



UiT The Arctic University of Norway

Faculty of Health Sciences

Department of Community Medicine

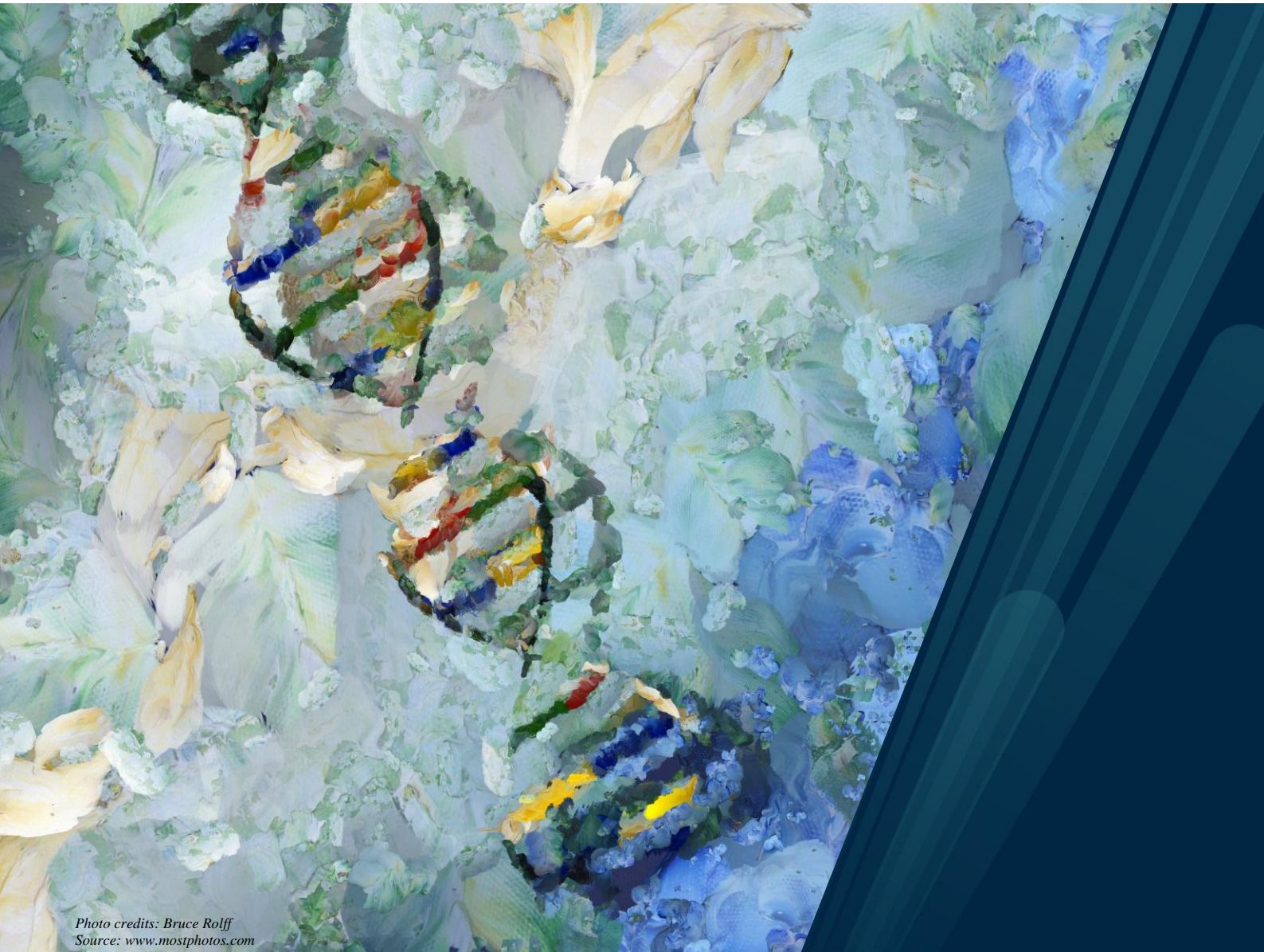
## **Gene expression in blood and cancer risk factors**

*Investigating associations of blood gene expression with Smoking, BMI, and Menopause in the Norwegian Women and Cancer (NOWAC) postgenome cohort*

**Nikita Baiju**

A dissertation for the degree of Philosophiae Doctor

June 2023





# **Gene expression in blood and cancer risk factors**

*Investigating associations of blood gene expression with Smoking, BMI, and Menopause in the Norwegian Women and Cancer (NOWAC) postgenome cohort*

**Nikita Baiju**

Department of Community Medicine

Faculty of Health Sciences

UiT The Arctic University of Norway

Tromsø, Norway

2023



*The fear of the LORD is the beginning of wisdom, and the knowledge of the holy is understanding. (Proverbs 9:10; KJV)*



## Acknowledgements

*"The process was overwhelming than most people would like to believe.....and it got worse before it got better again!"*

.....

My PhD journey was not an easy task. There were many highs and lows in both my academic and personal life. But the love, help, and support of numerous individuals played an indispensable role in making this journey achievable.

I would like to express my sincere thanks and indebtedness to my supervisors –Therese Haugdahl Nøst, Torkjel Sandanger, and Pål Sætrom. First, thank you Therese and Torkjel for giving me the opportunity to pursue this PhD. You comprehended my passion for research and eagerness to learn new things and gave me a chance to experience the whole new world of *Gene expression*. **Therese**..., mere words cannot adequately convey the depth of my appreciation for you. Thank you for believing in me more than I had believed myself. I deeply appreciate your expert guidance, constant supervision, and ever-abiding encouragement. Thank you for generously providing me the time and support that I required to grow as a researcher. Thank you for not just guiding me academically, but also for always being there and supporting me emotionally when I needed. I consider myself incredibly fortunate to have the opportunity to be the first PhD student that is supervised by you –I'm blessed! You have the best quality of being the main supervisor one could ever ask for –just perfect (neither too much nor too little!). **Torkjel**, you have been a great leader and invaluable source of support! Thank you for being a mentor and always asking if I am doing okay in work and life. It means a lot! **Pål**, I am always amazed by the immense knowledge you have! Thank you for all the small and big suggestions and insightful thoughts you had given regarding my work, which have turned the path of my work in a better direction. Also, thank you for opening doors for me to come to Trondheim and work with you. Truly, I have learned a lot from you all, but most importantly, I have learned how to excel yet be humble and compassionate. Thank you for everything!

I would also like to thank my co-authors, Charlotta Rylander and Marit Waaseth for your ideas, guidance, and assistance in my second and third papers. Your critical questions and suggestions prompted me to think better, which ultimately enhanced the quality of my papers. Thank you!

Dolley, thank you for being a great colleague, a friend and above all my extended family in Norway. Thank you for all the academic and life discussions. Love you!

Sairah, thank you for being a wonderful officemate. Your encouraging words "You can do it!" has always given me that extra kick to work harder yet relax mentally.

Femi, Runa, and Torill, thank you for sharing your experiences and good advice. Fjorida, Bahar, Andrew, Masoud, and Faith, thank you making the last year more fun...wish I have known you all a little earlier. And Andrew especial thanks to you for helping me out to solve difficult R codes. Marko, Karina, Kristin, Kajsa, Illona, and all other members in the Systems

Epidemiology Research Group, thank you for the valuable academic guidance, fun filled social gatherings, and encouraging words. It has been a privilege to be a part of such an intellectually robust group. I want to express my appreciation to all the fellow PhD students and colleagues at Department of Community Medicine whom I have had the pleasure of meeting throughout the years.

I would like to thank Tonje Braaten, Arne Bastian Wiik, Marko Lukic, and Nikita Shvetsov for providing the data set for my project. Special thanks to Jo Inge for helping with data wrangling and Trudy Perdrix-Thoma for English language editing.

Besides, I would like to express my sincere gratitude to all the women who participated in the Norwegian Women and Cancer Study (NOWAC). I would like to thank Bente Augdal and all other personnel responsible for the administration of the data collection and the biobank. Thank you Eiliv Lund for initiating the NOWAC study and giving opportunities to so many researchers like me to address vital research questions. Thank you UiT The Arctic University of Norway, for entrusting me with this project.

My acknowledgment would not be complete without thanking all my family members in Nepal, Pakistan, and India. I will always be indebted to your prayers, moral support, and suggestions.

Mom, Dad, Dada, Bhauju, and my little darling Rebu, thank you for all your love, support, and prayers for me. Thank you for supporting all my life decisions, encouraging me always, and teaching me to become a better human being. Ammi and Abbu, thank you, for believing in me and loving me so much. Abbu...you left us too soon (1941-2021) ...I really wish you could be with us to witness my achievement. Thank you: Suru, Supi, Bhai-Bhabhi(s) in Pakistan, Aunty-Uncle(s) in India. I hope I make you all a little prouder today...I love you all so much!

Lastly, I am immensely grateful to my support systems –my dear husband, Ali, and my sweetheart, my baby little boy, Robel! Ali..., who would know my journey better than you and who would have given me the amount of love and support like you did...?! Thank you! And thank you for your enduring patience and unwavering belief in me. I would never have been able to pursue my dreams without you and your support. I love you so much! And Robel...you are a miracle baby and a true gift of God. Thank you for your smiles and hugs that refreshed and energized me whenever I was tired and low. Mama and Baba love you so much!

Above all, I thank and praise the Lord for giving me His wisdom, guidance, and strength throughout this journey, and blessing my life with more than I deserve. Glory be to God!

.....



*This thesis is a dedication to my roots, my loving grandmothers -Late Mrs. Dwarika Bardewa and Late Mrs. Thuli Devi Baiju -what remarkable women both of you were!*



# Table of Contents

Acknowledgements .....	vii
List of Figures .....	xi
Summary .....	xiii
Visual Abstract.....	xv
List of Papers.....	xvii
Abbreviations .....	xix
1 Introduction .....	1
1.1 Cancer risk factors.....	2
1.1.1 Smoking status and smoking metrics .....	4
1.1.2 Body mass index and weight change.....	7
1.1.3 Menopausal status and hormone therapy use .....	9
1.2 Gene expression profiling .....	13
1.2.1 Background: From DNA to proteins .....	13
1.2.2 Measurement techniques .....	15
1.2.3 Target tissue for measurements .....	18
1.2.4 Dynamics of gene expression profiles.....	19
2 Aims of the thesis .....	21
3 Methods and materials.....	23
3.1 Study population.....	23
3.1.1 The Norwegian women and cancer study (NOWAC).....	23
3.1.2 The NOWAC postgenome cohort .....	25
3.1.3 The study design.....	26
3.2 The gene expression data .....	26
3.2.1 Laboratory analyses.....	26
3.2.2 Pre-processing of dataset.....	27
3.3 Selected exposure variables -formation and definition .....	27
3.3.1 Smoking status and smoking metrics (Paper I) .....	28
3.3.2 Body mass index and weight change (Paper II) .....	28
3.3.3 Menopausal status and hormonal therapy use (Paper III) .....	29
3.4 Statistical analyses.....	29
3.4.1 The selected exposure variables .....	30
3.4.2 White blood cell proportions .....	32

3.4.3	Covariates and adjustment models .....	32
3.4.4	Functional enrichment analyses .....	33
3.4.5	Quantitative replication .....	34
3.5	Ethical considerations.....	34
3.5.1	Data management .....	34
3.5.2	Data availability.....	35
4	Results – Summary of papers .....	37
4.1	Paper I .....	37
4.2	Paper II .....	38
4.3	Paper III.....	39
5	Discussion.....	41
5.1	Main results .....	41
5.1.1	Smoking status and smoking metrics .....	41
5.1.2	BMI and weight change.....	45
5.1.3	Menopausal status and HT use .....	49
5.1.4	Across the risk factors investigated –the broader picture.....	53
5.1.5	The novelties .....	60
5.1.6	Knowledge contributions to future cancer studies .....	61
5.2	Methodological considerations.....	63
5.2.1	Study design .....	63
5.2.2	Bias.....	66
5.2.3	Confounding.....	71
5.2.4	Interaction (effect modification).....	72
5.2.5	Generalizability .....	73
5.2.6	Gene expression analyses .....	73
5.2.7	Statistical analyses.....	78
6	Conclusions .....	83
7	Future perspectives .....	85
8	References .....	87
	Errata .....	99
	Appendices .....	101

## List of Figures

<b>Figure 1:</b> An illustration of risk factors of cancer. ....	2
<b>Figure 2:</b> The steps involved in the central dogma in a eukaryote. ....	15
<b>Figure 3:</b> Timeline for the Norwegian Women and Cancer Study (NOWAC). ....	24
<b>Figure 4:</b> A flow chart of the study populations. ....	26
<b>Figure 5:</b> Different adjustment models across Paper I-III. ....	33
<b>Figure 6:</b> Venn diagrams showing intersects for DEGs among smoking, BMI, menopause in Model-2 and Model-3 (Paper III). ....	56
<b>Figure 7:</b> An illustration of a three-level nested study design. ....	64



## Summary

**Background:** Despite the advances made in cancer research and treatment, the global burden of cancer continues to rise. It is important to acknowledge exogenous and endogenous risk factors and increase knowledge of relevant molecular signatures in studies of cancer biomarkers or molecular mechanisms. Linking blood gene expression and common cancer risk factors represents an intriguing approach for gaining valuable insights into the biological functions of genes reflecting processes related to the exposures or the development and progression of cancer and other diseases.

Smoking and obesity are the two most important modifiable cancer risk factors. Menopause is another important risk factor, and although the impact of menopause and hormonal factors on cancer risk is limited, their collective effect at the population level can be substantial as the population of postmenopausal women is growing. These risk factors affect the major physiological and biological processes in one's body. Still, there is limited or no research evaluating the associations of blood gene expression profiles with these risk factors –smoking, body mass index (BMI), and menopause. Moreover, studies utilizing large and extensive population-based samples to assess such relationships are rare.

**Aim:** This thesis aimed to evaluate differentially expressed genes (DEGs) among different levels of selected risk factors, specifically: smoking status and smoking metrics (Paper I), BMI and weight changes (Paper II), and menopausal status and hormone therapy (HT) use (Paper III); and to gain insights into their gene ontologies and pathways.

**Methods:** This thesis is based on studies using cross-sectional analyses nested within the prospective longitudinal NOWAC study and microarray-based gene expression profiles obtained from bio-banked whole-blood samples of women (N=1,716). Relevant information was obtained from up to three main questionnaires before and one at the blood collection time point. We used gene-wise linear regression models to identify DEGs, and functional enrichment analyses to determine their biological functions.

**Results:** We observed 911 and 1,082 DEGs when comparing current-vs-never and current-vs-former smokers, respectively. Few or no DEGs were observed when focusing on former smokers, passive smokers, or selected smoking metrics. We observed *LRRN3*-driven discrimination in all smoking exposures, suggesting that *LRRN3* could supplant self-reported

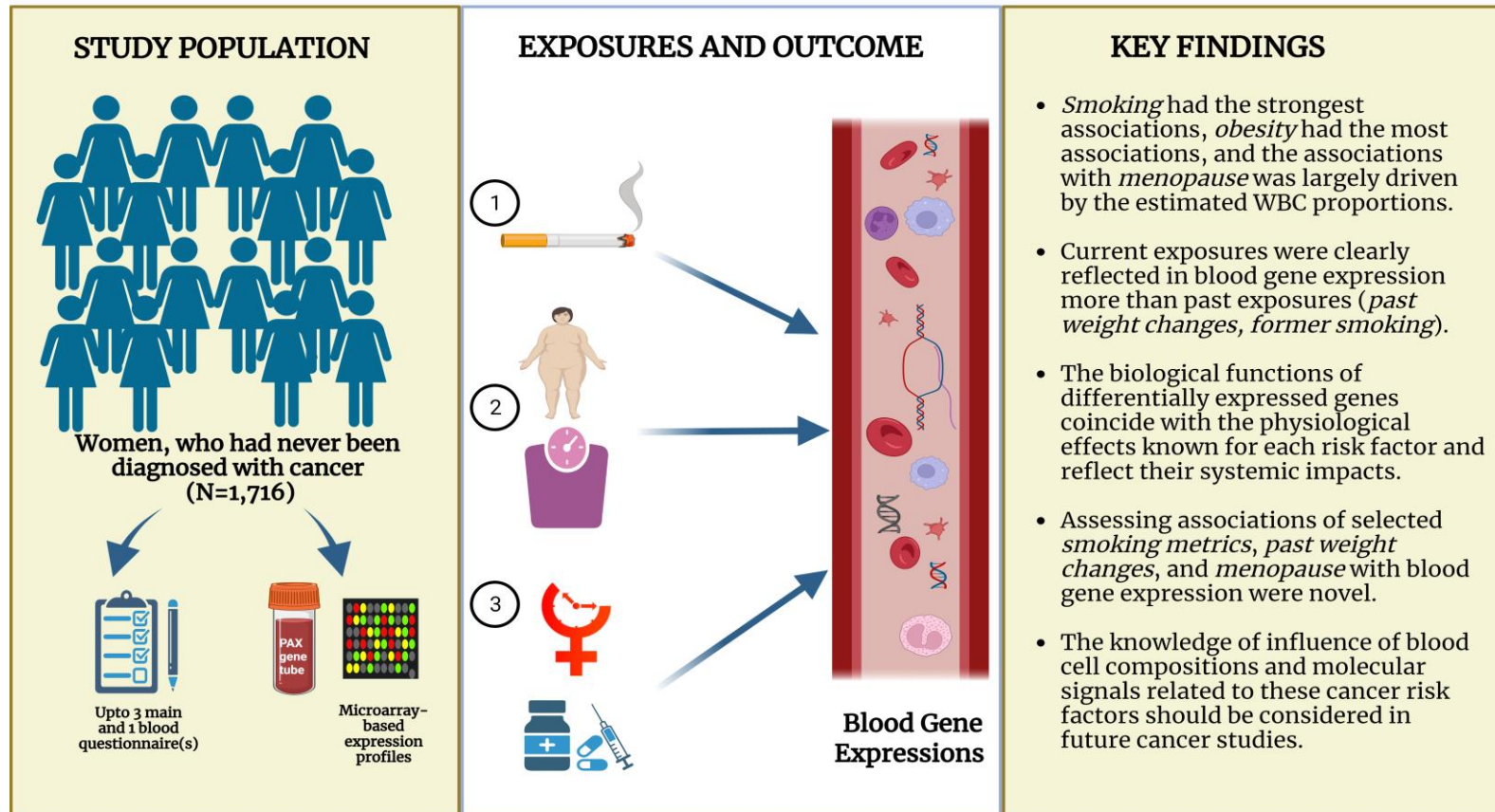
smoking data in future studies. We observed 2,394, 769, and 768 DEGs when comparing obesity-vs-normal-weight, obesity-vs-overweight, and overweight-vs-normal-weight, respectively. Only up to 169 DEGs were observed in past weight change analyses in interaction with BMI categories. There were 1,460 and 348 DEGs in relation to menopause and HT use, respectively, showing clear associations; yet when adjusted for the estimated white blood cell (WBC) proportions, the number reduced to 26 and 7, respectively. The biological functions of smoking-, BMI-, and menopausal-associated DEGs mainly revealed enriched terms like metabolic, immune, erythrocytes/reticulocytes related processes/functions.

**Conclusion:** The findings of this thesis conclude that there are overall associations of blood gene expression with cancer risk factors investigated among women in the NOWAC postgenome cohort. Among all the risk factors investigated, smoking had the strongest associations (in terms of effect sizes of the top-ranked DEGs), obesity had the most associations (in terms of number of DEGs), and the associations with menopause was largely driven by the relative presence of estimated WBCs. Further, the enriched gene ontologies and pathways of DEGs coincide with the physiological effects known for each risk factors and reflect their systemic impacts. In summary, current exposures were reflected in blood gene expression more than past exposures (former smoking status, past weight changes) and the influence of cell compositions on blood gene expression was important for interpretation of the results. The knowledge gained in this thesis is important for knowledge of molecular signals of the risk factors investigated.

# Visual Abstract

## Gene expression in blood and cancer risk factors

Investigating associations of blood gene expression with Smoking, BMI, and Menopause in the Norwegian Women and Cancer (NOWAC) postgenome cohort







# List of Papers

Below is the list of papers on which this thesis is based on:

## Paper I

Gene expression in whole-blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) Post-genome cohort.

Baiju N, Sandanger TM, Sætrom P, Nøst TH.  
*Scientific Reports*. 2021;11(1):1-13.

## Paper II

Associations of gene expression in blood with BMI and weight changes among women in the NOWAC postgenome cohort.

Baiju N, Rylander C, Sætrom P, Sandanger TM, Nøst TH.  
Accepted by *Obesity –A Research Journal*

## Paper III

Associations of blood gene expression profiles with menopausal status and hormone therapy use in the Norwegian Women and Cancer (NOWAC) postgenome cohort.

Baiju N, Waaseth M, Sætrom P, Sandanger TM, Nøst TH.  
Manuscript.



## Abbreviations

BMI	body mass index
cDNA	complementary DNA
CSI	comprehensive smoking index
DEG	differentially expressed gene
DNA	deoxyribonucleic acid
FDR	false discovery rates
GO	gene ontologies
GSEA	gene set enrichment analysis
HT	hormone therapy
KEGG	Kyoto encyclopaedia of genes and genomes
limma	linear models for microarray and RNA-seq data
logFC	log <sub>2</sub> fold-change
mRNA	messenger RNA
NOWAC	Norwegian Women and Cancer Study
ORA	over-representation analysis
PBMC	peripheral blood mononuclear cell
WHO	World Health Organization
PCA	principal component analysis
qPCR	quantitative polymerase chain reaction
Q1	baseline questionnaire
Q2	follow-up questionnaire
Q3	questionnaire at blood collection time point
RBC	red blood cell
REK	Regional Ethical Committee of North Norway
RNA	ribonucleic acid
RNA-seq	RNA sequencing
scRNA-seq	single-cell RNA Sequencing
TSC	time since smoking cessation
WBC	white blood cell
WC <sub>Q3-Q1</sub>	weight changes between Q1 and Q3
WC <sub>Q3-Q2</sub>	weight changes between Q1 and Q2



# 1 Introduction

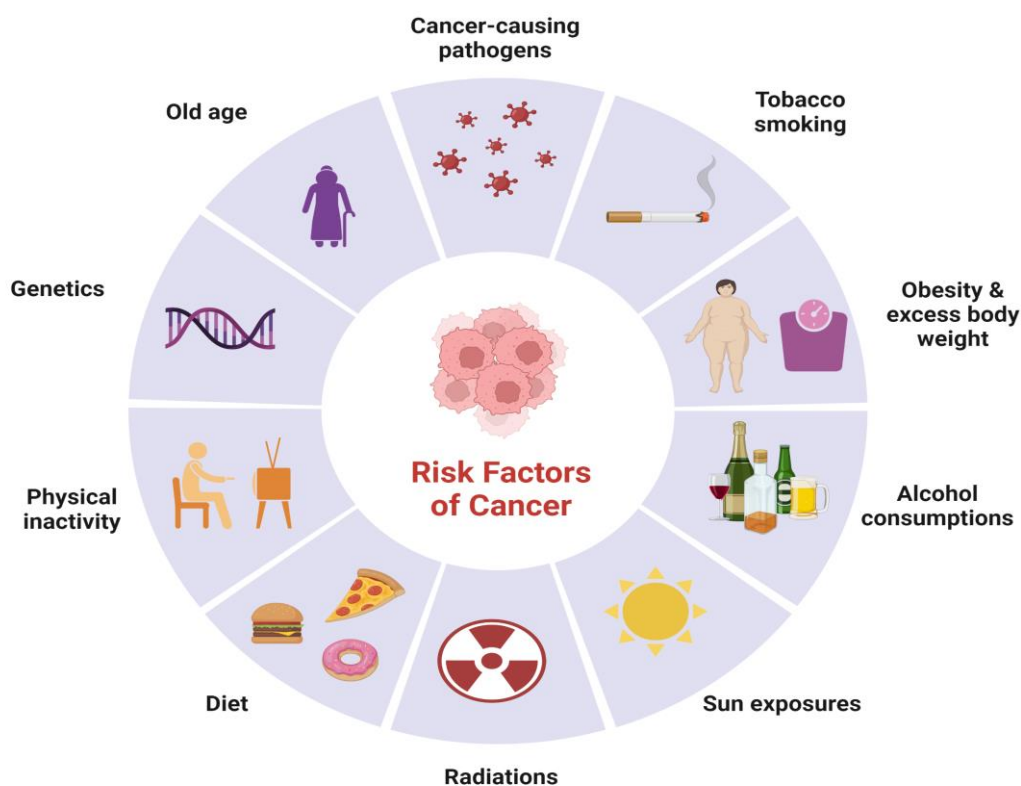
*The Human Genome Project* is one of the landmarks in the human history. It was a highly collaborative international effort, carried out in the years 1990-2003, to generate the first complete sequence of the human genome. It is one of the most important biomedical research projects that gave opportunity to explore novel areas of the biological world [1]. The period after the completion of this project is called the *postgenome era*, where the focus shifted from sequencing the human genome to analysing and interpreting the vast amount of genomic data generated. The birth of the Norwegian Women and Cancer Study (NOWAC) postgenome cohort (2003-2006) also occurred in this era.

Gene expression profiling/analyses give the opportunity to compare the expression levels for multiple genes simultaneously. Differentially expressed genes (DEGs), as observed through gene expression analysis, can lead to identification of potential molecular signals or biomarkers associated with exposure variables, disease progression, and/or specific diseases. Such analyses provide valuable insights into the involvement of DEGs in biological processes and pathways. Gene expression studies can be valuable resources, which have been widely utilized by biomedical research aiming to uncover the mechanisms underlying various diseases and contributing to the expression of the cells' phenotype. Indeed, one of the key features of cancer are deviated gene function and altered gene expression patterns [2]. Changes in gene expression can disrupt the normal function of cells, which can contribute to the development and progression of cancer [3]. Additionally, there are different risk factors of cancer that influence gene expression patterns, and investigating these could help in identifying DEGs in relation to cancer disease and provide knowledge on molecular signals related to different levels of these risk factors or exposure variables. Still, much research is yet to be conducted.

This thesis takes advantage of the availability large number of bio-banked whole-blood (or interchangeably referred to as 'blood' in this thesis) samples ready for gene expression analyses, that incorporates detailed and reliable lifestyle exposure information in a prospective design [4]. The cancer risk factors investigated in this thesis –mainly smoking, body mass index (BMI), and menopause – and their importance and associations with blood gene expression are introduced in Chapter 1.1. Gene expression technologies will be further introduced in Chapter 1.2.

## 1.1 Cancer risk factors

Any factor (internal and external) that can raise the likelihood of a person developing cancer is known as a *cancer risk factor* [5]. The key risk factors causing cancer are tobacco use, excess body weight, unhealthy diet, physical inactivity, infectious agents, and alcohol consumption; and many of these risk factors can be prevented [5,6], thus known as *modifiable risk factors* (*exogenous factors* or *external factors*). There are other risk factors such as age, personal or family history of cancer [5,6], which are unavoidable, and thus known as *non-modifiable risk factors* (*endogenous factors* or *internal factors*). An illustration of different cancer risk factors is given below (Figure 1).



**Figure 1:** An illustration of risk factors of cancer.

Template adaption from “Risk factors of Dementia”, by BioRender.com (Nordestgaard, L., 2022). Retrieved from: <https://app.biorender.com/biorender-templates/t-633de94d30ad4edb2a8ea4aa-risk-factors-of-dementia>

### *The global burden of cancer*

Cancer is the second leading cause of death globally accounting for 19 million cancer cases and 10 million deaths in 2020 [3,6,7]. The exposure to risk factors plays key role in the biology and burden of several cancer types [7]. The cancer types with most new cases in 2020 were breast,

lung, colon and rectum cancers, while those with most cancer deaths were lung, colon and rectum, liver, stomach, and breast cancers [6]. It has been estimated that 1/3<sup>rd</sup> of cancer deaths was due to tobacco use, high BMI, alcohol consumption, low fruit and vegetable intake, and lack of physical activity [6]. Epidemiological and laboratory studies conducted in high-income countries over the past 70 years provided evidence that around 40% of cancer burden can be attributed to identified risk factors –the two most important being smoking and obesity – highlighting that there are additional factors driving the remaining cancer burden [3].

There is extensive evidence suggesting that lifestyle factors play a crucial role in cancer risk and mortality, and making favorable, widespread lifestyle modifications can substantially diminish the cancer burden worldwide. Nevertheless, extensive modifications are required to achieve a substantial reduction in cancer incidence. By avoiding risk factors and implementing established prevention strategies, approximately 30-50% of cancers can currently be prevented [6]. Despite the advances made in cancer research and treatment, the global burden of cancer continues to rise. This highlights the need for innovative strategies in cancer prevention and control. It is important to acknowledge exogenous and endogenous risk factors and increase knowledge of relevant molecular signatures in studies of cancer biomarkers or molecular mechanisms that can lead to cancer prevention.

#### *Choice of risk factors (exposure variables) in this thesis*

Smoking has been and still is the foremost modifiable cancer risk factors, while obesity is the second on the list [3]. The associations of menopause and hormonal factors on cancer risk are relatively small at individual levels, but their collective effect at the population level can be substantial because all women experience these [8], and more importantly the population of postmenopausal women is growing globally [9]. The increase in lifetime number of monthly menstrual cycles has been linked to higher risk of breast, endometrial, and ovarian cancers. While the exact mechanisms remain unclear, one potential explanation for these relationships could be higher exposure to endogenous estrogen and progesterone levels [8]. Further, all these risk factors (both exogenous and endogenous in nature) affect the major biological processes in one's body and known for alteration in gene expression. Based on the relevance for studies on cancer and knowledge gaps based on previous studies on these risk factors and gene expression, *smoking, obesity, menopause* were the main exposure variables investigated in the papers

(Paper I-III) included in this thesis. Below we describe in more detail each of these risk factors and our rationale for their focus in relation to gene expression in blood in this thesis.

## **1.1.1 Smoking status and smoking metrics**

### **1.1.1.1 Background**

There are different forms of tobacco use and different products are available such as waterpipe tobacco, smokeless tobacco, cigars, cigarillos, roll-your-own tobacco, pipe tobacco, bidis, and kreteks; while cigarette smoking is the most common type of tobacco use globally [10]. Cigarette is a tube-shaped tobacco product, made of finely cut, cured tobacco leaves wrapped in thin paper [11]. A lit cigarette is typically smoked, and the smoke is usually inhaled into the lungs. Cigarettes contain nicotine and numerous cancer-causing chemicals, posing risks to both individuals who smoke and those who do not [11]. Common quantitative measures of smoking exposure (referred to as ‘smoking metrics’) are smoking intensity, smoking duration, time since smoking cessation (TSC), pack-years, and comprehensive smoking index (CSI) scores [12]. Passive smoking (or also called second-hand smoking) refers to inhaling the smoke exhaled by a smoker [13].

The history of tobacco smoking is long, dating back to 6,000 BC, when native Americans first started cultivating the tobacco plant. Around 1 BC, indigenous American tribes used tobacco for medicinal and religious purposes [14]. At early 16<sup>th</sup> century, approximately 50 years after Christopher Columbus’s first voyage to America, tobacco was introduced to Europe at the Portuguese court situated in Lisbon [15]. Towards the end of 19<sup>th</sup> century, due to mechanization and mass marketing, cigarette smoking was popularized [16], however the popularity reached their peak during the First and the Second World Wars in the 20<sup>th</sup> century, when tobacco companies dispatched countless packs of cigarettes to soldiers fighting on the front lines, creating hundreds of thousands of loyal and addicted consumers [14]. In the 20<sup>th</sup> century, health concerns related to cigarette smoking also began to rise with the studies linking smoking to lung cancer and other health problems [16].

Tobacco in all its forms is detrimental to health, and there is no level of exposure that can be deemed safe [10]. There have been comprehensive public health campaigns focusing on the dangers of smoking since early 20<sup>th</sup> century; and today, in many countries around the world, cigarette smoking is heavily regulated, with high pricing and taxes, bans on advertisements,



and support to smoking cessation [10]. However, prevention of smoking-related health injuries remains important for public health initiatives [17].

### **1.1.1.2 Prevalences and time trends**

The tobacco epidemic is one of the largest public health threats humankind has ever confronted [10]. The World Health Organization (WHO) recently estimated more than 8 million deaths every year due to tobacco use, including around 1.2 million deaths from passive smoking alone [10]. The burden of tobacco-related illness and death is the heaviest in low- and middle-income countries, where over 80% of the 1.3 billion tobacco users reside. This not only impacts health but also contributes to heavy economic drainage [10].

In 2020, it was estimated globally that 36.7% males and 7.8% females used some forms of tobacco. Among these, current smoking prevalence was 28.9% (with 26.4% cigarette smoking prevalence) and 5.2% (with 4.6% cigarette smoking prevalence), for males and females, respectively. These make 91% male and 88% female cigarette smokers [18]. The age-standardized rates of tobacco smoking prevalence have shown a consistent decline at a global level since at least the year 2000. In 2000, 27% of adults (aged 15 and older) were current tobacco smokers, while by 2020, this rate had decreased to 17%. This downward trend is anticipated to persist until at least year 2025, when the rate expected to reach 15% [18]. This decline in trend (2000-2025) is expected in all WHO regions. However, South-East Asia Region is anticipated to make the greatest progress in reducing the rates of smoking (29% in 2000 to 19% in 2010 and 13% in 2020), whereas Western Pacific Region (28% in 2000 to 25% in 2010, and 23% in 2020) and the Eastern Mediterranean Region (20% in 2000 to 17% in 2010 and 15% in 2020) are with the slowest progress. In Norway, in the early 1910s, nearly 30% Norwegian men smoked, while only a few women smoked [17]. Starting in the mid-1950s, significant decrease in smoking rates among men was observed, while the trend of increase continued among women, rendering equal proportions of men and women smokers by the end of the 1990s [17]. Today, daily smoking is equally common for men and women in Norway [17], with 7% men and 8% women who smoke daily in 2022 [19].

Further, smoking prevalence is not solely reliant on the proportions of smokers in a population but also on the intensity of smoking, which is measured by the average number of cigarettes smoked by individuals. Across Asia, Eastern Europe, North America, and Oceania, the average

number of cigarettes consumed per day by smokers is between 20-25, however, rates in Latin America, Africa, and Western Europe tend to be slightly lower [20].

### **1.1.1.3 Physiological and health effects**

Cigarette contains more than 7,000 chemicals, including nicotine, tar, and carbon monoxide [21], which contribute to several detrimental health effects on the physiology of the body. These effects might appear not long after smoking initiation or up to several decades after exposure [22,23]. When comparing individuals who smoke with those who never smoked, subclinical outcomes have been observed in smokers, such as increased oxidative stress, reduced antioxidant defences, increased inflammation, impaired immune status, and altered lipid profiles [22]. It also slows the wound healing process [24]. Additionally, smoking can dampen the immune system and alter a variety of immunological functions [25], making the body more susceptible to infections and reduce the ability to fight diseases.

Smoking mainly damages the respiratory system, including the lungs and airways. Smoking causes inflammation and irritation of the airways, increased mucus production, and narrowing of the air passages, making it harder to breathe [26]. Notably, respiratory symptoms caused by exposure to tobacco smoke have been observed more in women than men [27,28]. Further, chemicals inhaled during smoking can damage the heart and blood vessels, increase blood pressure (hypertension), and lead to atherosclerosis (build-up plaque in the arteries causing it to harden) [21]. Smoking can also have negative impact on the reproductive system [26], digestive systems [29], oral health [30], and accelerate the aging process and contribute to premature wrinkling and skin damage [31].

Smoking is the leading cause of lung cancer, but also increases the risk and mortality of other cancers, cardiovascular diseases, and chronic obstructive pulmonary disease, and lower respiratory tract infections [17,21,26].

Quitting smoking has significant health benefits and can help in reducing the risk of developing smoking-related diseases. The sooner an individual quits smoking, the greater the potential for improved health outcomes [26].

### **1.1.1.4 Associations with gene expression and research gap**

Multiple studies have reported that current exposure to tobacco smoke can lead to changes in the expression of various genes in blood, such as *LRRN3*, *CLDND1*, *GPR15*, *ATF4*, *SOD2*, and

*CDKN1C* [32-39]. Furthermore, altered gene expression in blood has been associated with smoking-related diseases [40]. However, most studies used smoking status as the smoking exposure measure and no study had conducted a more thorough investigation of the variability in gene expression profiles in whole-blood in relation to several quantitative measures of smoking exposure.

## **1.1.2 Body mass index and weight change**

### **1.1.2.1 Background**

The terms overweight and obesity refer to the atypical or surplus accumulation of body fat, which can have negative impacts on health [41]. BMI, previously called as the Quetelet index, is a widely used measure of body fat based on an individual's weight and height.

The BMI measure was developed by Adolphe Quetelet during the 19<sup>th</sup> century, became a popular measure of body composition because of its simplicity in measurement and calculation. It is widely used to determine the correlation between body weight –in proportion to height– and the risk of health problems at the population level. In the 1970s, it was found to be reliable proxy for adiposity and overweight-related problems [42]. It is the most useful population-level measure of overweight and obesity due to its similarity across both sexes and for all age groups of adults [41]. Nevertheless, like any other measure, it is not a perfect index. It has its limitations since it solely relies on height and weight and doesn't account for muscle mass, varying degrees of age-related adiposity, physical activity levels, or gender [42]. Further, weight gain during adulthood is primarily due to accumulation of fat, rather than lean tissue. Therefore, any change in weight could be a more accurate indication of adult adiposity as opposed to adult attained weight per se. Indicators such as adult weight gain, waist circumference, hip circumference, and waist-hip ratio provide insights into the distribution of adipose tissue [43]. Weight change guidelines depends on an individual's starting BMI and health status. People with overweight and obesity, who have cardiovascular risk factors, are recommended to change lifestyles to obtain modest, sustained weight loss; and only about 3-5% weight loss among them can bring meaningful health benefits [44]. It is also important to follow a proper diet plan to achieve weight loss and weight loss maintenance among people with obesity [44].

### **1.1.2.2 Prevalences and time trends**

Almost every country in the world is affected by the obesity epidemic today. According to the Global Burden of Disease study, in 2017, 4.7 million people died prematurely because of

obesity [45]. In 2016, globally, the number of overweight adults (aged 18 years and above) exceeded 1.9 billion, and of these over 650 million adults were with obesity [41]. In 2016, the prevalence of overweight was 39% among adults aged 18 years and above (39% of men and 40% of women), and in the same age group, the prevalence of obesity was 13% (11% of men and 15% of women). It is noteworthy that the prevalence of obesity worldwide almost tripled between 1975 and 2016 [41]. Except for certain regions in sub-Saharan Africa and Asia, the number of people with obesity surpasses those who are underweight [41]. In Norway, around 1-in-4 (25%) middle-aged men and 1-in-5 (20%) women are classified as having obesity [46]. In the last 40-50 years, the proportion of adults with overweight or obesity has increased in Norway, with variation by region and education level [46].

### **1.1.2.3 Physiological and health effects**

Obesity and excess body weight are associated with physiological changes in body affecting various systems and organs, and ultimately resulting in numerous medical conditions [47]. Obesity is closely linked to metabolic and endocrine disorders. It causes insulin resistance and impairs the body's ability to regulate blood sugar levels. It also disrupts the levels of hormones involved in appetite regulation such as leptin and ghrelin [48]. Impaired insulin resistance and oxidative stress are known to be caused by hyperglycemia (high blood sugar levels) in obesity [49]. Further, excess body weight can have negative impact on the heart and blood vessels, putting strain on the cardiovascular system [50]. Excess weight can lead to reduced lung capacity/volume and impaired respiratory function [51]. Obesity is recognized to impair the immune function and cell-mediated immune responses. Moreover, chronic inflammatory response stems from the connections between adipose tissue and the immune system [52]. Obesity adds additional stress on the bones, joints, and muscles due to excess body weight, affecting mobility of a person and causing challenge to his/her physical activity [53]. Further, obesity can cause hormonal imbalance and menstrual irregularity in women. It adversely affects reproductive function through alterations in the hypothalamic-pituitary-ovarian axis, oocyte quality, and endometrial receptivity [54].

Weight, weight gain, and obesity accounts for various cancer types, including postmenopausal breast and endometrium cancer in women [43,55,56], increase the risk of type 2 diabetes [48], and can lead to hypertension and other cardiovascular disease [50]. Further, obesity increases the risk of various respiratory [51], musculoskeletal [53], reproductive [54], and gastrointestinal [57] conditions and disorders. Obesity is an independent risk factor for liver damage, which can

contribute to liver fibrosis [58]. Further, obesity can lower one's self-esteem and body image issues affecting psychological well-beings, especially of children and adolescents. This can lead to depression, anxiety, and impact social interactions [59].

Managing obesity through a combination of healthy eating, regular physical activity, and, if necessary, medical interventions, can help mitigate these physiological effects and reduce the associated health risk.

#### **1.1.2.4 Associations with gene expression and research gap**

The understanding of the molecular mechanisms of multi-factorial conditions like obesity can be improved with gene expression studies [60]. Studies have reported associations between increased BMI and altered gene expression in blood [61-63] and adipose tissue [64], indicating changes in biological functions. Furthermore, previous studies have indicated that weight loss in individuals with obesity, following diet interventions, was linked to gene expression profiles in adipose tissue both before and after the diet interventions [65-71]. But no study had evaluated differences in blood gene expression related to past weight changes; and not many had investigated associations of BMI and blood gene expression in a large population-based sample.

### **1.1.3 Menopausal status and hormone therapy use**

#### **1.1.3.1 Background**

Menopause is a natural process of biological ageing that most women experience between 45-55 years of age. Menopause represents a crucial milestone in a women's life cycle, signifying the end of her reproductive years. After menopause, conception is no longer possible for a woman, unless specialized fertility treatments are employed in rare cases [9]. A confirmed diagnosis of menopause requires the absence of menstruation for a consecutive period of 12 months [72]. The menstrual transitional period, referred to as 'perimenopause', leads to various menopausal symptoms, which can last for 4-12 years [9,72].

Hormone replacement therapy, also called hormone therapy (HT), is a treatment used to relieve menopausal symptoms in women. The therapy usually contains estrogen (referred to as 'oestrogen' in the British English) hormone, but for women who still have their uterus, progestogen hormone is also added to protect the endometrial (lining of uterus). The estrogen hormone can be administered orally, intravaginally, or through transdermal means, while the

progestogen hormone can be taken orally, through transdermal means, or delivered via an intrauterine device [73].

The historical references and understanding of menopause have indeed evolved over time, and the earliest known references to menopause date back to Aristotle, who referred to age at menopause being 40 years. The term "menopause" was coined by a French physician in 1821. This marked a significant step in acknowledging and identifying the phenomenon. In the 1930s, menopause was described by some as a deficiency disease, resulting in advocacy of various replenishment therapies like testicular juice, crushed ovaries of animals. In the mid-19<sup>th</sup> century, medical interests in menopause began to increase considerably. In the 1970s, menopausal symptoms were attributed to estrogen deficiency, and HT was widely promoted to alleviate symptoms and improve the quality of life for middle-aged women [74].

In 1938, the development of synthetic estrogen marked a significant milestone. It opened the door for the medical industry, particularly pharmaceutical companies, to play a major role in addressing menopause-related concerns [74]. Nonetheless, HT use has very controversial history. It started in the 1960s and gained popularity during the 1990s. However, the release of the initial results of the Women's Health Initiative report in 2002 revealed that HT had more negative than positive effects. This received extensive media attention and changed the view on these drugs, causing a significant decline in HT use. In following years, the WHI trial was reanalysed, and new research emerged showing that the use of HT in younger women or those in the early postmenopausal stage had positive effects on the cardiovascular system, reducing the risk of coronary disease, and all-cause mortality [75]. Despite this, the public perception of HT remained rather negative, resulting in detrimental consequences for women's health and quality [75]. Still, it is recommended to consult one's doctor to assess individual HT-related risks and benefits before use [73].

### **1.1.3.2 Prevalences and time trends**

The global population of postmenopausal women is increasing. In 2021, women aged  $\geq 50$  years accounted for 26% of the total female population, marking an increase from 22% a decade ago [9]. Moreover, women are experiencing longer lifespans, as in a global context, a woman aged 60 years in 2019 could anticipate an average additional lifespan of 21 years [9]. With life expectancy for women on the rise globally, most of them will also undergo the menopausal transition and spend considerable years of their lives in the menopausal phase [76].

More than 80% of women are affected by symptoms associated with menopausal transition, and around 1/3<sup>rd</sup> of these women might experience severe symptoms [77]. Vasomotor symptoms (i.e., common symptoms experienced during menopausal transition) usually last for 1-6 years but might persist for 15 years or even more for 10-15% of women [76]. Prevalence rates of menopausal symptoms vary across different geographical regions: Europe (74%), North America (36-50%), Latin America (45-69%), and Asia (22-63%) [78]. The median age at menopause also varies across different geographical regions: Europe (50.1-52.8 years), North America (50.5-51.4 years), Latin America (43.8-53 years), and Asia (42.1-49.5 years) [78]. Further, according to a survey [79], the most common menopausal symptoms reported were feeling tired or worn out (Europe/US 74%, Japan 75%), aching in muscles and joints (Europe 69%, US 68%, Japan 61%), difficulty sleeping (Europe 69%, US 66%, Japan 60%), and hot flashes (Europe 67%, US 68%, Japan 62%). Around 50% of menopausal women experience the impact of vaginal estrogen deficiency [80] and 40% of women experience sleep difficulties/disorders [81].

There was high prevalence of menopausal HT use at its initial phase of development. However, studies showed sharp decline (with some variability) on its use after 2002 [82,83], likely due to the concerns for detrimental health effects of HT use demonstrated by a Women's Health Initiative report [83]. In 2002, the prevalence of menopausal HT use among women aged 45-69 in European countries exhibited variation across countries, with rates ranging from less than 5% to over 25% [82]. However, a substantial decline (50-77%) in menopausal HT use was observed in all European countries between 2002 and 2010. By the end of 2010, the estimated proportion of women aged 45-69 years using menopausal HT use had dropped to below 10% in all 17 European countries studied (including Norway), except for Finland [82].

### **1.1.3.3 Physiological and health effects**

Physiologically, the start of menopause is intricately linked to the cessation of ovarian function, resulting in a decrease in estrogen production by the follicles [84]. The hormonal fluctuations around menopause can trigger a range of symptoms affecting women's physical, emotional, mental, and social well-being, and overall quality of life [9]. Importantly, the immune system is affected by physiological changes that occur during menopause. These changes include a decrease in immune cells, which is not only attributed to the natural aging process but also to the deprivation of estrogen [85]. This is supported by studies that indicate an elevation in chronic systemic inflammation following menopause [86].

The absence of sex hormones can cause alterations in thermoregulation within the central nervous system. This gives rise to the most common symptoms women experience during menopausal transition –vasomotor symptoms (including hot flushes and night sweats) [76]. These symptoms can lead to sleep disturbances, tiredness, irritability, and low mood, impacting the overall quality of life of women. Severe and prolonged vasomotor symptoms are linked with an increased risk of future cardiovascular disease, likely attributed to the loss of protective effects provided by estrogen [76]. Further, menopause is also characterized by symptoms such as urogenital atrophy (including vaginal dryness, burning, irritation, tissue shrinkage, and painful intercourse), bladder symptoms (such as urgency and frequency), and pelvic organ prolapse. All of these if left untreated, increases the risk of urinary tract infections [87]. Reduced levels of estrogen hormone in the body can cause thinning of the vaginal mucosa, reduction in superficial cells, decreased glycogen and lactobacilli, and an increase in the pH of vaginal secretions (pH>5). Vaginal smooth muscle also changes due to reduction in collagen and elastin content [76]. Further, sexual dysfunction, including dyspareunia (discomfort/pain experienced during sexual intercourse), decreased libido (sex drive), difficulties with arousal, and problems achieving orgasm, can be attributed to the decrease in sex steroids associated with aging and the transition into menopause [88]. Low estrogen levels has been linked to sleep disturbances, hot flushes, anxiety, depressive symptoms, leading to depressive symptoms and depressive disorders [77,81]. Some also experience decline in cognitive function (referred to as "brain fog") [87]. Also, decreasing levels of estrogen can negatively impact connective tissue, joints, bone matrix, and skin [80], and muscle aging and loss of muscle mass [76].

Menopause and cancer risk is debatable. Menopause itself does not appear to directly cause cancer, but the likelihood of developing cancer rises with increasing age for women. A study reported that the risk factors for breast cancer were similar between pre-menopausal and post-menopausal women; however, slightly higher incidence of breast cancer in postmenopausal women was observed who had late menopause, potentially because of the longer duration of hormonal exposure [89].

HT treatment options are available that utilizes exogenous hormones (estrogen and/or progesterone) to alleviate these menopausal symptoms [73,80]. Among the available interventions, HT is considered the most effective treatment for managing troublesome vasomotor symptoms [73,76]. However, research on the risks and benefits of HT use suggests that various factors such as age at start of HT use, duration of use, type of hormone used [73,80].



Further, history of cancer related to systemic HT use is considered contradictory. HT (estrogen alone) is associated with little or no change in the risk of breast cancer while combined HT can be linked to an increased risk of endometrial cancer, and slightly increase the risk of developing ovarian cancer [80]. Further, HT is recommended as the primary treatment for preventing and treating osteoporosis in women with premature ovarian insufficiency and menopausal women under 60, particularly those experiencing menopausal symptoms [80]. Reports indicate that starting HT use before the age of 60 or within 10 years of menopause reduces atherosclerosis progression, coronary heart disease, cardiovascular mortality, and all-cause mortality [75].

#### **1.1.3.4 Associations with gene expression and research gaps**

Investigating molecular effects of menopausal status and HT use in the female body can provide valuable knowledge of this process. One study investigating gene expression profiles in peripheral blood monocytes among healthy pre- and postmenopausal women with a small sample size showed that the functional state of circulating monocytes can be influenced by menopause resulting in changes in gene expression profiles [84]. Still, no study has examined the relation between menopause and gene expression in whole-blood with a large sample size.

Transcriptional differences were observed in women receiving low-dose compared to higher conventional-dose of HT (17 $\beta$ -estradiol/norethisterone acetate) [90]. Other studies investigating the correlation between blood sex hormones and gene expression [91] and gene expression related to breast cancer in HT users-vs-non-users [92] have also identified several genes associated with different types of HT use. A study with small sample size (N=100) on HT use and blood gene expression found no significant difference between HT users-vs-non-users [93]. Yet, a population-based study with large sample size investigating HT use and blood gene expression had not been investigated.

## **1.2 Gene expression profiling**

### **1.2.1 Background: From DNA to proteins**

DNA (deoxyribonucleic acid) is a long, winding molecule that contains the biological instructions, which make each species unique [94]. Specifically, genes, the functional units of DNA, are considered as the fundamental hereditary units that are passed from parents to their offspring during reproduction, carrying the genetic information required to determine physical and biological characteristics [94,95]. In human, the complete DNA set, called 'genome', contains approximately 20,000-25,000 genes [94]

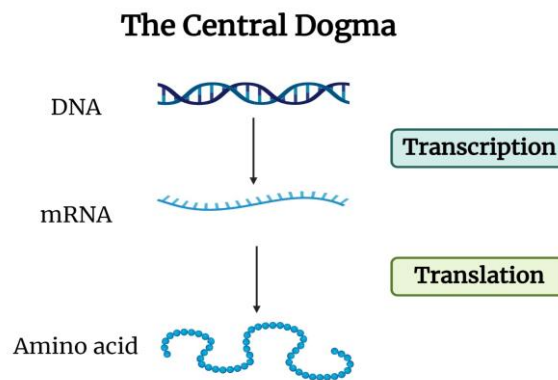
*Gene expression* refers to the process through which the encoded information within a gene is transformed into a functional outcome (i.e., proteins) [96]. It is based on directional flow of information from DNA → RNA → protein, called ‘the central dogma’ of molecular biology (Figure 2) [97]. Gene expression can be conceptualized as an "on/off switch" that governs when and where RNA (ribonucleic acid) molecules and proteins are made and as a "volume control" to regulate the quantity of these products [96]. Further, the full range of mRNA (messenger RNA) molecules expressed by an organism or in other words the collection of all the gene readouts present in a cell, is known as *transcriptome* [98].

In contrast with the genome, which is rather stable, the transcriptome is very dynamic, and continuously responds to various factors such as developmental stage, physiological, and environmental conditions [99]. *Gene expression analyses or gene expression profiling* represent measurements of the change in expression of thousands of genes at once, which provide a comprehensive picture of biological and cellular functions. Such profiles give the opportunity to answer various important research questions –from identifying phenotype-genotype relationships in genetic disorders to human molecular epidemiology. Further, such profiles have the potential to offer meaningful crucial insights into cellular biology, such as developmental state of cells, disease status, or response to environmental stimuli.

## The Central Dogma

The expression of gene that encodes a protein is carried out in two major steps:

1. **Transcription:** In this step, a gene's DNA sequence is replicated to produce an RNA molecule. This stage is named transcription since it involves copying or transcribing the DNA sequence into an RNA 'alphabet'. Eukaryotes require the RNA molecule to undergo processing to mature into mRNA.
2. **Translation:** In this step, the mRNA sequence is decoded to determine the proteins (or more precisely, polypeptides, a chain of amino acids). This stage is named translation as it involves the conversion of the mRNA's nucleotide sequence to amino acids that can build protein.



**Figure 2:** The steps involved in the central dogma in a eukaryote.  
(Created with [BioRender.com](https://www.biorender.com))

The directional flow of information from **DNA → RNA → protein** in a eukaryote explains the fundamental principle and is called *the central dogma* of molecular biology. This describes the basic process of how gene expression works.

### 1.2.2 Measurement techniques

Over the years, measurement techniques for gene expression analysis have undergone significant advancements. When gene expression analysis was at its infancy, the measurement techniques were also basic. Around 20 years ago, its measurement techniques started to undergo significant improvements, enabling higher throughput (the number of transcript samples measured per unit time), and more comprehensive analysis, allowing measurement of multiple genes simultaneously. Currently, due to rapid technological progress, these techniques offer

higher sensitivity, throughput, and enables the study gene expression at the single-cell level. Here, we provide a short overview of some of these techniques.

*Northern Blotting*: Northern blotting is hybridization-based assay technique used to determine the size and quantity of specific RNA molecules in a sample. It involved the separation of RNA molecules using gel electrophoresis, followed by their transfer onto nitrocellulose membrane, and then using labeled probes to hybridize with the target RNA. It is simple and inexpensive, but it is time-consuming, low throughput, and there is risk of RNA degradation during electrophoresis [100].

*Quantitative polymerase chain reaction (qPCR)*: qPCR is a method to quantify gene expression in real time, using a spectrophotometer. It is a PCR quantification-based assay technique. The initial RNA processing step includes reverse transcription of mRNA to complementary DNA (cDNA). It is easy to use and less time-consuming. However, it is not an exploratory method and is only used for quantification of known genes [100].

*Microarray*: Microarray technique has revolutionized gene expression analyses by enabling the simultaneous measurement of thousands of genes. It has been used to measure gene expression profiles for over 20 years now [100]. There are two types of microarrays: oligonucleotide microarrays and cDNA microarrays [101]. Microarray utilizes the principle of nucleic acid hybridization of cDNA strands to measure gene expression profiles [100]. Microarray experiment is a multistep process. First, RNA is extracted from tissue and converted to cDNA. The cDNA is then labeled and transferred to a microarray chip (a glass slide) for hybridization with immobilized probes. After hybridization step, the chips are washed in salt buffer, and the hybridized, tagged, fluorescent-labeled nucleic acid sequences remain on the chip. The intensity of the fluorescent signals is detected by scanning the hybridized microarray chip, that is reflected in the amount of transcribed RNA. The relative amount of RNA undergoes quality control steps and is usually preprocessed before it is ready for statistical analyses [100]. This technique became a preferred technology for analysing transcriptomes, due to its affordability, user-friendliness, and optimized framework of quality control [102]. Other advantages are that it can be used for exploratory and high throughput analyses. It is beneficial for genome-wide association studies, and prior knowledge of complete sequences is not required, still the probes need to be known. The limitations are that it can be time-consuming and may increase the

variability of output data. Also, it is dependent on specialized software for RNA quantification through image processing [100,103].

*RNA sequencing (RNA-seq):* RNA-seq has become a widely used technique and has today supplanted microarrays as the technology of choice for transcriptome-wide analyses. The initial RNA procession step includes reverse transcription of mRNA to cDNA or labelling of miRNA, tRNA, and rRNA, and subsequently sequencing it using high-throughput methods. RNA-seq provides accurate quantitative measurements of gene expression, enables the detection of alternative splicing events, and allows the identification of novel transcripts and non-coding RNAs. Like for the microarray technique, these can be high throughput analyses and useful for genome-wide association studies; but RNA-seq generally has less background noise and better detection than microarrays. In comparison to microarrays, RNA-seq has the capability to identify 30% higher number of DEGs. The limitations of this technique are that it is expensive and time-consuming. It also results in a high computational and data-storage burden [100].

