From The Department of Clinical Science and Education,
Södersjukhuset
Karolinska Institutet, Stockholm, Sweden

MOLECULAR SIGNATURES OF EARLY LIFE EXPOSURES AND COMPLEX DISEASES: APPLICATIONS USING EPIGENETICS AND TRANSCRIPTOMICS DATA

Simon Kebede Merid

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-814-4
Cover illustration: design by Kidus Kebede Merid

Molecular signatures of early life exposures and complex diseases: applications using epigenetics and transcriptomics data

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Simon Kebede Merid

The thesis will be defended in public at Sal Ihre, KI SÖS, Södersjukhuset, Sjukhusbacken 10, 118 83 Stockholm, Sweden

Friday 18th of November 2022, at 9.00

Principal Supervisor:
Professor Erik Melén
Karolinska Institutet
Department of Clinical Science and Education

Co-supervisors:
Associate Professor Olena Gruzieva
Karolinska Institutet
Institute of Environmental Medicine

Associate Professor Åsa Wheelock
Karolinska Institutet
Department of Medicine Solna

Professor Gerard Koppelman
University Medical Center Groningen
University of Groningen
Groningen Research Institute for Asthma and COPD

Opponent:
Associate Professor Craig P. Hersh
Harvard Medical School
Channing Division of Network Medicine,
Brigham and Women's Hospital

Examination Board:
Associate Professor Emma Andersson
Karolinska Institutet
Department of Cell and Molecular Biology

Professor Leonid Padyukov
Karolinska Institutet
Department of Medicine Solna

Associate Professor Carlos Guerrero Bosagna
Uppsala University
Department of Organismal Biology

To my mother

To my late father

To my doting wife and daughter

POPULAR SCIENCE SUMMARY OF THE THESIS

Both before and after birth, the human body undergoes rapid development including cell differentiation, organ formation, and organ system development. Environmental exposures and stressors during pregnancy and early life may influence these processes and thereby future health outcomes, potentially through molecular or epigenetic mechanisms. *Epigenetics* refers to changes in gene activity or function, without modification of the gene's deoxyribonucleic acid (DNA) sequence itself, and for example result in a gene being "switched on" or "switched off". This will in turn affect *gene expression* processes, i.e., whether and when a *protein is ultimately created* from the instructions contained in this gene. Epigenetic changes can be reversible or irreversible, and may also be heritable. The most extensively studied epigenetic mechanism is DNA methylation, where cytosine, one of the four DNA bases, is turned into 5-methylcytosine. DNA methylation has been proposed to constitute a link between genetic and environmental factors. Epigenetic patterns established early in life (already *in utero*) may affect gene expression over a lifetime and have been suggested to increase susceptibility to chronic diseases.

Chronic obstructive pulmonary disease (COPD) is a complex disease considered a major global health problem, and tobacco smoking is a major risk factor. Various other factors, such as genetics, air pollution, and repeated airway infections have also been shown to influence the risk of COPD. However, the role that DNA methylation might play in the pathogenesis of COPD has not been comprehensively studied.

Another complex disease, peanut allergy, is one of the most common food allergies and the leading cause of severe allergic reactions (anaphylaxis) among children. Peanut oral immunotherapy (pOIT) refers to ingestion of a small amount of peanut over time in a controlled manner and can lead to desensitization and tolerance. Combined treatment with the pharmaceutical omalizumab, an anti-immunoglobulin E (IgE) antibody that reduces the excessive immune response to e.g., peanuts, may facilitate initiation of pOIT in a safe way. However, the mechanisms of oral immunotherapy-induced tolerance are not well understood.

This thesis aimed to identify molecular signatures of early-life exposures, chronic respiratory disease, as well as allergy treatment responses.

In the first study, DNA methylation patterns at birth and later in childhood were investigated in relation to the child's age at birth, counted as gestational age in weeks. This large-scale investigation was based on data from 26 independent cohorts from an international PACE consortium and revealed that DNA methylation levels at several sites across the genome correlate strongly with gestational age. Further analyses of blood from the umbilical cord suggested that these changes were likely to capture fetal development across tissues. The absolute level of DNA methylation at the majority of the identified methylation sites changed during childhood, with many sites tending to catch up in methylation levels by school age and then stabilizing. However, we also identified a subset of methylation sites (17%) where differences in methylation levels related to gestation age were stable from birth to

adolescence. Bioinformatics analyses showed that the genes coupled to the identified methylation sites could be linked to human diseases, and are likely to be involved in biological processes essential for fetal development. Many of the methylation sites also affected the expression of nearby genes.

In the next study, the impact of outdoor exposure to particles of less than 2.5 micrometers in size ($PM_{2.5}$) at birth and current residential address on DNA methylation, and gene expression were assessed in childhood and adolescence. We found evidence suggesting that gene expression signatures in children and adolescents were associated with $PM_{2.5}$ exposure levels measured at the birth address. When we combined methylation and matched gene expression data to $PM_{2.5}$ exposure we found several examples where both methylation and expression levels were affected (called interactome hotspots). Some of the identified genes were associated with diseases known to be caused by or worsened by air pollution exposure.

In the third study, the methylation profiles in cells collected from the lower airways – primarily macrophages - were assessed in relation to COPD status and smoking in adults, with the aim of gaining further understanding of disease pathogenesis. We found several COPD-associated changes in the DNA methylation levels of the target cells, with a strong functional link to gene expression levels. Our analyses also suggest that both genetic and epigenetic mechanisms play important roles in COPD.

In the fourth study, the gene expression profiles before, during, and after pOIT were evaluated in adolescent patients with severe peanut allergy. Here, we found both up- and downregulation of immune-related genes in relation to pOIT and treatment with omalizumab. These results may shed light on mechanisms of allergen tolerance.

In conclusion, the presented results have increased our knowledge regarding the role of DNA methylation and gene expression in human development, pollution exposure effects, disease mechanisms, and treatment response. The findings may contribute in translational efforts bridging epidemiology, experimental research, and clinical care.

ABSTRACT

Environmental exposures and early life stressors may influence developmental processes and have long-term health consequences, potentially mediated by molecular mechanisms such as epigenetic modifications. The most extensively studied epigenetic mechanism is DNA methylation, which has been proposed to constitute a link between genetic and environmental factors. Epigenetic patterns established early in life (already *in utero*) may affect how a gene is expressed throughout life, and thereby increase susceptibility to chronic disease. Other factors like genetics and repeated airway infections also influence disease risk.

Chronic obstructive pulmonary disease (COPD) is a complex disease considered a major global health problem, with tobacco smoking being one of the main risk factors. The role that deoxyribonucleic acid (DNA) methylation might play in the pathogenesis of COPD has not been comprehensively studied. Bronchoalveolar lavage (BAL) cells from the airways and alveolar space are considered key targets for COPD. Peanut allergy is another complex disease – one of the most common food allergies and the leading cause of anaphylaxis among children. Peanut oral immunotherapy (pOIT) can lead to desensitization and tolerance, and combined treatment with anti-immunoglobulin E (IgE) using omalizumab may facilitate oral immunotherapy initiation. The mechanisms of oral immunotherapy-induced tolerance, including possible changes at the transcriptional level, are not well understood.

The main aim of this thesis was to identify molecular signatures of early-life exposures, chronic respiratory disease, as well as allergy treatment responses.

In **Study I**, the association between gestational age and DNA methylation patterns (at 5´-cytosine-phosphate-guanine-3´ sites, CpGs, across the genome) was investigated in newborns and older children from the large The Pregnancy And Childhood Epigenetics (PACE) consortium meta-analysis, including 11,000 participants in 26 independent cohorts. Changes in DNA methylation associated with gestational age were explored in additional pediatric cohorts at 4–18 years. The functional follow-up and correlation analyses between DNA methylation and gene expression were performed using cord blood. In addition, we evaluated DNA methylation profiles in other relevant tissues (fetal brain and lung) related to gestational age. We found numerous epigenome-wide differentially methylated CpGs related to gestational age at birth. Notably, many of the identified CpGs had not previously been associated with gestational age. Several CpGs affected the expression of nearby genes, displayed a strong functional link with human diseases, and were enriched in biological processes essential for fetal development. The epigenetic plasticity of fetal development across tissues was captured by many methylation sites. However, the majority of methylation levels underwent changes over time and stabilized after school age.

In **Study II**, the impact of outdoor exposure to particles of less than 2.5 micrometers in size ($PM_{2.5}$) at birth and current residential address on gene expression was explored in childhood and adolescence in the MeDALL consortium encompassing three European birth cohorts. In addition, the functional molecular patterns of $PM_{2.5}$ exposure were evaluated by integrating

protein-protein interaction and genome-wide gene expression with matched DNA methylation. We found evidence suggestive of gene signatures in children and adolescents associated with $PM_{2.5}$ exposure at birth. However, the integration of multi-omics profiles revealed several epigenetic deregulation gene module interactome hotspots where both methylation and expression levels were affected by $PM_{2.5}$ exposure at birth and current address. Some of the identified genes were associated with diseases known to be caused by or worsened by air pollution exposure.

In **Study III**, the pivotal role of DNA methylation profiles in BAL cells primarily macrophages was assessed in relation to COPD status and smoking in adults, to gain a further understanding of the disease pathogenesis. Several CpGs were associated with COPD in BAL cells, across the epigenome. Many of the identified CpGs displayed a strong functional link with gene expression and pathways enriched in cancer, various types of cell junctions, and cyclic adenosine monophosphate (cAMP) and Rap1 signaling. Notably, almost half of the CpGs co-located in the proximity of COPD-associated single nucleotide polymorphisms, which suggests that both genetic and epigenetic mechanisms are of importance at certain loci.

In **Study IV**, the blood gene expression profiles before, during, and after pOIT and Omalizumab (O, an anti-IgE monoclonal antibody) treatment were evaluated in adolescent patients with severe peanut allergy using high-throughput ribonucleic acid (RNA) sequencing. At the first two timepoints, baseline and pOIT start, we investigated if there was an effect of omalizumab treatment on gene expression. In addition, a longitudinal analysis was performed to evaluate the combined effect of pOIT with Omalizumab (pOIT+O). We also evaluated the overlap of pOIT+O-associated genes with genes associated with acute peanut allergic reactions in a previously published clinical study by Watson *et al*[1]. First, we showed that the blood gene expression of patients with peanut allergy was not altered by omalizumab treatment alone. However, the combined effect of pOIT+O showed up- and downregulation of several genes involved in T-cell functions and immune responses. Furthermore, comparing our findings with genes previously found to be affected during acute peanut allergic reactions suggested that pOIT+O may play a role in altering the same genes (in the opposite direction).

In conclusion, we demonstrated that DNA methylation profiles are related to gestational age at birth. The identified methylation sites were linked to human diseases and are likely to be involved in biological processes essential for fetal development. Most of the methylation sites also affect expression of nearby genes and reflect epigenetic plasticity of fetal development across tissues. We highlighted the added value of multi-omics analyses in relation to information on $PM_{2.5}$ exposure that may enhance the understanding of molecular mechanisms and biological responses induced by air pollutants. Moreover, we revealed COPD-associated methylation changes in macrophage-dense BAL cells with a strong functional link to different pathways and gene expression. Both genetic and epigenetic mechanisms play important roles at certain loci. We also provided insights into the transcriptome profiles during pOIT and combined treatment with omalizumab.

LIST OF SCIENTIFIC PAPERS

I. **Merid SK*,** Novoloaca A*, Sharp GC*, Küpers LK*, Kho AT*, Roy R, Gao L, Annesi-Maesano I, Jain P, Plusquin M, Kogevinas M, Allard C, Vehmeijer FO, Kazmi N, Salas LA, Rezwan FI, Zhang H, Sebert S, Czamara D, Rifas-Shiman SL, Melton PE, Lawlor DA, Pershagen G, Breton CV, Huen K, Baiz N, Gagliardi L, Nawrot TS, Corpeleijn E, Perron P, Duijts L, Nohr EA, Bustamante M, Ewart SL, Karmaus W, Zhao S, Page CM, Herceg Z, Jarvelin MR, Lahti J, Baccarelli AA, Anderson D, Kachroo P, Relton CL, Bergström A, Eskenazi B, Soomro MH, Vineis P, Snieder H, Bouchard L, Jaddoe VW, Sørensen TIA, Vrijheid M, Arshad SH, Holloway JW, Håberg SE, Magnus P, Dwyer T, Binder EB, DeMeo DL, Vonk JM, Newnham J, Tantisira KG, Kull I, Wiemels JL, Heude B, Sunyer J, Nystad W, Munthe-Kaas MC, Räikkönen K, Oken E, Huang RC, Weiss ST, Antó JM, Bousquet J, Kumar A, Söderhäll C, Almqvist C, Cardenas A, Gruzieva O, Xu CJ, Reese SE, Kere J, Brodin P, Solomon O, Wielscher M, Holland N, Ghantous A, Hivert MF, Felix JF, Koppelman GH, London SJ, Melén E. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. *Genome Med*. 2020 Mar 2;12(1):25

II. **Merid SK**, Bustamante M, Standl M, Sunyer J, Heinrich J, Lemonnier N, Aguilar D, Antó JM, Bousquet J, Santa-Marina L, Lertxundi A, Bergström A, Kull I, Wheelock ÅM, Koppelman GH, Melén E, Gruzieva O. Integration of gene expression and DNA methylation identifies epigenetically controlled modules related to PM2.5 exposure. *Environ Int.* 2021 Jan; 146:106248

III. Ström JE*, **Merid SK***, Pourazar J, Blomberg A, Lindberg A, Ringh MV, Hagemann-Jensen M, Ekström TJ, Behndig AF, Melén E. COPD is Associated with Epigenome-wide Differential Methylation in BAL Lung Cells. *Am J Respir Cell Mol Biol.* 2022 Mar 10; 66(6), 638–647

IV. Björkander S*, **Merid SK***, Brodin D, Brandström J, Fagerström-Billai F, van der Heiden M, Konradsen JR, Kabesch M, van Drunen CM, Golebski K, Maitland-van der Zee AH, Potočnik U, Vijverberg SJH, Nopp A, Nilsson C, Melén E. Transcriptome changes during peanut oral immunotherapy and omalizumab treatment. *Pediatric Allergy Immunol*. 2022 Jan;33(1): e13682

* These authors contributed equally.

