Disentangling the Impact of Lifestyle Factors on Health and Disease: A Molecular Epidemiology Approach

Irma Karabegović

Acknowledgments

The work presented in this thesis was performed at the Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands. The studies presented herein were conducted within the Rotterdam Study, Avon Longitudinal Study of Parents and Children (ALSPAC), Airwave Health Monitoring Study (Airwave), Atherosclerosis Risk in Communities Study (ARIC), Cardiovascular Health Study (CHS), European Prospective Investigation into Cancer and Nutrition Study (EPIC), Epidemiological investigations of the chances of preventing, recognizing early and optimally treating chronic diseases in an elderly population study (ESTHER), Framingham Heart Study (FHS), Cooperative Health Research in the Augsburg Region study (KORA), Lifelines study and the TwinsUK study. We gratefully acknowledge the contributions provided by the study participants, staff, data management, participating general practitioners and health professionals involved.

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

Layout and printed by: Gordan Zovko, Connect Design and Irma Karabegović
Printed by: Ridderprint BV | www.ridderprint.nl
Cover design: Suzan Renner and Irma Karabegović
ISBN: XX

Disentangling the Impact of Lifestyle Factors on Health and Disease: A Molecular Epidemiology Approach

De invloed van aanpasbare leefstijlfactoren op gezondheid en ziekte ontrafelen: Een moleculaire epidemiologische benadering

Thesis

To obtain the degree of Doctor from the

Erasmus University Rotterdam

by command of the rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on

Tuesday, 6th of June 2023 at 13:00 hrs

by

Irma Karabegović

born in Bihać, Bosnia and Herzegovina

Erasmus University Rotterdam

DOCTORAL COMMITTEE

Promotor:	Prof.dr. M. Arfan Ikram
Other members:	Prof.dr. F. Rivadeneira
	Prof.dr. L.H. Franke
	Dr. C.A.M. Cecil
Copromotor:	Dr. Mohsen Ghanbari

PARANYMPHS

Dr. Silvana Maas
Soroush Mohammadi Jouabadi

Za Ifatu, Izeta i Irfana

Jer ništa nisam morala,
ali sam u svim odlukama bila podržana.

Table of contents

MANUSCRIPTS THAT FORM THE BASIS FOR THIS THESIS

Chapter 2. DNA methylation signature of coffee and tea consumption

Karabegović, I., E. Portilla-Fernandez, Y. Li, J. Ma, S. C. E. Maas, D. Sun, E. A. Hu, B. Kuhnel, Y. Zhang, S. Ambatipudi, G. Fiorito, J. Huang, J. E. Castillo-Fernandez, K. L. Wiggins, N. de Klein, S. Grioni, B. R. Swenson, S. Polidoro, J. L. Treur, C. Cuenin, P. C. Tsai, R. Costeira, V. Chajes, K. Braun, N. Verweij, A. Kretschmer, L. Franke, J. B. J. van Meurs, A. G. Uitterlinden, R. J. de Knegt, M. A. Ikram, A. Dehghan, A. Peters, B. Schottker, S. A. Gharib, N. Sotoodehnia, J. T. Bell, P. Elliott, P. Vineis, C. Relton, Z. Herceg, H. Brenner, M. Waldenberger, C. M. Rebholz, T. Voortman, Q. Pan, M. Fornage, D. Levy, M. Kayser, and M. Ghanbari. "Epigenome-Wide Association Meta-Analysis of DNA Methylation with Coffee and Tea Consumption." Nat Commun 12, no. 1 (May 14 2021): 2830. https://dx.doi.org/10.1038/s41467-021-22752-6

Chapter 3. Lifestyle factors and their role in disease development through microRNAs

Karabegović, I., Y. Abozaid*, S. C. E. Maas*, J. Labrecque, D. Bos, R. J. De Knegt, M. A. Ikram, T. Voortman, and M. Ghanbari. "Plasma Microrna Signature of Alcohol Consumption: The Rotterdam Study." Journal of Nutrition (Sep 20 2022). https://dx.doi.org/6708378

