and cardiovascular disease. The Framingham study. JAMA. 1979;241(19):2035-8.
2. Collaborators GBDCoD. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736-88.
3. Wilson PW, D'Agostino RB, Parise H, Sullivan L, Meigs JB. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. Circulation. 2005;112(20):3066-72.
4. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet. 2014;94(2):223-32.
5. De Rosa S, Arcidiacono B, Chiefari E, Brunetti A, Indolfi C, Foti DP. Type 2 Diabetes Mellitus and Cardiovascular Disease: Genetic and Epigenetic Links. Front Endocrinol (Lausanne). 2018;9:2.
6. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009;19(1):92-105.
7. Stauffer BL, Russell G, Nunley K, Miyamoto SD, Sucharov CC. miRNA expression in pediatric failing human heart. J Mol Cell Cardiol. 2013;57:43-6.
8. Wang ZH, Sun XY, Li CL, Sun YM, Li J, Wang LF, et al. miRNA-21 Expression in the Serum of Elderly Patients with Acute Myocardial Infarction. Med Sci Monit. 2017;23:5728-34.
9. Rotllan N, Price N, Pati P, Goedeke L, Fernandez-Hernando C. microRNAs in lipoprotein metabolism and cardiometabolic disorders. Atherosclerosis. 2016;246:352-60.
10. Grasedieck S, Scholer N, Bommer M, Niess JH, Tumani H, Rouhi A, et al. Impact of serum storage conditions on microRNA stability. Leukemia. 2012;26(11):2414-6.
11. Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G, et al. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum Mutat. 2012;33(1):254-63.
12. de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol. 2017;35(9):872-8.
13. Huan T, Mendelson M, Joehanes R, Yao C, Liu C, Song C, et al. Epigenome-wide association study of DNA methylation and microRNA expression highlights novel pathways for human complex traits. Epigenetics. 2019:1-16.
14. Aure MR, Leivonen SK, Fleischer T, Zhu Q, Overgaard J, Alsner J, et al. Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. Genome Biol. 2013;14(11):R126.
15. Hijmans JG, Diehl KJ, Bammert TD, Kavlich PJ, Lincenberg GM, Greiner JJ, et al. Association between hypertension and circulating vascular-related microRNAs. J Hum Hypertens. 2018;32(6):440-7.
16. de Candia P, Spinetti G, Specchia C, Sangalli E, La Sala L, Uccellatore A, et al. A unique plasma microRNA profile defines type 2 diabetes progression. PLoS One. 2017;12(12):e0188980.
17. Saini HK, Griffiths-Jones S, Enright AJ. Genomic analysis of human microRNA transcripts. Proc Natl Acad Sci U S A. 2007;104(45):17719-24.
18. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68-73.

19. Nam JW, Kim J, Kim SK, Zhang BT. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Res. 2006;34(Web Server issue):W455-8.
20. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11.
21. Huan T, Rong J, Liu C, Zhang X, Tanriverdi K, Joehanes R, et al. Genome-wide identification of microRNA expression quantitative trait loci. Nat Commun. 2015;6:6601.
22. Nikpay M, Beehler K, Valsesia A, Hager J, Harper ME, Dent R, et al. Genome-wide identification of circulating-miRNA expression quantitative trait loci reveals the role of several miRNAs in the regulation of Cardiometabolic phenotypes. Cardiovasc Res. 2019.
23. Liu B, Shyr Y, Cai J, Liu Q. Interplay between miRNAs and host genes and their role in cancer. Brief Funct Genomics. 2018;18(4):255-66.
24. Lutter D, Marr C, Krumsiek J, Lang EW, Theis FJ. Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. BMC Genomics. 2010;11:224.
25. Hinske LC, França GS, Torres HA, Ohara DT, Lopes-Ramos CM, Heyn J, et al. miRIAD-integrating microRNA inter- and intragenic data. Database (Oxford). 2014;2014.
26. Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, et al. Distribution of miRNA expression across human tissues. Nucleic Acids Res. 2016;44(8):3865-77.
27. Panwar B, Omenn GS, Guan Y. miRmine: a database of human miRNA expression profiles. Bioinformatics. 2017;33(10):1554-60.
28. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
29. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131-8.
30. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. Eur J Epidemiol. 2017;32(9):807-50.
31. Touleimat N, Tost J. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012;4(3):325-41.
32. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692-702.
33. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2018;46(D1):D296-D302.
34. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.
35. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Magi R, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature. 2015;518(7538):187-96.
36. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012;44(6):659-69.
37. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, et al. Genetic variation in GIPR influences the glucose

and insulin responses to an oral glucose challenge. Nat Genet. 2010;42(2):142-8.
38. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes. 2011;60(10):2624-34.
39. Wheeler E, Leong A, Liu CT, Hivert MF, Strawbridge RJ, Podmore C, et al. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med. 2017;14(9):e1002383.
40. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42(2):105-16.
41. Scott RA, Scott LJ, Magi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66(11):2888-902.
42. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274-83.
43. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47(10):1121-30.
44. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9.
45. Ghanbari M, Sedaghat S, de Looper HW, Hofman A, Erkeland SJ, Franco OH, et al. The association of common polymorphisms in miR-196a2 with waist to hip ratio and miR-1908 with serum lipid and glucose. Obesity (Silver Spring). 2015;23(2):495-503.
46. Meerson A, Eliraz Y, Yehuda H, Knight B, Crundwell M, Ferguson D, et al. Obesity impacts the regulation of miR-10b and its targets in primary breast tumors. BMC Cancer. 2019;19(1):86.
47. Wang D, Xia M, Yan X, Li D, Wang L, Xu Y, et al. Gut microbiota metabolism of anthocyanin promotes reverse cholesterol transport in mice via repressing miRNA-10b. Circ Res. 2012;111(8):967-81.
48. Wang D, Wang W, Lin W, Yang W, Zhang P, Chen M, et al. Apoptotic cell induction of miR-10b in macrophages contributes to advanced atherosclerosis progression in ApoE-/- mice. Cardiovasc Res. 2018;114(13):1794-805.
49. Silbernagel G, Chapman MJ, Genser B, Kleber ME, Fauler G, Scharnagl H, et al. High intestinal cholesterol absorption is associated with cardiovascular disease and risk alleles in ABCG8 and ABO: evidence from the LURIC and YFS cohorts and from a meta-analysis. J Am Coll Cardiol. 2013;62(4):291-9.
50. Goedeke L, Rotllan N, Canfran-Duque A, Aranda JF, Ramirez CM, Araldi E, et al. MicroRNA-148a regulates LDL receptor and ABCA1 expression to control circulating lipoprotein levels. Nat Med. 2015;21(11):1280-9.
51. Gailhouste L, Gomez-Santos L, Hagiwara K, Hatada I, Kitagawa N, Kawaharada K, et al. miR-148a plays a pivotal role in the liver by promoting the hepatospecific phenotype and suppressing the invasiveness of transformed cells. Hepatology. 2013;58(3):1153-65.
52. Wagschal A, Najafi-Shoushtari SH, Wang L, Goedeke L, Sinha S, deLemos AS, et al. Genome-wide identification of microRNAs

regulating cholesterol and triglyceride homeostasis. Nat Med. 2015;21(11):1290-7.
53. Rockstroh D, Loffler D, Kiess W, Landgraf K, Korner A. Regulation of human adipogenesis by miR125b-5p. Adipocyte. 2016;5(3):283-97.
54. Wang X, Zheng Y, Ma Y, Du L, Chu F, Gu H, et al. Lipid metabolism disorder induced by up-regulation of miR-125b and miR-144 following beta-diketone antibiotic exposure to F0-zebrafish (Danio rerio). Ecotoxicol Environ Saf. 2018;164:243-52.
55. Zhang ZC, Liu Y, Xiao LL, Li SF, Jiang JH, Zhao Y, et al. Upregulation of miR-125b by estrogen protects against non-alcoholic fatty liver in female mice. J Hepatol. 2015;63(6):1466-75.
56. Ligthart S, Steenaard RV, Peters MJ, van Meurs JB, Sijbrands EJ, Uitterlinden AG, et al. Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. Diabetologia. 2016;59(5):998-1006.
57. Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. Clin Epigenetics. 2017;9:15.
58. Nano J, Ghanbari M, Wang W, de Vries PS, Dhana K, Muka T, et al. Epigenome-Wide Association Study Identifies Methylation Sites Associated With Liver Enzymes and Hepatic Steatosis. Gastroenterology. 2017;153(4):1096-106 e2.



Chapter 4.2

Genetic Polymorphism of miR-196a-2 is Associated with Bone Mineral Density (BMD)

Irma Karabegović, **Silvana Maas**, Carolina Medina-Gomez, Maša Zrimšek, Sjur Reppe, Kaare M. Gautvik, André G. Uitterlinden, Fernando Rivadeneira, and Mohsen Ghanbari

International Journal of Molecular Sciences 2017;18(12):2529

ABSTRACT

MicroRNAs (miRNAs) are small non-coding RNA molecules that post-transcriptionally regulate the translation of messenger RNAs. Given the crucial role of miRNAs in gene expression, genetic variants within miRNA-related sequences may affect miRNA function and contribute to disease risk. Osteoporosis is characterized by reduced bone mass, and bone mineral density (BMD) is a major diagnostic proxy to assess osteoporosis risk. Here, we aimed to identify miRNAs that are involved in BMD using data from recent genome-wide association studies (GWAS) on femoral neck, lumbar spine and forearm BMD. Of 242 miRNA-variants available in the GWAS data, we found rs11614913:C > T in the precursor miR-196a-2 to be significantly associated with femoral neck-BMD (p-value = 9.9×10^{-7}, $\beta = -0.038$) and lumbar spine-BMD (p-value = 3.2×10^{-11}, $\beta = -0.061$). Furthermore, our sensitivity analyses using the Rotterdam study data showed a sex-specific association of rs11614913 with BMD only in women. Subsequently, we highlighted a number of miR-196a-2 target genes, expressed in bone and associated with BMD, that may mediate the miRNA function in BMD. Collectively, our results suggest that miR-196a-2 may contribute to variations in BMD level. Further biological investigations will give more insights into the mechanisms by which miR-196a-2 control expression of BMD-related genes.

INTRODUCTION

Osteoporosis is characterized by reduced bone mass and micro-architectural degradation of bone tissue, resulting in increased bone fragility, with a consequent increase in fracture susceptibility (1). This is a common disease affecting one in three women and one in five men worldwide (2). Incidence and development of osteoporosis increases exponentially with age (3). The disease is diagnosed by common imaging modalities, and therefore, might be modifiable to prevent fractures (3,4). A major diagnostic proxy to assess osteoporosis risk in the clinical field is bone mineral density (BMD) measurements, especially in skeletal sites where osteoporotic fractures occur more frequently (i.e., lumbar spine, hip and forearm) (5). Genetic studies have estimated that 50–85% of the variance in BMD can be attributed to genetic factors (6). A number of protein-coding genes as well as non-coding genes have been posited to contribute to osteoporosis or decreased BMD (7,8,9,10). Functional genetics have also demonstrated eight genes that could explain up to 40% of BMD variation in postmenopausal osteoporosis and involve risk of fracture (11,12).

MicroRNAs (miRNAs) are small non-coding RNAs, approximately ~22 nucleotides long, which post-transcriptionally regulate gene expression. Together, they are estimated to regulate more than half of the genes in our genome (13). miRNAs' mode of action involves imperfect matching of the "seed region" (nucleotides 2–8 from the 5′ end of mature miRNA sequence) with a partially complementary sequence located at the 3′ UTR of target mRNA, resulting in translational inhibition and/or mRNA degradation (14). It has been shown that genetic variants in miRNAs contribute to disease risk (14,15,16,17). Polymorphisms in miRNA genes are presumed to alter miRNA biogenesis and consequently change the expression of the miRNA target genes (14,15). This altered gene expression might result in phenotypic variation (18). There are strong indications that miRNAs influence BMD levels by regulating several genes involved in bone-related pathways (19). For example, miR-146a has been shown to regulate TRAF6 and IRAK1 genes involved in apoptosis (20). In osteoclasts, these genes mediate IL-1β-induced activation of NF-κB signaling, which in turn promotes osteoclast activity and survival (21,22). Furthermore, previous candidate gene studies have shown that genetic variants within miRNA genes (e.g., miR-146, miR-125a, miR-27a, miR-433) are associated with osteoporosis and bone cell activity, possibly through altering the miRNA expression levels or function (9,23,24,25,26).

In the present study, we hypothesized that genetic variants in miRNAs affect miRNA-mediated regulation of genes involved in BMD. To test this hypothesis, we performed a genome-wide scan for miRNA variants associated with BMD using data from the recent genome-wide association studies (GWAS) on femoral neck, lumbar spine and forearm BMD (7). We found a genetic variant in pre-miR-196a-2 significantly associated with BMD. Subsequently, we performed in silico analyses to investigate whether miR-196a-2 and its putative target genes may contribute to BMD variation.

RESULTS

A Variant in miR-196a-2 Associates with BMD

A total of 2340 variants in miRNA-related sequences were collected by combination of a literature review and miRNASNP database (27). In parallel, we extracted summary statistics data from the recent GWAS meta-analysis on three BMD phenotypes, including femoral neck (FN-BMD), lumbar spine (LS-BMD) and forearm (FA-BMD), provided by Genetic Factors of Osteoporosis (GEFOS) consortium (7). Out of 2340 miRNA variants, 90 single-nucleotide polymorphisms (SNPs) were available in the GWAS data. Using the SNAP Web tool, we extracted the proxy SNPs (R2 > 0.8 and distance < 200 kb in 1000 Genomes project) for 152 of the unavailable variants. We studied the association of these 242 miRNA SNPs with BMD phenotypes. One of the SNPs passed the Bonferroni significance threshold of 2.1×10^{-4} (0.05/242). This includes rs11614913:C > T in miR-196a-2 which is significantly associated with FN-BMD (p-value = 9.9×10^{-7}, $\beta = -0.038$) and LS-BMD (p-value = 3.2×10^{-11}, $\beta = -0.061$). This analysis indicated that individuals carrying the rs11614913 minor allele T are more prone to have lower BMD. No significant association was identified between the miRNA variants and FA-BMD. A simplified scheme of the pipeline used for the identification of miRNA SNPs associated with the BMD phenotypes is shown in **Figure 1**.

The Potential Impact of rs11614913 on the miR-196a-2 Structure and Function

We generated the hairpin structures of hsa-miR-196a-2 containing either the major allele C or the minor allele T at rs11614913 site using the Vienna RNAfold algorithm (28). We observed 4.6 kcal/mol difference in the minimum free energy (MFE) of the thermodynamic predicted structure of pre-miR-196a-2 with the minor allele T compared to the wild type allele C (**Figure 2**). The analysis suggests that the investigated variant may affect the stability of miR-196a-2. In this line, it has been demonstrated previously that rs11614913-T decreases miR-196a-2 expression in different cell lines (29, 30).