CONTENTS

LIST OF ABBREVIATIONS

ADGRG6	Gene-Adhesion G Protein-Coupled Receptor G6
ALSPAC	Avon Longitudinal Study of Parents and Children
ASGR2	Asialoglycoprotein Receptor 2
ATS	American Thoracic Society
BAMSE	Children (Barn), Allergy, Environment (Miljö), Stockholm, Epidemiology
BAT	Basophil allergen test
BAL	Bronchoalveolar lavage
BioGRID	Biological General Repository for Interaction Datasets
ERS	European Respiratory Society
cAMP	Cyclic adenosine monophosphate
CBC	California Birth Cohort
CDKN2AIP	CDKN2A Interacting Protein
CD-sens	Basophil allergen threshold sensitivity
CDSN	Corneodesmosin
CHAMACOS	Center for the Health Assessment of Mothers and Children of Salinas
CHS	Children's Health Study
COPD	Chronic obstructive pulmonary disease
CO	Carbon monoxide
CPD	Carboxypeptidase D
CpGs	5´-cytosine-phosphate-guanine-3´sites
DL_{CO}	Diffusing capacity of the lungs for carbon monoxide
DNA	Deoxyribonucleic Acid
EDEN	Etude des Déterminants pré et post natals du développement et de la santé de l'Enfant
Envirogen	ENVIRomental influence ON early AGEing
ESCAPE	European Study of Cohorts for Air Pollution Effects
EWAS	Epigenome-wide association studies
FASTX	Food Allergen Suppression Therapy with Xolair®
FEM	Functional Epigenetic Module
FEV_1	Forced expiratory volume in 1 second
FLI1	Fli-1 Proto-Oncogene, ETS Transcription Factor
FOXA1-2	Forkhead Box A1-A2
FVC	Forced vital capacity
GDPR	General Data Protection Regulation
GECKO	Groningen Expert Center for Kids with Obesity
Gen3G	The Genetics of Glucose regulation in Gestation and Growth
GEO	Gene Expression Omnibus
GINIplus	German Infant Study on the influence of Nutrition Intervention PLUS environmental and genetic influences on allergy development
GNAI3	G Protein Subunit Alpha I3
GO	Gene Ontology
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GOYA	Genetics of Overweight Young Adults
GPBAR1	G Protein-Coupled Bile Acid Receptor 1
GREAT	Genomic Regions Enrichment of Annotations Tool
GWAS	Genome-wide association study
HINT	High-quality INTeractomes
HM13	Histocompatibility Minor 13
HPRD	Human Protein Reference Database
ICOS	Inducible T Cell Costimulator
IgE	Immunoglobulin E
ImmGen	Immunology Genome project
INMA	Infancia y Medio Ambiente
IoW	Isle of Wight Birth Cohort
IQR	Interquartile range
iRefWeb	Web interface to protein interaction data
JAZF1	JAZF Zinc Finger 1

KCTD15	Potassium Channel Tetramerization Domain Containing 15
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOLIN	Respiratory and Cardiovascular Effects in COPD
LUR	Land use regression
MAPK6	Mitogen-Activated Protein Kinase 6
MeDALL	Mechanisms of the Development of ALLergy
methQTL	Methylation quantitative trait locus
MINT	Molecular INTeraction database
MIPS	Munich Information Center for Protein Sequences
MiR-1296	MicroRNA 1296
MLST8	MTOR Associated Protein, LST8 Homolog
MOG	Myelin Oligodendrocyte Glycoprotein
MoBA	Norwegian Mother and Child Cohort Study
mRNA	Messenger ribonucleic acid
miRNA	MicroRNA
NCOR2	Nuclear Receptor Corepressor 2
NFBC1986	The Northern Finland Birth Cohorts 1986
NLRP3	NLR Family Pyrin Domain Containing 3
NO	Nitrogen oxide
NO_2	Nitrogen dioxide
NO_X	Nitrogen oxides
NR1I2	Nuclear Receptor Subfamily 1 Group I Member 2
OLIN	longitudinal Obstructive Lung disease In Northern Sweden
OIT	oral immunotherapy
PACE	Pregnancy And Childhood Epigenetics
PDB	Protein data bank
PIAMA	Prevention and Incidence of Asthma and Mite Allergy birth cohort
PiccoliPlus	Piccolipiù, a multicenter birth cohort in Italy
$PM_{2.5}$	Particles with an aerodynamic diameter less than 2.5 µm
PM_{10}	Particles with an aerodynamic diameter less than 10 µm
pOIT	Peanut oral immunotherapy
pOIT+O	Peanut oral immunotherapy with Omalizumab
POMC	Proopiomelanocortin
PPI	Protein-protein interaction
PREDO	Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction study
RAINE	The West Australian Pregnancy Cohort
RHEA	The mother-child Rhea cohort in Crete, Greece
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
SCARA3	Scavenger Receptor Class A Member 3
SCNN1A	Sodium Channel Epithelial 1 Subunit Alpha
SEMA7A	Semaphorin 7A
SNP	Single nucleotide polymorphisms
SOX30	SRY-Box Transcription Factor 30
TAF5	TATA-Box Binding Protein Associated Factor 5
TAF8	TATA-Box Binding Protein Associated Factor 8
TMEM176B	Transmembrane Protein 176B
TRIM69	Tripartite Motif Containing 69
TWAS	Transcriptome-wide association study
UCSC	University of California Santa Cruz
USP44	Ubiquitin Specific Peptidase 44
VC	Vital Capacity
WHO	World Health Organization
ZNF322	Zinc Finger Protein 322
ZNF609	Zinc Finger Protein 609

1 BACKGROUND

1.1 MOLECULAR BIOLOGY AND BIOINFORMATICS TOOLS

Molecular biology studies the chemical structure and different biological processes in the basic units of life, such as DNA, RNA, and proteins. It also evaluates the interaction of these macromolecules at a cellular level[2,3]. According to the central dogma of molecular biology, the information flow is consecutive starting from DNA to RNA and then to protein[4] (Figure 1). The chemical modifications of DNA by the environment without changing the sequence of DNA are called epigenetic, which may lead to the modification of genes at the transcriptomic level. In this thesis, epigenetic and transcriptomic molecular biological processes were explored, described in detail in the following sections.

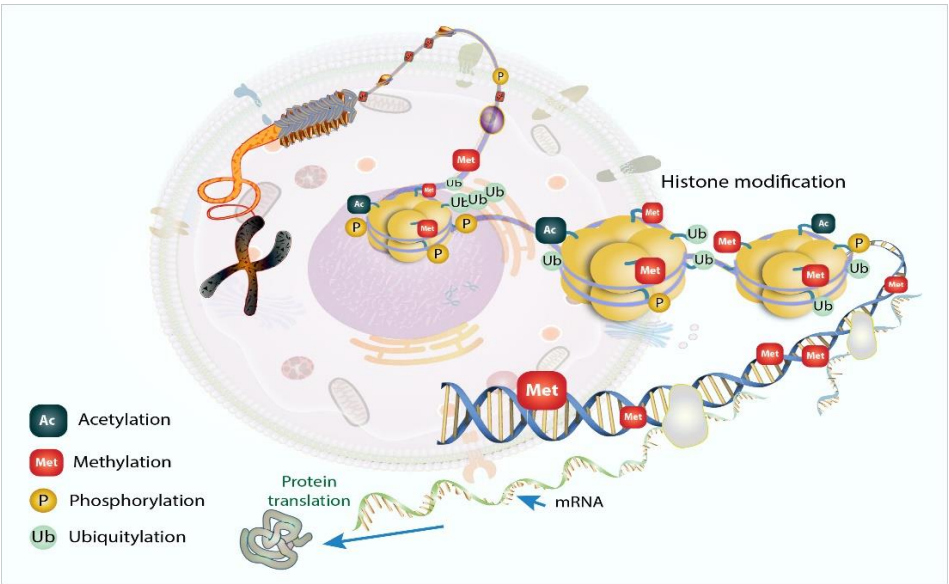

Figure 1. The central dogma of molecular biology information flow.

In recent years, numerous multi-omics data have been generated, and various bioinformatic tools have been developed to analyze such large-scale datasets. These approaches contribute to identifying potential biomarkers and give a functional interpretation of those markers. The application of bioinformatics is also discussed in the following section.

1.1.1 Epigenetics

Epigenetics studies heritable changes in gene activity or function that occur without modification of the DNA sequence. The epigenetic changes can occur during cell differentiation, X-chromosome inactivation, embryogenesis, and genomic imprinting[5,6]. DNA methylation, histone modifications, and non-coding RNAs are parts of the epigenetic processes[7]. The most extensively studied epigenetic mechanism is DNA methylation, where

5-methylcytosine occurs by transferring a methyl group onto the C5 position of cytosine. DNA methylation has been proposed to constitute a link between genetic and environmental factors[8]. In the event of disease, the abnormal response and behavior of the cell could be regulated by changes in DNA methylation patterns frequently induced by environmental factors, causing the alteration of the gene activity[9-11]. In addition, the change in DNA methylation level could transmit across human generations[12], and age-associated changes might occur even in the absence of disease[13]. Some of the studies included in this thesis focus on DNA methylation driven by early life exposure and disease status; details can be found in the upcoming sections.

1.1.2 Transcriptomics

A gene is usually subdivided into coding regions (exons) and non-coding regions (introns). Exons are the functional part of a gene where messenger RNA molecules are produced as the result of the transcription of the genetic sequence. However, in the splicing process, the introns are removed. During the gene activation, a set of mRNAs called transcripts are generated to form the transcriptome. The study of the transcriptome using mRNA expression is called transcriptomics. The expression levels can be assessed using both sequencing-based and microarray-based methods. In the upcoming section, detailed information on the transcriptomic studies included in this thesis can be found.

1.1.3 Application of bioinformatics tools

Substantial efforts have been made until today to develop appropriate statistical models and bioinformatics tools to get meaningful biological information out of the heterogeneous and high-dimensional data generated by high-throughput technologies. To unravel associations between DNA methylation and exposures or disease, many computational methods have been developed that can be applied in epigenome-wide association studies (EWAS)[14]. These bioinformatics aim to process and analyze the methylation data. Similarly, various methods have also been developed to process and analyze gene expression data[15]. Analyzing only single-level omics profiling can give important information about involved mechanisms, but will not fully elucidate a complex biological response. Combining and integrating various omics levels may therefore help to better understand the holistic underlying biological mechanisms[16]. To overcome these major challenges, different analytical protocols and computational tools have been developed to integrate the different layers of omics[17,18]. Many bioinformatic tools, computational capacity, and statistical methods have also been developed to provide the functional interpretation of the interaction between genes or proteins[19]. Those tools depend on already created databases by assembling thousands of reported literature interactions and enriching important biological information[20]. The most common biological databases used by most of the tools include Gene Ontology (GO) which characterizes and categorizes the functions of genes and their products according to biological processes, molecular functions, and cellular components[21], and functional databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG)[22] and Reactome pathway[23]. Utilizing different functional analysis tools can enhance our understanding of the underlying biological

mechanisms of a large list of genes or proteins generated from high-throughput genomics and proteomics technologies[24].

1.2 GESTATIONAL AGE

1.2.1 Health effects of preterm birth and low gestational age

Gestational age measures pregnancy length and the time a fetus grows inside the mother's womb, commonly assessed based on ultrasound estimations or the date of the last menstrual period[25]. Around 1 in 10 babies are born prematurely (birth before 37 weeks gestation), according to the World Health Organization, often with no apparent cause although several risk factors have been established (see below). Preterm birth is the leading cause of neonatal morbidity and mortality[26,27], as well as long-term morbidity and impaired health conditions[28-31]. Children surviving birth at a very low gestational age are more likely to have major health challenges, including retinopathy of prematurity, bronchopulmonary dysplasia, and cardiovascular and neurodevelopmental impairment[32-36]. Infants born moderately preterm (between 32 and 36 completed weeks of gestation) are also observed to have lower lung function compared to those born at term[37]. Moreover, learning disabilities, sensory defects, and respiratory illnesses are shown to be more prevalent in preterm-born children[38].

1.2.2 Gestational age and DNA methylation

It is well known that preterm birth is associated with many risk factors, including multiple pregnancies, smoking, maternal stress, and intrauterine infections, ethnicity, and genetic factors[39-44]. Many of these have also been associated with DNA methylation patterns, which may constitute an important link between exposure and outcome[45,46]. The epigenetic patterns related to such exposures potentially influence gene expression profiles that can lead to chronic diseases later in life[47-49]. Several previous studies have reported the link between gestational age and cord blood DNA methylation among both term and preterm births[50-55]. Although DNA methylation patterns related to gestational age have been identified at birth, little is known about the stability and persistence of these methylation changes over time. In a comparison of children born preterm and term, numerous loci identified as differentially methylated at birth were also associated with body size and height at school-age, with a potential link to bone mineralization processes[56]. A longitudinal EWAS comparing individuals born preterm to those born at term reported numerous methylation differences at birth. Interestingly, some CpG sites distinguishing preterm and term birth have been observed at 18 years of age, reflecting a lasting epigenetic effect[47]. Further knowledge of the role of DNA methylation and gene expression on the length of gestation will increase our understanding of the molecular basis of the aberrant process of prematurity and normal human development.

1.3 AIR POLLUTION EXPOSURE

1.3.1 Air pollution exposure and health effects

Air pollution constitutes a mixture of solid and liquid particles with different chemical and biological properties, such as particulate matter (PM), several gases, including ozone (O_3), nitrogen oxides (NO_x), and carbon monoxide (CO), as well as vapors like volatile organic compounds[57]. Particulate matter is a general term used for particles of different origins, sizes, shapes, and chemical and physical properties, suspended in air. PM are labeled depending on their size: particles having an aerodynamic diameter < 10 μm, PM_{10}; fine $PM_{2.5}$ representing particles less than 2.5 μm in diameter; and ultrafine PM with an aerodynamic diameter of <0.1 μm, $PM_{0.1}$. PM_{10} only reach the proximal airways and are usually removed by mucociliary clearance if the airway mucosa is intact. However, fine, and ultrafine PM are able to access the lower parts of the human airway passages and the circulation system as they bypass the alveolar wall[58]. NO_x has two components, nitrogen oxide (NO) and nitrogen dioxide (NO_2).

The major source of NO_x is automobile exhaust. The combustion processes are the primary source of fine and ultrafine PM, but some also as secondary particles from semi-volatile compounds. PM is mainly derived from re-suspended road dust, originated from the mechanical wear of tires and brake lining of the vehicles, road surface wear as well as sanding/salting of roads.

Outdoor air pollution has been estimated to cause 4.2 million premature deaths every year, according to WHO[59]. In epidemiological studies, air pollution exposure has also been linked to different health effects, such as adverse pregnancy outcomes[60], childhood airway disease[61], COPD[62], and neurodevelopmental disorders[63] (e.g., reduction in fundamental cognitive development[64] and autism spectrum disorder[65]). Previous studies have shown that children are particularly sensitive to the adverse effects of air pollution, and that timing of exposure plays a critical role in subsequent pathophysiological changes later in life. In the Swedish cohort BAMSE, exposure to air pollution during the first year of life has been associated with asthma, and lung function impairment in children up to school age and adolescence, as well as with chronic bronchitis and airflow limitation up to young adulthood[66-71]. Although the exact mechanisms underlying such adverse health effects of air pollution are unclear, several studies have suggested pollutant-induced oxidative stress and systemic inflammation as potential intermediate biological responses to exposure[72,73]. Overall, the present evidence indicates that traffic air pollution exposure is related to molecular changes through oxidative stress and systemic inflammation[74]. Yet, the exact mechanisms are not known.

1.3.2 Molecular changes induced by air pollution exposure

In the past, many single-level omics studies of environmental exposure, including air pollution effects on DNA methylation[75-80] and gene expression[81-84], were conducted. Moreover, the recent large-scale study of long-term $PM_{2.5}$ exposure on gene expression revealed several genes involved in cell signaling and immune response[85]. While analyses

using individual gene expression or DNA methylation profiling can give important information about either transcriptomic or epigenetic mechanisms, such an approach may be insufficient to reveal the complex biological response to air pollution exposure. In contrast, the integration of omics data may shed new light on gene modules or molecular pathways underlying negative health effects[86,87]. Combining and integrating multi-omics levels may help to better understand the holistic biological mechanisms[16]. In addition, it has been suggested that omics integration may enhance study power with an increased likelihood of identifying biologically relevant mechanisms[88,89]. For example, in the recent integrative study of methylome and transcriptome of human cardiomyocytes, multiple altered methylomes and transcriptome signatures in the cardiac disease-specific genes following $PM_{2.5}$ exposure have been reported[90]. At present, human studies of environmental exposures combining different -omics data are, however, scarce[91].

1.4 CHRONIC OBSTRUCTIVE PULMONARY DISEASE

1.4.1 Prevalence and definition of COPD

COPD is one of the complex diseases that constitute a major global health problem and is the third leading cause of death in the world[92]. About 384 million people suffer from COPD worldwide, and in Sweden, almost 500,000 people are affected by the disease[93-95]. Tobacco smoking is the major cause of COPD. Among the elderly tobacco smokers, approximately fifty percent develop COPD[96]. Yet, it is a puzzle why some but not all develop COPD. Both genetic and environmental exposures contribute to the risk of COPD. Indoor and outdoor air pollution, including occupational exposure to organic and inorganic dust and chemicals, are the key environmental risk factors for COPD[93].

Furthermore, children and young adults with impaired lung function have been shown to have an increased risk of COPD later in life[97,98]. There are different early life risk factors linked to the development of irreversible airflow limitation, which is a key feature in COPD, including parental asthma, asthma, childhood respiratory tract infections, and exposure to air pollution[66]. Prematurely born children often follow a low lung function trajectory and may not reach the peak lung function in early adulthood, which potentially contributes to the development of COPD later in life[99-101].

It has been demonstrated that the contribution of genetic factors is important in COPD. Several COPD-associated genetic variants have been identified in large genome-wide association studies[102]. In addition, a twin study has shown that 40-60% of COPD susceptibility is explained by genetic factors[103,104]. The accumulated gene-environment interaction that the individual encounter over the life span may also contribute to the development of COPD[105].

Historically, various definitions of COPD have been used. Recently, The Global Initiative for Chronic Obstructive Lung Disease (GOLD) harmonized and united the view and definition of COPD[93]. In this thesis, the GOLD guideline was used to define COPD (Table 1).

Table 1. GOLD stage classification of COPD[93].

All stages required to have the Post-bronchodilator $FEV_1/FVC < 0.7$	
Stage	%Predicted
GOLD 1: Mild	$FEV_1 \geq 80$
GOLD 2: Moderate	$50 \leq FEV_1 < 80$
GOLD 3: Severe	$30 \leq FEV_1 < 50$
GOLD 4: Very severe	$FEV_1 < 30$

GOLD: The Global Initiative for Chronic Obstructive Lung Disease,
FEV_1: Forced expiratory volume in 1 second and
FVC: Forced vital capacity

COPD is characterized by airflow limitation and persistent respiratory symptoms, including dyspnea. The observed airflow limitation can be caused by emphysema, inflammation of the central airways, and/or obstruction of the small peripheral airways[93]. COPD is also linked with other diseases, including lung cancer, cardiovascular disease, osteoporosis, skeletal muscle dysfunction, diabetes, and depression[106].