Karabegović I.*, S. C. E. Maas*, Y. Shuai, M. A. Ikram, J. Aerts, G. Brusselle, L. Lahousse, T. Voortman, M. Ghanbari, "Smoking-induced dysregulation of circulating microRNAs in plasma: the Rotterdam Study." Submitted

Chapter 4. Age related changes in elderly populations

Karabegović I., S. Maas, C. Medina-Gomez, M. Zrimsek, S. Reppe, K. M. Gautvik, A. G. Uitterlinden, F. Rivadeneira, and M. Ghanbari. "Genetic Polymorphism of Mir-196a-2 Is Associated with Bone Mineral Density (BMD)." Int J Mol Sci 18, no. 12 (Nov 25 2017). https://dx.doi.org/ijms18122529

Amiri, M., **Karabegović, I.***, A. C. van Westing*, Ajcf Verkaar, S. Beigrezaei, M. Lara, W. M. Bramer, and T. Voortman. "Whole-Diet Interventions and Cardiovascular Risk Factors in Postmenopausal Women: A Systematic Review of Controlled Clinical Trials." Maturitas 155 (Jan 2022): 40-53. https://dx.doi.org/S0378-5122(21)00293-0

* Denotes equal contribution

CHAPTER 1

GENERAL INTRODUCTION

INTRODUCTION

Behavioural lifestyle factors such as diet, smoking, physical activity, and alcohol consumption play a prominent role in general health and disease. For instance, ultra-processed diet, smoking, high alcohol consumption, and physical inactivity have been established as risk factors for many noncommunicable diseases (NCDs)[1], whereas the opposite, such as nutrient dense diets, smoking cessation, decrease in alcohol consumption and physical activity have a protective effect on overall human health[2]. With a steep increase in human life expectancy, notably in developed countries[3,4], researchers have noted high prevalence of NCDs. Although human population benefits from longer life span, it is not necessarily a healthy one, as we often live with disabilities and chronic diseases[5]. Fortunately, these behavioural lifestyle factors are modifiable, providing an early intervention for NCDs, rather than treatment of the disease. Considering that NCD currently account for more than one-half of the global burden of disease[6], various regulations and economic tools (eg. taxes) have been proposed as effective strategies in reducing tobacco and harmful alcohol use as behavioural determinants in adult mortality[4]. Large efforts have been made in attempt to decrease unhealthy commodities (tobacco, alcohol, and ultra-processed food and drinks), however, corporations which directly profit from sales of these unhealthy commodities use several strategies to undermine NCD prevention and control[7]. For instance, it has been shown that funding arising from food industry in nutrition-related scientific articles are likely to bias conclusions in favour of sponsor's products[8]. Given the reluctance of industry on self-regulation, and the fact that lifestyle factors (e.g. cigarette smoking, alcohol consumption, obesity and less physical activity) are modifiable risk factors of many NCDs, public health programs might benefit more on intervention of these modifiable lifestyle factors at global ranking. Although the link between lifestyle factors and NCDs is evident, the underlying molecular mechanisms are still not fully unravelled. Understanding these mechanisms might help us designing better strategies to combat NCDs. Noncommunicable diseases are a cumulative effect of interplay between one's genetic architecture, environmental effects, and lifestyle factors, **Figure 1**.

Epigenetics is one of the proposed mechanisms linking various lifestyle factors with disease risk. Epigenetic mechanisms depict an intersection between genetic and environmental influences on disease risk[9]. Epigenetics act through regulating gene expression without altering the DNA sequence and these comprise three main mechanisms: DNA methylation, histone modifications and non-coding RNAs.

Consideration of genetic, epigenetic and other omic layers as potential intermediate features of lifestyle factors and their impact on disease predispositions in later life will therefore provide a better understanding of the underlying mechanisms, and ultimately help to be more effective on unravelling novel biomarkers and therapeutic targets.