### ***Recent technologies***

The focus of current technologies is on single-cell and spatial transcriptomics; however, the technologies developed for single-cell and spatial analyses each rely on one or several of these basic measurement technologies, e.g., RNA-seq for drop-seq (including 10x's Chromium technology) and microarray-like slides for location barcoding combined with RNA-seq for spatial transcriptomics (10x's Visium technology) [104].

*Single-cell RNA Sequencing (scRNA-seq):* While traditional bulk sequencing has contributed significantly to our understanding of biology by analysing an average population of cells, the development of single-cell sequencing has revolutionized the field [105]. Single-cell transcriptomics, or scRNA-seq has emerged as the cutting-edge technology that enables the examination of cellular functions at a fine-grained level, unravelling the heterogeneity and complexity of RNA transcripts within individual cells. It also enables the identification of rare cell types and investigation of dynamic changes in gene expression within individual cells [106]. However, single-cell sequencing separates cells from their original tissue, losing the spatial context [105]. In situ hybridization is the most relevant method for validating cell-type-specific gene expression in solid tissue.

*Spatial analysis:* Spatial analysis encompasses two primary methods. The first method involves fluorescence in situ hybridization -FISH method, where transcripts are directly labeled within tissue sections to visualize the locations of single-cells [105]. The other method is spatially resolved transcriptomics, recognized as the Method of the Year by Nature in 2020 [107]. Expanding upon scRNA-seq, spatial transcriptomics employs oligonucleotide microarrays to capture RNA transcripts throughout a tissue section, which are subsequently sequenced using next-generation sequencing. This facilitates the generation of high-resolution tissue maps coupled with associated transcriptomic data [105]. Further, multiplexed in situ hybridization is an essential technology for some of the new technologies for spatial transcriptomics, such as MOSAICA [108] and 10x's Xenium platform [104].

*NanoString Technology:* NanoStrings are pioneers within the field of spatial biology, offering solutions, workflows and products for spatial transcriptomics, spatial proteomics, and single-cell spatial multi-omics [104]. It is multiplex nucleic acid hybridization technology that enables reliable and reproducible assessment of the expression of several genes in a single assay [109].

### **1.2.3 Target tissue for measurements**

Gene expression studies can be performed in various human tissue samples, such as whole-blood, peripheral blood leukocytes, breast biopsies, gut microbiomes. It is crucial to determine which tissue, for example, muscle biopsy or blood or which component of blood, is most suitable to address the research question. Further, the selection of tissue should have careful considerations with respect to tissue availability, sampling protocol, storage management, etc. [100]. The use of circulating blood (whole-blood or peripheral blood mononuclear cells (PBMCs)) as a tissue for gene expression analyses is widespread, as blood samples have relatively simple collection processes and could be stored for long-term use [100,110]. It has been claimed that blood could be regarded as an alternative to tissue samples for molecular profiling of human disease and disease risk [110]. Still, other studies have demonstrated differences between blood and organs affected by the disease [111].

Blood is a circulating "connective" tissue comprised of plasma (55%) and formed elements (i.e., cellular components) (45%) [110]. The cells found in blood comes from bone marrow: they begin as stem cells and mature into three main types of blood cells –red blood cells (RBCs or erythrocytes) (96%), white blood cells (WBCs or leukocytes) (1%), and platelets (thrombocytes) (3%) [110]. *RBCs* are the predominant cells found in whole-blood responsible

for oxygen transport. When RBCs matures, they lose their nucleus and organelles and therefore, it is generally believed that there is no RNA contribution to the overall blood RNA pool from these cells. Still, there is evidence that RBCs contribute a small fraction of the total RNA molecules in whole-blood [112,113]. Consequently, even though only a small fraction is contributed, the high proportion of RBCs in whole-blood combined with RBCs' high levels of globin RNAs, contribute to globin RNAs found highly expressed in bulk whole-blood transcriptomic profiles [114]. Further, the immature RBCs (*reticulocytes*) may contain residual nucleic acids, contributing to the total blood RNA pool. *Platelets* have vital function in the process of blood clotting. Similar to RBCs, mature platelets lack nucleus and organelles. However, there are immature platelets (reticulated platelets), which retain some RNA, that can contribute to total blood RNA [113]. *WBCs* are nucleated blood cells that are responsible for carrying out the immune functions within the body. WBCs are divided into mainly three types – monocytes, lymphocytes (B cells, T cells, and natural killer cells), and granulocytes (neutrophils, basophils, and eosinophils). WBCs are the most transcriptionally active cells in blood, making them the primary focus of gene expression studies involving blood [115]. Whole-blood consists of RBCs, WBCs, and platelets suspended in plasma, while PBMCs incorporate only mononucleated cells in blood i.e., monocytes and lymphocytes (e.g., T cells and B cells), which can be isolated from whole-blood.

One of the major obstacles linked with blood transcriptomes are their heterogenous cell populations, as the proportions of distinct cell types differ among different individuals. This heterogeneity results to be a source of variation in blood gene expression profiles [100,102], and this could also influence the average signal in blood. But technology such as scRNA-seq can solve the heterogeneity and complexity of blood transcriptome as it offers specific-cell-types analyses. However, they are expensive, laborious, and time-consuming.

#### **1.2.4 Dynamics of gene expression profiles**

One of the key aspects of gene expression profiles is their dynamic features. Gene expression is characterized by its remarkable dynamism, undergoing fluctuations influenced by a multitude of internal and external stimuli [99]. Further, researchers have observed variations in gene expression in various tissues, including whole-blood, over a 24-hr period and across different seasons [100,116]. Gene expression levels in whole-blood exhibit This issue can be considered in part technical, but it might as well be biological, as these variations during a day or different seasons might be due to the differences in 24-hour light-dark exposures and temperature cycles

and the difference in length of the day in different seasons, which are the circadian rhythm (“circadian” meaning “about a day”) [117]. Many physiological processes in our body exhibit daily fluctuations. These daily patterns in both behavior and bodily processes are not just acute responses to timing cues but are governed by an internal circadian timing system [117]. For instance, transcriptional response to hypoxia (low oxygen levels) and the specific time of day in humans has been observed [118]. Thus, it could be recommended that smaller studies should collect samples for gene expression analyses at the same time of the day or same season. Yet if not possible, the time and date should be recorded for the control of this potential source of variability [100].

These dynamics of gene expression can provide valuable insights into the regulation of biological processes, development, disease progression, and responses to external factors; and by their analyses, researchers can better understand the complex interactions and networks within cells and organisms. But it is also a challenge considering the technicalities. Expression studies measured at one time point represents a snapshot of gene expression in a set of samples. Thus, snapshot gene expression could not be expected to capture associations with past changes in exposure variables, if there are larger differences in time of the exposure variables and the gene expression profiles [119,120].

In recent years, significant progress has been made in understanding dynamic transcriptional responses. With increasing knowledge of functional genomics there has been a transition from a static view of transcription to a comprehensive understanding of its dynamic nature, both at the cellular and single-cell levels. Any transcriptional regulation cannot be fully understood based on the fixed snapshots of the process. Technological advancements, such as continuous monitoring of gene expression by fluorescent and luminescent reporters single-cell tracing, have greatly improved our ability to examine dynamic aspects of gene activation [99].



## **2 Aims of the thesis**

Linking gene expression in blood and common cancer risk factors or lifestyle factors is an intriguing approach for gaining valuable insights into the physiological and molecular processes of genes influenced by selected risk factors (exposure variables). Understanding these processes might ultimately help understand gene expression in the development and progression of cancer and other diseases. Especially, the main risk factors covered in this thesis – smoking, obesity, and menopause – are of particular interest due to limited or no research with these risk factors within blood gene expression. Moreover, studies utilizing large and extensive population-based samples to assess such relationships are rare.

To address these research gaps, this thesis aimed to unravel the associations of gene expression in blood with important cancer risk factors among women, who had never been diagnosed with cancer, in the NOWAC postgenome cohort; and to further explore in more detail the gene ontologies (GO) and pathways of genes that were differentially expressed between different levels of exposure variables.

Specific objectives are listed below:

To evaluate (1) the DEGs in blood of women and (2) their biological functions according to following risk factors:

- i. Smoking status and smoking metrics (Paper I)
- ii. BMI and weight changes (Paper II)
- iii. Menopausal status and HT use (Paper III)

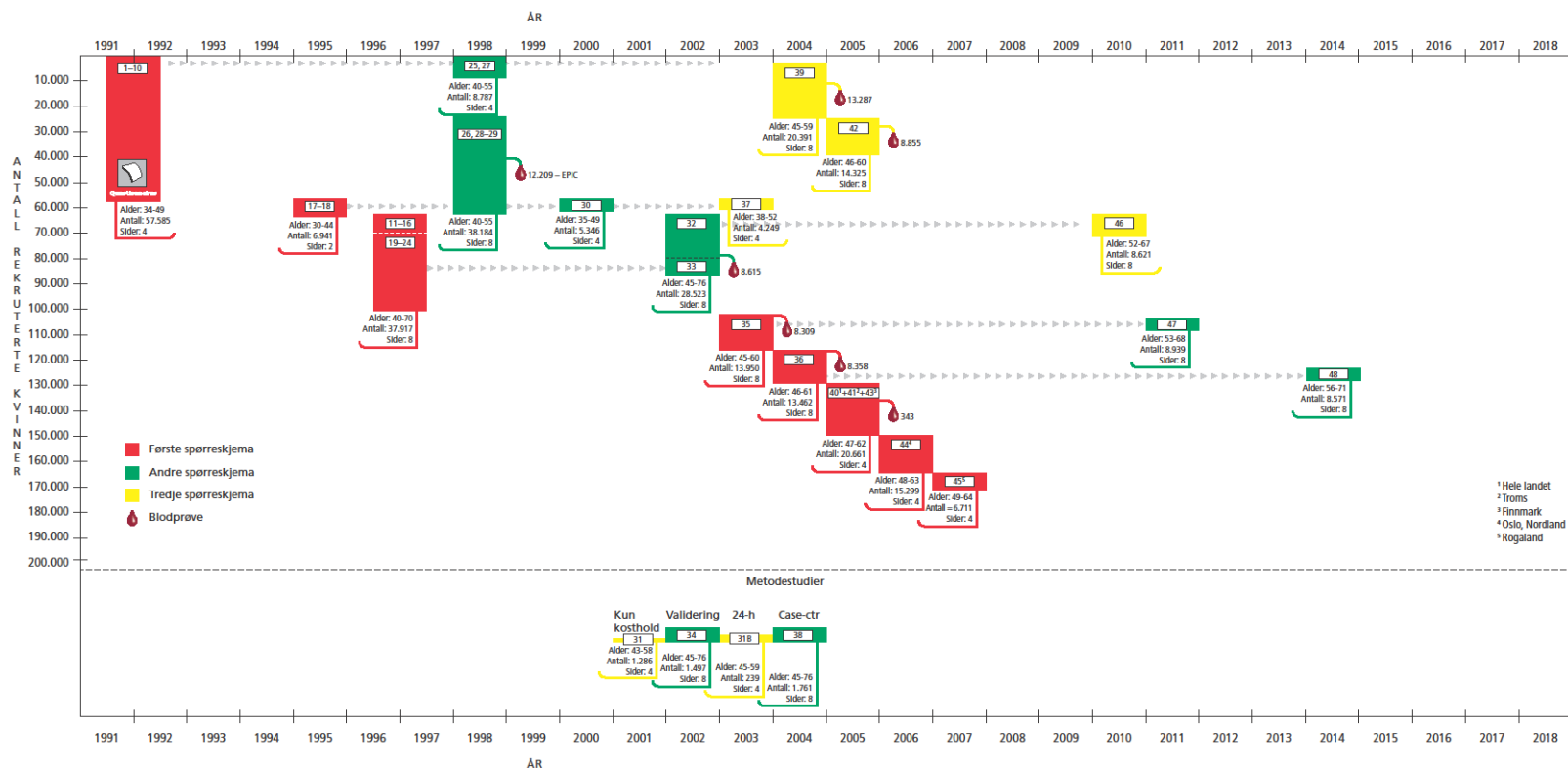


## **3 Methods and materials**

### **3.1 Study population**

#### **3.1.1 The Norwegian women and cancer study (NOWAC)**

The NOWAC study (*in Norwegian: Kvinne og Kreft studien*) is a nation-wide, population-based prospective cohort study, initiated in 1991. Its baseline purpose was to explore the original hypothesis of the use of oral contraceptives being a risk factor for breast cancer, as well as to explore other risk factors for breast cancer, but later it has expanded to target a wider range of scope by including biological material for the whole-genome expression profiling using microarray technique [4]. NOWAC study is based on questionnaires mailed to women aged 30-70 years, collecting detailed data related to lifestyle and health such as smoking exposure, height, weight, reproductive history, HT use, alcohol consumption, family history of breast cancer, dietary patterns, use of medication, and others. The participants were randomly selected from the National Population Register (*in Norwegian: Folkeregisteret*) of Norway. From 1991 up to June 2007, it incorporated approximately 172,000 women [121]. After the first invitation, each woman had answered between one and three follow-up questionnaires (main questionnaires) (Figure 3). The NOWAC study database is annually updated with information from the Cancer Registry of Norway (*in Norwegian: Kreftregisteret*), as well as the Norwegian Cause of Death Registry (*in Norwegian: Dødsårsaksregisteret*). Extensive descriptions about the NOWAC study are available in Lund et al. [4].



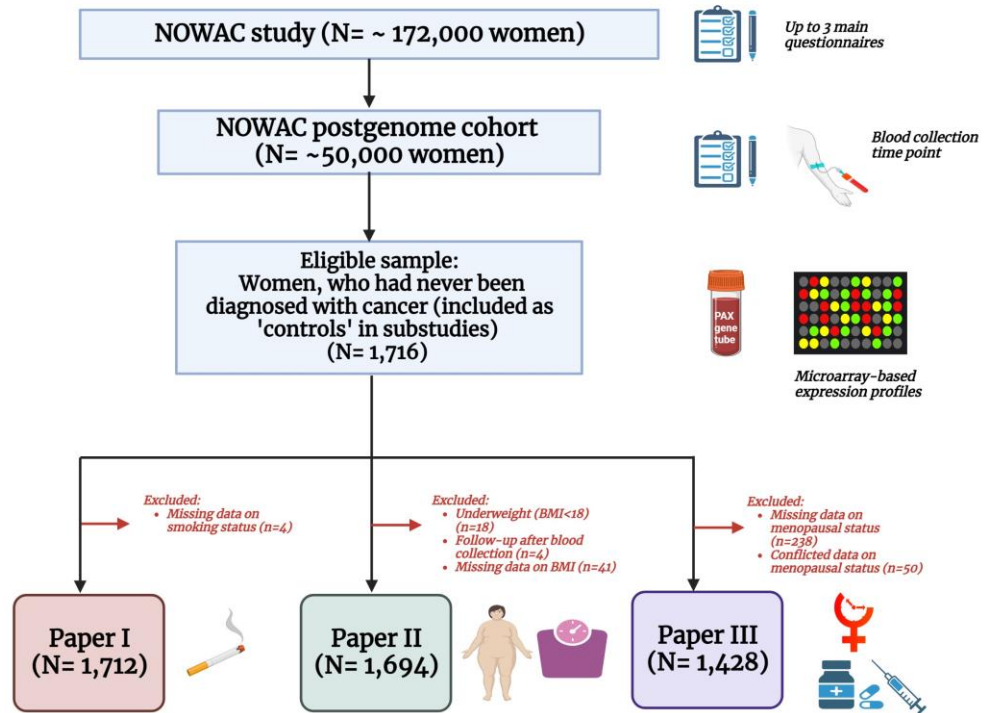
**Figure 3:** Timeline for the Norwegian Women and Cancer Study (NOWAC).

Number of women recruited during first (red boxes), second (green boxes) and third (yellow boxes) questionnaire mailings, and collection of blood samples (red blood drops) within the NOWAC study according to year of enrolment, age, and number of participants.

### **3.1.2 The NOWAC postgenome cohort**

The NOWAC postgenome cohort is a sub-cohort of the NOWAC study, that has collected biological materials, mainly peripheral blood but also some normal and malignant breast tissue, for the whole-genome expression profiling using microarray technique [4]. It consists of approximately 50,000 women, who between 2003 and 2006, had blood samples drawn using the blood collection kit, which were suitable for later mRNA (or transcriptomics or gene expression) analyses due to RNA stabilizing buffers [93,122]. The blood samples were drawn at the women's local general physician's offices. Of these, several have been selected for Illumina full-genome mRNA microarray analyses related to case-control studies within NOWAC. For RNA collection, PreAnalytiX (PAXgene™) Blood RNA tube was used, which is a BD Vacutainer™, containing a proprietary reagent that immediately stabilizes intracellular RNA. The samples were sent overnight to the Department of Community Medicine at UiT The Arctic University of Norway. The women were asked not to have their blood samples drawn on Thursdays and Fridays, to avoid weekend mail delays. The NOWAC biobank staff received the blood samples in most cases within 1-2 days (92%). PAXgene™ Blood RNA tubes were frozen at -20°C and then transferred to -70°C without being pre-processed within a week [121]. Additionally, the participants had answered a less extensive questionnaire about their lifestyle at the blood collection time point. Consequently, there are large number of samples from disease-free women that have been compiled, quality-assured, and pre-processed.

In this thesis, we included microarray-based expression profiles in bio-banked whole-blood samples from women in the NOWAC postgenome cohort. We obtained relevant questionnaire and registry information from NOWAC databases. There were several case-control studies (i.e., breast, lung, ovarian, and endometrial cancers, and diabetes) that had investigated gene expression profiles in the NOWAC postgenome cohort. The eligible participants were those women who were originally included as controls in these case-control studies, so that the study sample only included women who had never been diagnosed with cancer. Therefore, we excluded: (i) all cases from these case-control studies, (ii) those who were present in more than one sub-study, i.e., present in both the prospective and the post-diagnostic sample, (iii) who got cancer after being selected as a control until 2017, and (iv) those with missing information. Details about study participants with inclusion and exclusion criteria are presented in Figure 4 below. The final dataset considered for the all papers (Paper I-III) were based on 1,716 eligible women.



**Figure 4:** A flow chart of the study populations.  
(Created with [BioRender.com](https://www.biorender.com))

### 3.1.3 The study design

We compared gene expression profiles in whole-blood samples from women, who had never been diagnosed with cancer, according to the different levels of selected cancer risk factors. There were up to four time points for collection of questionnaire information, including up to three before blood collection (main questionnaires) and one at blood collection time point (blood questionnaires). Thus, this represents repeated measurements of exposure variables. However, the focus of this work was analyses of gene expression data, therefore, we had considered the blood collection time point as the main study time point or the reference point. Since, there was only a single time point for blood collection that were eligible for the transcriptomic analyses, capturing a snapshot of gene expression data. Therefore, the study design in all papers (Paper I-III) included in this thesis was a cross-sectional analysis nested within the prospective longitudinal NOWAC study.

## 3.2 The gene expression data

### 3.2.1 Laboratory analyses

The PAXgene™ Blood RNA kit protocol was followed for extraction and purification of total RNA from PAXgene™ Blood RNA tube samples. This was performed at the Genomics Core

Facility, Norwegian University of Science and Technology, Trondheim. RNA purity was evaluated using a NanoDrop ND 8000 spectrophotometer (ThermoFisher Scientific, Wilmington, DE, USA), and RNA integrity was evaluated using bio-analyser capillary electrophoresis (Agilent Technologies, Palo Alto, CA, USA). The Illumina TotalPrepT-96 RNA amplification kit was utilized to create the cDNA and then the cDNA was hybridized to Illumina Human WG-3 or HT-12 expression bead chip microarrays. Then, raw microarray images were processed using Illumina Genome Studio. The lab analyses were performed between January 2011 and January 2015.

### **3.2.2 Pre-processing of dataset**

The gene expression dataset was pre-processed prior to the main statistical analyses and the pre-processing included several steps as described here. First, technical outliers (identified by the laboratory quality measures) and any outliers that were detected from a standard operating procedure with *nowaclean* [123] were removed. This was performed for each case-control study sample set separately using principal component analysis (PCA) and density plots. Then, background correction, removal of bad quality probes, and filtering of probes detected in less than 20% of samples were performed, also for each case-control study sample set separately. Further, the sub-sets of data were merged and all women with cancer diagnoses until the end of follow up in 2017 were removed. The merged dataset was then processed further using  $\log_2$  transformation, quantile normalisation; and inspection of batch effects using PCA plots was performed. We then performed gene annotation using the Bioconductor packages '*lumi*', '*lumiHumanIDMapping*', and '*illuminaHumanv4.db*' [124-126]. As we had stringent filtering criteria for detection, it rendered in 9,095 probes. In case of more than one probe annotated to each gene, the probe with the largest inter-quartile range was kept. The final dataset contained 7,713 unique genes to be analysed as an outcome.

Our study sample did not have information on counts of reticulocytes, RBCs, or WBCs. However, during pre-processing, for the eligible women, we estimated the proportions of 22 populations of WBCs in the samples using an *in-silico* gene expression deconvolution software CIBERSORT and the LM22 signature matrix [127].

### **3.3 Selected exposure variables -formation and definition**

We obtained relevant questionnaire information from the NOWAC databases for the eligible women. The different exposure variables that were analysed in all the papers (Paper I-III) of

this thesis were based on information from all three main questionnaires and the questionnaire that were filled at the blood collection time point. Details about these exposure variables in each paper are given below:

### 3.3.1 Smoking status and smoking metrics (Paper I)

We had access to detailed information regarding past and current smoking exposures, including ages at smoking initiation and cessation, average number of cigarettes smoked per day across age intervals, and details about passive smoking. Smoking status and smoking metrics (i.e., smoking intensity, smoking duration, time since smoking cessation (TSC), pack-years, and comprehensive smoking index (CSI) scores) were created based on this information.

Smoking status was divided into three categories: current, former, and never smokers. Current smokers were defined as those who were currently smoking at the blood collection time point, former smokers were those who reported smoking cessation prior to the blood collection time point, and never smokers were those who reported they had never smoked either prior to or at the blood collection time point. Further, current smokers and former smokers combined represented ever smokers, and we defined passive smokers as those who were passively exposed to smoking at their homes as adults.

Smoking intensity was defined as the average number of cigarettes smoked per day during years of active smoking. Smoking duration was defined as the duration of active smoking in years. TSC was defined as the time since quitting smoking in years. Pack-years quantify individual, long-term exposure to tobacco smoking [128] and was calculated by the formula: *Number of pack-years = (smoking intensity/20) × smoking duration*. We considered 20 cigarettes in 1 pack, which is standard in the Norwegian context. CSI score is a cumulative measure of smoking exposure that incorporates smoking intensity (int), smoking duration (dur), and TSC (tsc). It was calculated using the formula [12]:

$CSI = (1 - 0.5^{\text{dur}^*/\tau})(0.5^{\text{tsc}^*/\tau}) \ln(\text{int} + 1)$ , where  $\tau$  is an estimated half-life parameter, and  $\delta$  is an estimated lag time parameter describing TSC and total duration as follows:

$\text{tsc}^* = \max(\text{tsc} - \delta, 0)$  and  $\text{dur}^* = \max(\text{dur} + \text{tsc} - \delta) - \text{tsc}^*$ .

### 3.3.2 Body mass index and weight change (Paper II)

We calculated BMI by dividing weight in kg by the square of height in meters, and the BMI values were then categorised according to the WHO standard (underweight: <18.5 kg/m<sup>2</sup>,



normal-weight: 18.5-24.9 kg/m<sup>2</sup>, overweight: 25.0-29.9 kg/m<sup>2</sup>, obesity:  $\geq 30.0$  kg/m<sup>2</sup>) [42]. For ease, in Paper II, we had represented the three questionnaires as: Q1 (baseline), Q2 (follow-up), and Q3 (blood collection). We calculated weight changes between Q1 and Q3 (WC<sub>Q3-Q1</sub>, mean interval 7 years), and between Q2 and Q3 (WC<sub>Q3-Q2</sub>, mean interval 1 year). We also defined weight change categories based on patterns of weight change between Q1-Q2 and Q2-Q3: consistent stable weight (women with stable weight (-2 to +2 kg) at Q1-Q2 and Q2-Q3); consistent weight gain (women with weight gain (above +2 kg) at Q1-Q2 and Q2-Q3); consistent weight loss (women with weight loss (below -2 kg) at Q1-Q2 and Q2-Q3); former weight gain (women with weight gain at Q1-Q2 and stable weight at Q2-Q3); former weight loss (women with weight loss at Q1-Q2 and stable weight at Q2-Q3); recent weight gain (women with stable weight at Q1-Q2 and weight gain at Q2-Q3); and recent weight loss (women with stable weight at Q1-Q2 and weight loss at Q2-Q3).

### **3.3.3 Menopausal status and hormonal therapy use (Paper III)**

Menopausal status was determined using data from all main questionnaires and the one filled at blood collection time point. Participants were classified into three categories: (i) premenopausal women, i.e., women who had regular menstrual cycles, (ii) perimenopausal women, i.e., women who had irregular menstrual cycles and possibly vasomotor symptoms, and (iii) postmenopausal women, i.e., women who had experienced 12 consecutive months without a menstrual cycle.

Self-reported HT use was further categorized as systemic HT use (oral or trans-dermal) and local HT use (vaginal cream or suppository).

## **3.4 Statistical analyses**

We performed all statistical analyses using the open-source software R [129], 3.2.1 and 3.6.2 versions for Paper I, 3.6.2 and 4.0.5 versions for Paper II, and 4.0.5 version for Paper III. We used the open-source Bioconductor package ‘*limma*’ (linear models for microarray and RNA-seq data) [130] for gene-wise linear models testing for DEGs in all papers (Paper I-III). Due to the high number of statistical tests, we considered using Benjamini-Hochberg correction to correct for false discovery rates (FDR), with a significance threshold of  $FDR \leq 0.05$  [131].

### **3.4.1 The selected exposure variables**

#### **3.4.1.1 Smoking status and smoking metrics (Paper I)**

Three comparisons of smoking status were used to determine the presence of DEGs: current-vs-never smokers, current-vs-former smokers, former-vs-never smokers. Further, we examined smoking metrics within current smokers and former smokers separately, as well as for adult passive smokers within never smokers. Eight women had missing information on smoking status, resulting in an analytical sample of 1,708 women for these analyses. We also used linear regression to examine the relationships between WBC proportions and smoking metrics.

*Sensitivity analyses:* Data on DNA methylation at the CpG site *AHRR* gene, cg05575921, was also available in a subset of participants (N=324) [132]. Thus, we used receiver operating characteristic (ROC) curves to compare the ability of the top-ranked gene in our analyses and a CpG site in the *AHRR* gene (cg05575921). The t-test and Wilcoxon rank sum tests were used to compare differences in average expression and log<sub>2</sub>fold-change (logFC) between DEG groups, respectively.

#### **3.4.1.2 Body mass index and weight change analyses (Paper II)**

##### ***BMI analyses***

For categorical BMI analyses, we evaluated the relationship between blood gene expression and BMI at blood collection time point (BMI<sub>Q3</sub>) categories in three different comparisons: obesity-vs-normal-weight, obesity-vs-overweight, and overweight-vs-normal-weight. For continuous BMI analyses, we modelled BMI<sub>Q3</sub> as a continuous standardised metric and scaled it using the R function '*scale*'. The final analytical sample for these analyses were 1,653 women (missing information for BMI: N=41).

##### ***Weight change analyses***

For categorical weight change analyses, we evaluated the relationship between blood gene expression and past weight changes in six comparisons: (i) consistent weight gain vs consistent stable weight, (ii) consistent weight loss vs consistent stable weight, (iii) former weight gain vs consistent stable weight, (iv) former weight loss vs consistent stable weight, (v) recent weight gain vs consistent stable weight, (vi) recent weight loss vs consistent stable weight. Here, we excluded weight-cyclers (N=160), i.e., women who reported decreased weight at Q1-Q2 and increased weight at Q2-Q3 and vice versa, and women with missing values (N=499). The final

analytical sample for these analyses were 1,035 women. *Sensitivity analyses*: we also conducted analyses restricted to women with <1 year between Q2 and Q3.

We modelled weight change as a continuous metric in two interaction models with BMI categories (WC-BMI interaction analyses) to assess trends across these categories. To address the differences in the intervals of  $WC_{Q3-Q1}$  and  $WC_{Q3-Q2}$ , we divided the absolute values of weight change (kg) by the number of years between Q3 and Q1 or Q2 (kg/year) before scaling it (R function 'scale'). The first interaction model included  $BMI_{Q1}$  or  $BMI_{Q2}$  and subsequent weight changes (i.e.,  $BMI_{Q1} * WC_{Q3-Q1}$  or  $BMI_{Q2} * WC_{Q3-Q2}$ ), while the second interaction model included  $BMI_{Q3}$  and prior weight changes (i.e.,  $BMI_{Q3} * WC_{Q3-Q1}$  or  $BMI_{Q3} * WC_{Q3-Q2}$ ). Thus, the two interaction models allowed for different approaches to the weight changes in terms of starting from the prior point or from the blood sampling point. We excluded 464 and 82 women with missing values for  $WC_{Q3-Q1}$  and  $WC_{Q3-Q2}$ , respectively, resulting in analytical samples of 1,230 and 1,612 women, respectively. *Sensitivity analyses*: To evaluate the influence of extreme weight change values, we conducted analyses in which we assigned weight change values that were below the 5<sup>th</sup> percentile and above the 95<sup>th</sup> percentile to the values of the 5<sup>th</sup> and 95<sup>th</sup> percentiles, respectively. Additionally, to evaluate the importance of the unit of weight changes, we conducted weight change analyses using the unit of BMI/year rather than kg/year.

#### **3.4.1.3 Menopausal status and hormonal therapy use (Paper III)**

We compared DEGs based on menopausal status in three comparisons: (i) post-vs-pre, (ii) post-vs-peri, and (iii) pre-vs-perimenopausal status. In the analyses of differences according to menopausal status, we excluded postmenopausal women who were HT users (N=265), as they could resemble premenopausal women in sex hormone status [121]. This resulted in an analytical sample of 1,163 women for these analyses.

We compared DEGs according to HT users-vs-non-users among the postmenopausal women only. The analytical sample for this analysis were 1,197 women. *Sensitivity analysis*: To avoid the potential influence of local HT users, we compared DEGs according to HT users-vs-non-users among postmenopausal women who reported using only systemic HT (N=1,170), where we excluded 20 local HT users and seven with missing information.

### **3.4.2 White blood cell proportions**

Gene expression profiles can be influenced by the composition of different types of blood cells [133]. Therefore, we used the estimated the proportions of 22 types of WBCs [127] in different models of adjustments, to identify differences in cell type composition related to our exposure variables.

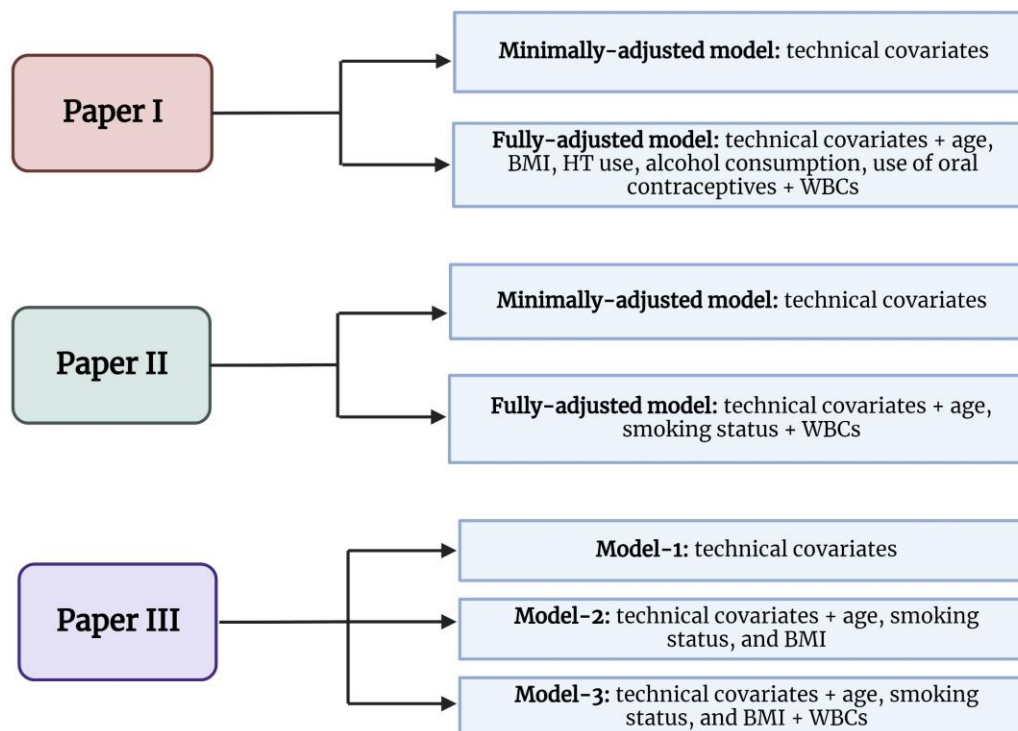
We evaluated WBC proportions that were significantly associated with smoking status (Paper I), BMI (Paper II), and menopausal status (Paper III) according to the Kruskal-Wallis test, and with overall gene expression data according to the Bioconductor package '*global test*' [134]. We then used these selected WBCs as covariates in limma analyses in different adjustment models.

### **3.4.3 Covariates and adjustment models**

We evaluated covariates that were significantly associated with smoking status (Paper I), BMI (Paper II), and menopausal status (Paper III) according to the Chi-square test (for categorical variables) or Kruskal-Wallis test (for continuous variables), and with overall gene expression data according to the Bioconductor package '*global test*' [134].

We considered laboratory plates (also called as: laboratory batch) and sample storage time as technical covariates. Further, we examined the distribution of following covariates across different exposure variables in different papers: age (years), smoking status (current/former/never smokers), BMI (normal-weight/overweight/obesity), and HT use at blood collection time point; alcohol consumption (units per week), parity (number of children), and use of oral contraception at baseline questionnaire. In addition, selected WBCs (as previously described) were included as covariates.

Possible confounders were addressed by adjusting for above mentioned covariates in the limma analyses in different models of adjustments. These models are demonstrated in Figure 5 below.



**Figure 5:** Different adjustment models across Paper I-III.  
(Created with [BioRender.com](https://BioRender.com))

### 3.4.4 Functional enrichment analyses

To investigate the biological functions of DEGs associated with different exposure variables, we performed functional enrichment analyses, specifically over-representation analysis (ORA). We used the open-source Bioconductor packages ‘*clusterProfiler*’ [135,136] and ‘*ReactomePA*’ [137]. For these analyses, in Paper I and II, we assessed DEGs from the fully-adjusted models, whereas in Paper III we assessed DEGs from both semi- (Model-2) and fully-adjusted models (Model-3). We conducted separate analyses for over-expressed ( $FDR \leq 0.05$  and  $\log_2\text{fold-change} (\log_2FC) > 0$ ), i.e., genes that are up-regulated (higher mean expression levels) in the reference group, and under-expressed genes ( $FDR \leq 0.05$  and  $\log_2FC < 0$ ), i.e., genes that are down-regulated (lower mean expression levels) in the reference group using the following databases: GO biological processes, GO molecular functions, GO cellular components, Kyoto encyclopaedia of genes and genomes (KEGG), and REACTOME pathways.

### **3.4.5 Quantitative replication**

For quantitative replication (the external validation), whenever possible, we compared our list of DEGs, fold-change values, and their directions associated with our exposure variables to the results from analyses of independent transcriptomic datasets from whole-blood and other relevant tissue types. This could help us to determine whether our results were consistent with previous research or represented as novel findings.

## **3.5 Ethical considerations**

The Regional Ethical Committee of North Norway (REK) has approved the collection and storage of data and human biological material in the NOWAC cohort and biobank (NOWAC postgenome cohort) (Reference numbers: 2010/2075/REK Nord and 2014/1605/REK Nord, respectively).

The invited participants were asked to sign if they did not want to give a blood sample, and thus avoid a reminder. The consent was given as a tick in the questionnaire to confirm that they had read the information and consented to participate. According to the assessment of the Norwegian centre of research data (*in Norwegian: Norsk Senter for Forskningsdata or NSD*) regarding the legal basis under the requirements in the Personal Data Protection Regulation today, it is appropriate that the basis for processing was public interest.

### **3.5.1 Data management**

The PhD candidate was given the access to the pre-processed data for analyses and was solely responsible for performing all the statistical gene expression analyses.

The NOWAC linkage to registries was performed using the personal identification number which is only stored at the Statistics Norway (*in Norwegian: Statistisk sentralbyrå or SSB*) and only they have access to it. All the participants had a unique LABNR in the databases, which served as the link between questionnaire data to gene expression data in this thesis. The data was stored and analysed in a safe environment (i.e., HUNT Cloud) which requires multi-factor authentication for access. HUNT Cloud (link: <https://www.ntnu.edu/mh/huntcloud>) is a data storage and computation facility offered by the Norwegian University of Science and Technology, Trondheim. Thus, the data were never transferred to personal computers or shared to other people or countries; and was only used to perform statistical analyses and make figures and tables.

### **3.5.2 Data availability**

Data could not be shared publicly because of local and national ethical and security policies. Data access for researchers will be conditional on adherence to both the data access procedures of the NOWAC study and the UiT The Arctic University of Norway (contact info: Tonje Braaten <[tonje.braaten@uit.no](mailto:tonje.braaten@uit.no)>) in addition to an approval from the local ethical committee. However, the project research findings were shared with the participants via NOWAC websites and in published articles.





## 4 Results – Summary of papers

### 4.1 Paper I

*Gene expression in whole-blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) Post-genome cohort [119].*

This study investigated associations of gene expression in blood with smoking status and smoking metrics among 1,708 cancer-free women (i.e., women who had never been diagnosed with cancer) from the NOWAC postgenome cohort. We performed gene-wise linear regression models to identify DEGs, and functional enrichment analyses to identify their biological functions.

Results showed that when compared to individuals who never smoked or formerly smoked, current smokers had 911 and 1,082 DEGs, respectively. The gene *LRRN3* showed the strongest association with current smoking status (logFC=1.01, FDR-adjusted p-value=1.52E-80). In the comparison between never smokers and all former smokers, there were no observed DEGs, but *LRRN3* was found to be differentially expressed when never smokers were compared to those former smokers who had quit smoking  $\leq 10$  years ago. Among current smokers, *LRRN3* was positively associated with smoking intensity, pack-years, and CSI scores, while among former smokers, it was negatively associated with TSC. The biological functions of the DEGs identified were linked to circulatory functions, translation, and immune responses.

In conclusion, many DEGs were observed associated with current smoking exposure, but few or no DEGs were observed in relation to former smoking exposures and/or smoking metrics. However, we observed that *LRRN3*-driven discrimination in all these comparisons; and thus, *LRRN3* expression in whole-blood could serve as a molecular signal of smoking exposure that could supplant self-reported smoking data in future studies focusing on blood-based markers related to the health effects of smoking. The biological functions of the identified DEGs explored in this study could indicate systemic impacts of obesity, as these functions coincide with its known physiological effects.

## 4.2 Paper II

*Associations of gene expression in blood with BMI and weight changes among women in the NOWAC postgenome cohort.*

This study investigated associations of gene expression in blood with BMI and weight changes among 1,694 women, who had never been diagnosed with cancer, from the NOWAC postgenome cohort. We used information gathered from three questionnaires (Q1: baseline, Q2: follow-up, and Q3: blood collection), and performed gene-wise linear regression models to identify DEGs, and functional enrichment analyses to identify their biological functions.

This study identified, 2,394, 769, and 768 DEGs according to the comparisons of obesity-vs-normal-weight, obesity-vs-overweight, and overweight-vs-normal-weight women, respectively. However, when investigating  $WC_{Q3-Q1}$  (mean interval=7 years, range=5.5-14 years), and  $WC_{Q3-Q2}$  (mean interval=1 year, range= <1 month-9 years) in interaction with BMI categories, only up to 169 DEGs were identified. Of these, between 1 and 169 genes were associated with the main effects of weight changes, and between 0 and 9 genes were associated with the interaction effects of BMI and weight changes. The biological functions of BMI-associated DEGs were linked to general metabolism, erythrocyte functions, oxidative stress, and immune processes, while weight change associated DEGs were linked to signal transduction.

In conclusion, many BMI-associated DEGs, but few weight change associated DEGs were identified in blood of women in Norway. The biological functions of the identified DEGs explored in this study could indicate systemic impacts of obesity, especially blood reticulocyte-erythrocyte ratio shifts; and these functions coincide with its known physiological effects.

### 4.3 Paper III

*Associations of gene expression in blood with menopausal status and hormone therapy use among women in the Norwegian Women and Cancer (NOWAC) postgenome cohort.*

This study investigated associations of gene expression in blood with menopausal status and HT use among 1,428 women, who never have been diagnosed with cancer, from the NOWAC postgenome cohort. We performed gene-wise linear regression models to identify DEGs, and functional enrichment analyses to identify their biological functions.

The study discovered 1,460 DEGs in blood samples when comparing postmenopausal with premenopausal women, showing a clear association of blood gene expression and menopausal status. Further, HT use among the postmenopausal women also revealed 348 DEGs. The biological functions of menopausal-associated DEGs were mostly linked to immune responses, cell-cell adhesion, cognition, muscle system process, and reproduction. HT-associated DEGs were linked to estrogen-dependent gene expression and RHO GTPases signalling. Nonetheless, when adjusted for the estimated WBC proportions, number of DEGs substantially reduced to 26 and 7, for the comparisons for menopause status and HT use, respectively. As a result, the enriched terms associated with immune processes in biological functions were no longer evident.

We concluded that many menopausal- and HT-associated DEGs were identified in blood of women, but when considering WBC proportions, most of these associations did not remain. Hence, the observed associations of menopausal status and HT use with blood gene expression seemed to be primarily influenced by the relative presence of blood cells. Further, we observed systemic impacts of menopause, as these functions correspond with their physiological effects.



## 5 Discussion

This thesis is comprised of three papers, which aimed to assess the association of cancer risk factors (in brief: smoking status and smoking metrics in Paper I, BMI and weight changes in Paper II, and menopausal status and HT use in Paper III) as exposure variables (or interchangeably referred to as ‘exposure(s)’ in the Discussion section) and gene expression profiles in blood as an outcome variable (or interchangeably referred to as ‘outcome’ in the Discussion section).

### 5.1 Main results

#### 5.1.1 Smoking status and smoking metrics

##### 5.1.1.1 Differentially expressed genes

We observed 911 and 1,082 DEGs when comparing current-vs-never smokers and current-vs-former smokers, respectively. Around 40% of these DEGs were over-expressed (genes with  $\log_{FC} > 0$ ) in both comparisons. *LRRN3* was the top-ranked gene (the gene with the lowest FDR-p-value) in both comparisons. These results were in line with similar previous studies [32-34,37,39,138]. Higher proportions of over-expressed genes have been observed most frequently [32-34,138], but some [37,39] reported a higher proportion of under-expressed genes. Further, when comparing the DEGs from the current-vs-never smoker comparisons, with genes (1,270 DEGs) identified in corresponding tests in a large meta-analysis containing 10,233 participants (51% women) [34], we observed 285 overlapping DEGs with 282 genes (98.94%) in same effect direction in both studies. This implies that the direction of the association to smoking was consistent for hundreds of genes between our study and the study by Huan et al. [34], demonstrating the comprehensive effects of current smoking exposures on gene expression in blood.

Among current smokers, only few significant genes were positively associated with increasing smoking exposure represented by CSI scores, smoking intensity, and pack-years, among which *LRRN3* was the top-ranked over-expressed gene. Among the former smokers, *LRRN3* was the only significant gene that was negatively associated with TSC. Nevertheless, we did not observe any DEG when comparing never smokers with all former smokers. This might probably be because TSC among the former smokers ranged from 1 year to over 40 years. However, when we limit the TSC for former smokers to  $\leq 10$  years ago, we observed one gene (*LRRN3*) was

differentially expressed between never smokers and former smokers. This highlights the need to consider TSC when analysing smoking effects in former smokers. Besides, among never smokers, no genes were observed to be associated with self-reported passive smoking when compared to those without passive smoking exposure. This might suggest that gene expression may be more influenced by active smoking in women themselves but could also indicate limitations in statistical power or imprecise exposure measurement.

These findings indicated the continued over-expression of *LRRN3* in relation to smoking exposures in our study. *LRRN3* has been consistently indicated to be over-expressed in the whole-blood of current smoker or former smokers in previous studies [32-34,36,37,39,139,140]. *LRRN3* was also the top-ranked gene in the overlap between our DEGs and the DEGs from Huan et al. [34], and we observed larger fold-change for *LRRN3* in our study (logFC=1.01) compared to them. Additionally, there are two new studies that are not covered in Paper I, one published during our submission process [141] and one published after Paper I was published [142]. Both these studies [141,142] show *LRRN3*-driven discrimination of smoking status. Further, our results showed that the *LRRN3* expression increases with ongoing smoking exposure but also in years after smoking cessation, it eventually (however takes around 20-30 years) reverts to levels like those of never smokers.

Further, the investigation of DNA methylation at specific CpG sites have shown promising abilities as markers of smoking status, that are capable of reflecting smoking exposure even long after smoking cessation [143,144]. Studies show that methylation at CpG sites in the *AHRR* gene is the top-ranked smoking marker [145,146]. Within a subset of our dataset, *LRRN3* exhibited a comparable ability to discriminate current smokers and former smokers (with  $\leq 10$  years TSC) from never smokers, as compared to methylation at CpG sites in the *AHRR* gene.

Nevertheless, several other genes like *PID1*, *RGL1*, *STAB1*, *NMRAL1* were also observed among the top-ranked genes associated with smoking status and smoking metrics in our study. The large overlap of DEGs of our study with study by Huan et al. [34], indicate that there are other genes consistently reported associated with smoking exposures but less clear than for *LRRN3*. It is crucial to explore the abilities of *LRRN3* expression as a quantitative marker for discrimination of smoking status in other population samples than those covered by research studies today and with the comparison to other markers, but investigating other genes in relation to smoking exposure should also be considered.

### 5.1.1.2 Functional enrichment analyses

Between current-vs-never and current-vs-former smokers comparison groups, we observed a large overlap on enriched categories of the over-expressed genes and under-expressed genes. However, there were more overrepresentation of ontologies and pathways when current smokers were compared to never smokers than when they were compared to former smokers. This could be because of the influence of smoking remained or was gradually decreasing in former smokers, which was also observed in our DEG findings of *LRRN3* expression. Additionally, the overall lack of overlap for enriched terms of the over-expressed and under-expressed genes likely demonstrated that these distinct gene groups are associated with separate pathways.

Over-expressed genes were mostly enriched for terms related to peptide metabolic and biosynthetic processes, protein formation and translation, humoral immune response, structural constituent of ribosome and molecule activity, ribosomal subunits, and adherence junction. On the other hand, under-expressed genes were enriched for terms related to response to wounding, regulation of blood vessels and tube size and diameter, neuron projection development, drug and hydrogen peroxide catabolic processes, heme binding, cell body, and hemoglobin complex. All these terms indicated that the smoking-associated genes were enriched for functions related to the well-known physiological effects of smoking on the human body; particularly linked to the circulatory and cardiovascular systems, as DEGs measured in blood could be directly influenced by such altered functions. Further, the top-ranked gene, *LRRN3*, is reported to be expressed in cardiomyocytes, which is the cell responsible for contraction of the heart muscle [147] and clearly indicates the link between smoking exposure and potential relevance of gene expression for heart disease as a health endpoint.

Carbon monoxide from smoking binds to haemoglobin in RBCs, thereby reducing the blood's oxygen-carrying capacity [148]. Accordingly, our results indicated that under-expressed genes related to smoking are involved in the haemoglobin complex, thereby potentially exacerbating smoking's negative effects on oxygen transport. Moreover, terms like heme binding, hemoglobin complex are RBCs-related processes/functions, reflecting the relative presence of RBCs; but we did not have the RBC estimates to adjust in the models. Further, smoking causes several negative vascular effects, including decreased coronary blood flow and myocardial oxygen delivery, as well as adverse effects on lipids, blood pressure, and insulin resistance [149]. In agreement with such processes, the under-expressed genes were enriched for blood

vessel size and diameter, and vascular processes in the circulatory system. We identified that oxidoreductase activity was under-expressed, which are in line with observations indicating that smokers experience observable and immediate oxidative damage, resulting in oxidative stress [22]. Additionally, we observed under-expressed genes overrepresented in wound healing and haemostasis processes, which is consistent with observations indicating a diminished ability to heal wounds among individuals who smoke [22,24]. Lastly, we could observe that terms related to immune responses were overrepresented among over-expressed genes. Several studies analysing gene expression associated with smoking also have revealed impacts on the regulation of processes within the immune system [32-34,36-39]. All these point towards a relation of smoking with the circulating immune system, which is also a well-known effect of smoking [22].

The enriched terms overrepresented by the DEGs in our study were largely consistent with those reported in the study by Huan et al., which primarily revealed enrichment in the activation of platelets and lymphocytes, immune response, and apoptosis [34]. Additionally, the expression of our top-ranked gene, *LRRN3*, has been associated with methylation of a CpG site on the *AHRR* gene [132], within a subset of our dataset (depicted by the ROC curve); and *AHRR* is connected to *AHR* and *CYP* proteins involved in detoxification mechanisms in the liver, as these genes might thus reflect features of a plausible physiological influence of smoking exposures.

Overall, smoking might influence a large spectrum of blood gene expressions and the enriched terms for the smoking-associated DEGs in our study indicate broad physiological effect of smoking exposures, mainly current smoking exposures. Still, considering the diverse molecules present in tobacco smoking, it can potentially influence multiple pathways, which was indeed observed in the GO categories indicated by our study.

### **5.1.1.3 Blood cell compositions**

Among the estimated proportions of 22 types of WBCs, we observed that CD8 T cells, naive CD4 T cells, resting NK cells, M0 macrophages, resting mast cells, and neutrophils were significantly associated with both smoking status and overall gene expression. We observed modest difference in number of DEGs in the model with WBC adjustments (fully-adjusted) and without WBCs adjustment (minimally-adjusted) and the logFC of the top-ranked genes did not change much. Further, our top-ranked gene, *LRRN3*, is reported to be expressed in T cells (naive



CD4 and CD8 T cells) [147]. We observed that WBC proportions and smoking metrics, especially resting NK cells but also CD8 T cells, resting mast cells, and neutrophils, were negatively associated with increasing smoking exposure. Naive CD4 T cells were positively associated with several smoking metrics. While resting NK cells were positively associated with TSC, indicating that if one quits smoking, the NK cells tend to increase, which increases the immune responses/regulations in the body [150].

However, we could observe that terms related to immune responses were overrepresented among current smokers. These results are in line with observations [32-34,36-39] that smoking may have adverse effects on the immune capacity of the body. In fact, smoking has been identified as a significant and reversible cause of elevated WBC counts in healthy adults [151]. Smoking can compromise the immune system and immune homeostasis as a whole [22]. It is a limitation that we could not adjust for RBCs or reticulocytes estimates/counts as we observed ontology categories related to RBC processes such as heme binding, hemoglobin complex was overrepresented among current smokers. Moreover, these results of blood cells compositions and smoking exposures are in line with research [152] revealing that continuous cigarette smoking has detrimental impacts on hematological parameters, such as hemoglobin, WBC counts, RBC counts, etc. These changes in hematological parameters may be linked to an increased susceptibility to conditions such as atherosclerosis, chronic obstructive pulmonary disease and/or cardiovascular diseases [152].

## **5.1.2 BMI and weight change**

### **5.1.2.1 Differentially expressed genes**

We observed 2,394, 769, and 768 DEGs for the obesity-vs-normal-weight, obesity-vs-overweight, and overweight-vs-normal-weight comparisons, respectively, and 3,106 DEGs in continuous BMI analyses. Thus, many associations of blood gene expression with BMI were observed. Across the models testing associations with BMI, 525 DEGs overlapped, and these can be considered as consistent genes associated with BMI. *FAM46C* was the top-ranked gene in all BMI analyses, except in the overweight-vs-normal-weight comparison, where the top-ranked gene was *SLC45A3*. *FAM46C* was positively associated with increasing BMI and *SLC45A3* was negatively associated. The BMI-associated DEGs (both 525 DEGs across all BMI-models with 3,106 DEGs from continuous BMI analyses) were largely consistent in terms of direction of effect with findings of previous studies in whole-blood [61,62], PBMCs [63],

and adipose tissue [64], although our top-ranked genes *FAM46C* and *SLC45A3* were only identified a study by Homuth et al. [62]. The overlap with the study in adipose tissue [64] was lower, and this was the only tissue that was negatively correlated with our estimation, as we observed positive correlation in whole-blood [61,62] and PBMCs [63]. This indicated that systemic patterns in expression in blood related to BMI differ from those in adipose tissues. Our results suggest that DEGs in whole-blood related to BMI in women could be applicable to both sexes and other blood samples, but gene expression profiles may be differently regulated in adipose tissue.

The categorical weight change analyses did not reveal any DEGs. In the weight change analyses as a continuous metric, we regarded two measurements in terms of intervals – a farther one or long-term ( $WC_{Q3-Q1}$ ), with an average time interval of 7 years (range=5.5-14 years), and a closer one or short-term ( $WC_{Q3-Q2}$ ), with an average time interval of 1 year (range=<1 month-9 years). When focusing on these weight changes after accounting for interactions with BMI categories (WC-BMI interaction analyses), a few associations with current gene expression were revealed. Between 1 and 169 genes were associated to the main effect of past weight changes and between 0 and 9 genes were associated with the interaction effect of past weight changes and current BMI (represented as  $BMI_{Q3}$ , i.e., at blood collection in Paper II). Here as well, the effect of obesity was prominent, as the few DEGs (9 and 1) in the weight change interaction effects were significant only among women with obesity. Further, we introduced two interaction models assuming that current BMI was a stronger predictor of current gene expression than past weight changes or past BMI, thus the second interaction model ( $BMI_{Q3} * WC$ ) could be expected to reveal more DEGs than the first ( $BMI_{Q1 \text{ or } Q2} * WC$ ). In accordance with these predictions, our findings demonstrated that the interaction effect of weight change and BMI was not significant in the first interaction model (where we had regarded past BMIs), whereas up to 9 genes were significant in the second interaction model (where we had regarded current BMIs). The expression of the top-ranked genes from the WC-BMI interaction analyses may suggest a positive correlation between  $WC_{Q3-Q1}$  and *CECR6*, and a negative correlation between  $WC_{Q3-Q2}$  and *STT3A* among women with obesity, though the observed trend was not very robust. Also, we observed that 21 DEGs among the weight change models (N=169) were overlapping with 525 DEGs across all BMI-models indicating that past weight changes to some extents were represented in current BMI models.

We could not attempt to replicate the results for past weight changes, as until now, no study had investigated the association between gene expression in blood and past weight changes. But there were some earlier studies in adipose tissue, which revealed that alterations in gene expression patterns within adipose tissue were linked to weight reduction in individuals with obesity following dietary interventions [65-71]. However, the follow-up time in these studies (4 weeks to 9 months) was shorter than the time intervals in our study (range=<1 month-14 years). The weight changes we observed could be too far in the past to have a major influence on blood gene expression. Still, sensitivity analyses restricted to women with <1 year between Q2 and Q3 did not show any significant DEGs. Future studies focusing on systemic signatures related to weight changes should likely include blood samples taken within months of the weight change occurring for transcriptomic signals to be detectable.

Overall, our results indicated that current BMI and past weight changes have, respectively, vast (>3,000 DEGs) and restricted overall influence (up to 169 DEGs) on blood gene expression in terms of number of DEGs.

#### **5.1.2.2 Functional enrichment analyses**

Functional enrichment analyses of BMI-associated DEGs revealed a wide range of functions in enriched ontologies and pathways and included general processes related to metabolic and blood homeostasis. In women with obesity, over-expressed genes were enriched for various catabolic (e.g., cofactor catabolic processes) and metabolic processes (e.g., hydrogen peroxide, tetrapyrrole metabolic processes), as well as erythrocyte homeostasis, haemoglobin binding, and ribosome structures. These align with previous studies that have analysed obesity and gene expression in whole-blood [61,62] or PBMCs [63].

The enriched terms erythrocyte differentiation, myeloid cell homeostasis, erythrocyte homeostasis, heme biosynthetic/metabolic processes suggest the overexpression of genes in erythrocytes or their precursors (reticulocytes) [153]. Of note, *FAM46C*, the top-ranked gene associated to BMI, and several other top-ranked genes (*HBD*, *GYPB*, and *ALAS2*) are primarily expressed in bone marrow, blood, and early erythroid cells [153,154]. Erythrocyte indices have been observed to be positively linked with obesity [155,156], possibly due to increased reticulocytes proliferation in the bone marrow [157,158] induced by the hormone leptin. However, people with obesity may experience a shorter half-life of erythrocytes in circulation due to impaired insulin resistance and oxidative stress caused by hyperglycemia [49]. This shift

in the reticulocyte-erythrocyte ratio is expected to be reflected in the whole-blood transcriptome because reticulocytes are also transcriptionally active [62]. Also, the top-ranked genes among 525 genes across BMI-models, were enriched for erythrocytes functions. Thus, these results indicate that BMI-associated DEGs identified in this study likely reflect a shift in the reticulocyte-erythrocyte ratio. Validation of changes in the reticulocyte/erythrocyte proportions in relation to BMI appeared outside the scope of our work based on transcriptomic profiles. However, the influence of this shift on blood gene expression by RBCs was indicated in a published study [62], which we have compared with as quantitative replication, and observed large consistency in terms of effect direction.