Associaton of miR-196a-2 Target Genes with BMD

Through leveraging the GEFOS GWAS data and using a candidate gene approach, we tested the association of genetic variants in 457 putative target genes of miR-196a-2 with FN-BMD and LS-BMD. **Table 1** shows the top ten target genes of miR-196a-2 with the most significant association with the BMD phenotypes. Using RNA-seq gene expression data of 86 hip bone (iliac crest) biopsies, we found evidence for expression of eight out of the ten highlighted target genes of miR-196a-2 in bone (**Figure 3**) (12). Among the bone-expressed targets, JAG1 passed the significance threshold, based on the number of variants in the tested miR-196a-2 target genes (**Table 1**). This analysis may suggest

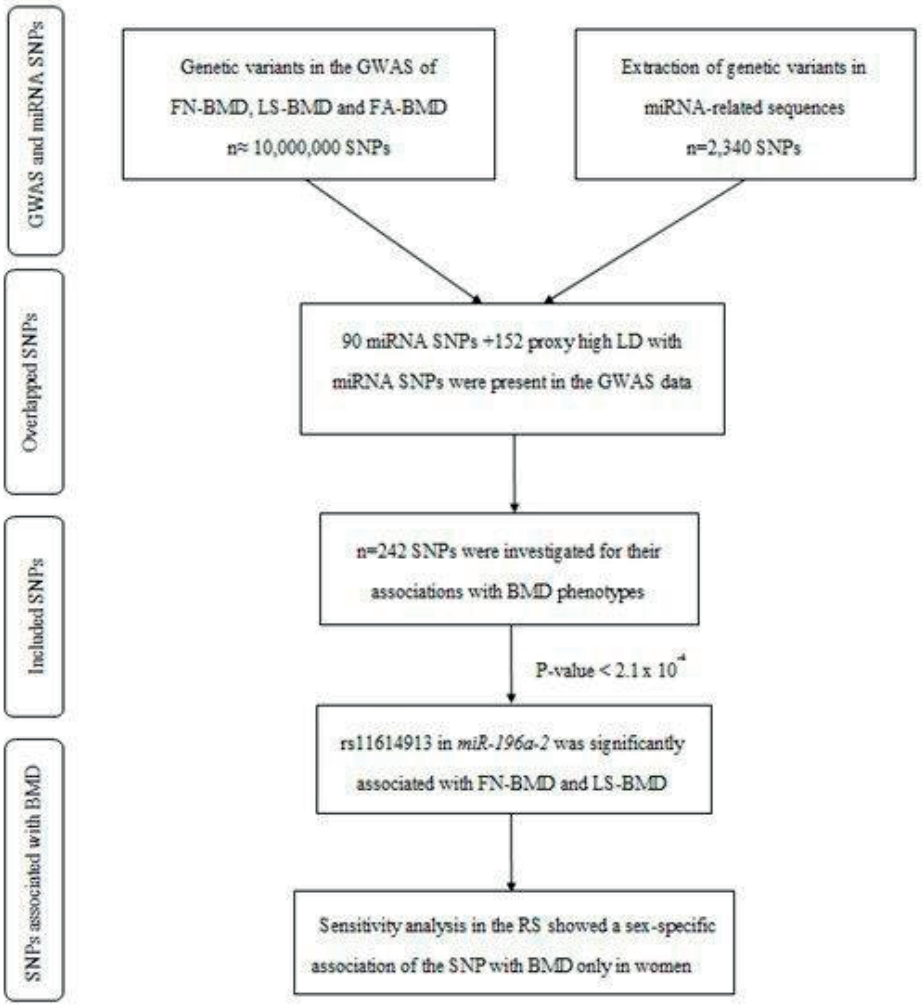


Figure 1. A simplified diagram of the pipeline used to identify miRNA genetic variants associated with BMD. FN-BMD: Femoral neck bone mineral density; LS-BMD: Lumbar spine bone mineral density; FA-BMD: Forearm bone mineral density; SNP: Single-nucleotide polymorphism; GWAS: Genome-wide association studies.

that JAG1 is more likely to mediate the downstream effect of miR-196a-2 in relation to BMD. Moreover, a number of genes have been demonstrated experimentally (i.e., by luciferase reporter assay, Western blot or qPCR) to be regulated by miR-196a-2. As shown in supplementary **Table S1** some of these genes are shown to be involved in either osteogenesis or bone function and may mediate the miR-196a-2 effect on BMD. We checked the correlation of rs11614913 with expression level of its surrounding genes as shown by GTEX portal (http://www.gtexportal.org/home/) and found the association of SNP with expression of HOXC8 and HOXC-AS1 across different tissues.

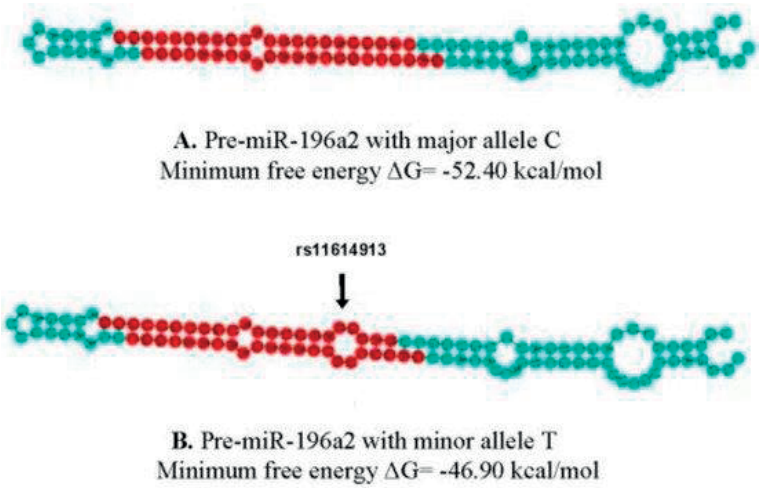


Figure 2. Schematic view of the predicted pre-miR-196a-2 hairpin structure containing the SNP major allele C or minor allele T. The minimum free energy (MFE) change of the thermodynamic ensemble (ΔG) is shown. The red part indicates mature sequence and the blue part shows the rest of pre-miRNA sequence.

Sensitivity Analyses for rs11614913 in miR-196a-2 Using the Rotterdam Study Data

Previous studies have reported sex-specific association of genetic variants with BMD (31,32). Furthermore, some studies have shown difference in sex response to musculoskeletal cell development, mediated by influence of steroid hormones (33,34). In order to investigate the potential difference in association between the miR-196a-2 variants and BMD across sexes, we performed a sensitivity analysis using the Rotterdam study (RS) data. The baseline characteristics of the RS participants are shown in **Table 2**. A total of 6,145 participants (3524 woman and 2621 men) from the three RS cohorts were eligible for this analysis (individuals with data available for rs11614913 and Dual X-ray Absorptiometry (DXA) imaging on FN-BMD and LS-BMD). Mixed linear regression analysis was carried out in sex-stratified data to investigate the association between rs11614913 and the BMD phenotypes (**Table 3**). In the basic model (adjusting for age, cohort, weight, waist to hip ratio and height) there was a significant association between rs11614913 and FN-BMD only in women (p-value = 0.003; β = 0.009; (95%Confidence Interval, CI) = 0.003, 0.014). The association remained significant for women in the second model (further adjusting for alcohol, smoking status and drugs used for treatment of bone diseases) (p-value = 0.003; β = 0.008; (95%CI) = 0.003, 0.014). We also tested the association between rs11614913 and LS-BMD and found, again, a clear significance only in women in the basic model (p-value = 0.023; β = 0.010; (95%CI) = 0.001, 0.019) and the second model (p-value = 0.026; β = 0.010; (95%CI) = 0.001, 0.018) (**Table 3**). Next, we further adjusted the second model for sex-hormones to see whether the miRNA variant is linked to sex-hormones (**Table 3**). The association in females remained significant after further

adjustment for five sex-hormones (Model 3) involved in the steroidogenesis pathway. These results suggest that there is sex specificity in the association of miR-196a-2 with BMD.

Table 1. Putative target genes of miR-196a-2 (3p and 5p) that are associated with FN-BMD and LS-BMD.

miRNA ID	Associated Phenotype	Associated Target Genes	*p*-Value in GWAS Data	Top SNP
miR-196a-3p	FN-BMD	*JAG1*	1.8×10^{-5}	rs2235811
		MACROD2	2.0×10^{-6}	rs365824
		SP1	4.2×10^{-5}	rs4759334
	LS-BMD	***JAG1***	**4.7×10^{-9}**	**rs2235811**
		ATF7	6.3×10^{-5}	rs1078358
		MACROD2	8.1×10^{-5}	rs6110288
miR-196a-5p	FN-BMD	*FRMD4B*	5.6×10^{-4}	rs1564757
		NEDD4L	9.6×10^{-4}	rs533502
		BIRC6	1.2×10^{-3}	rs6737916
	LS-BMD	*COL24A1*	2.6×10^{-3}	rs1359419
		RSPO2	3.1×10^{-3}	rs446454
		DIP2A	3.3×10^{-3}	rs2330593

Leading SNPs within each target gene associated with BMD in GEFOS GWAS data are shown. Significantly associated genes, after Bonferroni correction for multiple testing (p-value <7.0×10^{-6}), are depicted in bold.

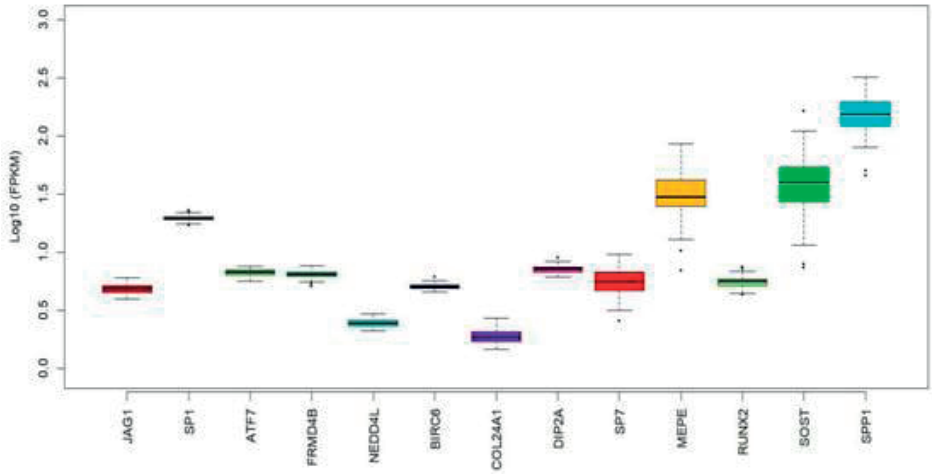


Figure 3. Expression of the highlighted miR-196a-2 target genes and positive controls (SP7, MEPE, RUNX2, SOST and SPP1) in RNA-seq data consisting of 86 hip bone (iliac crest) biopsies. The expression data are shown in the metric Log10 FPKM (fragments per kilobase of transcript per million mapped reads).

Table 2. Demographic characteristics of the Rotterdam study cohorts.

Variables		Men	Women
FN-BMD (g/cm^2)		0.95 (0.14)	0.87 (0.14)
LS-BMD (g/cm^2)		1.20 (0.19)	1.08 (0.19)
Age (years)		65.71 (10.45)	66.29 (10.61)
Weight (kg)		85.55 (12.85)	73.11 (13.09)
WHR		0.95 (0.07)	0.84 (0.07)
Height (cm)		176.41 (7.01)	162.73 (6.50)
Alcohol (g/day)		9.29 (3.57–20.00)	4.29 (0.54–10.00)
DHEA (nmol/L)		11.82 (7.32)	12.31 (7.65)
DHEAS (nmol/L)		3200.18 (1757.16)	2099.17 (1337.77)
Androstenedione (nmol/L)		3.24 (1.27)	2.70 (1.29)
Testosterone (nmol/L)		17.53 (5.78)	0.90 (0.45)
Estradiol (pmol/L)		96.93 (33.82)	38.86 (33.18)
Smoking	never smoker	1125 (42.9%)	2071 (58.8%)
	former smoker	1039 (39.7%)	841 (23.9%)
	current smoker	456 (17.4%)	612 (17.4%)
Bone drugs	no	2607 (99.5%)	3400 (96.5%)
	yes	13 (0.5%)	124 (3.5%)

Values are mean (standard deviation), numbers (percentages) or median (interquartile range (IQR)); used for alcohol only. FN-BMD: Femoral neck bone mineral density; LS-BMD: Lumbar spine bone mineral density; WHR: Waist to hip ratio; Bone drugs: drugs used for treatment of bone diseases; DHEA: dehydroepiandrosterone; DHEAS: dehydroepiandrosterone sulfate.

Table 3. Association between rs11614913 and BMD phenotypes in participants of the Rotterdam Study.

Phenotype	Model	Men			Women			Combined		
		β	95%CI	*p*-Value	β	95%CI	*p*-v	β	95%CI	*p*-Value
FN-BMD	M1	0.004	−0.003, 0.011	0.257	0.009	0.003, 0.014	0.003	0.007	0.003, 0.012	0.002
	M2	0.004	−0.003, 0.011	0.267	0.008	0.003, 0.014	0.003	0.007	0.003, 0.012	0.002
	M3	0.004	−0.004, 0.011	0.319	0.008	0.003, 0.014	0.003	0.007	0.002, 0.011	0.003
LS-BMD	M1	0.005	−0.006, 0.016	0.380	0.010	0.001, 0.019	0.023	0.009	0.002, 0.016	0.011
	M2	0.004	−0.007, 0.015	0.423	0.010	0.001, 0.018	0.026	0.009	0.002, 0.016	0.012
	M3	0.003	−0.008, 0.014	0.573	0.009	0.001, 0.018	0.038	0.008	0.001, 0.015	0.020

Model 1 (M1) is adjusted for age, cohort, weight, waist to hip ratio (WHR) and height. Model 2 (M2) is adjusted for M1 + alcohol, smoking status (current, former and never smoker) and drugs used for treatment of bone diseases. Model 3 (M3) is adjusted for M2 + estradiol, testosterone, androstenedione, DHEA, and DHEAS. "Combined" was additionally adjusted for sex.

DISCUSSION

Recent studies have shown that miRNAs are important regulators of genes linked to bone remodeling and osteoporosis development (35,36,37,38,39). Different approaches have been used in previous studies to identify miRNAs involved in osteoporosis, including

miRNA expression profiling (38,40) and candidate gene association studies (41). In this study, we have conducted a genome-wide scan investigating the association of miRNA genetic variants with BMD using GWAS data (7). This method represents a valuable, extended and complementary approach to previous methods used in the identification of miRNAs associated with BMD.

Our results showed that rs11614913 in the stem region of pre-miR-196a-2 is significantly associated with FN-BMD and LS-BMD. Lack of significant association between rs11614913 within pre-miR-196a-2 and forearm BMD could be attributed to the small sample size in GWAS (n = 8143) compared to FN-BMD (n = 32,735) or LS-BMD (n = 28,498) in the discovery cohorts (7), or differences in bone remodeling between anatomical sites. It has been shown that loaded and unloaded bone (forearm) have distinct transcriptional activities (42,43). The location of rs11614913 in pre-miR-196a-2 is likely to affect the miRNA processing by enzyme Dicer, and subsequently alter the expression of mature miR-196a-2 (44,45). Polymorphisms in pre-miRNA sequences have been shown to cause either a destabilization of the interaction due to changes in the free binding energy or a change in target accessibility due to alternations in the miRNA secondary structure (19,46,47). Our in silico analysis showed differences in the MFE between the predicted structure of pre-miR-196a-2 mutants and the wild type, suggesting the variant's minor allele may diminish the stability of pre-miR-196a-2. In agreement with this conjecture, previous studies have established the impact of rs11614913 polymorphism (C/T) on the miR-196a-2 expression levels (29,30,44,45,48). Zhibin Hu et al., have reported that rs11614913 wild-type allele (C) is associated with statistically significant increase in mature miR-196a-2 expression, while studying 23 human lung cancer tissue samples (30). They also showed that rs11614913 could affect binding of the mature miR-196a-2 to its candidate target mRNA (30). Furthermore, Zhao Hauanhuan et al., observed the same trend of rs11614913*CC genotype to increase the mature miR-196a-2 expression in different phenotypes of breast cancer (29). Likewise, Hoffman et al., experimentally demonstrated that rs11614913 mutant allele (T) is associated with statistically significant decrease in miR-196a-2 expression in breast cancer patients (44). Another study by Vinci et al., presented coherent results of rs11614913*TT decreasing miR-196a-2 expression levels in lung cancer patients (48). In addition, Xu et al., determined that rs11614913 affects the expression of miR-196a-2 and consequently, expression of its downstream target gene HOXB8 (49). They hypothesized that the variant might have an impact on miR-196a-HOXB8-Shh signaling pathway, and therefore, be associated with congenital heart disease susceptibility (49). In other studies, the miR-196a-2 polymorphism rs11614913 has been linked to various phenotypic variations, ranging from several types of cancer (30,44,45,50) to increased risk for cardiovascular disease (49,51,52,53,54). These data strongly suggest an important functional impact of rs11614913 on miR-196a-2 expression and function that in turn might affect the risk and/or progression of disease.

MiR-196a is shown to be expressed from HOX clusters loci in mammals and HOX genes in turn are shown to be targets of miR-196a (19,55). The HOX genes play critical roles in limb development and skeletal patterning (56,57). The miRNA has been also shown to play a role in brown adipogenesis of white fat progenitor cells through targeting HOXC8 (58). It has been proven that the miRNA regulates HOXC8 at both mRNA and protein levels (55). In an independent study, Kim et al., observed that adding miR-196-a inhibitors to osteoblast cells in culture causes a significant increase in HOXC8 protein levels, with subsequent increased proliferation and decrease in osteogenic differentiation (59). These data suggest upregulation of HOXC8 in the miR-196a-2 variant carriers, of significance for osteogenic differentiation. Accordingly, Dong-Li Zhu et al., have recently shown that miR-196a-2 is expressed in osteoblasts and experimentally demonstrated that FGF2, previously identified as a susceptibility gene for osteoporosis in Caucasians (60), is a direct target of miR-196a-2 in the Chinese population (8). Their experiments proved that miR-196a-2 had an influence on FGF2 mRNA in hFOB1 cells, which is a human fetal osteoblastic cell line (8).

In addition to previously validated targets of miR-196a-2 involved in osteogenesis, we highlighted a number of putative target genes associated with BMD with a potential to mediate the miR-196a-2 effect in BMD. Among them, JAG1 passed the significant threshold to be associated with BMD and is expressed in bone. The JAG1 gene has been previously reported to be associated with increased BMD and suggested as a candidate gene for BMD regulation in diverse ethnic groups (61). Future experimental studies are needed to explore the postulated miR-196a-2-mediated regulation of the gene in bone tissue or cell lines.