GENETIC AND MOLECULAR EPIDEMIOLOGY

1. Genomics

Since discovering of the DNA structure by Watson and Crick[10], with crucial contributions by Rosalind Franklin[11], the field of genetics has massively evolved. In genetics, one focuses either on a single gene or individual variants[12], however, as the technologies have advanced in the last decades, the field of genomics has enabled researchers to focus simultaneously on an entire set of loci and genome regions[12]. In comparison to the pre-defined set of molecules, "omics" approach provides unique setting to explore the information underlying a trait of interest in a hypothesis-free approach. In addition, as technology has advanced in last few decades, the completion of the original human genome was possible, enabling us to further understand distribution of genetic variation and their implication for common diseases[13]. From accumulating this large data, researchers have estimated that all human beings are 99.9% identical in their genetic data, with the remaining 0.1% not only being responsible for our different physical appearance, but also for our different susceptibilities to the diseases[14]. In order to investigate this 0.1%, high-throughput technologies have enabled researchers to develop arrays which simultaneously asses millions of changes in DNA sequence, referred as single-nucleotide polymorphisms (SNPs), which are mostly stable throughout life. Variations in one's DNA sequence can be related to disease onset[15]. In a genome-wide association study (GWAS), genetic variants (SNPs) that are more frequently present in individuals having the disease or trait of interest are identified, in comparison to a healthy or control group[16]. Although genomics has been established as the first discipline to emerge with cost-efficient and high-throughput analysis of molecules, it was soon followed by transcriptomics, epigenomics, proteomics, metabolomics and microbiomics[12], and unlike genomics, these omics are dynamic through life.

2. Transcriptomics

The central dogma in molecular biology refers to the flow of information in our organism[17]. The accepted flow of information in central dogma is DNA > RNA > protein. Although the DNA sequence is mostly stable throughout life (unless we are exposed to radiation, develop cancer or DNA replication error arises), the translation of this information to functional product changes during our lifetime. Not all genes are expressed at the same time or in the same tissue. For instance, we all have the genetic sequence (instructions / blueprint) to grow taller, however, these genes are only expressed in a certain period of time during our lifespan until peak height has been achieved. Furthermore, the gene expression can also change as a response to external stimulus (eg. stress, diet, air pollution). To make the situation more intriguing, genetic variants and epigenetic marks (DNA methylation, non-coding RNAs and histone modification) can change the expression of the genes. The

SNPs associated with changes in gene expression are referred to as expression quantitative trait loci (eQTL)[18]. Similarly to genomics, transcriptomics has made vast advances with development of high-throughput technology and thereby resulted in methods such as transcriptome-wide association studies (TWAS). Similarly to GWAS, TWAS uses a hypothesis-free approach in order to investigate complete transcriptome in association with trait of interest.

3. Epigenomics

Epigenomics is a rapidly growing field where molecular mechanisms act via regulating the gene expression without alternating the underlying DNA sequence. Furthermore, in comparison to its precedent genomics, where DNA sequence itself is discrete, epigenomics is quantitative[19]. Another interesting fact about epigenomics is its tissue specificity. For instance, our bone and blood tissue are different from each other, although cells comprising these tissues share identical DNA sequence. As previously explained, not all the genes for which we have information are expressed simultaneously – especially not in each tissue. One of the culprits responsible for tissue specific differential expression of genes is epigenetics[20], as one of the main regulators of gene expression. In addition, unlike stable genomics, epigenomics can be altered based on both: internal (such as hormones and disease state) and external stimuli (such as environmental and lifestyle factors) and is highly dynamic, yet partially reversible. The epigenetic mechanisms comprise DNA methylation, histone modifications and non-coding RNAs[21].

3.1. DNA methylation

DNA methylation is thus far the most studied epigenetic mechanism. The modus operandi involves a methyl group (-CH3) addition or removal from the cytosine nucleotide, which is followed by guanine nucleotide in the DNA sequence[22]. This addition or removal possibly results in alternation of gene expression. It is interesting to note that some of the changes in DNA methylation induced by unhealthy lifestyle such as smoking are reversible following cessation[23]. What is even more remarkable is that some of the smoking induced changes on methylation markers are not only heritable, but also might contribute to disease development at later life point in the offspring[24-26]. Given the fact that some of these behaviours are well known to be hazardous (e.g. smoking or drinking alcohol, especially during pregnancy), and therefore frowned upon in today's society, collecting data on these variables often results in social desirability bias. To put it simply, a future mother might lie on an interview if asked about smoking or alcohol consumption. However, the unique characteristic of DNA methylation markers is that they can be used as non-invasive blood derived biomarkers in order to infer whether the person is a smoker[27] or consumes alcohol[28]. As DNA methylation can be measured both in tissues or body fluids, this