Other enriched terms among the over-expressed genes were peptide chain elongation and eukaryotic translation termination/elongation which appeared related to protein synthesis [159]. Another study conducted in whole-blood [61] observed ribosome and protein synthesis pathways as top-ranked among women with obesity. These terms are in line with the physiological changes previously observed in people with obesity such as higher levels of oxidative stress [49,160], haemoglobin [161,162], and disrupted protein synthesis [163,164].

The under-expressed genes were enriched for terms like antigen binding, processing and presentation, peptide binding, and TNF signalling pathways, which suggest there could be altered blood immune processes in women with obesity. Similar processes have been observed among people with obesity in other transcriptomic studies [61,64,165]. Moreover, we observed overrepresentation of terms like influenza, asthma, antigen binding. Such altered immune-related terms in women with obesity could explain previously observed link between obesity and increased risk of co-morbidities and infectious diseases, like influenza and COVID-19, and increased viral shedding and transmission [166-168]. Overall, the enriched terms for the BMI-associated DEGs in our study indicate broad physiological effect of obesity/overweight. The assessment of past weight changes and blood gene expression was novel, yet revealed only a small number of DEGs, which could not strongly indicate specific biological functions. Nonetheless, some over-expressed genes (*RBP1/FZD2/OPRL1/CD14*) indicated relations between past weight changes and genes involved in signal transduction.

### **5.1.2.3 Blood cell compositions**

Literature shows that BMI and body weight have been positively correlated with WBC counts in apparently healthy young adults, and especially higher in women [169]. Among the estimated

WBC proportions, we observed naive B cells, memory B cells, naive CD4 T cells, and memory-activated CD4 T cells were significantly associated with both BMI and overall gene expression. After adjustments of WBCs in the model, we observed slight increase in the number of BMI-associated DEGs and the logFC of the top-ranked genes did not change much. The enrichment analyses showed some immune related terms, but we observed the erythrocyte terms were mostly overrepresented. Even the top overlapped DEGs across all BMI models were related to erythrocyte functions. This clearly indicated the limitation of not having counts or estimates of their presence to investigate in this study.

### **5.1.3 Menopausal status and HT use**

#### **5.1.3.1 Differentially expressed genes**

We observed 1,460 DEGs for the post-vs-premenopausal status comparison showing clear associations of menopausal status with gene expression in blood. However, after adjusting for the estimated proportions of WBCs in the blood samples (Model-3), only 26 DEGs remained, indicating the relevance of including WBCs as covariates when investigating these associations.

Further, the comparisons of gene expression profiles for post-vs-perimenopausal women showed no DEGs, while one DEG was identified in the pre-vs-perimenopausal women comparison, in models adjusted (Model-3) and not adjusted (Model-2) for WBC proportions. The limited number of DEGs in the perimenopausal group could be due to the wide range of physiological changes occurring in this group, which may have resulted large inter-individual variation in gene expression in this group. Another possible explanation could be the self-reported menopausal status, which may have resulted in misclassification.

As this study is one of the first studies to investigate gene expression profiles in blood in relation to menopause, it was not possible to compare the DEGs in our study with those of similar studies to ensure consistency of results. Additionally, it was not feasible to investigate the study findings by comparing them with the only study investigating menopausal association with gene expression in circulating monocytes [84] due to unavailability of that study data.

In the comparison of HT-users-vs-non-users, we observed 348 DEGs, indicating that HT use can influence gene expression in years after menopause, at least for users of systemic HT. Still, these DEGs were reduced to 7 DEGs after adjusted for the estimated WBCs proportions (Model-3). As HT-users could be assumed to resemble premenopausal women in terms of sex

hormone status [121], confirming this assumption with sex hormone levels in these women would assure this, but this information was not available in our dataset. Still, we compared HT-associated DEGs with post-vs-premenopausal status associated DEGs. Among the 348 HT-associated DEGs identified in Model-2, 231 (66.38%) overlapped with post-vs-premenopausal associated DEGs (1,460 DEGs; Model-2). But none of the seven HT-associated DEGs identified in Model-3 were among the 26 DEGs associated with post-vs-premenopausal status in Model-3, although five of them were observed among DEGs associated with post-vs-premenopausal status in Model-2. Our hypothesis about the sex hormone status of HT-users and the resemblance to premenopausal women was confirmed by observing that all overlapping DEGs between HT-users and post-vs-premenopausal status exhibited opposite effect directions (i.e., higher expression in HT users and premenopausal women).

The study conducted by Waaseth et al. [91] investigating the relationship between sex hormone status and gene expression in blood, observed a total of nine DEGs, with sex hormone concentrations. Four genes had an FDR below 0.28 and five had an FDR below 0.25. Similarly, Dumeaux et al. [170], investigating the use of HT and its impact on gene expression in blood, identified 12 DEGs among HT users compared to non-users, with a global FDR of 26.6%. When we compared their [91,170] DEGs with the HT-associated DEGs in our study, no overlapping genes were observed. This lack of overlap could be attributed to the relatively small sample sizes in these studies and the absence of adjustments for WBC proportions, which might have influenced the model estimates.

Overall, these findings revealed association of blood gene expression with menopause and HT use, but associations were likely driven by the relative proportion of WBCs in the samples.

### **5.1.3.2 Functional enrichment analyses**

Functional enrichment analyses of menopausal-associated DEGs revealed a diverse array of functions in enriched ontologies and pathways. Under-expressed genes in post- compared to premenopausal comparisons in models, where adjustment for WBCs was not included (Model-2), were enriched for terms such as neutrophil activation and degranulation, myeloid leukocyte activation, leukocyte activation. This appeared to be primarily linked to leukocytes and immune responses; and in line with WBCs associated with menopausal status in our study. This indicated a general downregulation of immune system processes in postmenopausal compared to premenopausal women. These terms could be in agreement with the influence of WBCs

associated with menopause. Indeed, these enriched terms were no longer significant after adjustment for WBCs (Model-3), and the remaining enriched terms were cell-cell adhesion and regulation of body fluid levels. These terms could reflect general physiological changes during menopause independent of immune cell changes.

Among the over-expressed genes in post-vs-premenopausal comparisons in Model-2, the few enriched terms were related to function of catalytic activity and protein binding. Further, enriched terms for the over-expressed genes in the same comparisons in Model-3 were linked to cell-cell adhesion, cognition or nervous system, response to stress, muscle system process, cell development and reproduction, and regulation of body fluid levels. These terms agree with terms indicated as influenced by menopause status in circulating monocytes previously [84]. Additionally, the top-ranked gene in the post-vs-premenopausal comparison (Model-2), *ADCY3*, were mainly associated with e.g., biosynthetic and metabolic process, cognition, renal system process, regulation of catalytic activity [153]. The top-ranked gene in post-vs-premenopausal comparison (Model-3), *CTTN*, were related to e.g., cell adhesion and cell junction organization, muscle system process [153]. These ontologies and pathways are general in nature but could also point towards chronic systemic inflammation. Studies have reported that increased visceral fat mass and decline in estrogen levels during menopause could lead to chronic systemic inflammation [86]. Such terms are contributing factors to metabolic diseases like insulin resistance, type 2 diabetes, and cardiovascular disease – the diseases that tend to become more prevalent after menopause [86]. Thus, overall, we observed changes in the systemic immune and inflammatory status.

The DEGs related to HT use in the model with WBC proportions adjustment (Model-3) did not indicate any enriched ontologies/pathways. When WBCs were not adjusted for (Model-2), the DEGs related to HT-use were enriched only for a few terms in REACTOME pathways, mainly related to estrogen-dependent gene expression and RHO GTPases Effectors. Further, the gene *HIST1H3D* (also named as *H3C4*) was over-expressed among HT-users and has known functions in regulation of metabolic processes, immune system, cellular developmental process, and cellular response to stimulus [153]. The top-ranked gene in the HT-users-vs-non-users comparison (Model-2), *ARHGEF7*, was under-expressed among HT-users and has known functions in signalling by epidermal growth factor receptor -EGFR, ephrin, and cell death [153].

The enriched terms and top-ranked genes contribute to the overall understanding of how menopause and postmenopausal HT-use influence systemic gene expression, with immune cell changes likely playing a significant role in mediating these associations.

### **5.1.3.3 Blood cell compositions**

Among the estimated proportions of 22 types of WBCs, we observed CD8 T cells, naive CD4 T cells, regulatory T cells, activated NK cells, monocytes, M1 macrophages, activated mast cells, eosinophils, and neutrophils were significantly associated with both menopausal status and overall gene expression. Accordingly, the functional enrichment analyses of the DEGs observed in models without adjustment of WBCs revealed ontology terms associated with immune-related pathways. This suggests that the variances in the relative proportions of these WBCs according to menopausal status might be reflected in the bulk gene expression profiles.

Among the different groups, postmenopausal women exhibited slightly higher mean proportions of CD8 T cells and activated NK cells. Perimenopausal women, on the other hand, had slightly higher mean proportions of naive CD4 T cells, regulatory T cells, monocytes, and M1 macrophages. Lastly, premenopausal women showed slightly higher mean proportions of eosinophils and neutrophils. In our observations, we noted that while there were minimal differences in the proportions of other significant WBC populations, the median levels of activated mast cells were twice as high in premenopausal women compared to postmenopausal women. These findings align with previous research [171] indicating that mast cells can be activated by female sex hormones. We observed lower proportions of the estimated neutrophils and monocytes in postmenopausal compared to premenopausal women whereas higher proportions of the estimated CD8 T cells and CD4 T cells in postmenopausal women. Studies [86,172] analysing blood samples have reported a correlation between menopause and changes in the proportions of WBC. Chen et al. [172] reported similar results to our observations, noting a decrease in neutrophil percentages and an increase in lymphocyte percentages in women around the age of 50, which is typically the age of menopause. Additionally, a study [86] investigating the altered distribution of T-cell subsets in postmenopausal women and observed higher counts of T-cell subtypes like in our study and elevated levels of circulating inflammatory markers such as TNF- $\alpha$ , IL-1 $\beta$ , and IL-6 in postmenopausal women compared to premenopausal women. Still, they observed higher monocytes counts among postmenopausal women [86], whereas we observed the highest mean monocytes count among the



perimenopausal women. This likely agrees with that the systemic immune status during the years of menopausal transition is subject to physiological changes. Consequently, when analysing blood molecular markers that are influenced by the underlying compositions of immune cells in samples, it is crucial to incorporate adjustments and considerations for these aspects in the study design and analyses.

#### **5.1.4 Across the risk factors investigated –the broader picture**

##### **5.1.4.1 The associations of risk factors with blood gene expression**

This thesis investigated associations of blood gene expression with different cancer risk factors with different importance and different physiological effects. Two of them –smoking and obesity –were the two most important modifiable cancer risk factors globally, while the third one –menopause –is another important risk factor. The impact of menopause and hormonal factors on cancer risk may seem minor for individual cases, but when considered collectively within a population, they can have a significant effect; particularly relevant due to the universal experience of these factors among all women [8]. Additionally, there is a global trend of an increasing postmenopausal female population due to the aging of many populations [9]. Moreover, HT use, the medication used to control different menopausal symptoms, might be an independent risk factor to cancer. Furthermore, the risk factors investigated (mainly: smoking, obesity, and menopause/HT use) range from exogenous to endogenous factors. Active smoking is an external factor and it a choice people make whether to consume or not, which makes it an exogenous risk factor of cancer. HT used to increase estrogen levels in women [121] is also an exogenous risk factor and its use could to some extent be a choice. On contrary, the menopausal transition arises naturally in a women’s life due to the reduction in endogenous estrogen levels. A menopause is not a choice women make; it is a natural process of physiological changes, which makes it an endogenous risk factor of cancer. Lastly, obesity is multifactorial condition, and it can be the result of life choices one makes such as excess dietary intake and inadequate physical activity but also with contribution from other factors like endocrine disruptions [173,174]. Obesity/overweight could thus, be both an exogenous and/or endogenous risk factor [175]. Obesity could be the result of sustained energy imbalance with a combination of other factors involved in its development such as genetic, behavioral, cultural, environmental, and economic factors, while obesity could also, on other hand, in some cases be regarded as a risk factor and a disease in itself [175]. All these risk factors affect the major physiological and biological processes in one’s body and despite the differences in importance

and physiological impact of these risk factors, this thesis observed overall association with blood gene expression for all, yet the number and strength of associations were different for the different risk factors investigated.

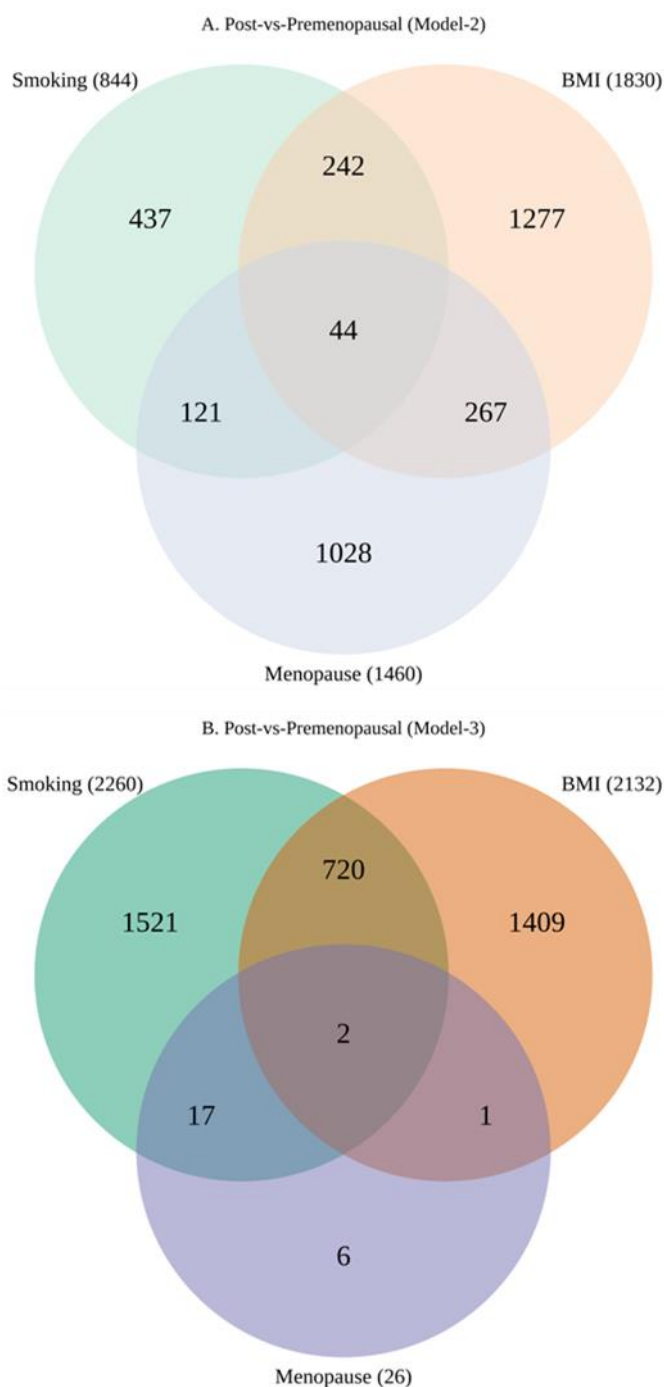
This thesis revealed large associations of blood gene expression and current exposure variables: current smoking status (Paper I) and current BMI (Paper II). Indeed, among the risk factors investigated in Paper I-III, the strongest association in terms effect sizes was observed for smoking, with the highest logFC (1.01) and lowest FDR adjusted p-value ( $1.52E-80$ ) of the top-ranked gene (*LRRN3*). These estimates indicated that the mean expression of *LRRN3* in current smokers was double compared to that in never smokers. The absolute logFC of all the other top-ranked genes in different comparisons in Paper II and III were modest ranging from 0.14 (weight changes) to 0.86 (BMI categories) in Paper II, and 0.07 (HT use) to 0.18 (menopause) in Paper III. Further, among the risk factors investigated, the most associations in terms of largest number of DEGs were observed for obesity with 2,394 DEGs when women with obesity were compared to women with normal-weight. The maximum number of DEGs in Paper I was 1,082 for current-vs-former smokers and was 1,460 for post-vs-premenopausal women in Paper III (when not adjusted for WBC proportions); demonstrating that the number of DEGs was still large in all papers (Paper I-III). We observed few or no DEGs in relation to past exposures: former smoking status (Paper I) and past weight changes (Paper II). This could be expected, as blood gene expression represents a snapshot, and past exposures such as weight changes are generally not strongly reflected [119,120]. Still, to our knowledge, this thesis concludes the first time that current BMI is reflected in blood gene expression more than past weight changes in terms of number of DEGs.

In Paper I, we observed *LRRN3*-driven discrimination in all analyses, be it with current smoking exposures or former smoking exposures or smoking metrics. Additionally, we demonstrated that *LRRN3* has similar capability as methylation status of CpG sites in the *AHRR* gene, the gene that is known for reflecting smoking exposures [143,144]. Thus, *LRRN3* expression in blood is a molecular signal of smoking exposure that could supplant self-reported smoking data in further research targeting blood-based markers related to the health effects of smoking. We did not observe signals of similar strength for the BMI (Paper II) and menopausal status (Paper III).

Paper III showed clear associations of blood gene expression with menopausal status and HT use in models without WBC adjustments (Model-2). The associations were more prominent with menopausal status with a greater number of DEGs and higher fold-changes than HT use. But when considering the estimated WBC proportions in the adjusted models, the number was notably reduced. Thus, the relative influence of WBCs appeared the strongest in menopausal status in Paper III, but we also acknowledge that RBCs could be important for the results observed in both Paper I and II although we could not adjust for them.

The associations observed for menopausal status could be regarded robust as the model included adjustments for smoking and BMI, both of which, with our knowledge from Paper I and II, had strong associations with blood gene expression. Figure 6 below shows intersects of DEGs among all three main risk factors in this thesis (smoking, BMI, menopause) from Model-2 (A) and Model-3 (B) of Paper III. This figure depicts the associations of blood gene expression with all three risk factors investigated in this thesis in a single model. It appears that there are numerous genes for which the modulated expression was specific to each factor, with only a small overlap in the middle. This suggests that there were only a few genes that are associated with both two risk factors e.g., smoking and BMI, simultaneously.

Overall, this thesis reflects the broad influence of smoking, BMI, and menopausal status on gene expression in the blood of women. Our studies were conducted with a large sample size which was further strengthened with the reliable and repeated measurements of questionnaire information. Therefore, the results of this thesis could represent robust knowledge about the molecular signals of the risk factors investigated.



**Figure 6:** Venn diagrams showing intersects for DEGs among smoking, BMI, menopause in Model-2 and Model-3 (Paper III).

#### 5.1.4.2 Biological insights

In this thesis, the DEGs revealed various ontologies and pathways. Paper I demonstrated that terms like metabolic process, immune responses, wound healing, oxidative stress, heme binding, haemoglobin complex enriched among smoking-associated DEGs. The presence of terms such as heme binding and hemoglobin complex in these findings may indicate processes

related with RBCs. Similarly, Paper II showed that terms like general metabolism, erythrocyte functions, oxidative stress, and immune processes were enriched among BMI-associated DEGs. But we could observe the dominance of processes related to RBCs, specifically the reticulocyte-erythrocyte ratio shift in blood. Even the top-ranked genes among the overlapping 525 genes across different BMI models in Paper II were enriched for terms describing RBC processes. The few weight change associated DEGs showed were linked to signal transduction. Likewise, Paper III showed that menopausal-associated DEGs were mostly linked to immune responses, cell-cell adhesion, cognition, muscle system process, and reproduction. HT-associated DEGs were related to estrogen-dependent gene expression and RHO GTPases signalling. But when adjusted for WBCs compositions, the biological functions related to immune responses disappears in menopausal model; presumably demonstrating the influence of WBCs on the associations of blood gene expression and menopause.

The smoking-, BMI-, and menopausal-associated DEGs all reflect the overrepresentation of immune responses/process and processes related to RBCs, in one or another form. Moreover, studies show increased oxidative stress caused by smoking [22] and obesity [49], which are related to RBC processes/functions. Furthermore, distortion of the immune system and change of many immunological functions and WBC counts, are one of the main physiological effects of smoking [25], obesity [52], and menopause [85]. These observations point towards the common effect of these risk factors to increase chronic systemic inflammation in body. Overall, the enriched terms coincide with the known physiological effects of these risk factors, and sheds light on the systemic impact of smoking, obesity, and menopausal status.

#### **5.1.4.3 Influence of blood cell compositions**

One of the main findings of this thesis is the influence of blood cell compositions on the associations of blood gene expression and measured risk factors.

We were aware that literature has shown that gene expression profiles in whole-blood can be influenced by the underlying composition of WBCs in the respective samples [39]. Thus, keeping in mind that skewed proportions of WBC have the potential to act as confounding factors when identifying gene expression differences associated with exposures such as smoking, obesity, or menopausal status, we regarded these as possible confounders and adjusted these selected WBCs in one of the adjustment models in all three papers. We observed rather moderate differences, in terms of number of DEGs and fold-changes, in models adjusted

(fully-adjusted) and not adjusted (minimally-adjusted) for WBCs in Paper I and II. Still, the adjustment for WBC proportions in assessment of menopause status groups was clearly important. In Paper III, the number of WBC proportions associated in the initial test was higher (nine) related to menopausal status than smoking status (six) in Paper I and BMI categories (four) in Paper II. Further, unlike Paper I and Paper II, we noticed a drastic difference in the number of DEGs from models adjusted (Model-3) and not adjusted (Model-2) for WBCs in relation to menopause in Paper III. Still, the top-ranked DEGs from Model-2 were observed also in Model-3, showing that the most DEGs still remained, and the fold-change did not change much. Even the menopausal-associated DEGs, which were enriched for terms related to immune processes, remained no longer significant after WBC adjustments. Thus, associations of blood gene expression and menopause seem to be largely driven by cell type composition rather than menopausal status itself.

WBCs play key role in inflammation and immunity. Neutrophils and monocytes are essential components of the innate immune system, while lymphocytes (B cells and T cells) play key roles in the adaptive immune system [176]. Furthermore, research indicate that smoking [25], obesity [52], and menopause [85] have been associated with immune system disruption and alteration of several immunological functions and WBC counts. Additionally, WBCs are the most transcriptionally active blood cell types [115]. Thus, our findings clearly show that these immune responses/processes are affected by smoking, obesity and menopause and are reflected in blood gene expression.

Of note, we only had access to estimated WBC proportions in our dataset and did not have actual WBC counts. The estimated WBC proportions in this thesis deviated from the expected range [177,178], also observed in other studies [179,180] giving rise to bias [179,181] (see Chapter 5.2.6.2). However, we do not think that the bias is systematic for any of the risk factor groups in this thesis and thus would represent a problem only if the deviations are related to the risk factor groups. The absolute differences when adjusting for WBCs in the number of DEGs related to smoking status (Paper I) and BMI (Paper II) were modest, and the top-ranked genes and their fold-changes identified in the models with and without the WBCs were similar in all papers (Paper I-III). This possibly indicates that these genes were not substantially confounded by distributions of WBCs in Paper I and II. However, it would be interesting to know how the results would have been had we had actual WBC counts instead of estimated WBCs proportions. Further, RBCs/reticulocytes counts were not possible to estimate and therefore not

available in our dataset, and their adjustment was thus not possible. We anticipate that if we had the opportunity to include models with RBCs/reticulocytes adjustments, it could reveal their possible influence both in number of DEGs and GO and pathways, especially in Paper II (related to obesity) but might also in Paper I (related to smoking) and Paper III (related to menopausal status).

Our holistic understanding considering the biological impacts of the risk factors investigated is that they are related to blood cell compositions, be in RBCs or WBCs. We observed strong influence of blood cell compositions (both WBC and RBCs/reticulocytes) on blood gene expression profiles. Future studies should consider adjustment of blood cell proportions in studies related to blood-based markers. Further, they should focus on cell-specific gene expression analyses, including RBCs/reticulocytes counts, so that they could assess whether it is the functional state, or it is the distribution of different cell types that would influence the association of exposure variables and gene expression profiles in blood.

#### **5.1.4.4 Dynamics of gene expression profiles**

In this thesis, we observed a large number of associations of blood gene expression with current exposure variables (i.e., current smoking status in Paper I, BMI in Paper II, and menopausal status and HT use in Paper III). Unlike exposure variables measured at the time of blood collection, we did not observe substantial associations of blood gene expression with past exposure variables (i.e., former smoking status in Paper I and past weight changes in Paper II). This could be expected because of the dynamics of gene expression profiles.

In this thesis, blood collection for gene expression analyses was only performed at one time point, representing a snapshot of the expression. It was not surprising that past exposures such as in this thesis (former smoking status and past weight changes) were generally not strongly reflected, also shown in other studies [120]. Still, if the interval between the reporting of exposure variables and gene expression profiles could be as short as possible, then we could probably observe more associations. Nevertheless, we had an advantage of having quantitative, reliable, and repeated measurements of questionnaire information. We utilized them to generate past and recent smoking exposures to investigate trends even in former smokers who quit smoking recently in Paper I and we could investigate the associations of short-term weight changes (mean interval=1 year) and long-term weight change (mean interval= 7 years) in Paper II.

Studies have shown that intra-individual gene expression profiles measured by microarray remained relatively stable over time, as less than 2% of genes analysed exhibited intra-subject differences over periods exceeding one month [182]. Additionally, if we had the repeated measurements of gene expression profiles (i.e., for more than one time point), it would enable us to measure the change in gene expression over time within the same individual, minimizing the intra-individual variability due to other factors than the risk factors. A study [116] has reported that by analysing multiple sequential samples obtained from the same individuals, they were able to identify unique and individual-specific patterns of gene expression. These findings contribute to the understanding of human individuality and establish a valuable database for comparing gene expression patterns associated with diseases [116].

Lastly, literature overall show gene expression changes over different periods (within hours or days) and different seasons [100,118]. Thus, information about date and time of sampling were considered during the pre-processing (Chapter 5.2.6.4). It was examined in the pre-processing of the data for these studies and that it was not a major influence but also that the variables were likely not very complete or reliable.

## **5.1.5 The novelties**

### **5.1.5.1 The novelty of the aims**

The novelty of this thesis lies in investigating associations of selected risk factors (or the exposure variables), which had never been assessed before, with gene expression in blood of women who had not had cancer. Overall, we had the opportunity to utilize large population-based samples from the NOWAC postgenome cohort, along with repeated measurements of exposure variables, which was unique in this research area.

Paper I assessed the association of gene expression in blood according to not only smoking status as had been done before but also different smoking metrics (duration, intensity, TSC, CSI scores) within ever smokers and former smokers. The novelty of Paper I lied in its utilization of quantitative, reliable, and repeated measurements of past and recent smoking exposures. Further, Paper II extensively investigated the association of blood gene expression with current BMI and past weight changes in a large sample of women. Exploring associations of past weight changes with gene expression in blood was novel. Lastly, Paper III broadly examined the association of blood gene expression with menopausal status and HT use among



postmenopausal women a large population-based sample of women. The investigation of association of menopausal status and gene expression in blood was also novel.

#### **5.1.5.2 The novelty of the results**

We discovered genes differentially expressed when assessing smoking metrics (Paper I), past weight changes (Paper II), and menopausal status (Paper III) that were novel. Further, to the best of our knowledge, Paper II is the first study to conclude that blood gene expression reflects current BMI more than past weight changes. Lastly, the importance of blood cell compositions, especially the blood erythrocytes-reticulocytes ratio shift associated with BMI in Paper II, and WBC proportions associated with menopausal status and HT use in Paper III were novel. These findings contributed to new knowledge on systemic responses of these risk factors.

#### **5.1.6 Knowledge contributions to future cancer studies**

Cancer is ever changing as it progresses through a series of histopathological stages, resulting in alterations in gene expression patterns [183]. Gene expression changes that take place from the early stages to advanced stages of cancer development can serve as indicators to monitor the progression of the disease [184]. Uncovering the specific genes and pathways involved in this process is crucial not only for advancing our understanding of the biology underlying cancer progression but also for identifying potential targets for early diagnosis and facilitating the development of effective treatments [183]. This could presumably aid to reduce the overall burden of cancer by prevention of occurring of new cases and by early detection [185].

Genomic techniques have proven successful in detecting chromosomal alterations and identifying disrupted genes in cancer. Additionally, gene expression profiling has enabled the categorization of tumors into distinct subtypes. Eventually, a combination of genomic and expression analysis approaches would be vital to validate genes within regions of DNA alteration and shed light on the downstream effects of these alterations [183]. By integrating the revolutionary new tools of genomics – at all levels of genome, transcriptome, and proteome – key pathways and functions can be defined, which ultimately can lead to breakthroughs in identification of new causes of cancer as well as early detection, and then may result in implementing strategies that limit exposure [185]. The potential impact of these techniques is large; however, their success will rely on international collaboration and strategic planning [185].

In this thesis, blood gene expression analyses were performed on samples of controls, i.e., we only included those women, who had never been diagnosed with cancer. Thus, we can presume that the molecular signals observed in the papers (Paper I-III) included in this thesis were solely due to the differences in levels of the exposure variables (smoking, obesity, or menopause) and not due cancer disease. In general, we assumed that women in these studies were healthy, because we only included cancer-free women. However, we cannot disregard the influence of other common chronic diseases the women might had have, which we were not aware of while collecting our data samples. Further, we investigated into GO and pathways during our analyses and did not focus on disease ontologies, especially because the study population were healthy. Still, knowing the genes that are differentially expressed related to the investigated cancer risk factors and pathways involved related to those genes could be useful in studies investigating DEGs in relation to cancers for which these exposures are risk factors. In case-control studies these factors are adjusted for and as little is known about the DEGs according to the risk factors, this study can add knowledge to those analyses. Especially when evaluating whether observed case-control differences in expression could be due to residual confounding by these risk factors.

Any new knowledge relating to cancer and exposure risk factors aid in the etiological knowledge of cancer disease and development which again can ultimately aid in future disease prevention and/or early indication of different diseases. The top-ranked genes from all the papers (Paper I-III) like *LRRN3*, *FAM46C*, *SLC45A3*, *ADCY3*, *CTTN*, *ARHGEF7*, *NCOA5* have been observed as related to different disorders and diseases including various cancer types in various publications [154]; for instance vascular skin diseases (*LRRN3*), blood protein disorders, plasma cell leukemia (*FAM46C*), cloacogenic carcinoma (*SLC45A3*), central corneal ulcer (*ADCY3*), larynx and breast cancer (*CTTN*), night blindness (*NCOA5*), or genetic diseases like Aarkskog-Scott Syndrome (*ARHGEF7*) [154]. These might be an indication that such genes may already be over- or under-expressed in relation to the exposure variables (risk factors) before the onset/development of diseases.

Further, the biological insights obtained from the papers (Paper I-III) included in this thesis increases our understanding of the processes the risk factors we investigated are involved in. This knowledge can be utilized by future observational studies but also clinical trials that test new pharmacological entities based on gene expression assessments as a tool. A broader

understanding of biological functions of these known risk factors can shed light to yet underappreciated factors contributing to cancer development. Lastly, we observed the influence of the immune cells on blood gene expression profiles in this thesis and this knowledge can be important to acknowledge and include in future gene expression and cancer studies.

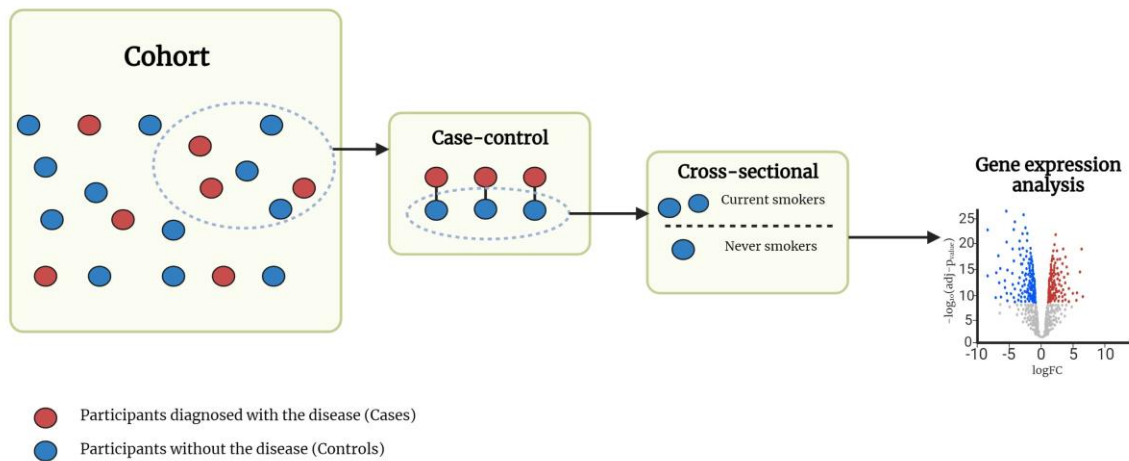
## **5.2 Methodological considerations**

### **5.2.1 Study design**

In this thesis, the blood collection for the gene expression analyses (the outcome variable) was performed at a single time point: a snapshot representation of blood gene expression. Further, the questionnaire information was collected at multiple time points (women participating had answered up to three main questionnaires and one questionnaire at the time of blood collection), thus we regarded repeated measurements of the exposure variables. The information at blood collection time point was our focus in all papers for comparisons of gene expression profiles between those who were exposed and those who were not exposed to various risk factors, for instance, comparing current to never smokers (Paper I), obesity to normal-weight (Paper II), and postmenopausal to premenopausal (Paper III). But we also used information from the preceding questionnaires to investigate exposures prior to the blood collection, like TSC in Paper I and past weight changes in Paper II. Therefore, the study design of all of papers (Paper I-III) comprised in this thesis was cross-sectional analysis nested within the prospective longitudinal NOWAC study.

This thesis used a study design that contained three levels of nested designs –a cohort, a nested case-control study, and a cross-sectional study including only controls (Figure 7). In this thesis, the 1<sup>st</sup> level nesting was the NOWAC postgenome cohort. Several case-control studies focusing on gene expression has been conducted within the NOWAC postgenome cohort (i.e., breast, lung, ovarian, and endometrial cancers, and diabetes), which was the 2<sup>nd</sup> level nesting. The participants of these studies have detailed information about various exposure variables such as smoking exposure, height, weight, HT use, etc. If we had included both cases and controls, distinguishing between the initial biological effects of cancer from the effects of exposure variables would have been challenging. But we could address our research question by identifying molecular signals or biomarkers for the selected exposure variables among the control groups. So, the 3<sup>rd</sup> level nesting was the cross-sectional study performed in this work, with only controls that were stratified based on the exposure variables (e.g., current smokers vs

never smokers) from the prospective questionnaires of the cohort study. Finally, to identify biological functions associated with these selected exposure variables, the gene expression profiles were analysed. A similar study design was also used by Barnung et al. 2018 [186] and the nesting of these designs was explained in a book by the initiator of the NOWAC cohort, Eiliv Lund [187].



**Figure 7:** An illustration of a three-level nested study design. (Created with [BioRender.com](https://www.biorender.com))

Adapted from Lund et al. 2020 under the Creative Commons CC BY 4.0 licence [187].

### *Strengths and weaknesses of the study design*

The foremost strength of cross-sectional studies is that they are usually feasible to conduct compared to other study designs and can provide a large sample of the population of interest. Only one time point is used to collect data for all variables, and multiple outcomes and exposures can be studied at the same time [188]. In this thesis, the study design has enabled us to have information on a relatively large sample size for transcriptomic analyses (N=1,716). In addition to the information on multiple variables collected at one time point (at blood collection), we had reliable, and comprehensive information of up to three main questionnaires (before the blood collection time point), strengthening our exposure variables because of the repeated measurements. This enabled us to investigate not only smoking status in Paper I, which was commonly used in previous studies investigating associations between smoking status and gene expression [32-34,37,39,138], but also to create and investigate quantitative measurements of smoking exposure i.e., the smoking metrics in Paper I, such as intensity,

duration, TSC, and CSI score. Likewise, in Paper II, we utilized the weight and height information of the participants from questionnaires at baseline, follow-up, and at the time of blood collection. These enabled us to calculate measures of past weight changes, including both long-term weight changes with seven years average mean interval and short-term weight changes with one-year average mean interval; and to create weight change categories (see Chapter 3.3.2). Further, this allowed us to calculate BMI at baseline, follow-up, and at the time of blood collection, that enabled us to introduce two interaction models including BMI and succeeding or preceding weight changes.

Another limitation in cross-sectional studies is that since the exposures and outcomes are measured at a single time point, it is difficult to make a causal inference, and researchers are unable to explore the temporal relationship between outcomes and exposures. In addition, the interpretation of identified associations might be challenging [188]. This has been a major limitation of all the papers (Paper I-III) included in this thesis as well, as we could not establish causal inference. For example, in Paper II, we do not know that whether BMI caused the gene expression to change or if altered gene expression could be representing factors that cause the variation in BMI of an individual.

As cross-sectional studies are observational studies, these are suitable for generating hypotheses and prevalence of outcomes and exposures [188]. This thesis can be considered exploratory research that can generate hypotheses for future studies. In cross-sectional studies, since study participants are not intentionally subjected to any exposure or treatment, there are not any ethical issues arising from intervention [188]. However, ethical issues related to the research questions or use of the information in later studies can cause ethical issues in this type of studies, therefore it is necessary to get approval from the concerned ethical committees. In this thesis, both NOWAC and NOWAC postgenome cohort are approved by the regional ethical committee- REK (see Chapter 3.5); and the participating women were informed beforehand that their blood samples could be used in gene expression/genetic studies later in future, and they had given consents for that purpose.

There are different potential errors and bias that may occur based on the study designs [188]. These issues in the context of this thesis are mainly discussed in the next sections.

## 5.2.2 Bias

Most studies based on samples taken from a population to estimate the present and future occurrence in the whole population are prone to errors [189]. Errors are broadly divided into *Random errors* and *Systematic errors* (also known as *Bias*); both of which lead to deviation of the estimate (over- or underestimate) from the true value or incorrect estimate of the true effect of an exposure on the outcome of interest [188-190]. Random errors are the portions of variation in a measurement that has no apparent connection to any other measurements or variables, and usually occurs due to chance [191]. Systematic errors (bias) are errors in the design, conduct, or analysis of the study [189,190] that is consistently deviating in a particular direction [191]. Potential biases and how to avoid them as much as possible should be carefully considered while planning a study, and the sources of bias and their impacts on the final results should be openly discussed to make valid conclusions [188,190,191]. In health research, bias can arise mainly due to (i) the approach/method used for selecting study participants, and (ii) the approach/method used for collecting or measuring data in the study [190]. There can be many systematic biases in a study, but for ease they can be further broadly divided into *Selection bias* and *Information bias* and these are discussed below [188-190].

### 5.2.2.1 Selection bias

*Selection bias* occurs in epidemiological studies from the procedures used to select individuals into the study or the analysis [191]. In interventional studies, selection bias can be largely minimized through random allocation to treatment and control groups [189]. This is not possible in observational studies like in the papers (Paper I-III) in this thesis. However, this thesis included women randomly sampled from the National Population Register of Norway. Further, an external validity study [192] conducted on the NOWAC study showed that the distributions of different exposure variables like smoking, weight, oral contraceptive use, parity, etc. had no statistical differences between the original responders and the non-responders. Thus, the proportions of exposure variables in each paper in this thesis are thought to represent those in the study population.

It is a known fact that 100% participation is never achieved in cross-sectional studies. The decision of whether to participate in a study or not is non-random and influenced by several factors, such as gender, age, socioeconomic status, substance abuse issues, etc. If any of these factors are associated with the exposure or outcome of interest, the sample being studied may not be truly representative of the background population [189]. Thus, *nonresponse bias* is one

of the common types of selection bias, which occurs when the characteristics of responders differ from non-responders. It is mostly encountered in cross-sectional studies that use mailed questionnaires for surveys. A large sample can still have a low response rate [188]. To avoid/minimize this bias, high response rates should be aimed for [189]. The relevant information collected in the papers (Paper I-III) of this thesis was based on questionnaires mailed to the participants, therefore this expected to lead to nonresponse bias. The external validity study [192] showed that the crude response rate for the NOWAC study was 57.1%. Lack of time and concern about privacy were the most important reasons among the non-responders for not returning the questionnaires [192]. This might have influenced the self-reporting of various variables in this thesis as well, for instance, smoking status and number of cigarettes, duration of smoking in Paper I or body weight in Paper II, or menopausal status in Paper III. But that study [192] also revealed that there was no major source of selection bias in NOWAC, which could utterly invalidate population attributable risk estimate. Further, responders in epidemiological studies tend to be healthier than those who do not participate, leading to an underestimation of disease prevalence in most surveys [189]. This thesis is based on blood sampling for gene expression data. The women were required to actively participate by visiting to their general physician for blood sampling. This might have led to selection bias, as these participants might be healthier or have easier access to health care than the non-responders. However, the overall response rate for NOWAC postgenome cohort, in which the blood sampling was performed, was approximately 74%, which is relatively high that might have minimized this bias [91]. Further, the gene expression data would likely not be influenced by selection bias or non-response bias. So, it is likely that our effect estimates have not been invalidated because of this bias in the exposure variables nor the gene expression profiles.

### **5.2.2.2 Information bias**

*Information bias* occurs when exposure, covariate, or outcome variables in a study are inaccurately measured, collected, or interpreted [188,191]. It is one of the most common sources of bias that impacts the accuracy of health research [190]. Some of the major information biases that has been observed in the papers included in this thesis are discussed below.

### 5.2.2.2.1 Self-reporting bias and recall bias

*Self-reporting bias* occurs when self-reporting of data is used, mostly in observational studies (e.g., cross-sectional, case-control, or cohort studies). Self-reporting includes questionnaires, surveys, or interviews and represent a common approach for collecting data in epidemiological studies, where participants respond to the questions without the interference of researchers. Compared to medical records or laboratory measurements, a self-reporting approach can be considered less reliable and more prone to self-reporting bias [190]. Study participants can provide incorrect responses depending upon their capacity to remember previous occurrences leading to *recall bias* [190,193]. However, one of the main strategies to access this bias could be internal or external validation of self-reporting instruments [190].

The issue of self-reporting bias, and to some extent recall bias, are one of the key limitations in the papers (Paper I-III) included in this thesis as these used self-reported/self-administered questionnaires for relevant information about the baseline and follow-up characteristics and exposure variables of interest. However, we included repeated measurements of exposures in a longitudinal study design, and by utilizing this, we were able to check for consistency across reported answers and thus strengthen our exposure variables internally, specifically: smoking status and smoking metrics (i.e., smoking intensity, smoking duration, TSC, pack-years, and CSI scores) in Paper I, and menopausal status in Paper III. Additionally, there are some reproducibility, reliability, and/or validation studies on self-reported smoking [194], BMI [195], menopausal status and HT use [121], which externally validated our exposure variables of interest.

In Paper I, we utilized these detailed and repeated information on past and recent smoking history such as TSC, duration and intensity of cigarettes smoked, etc. of the study participants to create comprehensive smoking metrics and to verify their reported smoking status at the blood collection time point. Most other studies with a similar aim was also based on self-reported smoking information from questionnaires [32,33,36-39]. But in addition to self-reported smoking status, several studies had measured concentrations of the metabolite of nicotine and/or cotinine in blood, urine, or saliva [32,37-39]. This could have added value for current smokers, but due to its relatively short half-life (16-19 hours) [196], it would not have provided valuable information for former smokers. Further, reliability study on self-reported smoking history demonstrated that current smokers had higher inconsistent reporting compared



to former smokers, however, self-reported tobacco use among smokers appear highly reliable over short time periods [194].

In Paper II, we could not use the repeated measurements to check for consistency, as the changes in weight between the repeated measurements was part of the analyses. The reported weight change could be actual weight change but also lead to some misreporting in our analyses. There were participants who had reported decreased weight at baseline and increased weight at follow-up questionnaires and vice versa; we defined them as weight-cyclers (N=160) and they might be representing self-reporting bias. So, to minimize this bias, we excluded the weight-cyclers from the analyses in Paper II. A validation study [195] had concluded that for middle-aged women in Norway, self-reported height and weight can still provide an accurate ranking of BMI.

In Paper III, we observed large variation in the self-reported menopausal status in the questionnaire before and at the blood collection time point. Again, the repeated measurements allowed for assessment of the inconsistency and create a new variable with menopausal status at the blood collection time point by merging information from both before and at the blood collection time point. Further, self-reported menopausal status and HT use defined by the different questionnaires in NOWAC study has been validated [121]. That study [121] showed that the questionnaire administered during the blood collection time point, exhibited a sensitivity of 92% (95% CI 89-96%) and a specificity of 73% (95% CI 64-82%) in determining menopausal status. On the other hand, the main questionnaires distributed at baseline and follow-up, yielded a sensitivity of 88% (95% CI 84-92%) and a specificity of 87% (95% CI 80-94%). Similarly, current HT use demonstrated a specificity of 100%.

#### **5.2.2.2.2 Misclassification bias**

*Misclassification bias* can occur when there is presence of systematic error in the information about exposure and/or outcome of interest [197]. This bias could be a result of the definition of cases, but also influence the interpretation of exposure measures, laboratory results, or other diagnostic procedures. In most studies, 100% sensitivity and specificity are not attained, and some degree of misclassification is expected to occur which will influence the conclusion of the study [198]. In this thesis, among the several sub-studies (case-control) in the NOWAC postgenome cohort, the cases and controls were verified from the Cancer Registry of Norway. Therefore, it is likely that the women we included, i.e., those who were never diagnosed with

cancer, in the papers (Paper I-III) included in this thesis were classified correctly. Further, misclassification of exposure variables could have been an issue in our papers in general as we used self-reported questionnaires. Particularly, misclassification of menopausal status could have been an issue, as we observed larger variation in responses of participants before and at the blood collection time point. This might be linked to differing definitions of menopause based on age or menstrual cycle. According to a study [199], improved classification of menopausal status can lead to more reliable findings and enhance the statistical power of health-related research. But this variation could also be because of variation in duration and symptoms from women-to-women. Nevertheless, the repeated measurements allowed us to assess the consistency of menopausal status of women in Paper III; and similarly, to create or identify smoking status, smoking metrics in Paper I. This has likely reduced the misclassification bias in this thesis. The NOWAC participants had a slight tendency to under-report weight and BMI, particularly among women with overweight/obesity [195], which means that there might be some women with higher BMI but misclassified in a lower BMI category, yet accurate ranking of BMI can be expected in NOWAC participants [195]. In our study, women with obesity were already the lowest in number (N=280) among the BMI categories (overweight: N= 622, normal-weight: N=751), still showed large number of DEGs in the obesity-vs-normal-weight and obesity-vs-overweight comparison (2,394 and 769 DEGs, respectively). Thus, although there would be some degree of misclassification and some women were in a lower BMI category, the BMI category for women with obesity did reveal DEGs and the BMI evaluated as a continuous variable is expected to be ranked right so our estimates should not have been influenced considerably or be slightly lower than if we had weight and height measured.

#### **5.2.2.2.3 Measurement bias**

*Measurement bias* (also known as *measurement error*, *instrumental error*, *measurement imprecision*, or *measurement error bias*) arises from inaccurate measurements of subjects on study variable(s), resulting in misleading conclusions [190,191]. It mainly occurs due to device inaccuracy, environmental conditions in the laboratory, or self-reported measurements [190]. This thesis incorporates measurements of blood gene expression profiles; and different processes were involved such as collection of blood from participants, shipping of the blood samples, freezing, thawing, and handling of blood samples for gene expression analyses at the laboratory. This might give rise to some degree of measurement bias, which could have influenced our estimates in the papers included in this thesis. However, various techniques

and/or measurements had been followed to reduce this bias. A detailed discussion on this topic is presented in section 5.2.6 and 5.2.7 below.

### 5.2.3 Confounding

In cross-sectional studies, *confounding* can arise when a variable is linked with the exposure and influences the outcome, i.e., for a variable to be identified as a confounder, it must satisfy three criteria – (i) associated with the exposure of interest, (ii) associated with the outcome of interest, and (iii) must not be the causal pathway between exposure and outcome [188,191]. It is important to account for possible *confounders* (also called *confounding variables*), as their presence may lead to alteration or misrepresentation of the actual association between the exposure and the outcome [191].

To identify the possible confounders, we assessed *a priori* selected covariates, if they were significantly associated –(i) with the different exposure(s) of interest (in brief: smoking status in Paper I, BMI in Paper II, and menopausal status in Paper III) according to Chi-square or Kruskal-Wallis tests, and (ii) with overall gene expression data according to the ‘Global test’ from Bioconductor package ‘*global test*’ [134]. Further, we considered information from the laboratory processing of the blood samples as technical covariates (i.e., laboratory plates/batches and sample storage time) and potential confounders.

As blood gene expression is influenced by the circulating blood cells [39], we considered the estimated WBC proportions in our data to be possible confounders. We assessed whether the estimated WBC proportions were significantly associated –(i) with the above-mentioned exposure(s) of interest, and (ii) with overall gene expression profiles. To do this we also used the Kruskal-Wallis test and Global test, respectively. However, we cannot disregard that the relative presence of these cells and it could still act be confounding factors as we used WBC estimates and that the observed associations could be results of residual confounding. Additionally, our data lacked information on RBCs and reticulocytes; and these cell types were not part of the CIBERSORT LM22 deconvolution matrix, so it was not feasible to estimate their proportions as well. Thus, such adjustments were not possible to observe their effects. But based on the results of Paper 2, it is a limitation that RBCs/reticulocytes could not be considered in these kinds of studies.

These possible confounders were then included in the limma analyses in different models that included different adjustments (See Figure 5 in Chapter 3.4.3). In Paper I and Paper II, we introduced two adjustment models –(i) minimally-adjusted (ii) fully-adjusted. Minimally-adjusted models included the technical covariates, while fully-adjusted models additionally included selected covariates and estimated WBCs as mentioned above. Considering consistency for all three papers, we initially defined the corresponding two adjustment models for Paper III as well. However, we noticed a considerable reduction in the number of DEGs, which prompted us to investigate covariates further and discovered that the WBC estimates were influencing the models more than in Paper I and II. Thus, we introduced three different adjustment models in Paper III –(i) Model-1 (ii) Model-2 (iii) Model-3. Model-1 was minimally-adjusted for the technical covariates, Model-2 was semi-adjusted that additionally included the selected covariates, while Model-3 was fully-adjusted including technical covariates, selected covariates, and estimated WBCs as mentioned above. These three layers of adjustment models helped us describe the influence of WBCs on the investigation of gene expression and menopausal status in Paper III. The relative meaning of WBCs was the strongest in Paper III, still it could be that RBCs could be important for both Paper I and II although we could not adjust for them.

#### **5.2.4 Interaction (effect modification)**

*Interaction*, also known as *effect modification*, occurs when the impact of one explanatory variable on the outcome depends on specific level or value of another explanatory variable. If an interaction effect is present between exposures, it means that these factors cannot be considered independent in causing a particular outcome [197]. In Paper II, where BMI and weight changes were exposure variables, we analysed association of gene expression and weight change within an interaction with BMI categories. If we had not defined an interaction model, we would have failed to observe that weight changes were associated with blood gene expression only among women with obesity, and not among women with overweight. It means that it was the women with the highest BMI driving the associations. Also, we introduced two different models for this interaction –(i) included BMI and succeeding weight changes, and (ii) included BMI and preceding weight changes. This enabled us to access that current BMI was stronger predictor than the past BMI.

### **5.2.5 Generalizability**

The NOWAC study is a national population-based cohort study, which has utilized Norway's existing population registers for sampling. Women in the corresponding age-groups are randomly selected using the National Population Register [121]. Thus, it can be considered as representative of the entire female population of Norway in those age-groups [192]. However, it is important to consider that the generalizability of these findings to women in the same age-groups today may not be entirely accurate, as women today may differ from the participants involved in this study conducted two decades ago. Additionally, since only ethnic Norwegians were invited to participate in the NOWAC study, the results from this thesis cannot be generalized to the entire population of Norway.

When possible, we had performed quantitative replication in all the papers (Paper I-III), comparing our results to results from similar studies, and observed overall consistency. Thus, if the population under study shares similar variables of exposure as those investigated in this thesis, the findings could potentially be applicable to broader populations in general.

### **5.2.6 Gene expression analyses**

We had analysed gene expression in whole-blood samples as an outcome variable associated with selected exposures in all papers (Paper I-III). Using gene expression profiles give rise to several challenges and limitations, especially related to its technology. Therefore, in this section we mainly discuss the considerations related to determination of gene expression profiles.

#### **5.2.6.1 'Blood' as a tissue for gene expression studies**

In this thesis, whole-blood (also simply called 'blood') was the target tissue for gene expression analyses in all papers (Paper I-III). Blood is a desirable sample material for biomarker research because of its easy access, minimal invasiveness, and relatively lower cost for collection [200]. It can uncover important characteristics that are significant for investigating features relevant to human health. These make blood an attractive sample for diagnostic purposes [200]. Different studies have found altered gene expression in blood and shown that blood is a suitable tissue for measuring exposure variables that have been investigated in this thesis.

The toxic components of tobacco smoke initially taken up by the lungs and subsequently enter the blood stream before being distributed throughout the body, which makes blood a suitable biological material to study the systemic influences of exposure to tobacco smoke [138].

Further, many studies have shown that current exposure to tobacco smoke is associated with altered gene expression in blood of several genes (*LRRN3*, *CLDND1*, *GPR15*, *ATF4*, *SOD2*, *CDKN1C*) [32-39]. Our study also showed strong associations of blood gene expression and smoking exposures, indicating that blood samples indeed reflect the systemic influence of smoking exposure.

In Paper II, adipose tissue would be biologically most relevant to investigate obesity [61] probably due to mechanisms related to obesity. Several studies also show that gene expression profiles in adipose tissue have been associated with obesity and weight loss [64-71,201]. However, there are studies that have examined gene expression related to obesity either in whole-blood [61,62] or PBMCs [63], and found associations between them. Further, our study observed large associations of blood gene expression and BMI, indicating that blood samples indeed reflect the systemic influence of obesity.

There is one study that investigated gene expression profiles in peripheral blood monocytes among healthy pre- and postmenopausal women that revealed that the functional state of circulating monocytes is influenced by menopause [84]. Related to HT use, there have been few transcriptomic studies in blood [90-92,170], showing small associations. However, our study (Paper III) demonstrated clear associations of blood gene expression with HT use and specifically with menopause, indicating that blood samples indeed reflect the systemic influence of menopause.

Thus, ‘blood’ can be regarded as an appropriate tissue in transcriptomic analyses for investigating systemic influence of various exposures like those investigated in the papers included in this thesis (in brief: for smoking status and smoking metrics in Paper I, obesity and weight change in Paper II, and menopause and HT use in Paper III).

#### **5.2.6.2 Blood cell compositions in blood**

Gene expression studies based on blood samples are challenged by its complicated biological system that consists of diverse cell types in various developmental stages [202]. Blood cell populations are heterogeneous, varying in terms of absolute and relative presence in the blood and differ across individuals. Their presence heavily influences the gene expression profiles in blood and is generally recognized as a major source of bias and variability [203]. It can lead to unexpected variation in gene expression levels between different samples [100].

In this thesis, the blood samples of participants were drawn at the offices of their local general physician into a tube provided to the participants by mail. These tubes, with the blood samples, were then returned to UiT The Arctic University of Norway by mail. Thus, flow cytometer counts were not planned as part of the establishment of this sample collection. Thus, our data lacked information on counts of WBC, RBCs, and reticulocytes. We estimated the WBCs proportions that were derived from the sample gene expression data using an *in-silico* gene expression deconvolution method CIBERSORT, and the LM22 signature matrix [127]. We then considered the estimated WBC proportions as possible confounders and have included them as adjustments in our models in all papers (Paper I-III) included in this thesis (See Figure 5 in Chapter 3.4.3). Still RBCs and reticulocytes were not part of the CIBERSORT LM22 deconvolution matrix and it was not possible to estimate their proportions. Thus, such adjustments were not possible to see their effects.

One potential shortcoming of using the CIBERSORT estimates is that these are based on the same data that we were later statistically analysing to identify DEGs. Ideally, the cell type estimates should have come from an independent measurement [127]. Further, the estimated proportions of WBCs in the papers (Paper I-III) deviated from the expected range [177,178]. However, other recent studies based on the NOWAC postgenome cohort [179,180] have also observed this variation. This could be due to a potential bias attributed to the deconvolution method or data pre-processing [179,181]. Still, absolute differences in estimated WBCs across the exposures we used in the papers (Paper I-III) were modest. Further, the identified top-ranked genes in different adjustment models in Paper I remained largely consistent, indicating that these genes were not substantially influenced by distributions of WBCs. In Paper II and III, we noticed that these cell populations have been reflected in the biological functions of DEGs; and especially, in Paper III, adjustment of WBC proportions influenced the models in terms of number of DEGs drastically. This is discussed in more detail in Chapter 5.1. Thus, this thesis demonstrates that blood cell compositions (WBCs, RBCs/reticulocytes), should be considered in the models while using blood as a tissue for gene expression analyses.