1.4.2 DNA methylation in COPD

The potential risk factors of COPD, such as cigarette smoking, and air pollution, may also induce DNA methylation changes following these exposures[107,108]. The overlap of DNA methylation network modules from fetal and adult lung tissues reveals the disease pathways linked with the exposure-related and age-associated developmental origin of COPD[109]. The role that DNA methylation might play in the pathogenesis of COPD has not yet been studied comprehensively. A systematic review by Machin *et al.* showed that COPD EWASs on peripheral blood have found no consistent differences[110]. However, studies based on cells from the main target organ of the disease (lung tissue and bronchial brushings) have identified consistent findings[111,112]; and previous transcriptomic studies on COPD also show rather strong associations[113]. Despite the great interest and intense research in DNA methylation patterns on target organs and surrogate tissue from COPD subjects and controls, no study has evaluated DNA methylation patterns in cells from airways on bronchoalveolar lavage (BAL) cells to date.

1.5 PEANUT ALLERGY AND ORAL IMMUNOTHERAPY

1.5.1 Peanut allergy

Peanut allergy is an IgE-mediated disease that affects approximately 2% of the population in high-income countries[114]. It is a complex disease that for many patients is a lifelong condition[115] starting early in life. However, about 20% of affected children may outgrow their peanut allergy[116]. In the first decade of the 21st century, the prevalence of peanut allergy has

increased in the United States[117]. A similar trend was also observed from nationwide United Kingdom (UK) records between 2001 and 2005[118]. In addition, the population birth cohort BAMSE in Stockholm (Sweden) showed an increase in self-reported peanut allergy and serum IgE-sensitization between 4- and 8-year-olds[119]. Some peanut-allergic individuals have a severe disease that may lead to anaphylaxis after peanut allergen ingestion. A Swedish study on emergency-care visits showed that about 19% of food-induced anaphylaxis in children was caused by peanuts[120], which is in line with data observed worldwide[121].

1.5.2 Oral immunotherapy

In recent years, oral immunotherapy (OIT) has emerged as a promising treatment for children with different IgE-mediated food allergies. It has a success rate of desensitizing (i.e., loss of IgE sensitization) up to 80% in children with persistent food allergies like peanut, milk, egg, or wheat[122-124]. However, safety issues and adverse events, such as severe allergic reactions, must be considered[125]. Oral immunotherapy protocols outline ingestion of the allergen in a controlled manner with gradually increasing dosages, rendering the individual to become desensitized and, if continued, may eventually result in tolerance. Specifically, peanut oral immunotherapy (pOIT) can induce desensitization and then tolerance[126]. Although the pathogenesis of food allergy is relatively well-studied[127], mechanisms of OIT-induced tolerance are not well understood. However, it is known that OIT impacts multiple cell types that can act together to adapt the immune response and lead to desensitization. Still, not all patients respond to OIT, for some the effect is very limited and short-lived[128,129].

Omalizumab (an anti-IgE monoclonal antibody), which is used as a treatment for severe allergic asthma and other IgE-driven allergies, can also facilitate OIT initiation[130]. However, little is known about the involved mechanisms, including possible changes at the transcriptional level.

1.5.3 Transcriptional changes in peanut oral immunotherapy

Even though we are beginning to understand the immunological events associated with desensitization in peanut OIT, studies on transcriptional changes may further increase our understanding of the underlining mechanisms and identify potential biomarkers of successful pOIT. This will be valuable both when using pOIT in the clinic and may help peanut-allergic individuals, especially those that are non- or low responders to OIT. Transcriptional changes associated with pOIT have not yet been studied extensively. A few studies have used singe cells RNA-seq to investigate peanut-specific CD4+ T cells and regulatory immune cell populations to understand successful desensitization in those undergoing pOIT[131,132]. However, little is known about the effect of pOIT on the blood transcriptome level.

2 AIMS

The overarching aim of this thesis was to identify molecular signatures of early-life exposures, chronic respiratory disease, and allergy treatment response by applying appropriate computational and statistical methods in a molecular epidemiological framework.

The specific aims of this thesis were:

I. To investigate the association of gestational age with DNA methylation patterns in newborns and older children
II. To assess the impact of air pollution exposure on gene expression and DNA methylation profiles in childhood and adolescence
III. To characterize DNA methylation patterns in BAL lung cells related to chronic obstructive pulmonary disease, COPD
IV. To explore changes in transcriptomic profiles during omalizumab treatment and oral immunotherapy for peanut allergy

3 MATERIALS AND METHODS

3.1 DATA SOURCES

Study I utilized individual participant data from 39 birth and child cohorts (including BAMSE described below) based in 14 countries in Europe, the United States of America (USA), Canada, and Australia, participating in the PACE consortium. **Study II** used data from the European MeDALL consortium that includes BAMSE and additional six European cohorts of younger children and six older cohorts (up to adolescence). **Study III** obtained data from the Swedish KOLIN study, and data from the Swedish FASTX study were analyzed in **Study IV**.

3.1.1 BAMSE birth cohort

The Barn (Children), Allergi (Allergy), Miljö (Environment), Stockholm, Epidemiologi (Epidemiology) (BAMSE) birth cohort is an ongoing prospective study from Stockholm, Sweden[133]. Enrollment occurred between February 1994 and November 1996 at the first child health visit from four predefined areas of Stockholm County (Järfälla, Solna, Sundbyberg, and parts of Stockholm inner-city), Figure 2. The study area represents the inner city, urban and suburban areas with different housing types, socio-demographic characteristics, and environmental exposures, including traffic-related air pollution levels.

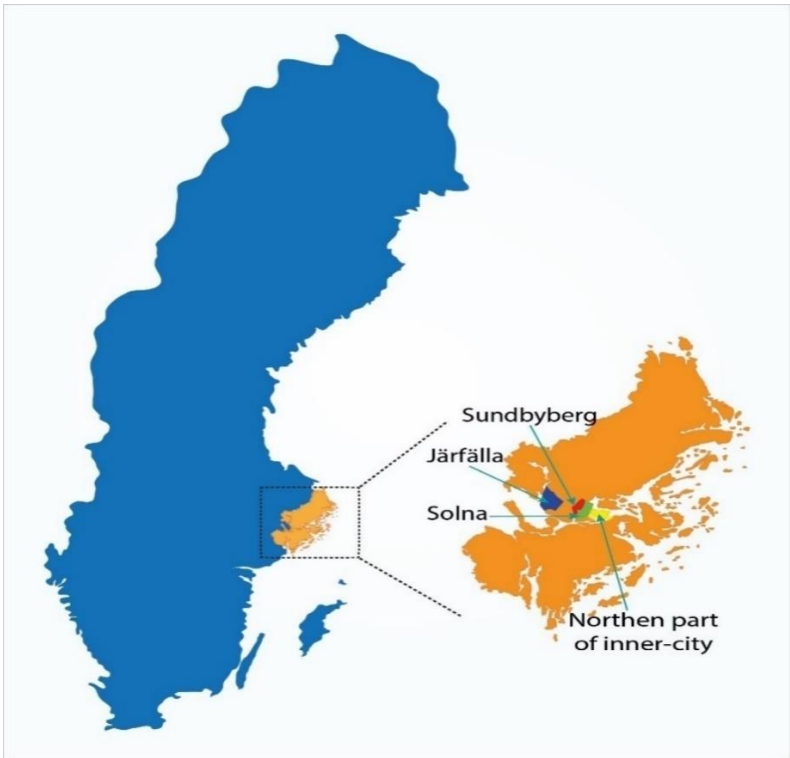

Figure 2. The four predefined recruitment areas included in the BAMSE birth cohort of Stockholm County, Sweden.

When the children were two months of age, their parents received a questionnaire on parental allergies and various exposures (e.g., housing characteristics and lifestyle factors). Out of the 7,221 newborns in the study area, 477 had unavailable addresses, 502 declined participation, 897 never answered the questionnaire, and 1,256 were actively excluded; either the family planned to move within a year, did not sufficiently understand Swedish, had a seriously ill child, or an older sibling already enrolled. In the end, a total of 4,089 children were included. At ages 1, 2, 4, 8, 12, and 16 years of the children, the parents received repeated questionnaires focused on symptoms of allergic disease in their children, as well as on various risk factors[134]. The survey response rates were 96%, 94%, 91%, 84%, 82%, and 78%, respectively. All children with completed questionnaires at 4, 8, and 16 years of age were invited to a clinical examination, including lung function testing and blood sampling. Blood samples are available for 2,605, 2,470, and 2,547 children at the age of 4, 8, and 16, respectively, including 1,699 children at all three clinical follow-ups. A summary of the follow-ups is illustrated in Figure 3. Additional BAMSE follow-ups have been completed in recent years (2016-2019 and 2020-2022), but data from these occasions were not included in the present thesis[67,135].

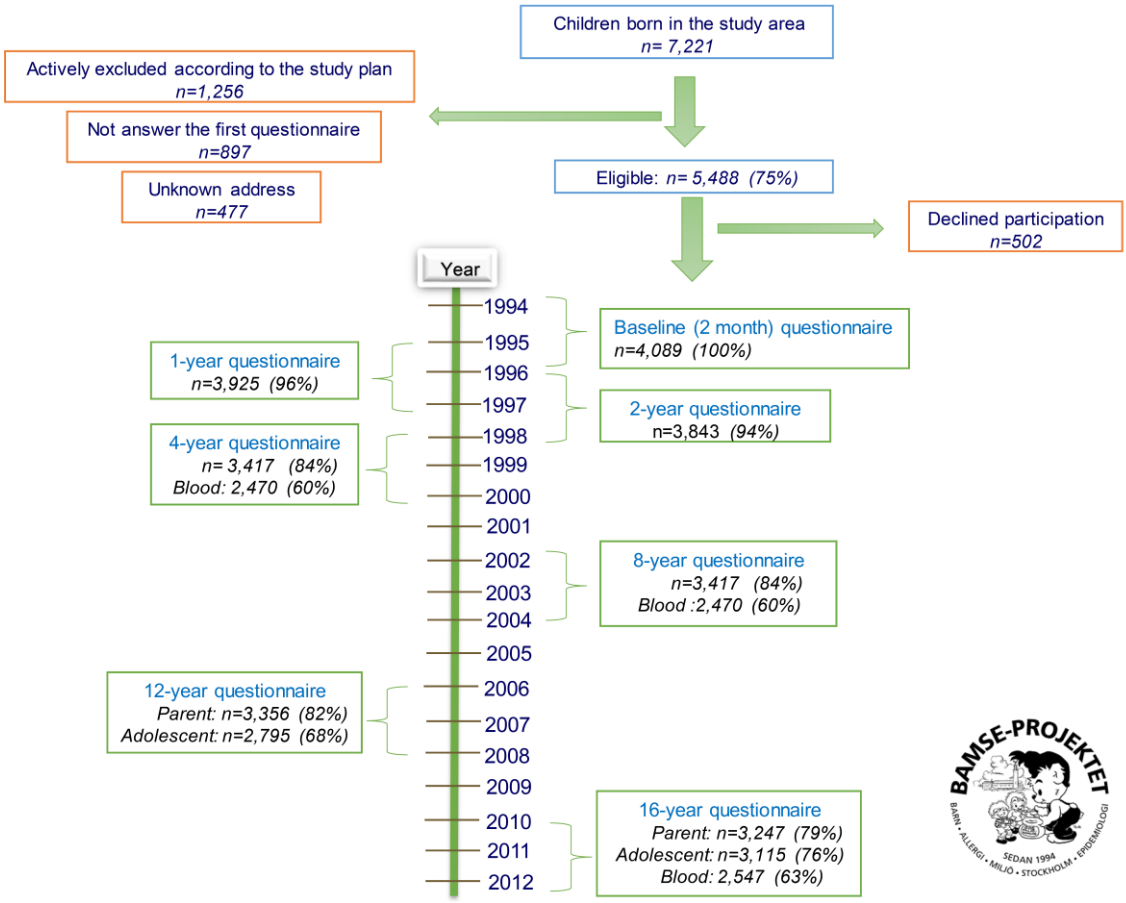

Figure 3. The flow chart of the BAMSE birth cohort recruitment and follow-up years of data collections.

3.1.2 PACE consortium

The Pregnancy And Childhood Epigenetics (PACE) consortium (http://www.niehs.nih.gov/research/atniehs/labs/epi/pi/genetics/pace/index.cfm)[136] is comprised of researchers around the world who are interested in studying the early life environmental impact on epigenetic changes and human disease. Through this ongoing collaboration, methylation data measured through Illumina HumanMethylation450 BeadChip from cord blood as well as older ages (up to adolescence) were available. The PACE consortium includes 39 birth and child cohorts from 14 counties in Europe, the USA, Canada, and Australia. The following cohorts were included in **Study I**, ALSPAC (UK), BAMSE (Sweden), CBC (USA), CHAMACOS (USA), CHS (USA), EDEN (France), EXPOsOMICS collaborative European project including (Envirogen (Belgium), PiccoliPlus (Italy), RHEA (Greece)), GECKO (The Netherlands), Gen3G (Canada), Generation R (The Netherlands), GOYA (Denmark), INMA (Spain), IoW (1st generation) (UK), IoW (2nd generation) (UK), MoBA1 (Norway), MoBA2 (Norway), MoBA3 (Norway), NFBC1986 (Finland), PIAMA (The Netherlands), PREDO (Finland), Project Viva (USA) and RAINE (Australia)[136]. A summary of included birth cohorts in **Study I** is presented in Table 2.

3.1.3 MeDALL consortium

Mechanisms of the Development of ALLergy (MeDALL) consortium[137] is an European Union-funded project (2010-2015) on the origin and mechanisms of IgE-associated allergic diseases such as asthma, allergic rhinitis, atopic dermatitis, and food allergy in children. The aim was to investigate how and which environmental factors influence the initiation of allergy. The MeDALL project includes six cohorts with younger children and seven older cohorts (up to adolescence) from 10 European countries. The following three European birth cohorts were included in **Study II**, BAMSE[133] (Sweden), GINIplus[138] (Germany), and INMA[139] (Spain). A summary of included birth cohorts in **Study II** is presented in Table 3.

3.1.4 KOLIN study

Respiratory and Cardiovascular Effects in COPD (KOLIN) is a mechanistic bronchoscopy-based cross-sectional study designed to investigate COPD, particularly the rapid lung function decline phenotype[140]. In the recruitment phase, KOLIN used data from the longitudinal Obstructive Lung disease In Northern Sweden (OLIN) COPD study to identify potential COPD subjects with rapid and non-rapid decline; ever-smoker and non-smoker controls with the aim to include 15 participants in each of the four groups. OLIN COPD is a prospective longitudinal population-based case-referent study within the OLIN studies[141]. The OLIN studies are part of an epidemiological research program including four adult cohorts; the first was founded in 1985 and the fourth in 1996 in Norrbotten, the northernmost part of Sweden. Initially, the project had a focus on asthma. However, it expanded to multiple research fields over the years, including COPD, allergy, and health economy. In 2002-2004, invitations for re-examination were sent to participants from four previous OLIN cohorts who had performed spirometry and/or completed structured interviews ($n \approx 4200$). A total of 993 subjects with airway obstruction ($FEV_1/VC < 0.7$) and sex- and age-matched controls were

included in OLIN COPD. In total n=1,986 individuals were followed every year from 2005 to measure spirometry, body mass index (BMI), and to answer structured interviews. At the follow-up in 2010, the first phase of KOLIN recruitment was conducted but did not identify enough COPD subjects. Thus, a second phase was conducted in 2012/2013. Participants who fulfilled the pre-determined criteria regarding lung function, annual FEV_1 decline, and smoking history in the OLIN COPD study population were specifically selected for potential bronchoscopy and in-depth clinical investigation and biomarker analyses. In total, 162 individuals were identified as potential KOLIN study subjects. Of these, 52 individuals were included in the final study population (Figure 4), but 15 non-smokers were excluded from **Study III** analysis, leaving 37 subjects to undergo bronchoscopy and other investigations.