We performed sex-stratified analysis using the Rotterdam study data to get insight into sex specificity for BMD variation on the miR-196a-2 polymorphism. In the sex-combined analysis, we observed significant association of rs11614913 with BMD phenotypes. However, sex-stratified analysis revealed that the association is mainly driven by women. We acknowledge that the observed association in women may have been driven by a lower number of men (our cohort contains 903 more women than men), however, sample size of 6145 should be sustainable to address sex difference. Notably, the miR-196a-2 polymorphism rs11614913 with combination of rs3746444 in miR-499a have been reported previously to be involved in the multiple sclerosis severity, where the association shows only female sex specificity (62). Multiple sclerosis and osteoporosis share a surprising number of risk factors (63,64,65) and genetics might be one of them, although the interplay of the two miRNA variants and their impacts on gene interaction should be taken in consideration when interpreting the results regarding sex specificity. Considering the sexual dimorphism of bone (31,66), these data might indicate a potential for further clinical and biological investigations regarding the role of miR-196a-2 underlying BMD variation.

This study has some strengths and limitations that need to be considered in interpretation of the reported results. The major strength of this study is leveraging genetic data from the recent GWAS of BMD phenotypes that enabled us legitimate statistical power for detection of miRNA-related variants associated with BMD. The main limitation that needs to be addressed is lack of experimental studies in relevant tissues or cell lines. MiRNA-related SNPs might be only utilitarian if the target mRNA is expressed in the same tissue (67). Thereby, further biological investigations warrant better insights into the mechanisms by which miR-196a-2 control expression of genes involved in BMD.

METHODS

Genome-Wide Association Studies on BMD Phenotypes

The summary statistics from the recent GWAS meta-analysis on FN-BMD (n = 32,735), LS-BMD (n = 28,498) and FA-BMD (n = 8143) provided by GEFOS consortium were extracted (7). The GEFOS consortium is a collective effort of numerous research groups combining GWAS data, in order to identify osteoporosis susceptibility alleles that regulate BMD and fracture risk (7). The GEFOS consortium performed meta-analysis of whole genome sequencing, whole exome sequencing and deep imputation of genotype data in order to determine low-frequency and rare variants associated with risk factors for osteoporosis. The collaboration within the GEFOS has resulted in producing files with summary statistics for approximately 10 million genetics variants (the 1000 Genomes/UK10K reference panel) in 53,236 individuals (7). More details on datasets and participants are described in detail elsewhere (7).

Identification of Genetic Variants in miRNA-Encoding Sequences

A dataset of single-nucleotide polymorphisms (SNPs) in miRNA-related sequences was created by combining miRNASNP (http://www.bioguo.org/miRNASNP/) (27) and the literature review (searching in PubMed for miRNA genetic variants). Precursor miRNA sequences (pre-miRNA) undergo cleavage by enzyme Dicer, yielding to mature miRNAs (13), therefore we screened all variants located in human pre-miRNA and mature miRNA sequences. The methodology was explained in details elsewhere (68). Variants with minor allele frequency (MAF) >0.01 were included. Variants with smaller MAF were illegible due to low imputation quality and issue of being underpowered in further studies. In total, 2340 miRNA variants were extracted. Of these, 242 variants were available in the GEFOS GWAS data and were therefore investigated further for their associations with BMD phenotypes.

miRNA Target Genes Associated with BMD Phenotypes

Once a miRNA variant was found to be significantly associated with BMD phenotypes, we searched for the miRNA target genes. We postulated that some of the miRNA target genes may mediate the downstream effect of miRNA in relation to BMD phenotypes. In order to identify target genes of miRNAs, putative target genes were extracted from combining TargetScan v7.1 (http://www.targetscan.org/vert_71/) and miRDB (http://mirdb.org/) database (69). Target genes present in both databases were selected for further investigation. Any supplementary information, such as miRNA conservation between species, host genes, miRNA sequences was collected from TargetScan (v7.1). Both context score and conserved target sites were used to rank the miRNA target genes. In addition, the online database, miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/) provides information on various functional experiments, such as microarrays, western blot, and reported assays performed between miRNAs and their target genes (70). We used miRTarBase to search for functional experiment confirming the putative interaction between miRNAs of interest and their target genes. A candidate gene approach was performed by leveraging the GWAS data on BMD phenotypes (7) and to investigate the association between genetic variants in the miRNA target genes and BMD. In addition, we evaluated the expression of selected target genes in the bone tissue. Dataset used for gene expression was created out of 86 iliac biopsies (12).

The Variant Effect on the Pre-miRNA Structure

The secondary structure of pre-miRNA is critical for the miRNA production. The Vienna RNAfold algorithm (ViennaRNA package 2.0) was used to predict the impact of miRNA variants on the hairpin stem-loop structure of pre-miRNAs (28). The ViennaRNA package 2.0 is available to the public domain and relies on numerous algorithms for prediction and analysis of RNA secondary structures (71). The program calculates the shift in minimum free energy (MFE) of the thermodynamic ensemble in the hairpin structure of miRNA (wild type and mutant) (72). The shift in MFE is likely to be related to the function, as it can result in instability of miRNA.

The Rotterdam Study Data

The Rotterdam study (RS) is a population-based cohort study, with main goal of identifying chronic disabling conditions of the middle aged and elderly people (73). Participants were interviewed at home and went through an extensive set of examinations, including bone mineral densitometry, sample collections for in-depth molecular and genetic analysis (73). The RS includes three sub-cohorts. We used the data from the baseline, second and third cohort (RS-I-4, RS-II-2, and RS-III-1). For all participants, DXA-based BMD measurements were collected for FN-BMD and LS-BMD. The RS does not include data on FA-BMD since this site is used for prediction of osteoporosis only when data is

not available for FN-BMD or LS-BMD due to numerous reasons (e.g., patients either being obese, men with hyperparathyroidism or receiving androgen-deprivation therapy (ADT) for prostate cancer) (74). Furthermore, determinants were assessed either by physical examinations, collection of blood samples, or by questionnaires. Participants were included if they had FN-BMD or LS-BMD measurements, which resulted in combination of three cohorts (RS-I-4, RS-II-2, and RS-III-1). We used multiple linear regression in sex-stratified dataset to examine the association between the candidate miRNA variant and BMD phenotypes (separately). Our analysis was adjusted for all potential confounders in three models.

CONCLUSIONS

The results of this study suggest that miR-196a-2 polymorphism (rs11614913:C > T) is associated with reduced FN-BMD and LS-BMD. We highlighted a number of target genes that may mediate miR-196a-2 function in influencing BMD. The identified miR-196a-2 might have a future implication in the clinical field related to diagnosis and treatment of osteoporosis. Future biological studies will give insight into the mechanisms by which miR-196a-2 may control expression of bone-related genes. Collectively, our study provides further understanding of the miRNA-mediated regulation of BMD.

SUPPLEMENTARY MATERIALS

Supplemental material for this chapter can be found in the online version of the paper via https://www.mdpi.com/1422-0067/18/12/2529/htm.

REFERENCES

1. Hernlund, E.; Svedbom, A.; Ivergard, M.; Compston, J.; Cooper, C.; Stenmark, J.; McCloskey, E.V.; Jonsson, B.; Kanis, J.A. Osteoporosis in the European Union: Medical management, epidemiology and economic burden. A report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA). Arch Osteoporos 2013, 8, 136.
2. Rivadeneira, F.; Makitie, O. Osteoporosis and bone mass disorders: From gene pathways to treatments. Trends Endocrinol. Metab. 2016, 27, 262–281.
3. Rivadeneira, F.; Styrkarsdottir, U.; Estrada, K.; Halldorsson, B.V.; Hsu, Y.H.; Richards, J.B.; Zillikens, M.C.; Kavvoura, F.K.; Amin, N.; Aulchenko, Y.S.; et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. Nat. Genet. 2009, 41, 1199–1206.
4. Kling, J.M.; Clarke, B.L.; Sandhu, N.P. Osteoporosis prevention, screening, and treatment: A review. J. Womens Health 2014, 23, 563–572.
5. Blake, G.M.; Fogelman, I. The role of DXA bone density scans in the diagnosis and treatment of osteoporosis. Postgrad. Med. J. 2007, 83, 509–517.
6. Stewart, T.L.; Ralston, S.H. Role of genetic factors in the pathogenesis of osteoporosis. J. Endocrinol. 2000, 166, 235–245.
7. Zheng, H.F.; Forgetta, V.; Hsu, Y.H.; Estrada, K.; Rosello-Diez, A.; Leo, P.J.; Dahia, C.L.; Park-Min, K.H.; Tobias, J.H.; Kooperberg, C.; et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 2015, 526, 112–117.
8. Zhu, D.L.; Guo, Y.; Zhang, Y.; Dong, S.S.; Xu, W.; Hao, R.H.; Chen, X.F.; Yan, H.; Yang, S.Y.; Yang, T.L. A functional SNP regulated by miR-196a-3p in the 3' UTR of FGF2 is associated with bone mineral density in the Chinese population. Hum. Mutat. 2017, 38, 725–735.
9. Dole, N.S.; Kapinas, K.; Kessler, C.B.; Yee, S.P.; Adams, D.J.; Pereira, R.C.; Delany, A.M. A single nucleotide polymorphism in osteonectin 3′ untranslated region regulates bone volume and is targeted by miR-433. J. Bone Miner. Res. 2015, 30, 723–732.
10. Guo, L.J.; Liao, L.; Yang, L.; Li, Y.; Jiang, T.J. MiR-125a TNF receptor-associated factor 6 to inhibit osteoclastogenesis. Exp. Cell Res. 2014, 321, 142–152.
11. Jemtland, R.; Holden, M.; Reppe, S.; Olstad, O.K.; Reinholt, F.P.; Gautvik, V.T.; Refvem, H.; Frigessi, A.; Houston, B.; Gautvik, K.M. Molecular disease map of bone characterizing the postmenopausal osteoporosis phenotype. J. Bone Miner. Res. 2011, 26, 1793–1801.
12. Reppe, S.; Refvem, H.; Gautvik, V.T.; Olstad, O.K.; Hovring, P.I.; Reinholt, F.P.; Holden, M.; Frigessi, A.; Jemtland, R.; Gautvik, K.M. Eight genes are highly associated with BMD variation in postmenopausal Caucasian women. Bone 2010, 46, 604–612.
13. Ha, M.; Kim, V.N. Regulation of microRNA biogenesis. Nat. Rev. Mol. Cell Biol. 2014, 15, 509–524.
14. Bushati, N.; Cohen, S.M. microRNA functions. Annu. Rev. Cell Dev. Biol. 2007, 23, 175–205.
15. Ardekani, A.M.; Naeini, M.M. The role of microRNAs in human diseases. Avicenna J. Med. Biotechnol. 2010, 2, 161–179.
16. Tufekci, K.U.; Oner, M.G.; Meuwissen, R.L.; Genc, S. The role of microRNAs in human diseases. Methods Mol. Biol. 2014, 1107, 33–50.
17. Zhang, Y.; Lu, Y.J.; Yang, B.F. Potential role of microRNAs in human diseases and the exploration on design of small molecule agents. Acta Pharm. Sin. 2007, 42, 1115–1121.

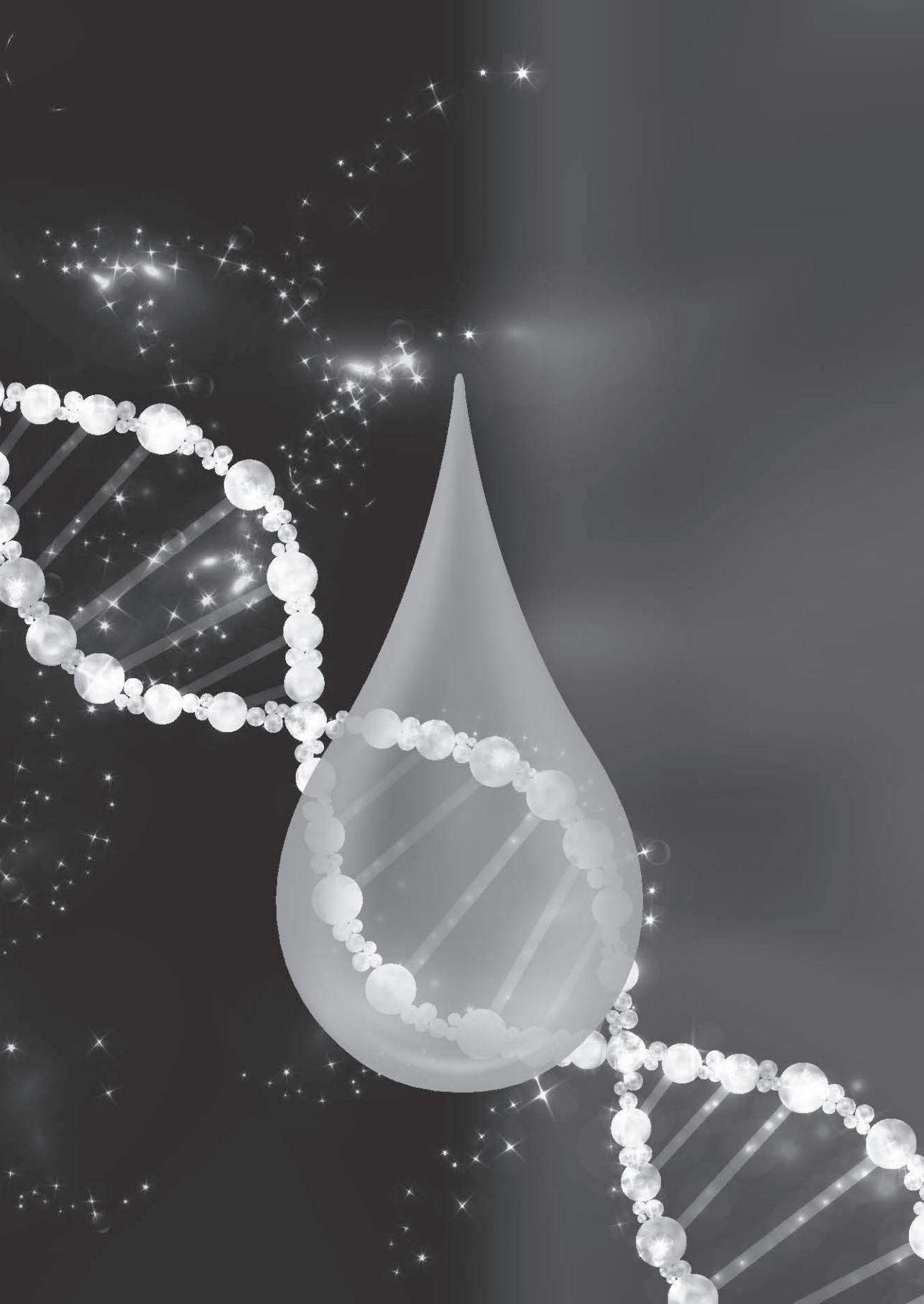
18. Lu, J.; Clark, A.G. Impact of microRNA regulation on variation in human gene expression. Genome Res. 2012, 22, 1243–1254.
19. Dole, N.S.; Delany, A.M. microRNA variants as genetic determinants of bone mass. Bone 2016, 84, 57–68.
20. Park, H.; Huang, X.; Lu, C.; Cairo, M.S.; Zhou, X. microRNA-146a and microRNA-146b regulate human dendritic cell apoptosis and cytokine production by targeting TRAF6 and IRAK1 proteins. J. Biol. Chem. 2015, 290, 2831–2841.
21. Gravallese, E.M.; Galson, D.L.; Goldring, S.R.; Auron, P.E. The role of TNF-receptor family members and other TRAF-dependent receptors in bone resorption. Arthritis Res. 2001, 3, 6–12.
22. Kim, J.H.; Jin, H.M.; Kim, K.; Song, I.; Youn, B.U.; Matsuo, K.; Kim, N. The mechanism of osteoclast differentiation induced by IL-1. J. Immunol. 2009, 183, 1862–1870.
23. Chen, P.; Wei, D.; Xie, B.; Ni, J.; Xuan, D.; Zhang, J. Effect and possible mechanism of network between microRNAs and RUNX2 gene on human dental follicle cells. J. Cell. Biochem. 2014, 115, 340–348.
24. Nakasa, T.; Shibuya, H.; Nagata, Y.; Niimoto, T.; Ochi, M. The inhibitory effect of microRNA-146a expression on bone destruction in collagen-induced arthritis. Arthritis Rheumatol. 2011, 63, 1582–1590.
25. Polesskaya, A.; Cuvellier, S.; Naguibneva, I.; Duquet, A.; Moss, E.G.; Harel-Bellan, A. Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency. Genes Dev. 2007, 21, 1125–1138.
26. Hassan, M.Q.; Gordon, J.A.; Beloti, M.M.; Croce, C.M.; van Wijnen, A.J.; Stein, J.L.; Stein, G.S.; Lian, J.B. A network connecting Runx2, SATB2, and the miR-23a~27a~24-2 cluster regulates the osteoblast differentiation program. Proc. Natl. Acad. Sci. USA 2010, 107, 19879–19884.
27. Gong, J.; Liu, C.; Liu, W.; Wu, Y.; Ma, Z.; Chen, H.; Guo, A.Y. An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. Database (Oxford) 2015, 2015.
28. Lorenz, R.; Bernhart, S.H.; Honer Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. Algorithms Mol. Biol. 2011, 6, 26.
29. Zhao, H.; Xu, J.; Zhao, D.; Geng, M.; Ge, H.; Fu, L.; Zhu, Z. Somatic mutation of the SNP rs11614913 and its association with increased miR-196a2 Expression in Breast Cancer. DNA Cell Biol. 2016, 35, 81–87.
30. Hu, Z.; Chen, J.; Tian, T.; Zhou, X.; Gu, H.; Xu, L.; Zeng, Y.; Miao, R.; Jin, G.; Ma, H.; et al. Genetic variants of miRNA sequences and non-small cell lung cancer survival. J. Clin. Investig. 2008, 118, 2600–2608.
31. Karasik, D.; Ferrari, S.L. Contribution of gender-specific genetic factors to osteoporosis risk. Ann. Hum. Genet. 2008, 72, 696–714.
32. Naganathan, V.; Macgregor, A.; Snieder, H.; Nguyen, T.; Spector, T.; Sambrook, P. Gender differences in the genetic factors responsible for variation in bone density and ultrasound. J. Bone Miner. Res. 2002, 17, 725–733.
33. Corsi, K.A.; Pollett, J.B.; Phillippi, J.A.; Usas, A.; Li, G.; Huard, J. Osteogenic potential of postnatal skeletal muscle-derived stem cells is influenced by donor sex. J. Bone Miner. Res. 2007, 22, 1592–1602.
34. Tosi, L.L.; Boyan, B.D.; Boskey, A.L. Does sex matter in musculoskeletal health? The influence of sex and gender on musculoskeletal health. J. Bone Joint Surg. Am. 2005, 87, 1631–1647.
35. Van Wijnen, A.J.; van de Peppel, J.; van Leeuwen, J.P.; Lian, J.B.; Stein, G.S.; Westendorf, J.J.; Oursler, M.J.; Im, H.J.; Taipaleenmaki, H.; Hesse, E.; et al. MicroRNA functions in osteogenesis and dysfunctions in osteoporosis. Curr. Osteoporos Rep. 2013, 11, 72–82.
36. Zhang, Y.; Gao, Y.; Cai, L.; Li, F.; Lou, Y.; Xu, N.; Kang, Y.; Yang, H. MicroRNA-221 is

involved in the regulation of osteoporosis through regulates RUNX2 protein expression and osteoblast differentiation. Am. J. Transl. Res. 2017, 9, 126–135.