### **5.2.6.3 Sample collection technique**

Before conducting gene expression analyses on blood samples in studies with larger samples, it is crucial to set up a reliable and reproducible method, however it can be challenging due to specific features of the samples and their collection methods [204]. There are two major

commercially available blood collection tube systems for isolation of high-quality RNA from blood –(i) the PAXgene™ blood RNA system (PreAnalytiX QIAGEN/BD, Hombrechtikon, Switzerland) and (ii) the Tempus™ blood RNA system (Applied Biosystems, Foster City, CA, USA) [200]. PAXgene™ collection tubes contain a proprietary blend solution that preserves the RNA molecules from degradation by immediately stabilizing them after collection using RNases enzymes and further prevents alteration of gene expression that can occur during sample handling and storage [204]. On the other hand, Tempus™ tubes incorporate a proprietary stabilizing solution that aids in preserving RNA and stability during the process of storage and transportation. These tubes are specifically designed to be compatible with various downstream analysis techniques, such as next-generation sequencing. By utilizing Tempus™ tubes, researchers can enhance the quality and sensitivity of RNA expression signatures acquired from the sample. Both have unique properties and features that can influence the quality and stability of the RNA in the sample, and thus when selecting between PAXgene™ and Tempus™ blood tubes for RNA expression analysis, the study's requirements and downstream analysis should be considered. Factors like storage time, sample purpose, and analysis type influence the tube choice. Additionally, validating results with controls and quality measures for accurate and reliable RNA expression signatures is crucial [200,205].

In this thesis, we used the PAXgene™ collection tubes to collect the blood samples for mRNA analyses in all the papers (Paper I-III). The blood samples in this thesis were collected, and the samples were mailed to UiT The Arctic University of Norway with a maximum of four days mail time to Tromsø that was registered. Then, the samples have been stored in freezer for a long time before analyses. PAXgene™ are specifically ideal for long-term storage or transport over long distances. The RNA preservation in PAXgene™ tubes is also relatively consistent, making the RNA expression signatures reproducible across various samples and conditions [205]Further, a study [204] observed strong variations in gene expression when different sampling methods and extraction kits were combined. We had employed PAXgene™ collection tubes and followed the PAXgene™ RNA blood kit protocol for RNA extraction; and following such protocol is considered to yield good quality and quantity of RNA [204].

#### **5.2.6.4 Time and date of sampling and storage duration**

Researchers have observed variation in gene expression over a 24-hr period and across different seasons [100,118]. Thus, it is recommended to collect the samples at the same time of day in



all participants, or if not possible, at least record the date and time of the sample collection; this may prevent potential sources of variability in downstream analyses [100]. In this thesis, time and date of sampling were self-reported, as women participating on the NOWAC postgenome cohort went to their local general physician's offices to draw blood at their own timing. Still, the time of blood collection was narrowed to typically 8am-16pm and there was no large pattern related to this factor in the pre-processing of the data. Thus, we did not adjust for time and date of sampling. Still when it was examined in the pre-processing of the data for these studies, it did not render a major influence but also that the variables were likely not very complete or reliable.

Additionally, as a technical covariate, we have adjusted for sample storage time as a proxy for duration of sample storage in freezers in our models, that might have possibly reduced the influence of sample storage time.

#### **5.2.6.5 Microarray as measurement technique**

There are several techniques for measurement of gene expression analyses; and each technique can introduce technical variability that can affect the accuracy and reproducibility of gene expression measurements [100]. Over time, there have been notable advancements in the measurement techniques used for gene expression analyses (Details in Chapter 1.2.2). At the initial phases, technologies were basic, but the rapid technological progress has enabled these techniques to offer higher sensitivity, throughput, and the ability to study gene expression at the single-cell level.

In this thesis, a microarray technique was used for gene expression profiles for all the papers (Paper I-III). The microarray technique was probably the best available technique around 20 years back when blood sample collection started for mRNA analyses in the NOWAC postgenome cohort (2003-2006), as microarrays were more economical option than RNA-seq and are also reliable for gene expression profiling [206]. But around 10 years ago, it was predicted that RNA-seq would be used more frequently [206]. Indeed RNA-seq has become a widely used technique today and replaced microarrays as the technology of choice for transcriptome-wide analyses. Current technologies focus on single-cell and spatial transcriptomics but future advancements in measurement techniques are expected and can expand research questions in the field.

Although the microarray results in this thesis are still reliable and overall comparable to RNA-seq results, non-coding RNAs and splice variants cannot be detected by microarrays. Another challenge when using the microarray technique was the probe-to-gene annotation, but this challenge might also be present when using other techniques such as sequencing-based techniques. We used the Bioconductor packages *'lumi'*, *'lumiHumanIDMapping'*, and *'illuminaHumanv4.db'* [124-126] for gene annotation, which resulted in 9,095 probes and 7,713 unique genes. The number of probes reflects a choice of stringent filtering criteria, and we would not have expected to detect a greater number of DEGs in our studies if we had more probes, as the excluded probes likely contained more random variability than those analysed. Further, microarray data can achieve similar detection as compared to RNA sequencing technologies depending on sequencing depth. Generally, microarrays can be quite sensitive for low-expressed probes when compared to sequencing technology [207]. Further, future studies using cell-type-specific gene expression analyses techniques such as scRNA-seq would add value to our findings of importance of blood cell compositions in blood gene expression studies.

## **5.2.7 Statistical analyses**

### **5.2.7.1 Pre-processing of data**

Technical outliers are observations that were altered in some way in the laboratory or during the sampling process. These observations, if not removed, may introduce bias and variance in later statistical analysis and distort the results [123]. There are challenges in removal of outliers, specifically, studies containing blood-based samples. Compared to tumor tissue-based studies, blood-based studies have greater biological variability in gene expression data and the blood-derived signals are weak and variable. Thus, distinguishing between outliers versus non-outliers, as well as signal versus noise, is more complex task in blood samples [123].

Before performing the main statistical analyses in this thesis, the gene expression data were pre-processed. Technical outliers that are identified by the laboratory quality measures were first removed. Then any outliers that were detected from a standard operating procedure with *nowaclean*, which were designed for large samples like in the NOWAC postgenome cohort [123], were removed. Then, background correction was performed, bad-quality probes were removed, and probes detected in <20% of samples were filtered.

Further,  $\log_2$  transformation, quantile normalisation, and inspection of batch effects using PCA plots were performed before extracting the controls in the different case-control datasets and

merged to a larger dataset. The *log<sub>2</sub> transformation* is the most common transformation technique for microarray data. It stabilizes the variance of high intensities while simultaneously increases the variance at low intensities [208]. Normalisation techniques are commonly used to re-distribute signal intensities across all samples to achieve a consistent distribution e.g., same mean and/or standard deviation, ultimately leading to removal/reduction of technical variabilities [209]. Many normalisation techniques exist, but *quantile normalisation* is particularly popular and exhibits very well-aligned distributions. It is widely used and is standard part of analysis pipelines for high-throughput analysis [209]. Several normalisation techniques (e.g., ComBat [210]) were evaluated in the pre-processing of the dataset in this thesis, and quantile normalisation was chosen because it exhibited the most well-aligned distributions. *Batch effects* refer to differences in gene expression measurements that are introduced when samples are processed on different days, in different groups, or different laboratory personnels. Batch effects are widely recognised as sources of latent (hidden or concealed) variation in genomic experiments [203]. These batch effects can confound the interpretation of gene expression data and reduce the ability to detect true biological differences. Thus, the goal of assessing batch effects was to evaluate whether there were strong effects resulting from microarray experiment variance rather than from biological variation among the samples. During the inspection of batch effects in our studies, no strong effects were indicated in PCA plots, still we included laboratory plates/batch in the adjustment models also as a proxy variable for case-control study origin.

### **5.2.7.2 Gene-level analyses**

One of the objectives of conducting gene expression analysis is to identify differences in the levels of gene expression between different groups, typically on a gene-by-gene basis or for sets of genes. Some of the statistical analyses tools we used for this analysis are discussed below.

#### *Linear models*

The primary statistical method to select DEGs between two groups is to use a t-test, but its use has been criticized in literatures. The computation of a t-tests can encounter challenges due to the potential skewing of variance estimates caused by genes with very low variance. Another limitation arises from its application on small sample sizes resulting in reduced statistical power. As a result, the effectiveness of t-tests and the significance of variance modeling have

been strongly questioned. This has prompted the development of numerous innovative alternatives, aiming to improve variance estimation accuracy and power; and among the different approaches, limma shows significant improvement over the t-test [211].

In this thesis, we used the limma package for the main analysis in all the papers (Paper I-III). Limma is a software package designed for analysing gene expression data obtained from microarray or RNA-seq technologies [212]. It is commonly used approach based on linear models that effectively incorporates various statistical principles for conducting large-scale expression studies [130]. One of the key strengths of limma is its capability to simultaneously analyse comparisons across numerous RNA targets. Moreover, it incorporates features that ensure stability in analyses, even when dealing with a small number of arrays, achieved through the borrowing of information across genes. Limma is specifically tailored to handle complex experiments encompassing various experimental conditions and predictors [212]. In this thesis, we have used its version for continuous variables (analyses of smoking metrics in Paper I and BMI in Paper II), categorical variables (smoking status in Paper I, BMI categories in Paper II, menopausal status in Paper III) and for interaction terms (weight change-BMI interaction models in Paper II). The analyses were based on a design matrix offered by limma, which made the definitions of tests very flexible.

### *Multiple testing challenges*

Examining multiple hypotheses within a single study is quite common in gene expression analyses, which can lead to increased probability of Type I errors [213] (i.e., false positives or falsely rejected hypotheses). To overcome this, statistical tools for multiple testing corrections are required. The simplest method is the Bonferroni correction, and this method controls the family-wise error rate, which is the likelihood of encountering Type I error in at least one of the many hypothesis tests. It calculates the significant threshold by simply dividing the desired alpha by the number of tests, i.e., if alpha is 0.05 then the threshold is  $0.05/\text{number of tests}$  [213]. In genetic studies involving several hundred thousand hypotheses, controlling for family-wise error rate seems too conservative [213]. An alternative is to control the FDR, which is the expected proportion of false positives among the rejected hypotheses. The most common method for false discovery rate (FDR) control is the Benjamini-Hochberg correction [213]. This method first sorts the p-values obtained from the statistical tests in ascending order, and then

calculates the critical value (q-value) for each p-value using the formula:  $q\text{-value} = (p\text{-value} * N) / k$ , where N is the total number of genes and k is the rank of the p-value [131].

In this thesis, we controlled for FDR, using the Benjamini-Hochberg correction [131]. Benjamini-Hochberg correction is less conservative than Bonferroni correction. Taking this thesis as an example, we had 7,713 unique genes. The Bonferroni correction threshold would be  $0.05/\text{number of tests}$ , i.e.,  $0.05/7713 = 6.48e-06$ . In Paper I, the top-ranked gene (*LRRN3*) from current-vs-never smokers had FDR adjusted-p-value= $1.52E-80$ . If we had used the Bonferroni threshold, we would still have detected *LRRN3* but would have missed to identify a lot of genes.

### *Effect estimates in gene expression analyses*

Early microarray publications [214,215] evaluated differential expression solely based on fold-change; where a cutoff of two-fold was typically considered significant [216]. The average log-ratio between two groups is evaluated and a gene is considered differentially expressed that vary by more than an arbitrary cutoff. It does not take the variance of the samples into account and therefore lacks a robust statistical foundation [211]. Instead of relying solely on a fold-change cutoff, one should prefer the use of statistical tests, as they incorporate variance when assessing differential expression [211]. Thus, we did not use the fold-change cutoffs but only used the FDR-adjusted p-value cutoffs to obtain the top-ranked genes between compared groups in all papers (Paper I-III) in this thesis. However, we still extracted fold-changes to represent effect estimates. A logFC value of one typically indicates a two-fold increase in gene expression compared to a reference group. The absolute value of the logFC indicates the magnitude of change in gene expression (or some other measured quantity), while the sign (positive or negative) indicates the direction of change, i.e., if  $\logFC > 0$  (or positive sign), then the gene was over-expressed, while  $\logFC < 0$  (or negative sign), then the gene was under-expressed compared to a reference group. For example, in Paper I, the logFC of *LRRN3* was 1.01 among current-vs-never smokers, it implies that the expression of *LRRN3* was approximately two times higher in current smokers compared to never smokers, and as it had positive sign, it represented over-expression of *LRRN3* among current smokers. We observed that the fold-changes of the top-ranked genes between different groups of interest in all the papers (Paper I-III) included in this thesis were modest (ranging from 0.07 to 0.86 in Paper I-III), except for *LRRN3* with 1.01 logFC. This might be because our study participants were a

large sample of healthy population. Other studies investigating blood gene expression and BMI also showed small effect estimates like logFC of up to 0.06 [62,63], but there are also studies that had large effect estimates (up to logFC=3.19) [61]. However, literature showed that most of the informative RNAs and differentially expressed transcripts can exhibit fold-changes less than two, and still reveal enrichment of biologically relevant functions [217]. Therefore, we believe that the modest fold-change differences we observed has not affected our enrichment analyses in all papers.

### **5.2.7.3 Pathway-level analyses**

To gain insight into biological functions of genes that are differentially expressed researchers have performed pathway analysis since long, as these are relatively straight forward and provide greater exploratory power to the results [218]. Functional enrichment analysis is widely used for interpreting gene lists or genome-wide regions of interest that are derived from different high-throughput studies [136]. These approaches utilize pathway knowledge in public repositories such as GO or KEGG, and therefore can be called *knowledge base–driven pathway analysis* [218]. In this thesis, we used over-representation analysis (ORA) as the functional enrichment analysis in all the papers (Paper I-III). Furthermore, we considered gene ontologies (GO) of biological processes, molecular functions, and cellular components. In addition to that, we also examined KEGG and REACTOME databases to increase our knowledge about well-driven pathways. Another functional enrichment analyses that could be used was gene set enrichment analysis (GSEA). GSEA uses the ranked list of genes and their effect estimates, while ORA simply compares the names of the gene sets with a background distribution of gene names. This can lead to more results indicated than from a GSEA, especially when analysing noisy or heterogenous datasets. We attempted both ORA and GSEA in Paper III, however, the indicated ontology terms were rather similar (results not shown).

## 6 Conclusions

In general, the findings of this thesis concludes that there are large associations of blood gene expression with cancer risk factors investigated, mainly –smoking, BMI, menopause –among women in the NOWAC postgenome cohort. Further, the differentially expressed genes (DEGs) indicated various gene ontologies and pathways, providing biological insights into their physiological and molecular mechanisms.

In more details, the specific conclusions are as following:

- Among all the risk factors investigated, the strongest association in terms effect sizes was shown by smoking status (the highest logFC exhibited by the top-ranked gene, *LRRN3*, when current smokers were compared with never smokers).
- Among all the risk factors investigated, the most associations in terms of largest numbers of DEGs were observed for investigations of obesity (2,394 DEGs when women with obesity were compared to women with normal-weight).
- Among all the risk factors investigated, the associations with menopause, and also HT use, were largely driven by the relative presence of estimated WBCs.
- Current exposures were clearly reflected in blood gene expression more than past exposures (past weight changes, former smoking). Thus, to the best of our knowledge, this thesis is the first to conclude that current BMI is reflected in blood gene expression more than past weight changes in terms of number of DEGs.
- The biological functions of smoking-, BMI-, and menopausal-associated DEGs mainly revealed enriched terms like metabolic, immune, and RBC-related processes/functions. These terms coincide with the physiological effects known for each risk factor and reflect their systemic impacts.
- The *LRRN3* expression increases with ongoing smoking exposure and reverts to levels like those of never smokers in years after smoking cessation. Because of the *LRRN3*-driven discrimination of smoking exposure, we concluded that *LRRN3* could supplant self-reported smoking data in future studies.
- This thesis provides knowledge on the influence of relative proportions of blood cells on the associations of blood gene expression with the risk factors investigated, reflecting its importance. Future studies should consider adjustment of blood cell proportions in studies related to blood-based markers.





## 7 Future perspectives

This thesis presents exploratory findings related to blood gene expression and important cancer risk factors. This work provides valuable insights in future investigations of how gene expression changes or responds to these factors and hypotheses regarding how exogenous and endogenous risk factors or a combination of these affect the overall health of women.

This thesis is the first to investigate associations of gene expression in blood with smoking metrics (Paper I), past weight changes (Paper II), and menopausal status (Paper III) in a large population-based samples. Some novel signals are reported should be validated by future studies using alternative targeted technologies, for instance, qPCR or NanoString, or utilizing larger, independent cohort samples.

Further, this thesis presented only a snapshot of gene expression, and thus could not capture the dynamics of gene expression. Temporal variation in associations could be addressed if future studies use repeated measurements for gene expression, along with questionnaire information, to assess changes in expression profiles.

Most importantly, this thesis depicts the influence of blood cell compositions on the associations of blood gene expression and risk factors investigated in this thesis, particularly the influence of WBCs on menopause (Paper III), and the influence of RBCs on smoking and BMI (Paper I and II). This, in itself, is the main open question to be addressed in future studies. But this question is also related to gene expression changes within individual cell types. More specifically, the two key questions for future studies are: (i) to what extent do the three exposure variables (smoking, BMI, and menopause) affect whole-blood cell type composition? (ii) to what extent do these exposure variables affect gene expression within individual cell types in whole-blood? If these are addressed in future, we might unravel novel insights into the biological mechanisms for how these three exposure variables are cancer risk factors.

We demonstrated that proper consideration and adjustments related to immune cell compositions are essential to incorporate in study design and data analyses when investigating blood molecular markers. Future investigations could design studies that can validate the findings related to differences in blood cell-type compositions. But we only considered the estimated proportions of WBCs and lacked information on other blood cell types, e.g., RBCs or reticulocytes. It would be of value to include measurements and not the estimates of the

blood cells in future studies. This could be done by analysing WBC compositions directly by using laboratory techniques like flow cytometry. Further, cell-type-specific gene expression analyses using methods such as scRNA-seq could be in future studies to investigate gene expression within individual blood cell types. This would help to access their individual influence, and to answer if it is functional state of cells (is the cell functioning differently?) or the distribution of cells (is there a difference in their absolute or relative presence?) that is driving the associations. But this would require completely new sample collection and large budget for the laboratory analyses.

## 8 References

1. National Human Genome Research Institute. Human Genome Project: [www.nih.gov](http://www.nih.gov); 2022 [Accessed 10 May, 2023]. Available from: <https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project>.
2. Jones PA, Baylin SB. The epigenomics of cancer. *Cell*. 2007;128(4):683-692.
3. Brennan P, Davey-Smith G. Identifying novel causes of cancers to enhance cancer prevention: new strategies are needed. *JNCI: Journal of the National Cancer Institute*. 2022;114(3):353-360.
4. Lund E, Dumeaux V, Braaten T, Hjartåker A, Engeset D, Skeie G, et al. Cohort profile: the Norwegian women and cancer study—NOWAC—Kvinner og kreft. *International journal of epidemiology*. 2007;37(1):36-41.
5. National Cancer Institute. Risk Factors for Cancer: <https://www.cancer.gov/>; 2015 [Accessed 12 May, 2023]. Available from: <https://www.cancer.gov/about-cancer/causes-prevention/risk>.
6. World Health Organization. Cancer: [www.who.int](http://www.who.int); 2022 [Accessed 10 April, 2023]. Available from: <https://www.who.int/news-room/fact-sheets/detail/cancer>.
7. Tran KB, Lang JJ, Compton K, Xu R, Acheson AR, Henrikson HJ, et al. The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. 2022;400(10352):563-591.
8. American Cancer Society. REPRODUCTIVE & HORMONAL FACTORS: <https://canceratlas.cancer.org/>; 2023 [Accessed 14 April, 2023]. Available from: <https://canceratlas.cancer.org/risk-factors/hormones/>.
9. World Health Organization. Menopause: [www.who.int](http://www.who.int); 2022 [Accessed 10 April, 2023]. Available from: <https://www.who.int/news-room/fact-sheets/detail/menopause>.
10. World Health Organization. Tobacco: [www.who.int](http://www.who.int); 2022 [Accessed 7 April, 2023]. Available from: <https://www.who.int/news-room/fact-sheets/detail/tobacco>.
11. National Cancer Institute. Cigarette: <https://www.cancer.gov/>; 2023 [Accessed 10 April, 2023]. Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cigarette>.
12. Leffondré K, Abrahamowicz M, Xiao Y, & Siemiatycki J. Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Statistics in medicine*. 2006;25(24):4132-4146.
13. National Cancer Institute. Secondhand Smoke and Cancer: <https://www.cancer.gov/>; 2023 [Accessed 11 April, 2023]. Available from: <https://www.cancer.gov/about-cancer/causes-prevention/risk/tobacco/second-hand-smoke-fact-sheet>.
14. Tobacco Free Life. History of Tobacco: <https://tobaccofreelife.org/>; 2016 [Accessed 11 April, 2023]. Available from: <https://tobaccofreelife.org/tobacco/tobacco-history/>.
15. Berridge V. Tobacco control in the WHO European Region. World Health Organization. Regional Office for Europe; 2020.
16. Proctor RN. The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco control*. 2012;21(2):87-91.
17. Vedøy TF, et al. Smoking and snus use in Norway. Public Health Report 2018: Norwegian Institute of Public health (NIPH); 2018.
18. Organization WH. WHO global report on trends in prevalence of tobacco use 2000–2025. 2021.

19. Statistics Norway. Tobacco, alcohol and other drugs: [www.ssb.no/en](http://www.ssb.no/en); 2023 [Accessed 20 April, 2023]. Available from: <https://www.ssb.no/en/helse/helseforhold-og-levevaner/statistikk/royk-alkohol-og-andre-rusmidler>.
20. Metrics IfH, Evaluation. Global Smoking Prevalence and Cigarette Consumption 1980–2012. Institute for Health Metrics and Evaluation (IHME) Seattle, United States; 2014.
21. Kondo T, Nakano Y, Adachi S, Murohara T. Effects of tobacco smoking on cardiovascular disease. *Circ J*. 2019;83(10):1980-1985.
22. Bonnie RJ, Kwan LY, & Stratton KR. Public health implications of raising the minimum age of legal access to tobacco products: National Academies Press Washington, DC; 2015. 91-123 p.
23. World Health Organization. WHO report on the global tobacco epidemic, 2008: the MPOWER package: World Health Organization; 2008.
24. McDaniel JC, Browning KK. Smoking, chronic wound healing, and implications for evidence-based practice. *Journal of wound, ostomy, and continence nursing: official publication of The Wound, Ostomy and Continence Nurses Society/WOCN*. 2014;41(5):415.
25. Sopori M. Effects of cigarette smoke on the immune system. *Nature Reviews Immunology*. 2002;2(5):372-377.
26. Centers for Disease Control and Prevention (US); National Center for Chronic Disease Prevention and Health Promotion (US); Office on Smoking and Health (US). How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta (GA): Centers for Disease Control and Prevention (US); 2010 [cited 2023 10 April]. Available from: <https://pubmed.ncbi.nlm.nih.gov/21452462/>.
27. Carey MA, Card JW, Voltz JW, Arbes Jr SJ, Germolec DR, Korach KS, et al. It's all about sex: gender, lung development and lung disease. *Trends in Endocrinology & Metabolism*. 2007;18(8):308-313.
28. Langhammer A, Johnsen R, Holmen J, Gulsvik A, & Bjermer L. Cigarette smoking gives more respiratory symptoms among women than among men The Nord-Trøndelag Health Study (HUNT). *J Epidemiol Community Health*. 2000;54(12):917-922.
29. NIDDK. Smoking and the Digestive System <https://www.cancer.gov/>; 2023 [Accessed 10 May, 2023]. Available from: [www.digestive.niddk.nih.gov](http://www.digestive.niddk.nih.gov).
30. Agnihotri R, Gaur S. Implications of tobacco smoking on the oral health of older adults. *Geriatr Gerontol Int*. 2014;14(3):526-540.
31. Yazdanparast T, Hassanzadeh H, Nasrollahi SA, Seyedmehdi SM, Jamaati H, Naimian A, et al. Cigarettes smoking and skin: a comparison study of the biophysical properties of skin in smokers and non-smokers. *Tanaffos*. 2019;18(2):163.
32. Beineke P, Fitch K, Tao H, Elashoff MR, Rosenberg S, Kraus WE, et al. A whole blood gene expression-based signature for smoking status. *BMC Med Genomics*. 2012;5:58.
33. Cheng X, Ferino E, Hull H, Jickling GC, Ander BP, Stamova B, et al. Smoking affects gene expression in blood of patients with ischemic stroke. *Annals of Clinical and Translational Neurology*. 2019;6(9):1748-1756.
34. 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. *Human Molecular Genetics*. 2016;25(21):4611-4623.
35. Lampe JW, Stepaniants SB, Mao M, Radich JP, Dai H, Linsley PS, et al. Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. *Cancer Epidemiology and Prevention Biomarkers*. 2004;13(3):445-453.
36. Martin F, Talikka M, Hoeng J, & Peitsch MC. Identification of gene expression signature for cigarette smoke exposure response--from man to mouse. *Hum Exp Toxicol*. 2015;34(12):1200-1211.

37. Na HK, Kim M, Chang SS, Kim SY, Park JY, Chung MW, et al. Tobacco smoking-response genes in blood and buccal cells. *Toxicology Letters*. 2015;232(2):429-437.
38. Van Leeuwen DM, van Agen E, Gottschalk RW, Vlietinck R, Gielen M, van Herwijnen MH, et al. Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. *Carcinogenesis*. 2007;28(3):691-697.
39. 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):550-560.
40. Arimilli S, Madahian B, Chen P, Marano K, & Prasad GL. Gene expression profiles associated with cigarette smoking and moist snuff consumption. *BMC Genomics*. 2017;18(1):156.
41. World Health Organization. Obesity and overweight: [www.who.int](http://www.who.int); 2021 [Accessed 23 July, 2021]. Available from: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
42. World Health Organization. Body mass index - BMI: [www.who.int](http://www.who.int); 2021 [Accessed 20 Aug 2021]. Available from: <https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi>.
43. World Cancer Research Fund. Body fatness and weight gain and the risk of cancer. World Cancer Research Fund International London; 2018.
44. Members EP, Jensen MD, Ryan DH, Donato KA, Apovian CM, Ard JD, et al. Executive summary: guidelines (2013) for the management of overweight and obesity in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Obesity Society published by the Obesity Society and American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Based on a systematic review from the The Obesity Expert Panel, 2013. *Obesity (Silver Spring)*. 2014;22(S2):S5-S39.
45. Roth G. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018. *The Lancet*. 2018;392:1736-1788.
46. Meyer HE, et.al. Overweight and obesity in Norway: Norwegian Institute of Public Health; 2017 [Accessed 23 July 2021]. Available from: <https://www.fhi.no/en/op/hin/health-disease/overweight-and-obesity-in-norway---/>.
47. Ortiz VE, Kwo J. Obesity: physiologic changes and implications for preoperative management. *BMC anesthesiology*. 2015;15:1-12.
48. Poddar M, Chetty Y, Chetty V. How does obesity affect the endocrine system? A narrative review. *Clinical obesity*. 2017;7(3):136-144.
49. Hurrle S, Hsu WH. The etiology of oxidative stress in insulin resistance. *Biomedical journal*. 2017;40(5):257-262.
50. Csige I, Ujvárosy D, Szabó Z, Lőrincz I, Paragh G, Harangi M, et al. The impact of obesity on the cardiovascular system. *J Diabetes Res*. 2018;2018.
51. Zammit C, Liddicoat H, Moonsie I, Makker H. Obesity and respiratory diseases. *Int J Gen Med*. 2010;335-343.
52. De Heredia FP, Gómez-Martínez S, Marcos A. Obesity, inflammation and the immune system. *Proceedings of the Nutrition Society*. 2012;71(2):332-338.
53. Anandacoomarasamy A, Caterson I, Sambrook P, Fransen M, March L. The impact of obesity on the musculoskeletal system. *Int J Obes (Lond)*. 2008;32(2):211-222.
54. Klenov VE, Jungheim ES. Obesity and reproductive function: a review of the evidence. *Current Opinion in Obstetrics and Gynecology*. 2014;26(6):455-460.

55. Wolin KY, Carson K, Colditz GA. Obesity and cancer. *The oncologist*. 2010;15(6):556-565.
56. da Silva M, Weiderpass E, Licaj I, Lissner L, Rylander C. Excess body weight, weight gain and obesity-related cancer risk in women in Norway: the Norwegian Women and Cancer study. *Br J Cancer*. 2018;119(5):646-656.
57. Ashburn DD, Reed MJ. Gastrointestinal system and obesity. *Critical care clinics*. 2010;26(4):625-627.
58. Del Gaudio A, Boschi L, Del Gaudio G-A, Mastrangelo L, Munari D. Liver damage in obese patients. *Obesity surgery*. 2002;12(6):802-804.
59. Wardle J, Cooke L. The impact of obesity on psychological well-being. *Best practice & research clinical endocrinology & metabolism*. 2005;19(3):421-440.
60. Kim K, Zakharkin SO, Allison DB. Expectations, validity, and reality in gene expression profiling. *Journal of clinical epidemiology*. 2010;63(9):950-959.
61. Ghosh S, Dent R, Harper ME, Gorman SA, Stuart JS, McPherson R. Gene expression profiling in whole blood identifies distinct biological pathways associated with obesity. *BMC Med Genomics*. 2010;3:56.
62. Homuth G, Wahl S, Muller C, Schurmann C, Mader U, Blankenberg S, et al. Extensive alterations of the whole-blood transcriptome are associated with body mass index: results of an mRNA profiling study involving two large population-based cohorts. *BMC Med Genomics*. 2015;8:65.
63. Vargas LB, Lange LA, Ferrier K, Aguet F, Ardlie K, Gabriel S, et al. Gene expression associations with body mass index in the Multi-Ethnic Study of Atherosclerosis. *Int J Obes (Lond)*. 2022:1-8.
64. Ronn T, Volkov P, Gillberg L, Kokosar M, Perfilyev A, Jacobsen AL, et al. Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Human Molecular Genetics*. 2015;24(13):3792-3813.
65. Bouchard L, Rabasa-Lhoret R, Faraj M, Lavoie M-È, Mill J, Pérusse L, et al. Differential epigenomic and transcriptomic responses in subcutaneous adipose tissue between low and high responders to caloric restriction. *The American journal of clinical nutrition*. 2010;91(2):309-320.
66. Capel F, Klimčáková E, Viguerie N, Roussel B, Vítková M, Kováčiková M, et al. Macrophages and adipocytes in human obesity: adipose tissue gene expression and insulin sensitivity during calorie restriction and weight stabilization. *Diabetes*. 2009;58(7):1558-1567.
67. Clément K, Viguerie N, Poitou C, Carette C, Pelloux V, Curat CA, et al. Weight loss regulates inflammation - related genes in white adipose tissue of obese subjects. *The FASEB Journal*. 2004;18(14):1657-1669.
68. Dahlman I, Linder K, Arvidsson Nordström E, Andersson I, Lidén J, Verdich C, et al. Changes in adipose tissue gene expression with energy-restricted diets in obese women. *The American journal of clinical nutrition*. 2005;81(6):1275-1285.
69. Johansson LE, Danielsson AP, Parikh H, Klintenberg M, Norström F, Groop L, et al. Differential gene expression in adipose tissue from obese human subjects during weight loss and weight maintenance. *The American journal of clinical nutrition*. 2012;96(1):196-207.
70. Kolehmainen M, Salopuro T, Schwab U, Kekäläinen J, Kallio P, Laaksonen D, et al. Weight reduction modulates expression of genes involved in extracellular matrix and cell death: the GENOBIN study. *Int J Obes (Lond)*. 2008;32(2):292-303.
71. Ma'ruquez-Quinones A, Mutch DM, Debarb C, Wang P, Combes M, Roussel B, et al. Adipose tissue transcriptome reflects variations between subjects with continued weight loss

- and subjects regaining weight 6 mo after caloric restriction independent of energy intake. *The American journal of clinical nutrition*. 2010;92(4):975-984.
72. Chisholm A. An Overview of Menopause: verywellhealth; 2022 [Accessed 10 August 2022]. Available from: <https://www.verywellhealth.com/menopause-4014690>.
  73. Hickey M, Elliott J, Davison SL. Hormone replacement therapy. *Bmj*. 2012;344.
  74. Singh A, Kaur S, Walia I. A historical perspective on menopause and menopausal age. *Bulletin of the Indian Institute of History of Medicine (Hyderabad)*. 2002;32(2):121-135.
  75. Cagnacci A, Venier M. The controversial history of hormone replacement therapy. *Medicina (B Aires)*. 2019;55(9):602.
  76. Talaulikar V. Menopause transition: Physiology and symptoms. *Best practice & research Clinical obstetrics & gynaecology*. 2022.
  77. O'Neill S, Eden J. The pathophysiology of menopausal symptoms. *Obstetrics, Gynaecology & Reproductive Medicine*. 2017;27(10):303-310.
  78. Palacios S, Henderson V, Siseles N, Tan D, Villaseca P. Age of menopause and impact of climacteric symptoms by geographical region. *Climacteric*. 2010;13(5):419-428.
  79. Nappi RE, Kroll R, Siddiqui E, Stoykova B, Rea C, Gemmen E, et al. Global cross-sectional survey of women with vasomotor symptoms associated with menopause: prevalence and quality of life burden. *Menopause (New York, NY)*. 2021;28(8):875.
  80. Hamoda H, Panay N, Pedder H, Arya R, Savvas M. The British Menopause Society & Women's Health Concern 2020 recommendations on hormone replacement therapy in menopausal women. *Post reproductive health*. 2020;26(4):181-209.
  81. Vivian-Taylor J, Hickey M. Menopause and depression: is there a link? *Maturitas*. 2014;79(2):142-146.
  82. Ameye L, Antoine C, Paesmans M, de Azambuja E, Rozenberg S. Menopausal hormone therapy use in 17 European countries during the last decade. *Maturitas*. 2014;79(3):287-291.
  83. Johansen LL, Thinggaard M, Hallas J, Osler M, Christensen K. Postmenopausal hormone therapy and mortality before and after the Women's Health Initiative study. *Sci*. 2023;13(1):539.
  84. Dvornyk V, Liu Y, Lu Y, Shen H, Lappe JM, Lei S, et al. Effect of menopause on gene expression profiles of circulating monocytes: a pilot in vivo microarray study. *J Genet Genomics*. 2007;34(11):974-983.
  85. Gameiro CM, Romão F, Castelo-Branco C. Menopause and aging: changes in the immune system—a review. *Maturitas*. 2010;67(4):316-320.
  86. Abildgaard J, Tingstedt J, Zhao Y, Hartling HJ, Pedersen AT, Lindegaard B, et al. Increased systemic inflammation and altered distribution of T-cell subsets in postmenopausal women. *PLoS One*. 2020;15(6):e0235174.
  87. Santoro N, Roeca C, Peters BA, Neal-Perry G. The menopause transition: signs, symptoms, and management options. *The Journal of Clinical Endocrinology & Metabolism*. 2021;106(1):1-15.
  88. Scavello I, Maseroli E, Di Stasi V, Vignozzi L. Sexual health in menopause. *Medicina (B Aires)*. 2019;55(9):559.
  89. Surakasula A, Nagarjunapu GC, Raghavaiah K. A comparative study of pre-and post-menopausal breast cancer: Risk factors, presentation, characteristics and management. *Journal of research in pharmacy practice*. 2014;3(1):12.
  90. Dahm AE, Eilertsen AL, Goeman J, Olstad OK, Ovstebo R, Kierulf P, et al. A microarray study on the effect of four hormone therapy regimens on gene transcription in whole blood from healthy postmenopausal women. *Thromb Res*. 2012;130(1):45-51.

91. Waaseth M, Olsen KS, Rylander C, Lund E, Dumeaux V. Sex hormones and gene expression signatures in peripheral blood from postmenopausal women-the NOWAC postgenome study. *BMC medical genomics*. 2011;4(1):29.
92. Hall P, Ploner A, Bjöhle J, Huang F, Lin C-Y, Liu ET, et al. Hormone-replacement therapy influences gene expression profiles and is associated with breast-cancer prognosis: a cohort study. *BMC Med*. 2006;4(1):1-13.
93. Dumeaux V, Børresen-Dale A-L, Frantzen J-O, Kumle M, Kristensen VN, & Lund E. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast cancer research*. 2008;10(1):R13.
94. National Human Genome Research Institute. Deoxyribonucleic Acid (DNA) Fact Sheet: [www.nih.gov](http://www.nih.gov); 2020 [Accessed 10 April, 2023]. Available from: <https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet>.
95. National Human Genome Research Institute. Gene: [www.nih.gov](http://www.nih.gov); 2023 [Accessed 10 April, 2023]. Available from: <https://www.genome.gov/genetics-glossary/Gene>.
96. National Human Genome Research Institute. Gene Expression: [www.nih.gov](http://www.nih.gov); 2023 [Accessed 10 April, 2023]. Available from: <https://www.genome.gov/genetics-glossary/Gene-Expression>.
97. National Human Genome Research Institute. Central Dogma: [www.nih.gov](http://www.nih.gov); 2023 [Accessed 15 April, 2023]. Available from: <https://www.genome.gov/genetics-glossary/Central-Dogma>.
98. National Human Genome Research Institute. Transcriptome Fact Sheet: [www.nih.gov](http://www.nih.gov); 2020 [Accessed 2 May, 2023]. Available from: <https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet>.
99. Pascual-Ahuir A, Fita-Torró J, Proft M. Capturing and understanding the dynamics and heterogeneity of gene expression in the living cell. *Int*. 2020;21(21):8278.
100. Singh KP, Miaskowski C, Dhruva AA, Flowers E, Kober KM. Mechanisms and measurement of changes in gene expression. *Biol Res Nurs*. 2018;20(4):369-382.
101. San Segundo-Val I, Sanz-Lozano CS. Introduction to the gene expression analysis. *Molecular genetics of asthma*. 2016:29-43.
102. McHale CM, Zhang L, Thomas R, Smith MT. Analysis of the transcriptome in molecular epidemiology studies. *Environmental and molecular mutagenesis*. 2013;54(7):500-517.
103. Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. Microarray and its applications. *Journal of pharmacy & bioallied sciences*. 2012;4(Suppl 2):S310.
104. Little L. Choosing single-cell and spatial analysis technologies.: <https://frontlinegenomics.com/>; 2023 [Accessed 25 May, 2023]. Available from: <https://frontlinegenomics.com/choosing-single-cell-and-spatial-analysis-technologies/>.
105. Little L. A guide to single-cell sequencing and spatial analysis.: <https://frontlinegenomics.com/>; 2023 [Accessed 25 May, 2023]. Available from: <https://frontlinegenomics.com/a-guide-to-single-cell-and-spatial-analysis/>.
106. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single - cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*. 2022;12(3):e694.
107. Xiaowei A. Method of the Year 2020: Spatially resolved transcriptomics. *Nat Methods*. 2021;18(1).
108. Vu T, Vallmitjana A, Gu J, La K, Xu Q, Flores J, et al. Spatial transcriptomics using combinatorial fluorescence spectral and lifetime encoding, imaging and analysis. *Nat Commun*. 2022;13(1):169.



109. Goytain A, Ng T. NanoString nCounter technology: high-throughput RNA validation. *Chimeric RNA: methods and protocols*. 2020:125-139.
110. Mohr S, Liew C-C. The peripheral-blood transcriptome: new insights into disease and risk assessment. *Trends in molecular medicine*. 2007;13(10):422-432.
111. Dumeaux V, Fjukstad B, Fjosne HE, Frantzen J-O, Holmen MM, Rodegerdts E, et al. Interactions between the tumor and the blood systemic response of breast cancer patients. *PLoS computational biology*. 2017;13(9):e1005680.
112. Kumar SD, Kar D, Akhtar MN, Willard B, Roy D, Hussain T, et al. Evidence for low-level translation in human erythrocytes. *Molecular biology of the cell*. 2022;33(12):br21.
113. Chen S-Y, Wang Y, Telen MJ, Chi J-T. The genomic analysis of erythrocyte microRNA expression in sickle cell diseases. *PloS one*. 2008;3(6):e2360.
114. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. *Science*. 2015;348(6235):660-665.
115. Olsen KS, Skeie G, Lund E. Whole-blood gene expression profiles in large-scale epidemiological studies: what do they tell? *Current Nutrition Reports*. 2015;4:377-386.
116. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, et al. Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences*. 2003;100(4):1896-1901.
117. Bollinger T, Schibler U. Circadian rhythms—from genes to physiology and disease. *Swiss medical weekly*. 2014;144(2930):w13984-w13984.
118. Manella G, Ezagouri S, Champigneulle B, Gaucher J, Mendelson M, Lemarie E, et al. The human blood transcriptome exhibits time-of-day-dependent response to hypoxia: Lessons from the highest city in the world. *Cell Reports*. 2022;40(7):111213.
119. Baiju N, Sandanger TM, Sætrum P, Nøst TH. Gene expression in blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) postgenome cohort. *Sci*. 2021;11(1):1-13.
120. Olsen KS, Lukic M, Borch KB. Physical activity and blood gene expression profiles: the Norwegian Women and Cancer (NOWAC) Post-genome cohort. *BMC Res Notes*. 2020;13(1):1-6.
121. Waaseth M, Bakken K, Dumeaux V, Olsen KS, Rylander C, Figenschau Y, et al. Hormone replacement therapy use and plasma levels of sex hormones in the Norwegian Women and Cancer Postgenome Cohort—a cross-sectional analysis. *BMC women's health*. 2008;8(1):1-11.
122. Lund E, Holden L, Bøvelstad H, Plancade S, Mode N, Günther C-C, et al. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC postgenome cohort as a proof of principle. *BMC medical research methodology*. 2016;16(1):28.
123. Bøvelstad HM, Holsbø E, Bongo LA, Lund E. A standard operating procedure for outlier removal in large-sample epidemiological transcriptomics datasets. *BioRxiv*. 2017:144519.
124. Du P, Feng G, Kibbe W, & Lin S. lumiHumanIDMapping: Illumina Identifier Mapping for Human. *R package version*. 2016;1(1).
125. Du P, Kibbe WA, & Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*. 2008;24(13):1547-1548.
126. Dunning M, Lynch A, & Eldridge M. illuminaHumanv4. db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4). *R package version*. 2015;1(0).
127. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*. 2015;12(5):453-457.

128. National Cancer Institute. Pack Years: [Accessed December 1, 2019]. NCI Dictionary of Cancer Terms]. Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pack-year>.
129. R Core Team. R: A language and environment for statistical computing. 2020.
130. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*. 2015;43(7):e47-e47.
131. Benjamini Y, & Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995;57(1):289-300.
132. Sandanger TM, Nost TH, Guida F, Rylander C, Campanella G, Muller DC, et al. DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian Women and Cancer cohort. *Sci*. 2018;8(1):16714.
133. Joehanes R, Johnson AD, Barb JJ, Raghavachari N, Liu P, Woodhouse KA, et al. Gene expression analysis of whole blood, peripheral blood mononuclear cells, and lymphoblastoid cell lines from the Framingham Heart Study. *Physiol Genomics*. 2012;44(1):59-75.
134. Goeman JJ, Van De Geer SA, De Kort F, & Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004;20(1):93-99.
135. Yu G, Wang L-G, Han Y, & He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*. 2012;16(5):284-287.
136. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*. 2021;2(3):100141.
137. Yu G, & He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*. 2016;12(2):477-479.
138. Paul S, & Amundson SA. Differential effect of active smoking on gene expression in male and female smokers. *Journal of carcinogenesis & mutagenesis*. 2014;5.
139. 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.
140. Obeidat M, Ding X, Fishbane N, Hollander Z, Ng RT, McManus B, et al. The Effect of Different Case Definitions of Current Smoking on the Discovery of Smoking-Related Blood Gene Expression Signatures in Chronic Obstructive Pulmonary Disease. *Nicotine Tob Res*. 2016;18(9):1903-1909.
141. Maas SC, Mens MM, Kühnel B, van Meurs JB, Uitterlinden AG, Peters A, et al. Smoking-related changes in DNA methylation and gene expression are associated with cardio-metabolic traits. *Clin Epigenetics*. 2020;12:1-16.
142. Nowak J, Dybska E, Adams A, Walkowiak J. Immune cell-specific smoking-related expression characteristics are revealed by re-analysis of transcriptomes from the CEDAR cohort. *Central European Journal of Immunology*. 2022;47(3):246-259.
143. Guida F, Sandanger TM, Castagné R, Campanella G, Polidoro S, Palli D, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human molecular genetics*. 2015;24(8):2349-2359.
144. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. *Circulation: cardiovascular genetics*. 2016;9(5):436-447.
145. Tantoh DM, Wu M-C, Chuang C-C, Chen P-H, Tyan YS, Nfor ON, et al. AHRR cg05575921 methylation in relation to smoking and PM 2.5 exposure among Taiwanese men and women. *Clin Epigenetics*. 2020;12:1-9.

146. Grieshober L, Graw S, Barnett MJ, Thornquist MD, Goodman GE, Chen C, et al. AHRR methylation in heavy smokers: associations with smoking, lung cancer risk, and lung cancer mortality. *BMC Cancer*. 2020;20:1-10.
147. The Human Protein Atlas. Explore genes/proteins: [proteinatlas.org](https://www.proteinatlas.org/); 2023 [Accessed May 16, 2023]. Available from: <https://www.proteinatlas.org/>.
148. Lee C-L, Chang W-D. The effects of cigarette smoking on aerobic and anaerobic capacity and heart rate variability among female university students. *International journal of women's health*. 2013;667-679.
149. Erhardt L. Cigarette smoking: an undertreated risk factor for cardiovascular disease. *Atherosclerosis*. 2009;205(1):23-32.
150. Vivier E, Tomasello E, Baratin M, Walzer T, Ugolini S. Functions of natural killer cells. *Nat Immunol*. 2008;9(5):503-510.
151. Higuchi T, Omata F, Tsuchihashi K, Higashioka K, Koyamada R, & Okada S. Current cigarette smoking is a reversible cause of elevated white blood cell count: cross-sectional and longitudinal studies. *Preventive medicine reports*. 2016;4:417-422.
152. Malenica M, Prnjavorac B, Bego T, Dujic T, Semiz S, Skrbo S, et al. Effect of cigarette smoking on haematological parameters in healthy population. *Medical Archives*. 2017;71(2):132.
153. UniProtKB/Swiss-Prot. Find Your Protein: [UniProt.org](https://www.uniprot.org/); 2023 [Accessed Feb 23, 2023]. Available from: <https://www.uniprot.org/>.
154. GeneCards. Explore a gene: [genecards.org](https://www.genecards.org/); 2023 [Accessed February 16, 2023]. Available from: <https://www.genecards.org/>.
155. Bird JK, Ronnenberg AG, Choi S-W, Du F, Mason JB, Liu Z. Obesity is associated with increased red blood cell folate despite lower dietary intakes and serum concentrations. *The Journal of nutrition*. 2015;145(1):79-86.
156. Kohsari M, Moradinazar M, Rahimi Z, Najafi F, Pasdar Y, Moradi A, et al. Association between RBC indices, anemia, and obesity-related diseases affected by body mass index in Iranian Kurdish population: Results from a cohort study in Western Iran. *International Journal of Endocrinology*. 2021;2021.
157. Umemoto Y, Tsuji K, Yang F-C, Ebihara Y, Kaneko A, Furukawa S, et al. Leptin stimulates the proliferation of murine myelocytic and primitive hematopoietic progenitor cells. *Blood, The Journal of the American Society of Hematology*. 1997;90(9):3438-3443.
158. Trottier MD, Naaz A, Kacynski K, Yenumula PR, Fraker PJ. Functional capacity of neutrophils from class III obese patients. *Obesity (Silver Spring)*. 2012;20(5):1057-1065.
159. QuickGO. GO annotations: EMBL-EBI; 2023 [Accessed February 16, 2023]. Available from: <https://www.ebi.ac.uk/QuickGO/annotations>.
160. Goyal R, Singhai M, Faizy AF. Glutathione peroxidase activity in obese and nonobese diabetic patients and role of hyperglycemia in oxidative stress. *Journal of mid-life health*. 2011;2(2):72.
161. Akter R, Nessa A, Sarker D, Yesmin M. Effect of Obesity on Hemoglobin Concentration. *Mymensingh Medical Journal: MMJ*. 2017;26(2):230-234.
162. Gozkaman A, Okuturlar Y, Mert M, Harmankaya O, Kumbasar A, editors. The relationship between haemoglobin and BMI in overweight and obese patients. *Endocrine Abstracts*; 2015: Bioscientifica.
163. Rui L. A link between protein translation and body weight. *The Journal of clinical investigation*. 2007;117(2):310-313.
164. National Human Genome Research Institute. Ribosome: NIH; 2021 [Accessed February 10, 2022]. Available from: <https://www.genome.gov/genetics-glossary/Ribosome>.

165. Jung UJ, Seo YR, Ryu R, Choi MS. Differences in metabolic biomarkers in the blood and gene expression profiles of peripheral blood mononuclear cells among normal weight, mildly obese and moderately obese subjects. *Br J Nutr.* 2016;116(6):1022-1032.
166. De Frel DL, Atsma DE, Pijl H, Seidell JC, Leenen PJ, Dik WA, et al. The impact of obesity and lifestyle on the immune system and susceptibility to infections such as COVID-19. *Frontiers in nutrition.* 2020:279.
167. Honce R, Schultz-Cherry S. Impact of obesity on influenza A virus pathogenesis, immune response, and evolution. *Front.* 2019:1071.
168. Mandal A. Obesity could increase severity of influenza study shows: NEWS-Medical Life Sciences; 2020 [Accessed February 8, 2022]. Available from: <https://www.news-medical.net/news/20200303/Obesity-could-increase-severity-of-influenza-study-shows.aspx>.
169. Ghannadiasl F. Associations between white blood cells count and obesity in apparently healthy young adults. *Nutrition & Food Science.* 2020.
170. Dumeaux V, Johansen J, Børresen-Dale A-L, Lund E. Gene expression profiling of whole-blood samples from women exposed to hormone replacement therapy. *Molecular cancer therapeutics.* 2006;5(4):868-876.
171. Zierau O, Zenclussen AC, Jensen F. Role of female sex hormones, estradiol and progesterone, in mast cell behavior. *Front.* 2012;3:169.
172. Chen Y, Zhang Y, Zhao G, Chen C, Yang P, Ye S, et al. Difference in leukocyte composition between women before and after menopausal age, and distinct sexual dimorphism. *PLoS One.* 2016;11(9):e0162953.
173. Keith SW, Redden DT, Katzmarzyk PT, Boggiano MM, Hanlon EC, Benca RM, et al. Putative contributors to the secular increase in obesity: exploring the roads less traveled. *Int J Obes (Lond).* 2006;30(11):1585-1594.
174. Symonds ME, Budge H, Frazier-Wood AC. Epigenetics and obesity: a relationship waiting to be explained. *Human Heredity.* 2013;75(2-4):90-97.
175. Saavedra JM. Obesity—a risk factor or a disease: What can exercise do for obese children? *The Indian journal of medical research.* 2014;139(5):661.
176. Tigner A, Ibrahim SA, Murray I. Histology, white blood cell: In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK563148/>.
177. Moses K, Brandau S, editors. Human neutrophils: their role in cancer and relation to myeloid-derived suppressor cells. *Seminars in immunology*; 2016: Elsevier.
178. Treffers LW, Hiemstra IH, Kuijpers TW, Van den Berg TK, Matlung HL. Neutrophils in cancer. *Immunological reviews.* 2016;273(1):312-328.
179. Jareid M, Snapkov I, Holden M, Busund L-TR, Lund E, Nøst TH. The blood transcriptome prior to ovarian cancer diagnosis: A case-control study in the NOWAC postgenome cohort. *Plos one.* 2021;16(8):e0256442.
180. Nøst TH, Holden M, Dønnem T, Bøvelstad H, Rylander C, Lund E, et al. Transcriptomic signals in blood prior to lung cancer focusing on time to diagnosis and metastasis. *Sci.* 2021;11(1):7406.
181. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun.* 2020;11(1):5650.
182. Karlovich C, Duchateau-Nguyen G, Johnson A, McLoughlin P, Navarro M, Fleurbaey C, et al. A longitudinal study of gene expression in healthy individuals. *BMC medical genomics.* 2009;2(1):1-16.
183. Garnis C, Buys TP, Lam WL. Genetic alteration and gene expression modulation during cancer progression. *Mol Cancer.* 2004;3(1):1-23.

184. Livesey M, Rossouw SC, Blignaut R, Christoffels A, Bendou H. Transforming RNA-Seq gene expression to track cancer progression in the multi-stage early to advanced-stage cancer development. *Plos one*. 2023;18(4):e0284458.
185. Brennan P, Wild CP. Genomics of cancer and a new era for cancer prevention. *PLoS genetics*. 2015;11(11):e1005522.
186. B Barnung R, H Nøst T, Ulven S, Skeie G, S Olsen K. Coffee consumption and whole-blood gene expression in the norwegian women and cancer post-genome cohort. *Nutrients*. 2018;10(8):1047.
187. Lund E. *Advancing Systems Epidemiology in Cancer: Exploring Trajectories of Gene Expression*: Scandinavian University Press; 2020.
188. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. *Chest*. 2020;158(1):S65-S71.
189. Henderson M, Page L. Appraising the evidence: what is selection bias? *BMJ Ment Health*. 2007;10(3):67-68.
190. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*. 2016:211-217.
191. Porta M. *A dictionary of epidemiology* (6th ed.). New York: Oxford University Press; 2014.
192. Eiliv L, Merethe K, Tonje B, Anette H, Kjersti B, Elise E, et al. External validity in a population-based national prospective study—the Norwegian Women and Cancer Study (NOWAC). *Cancer Causes Control*. 2003;14:1001-1008.
193. Szklo M, Nieto FJ. *Epidemiology: beyond the basics*: Jones & Bartlett Publishers; 2014.
194. Volk RJ, Mendoza TR, Hoover DS, Nishi SP, Choi NJ, Bevers TB. Reliability of self-reported smoking history and its implications for lung cancer screening. *Preventive Medicine Reports*. 2020;17:101037.
195. Skeie G, Mode N, Henningsen M, Borch KB. Validity of self-reported body mass index among middle-aged participants in the Norwegian Women and Cancer study. *Clinical epidemiology*. 2015;7:313.
196. Jarvis MJ, Russell M, Benowitz NL, & Feyerabend C. Elimination of cotinine from body fluids: implications for noninvasive measurement of tobacco smoke exposure. *Am J Public Health*. 1988;78(6):696-698.
197. Vetter TR, Mascha EJ. Bias, confounding, and interaction: lions and tigers, and bears, oh my! *Anesthesia & Analgesia*. 2017;125(3):1042-1048.
198. Pham A, Cummings M, Lindeman C, Drummond N, Williamson T. Recognizing misclassification bias in research and medical practice. *Family practice*. 2019;36(6):804-807.
199. Yap S, Vassallo A, Goldsbury DE, Salagame U, Velentzis L, Banks E, et al. Accurate categorisation of menopausal status for research studies: a step-by-step guide and detailed algorithm considering age, self-reported menopause and factors potentially masking the occurrence of menopause. *BMC Res Notes*. 2022;15(1):1-7.
200. Skogholt AH, Ryeng E, Erlandsen SE, Skorpen F, Schønberg SA, Sætrum P. Gene expression differences between PAXgene and Tempus blood RNA tubes are highly reproducible between independent samples and biobanks. *BMC Res Notes*. 2017;10(1):1-12.
201. 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):81-86.
202. Fan H, Hegde PS. The transcriptome in blood: challenges and solutions for robust expression profiling. *Current molecular medicine*. 2005;5(1):3-10.