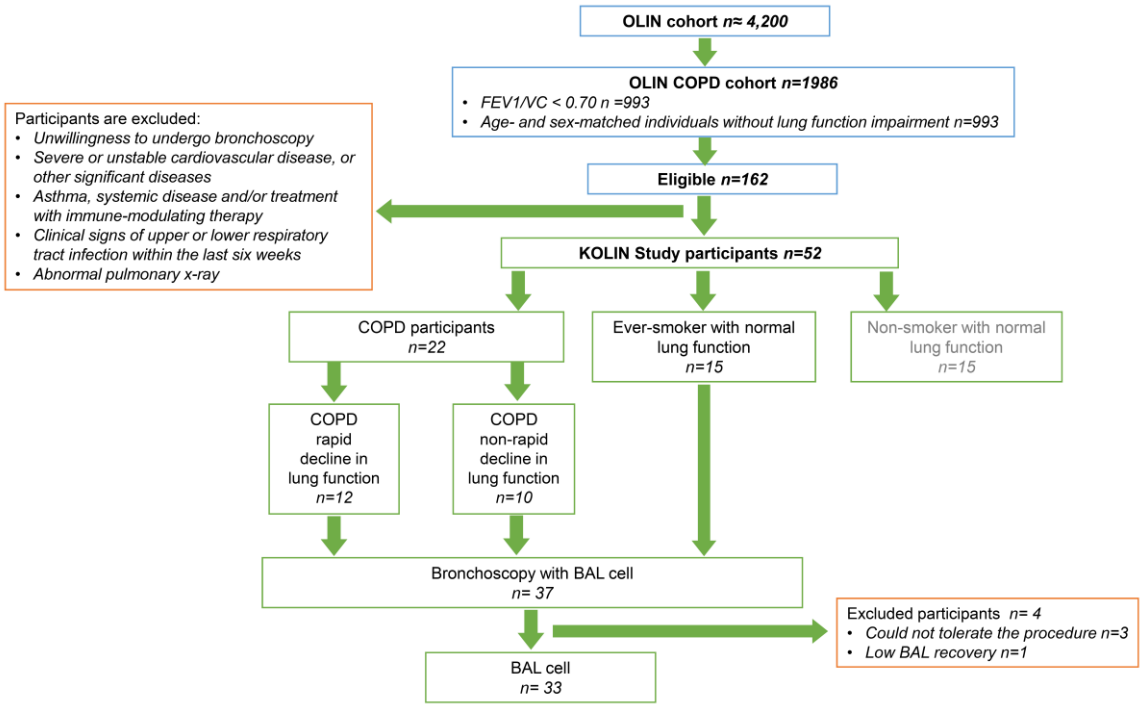

Figure 4. The flow chart of the KOLIN study, bronchoscopy and BAL cell collection.

The most common reason for non-participation among potential KOLIN subjects was an unwillingness to undergo bronchoscopy, severe or unstable cardiovascular disease, or other significant diseases like dementia, cancer, and porphyria. In addition, participants were excluded if they had asthma, systemic disease, and/or treatment with immune-modulating therapy, or clinical signs of upper or lower respiratory tract infection within the last six weeks. The participants underwent clinical examination, spirometry, electrocardiogram, routine blood tests, and bronchoscopy (for collecting BAL, bronchial wash (2 x 20 ml), and biopsies).

3.1.5 FASTX study

The FASTX (Food Allergen Suppression Therapy with Xolair®) study is an open one-armed exploratory phase-2 study of pOIT combined with omalizumab treatment. The study includes peanut-allergic adolescents (n=23) with documented anaphylactic reactions to peanuts within the last five years and with a positive peanut basophil activation test (BAT/CD-sens[142,143]). Twenty-three participants with severe allergy were selected out of 41 screened participants. The excluded 18 participants had very high or low total serum IgE levels, or negative BAT/CD-sens to peanut and concordant allergens. The recruitment of patients was performed at the outpatient allergy clinic at Sachs' Children and Youth Hospital or referred by pediatric allergists in the Stockholm area to the study team. However, some patients or parents contacted the study team directly after hearing or reading about the listed clinical trials on the clinicaltrials.gov website. In addition, all the patients were required to have a concomitant allergy to either pollen or pets. Information regarding included patients and clinical outcomes has previously been described in detail[144-146]. For the study outline, see Figure 5.

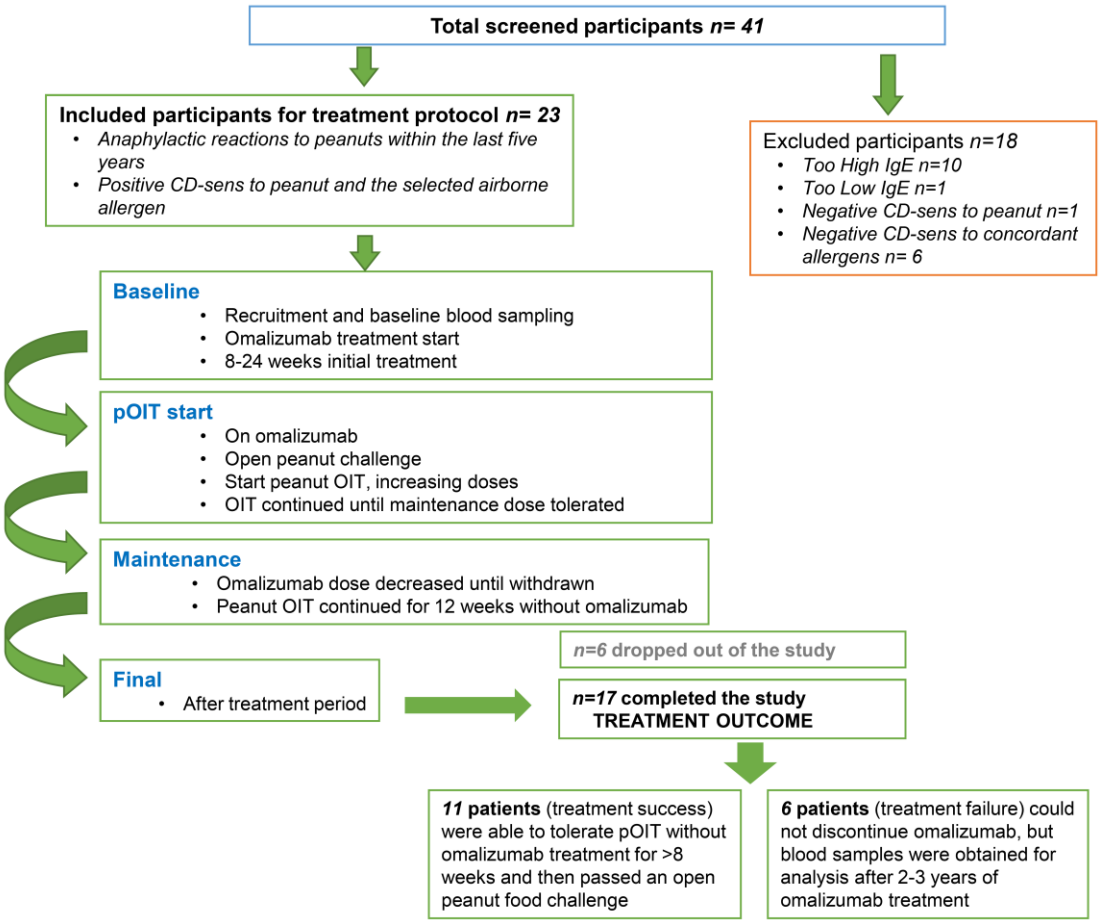

Figure 5. The flow chart of the FASTX study treatment protocol.

The study was divided into two parts. First, the patients were treated with omalizumab at individual recommended dosages (dosages for allergic asthma) for an initial 8-week period. For some patients, additional periods were needed until the BAT/CD-sens analysis showed suppressed reactivity to peanut[145,146]. In the second part, all patients underwent an open peanut challenge (pOIT start) before starting peanut OIT (pOIT) at 280 mg of peanut protein combined with omalizumab treatment. The peanut dose was gradually increased every two weeks, if tolerated, until reaching a maintenance dose of 2800 mg. After eight weeks on the maintenance dose, patients decreased the omalizumab dose by 50% (Maintenance). After that, they continued to reduce the omalizumab by 50 % every eighth week if pOIT was tolerated and BAT/CD-sens suppressed. Eleven patients (treatment success) were able to tolerate pOIT without omalizumab protection for >8 weeks and then pass an open peanut food challenge (Final). Six patients (treatment failure) could not discontinue omalizumab, but blood samples were obtained for analysis after 2-3 years of omalizumab treatment (Final). Six patients dropped out of the study (the treatment protocol is summarized in Figure 5).

3.2 STUDY POPULATIONS

The study population of **Study I** was based on the PACE consortium, including 11,000 participants in 26 independent cohorts with available cord and/or peripheral blood DNA methylation as well as information on gestational age. Women with multiple births and gestational age of more than 42 weeks (294 days) were excluded. In total, 20 cohorts with data on newborns (n = 6,885), four cohorts in early childhood (4–5 years; n = 736), five cohorts at school age (7–9 years; n = 1,445), and five cohorts in adolescence (16–18 years; n = 1,934) were included. However, after excluding participants with maternal complications (i.e., maternal pre-eclampsia, diabetes, or hypertension) and cesarean section delivery or delivery start with induction, 17 cohorts of newborns (n = 3,648), four cohorts in early childhood (n = 453), five cohorts in school-age (n = 899) and five cohorts in adolescence (n = 1,129) were included in the final analyses. Four of the included cohorts had data both at birth and at an older age (Table 2).

Table 2. A summary of included birth cohorts from the PACE consortium.

								After exclusion of mothers who experienced pregnancy complications and those in whom labor was induced or delivered by caesarian section			
Cohort	Country	Years of recruitment	Number of children at recruitment	Newborn (N)	Early childhood (N)	School -age(N)	Adolescence (N)	Newborn (N)	Early childhood (N)	School -age(N)	Adolescence (N)
ALSPAC	UK	1991-92	14,541	423		461	457	249		273	272
BAMSE	Sweden	1994-96	4,089		198	207	212		145	141	159
BAMSE EpiGene	Sweden	1994-96	4,089			302				232	
CBC (Hispanic)	USA	1982-2009	1,200	161				128			
CBC (Caucasian)	USA	1982-2009	1,200	175				132			
CHAMACOS	USA	1999-2000	601	372				110			
CHS	USA	1995-97	5,341	228				120			
EDEN	France	2003-06	2,002	150	130			100	89		
EXPOSOMICS				376				252			
EXPOSOMICS (Environage)	Belgium	2002-04	1,196								
EXPOSOMICS (PiccoliPlus)	Italy	2011-15	3,338								
EXPOSOMICS (RHEA)	Greece	2007-08	1,500								
GECKO	The Netherlands	2006-07	2,874	204							
GENERATION R	The Netherlands	2002-06	9,901	763				486			
Gen3G	Canada	2010-13	1,034	167							
GOYA	Denmark	1996-2002	91,387	497							
INMA	Spain	1997-2008	3,768	370	191			134	71		
IoW F1	UK	1989-90	1,456				102				97
IoW F2	UK	2012-17	420	121				93			
MoBa1	Norway	1999-2009	11,4479	964				749			
MoBa2	Norway	1999-2009	11,4479	609				460			
MoBa3	Norway	1999-2009	11,4479	213				177			
NFBC1986	Finland	1985-86	9,362				494				287
PIAMA	The Netherlands	1996-97	3,963		217	199			148	134	
PREDO	Finland	2006-10	1,079	775				308			
Project Viva	USA	1999-2003	2,128	317		276		150		119	
RAINE	Australia	1989-91	2,868				669				314
Total			278,527	6,885	736	1,445	1,934	3,648	453	899	1,129

In **Study II,** 656 participants from the MeDALL consortium of three European birth cohorts, BAMSE (n=244), GINIplus (n=247), and INMA (n=165), with available gene expression and $PM_{2.5}$ air pollution exposure were included. Moreover, matched DNA methylation data were available from BAMSE (n=240) and INMA (n=103) cohorts. (Table 3).

Table 3. A summary of included birth cohorts from the MeDALL consortium.

Cohort	Country	Years of recruitment	Number of children at recruitment	DNA methylation data in final analyses (N)	Matched gene expression in final analyses (N)
BAMSE	Sweden	1994-96	4,089	244	240
INMA	Spain	1997-2008	3,768	247	
GINIplus	Germany	1995-98	5,991	165	103
Total			13,848	656	343

Study III was based on the cross-sectional KOLIN study of 18 COPD subjects and 15 controls (ex- and current smokers with normal lung function) who underwent bronchoscopy to collect BAL cells for cellular analyses and DNA extraction.

The final **Study IV** comprised peanut-allergic adolescents (n=17) who had been treated with omalizumab for eight weeks, followed by a stepwise increase of daily peanut ingestion and subsequent withdrawal of omalizumab. Finally, an open peanut challenge was performed. Peripheral blood cells were collected before and three times during pOIT.

3.3 EXPOSURE AND OUTCOME ASSESSMENTS

3.3.1 Gestational age

In **Study I**, gestational age in days was obtained from each cohort based on birth certificates, and medical records using ultrasound estimation or last menstrual period date, (or combined estimate). Otherwise, this information was extracted from self-administrated questionnaires.

3.3.2 Outdoor air pollution

In **Study II**, annual average traffic-$PM_{2.5}$ concentrations were estimated at the home address at birth and at the time of bio-sampling for each study participant in the MeDALL cohorts using land-use regression (LUR) models developed through the ESCAPE (European Study of Cohorts for Air Pollution Effects) project[147]. A total of 20 PM sampling sites were selected in each study area to characterize the spatial distribution of air pollution and residential addresses of cohort participants in these areas. In each site, the measurements were performed three times during two weeks in the cold, warm, and intermediate seasons, and the results

were averaged to estimate the annual average after adjusting for temporal variation using a background reference site located centrally. LUR models were developed for $PM_{2.5}$ based on measured yearly average concentrations and additional predictors of geographic variables from both European and local databases. To explain the spatial variation of $PM_{2.5}$ concentrations in the model, supervised forward stepwise procedures were performed by including additional predictors such as digital road network, land use, population density, altitude, and study-specific local data. In addition, traffic intensity on the nearest street and traffic load on all major roads within a 100m buffer were considered in the model. Modeling was done locally at each center using a common exposure assessment manual (http://www.escapeproject.eu/manuals/) following harmonized procedures regarding air pollutants measurements, development of land-use regression models, and validation[147]. The models were then used to estimate annual average $PM_{2.5}$ concentrations at the birth addresses, as well as the current addresses at the time of blood sampling.

3.3.3 COPD

In **Study III**, the participants' COPD status was assigned using GOLD stage 2-3[148] ($FEV_1/VC < 0.70$ and FEV_1 30-80% of predicted) with a rapid decline (FEV_1 decline ≥60 ml/year) or non-rapid decline (FEV_1 decline ≤30 ml/year). Additionally, all the participants had a smoking history of at least ten pack-years at baseline. Conversely, the control groups had a similar smoking history at baseline but had normal lung function ($FEV_1/VC \geq 70\%$ and $FEV_1 \geq 80\%$ of predicted) and a decline in FEV_1<20 ml/year.

3.3.4 Peanut oral immunotherapy

In the final **Study IV**, pOIT combined with omalizumab was given to peanut-allergic adolescents with documented anaphylactic reactions to peanuts and positive for peanut-specific BAT. At baseline, the patients were first challenged with 0.1 mg of peanuts proteins and then, the dose was increased every 30 minutes (1 mg, 10 mg, 100 mg, 1 g, and 10 g) to confirm that they reacted to peanuts clinically. The challenge was only stopped if anaphylaxis was on the verge of breaking out. Moreover, the physician limited the dose or dosing intervals depending on the patient's situation. However, in the absence of anaphylaxis, no medication was given besides the common practice; instead, patients were under very close supervision with emergency treatment drugs readily available.

Detailed information about the treatment procedure can be found in section 3.1.5.

3.4 OTHER COVARIATES

The covariate variables for **Studies I-IV** are summarized in Table 4. The definitions of covariates in **Study I and II** were harmonized across all included cohorts in the PACE and MeDALL consortia, respectively. In **Study III** and **Study IV,** the covariates were defined according to KOLIN and FASTX study protocols, respectively.

Table 4. Definitions of covariates in studies I to IV.