37. Sun, M.; Zhou, X.; Chen, L.; Huang, S.; Leung, V.; Wu, N.; Pan, H.; Zhen, W.; Lu, W.; Peng, S. The Regulatory Roles of MicroRNAs in Bone Remodeling and Perspectives as Biomarkers in Osteoporosis. Biomed. Res. Int. 2016, 2016, 1652417.
38. De-Ugarte, L.; Yoskovitz, G.; Balcells, S.; Guerri-Fernandez, R.; Martinez-Diaz, S.; Mellibovsky, L.; Urreizti, R.; Nogues, X.; Grinberg, D.; Garcia-Giralt, N.; et al. MiRNA profiling of whole trabecular bone: Identification of osteoporosis-related changes in MiRNAs in human hip bones. BMC Med. Genom. 2015, 8, 75.
39. Hackl, M.; Heilmeier, U.; Weilner, S.; Grillari, J. Circulating microRNAs as novel biomarkers for bone diseases—Complex signatures for multifactorial diseases? Mol. Cell. Endocrinol. 2016, 432, 83–95.
40. Seeliger, C.; Karpinski, K.; Haug, A.T.; Vester, H.; Schmitt, A.; Bauer, J.S.; van Griensven, M. Five freely circulating miRNAs and bone tissue miRNAs are associated with osteoporotic fractures. J. Bone Miner. Res. 2014, 29, 1718–1728.
41. De-Ugarte, L.; Caro-Molina, E.; Rodriguez-Sanz, M.; Garcia-Perez, M.A.; Olmos, J.M.; Sosa-Henriquez, M.; Perez-Cano, R.; Gomez-Alonso, C.; Del Rio, L.; Mateo-Agudo, J.; et al. SNPs in bone-related miRNAs are associated with the osteoporotic phenotype. Sci. Rep. 2017, 7, 516.
42. Kalogeropoulos, M.; Varanasi, S.S.; Olstad, O.K.; Sanderson, P.; Gautvik, V.T.; Reppe, S.; Francis, R.M.; Gautvik, K.M.; Birch, M.A.; Datta, H.K. Zic1 transcription factor in bone: Neural developmental protein regulates mechanotransduction in osteocytes. FASEB J. 2010, 24, 2893–2903.
43. Varanasi, S.S.; Olstad, O.K.; Swan, D.C.; Sanderson, P.; Gautvik, V.T.; Reppe, S.; Francis, R.M.; Gautvik, K.M.; Datta, H.K. Skeletal site-related variation in human trabecular bone transcriptome and signaling. PLoS ONE 2010, 5, e10692.
44. Hoffman, A.E.; Zheng, T.; Yi, C.; Leaderer, D.; Weidhaas, J.; Slack, F.; Zhang, Y.; Paranjape, T.; Zhu, Y. microRNA miR-196a-2 and breast cancer: A genetic and epigenetic association study and functional analysis. Cancer Res. 2009, 69, 5970–5977.
45. Song, Z.S.; Wu, Y.; Zhao, H.G.; Liu, C.X.; Cai, H.Y.; Guo, B.Z.; Xie, Y.A.; Shi, H.R. Association between the rs11614913 variant of miRNA-196a-2 and the risk of epithelial ovarian cancer. Oncol. Lett. 2016, 11, 194–200.
46. Haas, U.; Sczakiel, G.; Laufer, S.D. microRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3′-UTR via altered RNA structure. RNA Biol. 2012, 9, 924–937.
47. Mahen, E.M.; Watson, P.Y.; Cottrell, J.W.; Fedor, M.J. mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS Biol. 2010, 8, e1000307.
48. Vinci, S.; Gelmini, S.; Pratesi, N.; Conti, S.; Malentacchi, F.; Simi, L.; Pazzagli, M.; Orlando, C. Genetic variants in miR-146a, miR-149, miR-196a2, miR-499 and their influence on relative expression in lung cancers. Clin. Chem. Lab. Med. 2011, 49, 2073–2080.
49. Xu, J.; Hu, Z.; Xu, Z.; Gu, H.; Yi, L.; Cao, H.; Chen, J.; Tian, T.; Liang, J.; Lin, Y.; et al. Functional variant in microRNA-196a2 contributes to the susceptibility of congenital heart disease in a Chinese population. Hum. Mutat. 2009, 30, 1231–1236.
50. Sun, M.; Liu, X.H.; Li, J.H.; Yang, J.S.; Zhang, E.B.; Yin, D.D.; Liu, Z.L.; Zhou, J.; Ding, Y.; Li, S.Q.; et al. miR-196a is upregulated in gastric cancer and promotes cell proliferation by downregulating p27(kip1). Mol. Cancer Ther. 2012, 11, 842–852.
51. Zhi, H.; Wang, L.; Ma, G.; Ye, X.; Yu, X.; Zhu, Y.; Zhang, Y.; Zhang, J.; Wang, B. Polymorphisms of miRNAs genes are associated

with the risk and prognosis of coronary artery disease. Clin. Res. Cardiol. 2012, 101, 289–296.
52. Zhou, B.; Rao, L.; Peng, Y.; Wang, Y.; Chen, Y.; Song, Y.; Zhang, L. Common genetic polymorphisms in pre-microRNAs were associated with increased risk of dilated cardiomyopathy. Clin. Chim. Acta 2010, 411, 1287–1290.
53. Su, Y.M.; Li, J.; Guo, Y.F.; Cai, F.; Cai, X.X.; Pan, H.Y.; Deng, X.T.; Pan, M. A Functional single-nucleotide polymorphism in pre-microRNA-196a2 is associated with atrial fibrillation in han chinese. Clin. Lab. 2015, 61, 1179–1185.
54. Ghanbari, M.; Sedaghat, S.; de Looper, H.W.; Hofman, A.; Erkeland, S.J.; Franco, O.H.; Dehghan, A. The association of common polymorphisms in miR-196a2 with waist to hip ratio and miR-1908 with serum lipid and glucose. Obesity (Silver Spring) 2015, 23, 495–503. Yekta, S.; Shih, I.H.; Bartel, D.P. microRNA-directed cleavage of HOXB8 mRNA. Science 2004, 304, 594–596.
56. Alexander, T.; Nolte, C.; Krumlauf, R. Hox genes and segmentation of the hindbrain and axial skeleton. Annu. Rev. Cell Dev. Biol. 2009, 25, 431–456.
57. Zakany, J.; Duboule, D. The role of Hox genes during vertebrate limb development. Curr. Opin. Genet. Dev. 2007, 17, 359–366.
58. Mori, M.; Nakagami, H.; Rodriguez-Araujo, G.; Nimura, K.; Kaneda, Y. Essential role for miR-196a in brown adipogenesis of white fat progenitor cells. PLoS Biol. 2012, 10, e1001314.
59. Kim, Y.J.; Bae, S.W.; Yu, S.S.; Bae, Y.C.; Jung, J.S. miR-196a regulates proliferation and osteogenic differentiation in mesenchymal stem cells derived from human adipose tissue. J. Bone Miner. Res. 2009, 24, 816–825.
60. Lei, S.F.; Papasian, C.J.; Deng, H.W. Polymorphisms in predicted miRNA binding sites and osteoporosis. J. Bone Miner. Res. 2011, 26, 72–78.
61. Kung, A.W.; Xiao, S.M.; Cherny, S.; Li, G.H.; Gao, Y.; Tso, G.; Lau, K.S.; Luk, K.D.; Liu, J.M.; Cui, B.; et al. Association of JAG1 with bone mineral density and osteoporotic fractures: A genome-wide association study and follow-up replication studies. Am. J. Hum. Genet. 2010, 86, 229–239.
62. Kiselev, I.; Bashinskaya, V.; Kulakova, O.; Baulina, N.; Popova, E.; Boyko, A.; Favorova, O. Variants of microRNA genes: Gender-specific associations with multiple sclerosis risk and severity. Int. J. Mol. Sci. 2015, 16, 20067–20081.
63. Hearn, A.P.; Silber, E. Osteoporosis in multiple sclerosis. Mult. Scler. J. 2010, 16, 1031–1043.
64. Sioka, C.; Kyritsis, A.P.; Fotopoulos, A. Multiple sclerosis, osteoporosis, and vitamin D. J. Neurol. Sci. 2009, 287, 1–6.
65. Gibson, J.C.; Summers, G.D. Bone health in multiple sclerosis. Osteoporos Int. 2011, 22, 2935–2949.
66. Estrada, K.; Styrkarsdottir, U.; Evangelou, E.; Hsu, Y.H.; Duncan, E.L.; Ntzani, E.E.; Oei, L.; Albagha, O.M.; Amin, N.; Kemp, J.P.; et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat. Genet. 2012, 44, 491–501.
67. Arnold, M.; Ellwanger, D.C.; Hartsperger, M.L.; Pfeufer, A.; Stumpflen, V. Cis-acting polymorphisms affect complex traits through modifications of microRNA regulation pathways. PLoS ONE 2012, 7, e36694.
68. Ghanbari, M.; Ikram, M.A.; de Looper, H.W.; Hofman, A.; Erkeland, S.J.; Franco, O.H.; Dehghan, A. Genome-wide identification of microRNA-related variants associated with risk of Alzheimer's disease. Sci. Rep. 2016, 6, 28387.
69. Agarwal, V.; Bell, G.W.; Nam, J.W.; Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. Elife 2015, 4, e05005.

70. Chou, C.H.; Chang, N.W.; Shrestha, S.; Hsu, S.D.; Lin, Y.L.; Lee, W.H.; Yang, C.D.; Hong, H.C.; Wei, T.Y.; Tu, S.J.; et al. miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. 2016, 44, D239–D247.
71. Hofacker, I.L. Vienna RNA secondary structure server. Nucleic Acids Res. 2003, 31, 3429–3431.
72. Will, S.; Jabbari, H. Sparse RNA folding revisited: Space-efficient minimum free energy structure prediction. Algorithms Mol. Biol. 2016, 11, 7.
73. Ikram, M.A.; Brusselle, G.G.O.; Murad, S.D.; van Duijn, C.M.; Franco, O.H.; Goedegebure, A.; Klaver, C.C.W.; Nijsten, T.E.C.; Peeters, R.P.; Stricker, B.H.; et al. The Rotterdam Study: 2018 update on objectives, design and main results. Eur. J. Epidemiol. 2017, 32, 807–850.
74. Watts, N.B.; Adler, R.A.; Bilezikian, J.P.; Drake, M.T.; Eastell, R.; Orwoll, E.S.; Finkelstein, J.S.; Endocrine, S. Osteoporosis in men: An Endocrine Society clinical practice guideline. J. Clin. Endocrinol. Metab. 2012, 97, 1802–1822.



Chapter 5

General discussion

5.1 MAIN FINDINGS AND INTERPRETATIONS

The overall aim of this thesis is to investigate epigenetic mechanisms as a possible biomarker for disease risk, as a possible mediator between lifestyle factors and disease risk, and for inferring lifestyle factors from human materials.

I investigated the inference of alcohol consumption and smoking habits from DNA methylation patterns. Also, I studied if smoking-related changes in DNA methylation and gene expression patterns are associated with each other and with cardio-metabolic traits in a three-way association. Using multi-omics data, I studied miRNAs that could play a role in cardio-metabolic traits and bone mineral density. In this chapter, the main findings of these studies are summarized, the main methodological considerations are discussed, and future perspectives are provided.

5.1.1 DNA methylation-based lifestyle inference

Negative lifestyle factors/variables are modifiable habits that are associated with diseases risk. Lifestyle information from patients or study participants is often obtained using self-reported questionnaires or during interviews, which are both error-prone. DNA methylation-based prediction models are proposed as an alternative to complement or overcome the use of self-reported questionnaires to obtain lifestyle information [1-11]. So far, developed DNA methylation-based prediction models suffer from statistical and population limitations [12]. **Chapter 2** of this thesis focuses on the development and/or validation of DNA methylation-based prediction models for alcohol consumption (**Chapter 2.1**) and smoking habits (**Chapter 2.2**). In **Chapter 2.1**, I validated the DNA methylation-based prediction models for alcohol consumption published by Liu *et al.* [2]. In the study by Liu *et al.*, participants were categorized into four alcohol consumption categories, of which subsets (2 or 3 categories) were inferred using seven prediction models implementing four DNA methylation marker sets. The model validation in four independent studies used unique marker-weights in each study by employing the same dataset for model training and testing, resulting in impressively high AUCs in both model building and model validation. In the study implemented in **Chapter 2.1**, I validated the four markers sets (5, 23, 78, and 144 CpGs) and the seven prediction models. By employing the same methods, of training and testing the models in the same dataset, I obtained very similar results compared to the original study. Nevertheless, I argued that using unique weights for each CpG and model per validation cohort rather than using the marker-weights obtained in model building overestimates the true prediction accuracy. In this line, I tested the transportability of the models using data independent from our model building dataset to test the prediction accuracies. Here, I obtained much lower AUCs compared to the originals study. Also, I obtained a high variance in AUCs between the different external validation datasets. The results obtained in this study

suggest that the previously published prediction markers are unreliable or provide only low accuracies. Therefore, the markers are not yet suitable for the accurate prediction of alcohol consumption, indicating that the previously published AUCs by Liu *et al.* are strongly overestimated.

In **Chapter 2.2**, I aimed to develop a robust prediction model for smoking status. Several studies have already shown the possibility to develop a prediction model for smoking status obtaining high AUCs, all with limitations [3-10]. Here, I selected 13 CpGs that were independently replicated and resulted in the highest prediction accuracy (AUC=0.90) to distinguish smokers from non-smokers (former and never smokers combined) in our model building dataset. Using internal validation and external validation, I obtained very similar results of 0.90±0.14 and 0.91, respectively. I used an arbitrary probability cut-off of 0.50 to categorize smokers, which resulted in the model building dataset in a sensitivity of 0.59 and a specificity of 0.98. The sensitivity of a prediction model reflects the true positive cases, while the specificity represents the true negative cases. The sensitivity and specificity of a binomial prediction model depend on a cut-off point (0.5 in our smoking prediction model). Therefore, they inversely change when using another cut-off, e.g., when the sensitivity increases, the specificity will decrease. The AUC reflects the overall model accuracy and provides a better understanding of how well the model performs [13-15].

Several other studies have investigated DNA methylation-based prediction of smoking status, all with low sample size or the exclusion of smoking categories from their models [3-6, 9, 10]. The top smoking-related CpG, cg05575921 (*AHRR*), was used in several previous smoking predictors [3, 5, 6, 8, 10]. For instance, Philibert *et al.* [5] obtained an AUC of 0.99 using only the methylation levels of this specific CpG (cg05575921; *AHRR*) employing 35 non-smokers and 26 smokers. The sole use of cg05575921 (*AHRR*), provided in our model building dataset an AUC of 0.88 distinguishing smokers (N=511) from non-smokers (3,764). These results support our original hypothesis that using a small sample size could provide highly overestimated prediction accuracies.