203. 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):882-883.
204. Meyer A, Paroni F, Günther K, Dharmadhikari G, Ahrens W, Kelm S, et al. Evaluation of existing methods for human blood mRNA isolation and analysis for large studies. *PLoS One*. 2016;11(8):e0161778.
205. Cabuzu D. How PAXgene® and Tempus™ blood tubes influence RNA expression signatures: ALITHEA Genomics; 2023 [Accessed 10 May, 2023]. Available from: <https://alitheagenomics.com/blog/how-paxgene-and-tempus-blood-tubes-influence-rna-expression-signatures>.
206. Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Medical science monitor basic research*. 2014;20:138.
207. illumina. Transitioning from Microarrays to mRNA-Seq2011 [cited 2023 2 June]. Available from: [https://www.illumina.com/content/dam/illumina-marketing/documents/icomunity/article\\_2011\\_12\\_ea\\_rna-seq.pdf](https://www.illumina.com/content/dam/illumina-marketing/documents/icomunity/article_2011_12_ea_rna-seq.pdf).
208. Ambroise J, Bearzatto B, Robert A, Govaerts B, Macq B, Gala J-L. Impact of the spotted microarray preprocessing method on fold-change compression and variance stability. *BMC Bioinformatics*. 2011;12:1-12.
209. Zhao Y, Wong L, Goh WWB. How to do quantile normalization correctly for gene expression data analyses. *Sci*. 2020;10(1):1-11.
210. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127.
211. Jeanmougin M, De Reynies A, Marisa L, Paccard C, Nuel G, Guedj M. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PloS one*. 2010;5(9):e12336.
212. Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. limma: linear models for microarray and RNA-Seq data user's guide. *Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia*. 2002.
213. Lydersen S. Adjustment of p-values for multiple hypotheses. *Tidsskrift for Den norske legeförening*. 2021.
214. DeRisi J, Penland L, Bittner M, Meltzer P, Ray M, Chen Y, et al. Use of a cDNA microarray to analyse gene expression. *Nat genet*. 1996;14:457-460.
215. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences*. 1996;93(20):10614-10619.
216. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*. 2009;25(6):765-771.
217. Laurent GS, Shtokalo D, Tackett MR, Yang Z, Vyatkin Y, Milos PM, et al. On the importance of small changes in RNA expression. *Methods*. 2013;63(1):18-24.
218. Khatri P, Sirota M, & Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*. 2012;8(2):e1002375.

# Errata

## Paper I

The labels in x-axis (Observed p-values) and y-axis (Expected p-values) should have been switched to x-axis (Expected p-values) and y-axis (Observed p-values) in the Supplementary Figure S4 (Q-Q plot).

There should have been “Complementary DNA c(DNA).....” instead of “Complementary RNA c(RNA).....” in the Method section under Laboratory analyses and pre-processing of the gene expression data.





## **Appendices**

1. Information letter (NOWAC)
2. Questionnaire at the blood collection time point (NOWAC 2003)





# KVINNER OG KREFT

Du sendte våren 2002 et utfylt spørreskjema til Institutt for samfunnsmedisin som del av den landsdekkende undersøkelsen "Kvinner og kreft". Spørsmålene var særlig rettet mot kosthold. Vi ønsker å studere hvilken betydning våre matvaner har for kreftutvikling hos kvinner. I følgeskrivet til spørreskjemaet informerte vi om at en del kvinner senere ville bli forespurt om de var villig til å avgi blodprøve. Blodprøvene vil bli aidentifisert ved ankomst Institutt for samfunnsmedisin.

Formålet med blodprøven vil være:

- Måle nivå av vitaminer, mineraler og andre stoffer i blodet som kan settes i forbindelse med kostholdet.
- I fremtiden kunne studere de såkalte genetiske markører dvs. egenskaper i arvestoffet som kan disponere for kreft.
- Teste nye ideer eller hypoteser som oppstår i fremtiden.


Det er frivillig om du vil delta. Du kan trekke deg uten begrunnelse, og du kan be om at opplysninger du har gitt blir slettet, uten at dette vil få konsekvenser for deg. Blodprøven vil kun bli benyttet til forskning og ingen resultater vil bli utlevert til deg eller noen andre. Blodprøven vil bli lagret i 30 år.

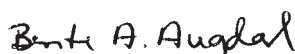
Ansvarlig for undersøkelsen er professor Eiliv Lund. Undersøkelsen er tilrådd av Regional komité for medisinsk forskningsetikk, Nord-Norge (REK NORD), og Datatilsynet har gitt konsesjon for oppbevaring av opplysninger.

Fremtidige forskningsprosjekter som vil benytte de lagrete blodprøvene vil forelegges Regional komité for medisinsk forskningsetikk, Nord-Norge (REK NORD).

Du kan finne mer informasjon om "Kvinner og kreft" og om forskningsresultatene på våre nettsider: [www.ism.uit.no/kk/](http://www.ism.uit.no/kk/)

Med vennlig hilsen

  
Eiliv Lund  
professor dr.med.

  
Bente A. Augdal  
prosjektmedarbeider



Ønsker du ikke å delta og vil slippe påminning pr. brev ber vi deg fyller ut svar-slippen og returnere denne sammen med utstyret tilbake til oss (forseglet utstyr **må** ikke åpnes).

Jeg ønsker **ikke** å delta i blodprøvetakingen. ....

Underskrift



# KVINNER OG KREFT

Følgende opplysninger fylles ut i forbindelse med blodprøvetaking.

DETTE SKJEMA **MÅ** FØLGE BLODPRØVEN!

Skjemaet skal leses optisk. Vennligst bruk blå eller sort penn. Du kan ikke bruke komma, bruk blokkbokstaver.

ID-nr:

LAB-kobling.

Jeg har lest informasjonen om blodprøveundersøkelsen og samtykker i å delta i denne:

Ja:

## PRØVETAKINGSDAGEN

Fyll inn tidspunkt når blodprøven er tatt: Dato:     dag mnd

Klokkeslett:

+

## MENSTRUASJONSFORHOLD

Har du menstruasjon?

Ja .....

Nei .....

Uregelmessig .....

Er gravid .....

*Hvis ja:*  
Angi dato for første dag i siste menstruasjon:     dag mnd

## MATINNTAK

Når spiste du siste måltid før blodprøven ble tatt: Dato:     dag mnd

Klokkeslett:

+

## RØYKEVANER SISTE UKEN

Har du røkt i løpet av siste uke?

Ja .....

Nei .....

*Hvis ja:* Hvor mange sigaretter røkte du?

Antall i går:

Antall i dag:

+

## VEKT OG HØYDE

Hvor mye veier du i dag? kg

Hvor høy er du? cm

Er disse målene tatt på legekantoret i dag?

Ja .....

Nei .....

+

## MEDISINER I LØPET AV SISTE UKE

Har du brukt P-piller i løpet av siste uke?

Ja .....   
Nei .....

*Hvis ja:*

Angi dato for siste tablett

dag		mnd	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Preparatnavn: .....  
  
(ikke skriv her)

Har du i løpet av siste uke brukt hormontabletter (østrogen, gestagen) for overgangsalderen?

Ja .....   
Nei .....

*Hvis ja:*

Angi dato for siste tablett

dag		mnd	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Preparatnavn: .....  
  
(ikke skriv her)

Preparatnavn: .....  
  
(ikke skriv her)

Preparatnavn: .....  
  
(ikke skriv her)



Har du brukt andre medisiner i løpet av siste uke?

Ja .....   
Nei .....

*Hvis ja:*

Angi dato for siste tablett

dag		mnd	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Preparatnavn: .....  
  
(ikke skriv her)

Preparat navn: .....  
  
(ikke skriv her)

Preparat navn: .....  
  
(ikke skriv her)

## BRUK AV KOSTTILSKUDD I LØPET AV SISTE UKE

Har du brukt tran (flytende) i løpet av siste uke?

Ja .....   
Nei .....

*Hvis ja:*

Angi dato du sist tok tran

dag		mnd	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Hvor mye tran tok du da?

1 ts     1/2 ss     1+ ss

Har du brukt trankapsler i løpet av siste uke?

Ja .....   
Nei .....

*Hvis ja:*

Angi dato du sist tok trankapsel

dag		mnd	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Hvor mange trankapsler tok du da?

1     2     3+

Navn på trankapselpreparatet du tok sist:

.....  
  
(ikke skriv her)

Har du brukt andre kosttilskudd (vitaminer/mineraler) i løpet av siste uke?

Ja .....   
Nei .....

*Hvis ja:*

Angi dato for siste tablett

dag		mnd	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Preparatnavn: .....  
  
(ikke skriv her)

Preparatnavn: .....  
  
(ikke skriv her)

Preparatnavn: .....  
  
(ikke skriv her)

## STILLING NÅR BLODPRØVEN BLE TATT

Sittende .....   
Liggende .....

Takk for hjelpen!

## **Paper I**

**Gene expression in whole-blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) Post-genome cohort.**

Baiju N, Sandanger TM, Sætrom P, Nøst TH.

*Scientific Reports.* 2021;11(1):1-13.







OPEN

# Gene expression in blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) postgenome cohort

Nikita Baiju<sup>1</sup>✉, Torkjel M. Sandanger<sup>1</sup>, Pål Sætrom<sup>2,3,4,5</sup> & Therese H. Nøst<sup>1,5</sup>

Active smoking has been linked to modulated gene expression in blood. However, there is a need for a more thorough understanding of how quantitative measures of smoking exposure relate to differentially expressed genes (DEGs) in whole-blood among ever smokers. This study analysed microarray-based gene expression profiles from whole-blood samples according to smoking status and quantitative measures of smoking exposure among cancer-free women (n = 1708) in the Norwegian Women and Cancer postgenome cohort. When compared with never smokers and former smokers, current smokers had 911 and 1082 DEGs, respectively and their biological functions could indicate systemic impacts of smoking. *LRRN3* was associated with smoking status with the lowest FDR-adjusted p-value. When never smokers and all former smokers were compared, no DEGs were observed, but *LRRN3* was differentially expressed when never smokers were compared with former smokers who quit smoking  $\leq 10$  years ago. Further, *LRRN3* was positively associated with smoking intensity, pack-years, and comprehensive smoking index score among current smokers; and negatively associated with time since cessation among former smokers. Consequently, *LRRN3* expression in whole-blood is a molecular signal of smoking exposure that could supplant self-reported smoking data in further research targeting blood-based markers related to the health effects of smoking.

Tobacco smoking is one of the major threats to public health, and it is currently responsible for more than 8 million deaths worldwide each year<sup>1</sup>. Exposure to tobacco smoke is a risk factor for many chronic diseases, such as cardiac and pulmonary diseases and several cancers. Further, smoking can suppress the immune system and modifies a range of immunological functions<sup>2</sup>. Subclinical outcomes, such as increased oxidative stress, reduced antioxidant defences, increased inflammation, impaired immune status, and altered lipid profiles, have been observed in smokers when compared to their counterparts who never smoked<sup>3</sup>. Notably, more respiratory symptoms caused by exposure to tobacco smoke have been observed in women than men<sup>4,5</sup>. Thus, tobacco smoking has several detrimental health effects, which might appear not long after smoking initiation or up to several decades after exposure<sup>3,6</sup>.

The toxic components of tobacco smoke are first absorbed in the lungs and then enter the blood stream before being distributed throughout the body, making blood an appropriate biological material to study the systemic influences of exposure to tobacco smoke<sup>7</sup>. In addition, the collection of whole-blood (or simply, 'blood

<sup>1</sup>Department of Community Medicine, Faculty of Health Sciences, UiT –the Arctic University of Norway, 9037 Tromsø, Norway. <sup>2</sup>Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway. <sup>3</sup>Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway. <sup>4</sup>Bioinformatics Core Facility, Norwegian University of Science and Technology, 7491 Trondheim, Norway. <sup>5</sup>Department of Public Health and Nursing, K.G. Jebsen Center for Genetic Epidemiology, Norwegian University of Science and Technology, 7491 Trondheim, Norway. ✉email: nikita.baiju@uit.no

samples') is easy and minimally invasive, and these samples can reveal features that are relevant for studies of human health effects<sup>8</sup>. Current exposure to tobacco smoke has been linked with modulated expression of many genes in blood, for example *LRRN3*, *CLDN1*, *GPR15*, *ATF4*, *SOD2*, and *CDKN1C*<sup>9–16</sup>. Altered gene expression in blood has also been linked to diseases for which smoking is a risk factor<sup>17</sup>. However, there is a need for a more thorough understanding of the variability in gene expression profiles in whole-blood in relation to quantitative measures of smoking exposure among ever smokers. Therefore, this cross-sectional analysis used data from 1708 cancer-free women participating in the prospective, population-based Norwegian Women and Cancer (NOWAC) postgenome cohort. Microarray-based gene expression profiles from bio-banked whole-blood samples were assessed according to smoking status and quantitative measures of smoking exposure (hereafter referred to as 'smoking metrics'), such as smoking intensity, smoking duration, time since smoking cessation (TSC), pack-years, and comprehensive smoking index (CSI) scores<sup>18</sup>. Enriched pathways and gene ontology (GO) categories of significant genes associated with smoking were also assessed.

## Results

**General characteristics of the study population.** The current study was based on microarray data from cancer-free women participating in the NOWAC postgenome cohort. The full cohort consists of approximately 50,000 women (mean age: 49.78 years; mean body mass index (BMI): 23.38 kg/m<sup>2</sup>), all of whom have given a blood sample. In total, 1708 of these women have been included as cancer-free controls in various studies and have gene expression profiles available for study, and only these women were included in the present analyses. All included women had completed up to three comprehensive questionnaires before blood collection (main questionnaires), and an additional questionnaire on lifestyle factors was completed at the time of blood collection. Thus, information was available for up to four time points in total. Smoking status and smoking metrics (smoking intensity, smoking duration, TSC, pack-years, and CSI scores) were based on information from all four questionnaires. Current smokers (CS) were defined as those who were currently smoking at the time of blood collection, former smokers (FS) were defined as those who reported smoking cessation prior to the time of blood collection, and never smokers (NS) were defined as those who reported they had never smoked either prior to or at the time of blood collection. CS and FS combined represented ever smokers. We defined passive smokers (PS) as those who were passively exposed to smoking at their homes as adults. Gene expression values were available for 7713 unique genes for all the women in this study.

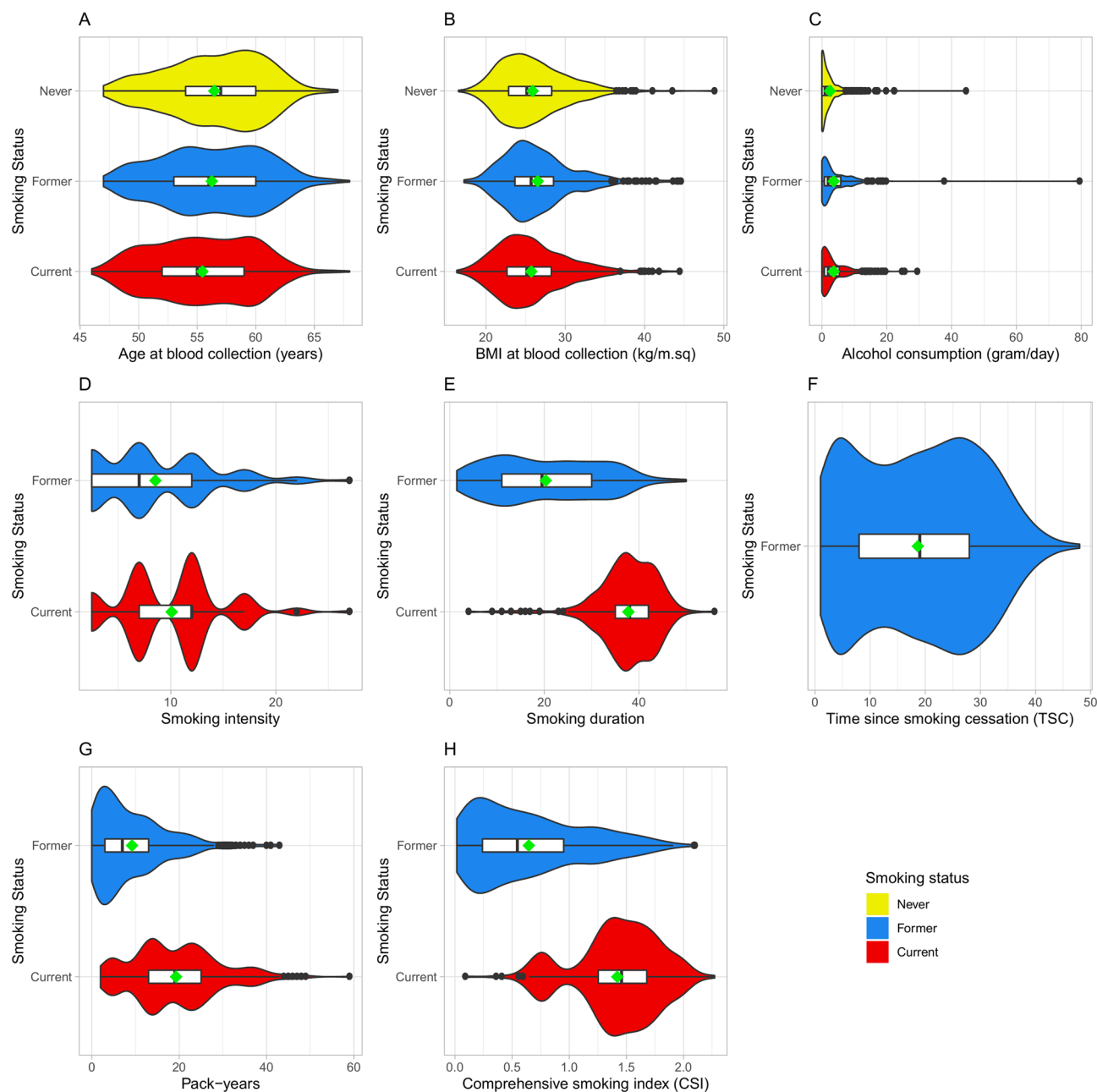
We investigated associations between smoking status and potential covariates, such as age and BMI at blood collection, and white blood cell (WBC) proportions, using Chi-square or Kruskal–Wallis tests. We then performed a 'global test' to indicate any association between these variables and the overall gene expression data. We considered variables that were significant in both of these tests as potential confounders and adjusted for these in further models (Supplementary Table S1).

There were 473, 613, and 622 CS, FS, and NS, respectively, among the 1708 women. The distributions of age and BMI at blood collection did not deviate markedly from normality, whereas the distribution of alcohol consumption was skewed (Fig. 1). Each of these distributions were similar across different categories of smoking status (Fig. 1A–C), but FS had the highest mean BMI and alcohol consumption, and NS had the highest mean age (Supplementary Table S1). Further, the smoking metrics—smoking intensity, smoking duration, pack-years, and CSI score had the highest means for CS as compared to FS (Fig. 1D–H). Finally, there were 192, 147, and 100 PS among CS, FS, and NS, respectively.

**Estimated white blood cell proportions.** We estimated proportions of 22 types of WBCs using an *in silico* gene expression deconvolution method. CD8 T cells, naive CD4 T cells, resting NK cells, M0 macrophages, resting mast cells, and neutrophils were significantly associated with both smoking status and overall gene expression (Supplementary Table S2 and Supplementary Fig. S1). Further, we used linear regression to assess the associations between WBC proportions and smoking metrics. We observed that CD8 T cells were negatively associated with pack-years and CSI score; naive CD4 T cells were positively associated with smoking intensity, smoking duration, pack-years, and CSI score; resting NK cells were negatively associated with smoking intensity, smoking duration, pack-years, and CSI score but positively associated with TSC; resting mast cells were negatively associated with smoking duration; and neutrophils were negatively associated with TSC (Supplementary Table S3).

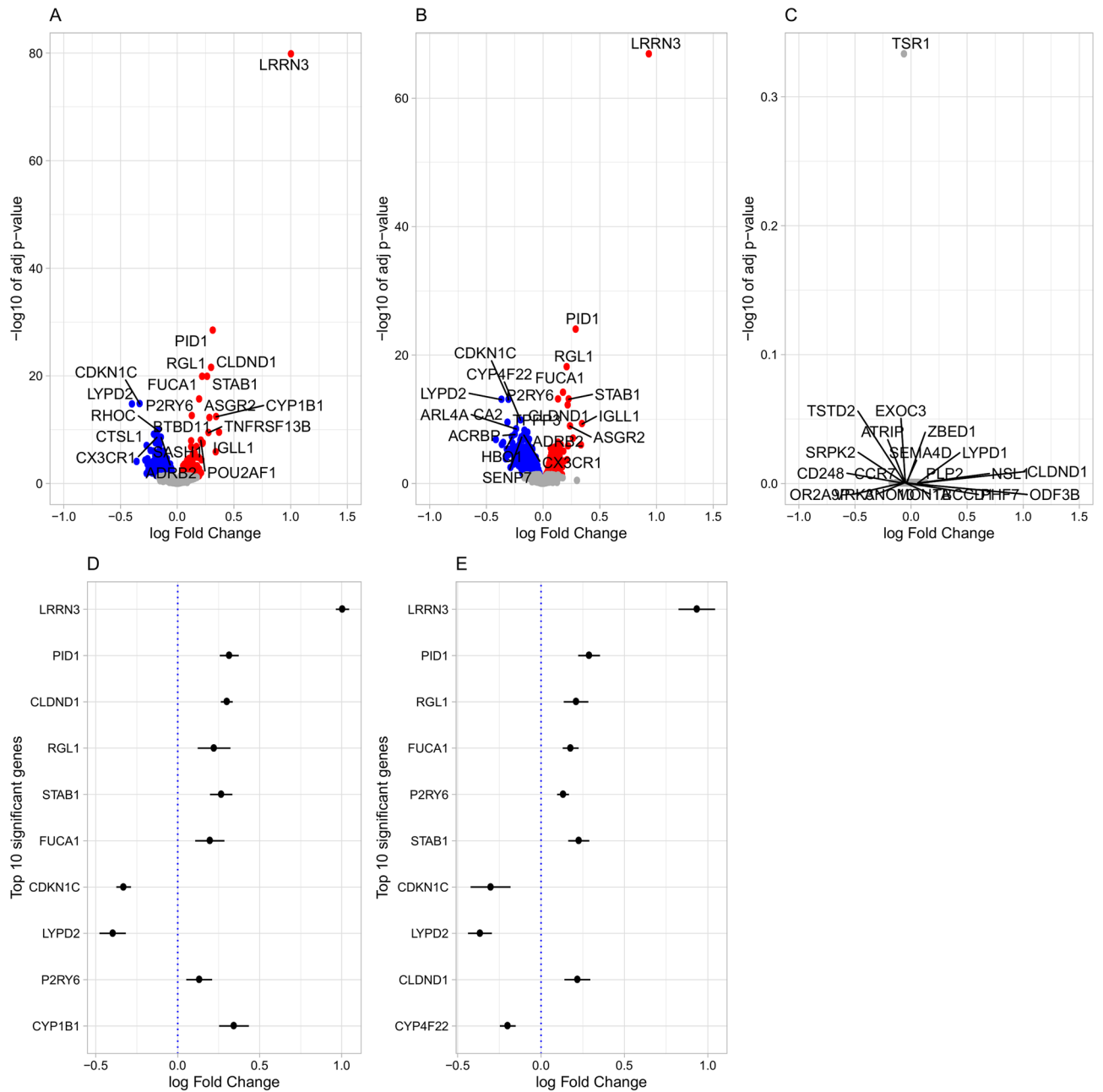
**Differentially expressed genes dependent on smoking status.** We used two adjusted (minimally- and fully-adjusted) models to assess the relationships between smoking status and gene expression profiles, using the 'limma' package for gene-wise linear models. In minimally-adjusted models, we adjusted for technical variables such as laboratory batch (laboratory plates) and sample storage time, while in fully-adjusted models, in addition to the technical variables, we included the following variables that were associated with both the exposure and the outcome: selected WBC proportions, age, BMI, and use of hormone replacement therapy at the time of blood collection, as well as information on alcohol consumption and use of oral contraceptives, which was taken from the main questionnaires. The presence of differentially expressed genes (DEGs) was determined by three comparisons of smoking status groups: CS-vs-NS, CS-vs-FS, and FS-vs-NS. We considered Benjamini–Hochberg false discovery rates (FDR) with the significance threshold  $FDR \leq 0.05$ .

In minimally-adjusted models, there were 1009 DEGs in the CS-vs-NS comparison; 427 up-regulated and 582 down-regulated genes. Correspondingly, in the CS-vs-FS comparison, there were 1371 DEGs (559 up-regulated, 812 down-regulated). In fully-adjusted models, there were 911 DEGs in the CS-vs-NS comparison (355 up-regulated, 556 down-regulated; Fig. 2A,D), and 1082 DEGs in the CS-vs-FS comparison (435 up-regulated, 647 down-regulated; Fig. 2B,E). The two adjusted models had 670 overlapping DEGs in the CS-vs-NS comparison



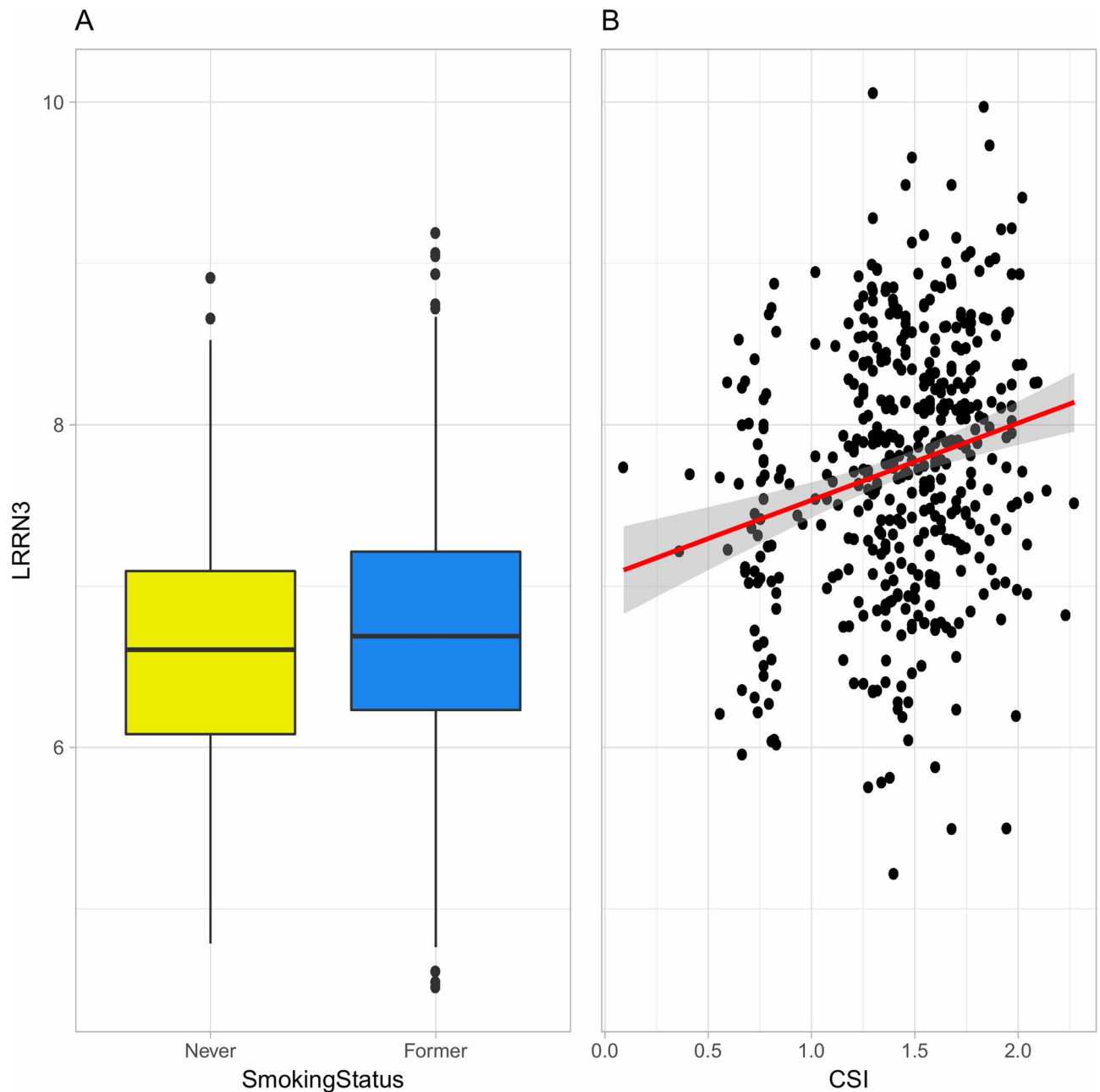
**Figure 1.** Descriptive statistics of study participants by smoking status for (A) age at blood collection, (B) body mass index (BMI) at blood collection, (C) alcohol consumption at baseline, (D) smoking intensity, (E) smoking duration, (F) time since smoking cessation (TSC), (G) pack-years, and (H) comprehensive smoking index (CSI) scores. Yellow, blue, and red coloured violin plots represent kernel density estimates for never, former, and current smokers, respectively. White boxes extend from the 25th to the 75th percentile, vertical bars inside the box represent the median, whiskers extend 1.5 times the length of the interquartile range right and left side of the 75th and 25th percentiles respectively, and outliers are represented as black dots. The green diamond shaped dot represents the respective mean.

(Supplementary Table S4) and 839 in the CS-vs-FS comparison (Supplementary Table S5). Similarly, the CS-vs-NS and CS-vs-FS comparisons had 776 and 652 overlapping DEGs in the minimally- and fully-adjusted models, respectively. In the fully-adjusted models, there were 230 up-regulated and 422 down-regulated genes that overlapped between the CS-vs-NS and CS-vs-FS comparison and displayed the same direction of effects. The top-ranked gene (i.e., the gene with the lowest FDR adjusted  $p$ -values) in all comparisons was *LRRN3* (Supplementary Fig. S2). Receiver operating characteristics (ROC) curve analyses showed that expression levels of *LRRN3*, as measured by the Illumina arrays, could strongly distinguish CS from NS and moderately distinguish FS (with  $\leq 10$  years TSC) from NS (Supplementary Fig. S3). Moreover, in a subset of our dataset, *LRRN3* expression showed similar discriminative power as DNA methylation at the *AHRR* CpG site (cg05575921), which is a known marker for smoking exposure<sup>19</sup>. There were no DEGs in the FS-vs-NS comparison in either



**Figure 2.** Volcano plots for the test statistics in fully-adjusted models from the tests of differentially expressed genes (DEGs) in comparisons of (A) current versus never smokers, (B) current versus former smokers, and (C) former versus never smokers; and forest plots for the 10 top-ranked DEGs in tests of DEGs in comparisons of (D) current versus never smokers and (E) current versus former smokers. In volcano plots (A–C), red dots display up-regulated genes, blue dots display down-regulated genes, while grey dots display genes with FDR > 0.05; the x-axis presents log<sub>2</sub> fold-changes and the y-axis presents –log<sub>10</sub> of FDR adjusted p-values; and gene names displayed are the 20 top-ranked DEGs in the respective tests. In forest plots (D and E), dots in the x-axis represent log<sub>2</sub> fold-changes and the y-axis represents DEGs with the lowest FDR adjusted p-values ranked from the top; the horizontal line for each gene represents their confidence interval; and the vertical blue dotted line represents no difference.

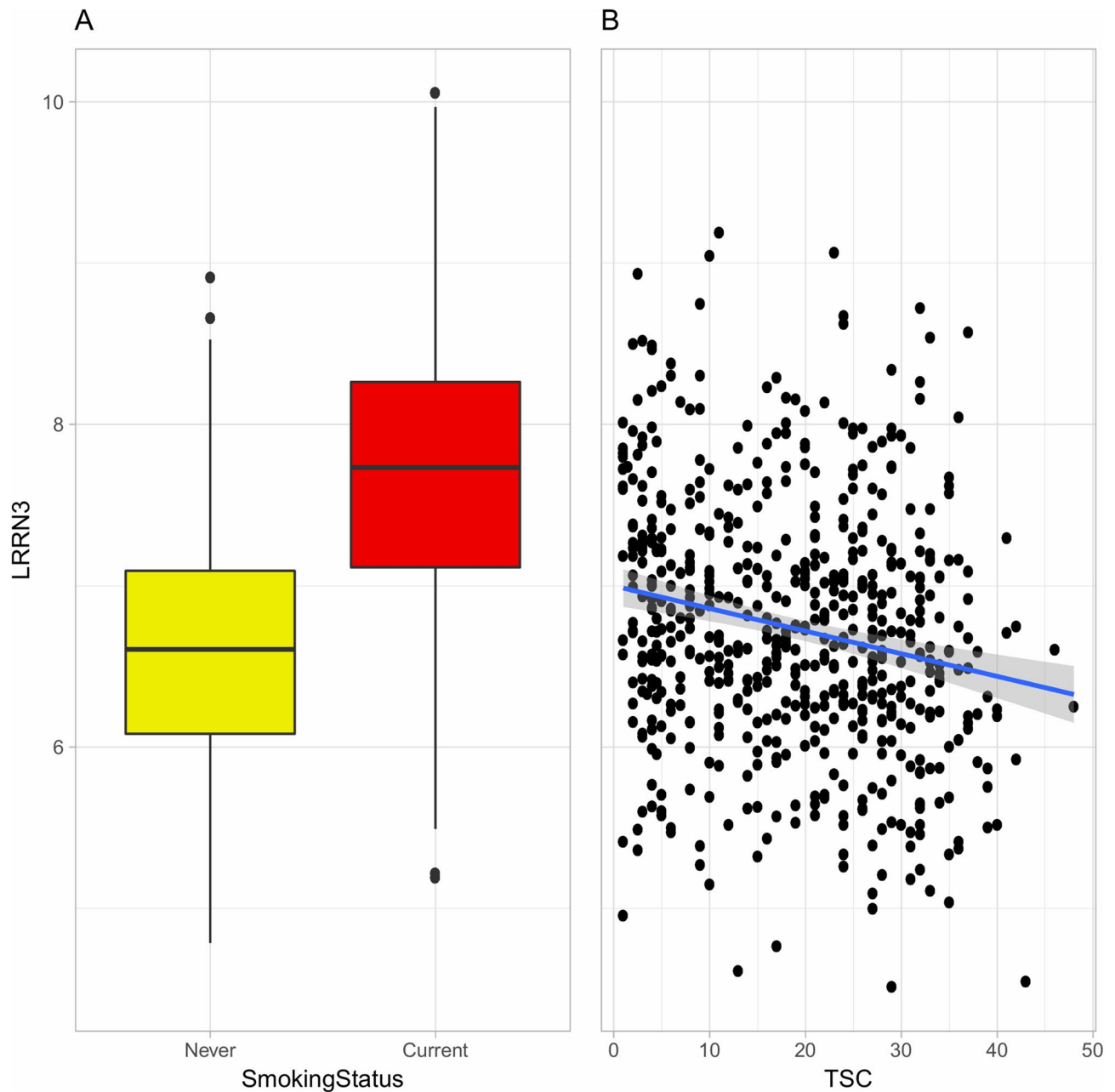
model (Fig. 2C). However, *LRRN3* was the only DEG that remained significant when we included only FS with TSC ≤ 10 years and compared it with NS in the minimally-adjusted model (with log<sub>2</sub> fold-change (logFC) = 0.34 and FDR = 3.63E–04). The p-values were uniformly distributed only in the FS-vs-NS comparison, but not in the other comparisons, as presented in quantile–quantile plots (Supplementary Fig. S4). Further, we used the ‘limma’ package to analyse the effects of passive smoking among NS, by contrasting all NS who were PS in adulthood with the other NS using the minimally-adjusted model. There were no DEGs when testing differences between PS (n = 100) and non-PS (n = 428).



**Figure 3.** Distributions of expression values for the top-ranked gene (*LRRN3*) (A) among never (yellow) and former (blue) smokers and (B) among current smokers according to comprehensive smoking index (CSI) scores. In figure (A), boxes extend from the 25th to the 75th percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75th and 25th percentiles, respectively, and outliers are represented as points. In figure (B), the red line represents the linear regression fit and the shaded grey area its standard error.

**Analyses of smoking metrics within ever smokers.** To identify genes associated with magnitude of smoking exposure, we used the ‘limma’ framework to identify genes for which the expression level correlated with the given smoking metrics among ever smokers. Specifically, we extended the minimally-adjusted model to include the given smoking metrics and analysed CS and FS separately.

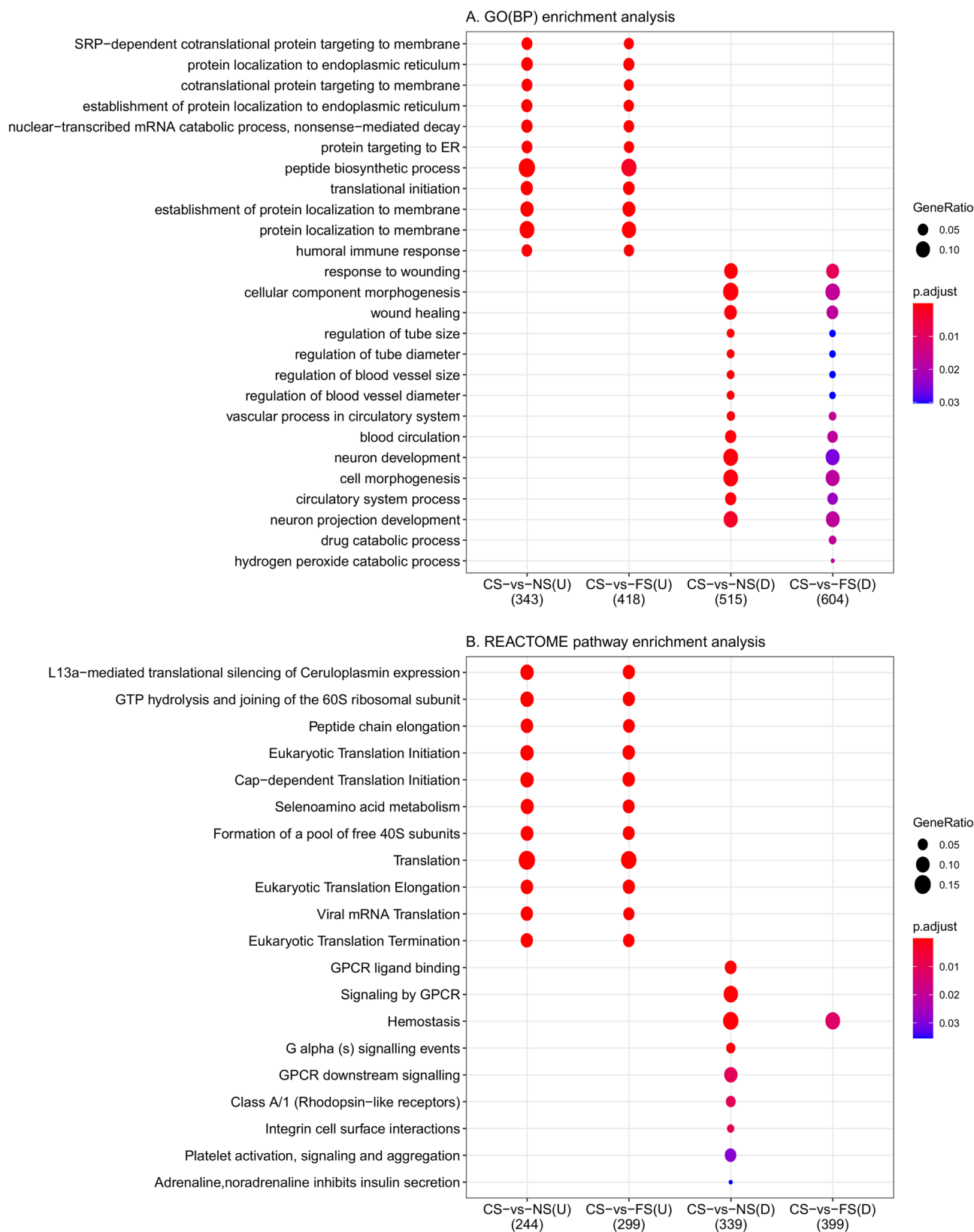
In analyses of CS, the top-ranked gene, *LRRN3* ( $\log_{2}FC = 0.60$ ,  $FDR = 4.70E-05$ ), was positively associated with CSI score (Fig. 3). Further, there were five genes positively associated and two genes negatively associated with smoking intensity (Supplementary Table S6), where *LRRN3* was the top-ranked gene, with a positive association (Supplementary Fig. S5). Likewise, there were three genes positively associated and two genes negatively associated with pack-years (Supplementary Table S7), where *LRRN3* was the top-ranked gene, with a positive association (Supplementary Fig. S6). There were no genes significantly associated with smoking duration among CS.



**Figure 4.** Distributions of expression values for the top-ranked gene (*LRRN3*) (A) among never (yellow) and current (red) smokers and (B) among former smokers according to time since smoking cessation (TSC). In figure (A), boxes extend from the 25th to the 75th percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75th and 25th percentiles, respectively, and outliers are represented as points. In figure (B), the blue line represents the linear regression fit and the shaded grey area its standard error.

In analyses of FS, the top-ranked gene, *LRRN3* ( $\log_{2}FC = -0.014$ ,  $FDR = 2.63E-03$ ), was negatively associated with TSC (Fig. 4). Correspondingly, *NMRAL1* ( $\log_{2}FC = -0.008$ ,  $FDR = 2.72E-02$ ) was negatively associated with pack-years (Supplementary Fig. S7). No genes were significantly associated with smoking intensity, smoking duration, or CSI scores among FS.

**Functional enrichment analyses.** To investigate the potential common functions of the identified DEGs affected by smoking, we performed functional enrichment analyses to identify GO biological processes (BP), GO molecular functions (MF), GO cellular components (CC), Kyoto encyclopaedia of genes and genomes (KEGG) pathways, and REACTOME pathways enriched for DEGs in the CS-vs-NS and CS-vs-FS comparisons (Supplementary Tables S8–12, Fig. 5, and Supplementary Fig. S8). Analyses were performed for DEGs in fully-



**Figure 5.** Summary of functional enrichment analyses for up- and down-regulated genes for the (A) GO(BP) and (B) REACTOME pathway databases. The colour of the dots indicates the adjusted p-value, where red dots represent the most enriched categories; the ‘GeneRatio’ indicates the proportion of genes overlapping between lists of differentially expressed genes (DEGs) and the genes in gene ontology categories. GO: gene ontology; BP: biological processes; CS-vs-NS: comparison of current smokers versus never smokers; CS-vs-FS: comparison of current smokers versus former smokers; U: Up-regulated genes; D: Down-regulated genes.

Database	CS-vs-NS		CS-vs-FS	
	Up-regulated genes (n = 355)	Down-regulated genes (n = 556)	Up-regulated genes (n = 435)	Down-regulated genes (n = 647)
GO(BP)	33	51	22	41
GO(MF)	4	0	6	5
GO(CC)	23	6	14	4
KEGG	1	7	1	0
REACTOME	31	9	34	1

**Table 1.** Number of enriched terms in different categories of enrichment analyses in comparisons of current versus never smokers (CS-vs-NS) and current versus former smokers (CS-vs-FS).

adjusted models and separately for up-regulated and down-regulated genes. The numbers of enriched terms in the respective categories are presented in Table 1.

For both up-regulated and down-regulated genes, enriched categories overlapped considerably for genes that were significant according to the FDR in the CS-vs-NS and CS-vs-FS comparisons. However, there were more enriched categories among genes in the CS-vs-NS comparison, except for GO(MF), where there were significant categories only in the CS-vs-FS comparison. Considering the terms themselves, up-regulated genes were mostly enriched for terms related to translation, such as ribosome (KEGG and GO(CC)), protein localisation to endoplasmic reticulum (GO(BP)), and translation (REACTOME). Terms were also related to immune responses, such as humoral immune response, inflammatory response, and B cell activation (GO(BP)). In contrast, down-regulated genes were enriched for many terms related to circulatory functions, including response to wounding and regulation of blood vessel size (GO(BP)), and extracellular signalling, such as G protein-coupled receptor ligand binding (REACTOME), and plasma membrane region (GO(CC)).

## Discussion

This study presents DEGs across categories of smoking status, as well as genes associated with different smoking metrics within ever smokers in the whole-blood of cancer-free women from the NOWAC postgenome cohort. These assessments, which used quantitative and repetitive smoking metrics, bring novel knowledge about the systemic responses to smoking exposure within ever smokers.

The study participants had similar proportions of CS, FS, and NS. They had comparable mean age and BMI at the time of blood collection as that of the full cohort, and to that of participants in other studies targeting the relation between smoking exposure and gene expression<sup>9–11</sup>. Among the 7713 genes assessed, 911 and 1082 genes were differentially expressed in CS-vs-NS and CS-vs-FS comparisons, respectively. When looking at the DEGs in the CS-vs-NS comparison and the significant genes indicated in corresponding tests in a large meta-analysis containing 10,233 participants (51% women), we found that among the 285 DEGs in our study that overlapped with the 1270 DEGs in that study, 282 genes had the same direction of effects<sup>11</sup>. Moreover, the mean expression levels for the 285 DEGs between CS-vs-NS that overlapped with DEGs identified in corresponding tests in the meta-analysis were higher (7.56) than those DEGs that did not overlap (6.83;  $t = 5.23$ ,  $p$ -value =  $2.63E-07$ ). Still, the average absolute logFC for the overlapped (0.12) and non-overlapped DEGs (0.09;  $W = 128,066$ ,  $p$ -value =  $1.29E-14$ ) were similar. This implies that the relation to smoking was consistent for hundreds of genes between these studies and demonstrates the comprehensive effects of smoking on gene expression in blood.

Around 40% of the genes were over-expressed in CS as compared to both NS and FS (i.e., 60% were under-expressed). Although higher proportions of up-regulated genes have been observed more frequently in other studies<sup>7,9–11</sup>, higher proportions of down-regulated genes have also been observed<sup>14,16</sup>. Interestingly, there could be sex differences in the directionality of observed DEGs, as one study comparing smokers and non-smokers observed that 29% of DEGs in men were down-regulated, compared to 62% in women<sup>7</sup>. However, only about 4% of the DEGs in our study were in X-chromosomes in both the CS-vs-NS and CS-vs-FS comparisons. Notably, differences in gene expression between adult men and women do not need to originate in genes on the X-chromosomes, but a meta-analysis of sex expression differences in blood found that 25% of DEGs do map to the sex chromosomes<sup>20</sup>. Thus, it is unlikely that the higher proportion of down-regulated genes in our study was due to the inclusion of women only.

Among NS, there were no genes associated with self-reported passive smoking in their homes as adults when compared to individuals with no passive smoking exposure. This could indicate that gene expression was more influenced by tobacco smoking of the women themselves. However, this could also be due to lack of statistical power or an imprecise exposure measure (lack of detailed information on timing, duration, and intensity of exposure).

Among CS, there were one, five, and three significant genes that were positively associated with CSI scores, smoking intensity, and pack-years, respectively. Among these, the top-ranked gene, *LRRN3*, was up-regulated in CS, which demonstrated that even within CS, *LRRN3* had a higher expression among those with a higher smoking exposure, as represented by increasing CSI scores, smoking intensity, and pack-years. Among FS, there was one significant gene that was negatively associated with TSC (*LRRN3*) and one that was negatively associated with pack-years (*NMRAL1*). This demonstrated that within FS, those who had quit smoking recently had a higher expression of *LRRN3* than those who had quit long ago, and FS with more pack-years had a lower expression of *NMRAL1* than those with fewer pack-years. Also, when restricting the FS-vs-NS comparison to recent



quitters (with TSC  $\leq 10$  years), *LRRN3* remained significant in minimally-adjusted models. This indicates that there are differences in gene expression related to ongoing smoking exposure in women that persist for *LRRN3* in those who recently stopped smoking. Thus, *LRRN3* expression increases during smoking exposure and years after smoking cessation, but it eventually reverts back to levels similar to those of NS. However, according to the results of our linear model (Fig. 4), it appeared to take approximately 20–30 years for *LRRN3* expression in FS to reach the average expression among NS. The difference in results from the overall FS-vs-NS comparison and those restricted to FS with TSC up to 10 years emphasises that TSC needs to be taken into account when analysing smoking effects in FS.

*LRRN3* was the top-ranked DEG in most comparisons, and its expression differences were large compared to the other DEGs identified. *LRRN3* has been consistently indicated to be over-expressed in the whole-blood of CS or FS in previous studies<sup>9–11,13,14,16,21,22</sup>. This gene is highly expressed in the adrenal glands, the brain, and the lungs, but also in 11 other tissues<sup>23</sup>, and *LRRN3* codes for a membrane protein. The GO database has little information on *LRRN3*'s potential functions, except that electronic annotations indicate that it is involved in the positive regulation of synapse assembly<sup>23,24</sup>. Notably, *LRRN3* has six known SNPs<sup>25</sup> but genetic variants in participants were not available in this study. Top-ranked DEGs other than *LRRN3* in the CS-vs-NS and CS-vs-FS comparisons were *PID1*, *RGL1*, and *STAB1*, and in the analyses of ever smokers was *NMRAL1*. These genes are expressed in various tissues that differed across genes. The main functions of the aforementioned genes are to increase the proliferation of pre-adipocytes (*PID1*)<sup>26</sup>, to be involved in probable guanine nucleotide exchange factor (*RGL1*)<sup>27</sup>, and to act as a scavenger receptor for acetylated low-density lipoprotein, bind to both gram-positive and gram-negative bacteria, and to play a role in the defence against bacterial infection (*STAB1*)<sup>28</sup>. However, the interpretation of the potential function of these genes in blood in relation to smoking is not clear.

We performed functional enrichment analyses for GO(BP), GO(MF), and GO(CC) categories; and for KEGG and REACTOME pathways. This gave insight into the underlying biology and provided knowledge of pathways for the identified DEGs<sup>29</sup>. The overlap in the enriched categories of the up-regulated and down-regulated genes in the CS-vs-NS and CS-vs-FS comparisons indicated that similar GO categories and pathways were enriched when current smoking exposure was compared to both FS and NS. Still, the enrichment was clearer when CS were compared to NS than to FS. The latter might be because the effect of smoking was not completely absent or was being slowly reduced in FS. In addition, the overall lack of overlap for enriched categories of the up-regulated and down-regulated genes likely demonstrated that these separate groups of genes are involved in different pathways.

The GO enrichment analysis indicated categories such as peptide metabolic and biosynthetic processes, protein formation and translation, humoral immune response, structural constituent of ribosome and molecule activity, ribosomal subunits, and adherens junction were up-regulated in CS. In contrast, processes such as response to wounding, circulatory system, regulation of blood vessels and tube size and diameter, neuron projection development, drug and hydrogen peroxide catabolic processes, heme binding, cell body, and hemoglobin complex were down-regulated. Categories indicated in the KEGG and REACTOME enrichment analyses were largely in line with those in GO analysis. In summary, these categories indicate that the DEGs we identified were enriched for functions related to the physiological effects of smoking on the human body, which are well documented in the literatures. This is particularly relevant for the physiological functions linked to the cardiovascular system, as DEGs measured in blood could be directly influenced by such altered functions. For example, carbon monoxide binds to haemoglobin, thereby reducing the blood's oxygen-carrying capacity<sup>30</sup>. Accordingly, our results indicated that smoking could also down-regulate genes involved in the haemoglobin complex, thereby potentially exacerbating smoking's negative effects on oxygen transport. Further, smoking causes several negative vascular effects, including decreased coronary blood flow and myocardial oxygen delivery, as well as adverse effects on lipids, blood pressure, and insulin resistance<sup>31</sup>. Thus, the down-regulated processes for blood vessel size and diameter, and vascular processes in the circulatory system. The general circulatory system processes indicated in whole-blood in this study could be related to these known physiological effects of smoking. We identified that oxidoreductase activity was down-regulated, which is in line with smokers experiencing measurable and immediate oxidative damage, resulting in oxidative stress<sup>3</sup>. We also observed down-regulated wound healing and haemostasis, which is in agreement with observations of a reduced capacity to heal wounds among smokers<sup>3,30</sup>. Lastly, categories related to immune responses were up-regulated in CS. Smoking can compromise the immune system and immune homeostasis as a whole<sup>3</sup>, and gene enrichment analyses of genes related to smoking in other studies have indicated effects on the regulation of immune system processes<sup>9,10,13–16</sup>. GO analyses in a large meta-analysis of genes related to smoking demonstrated enrichment mainly for activation of platelets and lymphocytes, immune response, and apoptosis<sup>11</sup>. The enriched terms for the DEGs in our study only were largely the same as for those for DEGs that overlapped between the meta-analysis and our study (results not presented). Further, the expression of *LRRN3* has been linked to the methylation of a CpG site on the *AHRR* gene<sup>19</sup> and *AHRR* is linked to AHR and CYP proteins, which represent detoxifying mechanisms in the liver. This can be a plausible physiological influence of smoking exposures. Still, considering the great variety of molecules in tobacco smoking, it can potentially influence multiple pathways, which was observed in the GO categories indicated.

In general, gene expression profiles in whole-blood are affected by the underlying composition of WBCs in the respective samples. Thus, skewed WBC proportions could act as confounders when identifying gene expression differences related to exposures like smoking, which can disturb WBC populations<sup>16</sup>. Neutrophils constituted a large fraction of estimated WBCs but was considerably lower as estimated from gene expression than what is typical in blood<sup>32,33</sup> as well as estimated from DNA methylation in a subset of the samples ( $n = 324$ )<sup>19</sup>. Still, we observed that WBC proportions and smoking metrics—especially resting NK cells but also CD8 T cells, resting mast cells, and neutrophils—were negatively associated with increasing smoking exposure. Further, naive CD4 T cells were positively associated with several smoking metrics. These results are in line with observations that smoking may have detrimental effects on the immune capacity of the body. Indeed, smoking has been shown to be a significant and reversible cause of elevated WBC counts in healthy adults<sup>34</sup>. These estimated cell proportions

were included in our fully-adjusted models when assessing DEGs. Still, the top-ranked genes identified in fully-adjusted models were similar to those from the minimally-adjusted models, indicating that these genes were likely not substantially confounded by the distributions of WBC.

The main strength of this study was its use of smoking metrics based on detailed, repeated information on past and recent smoking history of the study participants when assessing DEGs in blood between smoking status groups. Among the women we included in our study, 51%, 24%, and 25% had information available at four, three, and two time points, respectively. Still, this study was based on self-reported smoking information from questionnaires, as in most other studies<sup>9,10,13–16</sup>. Many studies have measured concentrations of the metabolite of nicotine, cotinine, in blood, urine or saliva in addition to self-reported smoking status<sup>9,14–16</sup>. However, due to its relatively short half-life (16–19 h)<sup>35</sup>, it would not have provided valuable information for FS. Further, DNA methylation at specific CpG sites have also showed promising abilities as markers of smoking status and could reflect smoking exposures even decades after cessation<sup>36,37</sup>. In a subset of our data, *LRRN3* demonstrated similar ability to discriminate CS and FS (with  $\leq 10$  years TSC) from NS as compared to methylation at the CpG site in the *AHRR* gene. Therefore, the abilities of *LRRN3* expression as a quantitative marker for discrimination of smoking status should be investigated in other population samples and with the comparison to other markers.

This study comprised a large number of women ( $n = 1708$ ), whereas most studies targeting associations between smoking exposure and gene expression in blood have been conducted in rather small samples, ranging from 9 to 219 participants<sup>9,10,13–15</sup>. The two exceptions are one population-based cohort study in the Netherlands with 3319 participants (65% women)<sup>16</sup> and a meta-analysis with 10,233 participants (51% women)<sup>11</sup>. As mentioned, our results are in line with those observed in these studies. The present study included only cancer-free women, although we cannot disregard influences of other common chronic diseases. Further, this study was based on whole-blood samples, which is a relevant tissue to investigate the effects of smoking, as it expresses a large proportion of the genes in the human genome<sup>16</sup>. Still, the current cross-sectional study results represent snapshots of gene expression in blood<sup>38</sup>. Lastly, although RNA-sequencing has become a routinely used technology, results from microarray technology, like those in this study, are still reliable and overall comparable to RNA-sequencing results<sup>39</sup>. However, RNA-sequencing technology would be relevant for studying the effects of smoking exposure on other genes not captured by the Illumina microarray technology, such as most non-coding RNAs.

In conclusion, our results demonstrated associations between smoking exposure and gene expression profiles in whole-blood of cancer-free women in the NOWAC postgenome cohort. The use of quantitative, reliable, and repeated measurements of past and recent smoking exposures was the novelty of this study, as it contributes new knowledge on systemic responses of smoking exposure. Close to a thousand DEGs in comparisons between CS and NS or FS, *LRRN3*, was the top-ranked gene. *LRRN3* was also associated with CSI score, smoking intensity, and pack-years among CS; and with TSC among FS. Consequently, *LRRN3* expression in blood is a molecular signal of smoking exposure that could supplant self-reported smoking data in gene expression studies of the association between smoking exposure and specific phenotypes. The biological functionality of the DEGs identified were linked to circulatory functions, translation, and immune responses, and could indicate systemic impacts of smoking. Genes that are differentially expressed depending on smoking exposure could be of interest in studies that focus on the effects of smoking exposure on health. This study has provided knowledge on the relationship of genes and pathways with detailed information on smoking exposure among cancer-free women.

## Methods

**Study population.** The NOWAC study is a nation-wide, population-based prospective cohort study initiated in 1991. Currently, it includes approximately 172,000 Norwegian women aged 30–70 years. Women were randomly selected from the Norwegian National Population Register and sent an invitation letter along with a first questionnaire, which included a detailed set of questions related to smoking exposure, height, weight, reproductive history, hormone replacement therapy, alcohol consumption, family history of breast cancer, dietary patterns, use of medication, and others. Since then, each woman has answered between one and three follow-up questionnaires (main questionnaires). The NOWAC study database takes information from the Cancer Registry of Norway, as well as national death and emigration registries. Details about the NOWAC study are available in Lund et al.<sup>40</sup>.