Variable	Variable definition	Study
Sex	Male Vs. female	I, II, III
Age	Participant's age at the follow-up	I, II, III
Maternal smoking	Maternal smoking during pregnancy has three categorizations: no smoking in pregnancy, stopped smoking in early pregnancy and smoking throughout pregnancy, but a binary categorization of any versus no smoking was also defined	three categories or binary in I; and only binary in II
Second-hand smoking	Parental smoking ≥1 cigarette daily at the time of the respective follow-up questionnaire	II
Active smoking status	Any smoking during adolescence	II
Pack-years	Number of cigarettes smoked per day/20 cigarettes multiplied by the number of years smoked	III
Smoking intensity	How many cigarettes/days smoked	III
Maternal age	Mother's age at delivery	I
Maternal socioeconomic status	Variable defined by each cohort of maternal education or income definition	I
Maternal education	Based on the highest educational level of the mother. Divided into low –elementary school or 2-year high school; medium – 3-year high school; high – University	II
Parity	Having no previous children, one or more previous children	I
Birth weight	Birthweight of the child in grams	I
BMI	BMI at the time of blood sample	II
Physical activity	Physical activity < 5 hours per week Vs. ≥5 hours per week	II
Season of blood sampling	The season when blood samples were taken (Winter, Spring, Summer, or Autumn)	II
Inhaled corticosteroid	inhaled corticosteroid usage: Yes/No	III
Treatment outcome	Treatment success declares if the participants were able to tolerate pOIT without omalizumab for more than eight weeks and then pass an open peanut food challenge. Otherwise, with treatment failure, the patients could not discontinue omalizumab for 2-3 years.	IV
Study region	The study regions included in BMASE were Stockholm, Järfälla, Solna, and Sundbyberg. GINIplus only contains Munich; INMA consists of Gipuzkoa and Sabadell regions.	II
Batch	Each cohort defines the batch variable or uses surrogate variable analysis for batch effect correction	I, II, III
Cell type	Measure/Estimate cell composition	I, II, III, IV
Additional covariates	Cohorts that have an ancestry and/or selection covariates	I

3.5 SPIROMETRY

The participants' spirometry was measured following the ATS/ERS guidelines[149] using a dry volume spirometer (Mijnhardt Vicatest 5, the Netherlands) in **Study III**. The Swedish spirometric reference values were applied[150]. The highest value of forced vital capacity or slow vital capacity was defined as vital capacity (VC). If a participant showed values FEV_1 < 80% of predicted or $FEV_1/VC < 0.70$, then a reversibility test (after administration of 400 micrograms of salbutamol) was performed. The highest value of pre- and post-bronchodilatation FEV_1 and VC was used to define COPD.

3.6 BRONCHOSCOPY

The same medical team performed bronchoscopies in **Study III** at Luleå Hospital (the Division of Respiratory Medicine and Allergy, Department of Medicine, Sunderby Central Hospital of Norrbotten Luleå, Sweden) and Umeå Hospital (the Division of Respiratory Medicine and Allergy, Department of Medicine, University Hospital, Umeå, Sweden). Premedication was given to the participant 30 min before the procedure with 1.0 mg of atropine, but some received midazolam 4–8 mg per os. Lidocaine was used to achieve local

anesthesia. A flexible video bronchoscope was inserted into the participant through the mouth via a mouthpiece in a supine position. BAL was performed by infusing three aliquots of 60 ml of sterile sodium chloride (0.9%), pH 7.3, at 37 °C in the middle or lingula lobe. The fluid was gently sucked back after each infusion and pooled into a tube placed in iced water. The recovered BAL fluid was transported to the laboratory for immediate analysis.

3.7 DNA METHYLATION ANALYSIS

Different technologies have been developed throughout the years to investigate DNA methylation. In this thesis, genome-wide DNA methylation was assessed using two Illumina array-based BeadChip platforms. Detailed descriptions of the two methods are in the following sections.

3.7.1 Illumina Infinium Methylation450 BeadChip

Cord and peripheral blood samples were collected and stored according to standard procedures in the PACE consortium for 11,000 participants from 26 cohorts in **Study I.** The extracted DNA was bisulfite converted using the Zymo EZ DNA MethylationTM kit (Zymo, Irvine, CA). Following conversion, the genome-wide methylation status of over 485,000 CpG sites was measured using the Illumina HumanMethylation450k BeadChip after randomizing the samples in the 96-well plates. GenomeStudio Software processed the raw methylation intensities. The level of methylation expressed as β value ($\beta = M/[c + M + U]$), the proportion of intensity of methylated (M) over the sum of methylated (M) and unmethylated (U) probes, c is a constant. Each cohort used its quality control, normalization, and batch correction pipeline from the standard Bioconductor R packages. In the same way, cord blood DNA methylation levels from 38 newborns (open access look-up dataset) were bisulfite converted and quality accessed using Illumina HumanMethylation450k BeadChip (see GEO platform: GSE62924, https://www.ncbi.nlm.nih.gov/geo/)[151]. More details about sample selection, quality controls, and normalization procedures, including batch correction, can be found in the original paper.

3.7.2 Illumina Infinium MethylationEPIC BeadChip

In **Study III**, the genome-wide methylation status of over 866 836 CpG sites was measured using DNA extracted from BAL cells of 18 COPD subjects and 15 controls (ex- and current smokers with normal lung function). The extracted DNA (500 ng) underwent bisulfite conversion using the Zymo EZ DNA MethylationTM kit (Zymo, Irvine, CA). The 96-well plates were used to place the sample after randomization and measures by Illumina MethylationEPIC BeadChip. The process of raw methylation intensity and measure of methylation level was similarly performed as stated in the previous 3.7.1 section. The standard Bioconductor R packages were used for quality control, normalization, and batch correction. Similarity, BAL cells from 19 smoker and non-smoker individuals' methylation (open-access look-up dataset) were processed with MethylationEPIC BeadChip (see GEO platform: GSE133062, https://www.ncbi.nlm.nih.gov/geo/)[152]. More detailed information about sample selection and quality controls can be found in the original paper.

3.8 GENE EXPRESSION ANALYSIS

A range of technologies has been developed for the purpose of assessing the gene expression profile. In this thesis, array-based and sequence-based technologies were used to determine the differential gene expression levels, as described in the following section.

3.8.1 Microarray

In **Study II**, whole blood was collected from three MeDALL consortium cohorts, i.e., BAMSE (n=244, mean age of 16.7 years), GINIplus (n=247, mean age 15.2 years), and INMA (n=165, mean age 4.5 years) by PAXgene tubes, and extraction of RNA was processed batch-wise using an extraction kit (QIAGEN, Courtaboeuf, France). Expression levels were quality assessed and hybridized by Human Transcriptome Array 2.0 Genechips (HTA 2.0, Affymetrix). In **Study I**, cord blood gene expression levels from 38 newborns (look-up) were determined using Affymetrix Human Gene 2.0 Array (see GEO platform: GSE48354, https://www.ncbi.nlm.nih.gov/geo/)[153]. More detailed information about sample selection and quality controls can be found in the original paper.

3.8.2 RNA-sequencing

In **Study IV**, peripheral blood cells from (n=17 adolescents, age 12-18 years) were collected before and three times during pOIT combined with omalizumab in the FASTX study using PAX gene blood RNA tubes (PreAnalytix/Qiagen) and stored at −80 °C. RNA was extracted by the PAXgene Blood miRNA kit (PreAnalytiX, QIAGEN, Inc., Germantown, MD, USA) and purified with TruSeq Stranded Total RNA using the Qiagen Fastselect human rRNA and Globin removal. RNA sequencing was performed in one batch by 150bp paired ends using the Illumina NovaSeq-6000 system. RNA sequencing reads were filtered based on their quality with FASTQC and trimmed for adapters using TrimGalore (cutadapt 2.8). The quantification of gene-level counts was performed by featureCount 1.5.1. In **Study III**, RNA extracted from BAL cells from 19 smoker and non-smoker individuals (look-up) was sequenced by 125 bp paired-end on an Illumina HiSeq 2500[152]. More detailed information about sample selection and quality controls can be found in the original paper.

3.9 ANNOTATIONS AND BIOINFORMATICS RESOURCES

3.9.1 Annotations

In **Studies I-III**, CpG sites were annotated to the target gene using Illumina's annotation based on the UCSC database. However, in **Study I,** we enhanced annotation for nearest genes within 10 Mb of each site by the program Snipper[154]. On the other hand, in **Study III,** for those CpG sites where no target gene was available, the nearest gene was annotated using GREAT version 4.0.4[155]. In **Study II**, the probes in Illumina 450k were assigned to the gene by taking the average value mapped within 200 bp of the TSS. If no probes were found within a 200 bp window, but with the first exon, the corresponding average of probes mapping to the 1st exon of the gene was assigned. Finally, for a gene with no 200 bp of the TSS and first exon probes, we used the average over probes within 1500 bp of the TSS.

The annotation of transcript clusters of gene expression in **Study II** was performed by NetAffx annotation version 36. Whereas RNA-seq in **Study IV,** the reads were aligned to the GRCh38 human reference genome assembly.

3.9.2 Bioinformatics resources

In this thesis, we used multiple available online bioinformatic resources. The High-quality INTeractomes (HINT) database (http://hint.yulab.org/)[156] was utilized to integrate DNA methylation and gene expression in **Study II**. This contains protein-protein interactions (PPIs) from 8 interactome resources (BioGRID, MINT, iRefWeb, DIP, IntAct, HPRD, MIPS, and the PDB). The two types of PPIs are binary physical and co-complex associations from different organisms. After removing duplicates and self-link, 150,199 PPIs remained.

In addition, functional enrichment analyses of identified genes from **Studies I-IV** were conducted making use of two functional bioinformatic resources based on gene ontology (GO, http://geneontology.org/)[21,157] and Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/)[22]. GO terms describe biological processes, molecular functions, and cellular components. The KEGG pathways include known knowledge of molecular interaction, reaction and network relation for human diseases, metabolism, cellular processes, genetic and environmental information processing, organismal system, and drug development[22]. Furthermore, the Immunological Genome Project RNA-seq dataset (ImmGen, https://www.immgen.org/Databrowser19/DatabrowserPage.html)[158] was explored in **Study III** to evaluate differentially methylated genes from BAL cells related to changes in macrophages.

3.10 ETHICAL CONSIDERATIONS

All studies included in this thesis followed the standard study protocols with Helsinki Declaration to adhere to high standards of ethical conduct.

In BAMSE, informed written consent was given by the parents at each of the follow-ups between ages of two months up to 16 years. Similarly, in the FASTX study, informed consent was provided by the parents, but if the participant was older than 18, informed consent was given by the participant him/herself. In KOLIN, the participants gave informed consent (oral and written). In all studies information ensuring participants' right to withdrawal and confidential treatment of their collected data was given. All data referring to personal identity in BAMSE and FASTX were separated from the original database and stored in a different database highly protected by firewalls to promote data security. Furthermore, no individual data on study subjects are available to persons outside the study team, and biological samples are stored in locked freezers and utilized according to standards set by Karolinska Institutet. Similarly, in the KOLIN study, consent forms and questionnaires were stored locally on paper, and datasets were stored on a select few local computers protected by firewalls in accordance with Umeå university guidelines, on a network-attached storage only accessible to authenticated users. The same guidelines apply to our collaborative cohorts from MeDALL and PACE (albeit with specific national/local rules – see below).

In BAMSE, the participants received a gift certificate of 500 SEK at the 16-year follow-up as reimbursement for the inconvenience and time during the clinical examination. The participants have received feedback on their clinical examination results (IgE tests, lung function and blood pressure) and in case of any abnormal results, they are advised to seek appropriate public health care for further consultation. No information about genetics, epigenetics, or transcriptomics analyses results were given to the participants, since these data were considered exploratory and not ready for interpretation or use at the individual level.

In the FASTX study, only severely allergic adolescents were included to get a homogenous study population and to offer potentially disease-modifying treatment to those who would benefit most. Moreover, adolescents, not younger children, were enrolled in the study to get willing participation as a good ethical practice, less trouble during the treatment, and a reliable report of potential adverse events as young children have difficulty reporting subjective symptoms. All allergic reactions were documented in the participant's medical records, according to clinical practice.

Clinical examination, spirometry, electrocardiogram, routine blood tests and bronchoscopy reports in KOLIN study were documented in the participant's medical records, according to clinical practice. All subjects received a study-specific identification code, and the code key was stored behind locked doors within the premises. To guarantee participants' confidentiality and anonymity, data files distributed to researchers were de-identified and included only the study-specific identification code. DNA methylation data has been uploaded to the Gene Expression Omnibus database repository (GSE198870) without the identification code. Participants who underwent bronchoscopy were offered financial compensation for taking part in time-consuming examinations. Undergoing the clinical examinations and the above different procedures might uncover diseases and/or medical conditions previously unknown to the subject. While knowledge about these conditions might be unwanted, the benefit of early diagnosis was assessed to exceed the possible harm. In KOLIN, pathological findings have been evaluated by physicians, and subjects were informed about the findings and offered referral to the appropriate public health care for further consultation.

In this thesis, we also used data from two big consortia, MeDALL and PACE. Each participating cohort followed local/national medical ethical guidelines on recruiting and follow-up of their participant with a described study design, inclusion criteria, enrolment, and data collection. The PACE consortium has a bottom-up approach with a core principle of scientific excellence, transparency, collaboration, disclosure with mutual trust, and confidentiality. Each researcher signed a confidentiality statement to keep all exchange information confidential. Moreover, each cohort performed independent analysis according to a common, pre-specified analysis plan and sent summary statistic results for meta-analysis for greater power and novel discovery, or they signed a data transfer agreement for analysis based on the General Data Protection Regulation (GDPR) principle of processing and storing of data. If a participant in any of the included cohorts in this thesis no longer wants his/her information to be available, individual-level data can be expunged.

In Studies I-II, ethical approval has been received for the BAMSE study from the regional ethics committee at Karolinska Institutet, Stockholm, Sweden (reference numbers: 93-189, 98-175,01-475, 02-420, 2010/1474-31/3, and 2011/2037-32). In **Study III**, the KOLIN study was approved by the Regional Ethical Review Board at Umeå University, Sweden (reference numbers: 2011-147-31M and 2017-91- 32M). Moreover, the KOLIN study was registered at ClinicalTrails.gov; NCT02729220. The ethical committee in Stockholm, Sweden approved FASTX **Study IV** (reference numbers: 2013/827-31/3, 2014/1980-32, 2016/1390-32, 2020–00807), and the Swedish Drug Agency also approved the study (reference numbers: 5.1–2013–46183). The FASTX study was registered at EudraCT: 2012–005625–78 and ClinicalTrails.gov; NCT02402231. For the **Study I** PACE consortium and **Study II** MeDALL consortium, local ethical committees in each country approved the study protocol and data collection.

3.11 STATISTICAL ANALYSIS

All analyses were carried out using R version 3.4.0 or later[159] and Bioconductor packages[160] unless stated otherwise.

3.11.1 Presenting the study characteristics

The study participant's characteristics were presented as frequency and percentage total in **Studies I-IV**. For continuous variables, mean, median, standard deviation, IQR, and minimum and maximum values were presented in **Studies I-IV**. Fisher exact test and the Mann-Whitney U test were used in **Study III** to assess the difference between COPD cases and controls.

3.11.2 Pearson correlation

Correlation analysis in **Study I** and **Study II** was performed using the Pearson correlation test.

3.11.3 Covariate assessment

The potential covariates in **Studies I-IV** were assessed using principal component regressions[161].

3.11.4 Epigenome-wide association study

The EWAS analysis was performed by robust linear regression (rlm in the MASS R package[162]) to estimate the association between DNA methylation and gestational age in **Study I**. Robust linear regression accounts for the potential heteroscedasticity in the methylation data. In the same way, the association between DNA methylation and COPD in **Study III** applied the same regression model. In both studies, potential confounders and cell types were adjusted for. Detailed information is shown above or can be found in the original papers.

3.11.5 Differentially methylated regions

Differentially methylated regions (DMR) in **Study I** were analyzed using comb-p[163] and DMRcate[164] for meta-analysis of EWAS results. In **Study III,** only DMRcate was applied to identify DMR from the associated EWAS result. Input parameters used for the DMR calling are provided in the original papers.

3.11.6 Transcriptome-wide association study

The differential gene expression analysis in **Study II** was performed using the limma package in R[165]. The RNA-seq data in **Study IV** used DESeq2[166] to identify differentially expressed genes. Detailed information about the adjusted variables is shown above; or can be found in the original papers.

3.11.7 Cell type correction

The relative proportions of white blood cells from DNA methylation in **Studies I-II** were estimated by the Houseman method[167] using the estimateCellCounts function in the Minfi package in R[168]. The cord blood analyses used the Bakulski reference[169] to estimate the proportions of 7 blood cell types (nucleated red blood cells, CD4+ T-lymphocytes, CD8+ T-lymphocytes, NK (natural killer) cells, B-lymphocytes, monocytes, and granulocytes). For the whole blood analyses, Reinius reference[170,171] was used to estimate the proportions of 6 cell types (CD4+ T-lymphocytes, CD8+ T-lymphocytes, B-lymphocytes, monocytes, granulocytes, and natural killer cells). In contrast, the RNA-seq data in **Study IV** estimated the cell composition via deconvolution through the CIBERSORT method[172]. In **Study III,** the actual cell type counts were measured for BAL cells which include macrophages, lymphocytes, neutrophils, eosinophils, and mast cells.