A more recent study by Sugden *et al.* [8] used 2,623 CpG previously identified by Joehanes *et al.* [1], obtaining an AUC of 0.77 for distinguishing never from ever smokers in participants at age 38. Although this study did include all smoking categories, the use of 2,623 CpGs risks overfitting of the models and missing values of one of the markers is likely to occur. Also, distinguishing smokers from non-smokers, as done in our model (AUC= 0.90), might provide more clinically relevant information. The similar results obtained between our model building dataset and the external validation shows the accuracy and robustness of our model. The use of all smoking categories provides a broad range of possible applications for our model, including the general population.

In summary, the results in **Chapter 2.1** and **Chapter 2.2** demonstrate the possibility of using DNA methylation markers (CpG sites) to develop prediction models for lifestyle

factors that are transportable to independent datasets. However, it also shows the importance of using replicated predictive markers, a large diverse model building dataset, and of implementing strict internal and external validation methods. Specifically, for the models to be transportable to independent datasets, it would be important to test the replication of the markers in independent data. The combination of markers that have been independently replicated in several studies will likely provide a more robust and transportable model compared to a model including markers discovered in one study. In this line, Liu *et al.* [2] selected their predictive markers from only one meta-analysis, which resulted in a non-transportable model. In contrast, I selected independently replicated predictive markers using data from 14 EWASs [3, 16-27] (**Chapter 2.2**), obtaining a model that provided similar AUCs in model building, internal validation, and external validation. I would, therefore, suggest using the 13 population-based prospective cohorts embedded in the study by Liu *et al.* [2] to select the CpGs that (suggestively) replicate in multiple studies. The independent replication in several studies will provide a finite but more robust marker set than the inclusion of CpGs that reach (suggestive) significance in one large meta-analysis. Also, a model that includes too many predictive markers compared to the number of observations is more likely to be over-fitted [28], which likely partially explains the poor prediction results obtained when applying the 78 and 144-CpG alcohol models to independent datasets (**Chapter 2.1**). Finally, the developed models by Liu *et al.* only included subsets of the alcohol categories, which are difficult to extrapolate to the general population. Similarly, category subsets have been used before in papers developing alcohol consumption and smoking status prediction models [7, 10, 11]. Studies should focus on models that would be beneficial in a wide range of settings, as was done in **Chapter 2.2**. Therefore, future studies are needed to develop new alcohol consumption prediction models using the correct statistical methods and all available categories and study participants [12].

In addition, future studies would be needed to test the generalizability of our smoking prediction models in participants of non-European ancestry. Also, by using the same 13 CpGs, I developed as first a model that can predict lifetime smoking information, including pack-years in current smokers, smoking cessation in former smokers, and the never smokers. Although I obtained very interesting results for this lifetime smoking model, further research would be needed using a much larger sample. Due to the use of five categories, only a limited number of participants were available in the different pack-year categories.

Most large cohort studies already collect blood samples from their participants, which could be used for DNA methylation measurements. Hence, I believe our model will especially be useful in these studies due to easy access. This emphasizes the possible importance for future development of DNA methylation-based prediction models for other lifestyle factors. In this thesis, I specifically focus on smoking habits and alcohol

consumption; however, it is important to note that several other lifestyle factors are also associated with disease risk [29]. For example, of all DALYs in 2019, 11.9% (95% CI 9.6–14.5) is attributed to the combined burden of diet quality, physical inactivity, and high BMI. This indicates how important diet and physical activity is for the current health state and future disease risk. For these lifestyle factors and for many more, questionnaires are most often used for data collection. It would also be important to develop alternative methods to collect this information. For instance, information on physical activity can be obtained via specific questionnaires (e.g. LASA Physical Activity Questionnaire (LAPAQ) [30]) or via an accelerometer. In the Rotterdam study, it was shown that overall total physical activity was underestimated in the questionnaire compared to the use of a triaxial accelerometer (GeneActiv; ActivinsightsLtd, Kimbolton, UK) [31]. These results show the importance of objective measurements, also in protective lifestyle factors. Future studies are needed to investigate the impact of other lifestyle factors on disease risk and DNA methylation and the possibility to infer these factors using DNA methylation patterns.

5.1.2 Smoking, epigenetics, and cardio-metabolic traits

In **Chapter 3**, I investigated if smoking-related alterations in DNA methylation and gene expression are associated with each other and with cardio-metabolic traits. I selected smoking-related CpGs and genes previously found to be associated with current *vs.* never smokers in the largest EWAS and TWAS to date [1, 32]. The use of previously identified smoking-related CpGs and genes provided a broader starting point compared studies that start with new marker-identification in a smaller sample size. For instance, a previous study only identified 42 smoking-related CpGs and genes in 542 subjects [33]. Using previously identified markers, I was able to select 2,549 smoking-CpGs and 443 smoking-gene expression probes as candidate markers. This made it possible to have a broader look into the expression quantitative trait methylation (eQTM) associations between these smoking-related CpGs and genes. In most studies, the eQTM associations are solely tested for CpGs located in the promoter, enhancer, the transcription start site, or the gene body of the CpG annotated gene, as this is associated with altered gene expression [34]. I tested the association between all smoking-CpGs with smoking-genes and showed several significant *cis*- and *trans*-eQTM associations.

For the eQTM associated CpGs and genes, I tested the association with cardio-metabolic traits and found 26 smoking-related CpGs and 19 smoking-related genes (21 probes) associated with a cardio-metabolic trait. Next, I looked for an overlap in the obtained results, indicating a three-way association in which a CpG is associated with a gene and both are associated with the same trait. I found for triglycerides two associations and for BMI I found several three-way associations. Several of these markers were previously identified to be associated with CVD-related outcomes.

In summary, by combining DNA methylation and gene expression data, I was able to identify several three-way associations for CpGs and genes that have previously been associated with CVD-related risk factors and death. Our results suggest that smoking-related changes in DNA methylation and gene expression are important molecular pathways in which smoking can affect cardio-metabolic traits. Nevertheless, future studies are needed to validate the results obtained in our study. Specifically, I identified several *cis*- and *trans*-eQTM associations; however, only 134 out of the 1,224 CpGs and 50 out of the 200 gene-probes passed the Bonferroni corrected significance threshold in independent data. In both discovery and replication datasets, around 700 participants were available with both gene expression and DNA methylation data. Future studies would benefit from a larger sample size, possibly via meta-analysis followed by independent replication. Due to the identification of several *trans*-eQTMs, I would propose a broadening of the current approach of investigating only the *cis*-regulatory effect. I believe that this approach might help to identify unknown molecular paths that are currently being ignored. Studies should focus on investigating the direction and causality of these eQTM associations to determine if the alterations in DNA methylation levels induce the changes in gene expression or vice versa. Also, it would be important to investigate further if the identified *trans*-eQTM associations are indeed *trans*-regulatory effects or possibly reflect a downstream effect. Finally, functional studies are needed to verify the observed impact of smoking on these CpGs and genes and to validate their downstream effect on cardio-metabolic traits.

5.1.3 Disease-related miRNAs

5.1.3.1 SNPs in miRNA-related sequences

In **Chapter 4**, I used summary statistics from large-scale GWAS to identify genetic variants in miRNA-related sequences in association with cardio-metabolic traits [35-45] (**Chapter 4.1**) and bone mineral density (BMD) traits (femoral neck, lumbar spine, and forearm BMD) [46] (**Chapter 4.2**). The use of publicly available GWAS strongly increased our study power to identify disease-related SNPs in miRNA-related sequences. However, mutations in miRNA-related sequences are rare and possibly population-specific and are therefore not all present in available GWAS summary statistics. This was also evident in the studies included in this thesis. In **Chapter 4.1**, I selected 23,990 SNPs in miRNA-related sequences (within +/- 2kb of primary miRNAs sequences). Only 2,358 SNPs were present in HapMap imputed GWAS statistics and 8,652 in the 1000 Genomes consortia imputed GWAS statistics [35-45]. In **Chapter 4.2**, I selected 2,340 variants within primary and mature miRNAs sequences. Only 90 SNPs were available in the BMD-GWAS data, which used a novel imputation reference panel generated by the UK10K and 1000 Genomes consortia [46-48]. The comprehensiveness and accuracy of genetic imputations are dependent on the reference information. Large efforts have already been done

to gain extensive haplotype information in several populations [47-49]. However, the occurrence of these haplotypes can vary between populations or might be population-specific. Thus, larger collaborations (trans-ethnic studies) and denser genotyping (such as TopMed reference panel) are needed to cover more sequences in the genome from different populations to improve the human haplotype map, and thereby the quality and accuracy of imputation. Although this limitation, the statistical power that comes with the inclusion of thousands of participants in each GWAS makes our approach a good initial step for marker selection to identify important miRNA-related SNPs associated with diseases.

5.1.3.2 MiRNAs in association with cardio-metabolic traits

Using these GWAS summary statistics data, I identified 180 SNPs annotated to 67 miRNAs in association with at least one cardio-metabolic trait (**Chapter 4.1**). When testing the DNA methylation of annotated CpGs and the expression of these miRNAs, I identified five miRNAs (including miR-10b-5p, miR-148a-3p, miR-100-5p, miR-125b-5p, and miR-6886), which had at least one annotated CpG and of which the expression was associated with the same group of cardio-metabolic traits in blood. Some of the identified miRNAs were previously identified in association with a cardio-metabolic related trait.

5.1.3.3 Disease-related DNA variation in precursor miR-196a-2

CVD and osteoporosis are both complex age-related disorders and share some common pathogenic mechanisms [50]. In this line, I identified the same mutation (rs11614913:C > T) in the precursor (pre-) miR-196a-2 positively associated with waist to hip ratio (WHR) (**Chapter 4.1**) and negatively associated with femoral neck-BMD and lumbar spine-BMD (**Chapter 4.2**). This would indicate that minor Allele T carriers have a lower BMD and a higher WHR. Overweight and obesity are protective for aging-related bone loss and a low BMI is associated with an increased risk for future fracture [51]. However, recent evidence suggests that an increase in abdominal obesity is associated with an increase in hip fracture [52, 53]. Our results possibly suggest that the identified SNP (rs11614913) might be involved in this WHR-related increased risk for osteoporosis. It is important to note that the associations of rs11614913 with WHR and BMD are obtained in cross-sectional data analysis. Therefore, it is impossible to determine any causality from these results. Future studies are needed to investigate if rs11614913 affects both WHR and BMD independently or if its association with BMD is a possible downstream effect of the increased WHR.

Rs11614913 is located in the stem region of pre-miR-196a-2 and is shown to affect the miRNA processing and subsequently alter the expression of mature miR-196a-2 [54, 55]. Also, rs11614913 has been reported to affect the binding of the mature miR-196a-2 to its target mRNA [55, 56]. One of these experimentally validated target genes is *Heme Oxy-*

genase 1 (*HMOX1)*. I identified a positive association between the expression of *HMOX1* and HDL and a negative association with triglycerides in our discovery dataset (**Chapter 3.1**). This same gene was previously identified with mesenchymal stem cell differentiation into osteoblasts, osteoclastogenesis, and bone resorption [57-59]. Studies have shown a positive association between BMD and triglycerides levels and a negative association with HDL levels [60, 61]. In addition, a large systematic review shows a positive association between total cholesterol and risk of bone fracture and that individuals with a decreased level of HDL (<40 mg/dL) have a lower risk of bone fracture compared with those with a normal level [62]. Future studies are needed to investigate miR-196a-2 as a potential regulator linking alterations in BMD and lipid profile. It would be important to further validate *HMOX1* as target gene of miR-196a-2 in both blood tissue and bone tissue and its regulatory effect in both BMD and lipid levels.

In summary, these results might indicate that the identified pre-miR-196a-2 plays an important role in the underlying mechanisms linking CVD and osteoporosis. Also, our results provide further evidence of the interaction between multi-omics layers in disease pathology, showing the importance of expanding the current approaches in disease studies of the sole use of one omics-layer. Finally, our results show that the publicly available GWAS data could be used as an important starting point for miRNA selection and should be further explored in future studies, possibly identifying more important miRNAs involved in different diseases.

5.2 METHODOLOGICAL CONSIDERATIONS

5.2.1 Study population

The studies in this thesis were embedded in large population-based cohorts. The specific study design of cohort studies provides the availability to a wide range of exposure variables and disease outcomes over a long follow-up period. This makes it possible to investigate risk factors for disease in both a cross-sectional and longitudinal manner. Also, with the inclusion of younger participants, the long-follow up period makes it possible to investigate causal factors in age-related diseases. However, this design is also subjected to selection bias in which the correlation between exposure and outcome varies between the participants included in the study and those that were eligible for the study [63]. For example, loss-to-follow-up bias can occur due to the long follow-up period and non-response bias when non-participants vary from participants in an essential manner. In this case, more healthy individuals will often participate in this type of studies compared to individuals with a harmful lifestyle [63-65].

In this thesis, I investigated alterations in DNA methylation levels and gene expression, including protein-coding genes and miRNAs, in cross-sectional associations.

This approach restricts us from defining any causal relationships between the studied variables. Moreover, the dynamic epigenome is influenced by both external and internal effects. Although the large efforts made to correct for confounding effects, as in any observational study, it is impossible to rule out residual confounding. The cumulative environmental exposure influence epigenetic markers, which possibly affect gene expression, and can contribute to the onset of complex diseases. The prospective design of cohort studies, as included in this thesis, provides large-scale environmental exposure data, which would make it possible to identify risk factors before the occurrence of the event of interest. Unfortunately, most cohort studies have limited DNA methylation and gene expression data and most often without follow-up measurements. Repeated measurements at different time points of DNA methylation and gene expression levels may provide further evidence if DNA methylation affect gene expression or vice versa. Similarly, repeated measurements of DNA methylation and gene expression levels could possibly provide insights into their involvement in the lifestyle-related health risk.

5.2.2 Epigenetics and gene expression tissue of choice

Genetic data is coherent across tissues, though, both epigenetic and transcriptomic data are tissue-specific [66-68]. As blood is more feasible to collect from study participants, it is the most often collected tissue. It is, therefore, used for DNA methylation and gene expression profiling in large cohort studies. Whole blood contains a mixture of red blood cells, white blood cells, and platelets. Of these, only the white blood cells, the leukocytes, have a nucleus and therefore DNA. A whole blood sample varies in leukocyte proportions, which have a different function and thus a different epigenetic pattern depending on the sample donor. While testing epigenetic associations, it is, therefore, important to correct for these cell-type proportions; otherwise, it might be possible that the observed associations actually reflect cell-type differences instead of the epigenetic alterations [69].

I used genetic, epigenetic, and transcriptomic data obtained from blood in the studies in this thesis. When I investigated the association between disease and epigenetic alterations, I adjusted the models for cell count measurement or houseman predicted cell composition [70]. In **Chapter 2.1**, I validated the previously used methods by Liu *et al.* [2] that included, among others, the correction for cell count. In **Chapter 2.2**, I developed an easy-to-access prediction model for smoking status in blood. As smoking is also associated with changes in cell count [71-73], I included a sensitivity analysis to test the impact of cell count in smoking inference. I obtained a slight increase in prediction AUC from 0.906 to 0.907. This suggests that for smoking inference, cell count might not be as important as for association analysis. In **Chapter 4.1**, I used DNA methylation and miRNA expression in blood to investigate possible miRNAs associated with cardiometabolic traits. The detection of these markers in blood could subsequently lead to

miRNAs as biomarkers for early disease diagnosis. In both the biomarker potential and the miRNA target prediction, blood might be the most promising tissue as it is a non-invasive method making the collection of a test sample feasible. Nevertheless, while investigating possible disease-related pathways, the associations obtained in blood samples should be further examined in relevant cells or disease-related tissues to validate the true mechanisms involved.

5.2.3 Data measurements and analysis in prediction modeling and association studies

More participants with more data and more measurements seem to be the future to answer most health-related research questions. However, the sole use of large datasets will not automatically provide robust associations that will be applicable in every population. In this line, the use of markers that have been identified in one large study but without independent replication does not automatically result in robust prediction markers (**Chapter 2.1**). Therefore, it is important that associations and prediction models are being replicated and validated using independent data. Several large studies have been conducted identifying CpGs with a large range of outcomes. Unfortunately, studies often only provide the results that reached Bonferroni corrected significance in the replication phase. This limits the possibility using the summary statistics in downstream analysis, meta-analysis, and for studies to look up their hits for independent replication. Another limitation in replication and validation is the between-study variability. Different cohorts have different methods for variable definitions, calculation, and measurements. For example, the use of population-specific food frequency questionnaires (FFQ), array data normalization, and measured *vs.* predicted white blood cell count.

5.2.4 Lifestyle factor information collection and implementation

5.2.4.1 Lifestyle factor assessment

Lifestyle factors are known to be associated with health outcomes and are, therefore, often studied as main exposure or to control for confounding effects. Due to the comprehensive data collection in large studies, lifestyle information is most often collected using self-reported questionnaires as they are cheap and fast. Unfortunately, they are also prone to underestimation of the true exposure when it comes to negative lifestyle factors, causing information bias [63, 74, 75]. Self-reporting bias is a crucial problem in the assessment of most observational research study designs. Bias can arise due to several reasons, including the recall period, selective recall, sampling approach, and social desirability [75]. Specifically, participants might not remember the true exposure (recall bias) or perhaps deliberately underestimate due to its socially stigmatized nature (social desirability). Therefore, misclassification and/or underestimation of the true exposure might influence the effect estimate [75].