The current study was based on data from the NOWAC postgenome cohort<sup>41,42</sup>, a sub-cohort of the NOWAC study. This consists of approximately 50,000 women who, between 2003 and 2006, had blood samples collected in PreAnalytiX (PAX) gene-tubes for gene expression analysis and, at the same time, answered a less extensive questionnaire about their lifestyle. The current study incorporated microarray-based expression profiles in bio-banked whole-blood samples from cancer-free women in the NOWAC postgenome cohort, who were originally enrolled as controls in several studies on breast, lung, ovarian, and endometrial cancers, and diabetes. We obtained relevant questionnaire and registry information from NOWAC databases and excluded those women that did not respond to any questions on smoking exposure, those who participated in more than one study, and those who were diagnosed with cancer before 2017. This resulted in a final analytical sample of 1708 women.

**Smoking status and smoking metrics.** The main questionnaires included detailed questions regarding past and current smoking exposures, including ages at smoking initiation and cessation, average number of cigarettes smoked per day across age intervals, and details about passive smoking. Smoking status and smoking metrics (smoking intensity, smoking duration, TSC, pack-years, and CSI scores) were based on information from all main questionnaires and the questionnaire completed at the time of blood collection. Smoking intensity was defined as the average number of cigarettes smoked per day during years of active smoking, smoking duration was the duration of active smoking in years, and TSC was the time since smoking cessation in years. Pack-years quantify individual, long-term exposure to tobacco smoking<sup>43</sup>; this variable was calculated by the formula:

*Number of pack-years* = (*smoking intensity*/20) × *smoking duration*. We considered 20 cigarettes in 1 pack, which is standard in the Norwegian context. CSI score is a cumulative measure of smoking exposure that incorporates smoking intensity (int), smoking duration (dur), and TSC (tsc). CSI scores were calculated using the formula<sup>18</sup>:  $CSI = (1 - 0.5^{dur/\tau})(0.5^{tsc/\tau}) \ln(int + 1)$ , where  $\tau$  is an estimated half-life parameter, and  $\delta$  is an estimated lag time parameter describing TSC and total duration as follows:

$$tsc^* = \max(tsc - \delta, 0) \text{ and } dur^* = \max(dur + tsc - \delta) - tsc^*.$$

**Laboratory analyses and pre-processing of the gene expression data.** Total RNA was extracted and purified from PAX gene-tube samples according to the PAX gene blood RNA kit protocol at the Genomics Core Facility, Norwegian University of Science and Technology (NTNU), Trondheim. A NanoDrop ND 8000 spectrophotometer (ThermoFisher Scientific, Wilmington, DE, USA) was used to assess RNA purity, and bio-analyser capillary electrophoresis (Agilent Technologies, Palo Alto, CA, USA) was used to assess RNA integrity. Complementary RNA (cRNA) was prepared using the Illumina TotalPrepT-96 RNA amplification kit, and hybridised to Illumina human WG-3 or HT-12 expression bead chip microarrays. The raw microarray images were processed in Illumina genome studio. The laboratory analysis date varied from January 2011 to January 2015.

For each study sample set separately, potential outliers were evaluated based on plots such as principal component analysis (PCA) plots and boxplots of probe signals displaying variation along with the laboratory quality measures<sup>44</sup>. We performed background correction, removed bad quality probes, and filtered probes detected in less than 20% of samples. Further, we performed  $\log_2$  transformation and quantile normalisation before all data were combined and inspected for batch effects using PCA plots. We performed gene annotation using the Bioconductor packages ‘lumi’, ‘lumiHumanIDMapping’, and ‘illuminaHumanv4.db’<sup>45–47</sup>. If there were more than one probe annotated to each gene, the probe with the largest inter-quartile range was kept, which resulted in 7713 unique genes in the data analysed. Estimates for the proportions of 22 populations of WBCs in samples were obtained using the CIBERSORT procedure<sup>48</sup>.

**Statistical analyses.** We considered covariates and WBC proportions as potential confounders if they were significantly associated with smoking status according to Chi-square or Kruskal–Wallis tests, and with overall gene expression data according to the ‘global test’ from the Bioconductor package ‘global test’<sup>49</sup>. We used two adjusted (minimally- and fully-adjusted) models to assess the relationship between smoking status and gene expression profiles. We also performed linear regression analysis between WBC proportions and smoking metrics to assess their associations.

We performed all the main analyses using R version 3.2.1 and 3.6.2<sup>50</sup>. We used the Bioconductor package ‘limma’<sup>51</sup> for the gene-wise linear models. The presence of DEGs was determined by three comparisons of smoking status groups: CS-vs-NS, CS-vs-FS, and FS-vs-NS, using a significance threshold of  $FDR \leq 0.05$ <sup>52</sup>. We performed analyses of smoking metrics within CS and FS separately, and for adult PS within NS. Further, data on DNA methylation at the CpG site *AHRR* gene, cg05575921, was available in a subset of participants ( $n = 324$ )<sup>19</sup>. Therefore, we compared the ability of the top-ranked gene in our analyses and CpG site in the *AHRR* gene (cg05575921) using ROC curves. Differences in average expression and  $\log_2FC$  between groups of DEGs were tested using t-test and Wilcoxon rank sum test, respectively. To evaluate common biological functions of results of the gene-wise tests, we performed functional enrichment analyses of all significant up-regulated genes and all significant down-regulated genes. We used the bioconductor packages ‘clusterProfiler’<sup>53</sup> and ‘ReactomePA’<sup>54</sup> to conduct functional enrichment analyses of GO(BP), GO(MF), and GO(CC) categories, and KEGG<sup>55</sup> and REACTOME pathways for DEGs from different smoking status groups.

**Ethical statement.** The Regional Ethical Committee of North Norway (REK) has approved the NOWAC study and the NOWAC postgenome cohort (Reference Numbers: 2010/2075/REK Nord and 2014/1605/REK Nord, respectively), and the collection and storage of human biological material, the individual case-control studies, and gene expression analyses that this project was constructed from. The women gave written informed consent for the blood collection and for gene expression analyses<sup>42</sup>. All methods were carried out in accordance with relevant guidelines and regulations in the manuscript for human.

## Data availability

Data cannot be shared publicly because of local and national ethical and security policy. Data access for researchers will be conditional on adherence to both the data access procedures of the Norwegian Women and Cancer Cohort and the UiT – The Arctic University of Norway (contact via Tonje Braaten <tonje.braaten@uit.no> and Arne Bastian Wiik <arne.b.wiik@uit.no>) in addition to an approval from the local ethical committee.

Received: 14 September 2020; Accepted: 15 December 2020

Published online: 12 January 2021

## References

1. World Health Organization. *Don't Let Tobacco Take Your Breath Away: Choose Health, not Tobacco: 31 May, World tobacco day* (accessed 10 August 2019). [https://www.who.int/docs/default-source/world-no-tobacco-day/wntb-2019-brochure.pdf?sfvrsn=deac371c\\_22](https://www.who.int/docs/default-source/world-no-tobacco-day/wntb-2019-brochure.pdf?sfvrsn=deac371c_22) (2019).
2. Sopori, M. Effects of cigarette smoke on the immune system. *Nat. Rev. Immunol.* **2**, 372–377 (2002).

3. Bonnie, R. J., Kwan, L. Y. & Stratton, K. R. *Public Health Implications of Raising the Minimum Age of Legal Access to Tobacco Products* 91–123 (National Academies Press, Washington, DC, 2015).
4. Carey, M. A. *et al.* It's all about sex: gender, lung development and lung disease. *Trends Endocrinol. Metab.* **18**, 308–313 (2007).
5. Langhammer, A., Johnsen, R., Holmen, J., Gulsvik, A. & Bjermer, L. Cigarette smoking gives more respiratory symptoms among women than among men The Nord-Trøndelag Health Study (HUNT). *J. Epidemiol. Community Health* **54**, 917–922 (2000).
6. World Health Organization. *WHO Report on the Global Tobacco Epidemic, 2008: The MPOWER Package* (World Health Organization, Geneva, 2008).
7. Paul, S. & Amundson, S. A. Differential effect of active smoking on gene expression in male and female smokers. *J. Carcinog. Mutag.* **5**, 1000198 (2014).
8. McHale, C. M., Zhang, L., Thomas, R. & Smith, M. T. Analysis of the transcriptome in molecular epidemiology studies. *Environ. Mol. Mutagen.* **54**, 500–517 (2013).
9. Beineke, P. *et al.* A whole blood gene expression-based signature for smoking status. *BMC Med. Genom.* **5**, 58 (2012).
10. Cheng, X. *et al.* Smoking affects gene expression in blood of patients with ischemic stroke. *Ann. Clin. Transl. Neurol.* **6**, 1748–1756 (2019).
11. Huan, T. *et al.* A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. *Hum. Mol. Genet.* **25**, 4611–4623 (2016).
12. Lampe, J. W. *et al.* Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. *Cancer Epidemiol. Prev. Biomark.* **13**, 445–453 (2004).
13. Martin, F., Talikka, M., Hoeng, J. & Peitsch, M. C. Identification of gene expression signature for cigarette smoke exposure response—from man to mouse. *Hum. Exp. Toxicol.* **34**, 1200–1211 (2015).
14. Na, H. K. *et al.* Tobacco smoking-response genes in blood and buccal cells. *Toxicol. Lett.* **232**, 429–437 (2015).
15. Van Leeuwen, D. M. *et al.* Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. *Carcinogenesis* **28**, 691–697 (2007).
16. Vink, J. M. *et al.* Differential gene expression patterns between smokers and non-smokers: cause or consequence?. *Addict. Biol.* **22**, 550–560 (2017).
17. Arimilli, S., Madahian, B., Chen, P., Marano, K. & Prasad, G. L. Gene expression profiles associated with cigarette smoking and moist snuff consumption. *BMC Genom.* **18**, 156 (2017).
18. Lefondré, K., Abrahamowicz, M., Xiao, Y. & Siemiatycki, J. Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Stat. Med.* **25**, 4132–4146 (2006).
19. Sandanger, T. M. *et al.* DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian Women and Cancer cohort. *Science* **8**, 16714 (2018).
20. Bongen, E. *et al.* Sex differences in the blood transcriptome identify robust changes in immune cell proportions with aging and influenza infection. *Cell Rep.* **29**, 1961–1973 (2019).
21. Charlesworth, J. C. *et al.* Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC Med. Genom.* **3**, 29 (2010).
22. Obeidat, M. *et al.* The effect of different case definitions of current smoking on the discovery of smoking-related blood gene expression signatures in chronic obstructive pulmonary disease. *Nicotine Tob. Res.* **18**, 1903–1909 (2016).
23. National Center for Biotechnology Information. *LRRN3 Leucine Rich Repeat Neuronal 3* [Homo sapiens (human)]: NCBI. <https://www.ncbi.nlm.nih.gov/gene/54674#gene-expression>. Accessed May 28, 2020 (2020).
24. UniProt consortium. *LRRN3: uniprot.org*. <https://www.uniprot.org/uniprot/Q9H3W5>. Accessed Apr 17, 2020 (2020).
25. GeneCardsSuite. *LRRN3 Gene: genecards.org*. <https://www.genecards.org/cgi-bin/carddisp.pl?gene=LRRN3>. Accessed Nov 5, 2020 (2020).
26. Wang, B. *et al.* Identification and characterization of NYGGF4, a novel gene containing a phosphotyrosine-binding (PTB) domain that stimulates 3T3-L1 preadipocytes proliferation. *Gene* **379**, 132–140 (2006).
27. UniProt consortium. *RGL1: uniprot.org*. <https://www.uniprot.org/uniprot/Q9NZL6>. Accessed Apr 17, 2020 (2020).
28. Adachi, H. & Tsujimoto, M. FEEL-1, a novel scavenger receptor with in vitro bacteria-binding and angiogenesis-modulating activities. *J. Biol. Chem.* **277**, 34264–34270 (2002).
29. Khatri, P., Sirota, M. & Ten, B. A. J. years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).
30. Silverstein, P. Smoking and wound healing. *Am. J. Med.* **93**, S22–S24 (1992).
31. Erhardt, L. Cigarette smoking: an undertreated risk factor for cardiovascular disease. *Atherosclerosis* **205**, 23–32 (2009).
32. Moses, K. & Brandau, S. (eds) *Human Neutrophils: Their Role in Cancer and Relation to Myeloid-Derived Suppressor Cells. Seminars in Immunology* (Elsevier, Amsterdam, 2016).
33. Treffers, L. W., Hiemstra, I. H., Kuijpers, T. W., Van den Berg, T. K. & Matlung, H. L. Neutrophils in cancer. *Immunol. Rev.* **273**, 312–328 (2016).
34. Higuchi, T. *et al.* Current cigarette smoking is a reversible cause of elevated white blood cell count: cross-sectional and longitudinal studies. *Prev. Med. Rep.* **4**, 417–422 (2016).
35. Jarvis, M. J., Russell, M., Benowitz, N. L. & Feyerabend, C. Elimination of cotinine from body fluids: implications for noninvasive measurement of tobacco smoke exposure. *Am. J. Public Health* **78**, 696–698 (1988).
36. Guida, F. *et al.* Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum. Mol. Genet.* **24**, 2349–2359 (2015).
37. Joehanes, R. *et al.* Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).
38. Carlson, M. D. & Morrison, R. S. Study design, precision, and validity in observational studies. *J. Palliat. Med.* **12**, 77–82 (2009).
39. Mantione, K. J. *et al.* Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med. Sci. Monit. Basic Res.* **20**, 138 (2014).
40. Lund, E. *et al.* Cohort profile: the Norwegian women and cancer study—NOWAC—Kvinner og kreft. *Int. J. Epidemiol.* **37**, 36–41 (2007).
41. Dumeaux, V. *et al.* Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast Cancer Res.* **10**, R13 (2008).
42. Lund, E. *et al.* A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC postgenome cohort as a proof of principle. *BMC Med. Res. Methodol.* **16**, 28 (2016).
43. National Cancer Institute. *Pack Years. NCI Dictionary of Cancer Terms*. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pack-year>. Accessed Dec 1, 2019.
44. Bøvelstad H. M., Holsbø E., Bongo L. A., Lund E. A standard operating procedure for outlier removal in large-sample epidemiological transcriptomics datasets. *BioRxiv*. 144519 (2017).
45. Du P., Feng G., Kibbe W., & Lin S. lumiHumanIDMapping: illumina identifier mapping for human. *R package version*. **1** (2016).
46. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
47. Dunning M., Lynch A., & Eldridge M. illuminaHumanv4.db: illumina HumanHT12v4 annotation data (chip illuminaHumanv4). *R package version*. **1** (2015).
48. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

49. Goeman, J. J., Van De Geer, S. A., De Kort, F. & Van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99 (2004).
50. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015 (ver 3.2.1) and 2019 (ver 3.6.2)). <https://www.R-project.org>.
51. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl. Acids Res.* **43**, e47–e (2015).
52. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).
53. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omic J. Integr. Biol.* **16**, 284–287 (2012).
54. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
55. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30 (2000).

## Acknowledgements

The authors gratefully extend their gratitude to all the Norwegian women participating in the NOWAC study. We thank B.A. and all the other personnel, who have handled the administration of the data collection and the biobanks. We thank T.B., M.L. and N.S. for providing the data set. Special thanks to J.I. for helping with data wrangling and T.P.-T. for editing the language of the manuscript. This research received no external funding. The publication charges for this article have been funded by a grant from the publication fund of UiT —The Arctic University of Norway, Tromsø, Norway.

## Author contributions

N.B. made the contributions as first authors. T.H.N. and T.M.S. conceptualized the main research idea. T.H.N. curated the data. T.H.N., P.S., T.M.S., and N.B. designed the research methodology. N.B. performed the formal data analysis, wrote the manuscript, and Supplementary Information with significant contribution from T.H.N. All the co-authors discussed the results and reviewed the manuscript and Supplementary Information.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-020-80158-8>.

**Correspondence** and requests for materials should be addressed to N.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

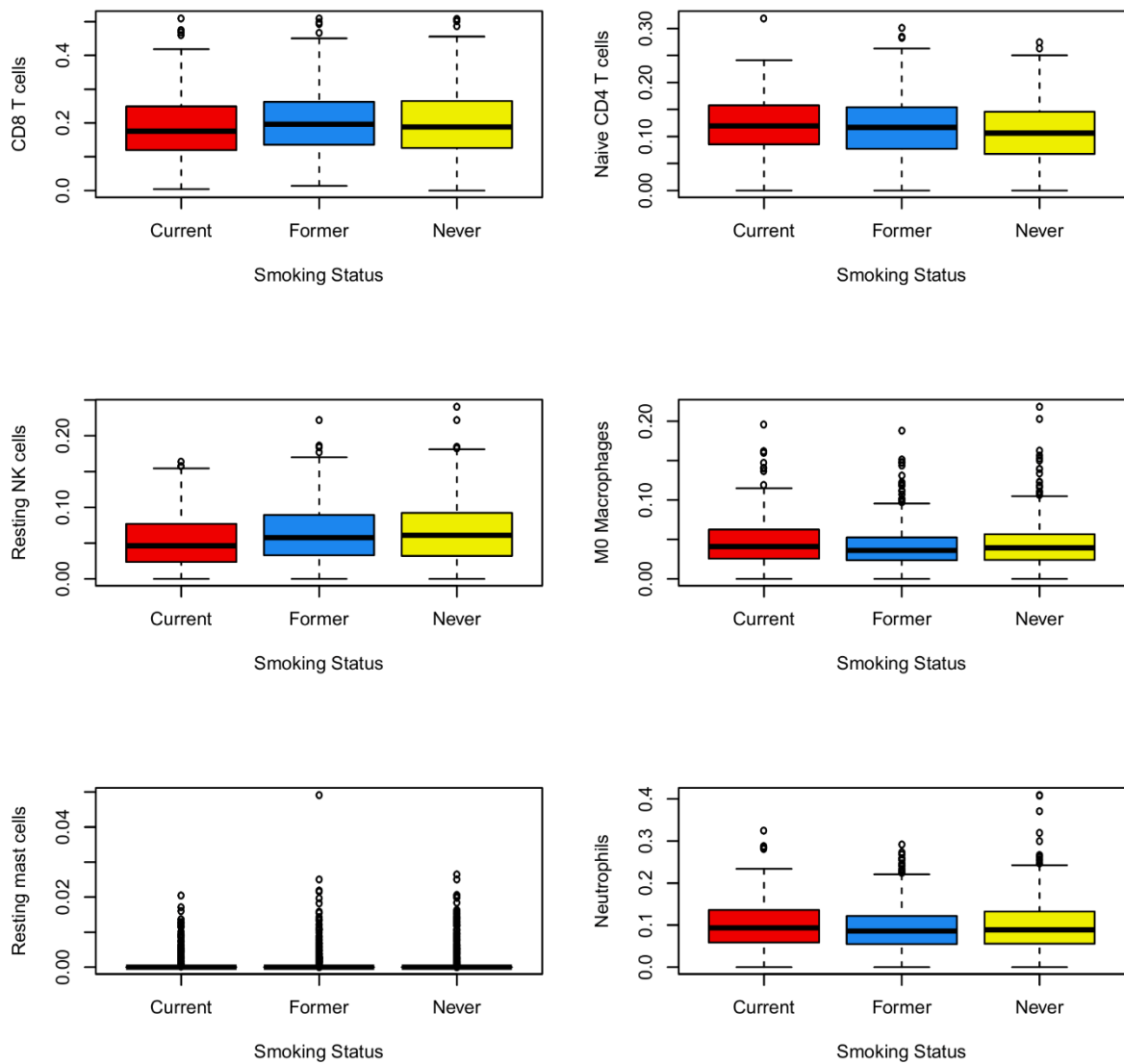


**Gene expression in blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) postgenome cohort**

**Nikita Baiju<sup>1,\*</sup>, Torkjel M. Sandanger<sup>1</sup>, Pål Sætrum<sup>2, 3, 4, 5</sup>, and Therese H. Nøst<sup>1, 5</sup>**

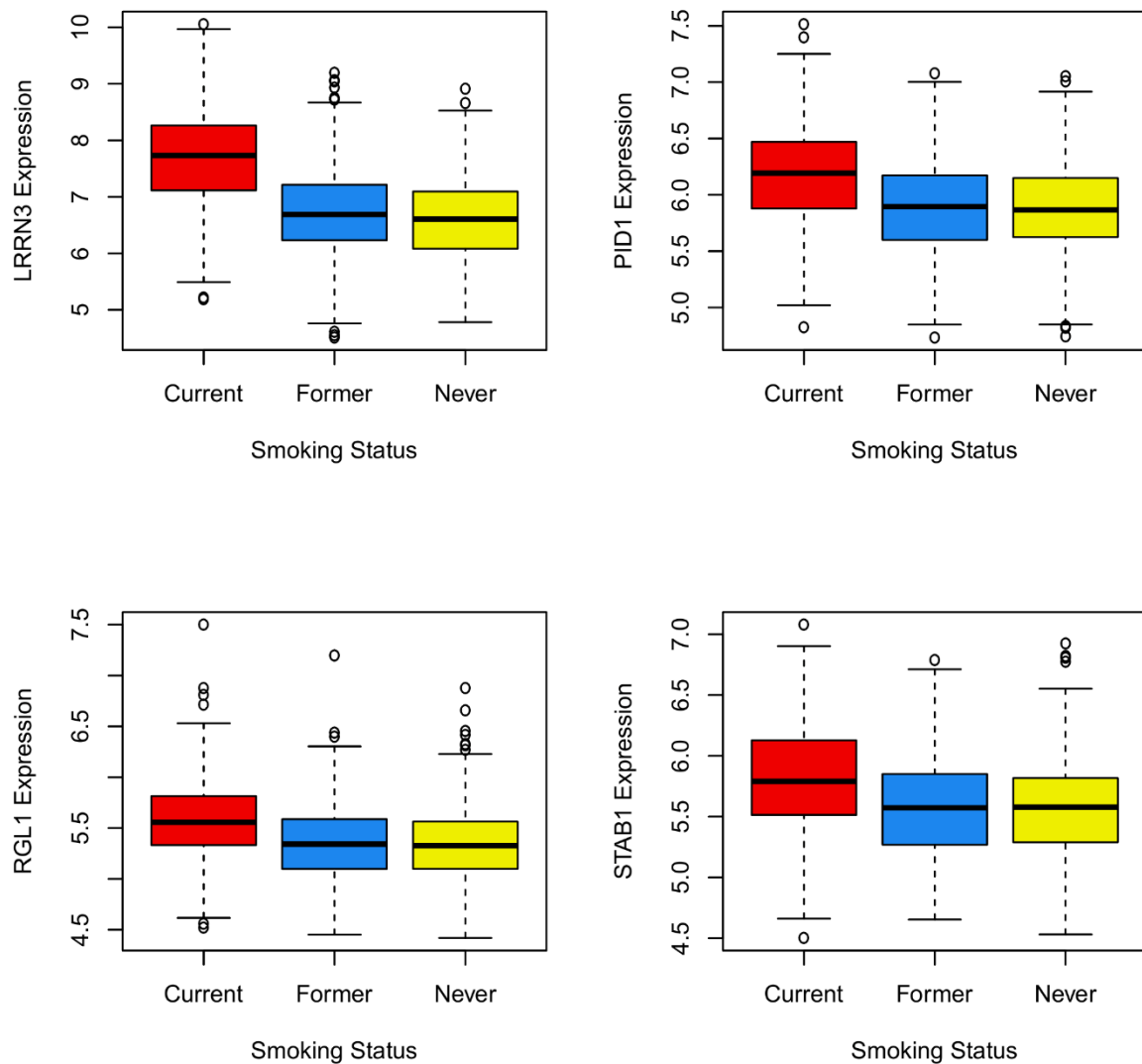
**Supporting information**

**Supplementary Figures**

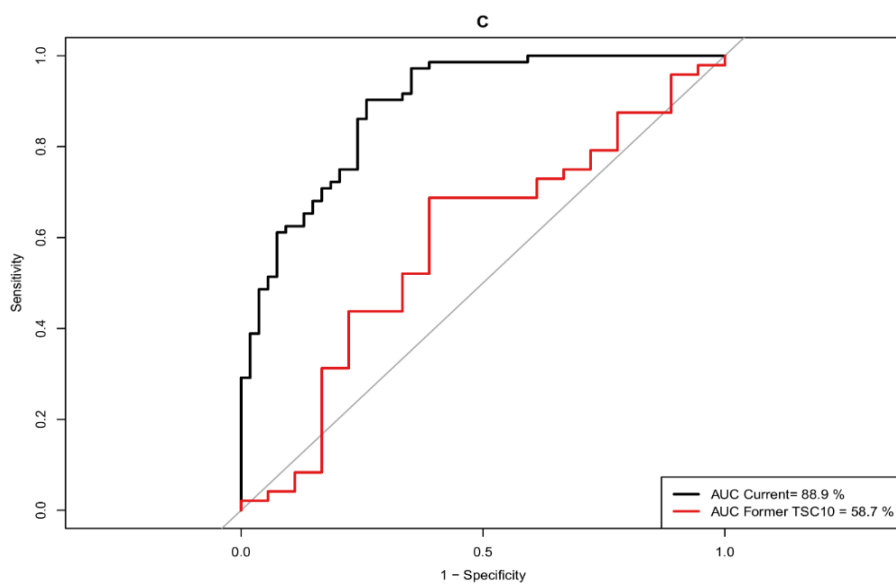
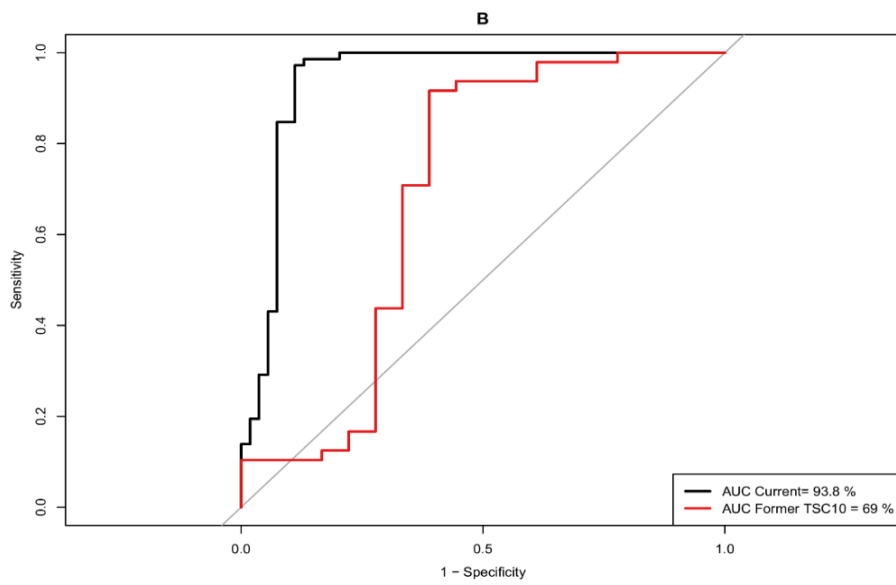
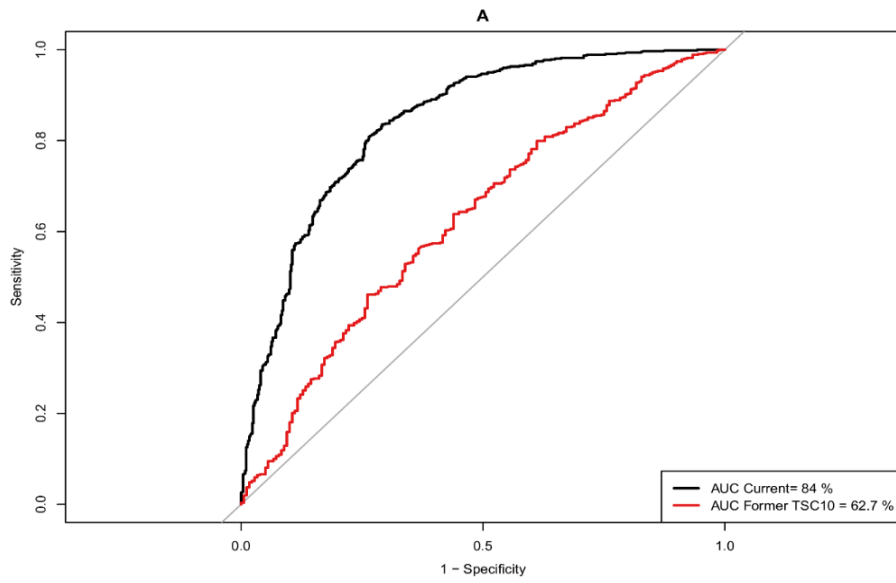


**Supplementary Figure S1. Boxplots for white blood cell (WBC) proportions that were different among current smokers (red), former smokers (blue), and never smokers (yellow).** The X-axis displays the smoking status and the Y-axis displays the proportions of WBCs. Boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as small circles. (TIFF)

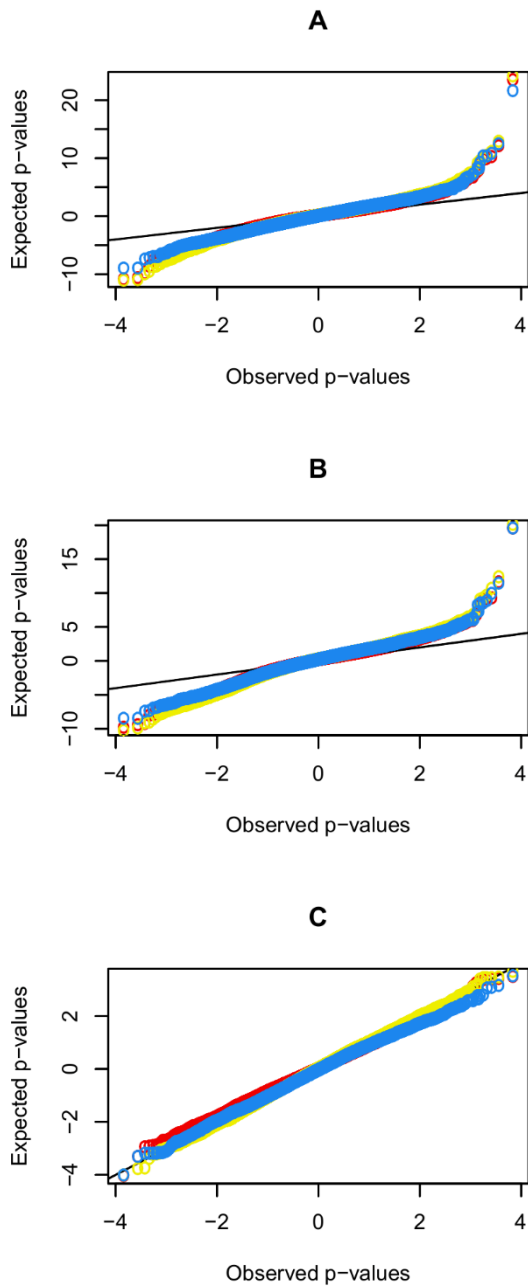




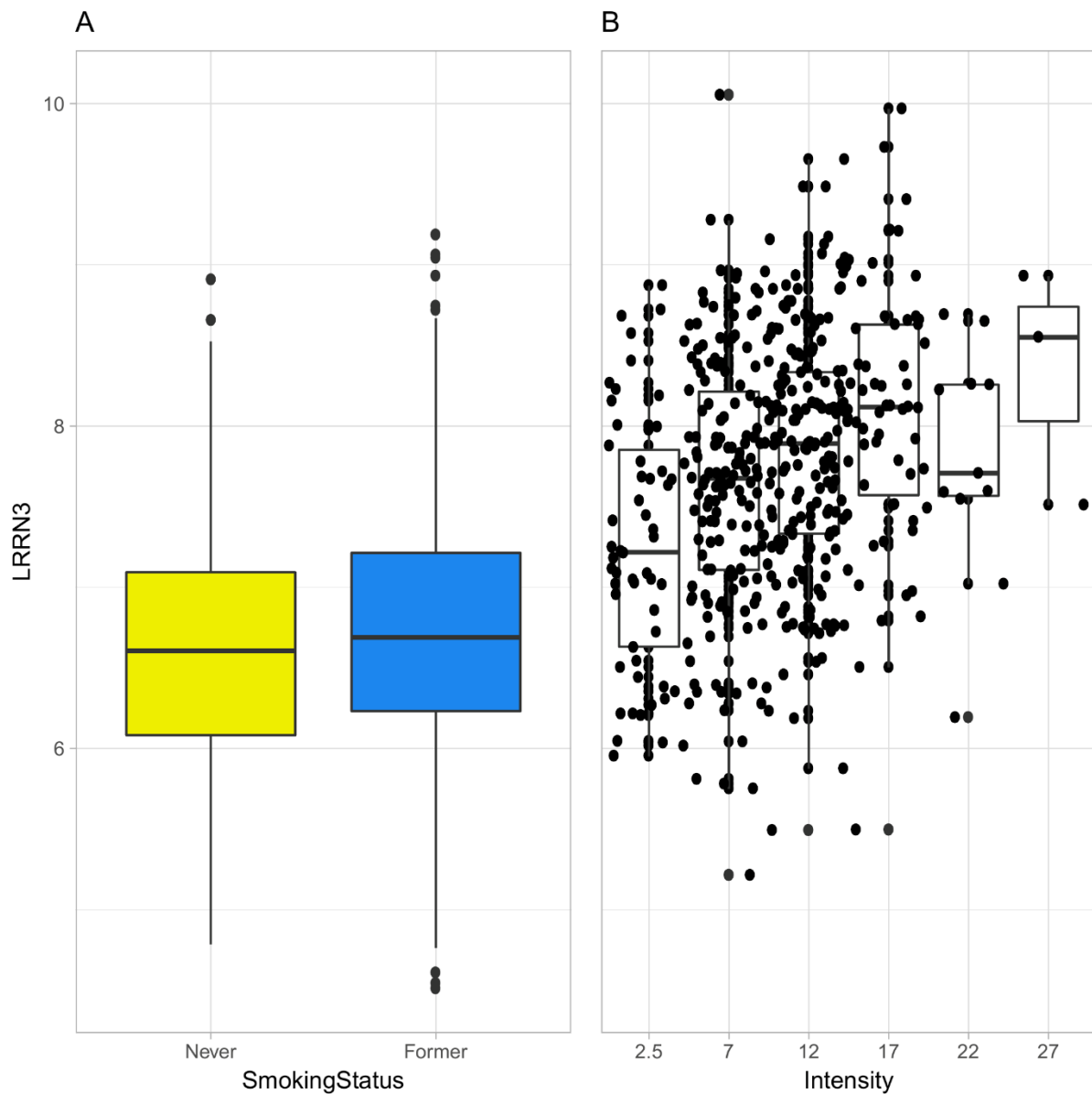
**Supplementary Figure S2. Boxplots for expression values of the four top-ranked genes in comparisons of current vs never smokers, and current vs former smokers that were different among current smokers (red), former smokers (blue), and never smokers (yellow).** The X-axis displays the smoking status and the Y-axis displays the gene expression values. Boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as small circles. (TIFF)



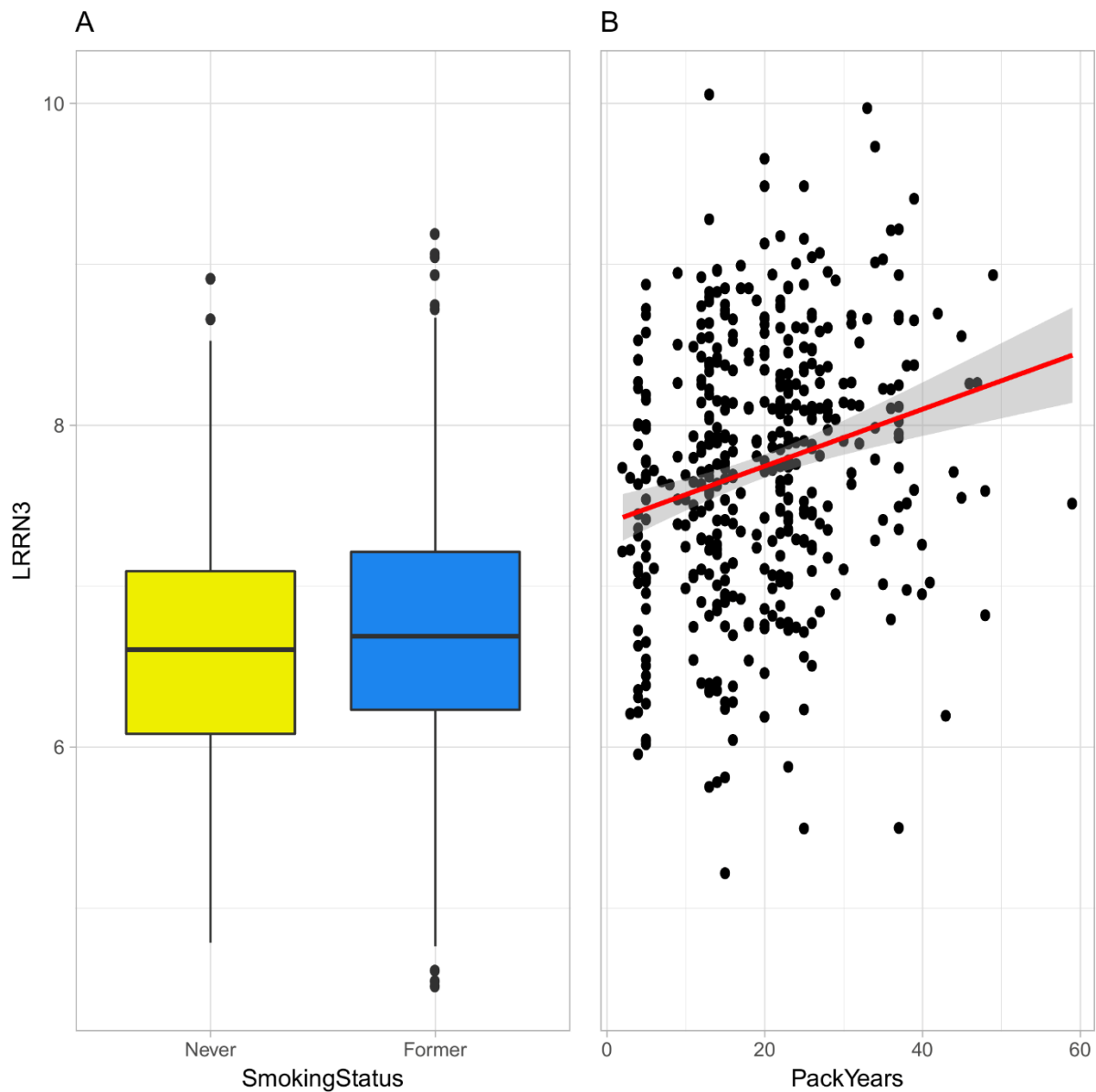
**Supplementary Figure S3. Receiver operating characteristics (ROC) curve for (A) presents the ability of *LRRN3* to discriminate between current from never smokers (black line, n=1095) and former (with TSC $\leq$ 10 years) smokers from never smokers (red line, n=810), (B) presents the ability of cg05575921, a CpG on the *AHRR* gene, for the same discrimination in a subset of samples n=126 and 74, respectively), and (C) presents the ability of *LRRN3*, for the subset of samples that had DNA methylation data available (n=126 and 74, respectively), for the same discrimination. The X-axis presents the specificity of the model and the Y-axis presents the sensitivity.**



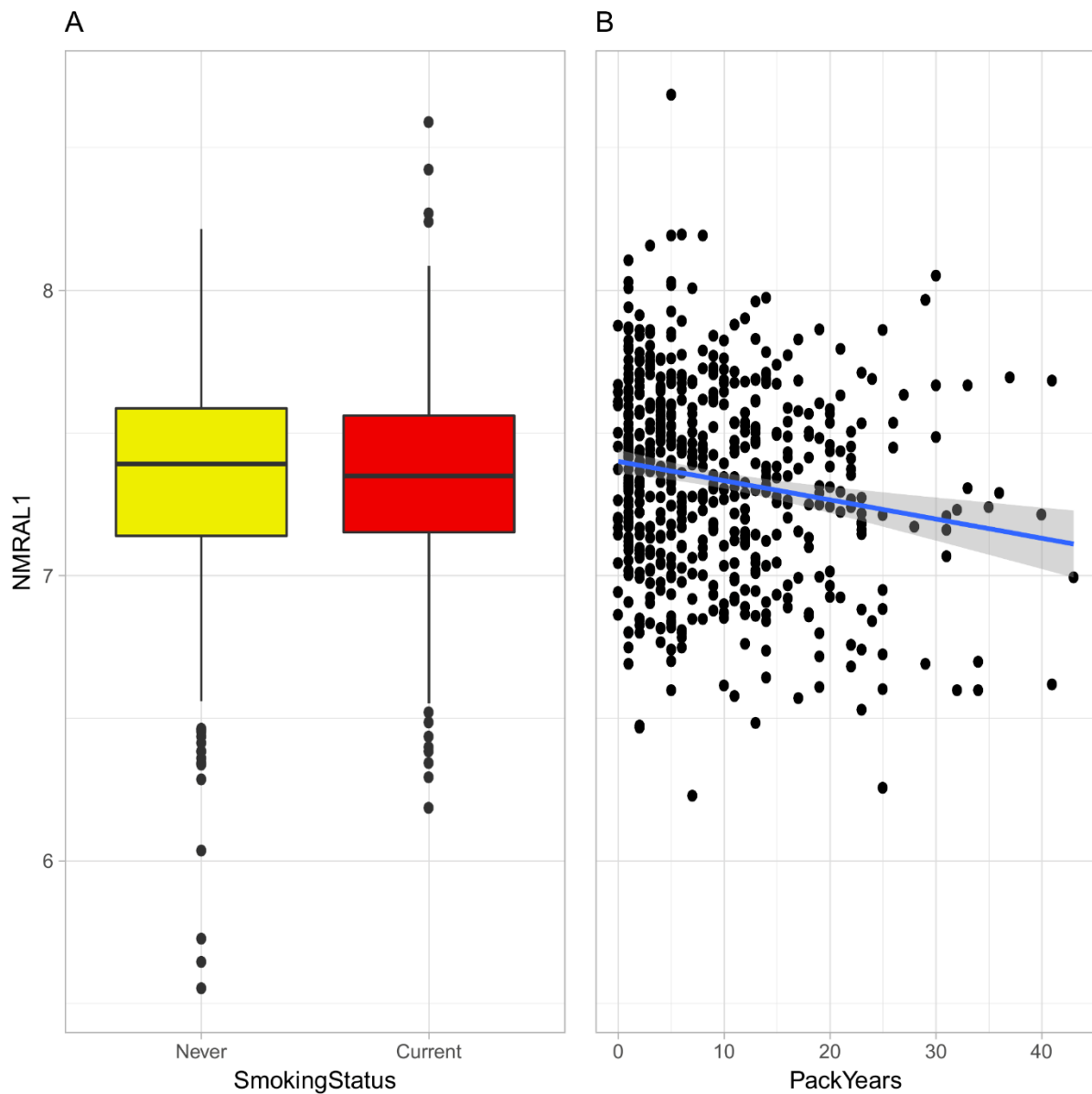
**Supplementary Figure S4. Quantile-quantile plots for comparisons of (A) current vs never smokers, (B) current vs former smokers, and (C) former vs never smokers in unadjusted models (red), minimally-adjusted models (yellow), and fully-adjusted models (blue). The X-axis shows the ‘observed p-values’ and the Y-axis shows the ‘expected p-values’. (TIFF)**



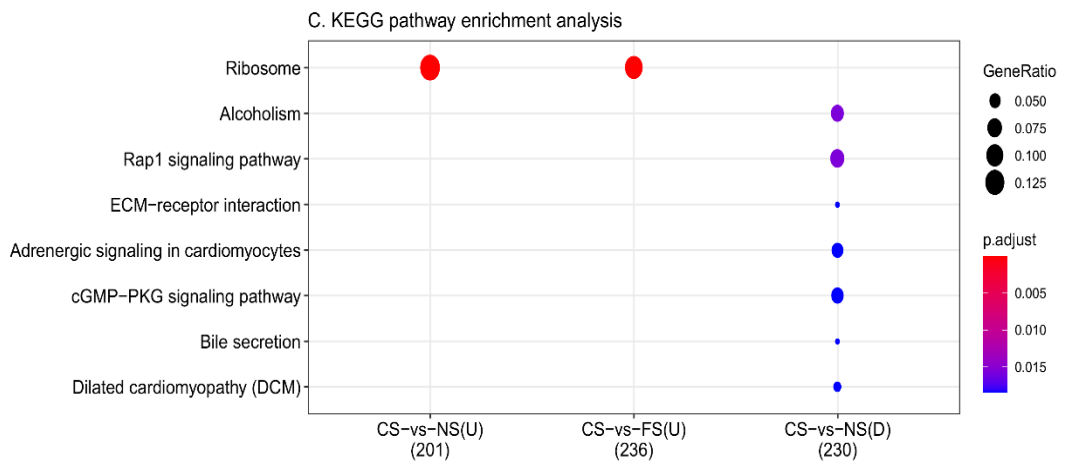
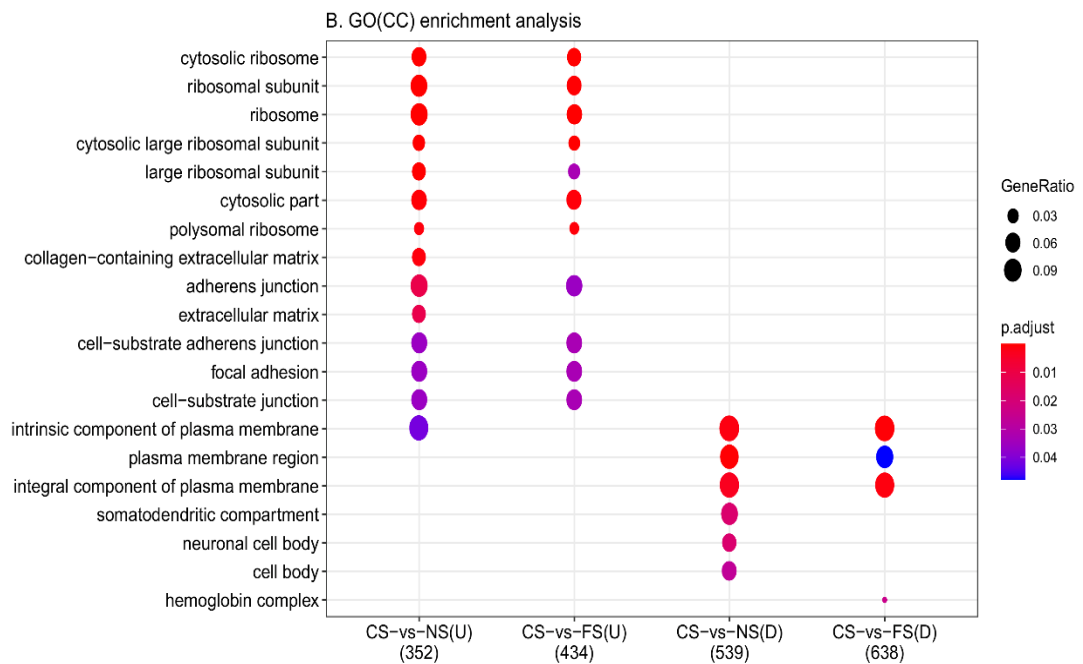
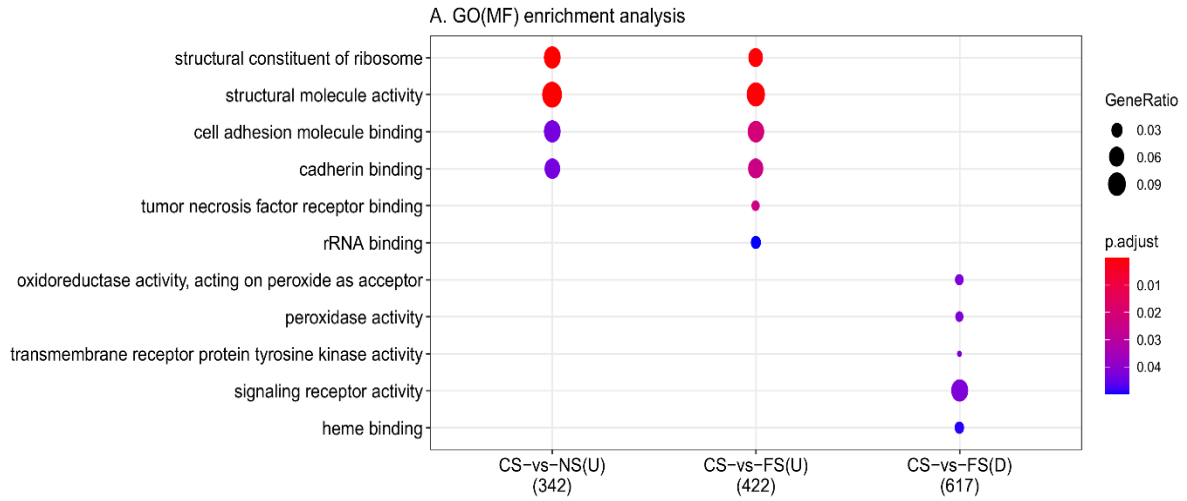
**Supplementary Figure S5. Distributions of expression values for the top-ranked significant gene (*LRRN3*) (A) among never and former smokers and (B) among current smokers according to smoking intensity.** In figure A: yellow colour represents never smokers and blue colour represents former smokers; boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as points. (TIFF)



**Supplementary Figure S6. Distributions of expression values for the top-ranked significant gene (*LRRN3*) (A) among never and former smokers and (B) among current smokers according to pack-years.** In figure A: yellow colour represents never smokers and blue colour represents former smokers; boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as points. In figure B: the red line represents the regression line with a shaded grey area representing the standard error. (TIFF)



**Supplementary Figure S7. Distributions of expression values for the top-ranked significant gene (*NMRAL1*) (A) among never and current smokers and (B) among former smokers according to pack-years.** In figure A: yellow colour represents never smokers and red colour represents current smokers; boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as points. In figure B: the blue line represents the regression line with a shaded grey area representing the standard error. (TIFF)





**Supplementary Figure S8. Summary of functional enrichment analyses for up- and down-regulated genes for the (A) GO(MF), (B) GO(CC), and (C) KEGG pathway databases.** The colour of the dots indicates the adjusted p-value, where red dots represent the most enriched categories; the ‘GeneRatio’ indicates the proportion of genes overlapping between lists of differentially expressed genes (DEGs) and the genes in gene ontology categories. GO: gene ontology; MF: molecular functions; CC: cellular components; KEGG: Kyoto Encyclopedia of Genes and Genomes; CS-vs-NS: comparison of current smokers vs never smokers; CS-vs-FS: comparison of current smokers vs former smokers; U: Up-regulated genes; D: Down-regulated genes.

The supplementary information of this paper comprised of large tables that are not suitable for print and are not included here. These files can be accessed online at following link:  
<https://www.nature.com/articles/s41598-020-80158-8>.

## **Paper II**

**Associations of gene expression in blood with BMI and weight changes among women in the NOWAC postgenome cohort.**

Baiju N, Rylander C, Sætrum P, Sandanger TM, Nøst TH.

*Accepted by Obesity –A Research Journal*



# **Associations of gene expression in blood with BMI and weight changes among women in the NOWAC postgenome cohort**

**Nikita Baiju<sup>1\*</sup>, Charlotta Rylander<sup>1</sup>, Pål Sætrum<sup>2,3,4,5</sup>, Torkjel M. Sandanger<sup>1</sup>,**

**Therese H. Nøst<sup>1,5</sup>**

<sup>1</sup>Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, NO 9037 Tromsø, Norway

<sup>2</sup>Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, NO 7491 Trondheim, Norway

<sup>3</sup>Department of Computer Science, Norwegian University of Science and Technology, NO 7491 Trondheim, Norway

<sup>4</sup>Bioinformatics Core Facility, Norwegian University of Science and Technology, NO 7491 Trondheim, Norway

<sup>5</sup>Department of Public Health and Nursing, K.G. Jebsen Center for Genetic Epidemiology, Norwegian University of Science and Technology, NO 7491 Trondheim, Norway

\*Corresponding author:

## **Contact Info:**

### **Nikita Baiju**

Department of Community Medicine, Faculty of Health Sciences,  
UiT The Arctic University of Norway, NO-9037 Tromsø, Norway

Email ID: [nikita.baiju@uit.no](mailto:nikita.baiju@uit.no)

Telephone No.: +47 77 64 48 04

**Keywords:** gene expression, body mass index-BMI, obesity, weight change, women

**Running title:** BMI, weight changes, and gene expression in NOWAC

**Word count:** 4,495

**Funding:** This research received no external funding. The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway, Tromsø, Norway.

**Disclosure:** The authors declare no competing interests. There were not any competing financial interests in relation to the work described.

**Author contributions:** N.B. made the contributions as first author. T.H.N. and C.R. conceptualised the main research idea. T.H.N. curated the data. T.H.N., P.S., C.R., T.M.S., and N.B. designed the research methodology. N.B. performed the formal data analyses, wrote the manuscript and the supplementary information with significant contribution from T.H.N. All the co-authors discussed the results and reviewed the manuscript and supplementary information.

## **Study Importance**

### **What is already known?**

- Obesity has been linked to altered gene expression in whole-blood, yet few studies have investigated the association between blood gene expression and BMI in a large sample of women.
- No study has assessed the association between blood gene expression and past WCs.

### **What are the new findings in this manuscript?**

- A large number of BMI-associated DEGs, but few WC-associated DEGs (i.e., >700 and <168 DEGs, respectively) were identified in blood of women in Norway.
- The biological functions of BMI-associated DEGs were linked to general metabolism, erythrocyte functions, oxidative stress, and immune processes, whereas WC-associated DEGs were linked to signal transduction.
- This is the first study to our knowledge to conclude that blood gene expression reflects current BMI more than past WCs.

### **How might these results change the direction of research or the focus of clinical practice?**

The study results likely reflect systemic impacts of obesity, especially reticulocyte-erythrocyte ratio shifts in blood, as these functions coincide with its known physiological effects. This knowledge is relevant for further research related to the health effects of BMI and WC, especially those that focus on blood-based markers.

## Abstract

**Objective:** We aimed to evaluate associations between blood gene expression profiles and 1) current body mass index (BMI), and 2) past weight changes (WCs) among women, who had never been diagnosed with cancer, in the NOWAC postgenome cohort.

**Methods:** This cross-sectional study (N=1,694) used gene expression profiles and information from three questionnaires: Q1 (baseline), Q2 (follow-up), and Q3 (blood collection). We performed gene-wise linear regression models to identify differentially expressed genes (DEGs), and functional enrichment analyses to identify their biological functions.

**Results:** When assessing BMI<sub>Q3</sub>, we observed 2,394, 769, and 768 DEGs for the obesity-vs-normal-weight, obesity-vs-overweight, and overweight-vs-normal-weight comparisons, respectively. Up to 169 DEGs were observed when investigating WC<sub>Q3-Q1</sub> (mean=7 years, range=5.5-14 years), and WC<sub>Q3-Q2</sub> (mean=1 year, range=<1 month-9 years) in interaction models with BMI categories, of which, 1-169 genes were associated with WCs, and 0-9 were associated with interaction effects of BMI and WCs. Biological functions of BMI-associated DEGs were linked to metabolism, erythrocytes, oxidative stress, and immune processes, whereas WC-associated DEGs were linked to signal transduction.

**Conclusions:** Many BMI-associated, but few WC-associated DEGs were identified in blood of women in Norway. The biological functions of BMI-associated DEGs likely reflect systemic impacts of obesity, especially blood reticulocyte-erythrocyte ratio shifts.



## **Introduction**

Overweight and obesity are states of abnormal or excessive fat accumulation that cause risks to health<sup>1</sup>. These states are widespread, and global projections indicate more than 2.16 and 1.12 billion individuals will have overweight and obesity by 2030, respectively<sup>2</sup>. Although the primary causes of obesity are considered to be excess dietary intake and inadequate physical activity, other factors (e.g., endocrine disruptions, smoking cessation) may also contribute<sup>3,4</sup>. Obesity and overweight are major risk factors for non-communicable diseases, such as diabetes, cardiovascular diseases, musculoskeletal disorders, and several cancers<sup>1,5</sup>. Additionally, independent of body composition, weight gain in adulthood is a risk factor for several cancers, including postmenopausal breast cancer<sup>5,6</sup>.

Gene expression profiles can improve our understanding of the molecular mechanisms of multifactorial conditions like obesity<sup>7</sup>. In cross-sectional studies, increased body mass index (BMI) has been associated with differences in the expression of several genes<sup>8,9</sup> that indicated alterations in biological functions related to the regulation of body mass, metabolism, and cellular function<sup>8,10</sup>. Insulin resistance, oxidative stress, and liver damage markers were overexpressed in people with obesity compared to those without obesity<sup>11</sup>. Gene expression profiles in adipose tissue have been associated with obesity and weight loss in several studies<sup>9,12-19</sup>. This is expected as, biologically, adipose tissue is most relevant to obesity<sup>10</sup>. Few studies have examined gene expression related to obesity either in whole-blood<sup>10,20</sup> or peripheral blood mononuclear cells (PBMCs)<sup>21</sup>, and found differences. However, no study has assessed this relationship in a large, population-based sample. Furthermore, no study has yet evaluated differences in blood gene expression related to weight changes (WCs).