makes methylation as potential non-invasive biomarkers for many diseases [29,30]. In addition, methylation markers can be associated with other omics layers, including CpGs that are linked with changes in expression levels of the genes (transcriptomics) –referred as expression quantitative trait methylation (eQTMs)[31]. Other examples are CpGs that have underlying genetic variants (genomics) which might influence their methylation patterns either nearby (cis-) or far away (trans-methylation quantative trait loci (meQTLs)), **Figure 2.** Just like its precedent genomics, the epigenomics field has utilized the high-throughput technology and arrays which resulted in epigenome-wide association studies (EWAS) – a hypothesis-free approach to investigate epigenome landscape in association with either a disease, environmental or lifestyle exposure.

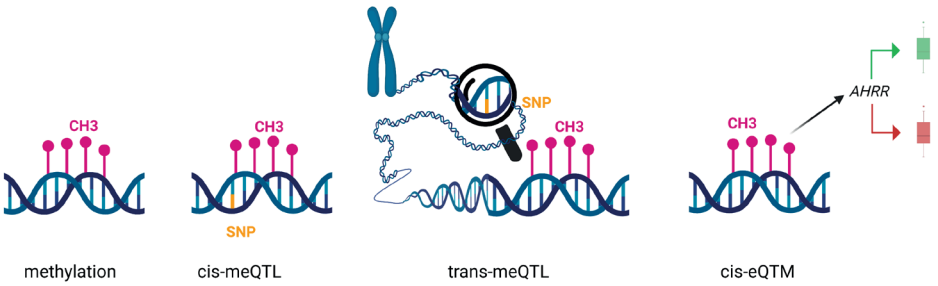

Figure 2. DNA methylation markers in association with other omics layers.

DNA methylation marker can be influenced by underlying SNPs that are located either nearby (*cis*-meQTLs) or far away (*trans*-meQTLs). Furthermore, DNA methylation might also be associated with changes of expression of their respective annotated genes (*cis*-eQTMs).

3.2. Gene regulation by microRNAs

microRNAs (miRNAs) are small non-coding RNAs (~22 nucleotide long) which post-transcriptionally regulate gene expression. Similarly to other epigenetic makers, miRNAs are also tissue specific. Their mechanism of action relies on targeting their complementary messenger RNA (mRNA). Once the targeted mRNA is recognized, it can either result in perfect matching with mRNA- therefore cleaving the mRNA, or imprecise matching with mRNA - resulting in repressed translation of mRNA[32], **Figure 3**. miRNAs can also be found extracellularly in body fluids such as serum, saliva, urine and plasma, making them a potential biomarker for many diseases[33]. There are many detection methods that could be used to assess miRNAs. While qRT-PCR is generally accepted in the field, as well as cost-effective, it has limited multiplex capacity. Newly developed targeted microarray platforms and novel next-generation sequencing technology has allowed researchers to investigate miRNAs in a wider landscape[34].

LIFESTYLE FACTORS IN NONCOMMUNICABLE DISEASES

Lifestyle factors have been known to play a role on noncommunicable diseases (NCDs)[1,7]. Researchers have estimated that observed increase in NCDs could be attributed to longer lifespan and population growth, however the main modifiable risk factors contributing to NCDs are unhealthy lifestyle factors such as diets rich in ultra-processed foods, physical inactivity, tobacco consumption, and alcohol misuse[35,36]. One of the proposed mechanisms through which modifiable lifestyle factors might contribute to NCDs is epigenetics[37]. For instance, smoking is one of the firmly established lifestyle factors having a strong impact on DNA methylation patterns[38], while some changes observed due to impact of tobacco smoking were irreversible following cessation. Smoking induced changes on methylation patterns have been causally linked to inflammation[39]. In addition to this, exposure to maternal smoking has provided evidence for causal role on increased risk for schizophrenia and inflammatory bowel disease in offspring[24]. Similarly to smoking, diet quality has also been associated with DNA methylation markers[40,41] as well as alcohol consumption[42], and other lifestyle factors[37,43]. Some other omics layers were investigated in association with lifestyle factors, for instance large scale GWASs have been performed on coffee consumption[44-46]. Other omics layers including metabolomics [47] and proteomics[48] have also been investigated in relation to coffee consumption.