3.11.8 Meta-analysis

The study-specific effect estimates could be combined in a meta-analysis using a weighted inverse of the variance. The combined estimates can be calculated either by fixed effect with METAL[173] or random effect with METASOFT[174] tools. In the fixed-effect meta-analysis, a similar true effect common to all the studies is assumed whereas, the random effect meta-analysis considers within- and between-studies variation while estimating the mean of a distribution of effects. In **Studies I** and **II**, effect estimates from cohort-specific EWAS results were subsequently included in a fixed-effect meta-analysis. In the same way, cohort-specific results of the TWAS were also included in a fixed-effect meta-analysis in **Study II**. Additionally, in **Study I**, a random-effects model was performed considering between studies' heterogeneity by the I^2 statistic[175]. $I^2 > 50\%$ was defined as a high level of between-study variation and replaced fixed-effect values with random-effects estimates.

3.11.9 Longitudinal analysis

A linear mixed model was applied for longitudinal DNA methylation with gestational age in **Study I** and RNA-sequence of pOITs in **Study IV** by considering the within-person time

effect. The models were adjusted for potential confounders in each study shown above or can be found in the original papers. In **Study I**, an interaction term with time was included to assess the impact of methylation change over time per day increase in gestational age at delivery.

3.11.10 Epigenetic aging

In **Study IV**, the DNA methylation age was calculated using an open-access tool developed by Horvath[176]. The recommended normalization (preprocessQuantile)[168] and analysis options stated in the software tutorial were applied. A set of variables, including different measures of biological age and epigenetic age acceleration, was returned by Horvath's epigenetic age calculator.

3.11.11 Omics analysis

In recent years, different methodological approaches have been developed to integrate many omics platforms and assess their biological meaning. In this thesis, two approaches were used to find the link between DNA methylation and gene expression and describe their biological meaning. Moreover, the colocalization of identified CpGs with SNPs is also stated in the following sections.

3.11.11.1 Correlation of DNA methylation and gene expression

DNA methylation and gene expression correlations were tested using publicly available paired measured mRNA gene expression (Affymetrix Human Transcriptome Array 2.0) and DNA methylation (Illumina Infinium® HumanMethylation450 BeadChip assay) of cord blood[151,153] in **Study I** and mRNA gene expression (RNA-seq) and DNA methylation (Illumina Infinium® MethylationEPIC BeadChip assay) of BAL cell[152] in **Study III** respectively. The transcript levels of genes within the 500kb region of the significant CpGs were tested (250 kb upstream and 250 kb downstream). First, the residuals for mRNA expression and residuals for DNA methylation were created, and then correlations between expression residuals and DNA methylation residuals were evaluated using Pearson correlation and linear regression. These residual models were adjusted for potential confounders, as stated in the original papers.

3.11.11.2 Integration of DNA methylation and gene expression

In **Study II**, an integrated genome-wide DNA methylation meta-analysis with matched transcriptome-wide gene expression meta-analysis on $PM_{2.5}$, along with PPI, was performed using the Functional Epigenetic Module (FEM) algorithm in R package[87]. The FEM algorithm first constructs an integrated network with weights on the network edges from the associations' analysis of $PM_{2.5}$ and both gene expression and DNA methylation. Afterward, the inference of the FEMs as heavy subgraphs on this weighted network[87]. The integration of DNA methylation and gene expression profiling was performed after constructing the PPI network for hub gene identification from the HINT database.

3.11.11.3 Colocalization of differentially methylated CpGs with GWAS

In **Study IV**, the evidence of colocalization of genetic and epigenetic variation was evaluated by comparing COPD-associated CpGs with the 82 SNPs from a recently published large-scale GWAS of 35,735 COPD cases and 222,076 controls[102]. The COPD-associated CpGs were located within a 1Mb window (500 kb upstream and 500 kb downstream) surrounding the 82 SNPs. In addition, SNPs and CpGs located in the same gene were checked.

3.11.12 Functional and enrichment analysis

To gain insight into the functional and biological relevance of differentially methylated and expressed genes in **Studies I-IV**, enrichment analysis of gene ontology and KEGG pathway was performed. Overrepresentation analysis (ORA) in **Study I** and **Study III** was applied using ConsensusPathDB tool[177,178] (http://consensuspathdb.org). In **Study II,** a network-based pathway annotation tool BinoX[179] in PathwAX II web server (http://pathwax.sbc.su.se/)[180] was used to perform the KEGG pathway enrichment analysis. Enrichplot R package[181] was employed when visualizing the functional enrichment of GO biological processes in **Study IV**. ORA p-values were calculated using a hypergeometric test, and network-based pathway enrichment used binomial distribution to calculate the p-value of interactions between genes and pathways after randomizing the network 1,000 times.

The two-sided doubling mid p-value of the hypergeometric test was also used to assess the enrichment of colocalization deferentially methylated CpGs for several biologic annotations provided in Illumina Array in **Study I** and **Study III**. Likewise, the enrichment of genes associated with pOIT and genes associated with peanut allergic reactions in Watson *et al.*[1] in **Study IV**. Furthermore, the enrichment analysis of birth weight EWAS findings[182] and the gestational age EWAS in **Study I** used the same method.

3.11.13 Multiple testing

In statistical hypothesis testing, there is a 5% chance of incorrectly rejecting the null hypotheses (Type I error), even though all the null hypotheses are true. The probability of false positive results increases with an increase in the number of null hypothesis tests. When controlling for multiple comparisons, the number of false positives reduces with the cost of power, which means an increase in false negatives (Type II error). Several methods currently exist to control multiple testing by maximizing power and ensuring an acceptable Type I error rate. The Bonferroni correction is a method for controlling multiple tests by dividing the nominal p-value by the number of tests performed. Bonferroni correction is shown to be too conservative in many settings. It assumes independence between the tested associations, which is usually not the case in omics analysis where the correlation between tested variables is more common. A slightly less conservative method, a one-step Šidák correction[183] that controls the family-wise error rate, was also developed. Even a more relaxed framework like Benjamini-Hochberg FDR correction[184] has frequently been used in omics analyses.

In **Study I** and **Study III**, a strict Bonferroni correction for differential methylation analysis and the correlations between DNA methylation and gene expression was applied. However, the DMR analysis used a one-step Šidák correction for Comb-p and Benjamini-Hochberg FDR correction for DMRcate. The differential gene expression analysis in **Study II** and **Study IV** and integration analysis of genome-wide DNA methylation with matched gene expression in **Study II** applied FDR correction. Similarly, the enrichment p-values in **Studies I-IV** were adjusted for multiple testing using FDR. The multiple testing p-value <0.05 was considered statistically significant.

4 RESULTS AND DISCUSSION

4.1 STUDY I: THE INFLUENCE OF GESTATIONAL AGE ON DNA METHYLATION

In this study, we investigated the influence of gestational age in days on blood DNA methylation in the large PACE consortium meta-analysis. In total, 3,648 newborns from 17 cohorts and 2,481 older children from ten cohorts (aged 4-18 years), including four cohorts with data both at birth and at an older age, were included in the main "no complication model" after excluding participants having pregnancy complications like maternal diabetes, hypertension or pre-eclampsia, or cesarean section delivery or delivery starting with induction. Detailed information about the participating cohorts and their characteristics, including the gestational age range, can be found in **Study I** (Table 1).

We identified 8,899 differentially methylated CpGs annotated to 4,966 genes in cord blood DNA methylation associated with gestational age at birth across the genome at Bonferroni-significance $p<0.05$ with a somewhat more negative (60%) than positive (40%) direction of effect. For functional downstream analyses, we selected the loci that included at least three adjacent CpGs that survived Bonferroni correction, and those fulfilling the section criterion were in total 1,276 CpGs annotated to 325 unique genes. Similarly, we observed a slightly more frequent negative (55%) than positive (45%) direction of effect (Figure 6).

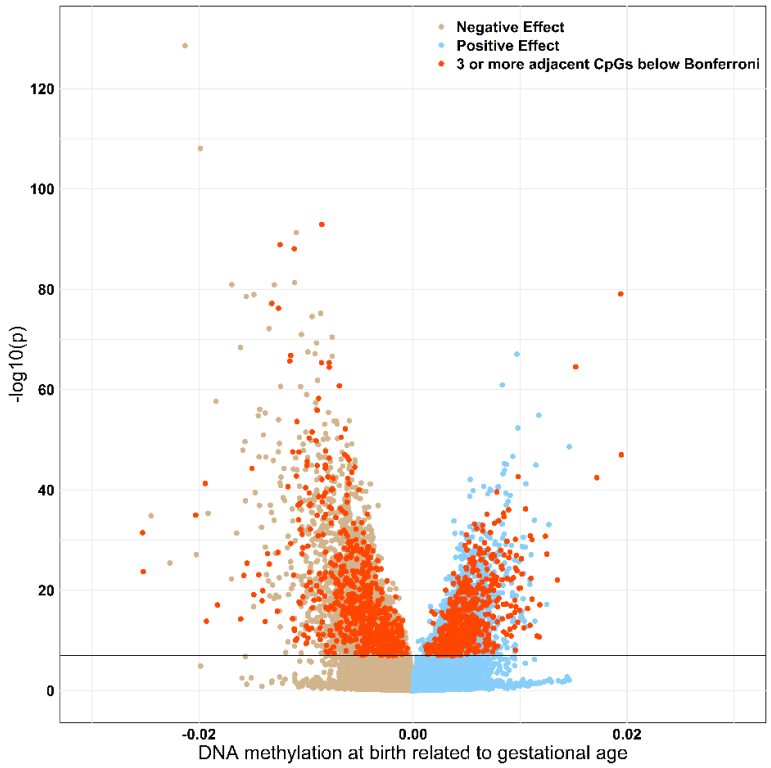

Figure 6. Volcano plot for the meta-analysis of DNA methylation at birth related to gestational age per week. X-axis: effect size estimates in weeks; y-axis: $\log_{10}$(p-value).

The DMR analysis using two independent approaches (comb-p and DMRcate) revealed numerous unique regions associated with gestational age at birth. Moreover, the identified CpGs were concordant with the larger dataset of 6,885 participants from 20 cohorts "all births model" without excluding maternal complications and cesarean section delivery or induced delivery. In addition, when the data were restricted to term birth only, we observed consistent results, which implies that preterm birth was not driving the majority of the association results. We also found a strong enrichment of CpG island shores, enhancers, and DNase I hypersensitive sites among the significant CpGs that may suggest the functional importance of identified CpGs.

Notably, many of the 8,899 CpGs overlapped with previously known EWAS of gestational age-associated CpGs[47,51,52,54,55,185]. Importantly, we also found 3,343 novel CpGs annotated to 2,577 genes that had not been linked with gestational age in previous studies.

To explore whether the detected gestational age-associated CpGs in cord blood were also differentially methylated in other fetal tissues, we used two publicly available datasets based on collected prenatally fetal lung[186] and fetal brain[187]. We observed a significant overlap between CpGs in cord blood and fetal brain and lung tissues in relation to gestational age. Thus, the epigenomic plasticity of prenatal development across tissues was partly captured by the cord blood findings. One of the genes in gestational age-associated CpGs with the most prominent negative effect estimated in cord blood, *NCOR2* was also significant in brain and lung fetal tissues. *NCOR2* has previously been linked with lung function in GWAS through vitamin A metabolism[188]. It has been suggested that Vitamin A supplementation declines the risk of bronchopulmonary dysplasia in extremely preterm-born children[189]. In addition, *NCOR2* in neurons is related to aging in methylation levels[190].

The persistence effect of identified gestational age-associated CpGs in cord blood was examined using cross-sectional whole blood from older children, including early childhood, school age, and adolescence. The cord blood findings were generally not persisted into childhood and adolescence. Only cg26385222 CpGs annotated to *TMEM176B* were related to gestational age at birth, childhood, and adolescence. *TMEM176B* gene was associated with gestational age in cord blood in a previous study[52] and has been indicated as a potential biomarker for various cancers[191]. The lack of stability at older ages could be explained by the smaller sample size, or the later exposures or development may influence the association. Similarly, the longitudinal analysis showed that many gestational age-related CpGs at birth undergo dynamic changes during early childhood and tend to stabilize in methylation levels by school age. However, some of the gestational age-related CpGs at birth (17%) were stable over time across childhood and into adolescence (Figure 7).

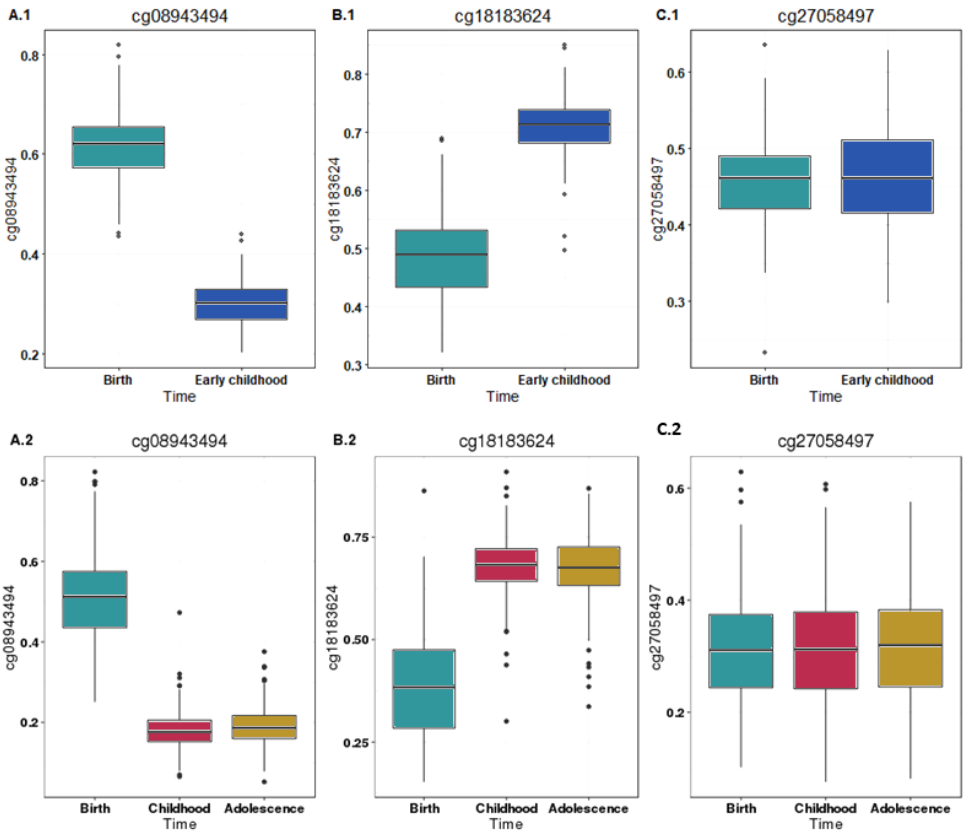

Figure 7. Three selected single representative CpGs with methylation change from birth to adolescence in association with gestational age. (1) Birth to early childhood in the INMA cohort and (2) Birth to adolescence in the ALSPAC cohort; (A) The methylation level decreases and stabilize after school age. (B) The methylation level increases and stabilize after school age. (C) The methylation level is stable from birth to school age/adolescence. The selected methylation beta values are shown on the y-axis for each age group (x-axis). Turquoise: birth; blue: early childhood; red: childhood school age and golden yellow: adolescence.

The genes annotated to gestational age-related CpGs demonstrated a functional enrichment of Gene Ontology biological processes, including embryonic development, regulation of cellular and biological processes, and immune system development. Likewise, the subset of gestational age-related genes was enriched in pathways of various diseases where low gestational age was a known risk factor, such as asthma[192], cancer[193], inflammatory bowel disease[194], and type I and type II diabetes[195] [71]. Notably, genes annotated to stable CpGs across time up to adolescence revealed enrichment of infection- and immune-related conditions.

In addition, the potential functional impact of significant CpGs was explored using paired correlation of DNA methylation and gene expression in 38 cord blood samples[151,153] within a ±250-kb window of a transcript. We found 367 CpG-transcript associations; 246 were unique CpGs significant at Bonferroni $p<0.05$. Of these, 46% were negatively correlated, and the

most significant was cg01332054 and *SEMA7A*, and the most prominent effect estimate was for cg26179948 and *JAZF1*. Similarly, among the 44% positively correlated CpG-transcript associations, the most significant was cg20139800 and *MOG*, and the most prominent effect estimate was for cg03665259 and *CDSN* (Figure 8). These strong correlations of cis-effects imply that the gestational age-related CpGs are most likely to have a direct functional impact on newborns.