We aimed to assess the associations between gene expression profiles in whole-blood and 1) current BMI and 2) past WCs in a large, population-based sample of women, who never have been diagnosed with cancer, and to assess the biological functions of differentially expressed genes (DEGs).

## **Materials and Methods**

### **Study design and sample**

We used a cross-sectional study design based on microarray data from women participating in the prospective, population-based Norwegian Women and Cancer (NOWAC) postgenome

cohort. This subcohort of the NOWAC study consists of approximately 50,000 women (mean age: 49.78 years; mean BMI: 23.38 kg/m<sup>2</sup>) who had blood samples collected during 2003-2006 for gene expression analysis. Samples were collected in PreAnalytiX (PAX) gene-tubes, and details about the study sample and women selected for gene expression analyses are available in Baiju et al.<sup>22</sup>. Several case-control studies have investigated gene expression profiles in the NOWAC postgenome cohort; here, we only included the controls from these studies and further excluded women who had later been diagnosed with cancer, so that the study sample only included women who had never been diagnosed with cancer (N=1,694) (details about inclusion and exclusion criteria are presented in Figure 1A). All included women completed a baseline questionnaire (Q1), many completed a follow-up questionnaire (Q2), and all completed a third questionnaire at the time of blood collection (Q3). Mean interval between Q1 and Q2 (Q1-Q2) was approximately 6 years, between Q2 and Q3 (Q2-Q3) approximately 1 year, and between Q1 and Q3 (Q1-Q3) approximately 7 years (Figure 1B). We obtained relevant questionnaire and registry information from the NOWAC study databases. The Regional Ethical Committee of North Norway (REK) has approved the collection and storage of data and human biological material in the NOWAC cohort and biobank (NOWAC postgenome cohort) (Reference numbers: 2010/2075/REK Nord and 2014/1605/REK Nord, respectively).

### **Laboratory analyses and pre-processing of gene expression data**

Laboratory analyses were performed between January 2011 and January 2015 at the Genomics Core Facility, NTNU, Trondheim. Total RNA was extracted and purified from PAX gene-tube samples following the PAX gene blood RNA kit protocol. RNA purity and RNA integrity were assessed using a NanoDrop ND 8000 spectrophotometer (ThermoFisher Scientific, Wilmington, DE, USA) and bio-analyser capillary electrophoresis (Agilent Technologies, Palo Alto, CA, USA), respectively. Complementary DNA was prepared using the Illumina TotalPrepT-96 RNA amplification kit and hybridised to Illumina Human WG-3 or HT-12 expression bead chip microarrays. The raw microarray images were processed in Illumina Genome Studio.

Details about the pre-processing of gene expression data are available in Baiju et al.<sup>22</sup>. Briefly, we performed background correction, removed bad-quality probes, and filtered probes detected in <20% of samples. Further, we performed log<sub>2</sub> transformation and quantile normalisation before all data were combined and inspected for batch effects using principal component analysis plots. These stringent filtering criteria rendered 9,095 probes, and the probe with the

highest interquartile range was selected per gene, which resulted in 7,713 unique genes in our dataset.

### **Body mass index and weight changes**

BMI at Q1 ( $BMI_{Q1}$ ), Q2 ( $BMI_{Q2}$ ), and Q3 ( $BMI_{Q3}$ ) was calculated by dividing weight in kg by the square of height in m, and then categorised according to the World Health Organisation standard (underweight:  $<18.5$  kg/m<sup>2</sup>, normal-weight: 18.5-24.9 kg/m<sup>2</sup>, overweight: 25.0-29.9 kg/m<sup>2</sup>, obesity:  $\geq 30.0$  kg/m<sup>2</sup>).

We calculated WCs between Q1 and Q3 ( $WC_{Q3-Q1}$ , mean interval 7 years), and between Q2 and Q3 ( $WC_{Q3-Q2}$ , mean interval 1 year). We also defined WC categories based on patterns of WC between Q1-Q2 and Q2-Q3 : consistent stable weight (CSW, women with stable weight (-2 to +2 kg) at Q1-Q2 and Q2-Q3); consistent weight gain (CWG, women with weight gain (above +2 kg) at Q1-Q2 and Q2-Q3); consistent weight loss (CWL, women with weight loss (below -2 kg) at Q1-Q2 and Q2-Q3); former weight gain (FWG, women with weight gain at Q1-Q2 and stable weight at Q2-Q3); former weight loss (FWL, women with weight loss at Q1-Q2 and stable weight at Q2-Q3); recent weight gain (RWG, women with stable weight at Q1-Q2 and weight gain at Q2-Q3); and recent weight loss (RWL, women with stable weight at Q1-Q2 and weight loss at Q2-Q3).

### **Statistical analyses**

We performed all analyses using R version 3.6.3. We used the Bioconductor package ‘*limma*’ for gene-wise linear models to identify DEGs and considered a significance threshold of false discovery rate (FDR)  $\leq 0.05$ .

#### *Body mass index analyses*

We assessed the association between blood gene expression and  $BMI_{Q3}$  modelled as a categorical variable (categorical BMI analyses) in three comparisons: obesity-vs-normal-weight, obesity-vs-overweight, and overweight-vs-normal-weight. To assess incremental associations, we also modelled  $BMI_{Q3}$  as a continuous standardised metric (continuous BMI analyses) and scaled it using the R function ‘*scale*’, which for each observation subtracts the mean and divides by the standard deviation. Forty-one women had missing information on  $BMI_{Q3}$ , resulting in an analytical sample of 1,653 women for these analyses.

### *Weight change analyses*

We assessed the association between blood gene expression and past WC modelled as a categorical variable (categorical WC analyses) in six comparisons: CWG-vs-CSW, CWL-vs-CSW, FWG-vs-CSW, FWL-vs-CSW, RWG-vs-CSW, and RWL-vs-CSW. We then conducted sensitivity analyses restricted to women with <1 year between Q2 and Q3. We excluded 160 women who reported decreased weight at Q1-Q2 and increased weight at Q2-Q3 and vice versa, i.e., weight-cyclers, and 499 women with missing values, resulting in an analytical sample of 1,035 women in these analyses.

We also assessed the association between gene expression and past WC modelled as a continuous metric in two interaction models that included BMI category to assess trends across these categories (WC-BMI interaction analyses). To account for differences in the intervals of  $WC_{Q3-Q1}$  and  $WC_{Q3-Q2}$ , we divided the absolute values of WC (kg) by the number of years between Q3 and Q1 or Q2 (kg/year) before scaling it (R function *'scale'*). The first interaction model included  $BMI_{Q1}$  or  $BMI_{Q2}$  and succeeding WCs (i.e.,  $BMI_{Q1} * WC_{Q3-Q1}$  or  $BMI_{Q2} * WC_{Q3-Q2}$ ); the second included current BMI ( $BMI_{Q3}$ ) and preceding WCs (i.e.,  $BMI_{Q3} * WC_{Q3-Q1}$  or  $BMI_{Q3} * WC_{Q3-Q2}$ ). We excluded 464 and 82 women with missing values for  $WC_{Q3-Q1}$  and  $WC_{Q3-Q2}$ , respectively, resulting in respective analytical samples of 1,230 and 1,612 women.

To evaluate the influence of extreme WC values, we performed sensitivity analyses in which we assigned WC values that were under the 5<sup>th</sup> percentile and over the 95<sup>th</sup> percentile the values of the 5<sup>th</sup> and 95<sup>th</sup> percentiles, respectively. Additionally, we performed sensitivity analyses using the WC unit of BMI/year instead of kg/year.

### *White blood cell proportions*

Blood cell type composition affects gene expression profiles<sup>23</sup> and reticulocytes, erythrocytes, and white blood cells (WBCs) counts were not available for our study sample. However, we estimated the proportions of 22 populations of WBCs in the samples using an *in-silico* gene expression deconvolution method CIBERSORT, and the LM22 signature matrix<sup>24</sup>. To distinguish changes in gene expression related to WBC composition from those related to BMI, we adjusted for WBC proportions that were significantly associated with  $BMI_{Q3}$  according to the Kruskal-Wallis test, and with overall gene expression data according to the *'global test'* from the Bioconductor package *'global test'*.

### *Covariates*

We assessed the distribution of the following self-reported covariates by BMI<sub>Q3</sub> categories: age at Q3 (years), physical activity/day at either Q2 or Q1 (not available from Q3; stated on a scale of 1-10, where 1 represented ‘not active’ and 10 represented ‘extremely active’), total energy intake at either Q2 or Q1 (not available from Q3; kJ/day), and smoking status at Q3 (current/former/never smokers). We considered laboratory batch (laboratory plates) and sample storage time as technical covariates. We employed two adjustment models (minimally-adjusted and fully-adjusted) for all analyses. Minimally-adjusted models included technical covariates only, while fully-adjusted models included technical covariates, selected WBC proportions (described above), age, and smoking status at Q3. In categorical WC analyses, we additionally adjusted for BMI<sub>Q1</sub> in the fully-adjusted models. Further, sensitivity analyses related to BMI analyses were additionally adjusted for physical activity and total energy intake in the fully-adjusted models.

### *Functional enrichment analyses*

We investigated the biological functions of the BMI- and WC-associated DEGs identified in the fully-adjusted models by functional enrichment analyses. Analyses were conducted separately for over-expressed ( $\log_2$  fold-change ( $\logFC$ )>0) and under-expressed genes ( $\logFC$ <0); using the Bioconductor packages ‘*clusterProfiler*’ and ‘*ReactomePA*’ of gene ontology (GO) biological processes (BP), GO molecular functions (MF), GO cellular components (CC), Kyoto encyclopaedia of genes and genomes (KEGG) pathways, and REACTOME pathways.

### *Quantitative replication*

To assess whether our results were in line with previous results or novel findings, we compared our BMI-associated DEG results to results of analyses in external/independent transcriptomic datasets in whole-blood and other relevant tissues.

## **Results**

BMI<sub>Q3</sub> distribution in our study sample was 751 (45%), 622 (38%), and 280 (17%) for normal-weight, overweight, and obesity, respectively. There were no substantial differences in the distribution of most covariates across BMI<sub>Q3</sub> categories, but women with obesity were older,

and normal-weight women reported the highest mean physical activity level and total energy intake (Figure 2, Table S1). WC category distribution was 263 (25%), 138 (13%), and 14 (1%) for CSW, CWG, and CWL; and 396 (38%), 91 (9%), 101 (10%), and 32 (3%) for FWG, FWL, RWG, and RWL, respectively.

### **Estimated white blood cell proportions**

Among the estimated WBC proportions, naive B cells, memory B cells, naive CD4 T cells, and memory activated CD4 T cells were significantly associated with both BMI<sub>Q3</sub> category and overall gene expression (Table S2). Women with obesity had slightly higher mean proportions of naive B cells and lower mean proportions of the three other cell types listed, compared to women with normal-weight (Figure S1).

### **Body mass index-associated differentially expressed genes**

Table 1 presents the number of BMI-associated DEGs in the minimally- and fully-adjusted models. In the fully-adjusted model of categorical BMI analyses, the top-ranked gene (i.e., the gene with the lowest FDR-adjusted p-value) in the obesity-vs-normal-weight comparison (Figures 3A, 3D) and the obesity-vs-overweight comparison (Figures 3B, 3E) was *FAM46C* (renamed: *TTENT5C*) (logFC=0.86, FDR=6E-32). *FAM46C* (logFC=0.34, FDR=1E-45; Figure S2) was also the top-ranked gene in continuous BMI analyses. *FAM46C* expression was higher in women with higher BMI (Figure S3). In the overweight-vs-normal-weight comparison (Figures 3C, 3F), the top-ranked gene was *SLC45A3* (logFC=-0.28, FDR=1E-14), and its expression was lower in women with higher BMI (Figure S3).

The distributions of p-values varied across comparison groups, and the observed and expected distributions deviated the most for the obesity-vs-normal-weight comparison (Figure S4). Many DEGs overlapped in the minimally- and fully-adjusted models (2,080, 522, 580, and 2,705 overlapping DEGs in the obesity-vs-normal-weight, obesity-vs-overweight, overweight-vs-normal-weight comparisons, and continuous BMI analyses, respectively; Tables S3-S6). Further, results from sensitivity analyses, which were additionally adjusted for physical activity and total energy intake, did not alter the overall results (99%, 94%, 96%, and 99% of DEGs overlapped in the fully-adjusted models for the obesity-vs-normal-weight, obesity-vs-overweight, overweight-vs-normal-weight comparisons, and continuous BMI analyses, respectively); and the logFC for the top-ranked *FAM46C* changed  $\leq 2\%$ , while it did not change for *SLC45A3* (results not shown). Fully-adjusted categorical BMI analyses revealed a

cumulative total of 169 over-expressed and 72 under-expressed genes that overlapped in the different comparisons.

Further, 525 DEGs overlapped across all models testing association with BMI (obesity-vs-normal-weight, overweight-vs-normal-weight comparisons, and continuous BMI analyses; both adjustment models; Tables S3-S6). The overall gene expression of these 525 DEGs largely clustered according to BMI status (Figure S5). Among the 50 genes with the lowest p-values in the same models, 33 DEGs overlapped.

### **Weight change-associated differentially expressed genes**

No WC-associated DEGs were identified in any of the categorical WC analyses, be it the minimally-adjusted model, the fully-adjusted model, or the sensitivity analyses restricted to women with <1 year between Q2 and Q3 (N=657). However, a few DEGs were identified in the WC-BMI interaction analyses (Table 1). In the first interaction model ( $BMI_{Q1orQ2} * WC$ ), the main effect of  $WC_{Q3-Q1}$  or  $WC_{Q3-Q2}$  had 3 and 168 overlapping DEGs between minimally- and fully-adjusted models, respectively (Table S7); in the second interaction model ( $BMI_{Q3} * WC$ ) they had 5 and 1 overlapping DEGs, respectively (Table S8). The interaction effect of BMI and WC was not significant in the first interaction model (Tables: 1, S7), but it was significant in the second, indicating DEGs for each 1-unit increase in  $WC_{Q3-Q1}$  or  $WC_{Q3-Q2}$ , but only among women with obesity at Q3, (Tables: 1, S8). The top-ranked genes from the interaction effect of BMI and WC in the second interaction model were *CECR6* (renamed: *TMEM121B*) ( $\log FC=0.19$ ,  $FDR=9.91E-03$ ; Figure 4A) for  $WC_{Q3-Q1}$ , and *STT3A* for  $WC_{Q3-Q2}$  ( $\log FC=-0.09$ ,  $FDR=1.35E-02$ ; Figure 4B). All DEGs identified in the minimally-adjusted second interaction model overlapped with those in the fully-adjusted model (Table S8). Among the 169 DEGs from the main effect of  $WC_{Q3-Q2}$  in the first interaction model (fully-adjusted), 21 (12%) overlapped with the 525 DEGs across all BMI-models (Table S7). The overall gene expression of the 169 DEGs did not show apparent clustering in relation to WC (Figure S6), which could be because of the low  $\log FC$ -values observed for these genes. Differing trends in expression of the top-ranked genes across  $BMI_{Q3}$  categories could indicate slightly increasing expression with increasing weight from Q1 to Q3 for *CECR6* (Figures S7:A-C) and decreasing expression with increasing weight from Q2 to Q3 for *STT3A* (Figures S7:D-F) for women with obesity. The sensitivity analyses for WC-BMI interaction analyses, one that replaced extreme WC values with values of the 5<sup>th</sup> and 95<sup>th</sup> percentiles, and one that included WC as BMI/year, rendered the same results as the fully-adjusted models (results not shown). There were high correlations

(>0.93) between WC variables measured in kg/year and BMI/year. Further p-values were distributed uniformly in all comparisons of WC as a continuous metric (Figure S8).

### **Functional enrichment analyses**

Over-expressed genes identified in the categorical BMI analyses were enriched for terms largely related to metabolic and catabolic processes, cellular response to toxic substances, erythrocyte homeostasis, and development (GO-BP); cellular oxidant detoxification and blood protein bindings (GO-MF, GO-CC, and KEGG); ribosome structure and haemoglobin complexes (GO-CC); and metabolism of amino acids and translation (REACTOME). Under-expressed genes were enriched in fewer categories but included peptide and antigen bindings (GO-MF, KEGG); lysosome and vacuoles components (GO-CC); and asthma, tuberculosis, and influenza A (KEGG) (Figure 5, Table S9). The enriched terms for BMI-associated DEGs in continuous BMI analyses (Table S10, Figure S9) and for the 525 DEGs across all BMI models (Table S11, Figure S10) were largely like those indicated for the categorical BMI analyses. The enriched terms for the 33 DEGs overlapping across the 50 genes with lowest p-values in BMI-models were related to erythrocytes functions (Table S12).

There were few WC-associated DEGs in the WC-BMI interaction analyses. Still, the terms signalling receptor and molecular transducer activities (GO-MF) were overrepresented by 4/9 over-expressed genes identified in the WC<sub>Q3-Q1</sub> interaction model (Table S13).

### **Quantitative replication**

When comparing the 525 DEGs across BMI-models with DEGs reported in similar published studies in whole-blood (3,762<sup>20</sup> and 144<sup>10</sup> DEGs), in PBMCs (1,864 DEGs<sup>21</sup>), and in adipose tissue (only males, 2,936 DEGs<sup>9</sup>), there were 396 (75.42%), 19 (3.6%), 77 (14.66%), and 93 (17.71%) overlapping DEGs, respectively. The corresponding effect directions were 99.74%, 100%, 85.71%, and 25.8% overlapping, respectively. Further, among the 3,106 DEGs in the continuous BMI analyses (fully-adjusted), 1,552 (49.96%), 42 (1.35%), 337 (10.84%), and 538 (17.32%) DEGs overlapped with the above mentioned studies (of which 97.29%, 95.23%, 73.59%, and 39.77% had corresponding effect directions), respectively (Table S14). Finally, the effect estimates for BMI association showed strong positive correlations with those from previous studies in whole-blood<sup>10,20</sup> and PBMCs<sup>21</sup>, but negative correlation with those from adipose tissue<sup>9</sup> (Figure S11).



## Discussion

This was the first study to extensively investigate the association between blood gene expression and 1) current BMI and 2) past WCs in a large sample of women. We showed that blood gene expression is a good reflection of current BMI (here represented by BMI<sub>Q3</sub>, i.e., at blood collection), but not of past WC. BMI<sub>Q3</sub> was clearly associated with blood gene expression, as >2,000 DEGs were identified in the obesity-vs-normal-weight comparison. Further, >700 DEGs were identified in the obesity-vs-overweight and overweight-vs-normal-weight comparisons. Across the models testing associations with BMI, 525 DEGs overlapped. *FAM46C* was the top-ranked gene in all BMI analyses, except in the overweight-vs-normal-weight comparison, where the top-ranked gene was *SLC45A3*. Our results show that *FAM46C* was positively associated with increasing BMI, whereas *SLC45A3* was negatively associated. In contrast, when focusing on WC within the preceding 7 years (range=5.5-14 years) or 1 year (range=<1 month-9 years) and accounting for interactions with BMI categories, we found limited associations with current gene expression, as between 1 and 169 genes were associated to the main effect of past WCs and between 0 and 9 genes were associated with the interaction effect of past WC and current BMI.

Our categorical WC analyses did not reveal any DEGs, but after introducing an interaction with BMI categories (WC-BMI interaction analyses), up to 169 DEGs were identified. The assumption for the two interaction models was that current BMI was a stronger predictor of current gene expression than past WC/BMI, thus the second interaction model (BMI<sub>Q3</sub>\*WC) would be more accurate than the first (BMI<sub>Q1orQ2</sub>\*WC). Our results agreed with these predictions, as the interaction effect of WC and BMI was not significant in the first interaction model, whereas up to 9 genes were significant in the second interaction model. The expression of top-ranked genes from the WC-BMI interaction analyses could indicate a positive and negative association of WC<sub>Q3-Q1</sub> and WC<sub>Q3-Q2</sub> among women with obesity for *CECR6* and *STT3A*, respectively; however, the trend was not very strong. The 21 DEGs among the WC-models (N=169) that overlapped with 525 DEGs across all BMI-models indicate that past WC were represented in current BMI models to some extent, however, they were not among the top-ranked genes.

Comparing the 525 DEGs across BMI-models and 3,106 DEGs from continuous BMI analyses (fully-adjusted) with similar previous studies conducted in whole-blood<sup>10,20</sup> and PBMCs<sup>21</sup> showed that these and our results were largely consistent, although our top-ranked genes

*FAM46C* and *SLC45A3* were only identified in Homuth et al.<sup>20</sup>. The overlap with the study in adipose tissue<sup>9</sup> was less than expected by chance (p-value=4.1E-03; OR=0.58; Table S14). Further, the correlation between the estimates for associations being strongly positive with other studies in whole-blood<sup>10,20</sup>, and in PBMCs<sup>21</sup>, but negative in adipose tissue<sup>9</sup> indicate that systemic signals in blood related to BMI differ from those in adipose tissues. Thus, DEGs in whole-blood related to BMI in women in this study could be generalizable to both sexes and other blood samples but gene expression profiles are differently regulated in adipose tissue.

Functional enrichment analyses of BMI-associated DEGs indicated a broad range of functions in enriched pathways. For genes over-expressed in women with obesity, terms describing various catabolic (e.g., cofactor catabolic processes) and metabolic processes (e.g., hydrogen peroxide, heme, tetrapyrrole metabolic processes), as well as erythrocyte homeostasis, haemoglobin binding, and ribosome structures were enriched. These findings align with previous studies in whole-blood<sup>10,20</sup> and PBMCs<sup>21</sup>. The enriched terms erythrocyte differentiation, myeloid cell homeostasis, erythrocyte homeostasis, heme biosynthetic/metabolic process indicate overexpression of genes in processes in erythrocytes or their precursors (reticulocytes)<sup>25</sup>. Notably, the BMI-associated top-ranked gene, *FAM46C*, and several other top-ranked genes (*HBD*, *GYPB*, and *ALAS2*) are primarily expressed in bone marrow, blood, and early erythroid cells<sup>25</sup>. Erythrocyte indices have been observed as positively associated with obesity<sup>26,27</sup>, and could be explained by proliferation reticulocytes in the bone marrow<sup>28,29</sup> induced by the hormone leptin, released by bone marrow-resident adipocytes. In contrast, erythrocytes in people with obesity have a shorter half-life in circulation due to impaired insulin resistance and pronounced oxidative stress resulting from hyperglycemia<sup>30</sup>. This reticulocyte-erythrocyte ratio shift is expected to be reflected in the whole-blood transcriptome because reticulocytes are also transcriptionally active<sup>20</sup>. Therefore, DEGs identified in this study likely reflect a shift in the reticulocyte-erythrocyte ratio associated with higher BMI. However, as erythrocyte/reticulocyte cell counts were not available in NOWAC, such adjustments in the statistical analyses were not feasible.

Enriched terms for over-expressed genes further included the terms peptide chain elongation and eukaryotic translation termination/elongation which appeared related to protein synthesis<sup>31</sup>. In line with these terms, another study investigating gene expression related to BMI in whole-blood observed ribosome and protein synthesis pathways as top-ranked among women<sup>10</sup>. These enriched terms indicated physiological changes previously observed for people with obesity, e.g., higher levels of oxidative stress<sup>30,32</sup>, haemoglobin<sup>33,34</sup>, and disrupted protein synthesis<sup>35</sup>.

Enriched terms among under-expressed genes included antigen binding, processing and presentation, peptide binding, and TNF signalling pathways, which suggest there could be altered blood immune responses in women with obesity, something that has been observed among participants with obesity in previous transcriptomic studies<sup>9-11</sup>. Furthermore, altered immune response/function (e.g., related to the terms influenza, asthma, antigen binding) in persons with obesity could explain previously observed associations between obesity and increased risk of co-morbidities and infectious diseases, like influenza and COVID-19, and increased viral shedding and transmission<sup>36,37</sup>. Overall, general metabolism and blood processes were enriched, which likely reflects the broad influence of BMI on systemic gene expression.

The evaluation of past WCs and gene expression profiles in blood was novel but indicated few DEGs and thus related biological functions were not strongly indicated. Still, the over-expressed genes *RBPI/FZD2/OPRL1/CD14* indicated a relation between past WC and genes involved in signal transduction.

In general, our results indicate that current BMI and past WC have, respectively, large and small effects on blood gene expression. This could be expected, as blood gene expression represents a snapshot, and past exposures such as WCs are generally not strongly reflected<sup>22,38</sup>. Still, until now, no study had investigated the association between blood gene expression and past WC. Previous studies reported that weight reduction in individuals with obesity after diet interventions was associated with gene expression profiles in adipose tissue before and after the interventions<sup>13-19</sup>. However, as follow-up time in these studies (4 weeks to 9 months) was shorter than the time intervals in our study (range= $\leq$ 1 month-14 years), the WCs we observed could be too far in the past to have a major influence on blood gene expression. Still, sensitivity analyses restricted to women with  $<1$  year between Q2 and Q3 did not show any significant DEGs. Future studies focusing on systemic signatures related to WCs should likely include blood samples taken within months of the WC occurring for transcriptomic signals to be detectable. DEGs related to obesity/WC might be expected in adipose/muscle tissue, but a study has demonstrated that blood samples can be another informative, accessible tissue to explore circulating features of the state of obesity<sup>10</sup>.

We observed an association between current BMI and naive B cells, memory B cells, naive CD4 T cells, and memory-activated CD4 T cells, possibly because BMI and body weight have been positively correlated with WBC counts in apparently healthy young adults (higher in women)<sup>39</sup>. As skewed WBC proportions due to differences in BMI could have influenced our

BMI analyses, we included these estimated cell proportions in our fully-adjusted models<sup>40</sup>. The estimated proportions of WBCs in our study deviated from the expected range, but this deviation has been observed also in other recent studies based on the NOWAC postgenome cohort<sup>41,42</sup>. This indicates a bias which could be explained by the deconvolution technique or data pre-processing<sup>41,43</sup>. Still, absolute differences in estimated WBCs across BMI categories were modest, and the top-ranked genes identified in our models were very similar, indicating that these genes were not substantially influenced by distributions of WBCs. Lastly, erythrocytes/reticulocytes counts were not available, and their adjustment was thus not possible.

The main strength of this study was the large study sample (1,694 women). Indeed, previous studies on BMI and blood gene expression have been rather small (32-190 participants)<sup>8-11</sup>, with the exception of one large population-based cohort study (1,048 participants, 53% women)<sup>20</sup>. Another strength was that our study was based on repeated measurements, thus we were able to generate BMI and WC variables for all women at different time points. Still, individual intervals varied, and we standardised WC by dividing it by individual time differences. Additionally, this study was based on self-reported questionnaire information, which could be influenced by measurement and recall bias. A validation study of self-reported BMI among NOWAC study participants found a slight, but statistically significant, under-reporting of weight and self-reported BMI, especially among women with overweight and obesity, but they concluded that, for middle-aged Norwegian women, self-reported weight and height provide a valid ranking of BMI<sup>44</sup>. The present study included only women who had never been diagnosed with cancer, but we cannot disregard the influence of other common chronic diseases. Furthermore, the current cross-sectional study results only represent snapshots of blood gene expression and cannot indicate causality. Lastly, although RNA-sequencing has become a routinely used technology, results from microarray technology, like those in this study, are still reliable and overall comparable to RNA-sequencing results<sup>45</sup>, although non-coding RNAs and splice variants cannot be detected. Future studies could validate gene expression findings, especially related to WC, using alternative targeted technologies (e.g., qPCR or NanoString), or investigate cell-type specific gene expression using single cell RNA-sequencing, but that would require new sample collection.

## **Conclusion**

Many BMI-associated DEGs, but few WC-associated DEGs were identified in blood of women in Norway. This is the first study to our knowledge to conclude that blood gene expression

reflects current BMI more than past WCs. The biological functions of BMI-associated DEGs were linked to metabolism, erythrocyte, oxidative stress, and immune processes. These likely reflect systemic impacts of obesity, especially reticulocyte-erythrocyte ratio shifts in blood, as these functions coincide with its known physiological effects. Further, the biological functions of WC-associated DEGs were linked to signal transduction. This knowledge is relevant for further research related to the health effects of BMI and WC, especially those that focus on blood-based markers.

## **Acknowledgements**

The authors extend their gratitude to **all the women** participating in the NOWAC study. We thank **Bente Augdal** and all other personnel responsible for the administration of the data collection and the biobank. We thank **Tonje Braaten, Arne Bastian Wiik, Marko Lukic,** and **Nikita Shvetsov** for providing the data set. Special thanks to **Jo Inge** for helping with data wrangling and **Trudy Perdrix-Thoma** for English language editing.

## **Data availability**

Data cannot be shared publicly because of local and national ethical and security policies. Data access for researchers will be conditional on adherence to both the data access procedures of the NOWAC study and the UiT The Arctic University of Norway (contact: Tonje Braaten <[tonje.braaten@uit.no](mailto:tonje.braaten@uit.no)>) in addition to approval from the local ethical committee.

## **Reporting checklist**

We have used *STrengthening the REporting of Genetic Association studies (STREGA)* reporting recommendations, extended from STROBE Statement, as a reporting checklist, and provided it as a separate supplementary file in page 52.

## References

1. World Health Organization. Obesity and overweight: [www.who.int](http://www.who.int); 2021 [Accessed 23 July, 2021]. <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
2. Herrera BM, Keildson S, Lindgren CM. Genetics and epigenetics of obesity. *Maturitas*. 2011;69(1):41-49.
3. Keith SW, Redden DT, Katzmarzyk PT, Boggiano MM, Hanlon EC, Benca RM, et al. Putative contributors to the secular increase in obesity: exploring the roads less traveled. *Int J Obes (Lond)*. 2006;30(11):1585-1594.
4. Symonds ME, Budge H, Frazier-Wood AC. Epigenetics and obesity: a relationship waiting to be explained. *Human Heredity*. 2013;75(2-4):90-97.
5. da Silva M, Weiderpass E, Licaj I, Lissner L, Rylander C. Excess body weight, weight gain and obesity-related cancer risk in women in Norway: the Norwegian Women and Cancer study. *Br J Cancer*. 2018;119(5):646-656.
6. World Cancer Research Fund. Body fatness and weight gain and the risk of cancer. World Cancer Research Fund International London; 2018.
7. Kim K, Zakharkin SO, Allison DB. Expectations, validity, and reality in gene expression profiling. *Journal of clinical epidemiology*. 2010;63(9):950-959.
8. de Luis DA, Almansa R, Aller R, Izaola O, Romero E. Gene expression analysis identifies a metabolic and cell function alterations as a hallmark of obesity without metabolic syndrome in peripheral blood, a pilot study. *Clin Nutr*. 2018;37(4):1348-1353.
9. Romn T, Volkov P, Gillberg L, Kokosar M, Perfilyev A, Jacobsen AL, et al. Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Human Molecular Genetics*. 2015;24(13):3792-3813.
10. Ghosh S, Dent R, Harper ME, Gorman SA, Stuart JS, McPherson R. Gene expression profiling in whole blood identifies distinct biological pathways associated with obesity. *BMC Med Genomics*. 2010;3:56.
11. Jung UJ, Seo YR, Ryu R, Choi MS. Differences in metabolic biomarkers in the blood and gene expression profiles of peripheral blood mononuclear cells among normal weight, mildly obese and moderately obese subjects. *Br J Nutr*. 2016;116(6):1022-1032.
12. 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):81-86.
13. Bouchard L, Rabasa-Lhoret R, Faraj M, Lavoie M-È, Mill J, Pérusse L, et al. Differential epigenomic and transcriptomic responses in subcutaneous adipose tissue between low and high responders to caloric restriction. *The American journal of clinical nutrition*. 2010;91(2):309-320.
14. Capel F, Klimčáková E, Viguerie N, Roussel B, Vítková M, Kováčiková M, et al. Macrophages and adipocytes in human obesity: adipose tissue gene expression and insulin sensitivity during calorie restriction and weight stabilization. *Diabetes*. 2009;58(7):1558-1567.
15. Clément K, Viguerie N, Poitou C, Carette C, Pelloux V, Curat CA, et al. Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *The FASEB Journal*. 2004;18(14):1657-1669.
16. Dahlman I, Linder K, Arvidsson Nordström E, Andersson I, Lidén J, Verdich C, et al. Changes in adipose tissue gene expression with energy-restricted diets in obese women. *The American journal of clinical nutrition*. 2005;81(6):1275-1285.

17. Johansson LE, Danielsson AP, Parikh H, Klintonberg M, Norström F, Groop L, et al. Differential gene expression in adipose tissue from obese human subjects during weight loss and weight maintenance. *The American journal of clinical nutrition*. 2012;96(1):196-207.
18. Kolehmainen M, Salopuro T, Schwab U, Kekäläinen J, Kallio P, Laaksonen D, et al. Weight reduction modulates expression of genes involved in extracellular matrix and cell death: the GENOBIN study. *Int J Obes (Lond)*. 2008;32(2):292-303.
19. Ma´rquez-Quinones A, Mutch DM, Debard C, Wang P, Combes M, Roussel B, et al. Adipose tissue transcriptome reflects variations between subjects with continued weight loss and subjects regaining weight 6 mo after caloric restriction independent of energy intake. *The American journal of clinical nutrition*. 2010;92(4):975-984.
20. Homuth G, Wahl S, Muller C, Schurmann C, Mader U, Blankenberg S, et al. Extensive alterations of the whole-blood transcriptome are associated with body mass index: results of an mRNA profiling study involving two large population-based cohorts. *BMC Med Genomics*. 2015;8:65.
21. Vargas LB, Lange LA, Ferrier K, Aguet F, Ardlie K, Gabriel S, et al. Gene expression associations with body mass index in the Multi-Ethnic Study of Atherosclerosis. *Int J Obes (Lond)*. 2022:1-8.
22. Baiju N, Sandanger TM, Sætrom P, Nøst TH. Gene expression in blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer (NOWAC) postgenome cohort. *Sci*. 2021;11(1):1-13.
23. Joehanes R, Johnson AD, Barb JJ, Raghavachari N, Liu P, Woodhouse KA, et al. Gene expression analysis of whole blood, peripheral blood mononuclear cells, and lymphoblastoid cell lines from the Framingham Heart Study. *Physiol Genomics*. 2012;44(1):59-75.
24. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*. 2015;12(5):453-457.
25. UniProtKB/Swiss-Prot. Find Your Protein: UniProt.org; 2023 [Accessed Feb 23, 2023]. <https://www.uniprot.org/>.
26. Bird JK, Ronnenberg AG, Choi S-W, Du F, Mason JB, Liu Z. Obesity is associated with increased red blood cell folate despite lower dietary intakes and serum concentrations. *The Journal of nutrition*. 2015;145(1):79-86.
27. Kohsari M, Moradinazar M, Rahimi Z, Najafi F, Pasdar Y, Moradi A, et al. Association between RBC indices, anemia, and obesity-related diseases affected by body mass index in Iranian Kurdish population: Results from a cohort study in Western Iran. *International Journal of Endocrinology*. 2021;2021.
28. Umemoto Y, Tsuji K, Yang F-C, Ebihara Y, Kaneko A, Furukawa S, et al. Leptin stimulates the proliferation of murine myelocytic and primitive hematopoietic progenitor cells. *Blood, The Journal of the American Society of Hematology*. 1997;90(9):3438-3443.
29. Trottier MD, Naaz A, Kacynski K, Yenumula PR, Fraker PJ. Functional capacity of neutrophils from class III obese patients. *Obesity (Silver Spring)*. 2012;20(5):1057-1065.
30. Hurrle S, Hsu WH. The etiology of oxidative stress in insulin resistance. *Biomedical journal*. 2017;40(5):257-262.
31. QuickGO. GO annotations: EMBL-EBI; 2023 [Accessed February 16, 2023]. <https://www.ebi.ac.uk/QuickGO/annotations>.
32. Goyal R, Singhai M, Faizy AF. Glutathione peroxidase activity in obese and nonobese diabetic patients and role of hyperglycemia in oxidative stress. *Journal of mid-life health*. 2011;2(2):72.
33. Akter R, Nessa A, Sarker D, Yesmin M. Effect of Obesity on Hemoglobin Concentration. *Mymensingh Medical Journal: MMJ*. 2017;26(2):230-234.



34. Gozkaman A, Okuturlar Y, Mert M, Harmankaya O, Kumbasar A, editors. The relationship between haemoglobin and BMI in overweight and obese patients. *Endocrine Abstracts*; 2015: Bioscientifica.
35. Rui L. A link between protein translation and body weight. *The Journal of clinical investigation*. 2007;117(2):310-313.
36. De Frel DL, Atsma DE, Pijl H, Seidell JC, Leenen PJ, Dik WA, et al. The impact of obesity and lifestyle on the immune system and susceptibility to infections such as COVID-19. *Frontiers in nutrition*. 2020;279.
37. Honce R, Schultz-Cherry S. Impact of obesity on influenza A virus pathogenesis, immune response, and evolution. *Front*. 2019:1071.
38. Olsen KS, Lukic M, Borch KB. Physical activity and blood gene expression profiles: the Norwegian Women and Cancer (NOWAC) Post-genome cohort. *BMC Res Notes*. 2020;13(1):1-6.
39. Ghannadiasl F. Associations between white blood cells count and obesity in apparently healthy young adults. *Nutrition & Food Science*. 2020.
40. 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):550-560.
41. Jareid M, Snapkov I, Holden M, Busund L-TR, Lund E, Nøst TH. The blood transcriptome prior to ovarian cancer diagnosis: A case-control study in the NOWAC postgenome cohort. *Plos one*. 2021;16(8):e0256442.
42. Nøst TH, Holden M, Dønnem T, Bøvelstad H, Rylander C, Lund E, et al. Transcriptomic signals in blood prior to lung cancer focusing on time to diagnosis and metastasis. *Sci*. 2021;11(1):7406.
43. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*. 2020;11(1):5650.
44. Skeie G, Mode N, Henningsen M, Borch KB. Validity of self-reported body mass index among middle-aged participants in the Norwegian Women and Cancer study. *Clinical epidemiology*. 2015;7:313.
45. Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Medical science monitor basic research*. 2014;20:138.

**Table 1.** Number of DEGs identified in BMI and WC analyses.

		Minimally-adjusted models <sup>1</sup>			Fully-adjusted models <sup>2</sup>		
		DEGs (FDR ≤0.05)	Over-expressed (logFC>0)	Under-expressed (logFC<0)	DEGs (FDR ≤0.05)	Over-expressed (logFC>0)	Under-expressed (logFC<0)
<b>BMI models</b>	<b>Obesity-vs-Normal-weight</b>	<b>2,294</b>	1,011	1,283	<b>2,394</b>	1,057	1,337
	<b>Obesity-vs-Overweight</b>	<b>553</b>	370	183	<b>769</b>	475	294
	<b>Overweight-vs-Normal-weight</b>	<b>629</b>	285	344	<b>768</b>	315	453
	<b>BMI<sup>3</sup></b>	<b>2,970</b>	1,257	1,713	<b>3,106</b>	1,293	1,813
<b>WC models</b>							
<b>First interaction model (BMI<sub>Q1orQ2</sub>*WC)</b>							
	<b>WC<sup>4,5</sup><sub>Q3-Q1</sub></b>	<b>43</b>	34	9	<b>3</b>	1	2
	<b>BMI<sub>Q1</sub>-Overweight: WC<sup>4,6</sup><sub>Q3-Q1</sub></b>	<b>0</b>	0	0	<b>0</b>	0	0
	<b>BMI<sub>Q1</sub>-Obesity: WC<sup>4,6</sup><sub>Q3-Q1</sub></b>	<b>0</b>	0	0	<b>0</b>	0	0
	<b>WC<sup>4,5</sup><sub>Q3-Q2</sub></b>	<b>217</b>	31	186	<b>169</b>	23	146
	<b>BMI<sub>Q2</sub>-Overweight: WC<sup>4,5</sup><sub>Q3-Q2</sub></b>	<b>0</b>	0	0	<b>0</b>	0	0
	<b>BMI<sub>Q2</sub>-Obesity: WC<sup>4,6</sup><sub>Q3-Q2</sub></b>	<b>0</b>	0	0	<b>0</b>	0	0
<b>Second interaction model (BMI<sub>Q3</sub>*WC)</b>							
	<b>WC<sup>4,5</sup><sub>Q3-Q1</sub></b>	<b>17</b>	3	14	<b>5</b>	0	5
	<b>BMI<sub>Q3</sub>-Overweight: WC<sup>4,6</sup><sub>Q3-Q1</sub></b>	<b>0</b>	0	0	<b>0</b>	0	0
	<b>BMI<sub>Q3</sub>-Obesity: WC<sup>4,6</sup><sub>Q3-Q1</sub></b>	<b>27</b>	25	2	<b>9</b>	9	0
	<b>WC<sup>4,5</sup><sub>Q3-Q2</sub></b>	<b>4</b>	2	2	<b>1</b>	0	1
	<b>BMI<sub>Q3</sub>-Overweight: WC<sup>4,6</sup><sub>Q3-Q2</sub></b>	<b>0</b>	0	0	<b>0</b>	0	0
	<b>BMI<sub>Q3</sub>-Obesity: WC<sup>4,6</sup><sub>Q3-Q2</sub></b>	<b>1</b>	0	1	<b>1</b>	0	1

<sup>1</sup>Adjusted for laboratory batch (laboratory plates) and sample storage time

<sup>2</sup>Adjusted for minimally-adjusted model plus selected white blood cell proportions, age, and smoking status at Q3

<sup>3</sup>BMI was included in the model as a scaled continuous metric

<sup>4</sup>WC was included in the model as a scaled continuous metric

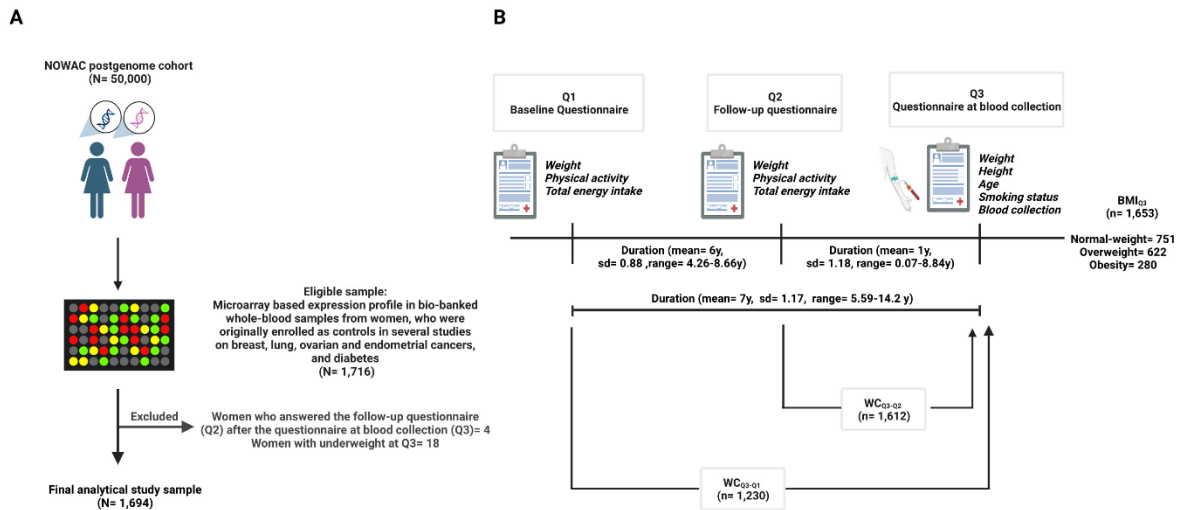
<sup>5</sup>These results represent those for the main effect of WCs in the interaction model

<sup>6</sup>These results represent those for the interaction effect of BMI and WCs in the interaction model

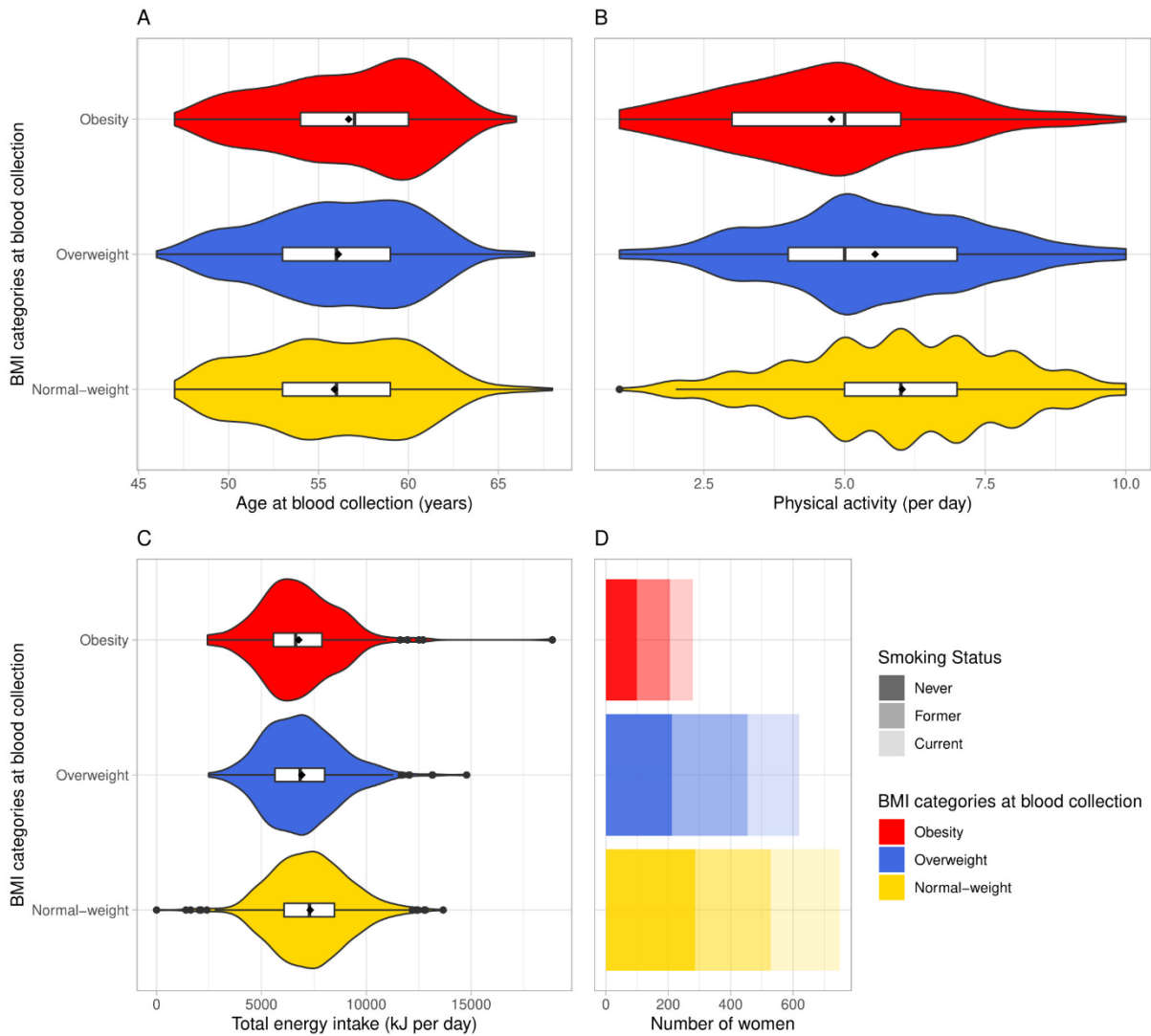
Note: No DEGs were observed in models comparing WC categories (CWG-vs-CSW, CWL-vs-CSW, FWG-vs-CSW, FWL-vs-CSW, RWG-vs-CSW, and RWL-vs-CSW), hence they are not presented here.

DEGs: differentially expressed genes; BMI: body mass index; WC: weight change; logFC: log<sub>2</sub> fold-change, FDR: false discovery rate; CSW: consistent stable weight; CWG: consistent weight gain; CWL: consistent weight loss; FWG: former weight gain; FWL: former weight loss; RWG: recent weight gain; RWL: recent weight loss; Q1: baseline questionnaire; Q2: follow-up questionnaire; Q3: questionnaire at blood collection.

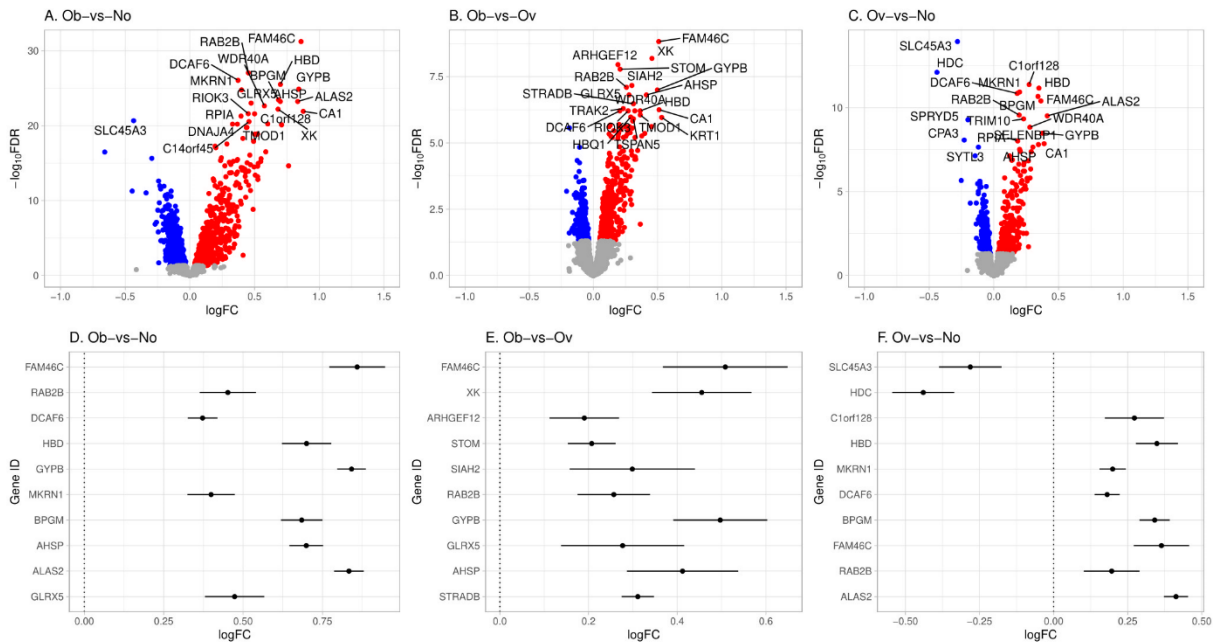
# Main Figures



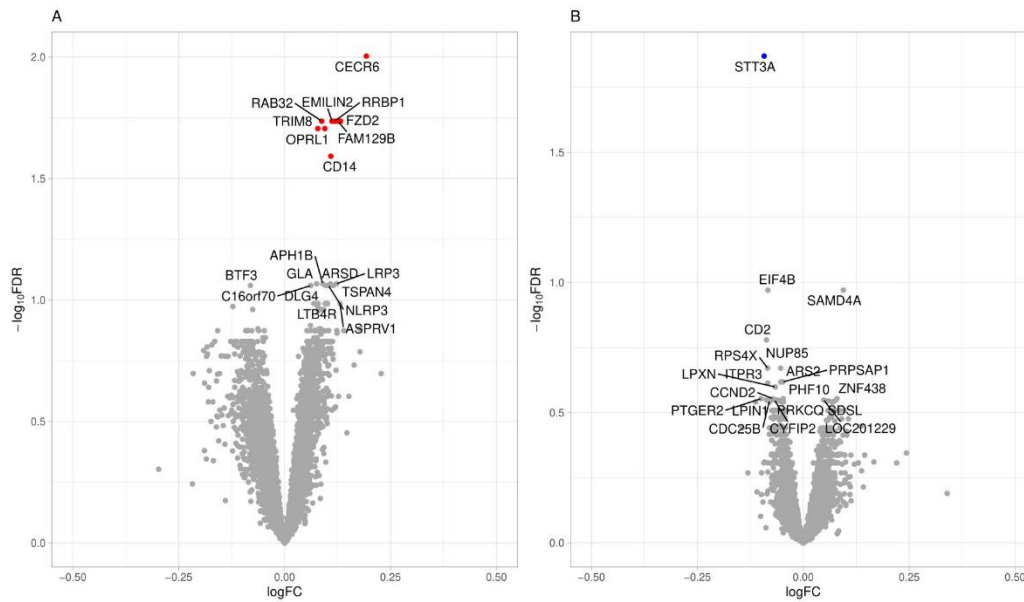
**Figure 1. Flowchart of study sample (A) and timeline of questionnaires (B) in the NOWAC study** (Created with [BioRender.com](https://www.biorender.com)). Q1: baseline questionnaire; Q2: follow-up questionnaire, Q3: questionnaire at blood collection; sd: standard deviation; BMI<sub>Q3</sub>: body mass index categories at Q3; WC<sub>Q3-Q1</sub>: weight change between Q1 and Q3; WC<sub>Q3-Q2</sub>: weight change between Q2 and Q3; y: year(s); NOWAC: The Norwegian Women and Cancer study.



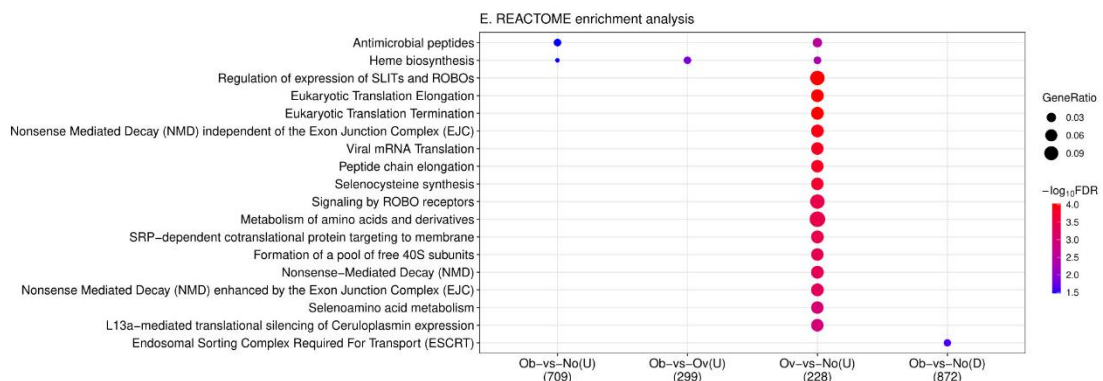
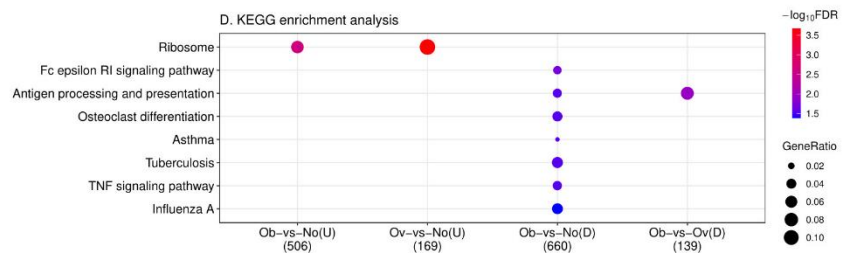
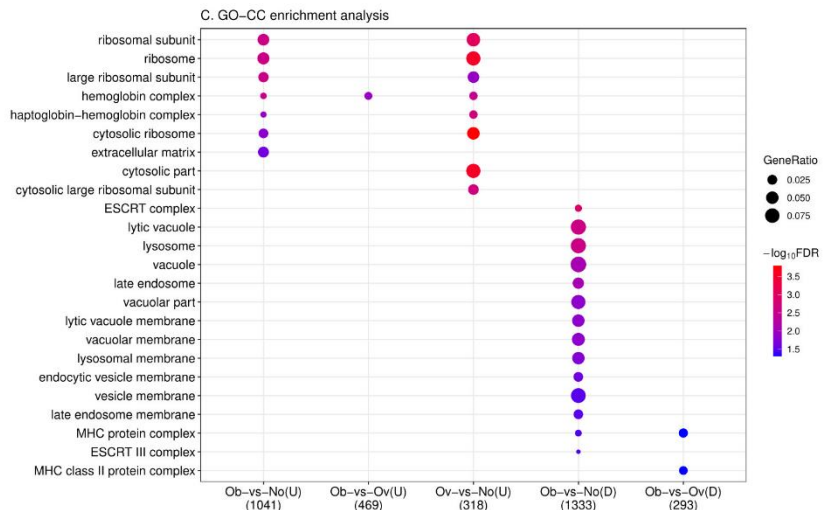
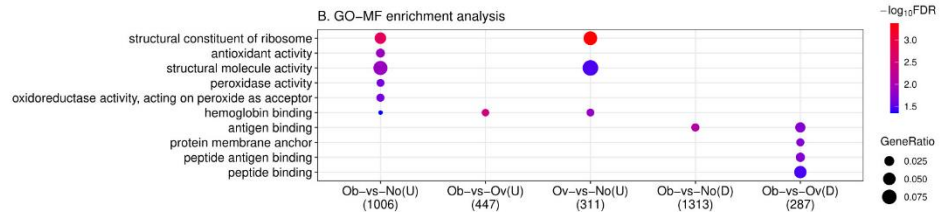
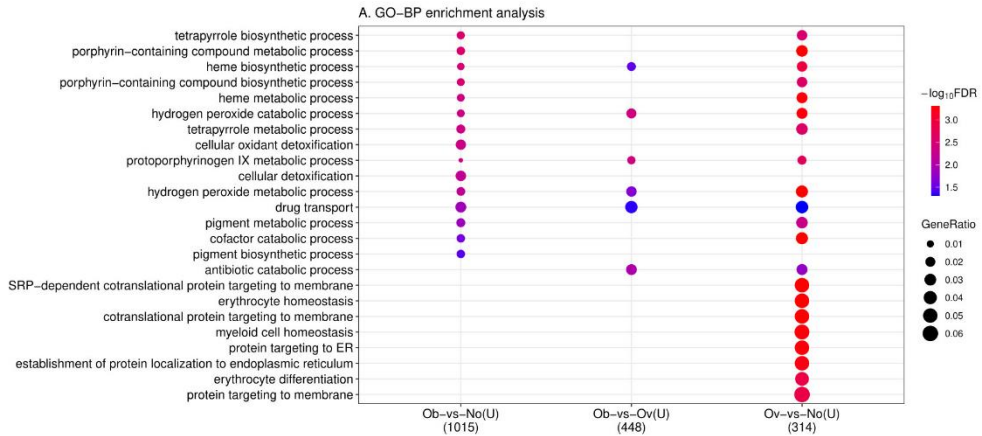
**Figure 2. BMI category at blood collection by age at blood collection (years) (A), physical activity (per day) (B), total energy intake (kJ per day) (C), and smoking status at blood collection (D).** In Figure A-C: The violin plots represent the kernel density estimates for women with obesity, overweight, and normal-weight. White boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, vertical bars inside the box represent the median, whiskers extend 1.5 times the length of the interquartile range to the right and left side of the 75<sup>th</sup> and 25<sup>th</sup> percentiles respectively, and outliers are represented as black dots. The black diamond-shaped dot represents the respective mean. In Figure D: the colour lightness represents smoking status (never, former, current smokers) at blood collection. Here, physical activity and total energy intake were from either Q2 or Q1 (not available for Q3). BMI: body mass index; Q1: baseline questionnaire; Q2: follow-up questionnaire, Q3: questionnaire at blood collection.