Given that the lifestyle factors are the main modifiable risk factors contributing to NCDs, we wanted to further expand the knowledge on how these lifestyle factors might be linked with NCDs through exploring their mechanism of action revolving around omics data.

AIM OF THIS THESIS AND OUTLINE

In this thesis I sought to elucidate the potential underlying mechanisms of lifestyle factors which might explain their association with noncommunicable diseases (NCDs) using population-based omics data. In **Chapter 2**, I aimed to identify two potential underlying mechanisms of coffee and tea consumption. Specifically, in **chapter 2.1** I investigated DNA methylation signatures with coffee and tea consumption, then I focused on exploring if the identified DNA methylation markers were linked with risk of metabolic diseases such as fatty liver disease.

Chapter 3 aimed to dissect the role of lifestyle factors on miRNA levels and their potential impact on noncommunicable diseases. Notably, **chapter 3.1** focuses on exploring miRNAs as potential signature of alcohol consumption, then if these miRNA could mediate the association between alcohol consumption and fatty liver disease. In **chapter 3.2**, I aimed to explore the impact of smoking-induced dysregulation of miRNA expression, as well as the impact of smoking cessation on these miRNAs. Additionally, I explored the link of smoking-associated miRNAs with lung cancer.

Chapter 4 focuses on exploring age related disease in elderly populations. Specifically, **chapter 4.1** investigates the association between genetic variants within miRNA sequences and bone mineral density (BMD), using publically available data. Next, I explored target genes of the identified miRNAs and sex-specific associations given the sex-specific nature of BMD. Moreover, **chapter 4.2** was conducted to review the role of whole-diet interventions and cardiovascular risk factors in postmenopausal women.

STUDY DESIGN AND POPULATIONS

This thesis was based on original studies, publicly available data and a systematic review study. The original studies conducted within this thesis were embedded in multiple prospective cohort studies. In prospective cohort studies, individuals undergo baseline examinations where the information is collected from all the participants equally, including data collection methods or same questions. In this study design, the method collection is designed before certain disease occurs (eg. fatty liver disease), and individuals are followed for a long period of time, so the exposure of interest (eg. coffee consumption) or other potential risk factors (eg. smoking) is collected with accurate information[49]. The individuals undergo re-examination where the information is updated. Although costly and time-consuming, this study design provides unique setting to explore exposure to environmental factors, lifestyle factors or gene-environment impact on complex human diseases[50]. The prospective cohort study contributing with original studies presented in this thesis was mainly embedded in the Rotterdam Study (RS)[51]. Rotterdam Study is a Dutch population based cohort where participants were recruited from the Ommord District, Rotterdam. Briefly, participant who were >45 years underwent physical examinations and questionnaires. RS was initiated in 2008, where in total n=14,926 participants were enrolled in three separated recruitment periods. With the most recent extension and inclusion of fourth cohort in 2016 (RS-IV, n=3005), the sample size totals to n=17,931.

In **Chapter 2.1** of this thesis, the data was utilized from the prospective cohort studies embedded in the Cohort for Heart and Aging Research in Genomic Epidemiology (CHARGE consortium) (https://www.chargeconsortium.com/)[52] and additional participating cohorts. CHARGE consortium was initiated in order to extenuate collaboration with genome-wide association study meta-analyses and replications across many large international longitudinal cohort studies [52]. The five founding member cohorts include the Framingham Heart Study (FHS)[53], the Rotterdam Study (RS), the Age, Gene/Environment Susceptibility-Reykjavik Study, the Atherosclerosis Risk in Communities Study(ARIC)[54] and the Cardiovascular Health Study (CHS)[55]. Later on, the consortium got extended with additional cohorts such as the Multi-Ethnic Study of Atherosclerosis (MESA), the Coronary Artery Risk Development in Young Adults, the Health, Aging, and Body Composition Study, the Jackson Heart Study and the Family Heart Study. Cohorts who had data available on the DNA methylation, as well as coffee and tea consumption, were included in