5.2.4.2 Smoking assessment and implementation

Smoking is the best-studied lifestyle factor as it is a major modifiable risk factor for several diseases [76]. As there is no standardized smoking categorization, different definitions for smoking exposure are used. For instance, smokers *vs.* never-smokers is often used as exposure variable, as this results in the largest effect estimate. Moreover, sometimes ever smokers (current and former smokers) are tested against never smokers or current smokers *vs.* non-smokers (former and never smokers). In addition, long-term heavy smoking has a stronger effect on health and epigenetic markers compared to social smoking; nonetheless, both participants will be categorized as smokers, which results in a wide variation of smoking effect [4]. Using pack-years for current smokers could be a solution; however, the formula used to calculate pack-years also suffers from limitations, as it does not consider the variability in the number of cigarettes smoked per day over the years and the difference in the tobacco content per cigarette.

5.2.4.3 Alcohol consumption assessment and implementation

Alcohol consumption information is most often acquired using food frequency questionnaires (FFQ), in which participants provide the average consumption of alcohol beverages per time-period. Most studies use alcohol as a continuous variable as alcohol consumption in grams/day, or alcohol consumption categories derived from this continuous variable. A standardized method can be used to obtain the alcohol consumption in grams/day variable from the FFQ data, in which it is assumed that every drink contains the same amount of alcohol when the appropriate glass is used. A more elaborate translation of the FFQ information is via the use of a food composition table. The use of different calculation methods for the continuous variable leads to a variation between studies. In addition, this could lead to misclassification during categorization and could affect the external validity, the generalizability of the obtained results.

5.2.4.5 Underestimation of lifestyle factors

A main issue of data collection using self-reported questionnaires and interviews is the underreporting of undesired behaviors and the overreporting of desired behaviors. For instance, a smoker is more likely to report the use of ten cigarettes per day, while it is in reality 20 cigarettes, compared to reporting to be a non-smoker. This will likely not affect the epigenetic inference of smoking status categorized as current, former, and never smokers. However, this might affect pack-year calculations and subsequently in-accurate inference. Alcohol consumption is often used as a continuous variable; therefore, quantitative underestimation will greatly impact the inner-variable variance. In addition, underestimation of the alcohol consumption might lead to misclassification of the true alcohol consumption category.

It is important to note that, in addition to smoking habits and alcohol consumption, most lifestyle factors are associated with disease risk [29]. Unfortunate, information regarding these lifestyle factors is also often obtained using questionnaires. It would be important to investigate the possible underestimation of these factors and subsequently develop alternative data collection methods.

5.3 POTENTIAL IMPLICATIONS AND FUTURE DIRECTIONS

5.3.1 Early life follow-up for disease biomarker discovery

Non-communicable diseases (NCDs) are the leading cause of death worldwide, contributing to 73.4% (95% CI 72.5–74.1) of the total deaths in 2017 [77]. Although extensive research has been done on NCDs, incidence rates keep rising worldwide [77]. A large number of population-based cohort studies have been initiated to investigate the underlying mechanisms of these diseases. Most studies are focused on the middle-aged and elderly population, but more recently also the number of birth cohorts increase. New markers are needed to be able to detect participants at risk for diseases, preferably during the pre-disease stage or even at the healthy stage. At that point, patients with a high-risk profile might still be able to make lifestyle changes to avoid disease onset, with that enhance their quality of life, and reduce their economic burden and its social impact. The possibility of following participants from the pre-natal stage throughout life will provide unique opportunities to detect early markers leading to disease in later life. Also, the inclusion of new cohorts with younger participants at baseline in existing population-based studies will open the possibility to follow participants during the healthy stage, the pre-disease stage until disease onset, and thereafter.

5.3.2 Underlying molecular mechanism involved in non-communicable diseases

It is well known that most NCDs have an underlying genetic compound. Large efforts have been done so far in thousands of individuals providing several disease-related genetic mutations. However, the total variance explained in most complex diseases is still limited, and there is a substantial missing heritability of complex diseases yet to be discovered. In addition, several mutations identified by GWAS are silent mutations or non-functional, e.g. they have no impact on the protein sequence, providing evidence that other functional variants (LD proxies) or mechanisms play a role in disease pathology. In both published work and as presented in this thesis, evidence shows the importance of epigenetics as a molecular mechanism underlying disease pathology.

Unlike genetics, epigenetic modifications are dynamic and can change over time due to environmental exposure or disease state. Although most cohort studies already

include epigenetic data of participants, this data is most often only present in a small sample size, at a single time point, or with a limited follow-up time. Extending this data by including epigenetic and transcriptomic data from the same participants over a long follow-up time could make it possible to identify alterations useful as disease biomarkers. For example, it would be possible to identify epigenetic markers that already show alterations while patients do not yet suffer any disease-related symptoms. This information could, after thorough validation, in the future be used in clinical practice together with standard health biomarkers to provide disease risk assessments to patients. The combination of genetic and epigenetic profiling in personalized medicine might be the future in clinical practice to provide patients disease-risk assessments and possibly provide a more personalized warning for future events.

5.3.3 Epigenetics in clinical practice

The increasing evidence showing the important influence of epigenetics in disease pathology also raises the hypothesis that these markers could be used to treat the disease. Although the large number of CpGs and miRNAs identified in association with NCDs, there are also several limitations for translating these associations to therapeutic targets that should be overcome. As mentioned above, most studies have been conducted in a single time point with limited validation and without causality testing. It is important to note that the sole use of a cross-sectional epigenetic association with a disease cannot rule out reverse causation. Application of causal inference methods (Mendelian randomization) could also help confirm the direction and causality of identified DNA methylation sites and miRNAs in the development of diseases. In addition, most miRNA investigations are done in small sample size, troubling the translation to population-based data. Future studies should invest in large-scale collaborations, as is done for GWAS, to increase study power. For this, it would be important to reduce the between-study variability. It would be beneficial to expand the current knowledge regarding using different normalization and statistical methods in each cohort. Most large cohorts already use the standardized Illumina HumanMethylation450K or its successor, the Human MethylationEPIC BeadChip, for DNA methylation quantification. However, several normalization methods are available. Similarly, there are several methods available to quantify circulating miRNAs and their normalization [78]. More research is needed to study the impact of different normalization methods on the replication of the results. Another important issue is that a CpG can be broadly methylated and a miRNA expressed, which might have several adverse effects in different tissues. Therefore, total inhibition of a certain marker will have severe effects on other mechanisms than solely on the intended target. More research is needed into a tissue-specific delivery system and the identification of markers that are only active in the intended target tissue. Moreover, close collaborations between dry and wet-lab teams would make it

possible to instantly investigate the identified epigenetic markers in cell lines providing confirmative findings.

5.3.4 Lifestyle-induces alterations in epigenetic mechanisms

There is only limited knowledge at which rate epigenetic markers are affected by lifestyle factors, nor is it clear how the adaptation to a healthier lifestyle impacts these patterns. In this context, longitudinal research is needed to investigate the impact of lifestyle changes on epigenetic markers and how this subsequently reflects the impact on disease susceptibility, e.g. is the damage already done or can we alter our epigenetic health. Long-term follow-up data would provide a better insight into these mechanisms and subsequently into the epigenetic mechanisms linking lifestyle to disease risk.

5.3.5 DNA methylation-based lifestyle inference in forensic investigations

In large cohort studies, DNA methylation measurements are often available from, a subset of, the participants. However, in clinical settings or forensic cases, DNA methylation measurement using arrays is not cost-effective. Also, the obtained DNA in forensic cases is often of low quantity and/or degraded, lowering the possibility of accurate DNA methylation quantification using arrays. I have provided a robust DNA methylation-based prediction model for smoking habits using only 13 CpGs, which would be a good starting point for a laboratory tool for lifestyle inference. A specific laboratory tool that only quantifies these 13 CpGs improves cost-efficiency and increases the possibility of accurate quantification in crime scene samples. Currently, forensic applications omit age, bio-geographic ancestry, and appearance, including eye, hair, and skin color [79, 80]. Lifestyle information could help to give a more complete picture of the unknown donor of a crime scene sample that may allow tracing the suspect and thus allow forensic DNA profiling to find out if he/she is the sample donor [81]. In addition, lifestyle factor inference for forensic application could benefit from transportability to different tissues, as a blood sample is not always available in this setting. Consequently, future studies are needed to investigate the possible translation of DNA methylation array data to a technology that can handle low quality and/or quantity DNA and for model translation to different tissues.

REFERENCES

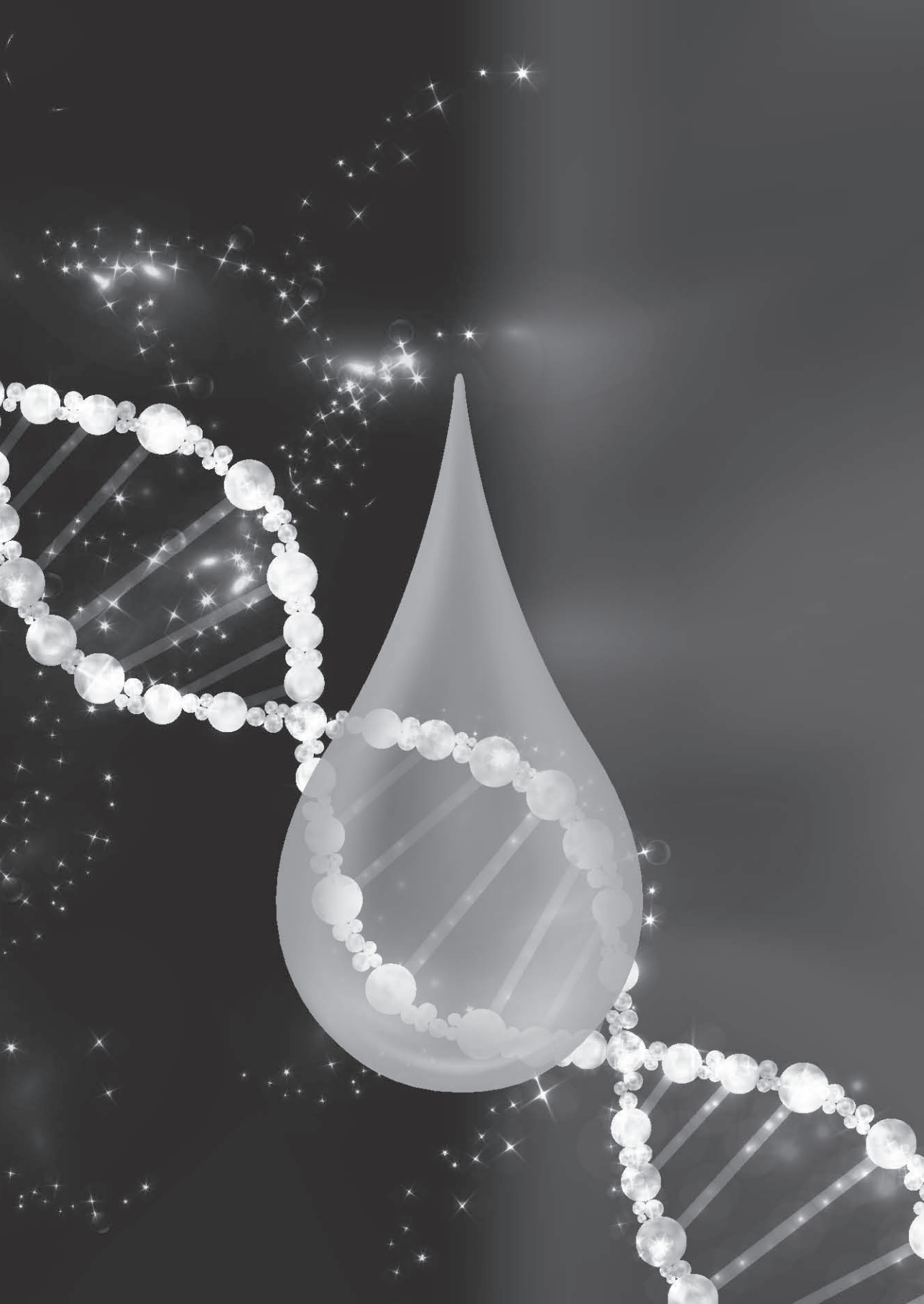
1. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;9(5):436-47.
2. Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23(2):422-33.
3. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6(1):4.
4. Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, et al. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology. 2013;24(5):712-6.
5. Philibert R, Hollenbeck N, Andersen E, Osborn T, Gerrard M, Gibbons FX, et al. A quantitative epigenetic approach for the assessment of cigarette consumption. Front Psychol. 2015;6:656.
6. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res. 2016;146:395-403.
7. Kondratyev N, Golov A, Alfimova M, Lezheiko T, Golimbet V. Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation. Clin Epigenetics. 2018;10(1):130.
8. Sugden K, Hannon EJ, Arseneault L, Belsky DW, Broadbent JM, Corcoran DL, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Transl Psychiatry. 2019;9(1):92.
9. Endo K, Li J, Nakanishi M, Asada T, Ikesue M, Goto Y, et al. Establishment of the MethyLight Assay for Assessing Aging, Cigarette Smoking, and Alcohol Consumption. Biomed Res Int. 2015;2015:451981.
10. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136.
11. Philibert R, Miller S, Noel A, Dawes K, Papworth E, Black DW, et al. A Four Marker Digital PCR Toolkit for Detecting Heavy Alcohol Consumption and the Effectiveness of Its Treatment. J Insur Med. 2019;48(1):90-102.
12. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130(6):515-24.
13. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin Biochem Rev. 2008;29 Suppl 1(Suppl 1):S83-7.
14. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978;8(4):283-98.
15. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006(27):861-74.
16. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011;88(4):450-7.
17. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22(5):843-51.
18. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One. 2013;8(5):e63812.

19. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environ Health Perspect. 2014;122(7):673-8.
20. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9(10):1382-96.
21. Allione A, Marcon F, Fiorito G, Guarrera S, Siniscalchi E, Zijno A, et al. Novel epigenetic changes unveiled by monozygotic twins discordant for smoking habits. PLoS One. 2015;10(6):e0128265.
22. Besingi W, Johansson A. Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet. 2014;23(9):2290-7.
23. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.
24. Sayols-Baixeras S, Lluís-Ganella C, Subirana I, Salas LA, Vilahur N, Corella D, et al. Corrigendum. Identification of a new locus and validation of previously reported loci showing differential methylation associated with smoking. The REGICOR study. Epigenetics. 2016;11(2):174.
25. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8(5):599-618.
26. Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2012;120(10):1425-31.
27. Zhu X, Li J, Deng S, Yu K, Liu X, Deng Q, et al. Genome-Wide Analysis of DNA Methylation and Cigarette Smoking in a Chinese Population. Environ Health Perspect. 2016;124(7):966-73.
28. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
29. Collaborators GBDRF. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396(10258):1223-49.
30. Stel VS, Smit JH, Pluijm SM, Visser M, Deeg DJ, Lips P. Comparison of the LASA Physical Activity Questionnaire with a 7-day diary and pedometer. J Clin Epidemiol. 2004;57(3):252-8.
31. Koolhaas CM, van Rooij FJ, Cepeda M, Tiemeier H, Franco OH, Schoufour JD. Physical activity derived from questionnaires and wrist-worn accelerometers: comparability and the role of demographic, lifestyle, and health factors among a population-based sample of older adults. Clin Epidemiol. 2018;10:1-16.
32. Huan T, Joehanes R, Schurmann C, Schramm K, Pilling LC, Peters MJ, et al. A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. Hum Mol Genet. 2016;25(21):4611-23.
33. Tsai PC, Glastonbury CA, Eliot MN, Bollepalli S, Yet I, Castillo-Fernandez JE, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. Clin Epigenetics. 2018;10(1):126.
34. Wolffe AP, Matzke MA. Epigenetics: regulation through repression. Science. 1999;286(5439):481-6.
35. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of

body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.
36. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Magi R, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature. 2015;518(7538):187-96.
37. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012;44(6):659-69.
38. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet. 2010;42(2):142-8.
39. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes. 2011;60(10):2624-34.
40. Wheeler E, Leong A, Liu CT, Hivert MF, Strawbridge RJ, Podmore C, et al. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med. 2017;14(9):e1002383.
41. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42(2):105-16.
42. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66(11):2888-902.
43. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274-83.
44. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47(10):1121-30.
45. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9.
46. Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature. 2015;526(7571):112-7.
47. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82-90.
48. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
49. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
50. Lampropoulos CE, Papaioannou I, D'Cruz DP. Osteoporosis--a risk factor for cardiovascular disease? Nat Rev Rheumatol. 2012;8(10):587-98.
51. De Laet C, Kanis JA, Odén A, Johanson H, Johnell O, Delmas P, et al. Body mass index as a predictor of fracture risk: a meta-analysis. Osteoporos Int. 2005;16(11):1330-8.
52. Meyer HE, Willett WC, Flint AJ, Feskanich D. Abdominal obesity and hip fracture: results from the Nurses' Health Study and the Health Professionals Follow-up Study. Osteoporos Int. 2016;27(6):2127-36.