**Figure 3. Volcano plots of the test statistics for BMI categories at blood collection in fully-adjusted models from the tests of DEGs (A-C) and forest plots for the 10 top-ranked genes in tests of DEGs (D-F).** In volcano plots (A-C), red dots display over-expressed genes ( $\text{FDR} \leq 0.05$  and  $\log_2(\text{FC}) > 0$ ), blue dots display under-expressed genes ( $\text{FDR} \leq 0.05$  and  $\log_2(\text{FC}) < 0$ ), while grey dots display genes with  $\text{FDR} > 0.05$ ; and gene names displayed are the top 20 DEGs in the respective tests. In forest plots (D-F), the gene ID of DEGs are presented with the lowest FDR-adjusted p-values ranked from the top; the horizontal line for each gene represents their confidence interval; and the vertical dotted line represents the line of no difference. BMI: body mass index; DEGs: differentially expressed genes; FDR: false discovery rate;  $\log_2(\text{FC})$ :  $\log_2$  fold-changes; Ob-vs-No: comparison of women with obesity versus normal-weight; Ob-vs-Ov: comparison of women with obesity versus overweight; Ov-vs-No: comparison of women with overweight versus normal-weight.



**Figure 4. Volcano plots for the test statistics of DEGs in fully-adjusted models WC-BMI interaction analyses according to WC from Q1 to Q3 (A) and from Q2 to Q3 (B).** Red dots display over-expressed genes ( $\text{FDR} \leq 0.05$  and  $\log_2(\text{FC}) > 0$ ), blue dots display under-expressed genes ( $\text{FDR} \leq 0.05$  and  $\log_2(\text{FC}) < 0$ ), while grey dots display genes with  $\text{FDR} > 0.05$ ; and gene names displayed are the significant DEGs in the respective tests. Here, WC is modelled as scaled continuous metric. The results are of the interaction effect of WC and BMI from the second interaction model (i.e.,  $\text{BMI}_{\text{Q3}} * \text{WC}_{\text{Q3-Q1}}$  or  $\text{BMI}_{\text{Q3}} * \text{WC}_{\text{Q3-Q2}}$ ). BMI: body mass index; DEGs: differentially expressed genes; WC: weight change; FDR: False Discovery Rate;  $\log_2(\text{FC})$ :  $\log_2$  fold-changes; Q1: baseline questionnaire; Q2: follow-up questionnaire; Q3: questionnaire at blood collection,  $\text{WC}_{\text{Q3-Q1}}$ : weight change between Q1 and Q3;  $\text{WC}_{\text{Q3-Q2}}$ : weight change between Q2 and Q3.



**Figure 5. Summary of functional enrichment analyses for BMI categories at blood collection for over-expressed (Up-regulated) and under-expressed (Down-regulated) genes for the GO-BP (A), GO-MF (B), GO-CC (C), KEGG (D), and REACTOME pathway (E) databases.** The colour of the dots indicates  $-\log_{10}\text{FDR}$ , where red dots represent the most enriched categories (i.e., ones with the lowest  $-\log_{10}\text{FDR}$ ); the ‘GeneRatio’ indicates the proportion of genes overlapping between lists of DEGs and the genes in GO categories. BMI: body mass index; DEGs: differentially expressed genes; FDR: false discovery rate; logFC:  $\log_2$  fold-changes; GO: gene ontology; BP: biological processes; MF: molecular functions; CC: cellular components; KEGG: Kyoto Encyclopedia of Genes and Genomes; Ob-vs-No: comparison of women with obesity versus normal-weight; Ob-vs-Ov: comparison of women with obesity versus overweight; Ov-vs-No: comparison of women with overweight versus normal-weight; U: Over-expressed genes (Up-regulated,  $\text{FDR} \leq 0.05$ ,  $\log\text{FC} > 0$ ); D: Under-expressed genes (Down-regulated,  $\text{FDR} \leq 0.05$  and  $\log\text{FC} < 0$ ).

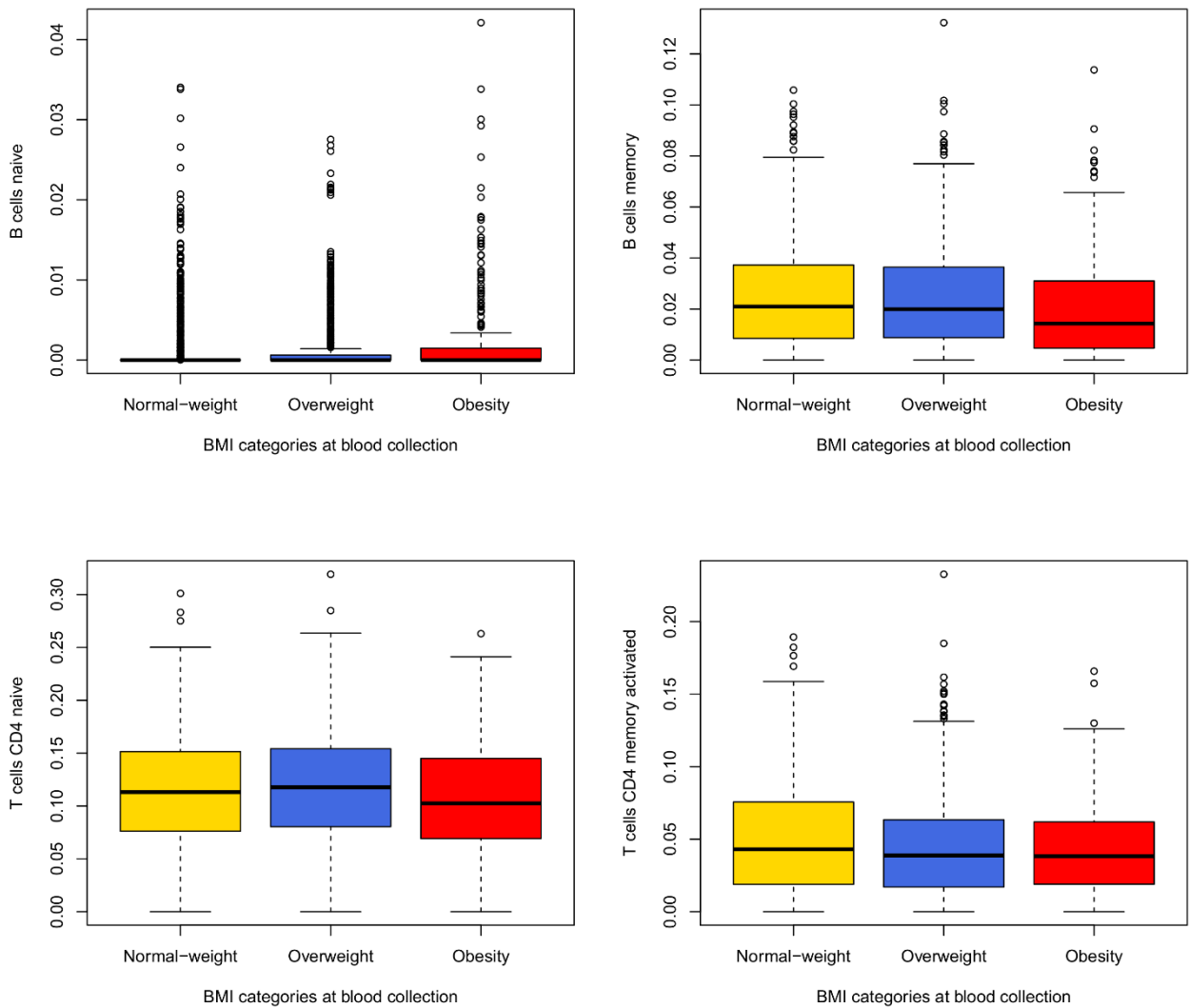


**Associations of gene expression in blood with BMI and weight changes  
among women in the NOWAC postgenome cohort**

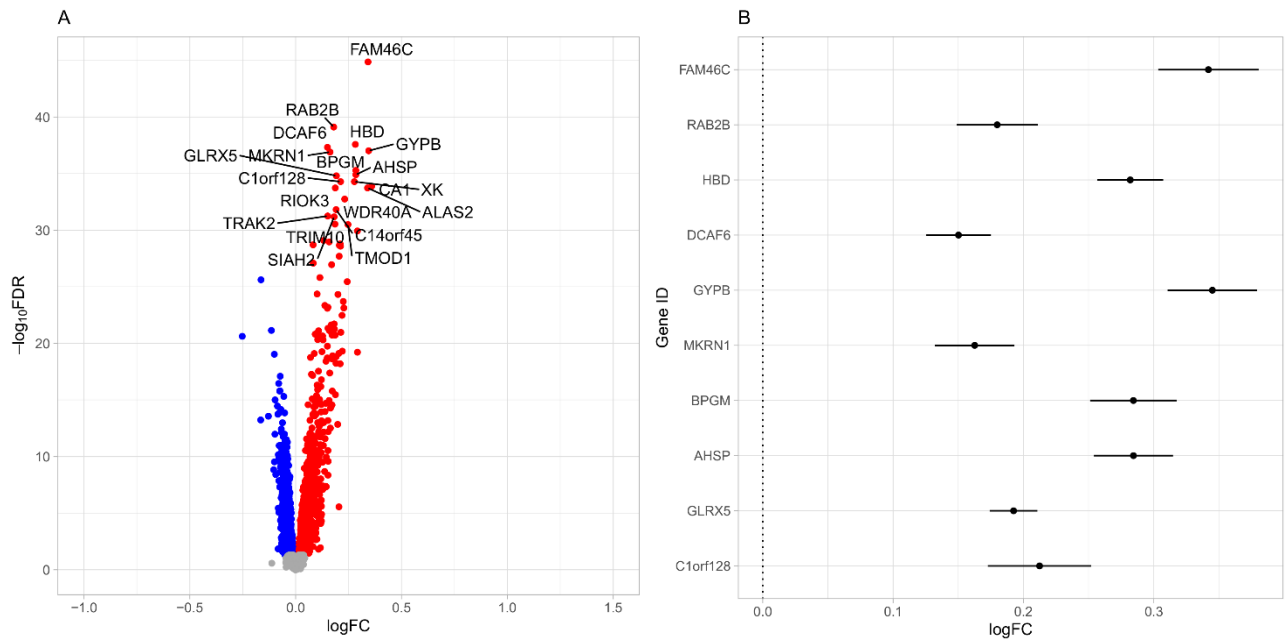
**Nikita Baiju<sup>1\*</sup>, Charlotta Rylander<sup>1</sup>, Pål Sætrom<sup>2, 3, 4, 5</sup>, Torkjel M. Sandanger<sup>1</sup>,  
Therese H. Nøst<sup>1, 5</sup>**

**Supporting information**

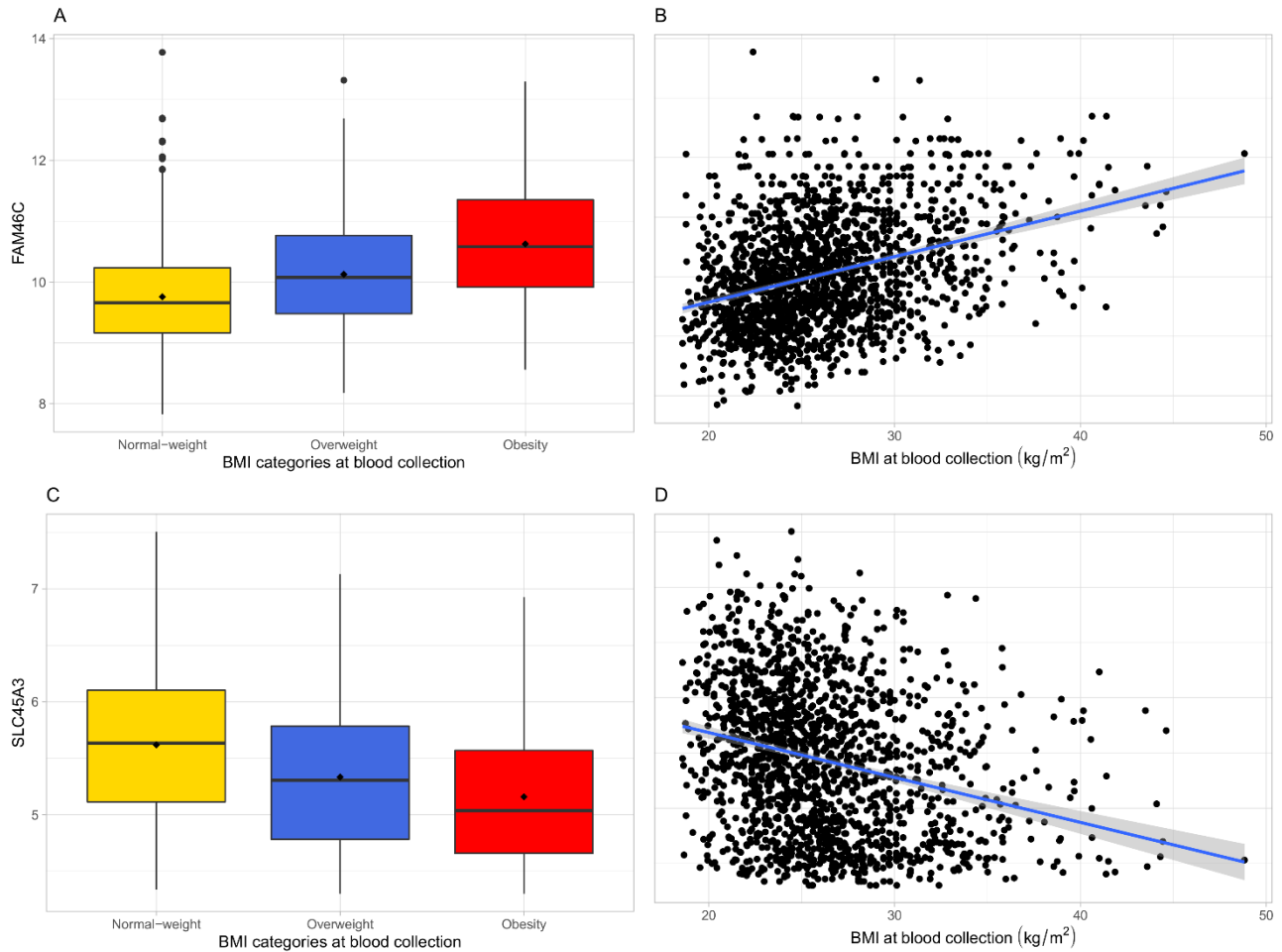
**Supplementary Figures**



**Figure S1. Estimated WBC proportions that differed among women with normal-weight, overweight, or obesity.** Boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as small circles. BMI: body mass index; WBC: white blood cell.

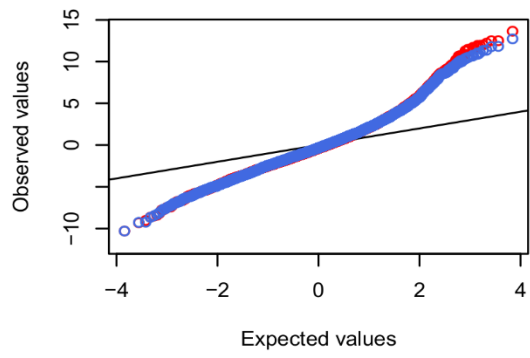


**Figure S2. Volcano plot of the test statistic for BMI at blood collection as a continuous metric in fully-adjusted model from the test of DEGs (A) and forest plot for the 10 top-ranked genes in test of DEGs (B).** In volcano plot (A), red dots display over-expressed genes ( $FDR \leq 0.05$  and  $\log_2FC > 0$ ), blue dots display under-expressed genes ( $FDR \leq 0.05$  and  $\log_2FC < 0$ ), while grey dots display genes with  $FDR > 0.05$ ; and gene names displayed are the top 20 DEGs in the test. In forest plot (B), the gene ID of DEGs are presented with the lowest FDR-adjusted p-values ranked from the top; the horizontal line for each gene represents their confidence interval; and the vertical dotted line represents the line of no difference. Here, BMI is modelled as scaled continuous metric. BMI: body mass index; DEGs: differentially expressed genes; FDR: false discovery rate;  $\log_2FC$ :  $\log_2$  fold-changes.

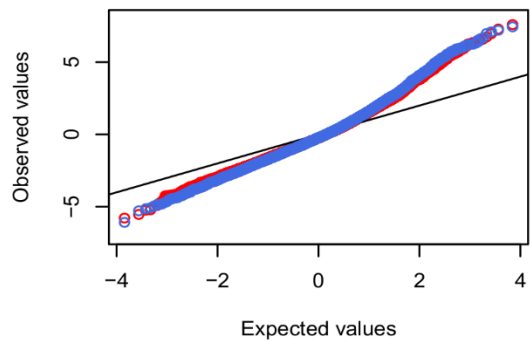


**Figure S3. Distributions of expression values for the top-ranked significant genes in BMI models: *FAM46C* (A and B) and *SLC45A3* (C and D) according to BMI categories (A and C) and BMI as continuous metric (B and D) at blood collection.** In figures A and C: boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as points. The black diamond-shaped dot represents the respective mean. In figures B and D: the blue line represents the regression line with a shaded grey area representing the standard error. BMI: body mass index.

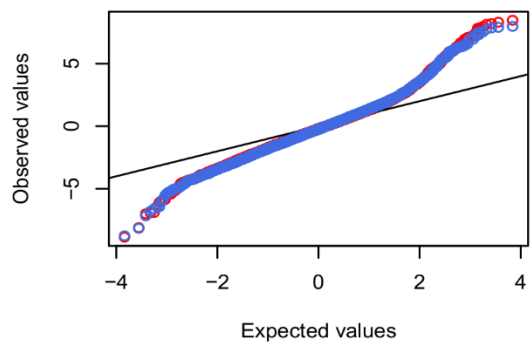
A. Ob -vs -No



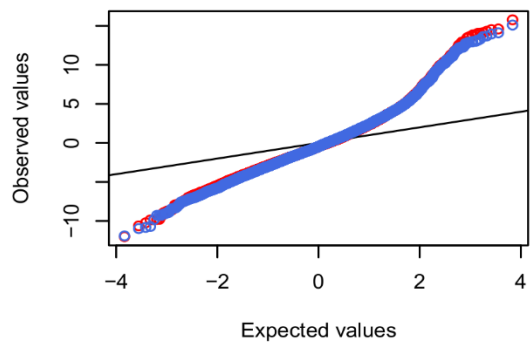
B. Ob -vs -Ov



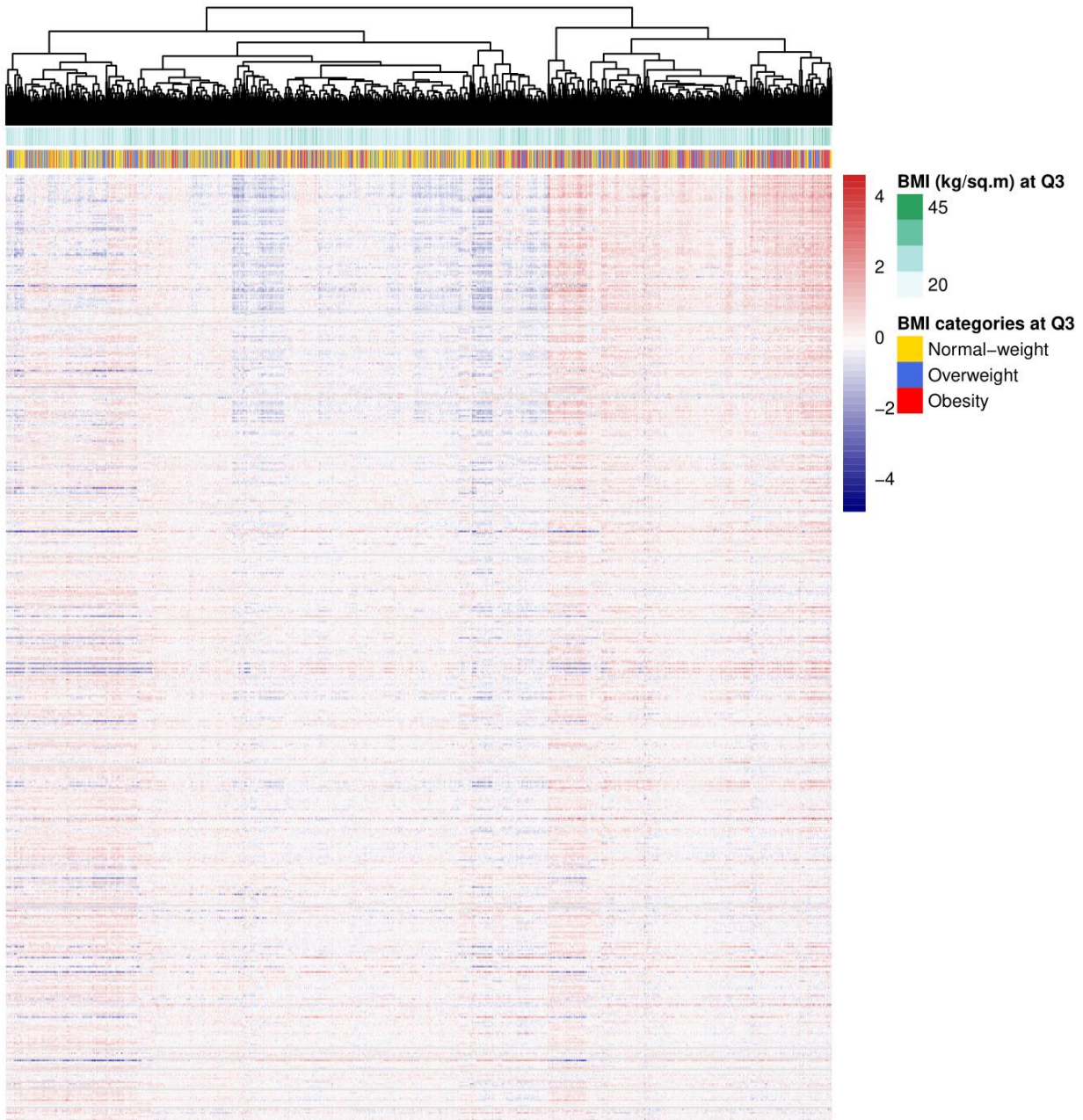
C. Ov -vs -No



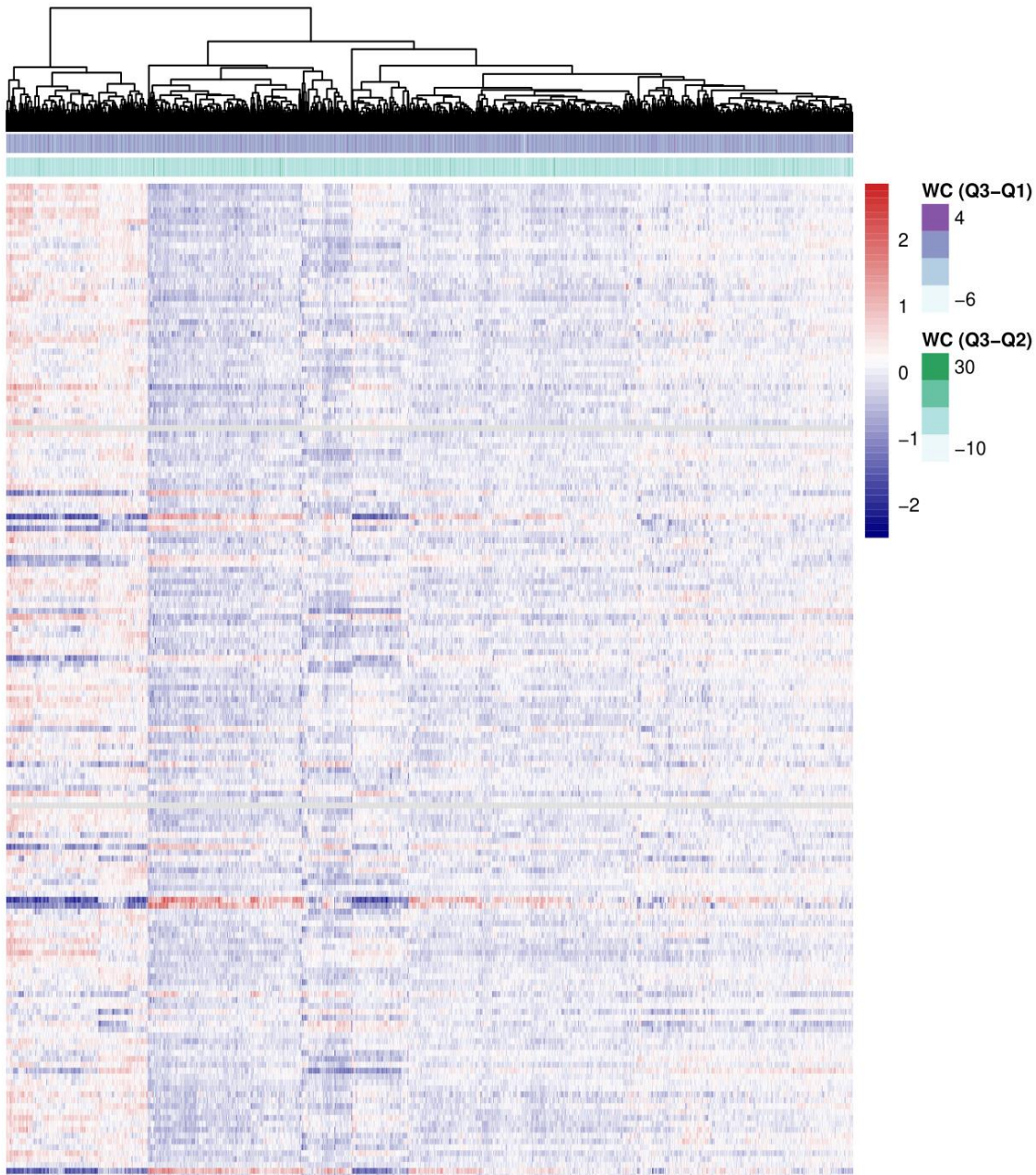
D. BMI



**Figure S4. Quantile-quantile plots for comparisons of BMI categories (A- C) and for BMI as a continuous metric (D) at blood collection in minimally-adjusted models (red), and fully-adjusted models (blue).** In figure D, BMI was modelled as scaled continuous metric. BMI: body mass index; Ob-vs-No: comparison of women with obesity versus normal-weight; Ob-vs-Ov: comparison of women with obesity versus overweight; Ov- vs-No: comparison of women with overweight versus normal-weight.

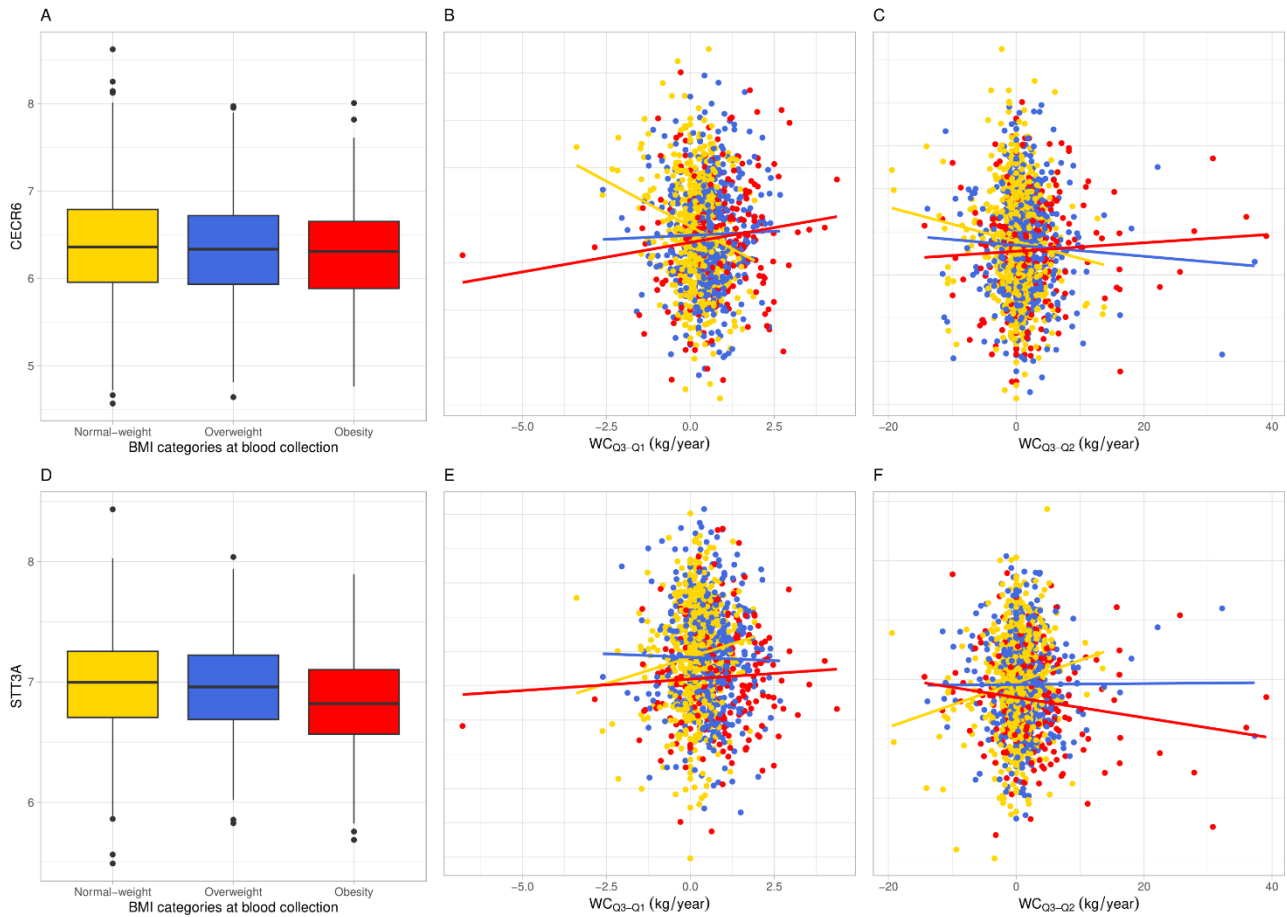


**Figure S5. Clustering hierarchical heatmap according to BMI categories at Q3 and BMI as a continuous metric at Q3 for the genes (N=525) overlapping across all models testing associations to BMI (obesity-vs-normal-weight, overweight-vs-normal-weight comparisons, and association to BMI as a continuous metric; both minimally- and fully-adjusted models).** Here, the rows represent genes, and the columns represent women. Expression levels are centered; red color represents higher expression, while blue color represents lower expression. BMI: body mass index, Q3: questionnaire at blood collection.



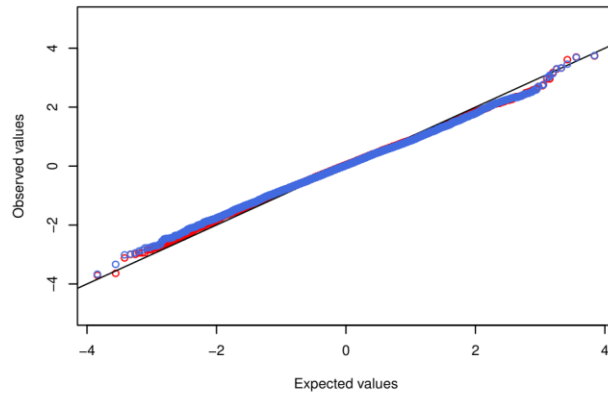
**Figure S6. Clustering hierarchical heatmap according to  $WC_{Q3-Q2}$  and  $WC_{Q3-Q1}$  for the genes (N=169) in the WC model (from the first interaction model).** Here, the rows represent genes, and the columns represent women. Expression levels are centered; red color represents higher expression, while blue color represents lower expression. Q1: baseline questionnaire; Q2: follow-up questionnaire; Q3: questionnaire at blood collection; WC: weight change;  $WC_{Q3-Q1}$ : weight change between Q1 and Q3;  $WC_{Q3-Q2}$ : weight change between Q2 and Q3.



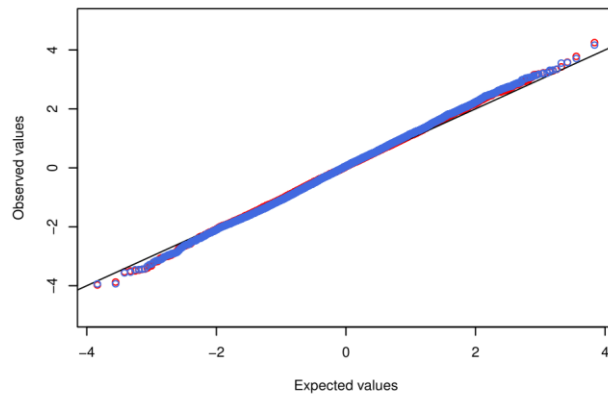


**Figure S7. Distributions of expression values for the top-ranked significant genes in WC analyses: *CECR6* (A-C), and *STT3A* (D-F) according to BMI categories at blood collection (A and D),  $WC_{Q3-Q1}$  (kg/year) (B and E), and  $WC_{Q3-Q2}$  (kg/year) (C and F). In figures A and D: Boxes extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile, horizontal bars represent the median, whiskers extend 1.5 times the length of the interquartile range above and below the 75<sup>th</sup> and 25<sup>th</sup> percentiles, respectively, and outliers are represented as points. The black diamond-shaped dot represents the respective mean. In figures B, C, E, and F, the lines represent the regression lines for different BMI categories at blood collection. Here, WC is modelled as scaled continuous variable. The top-ranked genes are results of the interaction effect of WC and BMI from the second interaction model (i.e.,  $BMI_{Q3} * WC_{Q3-Q1}$  or  $BMI_{Q3} * WC_{Q3-Q2}$ ). BMI: body mass index; WC: weight change; Q1: baseline questionnaire; Q2: follow-up questionnaire; Q3: questionnaire at blood collection;  $WC_{Q3-Q1}$ : weight change between Q1 and Q3;  $WC_{Q3-Q2}$ : weight change between Q2 and Q3.**

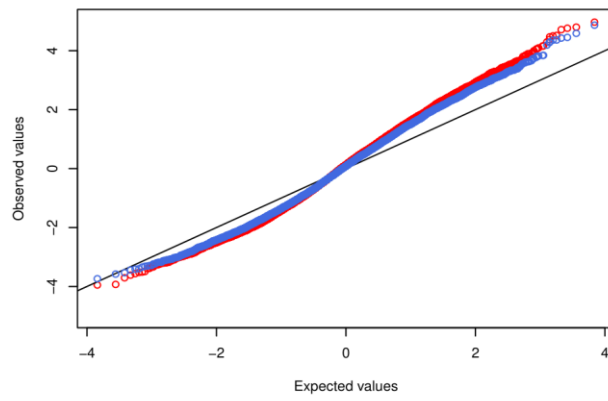
A.  $WC^1_{Q3-Q1}$



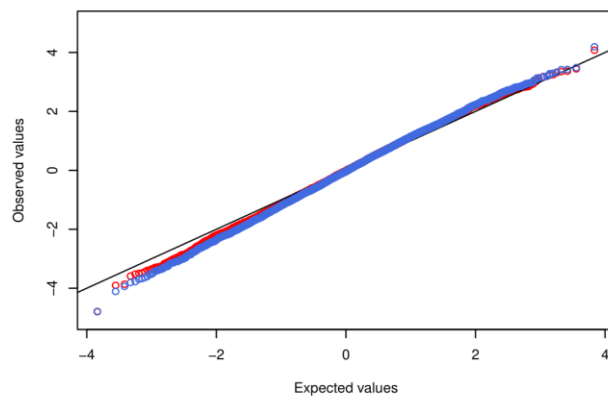
B.  $WC^1_{Q3-Q2}$



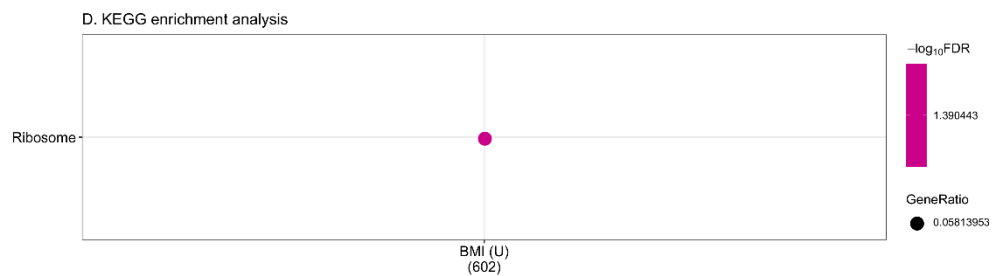
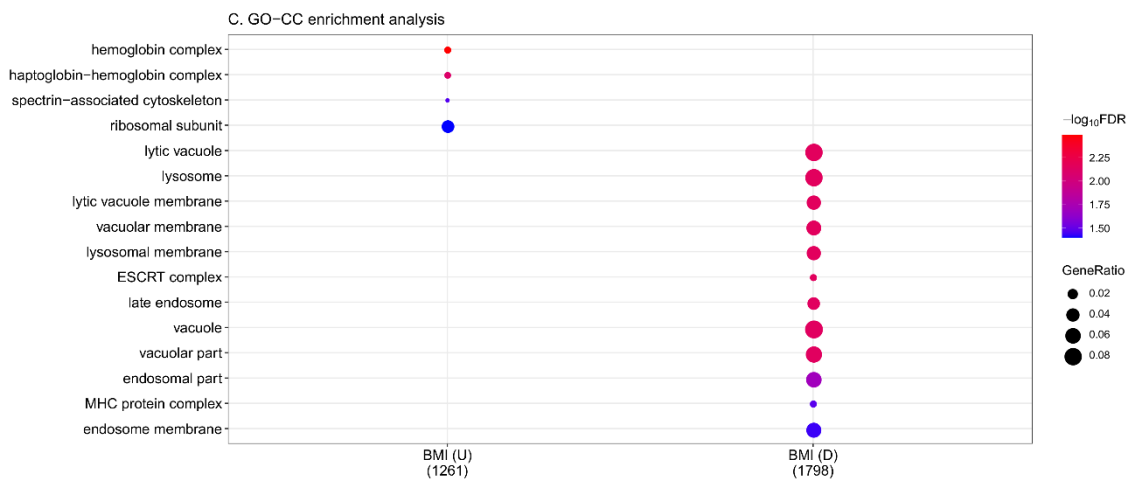
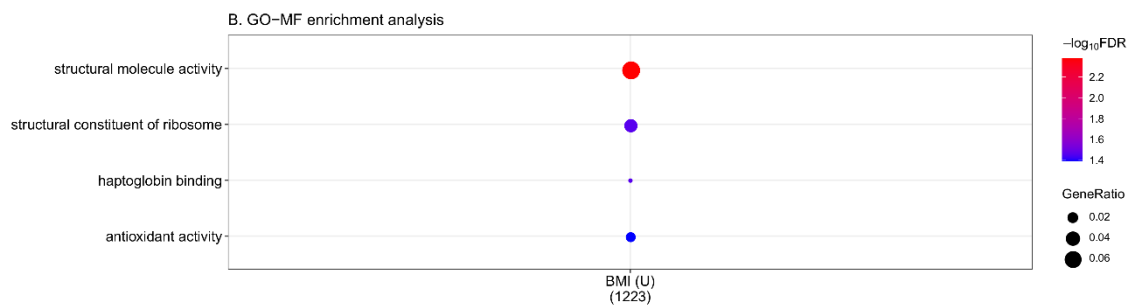
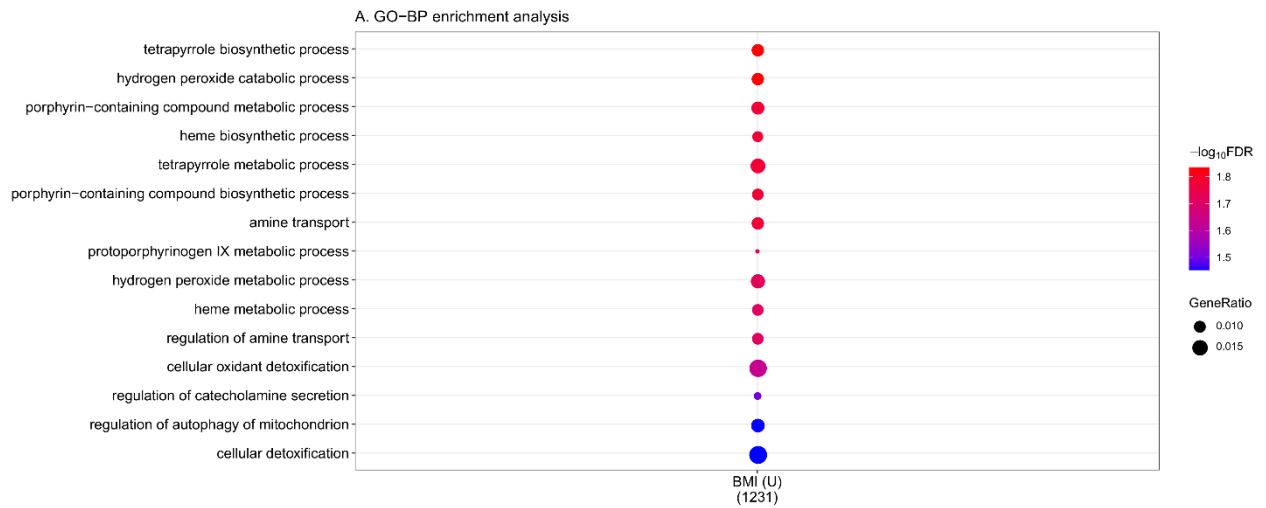
C.  $WC^2_{Q3-Q1}$



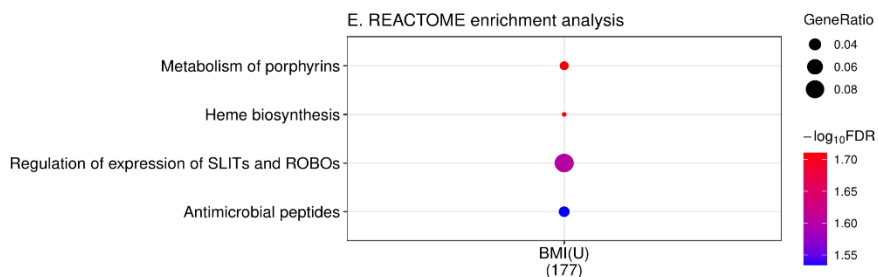
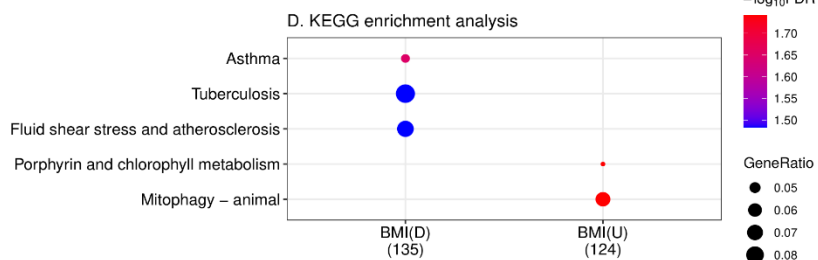
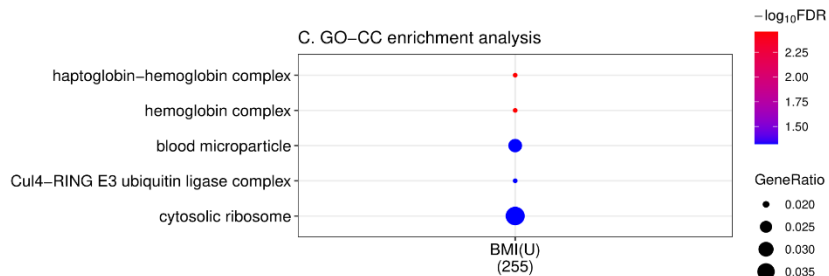
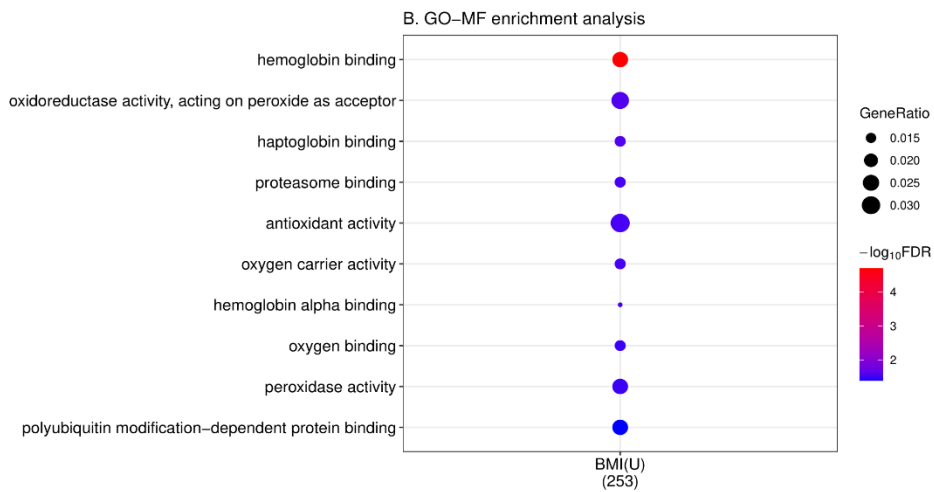
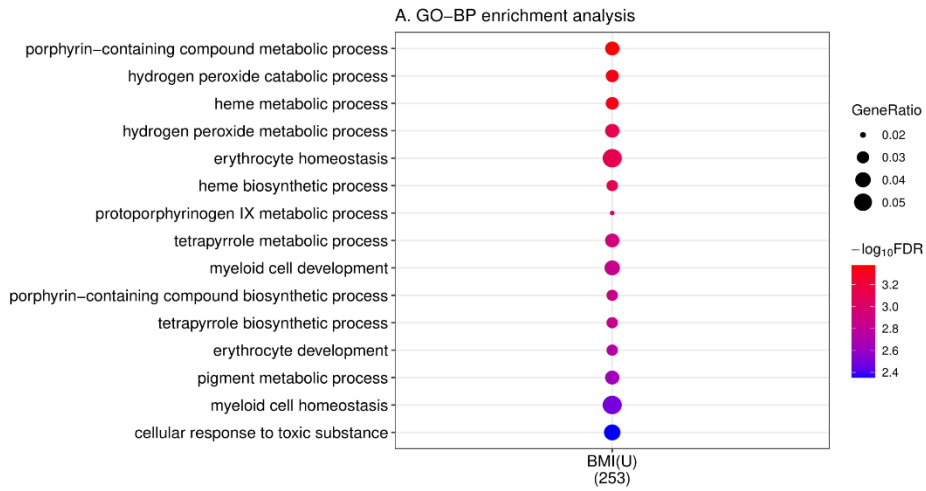
D.  $WC^2_{Q3-Q2}$



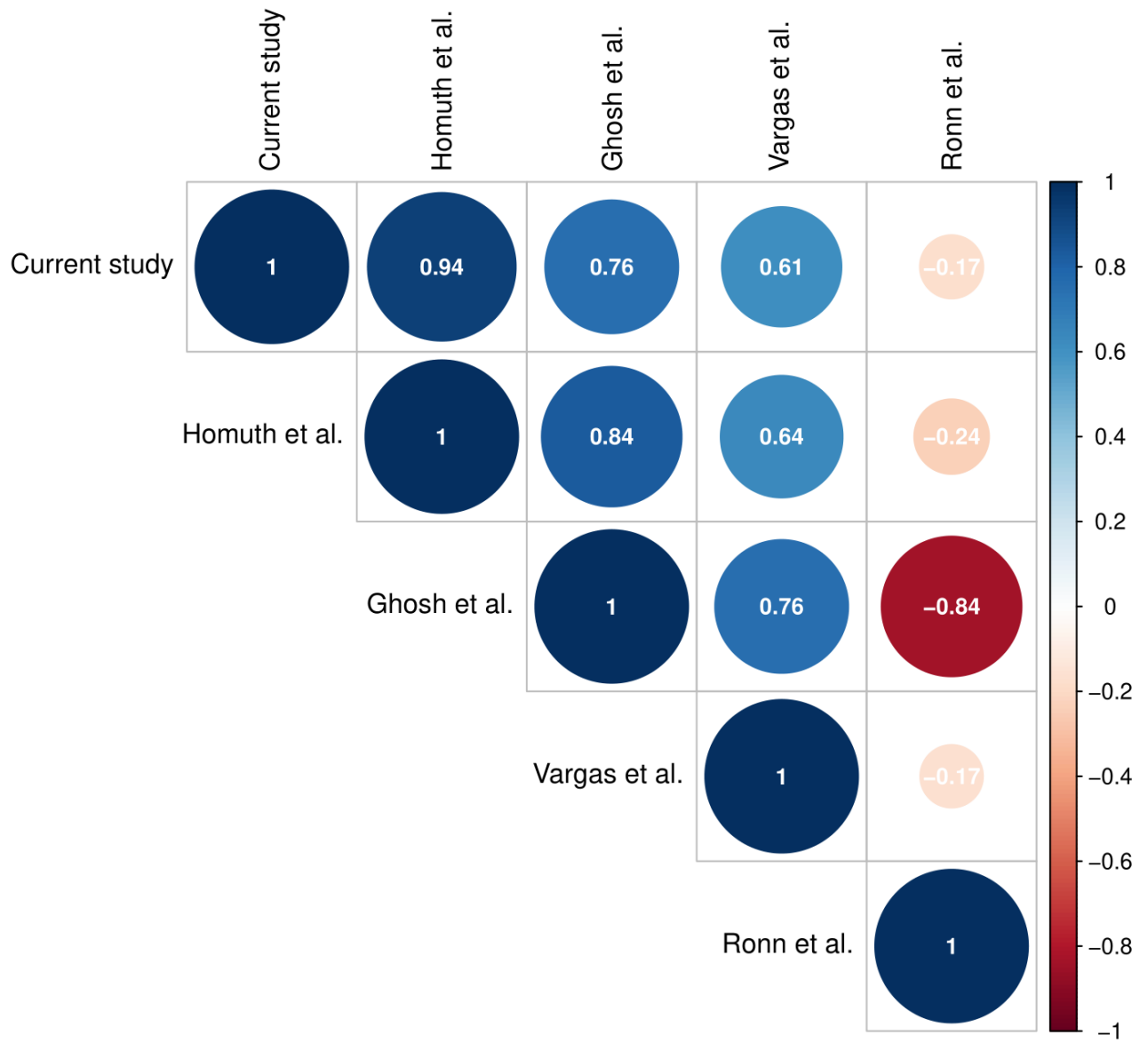
**Figure S8. Quantile-quantile plots for comparisons of  $WC_{Q3-Q1}$  (A), and  $WC_{Q3-Q2}$  (B) as continuous metrics in minimally-adjusted models (red), and fully-adjusted models (blue).** Here, WCs were modelled as a scaled continuous metric. Q1: baseline questionnaire; Q2: follow-up questionnaire; Q3: questionnaire at blood collection;  $WC_{Q3-Q1}$ : weight change between Q1 and Q3;  $WC_{Q3-Q2}$ : weight change between Q2 and Q3;  $WC^1$ : Results from the first interaction model (i.e.,  $BMI_{Q1} * WC_{Q3-Q1}$  or  $BMI_{Q2} * WC_{Q3-Q2}$ );  $WC^2$ : Results from the second interaction model (i.e.,  $BMI_{Q3} * WC_{Q3-Q1}$  or  $BMI_{Q3} * WC_{Q3-Q2}$ ).



**Figure S9. Summary of functional enrichment analyses for BMI at blood collection as a continuous metric for over-expressed (Up-regulated) and under-expressed (Down-regulated) genes for the GO-BP (A), GO-MF (B), GO-CC (C), and KEGG (D) pathway database.** The colour of the dots indicates the  $-\log_{10}\text{FDR}$ , where red dots represent the most enriched categories (i.e., ones with the lowest  $-\log_{10}\text{FDR}$ ); the 'GeneRatio' indicates the proportion of genes overlapping between lists of DEGs and the genes in GO categories. Here, BMI was modelled as scaled continuous metric. BMI: body mass index; DEGs: differentially expressed genes; FDR: false discovery rate; logFC:  $\log_2$  fold-changes; GO: gene ontology; BP: biological processes; MF: molecular functions; CC: cellular components; KEGG: Kyoto Encyclopedia of Genes and Genomes; U: Over-expressed genes (Up-regulated,  $\text{FDR} \leq 0.05$ ,  $\log\text{FC} > 0$ ); D: Under-expressed genes (Down-regulated,  $\text{FDR} \leq 0.05$  and  $\log\text{FC} < 0$ ).



**Figure S10. Summary of functional enrichment analyses for genes (N=525) overlapped across all models testing association to BMI (obesity-vs-normal-weight, overweight-vs-normal-weight comparisons, and association to BMI as a continuous metric; both minimally- and fully-adjusted models) for over-expressed (Up-regulated) and under-expressed (Down-regulated) genes for the GO-BP (A), GO-MF (B), GO-CC (C), and KEGG (D) pathway database.** The colour of the dots indicates the  $-\log_{10}\text{FDR}$ , where red dots represent the most enriched categories (i.e., ones with the lowest  $-\log_{10}\text{FDR}$ ); the ‘GeneRatio’ indicates the proportion of genes overlapping between lists of DEGs and the genes in GO categories. Here, BMI was modelled as scaled continuous metric. BMI: body mass index; DEGs: differentially expressed genes; FDR: false discovery rate; logFC:  $\log_2$  fold-changes; GO: gene ontology; BP: biological processes; MF: molecular functions; CC: cellular components; KEGG: Kyoto Encyclopedia of Genes and Genomes; U: Over-expressed genes (Up-regulated,  $\text{FDR} \leq 0.05$ ,  $\log\text{FC} > 0$ ); D: Under-expressed genes (Down-regulated,  $\text{FDR} \leq 0.05$  and  $\log\text{FC} < 0$ ).





**Figure S11. Correlation plot for logFC in the current study (N=3,106 DEGs) and estimates presented in four published studies investigating gene expression and the association with BMI.** Blue color indicates positive correlation, and red indicates negative correlation, whereas the size and number in the circles represent the strength of correlation. Homuth et al. (N=3,762) and Ghosh et al. (N=144) performed studies using whole-blood (overlapping DEGs with the current study: N=1,552 and N=42, respectively); Vargas et al. (N=1,864) was based on PBMCs (overlapping DEGs with the current study: N=337); and Ronn et al. (N=2,936) conducted the study using adipose tissue (overlapping DEGs with the current study: N=538).

The supplementary information of this paper comprised of large tables that are not suitable for print and are not included here. These files have been shared with the evaluation committee and will be made accessible when the manuscript is published and available online.

## **Paper III**

### **Associations of blood gene expression profiles with menopausal status and hormone therapy use in the Norwegian Women and Cancer Study (NOWAC) postgenome cohort.**

Baiju N, Waaseth M, Sætrom P, Sandanger TM, Nøst TH.

Manuscript.