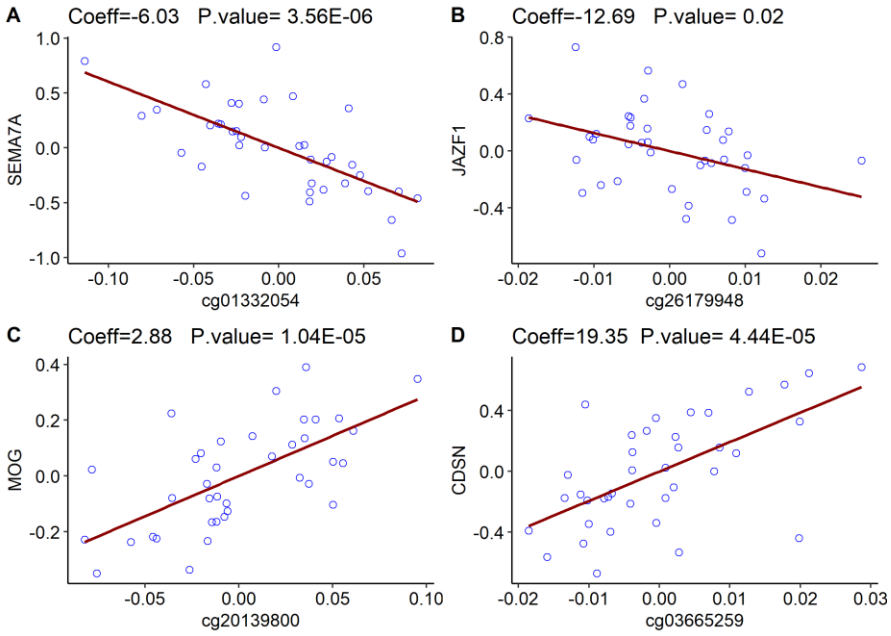

Figure 8. Correlations between selected pairs of DNA methylation and gene expression levels. (A) *SEMA7A* and *cg01332054*. (B) *JAZF* and *cg26179948*. (C) *MOG* and *cg20139800*. (D) *CDSN* and *cg03665259*. The effect size estimate and p-value are shown on top of each plot.

In conclusion, we found numerous epigenome-wide differentially methylated CpGs related to gestational age at birth, including novel CpGs that had not previously been associated with gestational age. Many identified CpGs displayed a strong functional link with human diseases and were enriched with biological processes essential for fetal development. The epigenetic plasticity of fetal development across tissues was captured by many methylation sites. However, the majority of CpGs underwent changes over time and stabilized after school age. In all, our study highlights new knowledge in relation to epigenetics, preterm birth, and gestational age.

4.2 STUDY II: MOLECULAR CHANGES ASSOCIATED WITH AIR POLLUTION EXPOSURE

In **Study II**, we explored the genome-wide gene expression associated with birth and current $PM_{2.5}$ exposure in 656 participants from the MeDALL consortium, including three European birth cohorts of pre-school children (INMA), and adolescents (BAMSE and GINIplus). In

addition, we investigated the integration of DNA methylation with matched gene expression in relation to both early-life and current $PM_{2.5}$ exposure, along with a protein-protein interaction network analysis. The characteristics of the study subjects are presented in (**Study II**; Table1). The $PM_{2.5}$ exposure levels were on average lowest for BAMSE and highest for INMA cohort (Figure 9). Moreover, we observed a moderate correlation between $PM_{2.5}$ exposure levels at birth and current residential address (0.45, 0.32, 0.46) in BAMSE, GINIplus, and INMA, respectively.

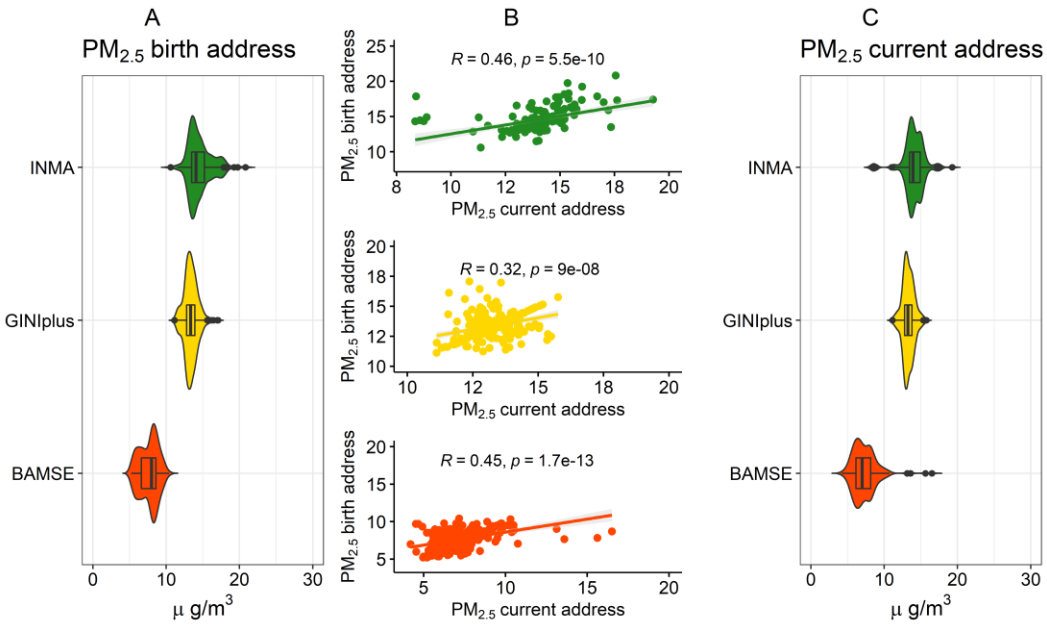

Figure 9. Birth and current address particulate matter ($PM_{2.5}$) $\mu g/m^3$ exposures in the different cohorts INMA, GINIplus and BAMSE. (A) Density plot with box plot at birth address. (B) The correlation between birth and current address $PM_{2.5}$ levels. Correlation coefficient and p-value on top of each plot (C) Density plot with box plot at current address.

We identified two genome-wide significant (FDR p<0.05) differentially expressed transcript clusters TC10001332.hg.1 (annotated to *MiR-1296* gene) and TC14001976.hg.1 (long non-coding RNA located near *FOXA1-2* (chr14:38066368-38067552)) in relation to $PM_{2.5}$ exposure levels at birth Figure 10. However, no significant association was found with current $PM_{2.5}$ exposure levels.

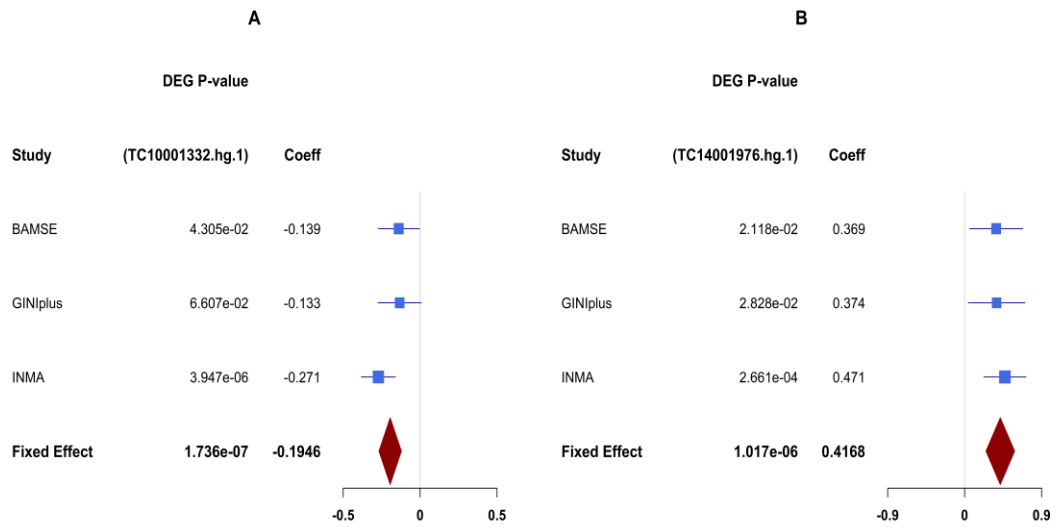

Figure 10. Forest plot of meta-analysis significant gene expression levels associated with birth $PM_{2.5}$ exposure in children and adolescents. (A) *TC10001332.hg.1* (B) *TC14001976.hg.1*

Nevertheless, out of the top 100 significant differentially expressed genes related to birth $PM_{2.5}$ exposure levels, 18 were also significant at nominal p-value ($p<0.05$) with the same direction of effect in the analysis with current $PM_{2.5}$ exposure levels.

One of the newly identified differentially expressed genes, *MIR-1296* (coding for miRNA), was found to link with different types of cancer, including breast, hepatocellular, colorectal, prostate, and lung cancer[196-200]. In addition, it has also been shown to be a potential prognostic marker of heart failure[201]. *FOXA1-2* plays a key role in lung alveolar and respiratory endoderm morphogenesis and differentiation, including α-cells in the endocrine pancreas, liver, and prostate luminal ductal epithelia[202-205].

Further, to understand the mechanisms of biological responses to $PM_{2.5}$ exposure, the functional molecular pattern was explored using the multi-omics profile of genome-wide gene expression and DNA methylation. The integration analysis of matched genome-wide gene expression and DNA methylation along with protein interaction network revealed 9 and 6 significant functional epigenetically deregulated modules associated with birth and current $PM_{2.5}$ exposure, respectively, at FDR level (FDR $p<0.05$) Table 5.

Table 5. The list of significant functional epigenetic deregulated modules of 9 and 6 hotspots from Functional Epigenetics Modules (FEM) algorithm in relation to birth and current $PM_{2.5}$ exposure.

Exposure to $PM_{2.5}$	Seed	Size (number of genes)	Modularity	FDR
Birth	*NR1I2*	11	2.12	0.001
	SH3GL2	26	2.00	0.005
	TENT5A	24	1.94	0.006
	MAPK6	64	1.95	0.010
	UBE2W	61	1.35	0.012
	KCTD15	10	1.86	0.014
	MLST8	10	2.11	0.018
	RPP40	12	1.61	0.024
	GGA1	14	1.59	0.025
Current	*TAF8*	20	2.17	<0.0001
	TAF5	20	2.17	0.001
	GNAI3	31	2.07	0.002
	ISLR	22	2.05	0.024
	TRIM69	10	1.84	0.032
	SCARA3	19	1.84	0.040

Size: number of gene found in FEM, Modularity: shows the average of the edge-weights, FDR: p-value < 0.05 is significant.

We observed simultaneous hypomethylation and overexpression of the top significant gene module centered around *NR1I2* related to $PM_{2.5}$ exposure at birth (Figure 11). *NR1I2* (Nuclear Receptor Subfamily 1 Group I Member 2) has been considered a potential target for asthma therapy[206].

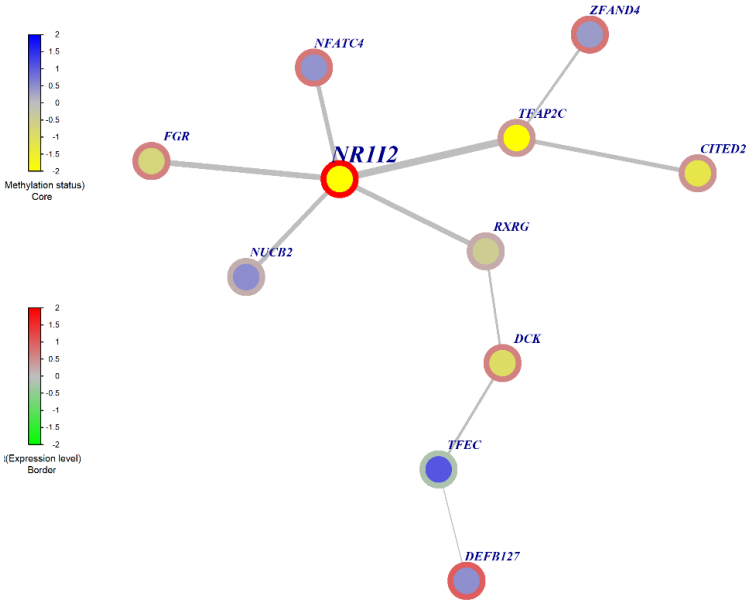

Figure 11. Functionally epigenetic deregulated top significant hotspot with the core gene *NR1I2* related to $PM_{2.5}$ exposure at birth address. DNA methylation status relates to node color with hypermethylation in blue and hypomethylation in yellow. Expression patterns are shown in circles around the nodes as upregulation in red and downregulation in green. Edge widths show the average statistics of gene-gene interaction in the network.

Another module that had the largest subnetwork size and contained 64 genes was centered around *MAPK6*. *MAPK6* gene was previously associated with air pollution *in vivo*[207] and *in vitro*[91].

The identified six hotspot modules in relation to current $PM_{2.5}$ exposure did not overlap with modules associated with exposure at birth, partly because of differences in exposure levels as indicated by a moderate correlation between birth and current exposure, as well as because exposure during different periods of life may result in different molecular responses. The two top gene hubs, *TAF8* and *TAF5* associated with current $PM_{2.5}$ exposure, contain the same 20 genes in their hubs. *TAF8* gene has in other studies been linked to prostate cancer[208]. The largest gene module among the six hotspot modules in relation to current $PM_{2.5}$ exposure was centered around the *GNAI3* gene, which contains 31 genes. *GNAI3* genes play a key role in lung adenocarcinoma[209]. Another epigenetically deregulated hotspot centered around the *SCARA3* gene (Figure 12) with simultaneous hypomethylation and overexpression has previously reported that oxidative stress induces the alteration of this gene expression[210]. Oxidative stress has been identified as one of the key mechanisms responsible for the adverse health effects of air pollution. Further, the gene was related to the progression of type 2 diabetes mellitus at methylation level[211].

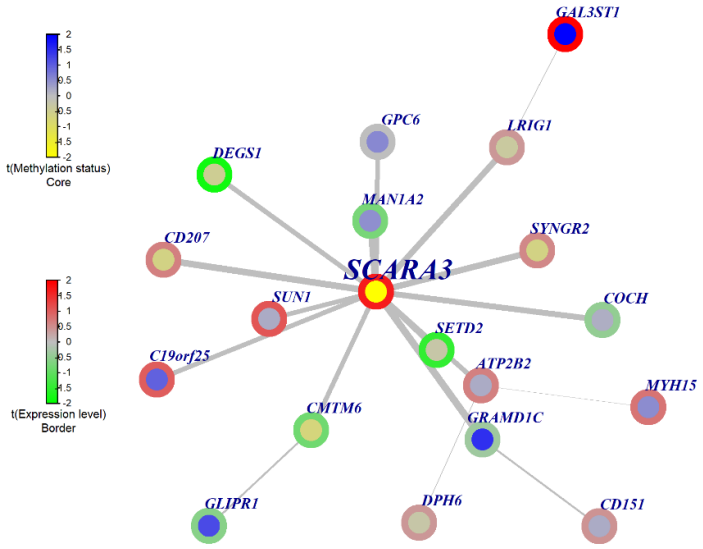

Figure 12. Functionally epigenetic deregulated top significant hotspot with the core gene *SCARA3* related to $PM_{2.5}$ exposure at birth address. DNA methylation status relates to node color with hypermethylation in blue and hypomethylation in yellow. Expression patterns shown in circles around the nodes as upregulation in red and downregulation in green. Edge widths show the average statistics of gene-gene interaction in the network.

Among other identified modules, the central genes *MLST8* was previously linked with type 2 diabetes[212] and *KCTD15* was associated with both obesity and type 2 diabetes[213,214]. In addition, *TRIM69* gene has been associated with cardiovascular diseases[215]. Air pollution, especially $PM_{2.5}$ exposure, is a previously known risk factor for obesity, type 2 diabetes as well as cardiovascular diseases[216-218].

Notably, the KEGG pathways analysis revealed that the differential gene expression associated with $PM_{2.5}$ exposure at birth address is linked to ribosome activity, olfactory transduction, complement and coagulation cascades, and systemic lupus erythematosus. On the other hand, genes associated with $PM_{2.5}$ exposure at the time of bio-sampling were linked to ribosome-related pathways only. Moreover, the identified different genes in the subnetwork modules were strongly linked to KEGG pathways, including adherens junction, cell cycle, thyroid hormones signaling and notch signaling, fatty acid degradation, and oxidative phosphorylation. They are also related to diseases where air pollution is considered a risk factor, such as Huntington, Alzheimer, Parkinson, cancer as well as rheumatoid arthritis.

Besides the suggestive evidence of differential gene expression in children and adolescents associated with $PM_{2.5}$ exposure at birth, the integration of DNA methylation with matched

gene expression identified several interactome hotspots of epigenetic deregulation gene modules in relation to $PM_{2.5}$ exposure both at birth and current address. Thus, our study emphasizes added value of additional layers of omics information in the environmental exposure data to enhance the understanding of molecular mechanisms of biological responses to harmful exposure.