53. Søgaard AJ, Holvik K, Omsland TK, Tell GS, Dahl C, Schei B, et al. Abdominal obesity increases the risk of hip fracture. A population-based study of 43,000 women and men aged 60-79 years followed for 8 years. Cohort of Norway. J Intern Med. 2015;277(3):306-17.
54. Song ZS, Wu Y, Zhao HG, Liu CX, Cai HY, Guo BZ, et al. Association between the rs11614913 variant of miRNA-196a-2 and the risk of epithelial ovarian cancer. Oncol Lett. 2016;11(1):194-200.
55. Hoffman AE, Zheng T, Yi C, Leaderer D, Weidhaas J, Slack F, et al. microRNA miR-196a-2 and breast cancer: a genetic and epigenetic association study and functional analysis. Cancer Res. 2009;69(14):5970-7.
56. Hu Z, Chen J, Tian T, Zhou X, Gu H, Xu L, et al. Genetic variants of miRNA sequences and non-small cell lung cancer survival. J Clin Invest. 2008;118(7):2600-8.
57. Vanella L, Kim DH, Asprinio D, Peterson SJ, Barbagallo I, Vanella A, et al. HO-1 expression increases mesenchymal stem cell-derived osteoblasts but decreases adipocyte lineage. Bone. 2010;46(1):236-43.
58. Zwerina J, Tzima S, Hayer S, Redlich K, Hoffmann O, Hanslik-Schnabel B, et al. Heme oxygenase 1 (HO-1) regulates osteoclastogenesis and bone resorption. Faseb J. 2005;19(14):2011-3.
59. Barbagallo I, Vanella A, Peterson SJ, Kim DH, Tibullo D, Giallongo C, et al. Overexpression of heme oxygenase-1 increases human osteoblast stem cell differentiation. J Bone Miner Metab. 2010;28(3):276-88.
60. von Muhlen D, Safii S, Jassal SK, Svartberg J, Barrett-Connor E. Associations between the metabolic syndrome and bone health in older men and women: the Rancho Bernardo Study. Osteoporosis International. 2007;18(10):1337-44.
61. Dennison EM, Syddall HE, Aihie Sayer A, Martin HJ, Cooper C, Hertfordshire Cohort Study G. Lipid profile, obesity and bone mineral density: the Hertfordshire Cohort Study. Qjm. 2007;100(5):297-303.
62. Ghorabi S, Shab-Bidar S, Sadeghi O, Nasiri M, Khatibi SR, Djafarian K. Lipid Profile and Risk of Bone Fracture: A Systematic Review and Meta-Analysis of Observational Studies. Endocr Res. 2019;44(4):168-84.
63. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. Nephron Clin Pract. 2010;115(2):c94-9.
64. Leening MJ, Heeringa J, Deckers JW, Franco OH, Hofman A, Witteman JC, et al. Healthy volunteer effect and cardiovascular risk. Epidemiology. 2014;25(3):470-1.
65. Lindsted KD, Fraser GE, Steinkohl M, Beeson WL. Healthy volunteer effect in a cohort study: temporal resolution in the Adventist Health Study. J Clin Epidemiol. 1996;49(7):783-90.
66. Lokk K, Modhukur V, Rajashekar B, Märtens K, Mägi R, Kolde R, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biol. 2014;15(4):r54.
67. Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, et al. Understanding Tissue-Specific Gene Regulation. Cell Rep. 2017;21(4):1077-88.
68. Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, et al. Distribution of miRNA expression across human tissues. Nucleic Acids Res. 2016;44(8):3865-77.
69. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9(5):757-68.
70. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
71. Su D, Wang X, Campbell MR, Porter DK, Pittman GS, Bennett BD, et al. Distinct Epigenetic Effects of Tobacco Smoking in

Whole Blood and among Leukocyte Subtypes. PLoS One. 2016;11(12):e0166486.
72. Bauer M, Fink B, Thürmann L, Eszlinger M, Herberth G, Lehmann I. Tobacco smoking differently influences cell types of the innate and adaptive immune system-indications from CpG site methylation. Clin Epigenetics. 2015;7:83.
73. Bauer M, Linsel G, Fink B, Offenberg K, Hahn AM, Sack U, et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin Epigenetics. 2015;7(1):81.
74. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res. 2009;11(1):12-24.
75. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7.
76. Collaborators GBDRF. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1345-422.
77. Collaborators GBDCoD. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736-88.
78. de Planell-Saguer M, Rodicio MC. Detection methods for microRNAs in clinic practice. Clin Biochem. 2013;46(10-11):869-78.
79. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Sci Int Genet. 2017;28:225-36.
80. Xavier C, de la Puente M, Mosquera-Miguel A, Freire-Aradas A, Kalamara V, Vidaki A, et al. Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA. Forensic Sci Int Genet. 2020;48:102336.
81. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol. 2017;18(1):238.



Chapter 6

Summary/ Samenvatting

ENGLISH SUMMARY

Lifestyle factors are associated with an increased risk of CVD, leading to almost 18 million deaths worldwide each year. Epigenetics is proposed as a possible mechanism in which these lifestyle factors could lead to disease and their risk factors. In this thesis, I used prediction methods to investigate the possibility of DNA methylation-based inference of lifestyle factors, such as cigarette smoking and alcohol drinking, and applied multi-omics approaches to investigate DNA methylation markers and microRNAs in association with health outcomes such as cardio-metabolic traits and bone mineral density.

In **Chapter 2.1**, I validated the statistical methods used to obtain previously published alcohol prediction models by Liu *et al*., which did not implement any internal or external validation methods. I found that when implementing a ten-fold cross-validation scheme as internal validation, the models including fewer CpGs obtained higher AUCs compared to the full CpG models, suggesting overfitting of the published models. The application of these models to independent datasets yielded much lower AUCs than previously published and with high variance between the four validation datasets with an overall lower AUC in the TwinsUK datasets. These results indicate the overestimation of the published AUCs and, with that, the need for a new prediction model for alcohol consumption before epigenetic inference of alcohol consumption can be considered for practical applications.

In **Chapter 2.2**, I investigated the possibility of DNA methylation-based inference of a person's smoking status. I identified 13 CpGs that can distinguish smokers from non-smokers with comparable accuracy as cotinine, a generally accepted smoking biomarker. The same 13 CpGs were able to infer cessation time in former smokers and pack-years in current smokers. Finally, I showed that these markers could infer lifetime smoking information. These models obtained high AUCs in both the model building data set and in two independent cohorts, showing the possibility for DNA methylation-based lifestyle inference.

In **Chapter 3.1**, I implemented a multi-omics approach to investigate the association for smoking-related changes in DNA methylation and gene expression with cardio-metabolic traits. I identified several significant *cis*- and *trans*-eQTM associations. Of these, I found 26 smoking-related CpGs and 21 smoking-related probes (19 genes) associated with a cardio-metabolic trait. I also identified three-way associations in which a CpG is associated with a gene and both are associated with the same trait. Specifically, I found for triglycerides a three-way association with two CpGs and two genes and for BMI with six CpGs and two genes. Finally, I implemented mediation analysis to investigate these three-way associations further and found a mediating effect for four CpGs in the association between smoking and changes in *LRRN3* expression. These results suggest

an important role for alterations in the epigenome and transcriptome in the association between smoking and cardio-metabolic traits.

In **Chapter 4.1**, I implemented a multi-omics approach to investigate miRNAs that are associated with cardio-metabolic traits. I used summary statistics from previous large-scale GWASs and identified 180 SNPs annotated to 67 miRNAs associated with at least one cardio-metabolic trait. Then, I identified 278 CpGs annotated to 64 miRNAs, and the expression of 22 miRNAs associated with a cardio-metabolic trait. In total, I identified an overlap for five miRNAs, including miR-10b-5p, miR-148a-3p, miR-100-5p, miR-125b-5p, and miR-6886, with at least one annotated CpG, and of which the expression were both associated with the same cardio-metabolic trait. Our results provide five potential biomarkers that could be of great interest in future studies.

In **Chapter 4.2**, I used summary statistics from large-scale GWASs on BMD and identified rs11614913 located in miR-196a-2 to be associated with femoral neck-BMD and lumbar spine-BMD. In addition, I showed a sex-specific association for rs11614913 with BMD in women. Finally, I showed an association for *JAG1* expression, a potential target gene of miR-196a-2, with BMD variation in the hipbone. These results suggest that miR-196a-2 changes are associated BMD, possibly via altered *JAG1* expression.

NEDERLANDSE SAMENVATTING

Leefstijlfactoren zijn geassocieerd met een verhoogd risico op hart- en vaatziekten, wat jaarlijks wereldwijd leidt tot bijna 18 miljoen sterfgevallen. Epigenetica wordt gesuggereerd als een mogelijk mechanisme waardoor deze leefstijlfactoren kunnen leiden tot ziektes en hun risicofactoren. In dit proefschrift heb ik gebruik gemaakt van predictiemodellen om te onderzoeken of het mogelijk is om op basis van DNA-methylatie iemands leefstijlfactoren, zoals roken en alcoholconsumptie, vast te stellen. Daarnaast heb ik door middel van multi-omics methodes DNA-methylatie markers en microRNAs onderzocht in associatie met cardio-metabole risicofactoren en botdichtheid.

In **Hoofdstuk 2.1** heb ik de statistische methoden gevalideerd die door Liu *et al.* werden gebruikt in eerder gepubliceerde alcoholpredictiemodellen waarin geen interne en externe validatiemethoden waren geïmplementeerd. Ik ontdekte dat bij het implementeren van een tienvoudige kruisvalidatie als interne validatie de modellen met minder CpGs hogere AUCs behaalden in vergelijking met de modellen waarin alle CpGs zijn inbegrepen. Dit wijst op over-fitting van de gepubliceerde modellen. Het toepassen van deze modellen op onafhankelijke datasets leverde veel lagere AUCs op dan eerder gepubliceerd en met een hoge variatie tussen de vier validatiedatasets met een algeheel lagere AUC in de TwinsUK-datasets. Deze resultaten bevestigen de overschatting van de gepubliceerde AUCs en daarmee de noodzaak voor de ontwikkeling van een nieuw predictiemodel voor alcohol inname voordat op DNA-methylatie gebaseerde inferentie van alcoholconsumptie kan worden overwogen voor toepassingen in de praktijk.

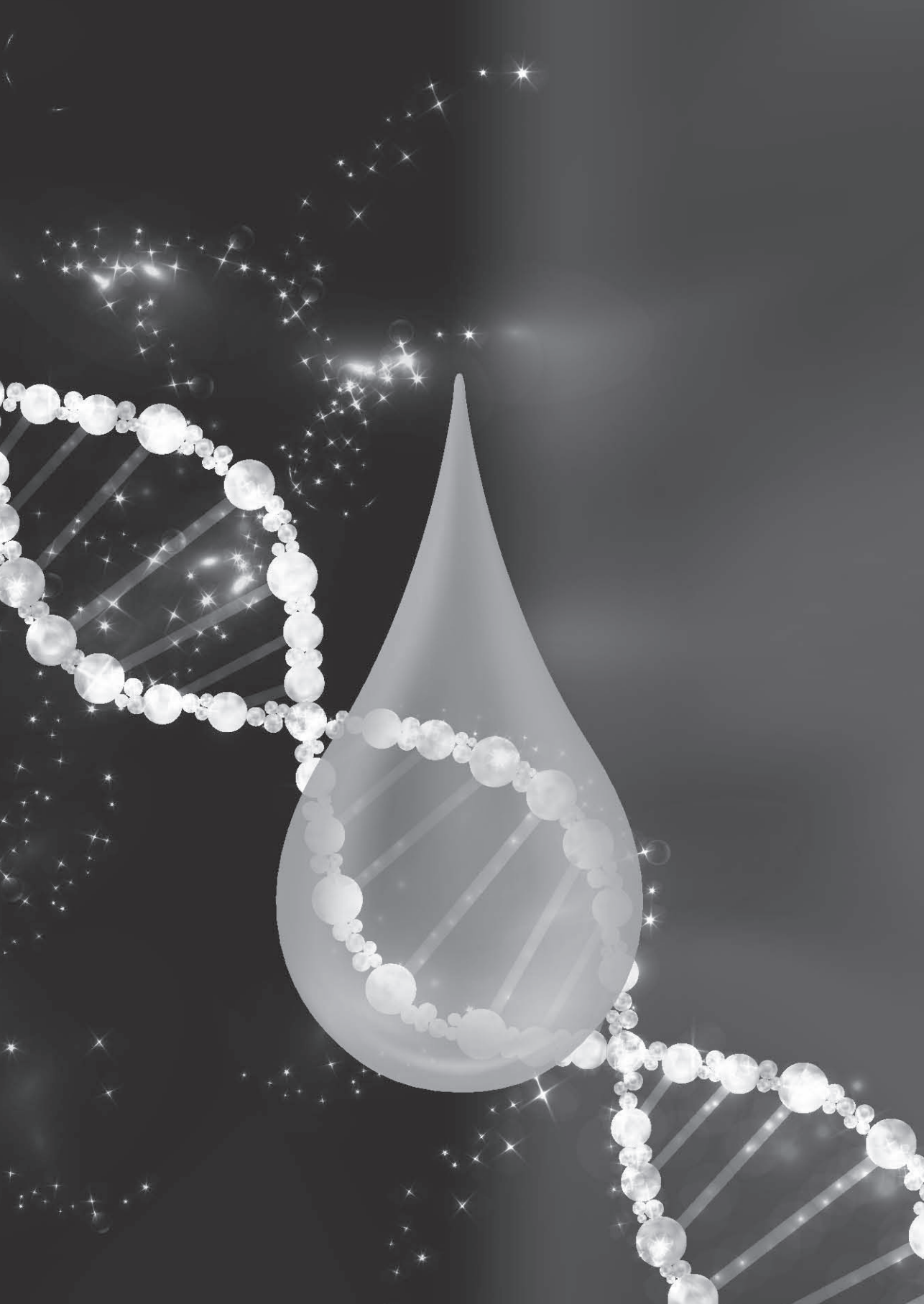
In **Hoofdstuk 2.2** onderzocht ik de mogelijkheid van op DNA-methylatie gebaseerde inferentie van iemands rookstatus. Ik identificeerde 13 CpGs die rokers van niet-rokers kunnen onderscheiden met een vergelijkbare nauwkeurigheid als cotinine, een algemeen geaccepteerde biomarker voor roken. Dezelfde 13 CpGs waren in staat om de tijd sinds stoppen te bepalen in voormalige rokers en pakjaren in rokers. Ten slotte toonde ik aan dat deze markers informatie over levenslang rookgedrag kunnen vaststellen. Deze modellen behaalden hoge AUCs in zowel de trainingsdataset als in twee onafhankelijke cohorten wat de mogelijkheid aantoont voor op DNA-methylatie gebaseerde leefstijlinferentie.

In **Hoofdstuk 3.1** heb ik een multi-omics methode geïmplementeerd om aan roken gerelateerde veranderingen in DNA-methylatie en genexpressie te onderzoeken in associatie met cardio-metabolische risicofactoren. Ik heb verschillende significante *cis*- en *trans*-eQTM associaties geïdentificeerd, waarvan 26 aan roken gerelateerde CpGs en 21 aan roken gerelateerde probes (19 genen) geassocieerd zijn met een cardio-metabole risicofactor. Daarnaast heb ik drieweg-associaties geïdentificeerd waarin een CpG is geassocieerd met een gen en beide zijn geassocieerd met dezelfde risicofactor. Ik vond voor triglyceriden een drieweg-associatie met twee CpGs en twee genen en voor BMI met

zes CpGs en twee genen. Ten slotte heb ik mediatie-analyses geïmplementeerd om deze drieweg-associaties verder te onderzoeken en een mediërend effect gevonden voor vier CpGs in de associatie tussen roken en veranderingen in *LRRN3* expressie. Mijn resultaten suggereren een belangrijke rol voor veranderingen in het epigenoom en transcriptoom in de associatie tussen roken en cardio-metabole risicofactoren.

In **Hoofdstuk 4.1** heb ik een multi-omics methode geïmplementeerd om miRNAs te onderzoeken in associatie met cardio-metabole risicofactoren. Hiervoor gebruikte ik samenvattende statistieken van eerder gepubliceerde grootschalige GWASs en identificeerden 180 SNPs geannoteerd aan 67 miRNAs in associatie met ten minste één cardio-metabole risicofactor. Vervolgens identificeerde ik 278 CpGs geannoteerd aan 64 miRNAs en de expressie van 22 miRNAs in associatie met een cardio-metabole risicofactor. In totaal heb ik een overlap geïdentificeerd voor vijf miRNAs, waaronder miR-10b-5p, miR-148a-3p, miR-100-5p, miR-125b-5p en miR-6886, met ten minste één geannoteerde CpG en waarvan de expressie was geassocieerd met dezelfde cardio-metabole risicofactor. Mijn resultaten presenteren vijf potentiële biomarkers die van groot belang kunnen zijn in toekomstige studies.

In **Hoofdstuk 4.2** gebruikte ik samenvattende statistieken van grootschalige GWASs over botdichtheid en identificeerden rs11614913 in miR-196a-2 in associatie met femurhals- botdichtheid en lumbale wervelkolom- botdichtheid. Daarnaast toonden ik een geslacht specifieke associatie aan voor rs11614913 met botdichtheid in vrouwen. Ten slotte toonden ik een associatie aan voor *JAG1*-expressie, een potentieel doelwit gen van miR-196a-2e, met variatie in botdichtheid in het heupbeen. Mijn resultaten suggereren dat veranderingen in miR-196a-2 geassocieerd zijn met botdichtheid, mogelijk via wijzigingen in *JAG1*-expressie.



Chapter 7

Appendices

LIST OF MANUSCRIPTS

Silvana C.E. Maas, Athina Vidaki, Alexander Teumer, Ricardo Costeira, Rory Wilson, Jenny van Dongen, Marian Beekman, Uwe Völker, Hans J. Grabe, Sonja Kunze BIOS Consortium, Karl-Heinz Ladwig, Joyce B.J. van Meurs, André G. Uitterlinden, Trudy Voortman, Dorret I. Boomsma, P. Eline Slagboom, Diana van Heemst, Carla J.H. van der Kallen, Leonard H. van den Berg, Melanie Waldenberger, Henry Völzke, Annette Peters, Jordana T. Bell, M. Arfan Ikram, Mohsen Ghanbari*, Manfred Kayser*. Validating biomarkers and models for epigenetic inference of alcohol consumption from blood. *Clinical Epigenetics* 2021;13(1):198. Doi: 10.1186/s13148-021-01186-3.

Irma Karabegović, Eliana Portilla-Fernandez, Yang Li, Jiantao Ma, **Silvana C.E. Maas**, Daokun Sun, Emily A. Hu, Brigitte Kühnel, Yan Zhang, Srikant Ambatipudi, Giovanni Fiorito, Jian Huang, Juan E. Castillo-Fernandez, Kerri L. Wiggins, Niek de Klein, Sara Grioni, Brenton R. Swenson, Silvia Polidoro, Jorien L. Treur, Cyrille Cuenin, Pei-Chien Tsai, Ricardo Costeira, Veronique Chajes, Kim Braun, Niek Verweij, Anja Kretschmer, Lude Franke, Joyce B.J. van Meurs, André G. Uitterlinden, Robert J. de Knegt, M. Arfan Ikram, Abbas Dehghan, Annette Peters, Ben Schöttker, Sina A. Gharib, Nona Sotoodehnia, Jordana T. Bell, Paul Elliott, Paolo Vineis, Caroline Relton, Zdenko Herceg, Hermann Brenner, Melanie Waldenberger, Casey M Rebholz, Trudy Voortman, Qiuwei Pan, Myriam Fornage, Daniel Levy, Manfred Kayser, Mohsen Ghanbari. Epigenome-wide association meta-analysis of DNA methylation with coffee and tea consumption. *Nature Communication* 2021;12(1):2830. Doi: 10.1038/s41467-021-22752-6.

Silvana C.E. Maas, Michelle M.J. Mens, Brigitte Kühnel, Joyce B.J. van Meurs, André G. Uitterlinden, Annette Peters, Holger Prokisch, Christian Herder, Harald Grallert, Sonja Kunze, Melanie Waldenberger, Maryam Kavousi, Manfred Kayser, Mohsen Ghanbari. Smoking-related changes in DNA methylation and gene expression are associated with cardio-metabolic traits. *Clinical Epigenetics* 2020;12(1):157. Doi: 10.1186/s13148-020-00951-0.

Michelle M.J. Mens, **Silvana C.E. Maas**, Jaco Klap, Gerrit Jan Weverling, Paul Klatser, Just P.J. Brakenhoff, Joyce B.J. van Meurs, André G. Uitterlinden, M. Arfan Ikram, Maryam Kavousi, Mohsen Ghanbari. Multi-Omics Analysis Reveals MicroRNAs Associated With Cardiometabolic Traits. *Frontiers in Genetics* 2020;11:110. Doi: 10.3389/fgene.2020.00110.

Silvana C.E. Maas, Athina Vidaki, Rory Wilson, Alexander Teumer, Fan Liu, Joyce B.J. van Meurs, André G. Uitterlinden, Dorret I. Boomsma, Eco J.C. de Geus, Gonneke Willemsen, Jenny van Dongen, Carla J.H. van der Kallen, P. Eline Slagboom, Marian Beekman, Diana

van Heemst, Leonard H. van den Berg, BIOS Consortium, Liesbeth Duijts, Vincent W.V. Jaddoe, Karl-Heinz Ladwig, Sonja Kunze, Annette Peters, M. Arfan Ikram, Hans J. Grabe, Janine F. Felix, Melanie Waldenberger, Oscar H. Franco, Mohsen Ghanbari*, Manfred Kayser*. Validated inference of smoking habits from blood with a finite DNA methylation marker set. *European Journal of Epidemiology* 2019;34(11):1055-74. Doi: 10.1007/s10654-019-00555-w.

Irma Karabegović, **Silvana Maas**, Carolina Medina-Gomez, Maša Zrimšek, Sjur Reppe, Kaare M. Gautvik, André G. Uitterlinden, Fernando Rivadeneira, Mohsen Ghanbari. Genetic Polymorphism of miR-196a-2 is Associated with Bone Mineral Density (BMD). *International Journal of Molecular Sciences* 2017;18(12):2529. Doi: 10.3390/ijms18122529

*Denotes equal contribution

PHD PORTFOLIO

Name PhD student:	Silvana Christina Elizabeth Maas
Erasmus MC Department:	Dept. of Epidemiology & Dept. of Genetic Identification
Research School:	Netherlands Institute for Health Sciences
PhD period:	August 2016 – Februari 2022
Promotors:	Prof. Dr. Manfred H. Kayser
	Prof. Dr. M. Arfan Ikram
Co-promotors:	Dr. Mohsen Ghanbari
	Dr. Athina Vidaki

PhD training	Year	Workload (ECTS)
Master of Science in Health Sciences, NIHES	2016-2018	120
Courses		
Study Design	2016	4.3
Biostatistical Methods I: Basic Principles	2016	5.7
Biostatistical Methods II: Classical Regression Models	2016	4.3
Principles of Research in Medicine and Epidemiology	2016	0.7
Principles of Genetic Epidemiology	2016	0.7
Genomics in Molecular Medicine	2016	1.4
Advances in Genomics Research	2016	0.4
Genetic-epidemiologic Research Methods	2016	5.1
SNPs and Human Diseases	2016	1.4
Linux for Scientists	2016	0.6
Genome-wide Association Studies	2016	0.7
Human Epigenomics	2016	0.7
Advances in Genome-Wide Association Studies of Complex Genetic Disorders	2017	1.4
Family-based Genetic Analysis	2017	1.4
An Introduction to the Analysis of Next-Generation Sequencing Data	2017	1.4
Markers and Prognostic Research	2017	0.7
Logistic Regression	2017	1.4
Introduction to Bayesian Methods in Clinical and Epidemiological Research	2017	1.4
Causal Mediation Analysis	2017	0.7
Cardiovascular Epidemiology	2017	0.9
Epigenesis and Epigenetics	2017	0.8
BBMRI-omics course	2017	0.6
Scientific Writing in English for publication	2017-2018	2.0
Repeated Measurements in Clinical Studies	2018	1.4
Topics in Meta-analysis	2018	0.7
Health Economics	2018	0.7
Joint Models for Longitudinal and Survival Data	2018	0.7
Quality of Life Measurement	2018	0.9

Missing Values in Clinical Research	2018	1.4
Intermediate Course in R	2018	1.4
Research integrity	2019	0.3
Seminars and meetings		
Cardiometabolic EPI meetings, Erasmus MC	2016-2019	1
MolEpi meetings, Erasmus MC	2016-2019	1
Genetic Identification lab meetings, Erasmus MC	2016-2018	1
Seminars at the department of Epidemiology, Erasmus MC	2016-2019	1
Erasmus MC PhD day	2019 and 2021	0.5
2020 meeting, department of Epidemiology, Erasmus MC	2016-2019	1
Conferences and presentations		
21st Molecular Medicine day, Erasmus MC – Attendance	2017	0.3
CHARGE Consortium meeting, Rotterdam – Attendance	2018	0.3
23rd Molecular Medicine day, Erasmus MC – Poster presentation	2019	0.5
Health sciences research day, Rotterdam – Oral presentation	2019	0.5
Other		
Peer review for *Clinical Epigenetics*	2021	
1 ECTS (European Credit Transfer System) equal to workload of 28 hours		

ABOUT THE AUTHOR

Silvana Christina Elizabeth Maas was born on October 15th 1989 in Breda, the Netherlands. She completed her Bachelor of Science degree in Biology and Applied Medical Laboratory Technology with a major in Forensic laboratory research in 2016 at the Avans University of Applied Sciences, Breda, the Netherlands. As part of this program, she completed internships in the Forensic Department of the Western Carolina University, Cullowhee, North Carolina, United States and at the Departments of Epidemiology and Immunology of the Erasmus University Medical Center, Rotterdam, the Netherlands.

Because of her research interest in combining lifestyle, epigenetics, and health, she continued with a Master of Science in Genetic Epidemiology at the Netherlands Institute of Health Sciences (NIHES), which she completed in 2018. Silvana expanded her work at the Department of Epidemiology and Genetic Identification of the Erasmus University Medical Center as a PhD student.

The results of her PhD research, entitled "*Epigenetic regulation and inference of lifestyle factors and health*", are presented in this dissertation.

Silvana will go on to work as a postdoctoral researcher at the Cancer Computational Biology Group, Vall d´Hebron Institute of Oncology, Barcelona, Spain, where she will work with Dr. José A. Seoane on the genetics and epigenetics of cancer. Silvana is married to Marcelo Ortiz and is the proud mother of Mateo-Millan (2).

DANKWOORD

Toen ik in 2015 mijn bachelor stage begon bij het Erasmus MC had ik nooit durven dromen dat ik een jaar later met mijn PhD zou mogen beginnen en nu vijf jaar later mijn proefschrift zou afronden. Het voltooien van dit proefschrift was niet mogelijk geweest zonder de hulp en steun van vele.

I would like to start by thanking my team of promotors and co-promotors, Prof. Dr. Manfred Kayser, Prof. Dr. M. Arfan Ikram, Dr. Mohsen Ghanbari, and Dr. Athina Vidaki, for the guidance and support you have provided over the past years. I have learned so much from each one of you and I have so much to thank you for. Working in two departments, Epidemiology and Genetic Identification, gave me the unique opportunity to learn several new techniques from leading experts in both fields. Thank you for welcoming me into your departments and for believing in me.

Daarnaast wil ik alle deelnemers en medewerkers van de Rotterdam Studie, het Nederlands Twin Register, Cohort on Diabetes and Atherosclerosis Maastricht, Prospective ALS Study Netherlands, Leiden Longevity Study, the Cooperative Health Research in the Region of Augsburg- F4 study, Study of Health in Pomerania- Trend, TwinsUK, LifeLines DEEP en Generation R bedanken voor hun onmisbare bijdrage. Ook wil ik graag Jolanda, Frank en Nano bedanken voor jullie hulp bij alle computerproblemen, alle ERGO gerelateerde vragen en voor de gezelligheid. Daarnaast natuurlijk ook Mirjam en Maaike voor jullie hulp bij eigenlijk alles, zonder jullie hulp was dit proefschrift nooit tot een goed einde gekomen en waren de afgelopen jaren niet zo gezellig geweest.

Prof. Dr. Oscar H. Franco, Dr. Abbas Deghan, and Dr. Mohsen Ghanbari, thank you for welcoming me into your team all the way back in 2015. It was during this internship that I fell in love with the world of (epi-)genetic epidemiology. Thank you for your guidance during this time and for believing in my potential for pursuing my PhD.

Prof. Joyce van Meurs, Prof. Bas Heijmans, and Dr. Abbas Dehghan, thank you for agreeing to be members of the reading committee and for taking the time to review my dissertation, providing feedback, and for joining my defense ceremony. Prof. Harold Snieder and Dr. Melanie Waldenberger, thank you for agreeing to serve on my doctoral committee and for participating in my PhD defense. It is a great honor for me to have you all as members of my committee and I truly appreciate the time and expertise of all the committee members in the assessment of my dissertation.

As research is seldom done alone, my gratitude extends to all the co-authors of the studies included in this dissertation. Thank you for your valuable contribution. I also want to extend my gratitude to all my colleagues at the departments of Epidemiology and Genetic Identification. Thank you for our constructive meetings, your comments and suggestions, but maybe even more for the great not-so-work-related meetings during the Christmas and Sinterklaas celebrations, lunches, coffees, cookie breaks, lunch at Het Park, museum visits, and many many more. Our time together has made the last five years not only an educational journey but also a time I look back at with great joy. A special thanks to Carolina, Irma, Paloma, Hamid, Arash, Banafsheh, Anh Nhi, Aline, Michelle, Eralda, Eliana, Marija, Arjola, Elif, Lyda, Oscar(s), Chantal, Janine, Marlou, Niels, Zhangling, Amy, Sven, Valentina, Jana, Mirjam, Maaike, Kim, Blerim, Jelena, Symen, Paul, Klodian, Mohsen, Adela, Maryam, Trudy, Layal, Abbas, Joyce, Maud, Sadaf, Vincent, Amanda, Athina, Arwin, Celia, Delano, Diego, Gabriela, Hedayat, Vivian, Leroy, Alex, Faida, Rochelle, Dion, Benjamin, Fan, Hilda, and all whose names I might have missed (sorry). It was an honor to work together with such amazing teams!

Michelle and Irma, thank you for inviting me onto your journey into the fascinating world of miRNAs as your co-author. It has been a great experience to work closely with you and to grow together in our knowledge and friendship.

Carolina and Irma, thank you for being such amazing friends, for your kindness and support, and for being such great companions during our MSc and PhD journey. Most of all, I want to thank you for giving me the honor of standing (virtually) behind me during my defense as my paranymphs. What a journey it has been, and what a beautiful journey we have ahead of us.

“The bond that links your true family is not one of blood, but of respect and joy in each other’s life. Rarely do members of one family grow up under the same roof” - Richard Bach. With this I want to give a very special thanks to my Rotterdam Family; Carolina, Alejandro, Irma, Paloma, Diego, Anh Nhi, Alice, Hamid, Banafsheh, and Arash. Over the last years, you have all grown to be a big part of my life and I could not have been more blessed than with being able to call you my friends.

Ook buiten het Erasmus MC zijn er een hele hoop mensen die ik enorm dankbaar ben.

Mijn lieve vrienden, Daphne, Ylona, Marlyn, Marlies, Ilse, Margot, Mariska, Maaike, en Luca, wat ben ik dankbaar voor jullie jaren lange vriendschap. Bedankt voor het bieden van een luisterend oor en voor de soms hardnodige afleiding tijdens onze stapavondjes en festivals, dinnerdates, video-cocktail meetings, dagjes weg en bioscoopdates.

Mijn lieve familie, ik wil jullie bedanken voor jullie constante interesse in hoe het nu op school gaat en de oneindig aanmoedigende woorden. Jullie onvoorwaardelijk steun heeft mij enorm geholpen door de jaren heen.

A mi familia Chilena, gracias por darme la bienvenida a su vida, por su bondad incondicional, amor, cuidado. Sobre todo, gracias por darme el amor de mi vida.

And finally, the biggest thanks goes out to the most important men in my life: Marcelo and Mateo-Millan. There are no words to describe how grateful I am for our little family. Thank you for your love, support, and encouragement. With you by my side is everything easier ♥ Te Quiero.

PROPOSITIONS

Epigenetic Regulation and Inference of Lifestyle Factors and Health

1. DNA methylation-based prediction models should be developed in such a way that they are easily extrapolated towards the general public. *(This thesis)*

2. Using DNA methylation data, it is possible to infer someone's smoking habits with reasonable accuracy. *(This thesis)*

3. Smoking-induced epigenetic and gene expression alterations affect cardio-metabolic health. *(This thesis)*

4. Investigating multi-omics layers concurrently in relation to health and disease outcomes will provide better insights into underlying molecular pathways. *(This thesis)*

5. Publicly available summary statistics from genome-wide association studies provide a valuable resource for the identification of microRNAs involved in complex traits. *(This thesis)*

6. The impact of positive lifestyle changes towards a healthier life should have a more prominent role in the medical practice.

7. Epigenetic markers have the potential to be used as disease biomarkers and in precision medicine.

8. The only impossible journey is the one you never begin. *(Tony Robbins)*

9. Realize that everything connects to everything else. *(Leonardo da Vinci)*

10. To do things right, first you need love, then technique. *(Antoni Gaudí)*

11. Your work is going to fill a big part of your life and the only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do. *(Steve Jobs)*