4.3 STUDY III: DNA METHYLATION PATTERNS RELATED TO COPD

COPD-associated changes in DNA methylation on macrophage-dense BAL cells from 18 COPD subjects and 15 controls with normal lung function who had current and previous smoking history were analyzed in **Study III**. The COPD cases and controls had a close match with the potential confounders age, sex, and BMI. Pack-years were however higher in the COPD group. Similarly, smoking intensity among the current smokers was higher in the COPD group. Moreover, COPD subjects were the only ones who used inhaled corticosteroids, and had a lower BAL recovery (Table 6). These potential confounders were adjusted for in the model.

Table 6. Study characteristics

	COPD (n = 18)	Smokers with normal Lung function (n = 15)	P-value
Female/Male	4/14	8/7	NS
Age	63 [62-72]	65 [63-71.5]	NS
BMI	26.2 [23.7-29.0]	26.0 [24.2-27.8]	NS
Current/Ex-smokers	10/8	3/12	NS
Smoking intensity	14 [10-20]	8 [7-10]	0.049
Pack-years	36 [30-40]	15 [13.5-21.5]	< 0.0001
FEV_1% predicted	68.3 [44.1-72.3]	110.2 [94.4-117.9]	< 0.0001
FEV_1/VC	0.53 [0.43-0.62]	0.74 [0.70-0.77]	< 0.0001
Use of inhaled corticosteroids: Yes/No	6/12	0 /15	0.02
BAL recovery (%)	43.6 [32.2-55.6]	61.1 [56.1-66.7]	0.001

Median and IQR are used for continuous values. Fischer's Exact Test was used for ratios, Mann-Whitney U-test for remaining comparisons; Significant P-value < 0.05; NS: Not significant. Smoking intensity: cigarettes smoked per day. Pack-years: (cigarettes smoked per day/20) × years smoked.

This first-time EWAS of COPD on BAL cells using the latest Illumina EPIC BeadChip revealed 1,155 Bonferroni-significant CpGs annotated to 1,089 genes that spread across all chromosomes (Figure 13). We also found a strong enrichment of CpG islands among the significant CpGs. More than half of the identified CpGs had higher mean methylation in COPD cases than controls. Furthermore, the DMR analysis identified numerous unique regions related to COPD, where the top significance located in chr15:64790751-64791797 consisted of 5 CpGs and overlapped with the promotor of the *ZNF609* gene.

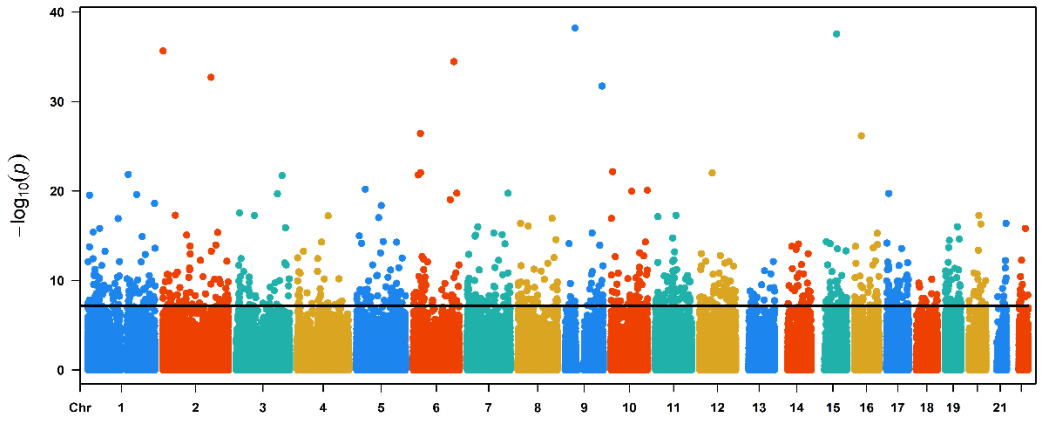

Figure 13. Manhattan plot of BAL DNA methylation in relation to COPD after adjustment for covariates and cell types. The horizontal line indicates Bonferroni-corrected significance level. Chr = chromosome in the x-axis and $\log_{10}$(p-value) in the y-axis.

The genes mapped to the significant CpGs with large DNA methylation change were previously linked with COPD or disease severity, such as *POMC*, *NLRP3*, *SCNN1A*, *ZNF322*, and *SOX30*[102,219-222]. Moreover, other novel genes not previously associated with COPD were identified. As the BAL fluid was dominated by macrophages, we investigated whether the identified COPD-associated genes were expressed in macrophages or other cell types using the Immunological genome project RNAseq dataset[158]. We found only an 8.4% overlap between our COPD-associated genes and genes expressed exclusively by cell types other than macrophages. Therefore, the identified changes observed in this study were most probably linked to macrophages. This is in line with a previous study that showed an epigenetic modification of alveolar macrophage function in COPD[223].

The enrichment of COPD-associated genes in both CpGs and DMRs shared several similar Gene Ontology biological processes, including organ or body structure development, nervous system, and metabolic processes. In addition, various molecular functions, such as transcription factor activity and growth factor binding, and different cellular compartments were highlighted. Enrichment for transcription factors was previously reported in EWAS of lung tissues from COPD cases and control subjects with normal lung function[111]. Likewise, COPD-associated enrichment for KEGG pathways revealed various cancer pathways, which is not surprising, as COPD is a known risk factor for several cancer types independent of smoking[224]. Moreover, different types of cell junctions, as well as cAMP and Rap1 signaling

pathways, were also identified. The two non-cancer pathways, cAMP and Rap1 signaling were targeted with COPD treatment drugs[225-229].

Normal aging processes in COPD lungs are found to be perturbed in many ways, such as increased oxidative stress, stem cell exhaustion, and cellular senescence[230]. This makes COPD a condition where accelerating aging may be of importance[231]. To investigate whether DNA methylation age and chronological age were different between COPD subjects compared to controls, we calculated DNA methylation aging using Horvath's epigenetic clock. In these analyses, we found however no significant difference between methylation and chronological age, which implies that either age as calculated by DNA methylation is not accelerated in COPD, or that epigenetic alteration involved in the accelerated aging of BAL cells was through mechanisms other than DNA methylation. Alternatively, the Horvath clock may not be optimal for BAL cell epigenetics.

The potential functional effects of significant CpGs were explored using paired correlation of BAL cell DNA methylation and gene expression data from Ringh *et al.*[152] within a ±250-kb window of a transcript. We found 101 CpG- transcript associations; 79 were unique CpGs at Bonferroni p<0.05. Of these, 54% were negatively correlated, and the most significant with the largest effect estimate was cg18196647 and *CPD*. Similarly, among the positively correlated CpG- transcript associations, the most significant and largest effect estimate was cg13267718 and *FLI1* (Figure 14). *FLI1* gene was previously associated with lung cancer[232,233]. Similarly, *CPD* is known to be involved in lung cancer[234]. In addition, the strong correlation between DNA methylation and gene expression implies a direct functional impact on identified COPD-related CpGs.

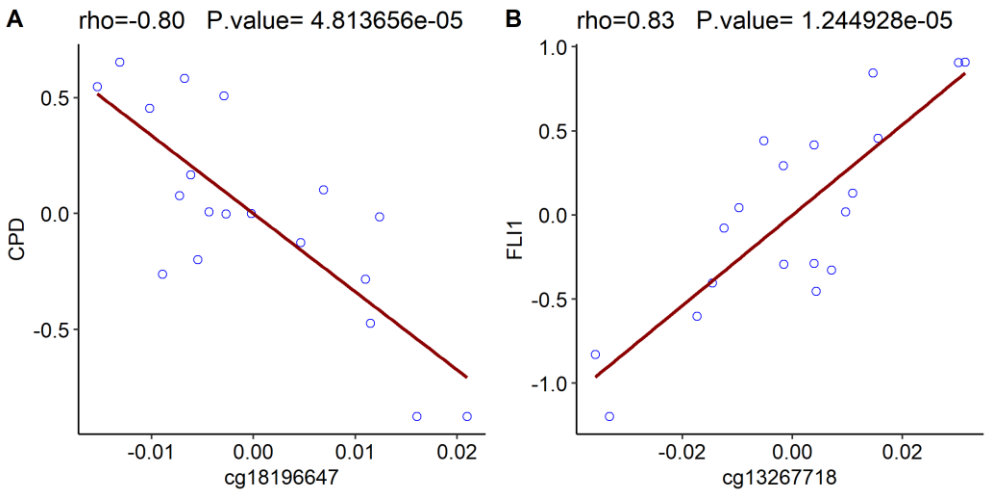

Figure 14. Correlations between selected pairs of DNA methylation and gene expression levels. (A) *CPD* and *cg18196647*. (B) *FLI1* and *cg13267718*. The correlation coefficient and p-value are shown on top of each plot.

Finally, we investigated the co-localization of CpGs in the present study with the genetics of recently identified 82 loci associated with COPD by Sakornsakolpat *et al.*[102]. We found that 38.7% of the CpGs were co-localized within a ±500 kb window of one or more COPD-associated SNPs; 10% were within a ±100 kb window of one or more SNPs. Of these, ten were annotated to the same gene, and only *ADGRG6/GPR126* SNPs were annotated to more than one CpGs. *ADGRG6/GPR126* has been associated with COPD in previous EWAS, GWA, and gene expression studies, and DL_{CO}/VA[112,235,236].

Thus, this study found several epigenome-wide differentially methylated CpGs related to COPD in BAL cells. Many of the identified CpGs displayed a strong functional link with gene expression and pathways enriched for cancer, different types of cell junctions, and cAMP and Rap1 signaling. Almost half of the CpGs co-locate in the proximity of COPD-associated SNPs, which indicates that both genetic and methylation-related mechanisms are of importance in these gene regions. Possibly, these SNPs may even act as methQTLs, thus influencing methylation levels, although we were not able to test this directly in our study.

4.4 STUDY IV: GENE EXPRESSION DURING PEANUT ORAL IMMUNOTHERAPY AND OMALIZUMAB TREATMENT

In **Study IV**, we explored gene expression changes in whole blood of 17 peanut-allergic adolescents (age 12-18 years) using RNA-sequencing profiles during pOIT and omalizumab treatment where the blood samples were taken at baseline, pOIT start, maintenance, and final timepoints. The characteristics of the study subject, including any concomitant asthma, conjunctivitis, rhinitis, or eczema, are presented in (Study IV Table S1).

At first, we investigated if there was an effect of omalizumab treatment on gene expression using the two timepoints, baseline and pOIT start. We found no significant difference, which suggests that omalizumab treatment alone does not induce alteration in peripheral blood gene expression. It should be noted that at baseline, the participants were not exposed to any peanut allergens, and the concomitant allergies or asthma were under control. This likely contributed to our observation of no expression effect of omalizumab treatment alone.

Secondly, the combined effect of pOIT with omalizumab (pOIT+O) was examined in the longitudinal analysis of three timepoints (pOIT start, maintenance, and final). We identified 680 genes (337 upregulated / 343 downregulated) associated with pOIT+O at nominal $p < 0.005$. Only 16 genes were differentially expressed at the FDR level ($FDR < 0.05$). The three largest most down-regulated genes: *ASGR2*, *GPBAR1*, and *HM13* and upregulated genes: *CDKN2AIP*, *ICOS*, and *USP44*, are presented in Figure 15. For example, *ICOS* expression plays a role in T-cell differentiation during inflammatory conditions[237] and may be involved in allergic disease mechanisms[238].

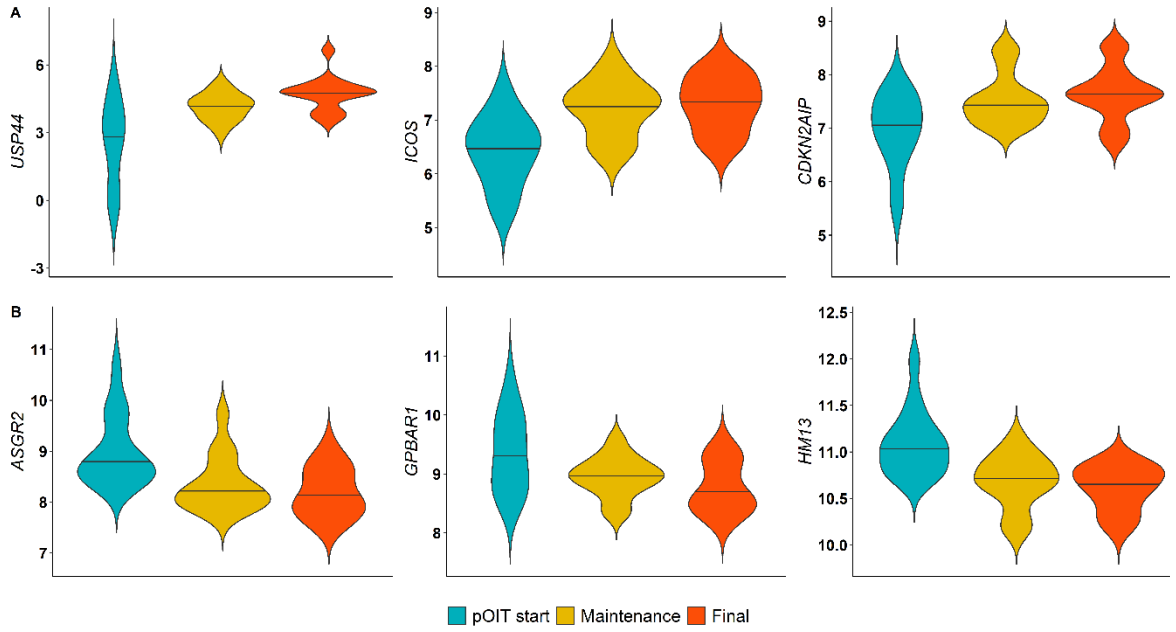

Figure 15. Violin plot of selected pOIT log_2 gene expression levels (A) Upregulated (B) Downregulated genes at FDR p-value < 0.05 for three treatment protocol steps. Turquoise violin: pOIT start, yellow violin: Maintenance, red violin: Final.

Pathway analysis revealed a strong overlap of Gene Ontology biological processes in 343 downregulated genes that converge to neutrophil degranulation, immune response, phagocytosis, and metabolic process, while upregulation of 337 genes linked to protein regulation and modification (Study IV Figure 1). Moreover, we evaluated the enrichment of our 680 pOIT+O-associated genes with genes linked to acute peanut allergic reactions in a recently published clinical study by Watson *et al.*[1] at the same p-value cut-off ($p < 0.005$) and found that 108 genes overlapped, mostly with opposite direction. (Figure 16). This comparison indicates that pOIT+O could alter genes affected during acute peanut allergic reactions.

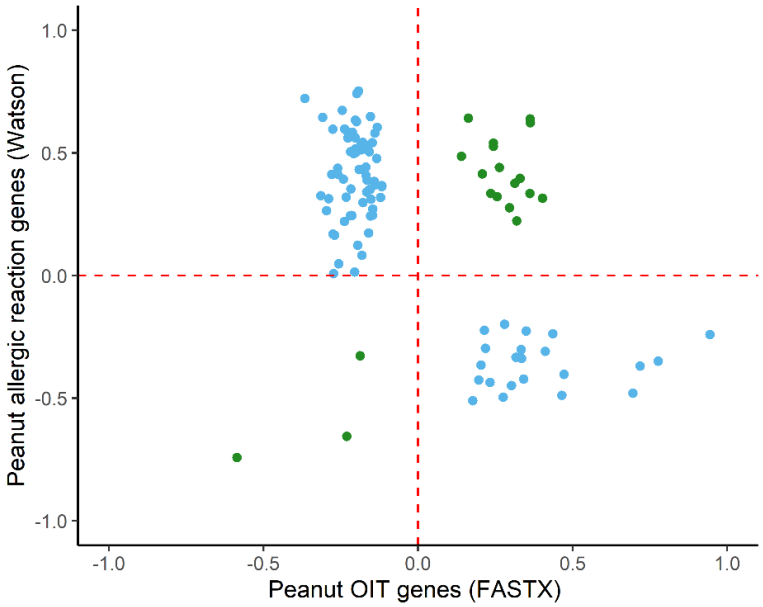

Figure 16. The overlap of FASTX pOIT genes with Watson *et al.* peanut-related genes at nominal p<0.005 (108 genes). The x-axis: Effect size estimates from the longitudinal FASTX pOIT. The y-axis: mean change of gene expression at baseline and after 4-hour peanut challenge in Watson *et al.* Blue: opposite direction. Green: same direction.

In conclusion, we demonstrated that the gene expression levels in blood of patients with peanut allergy did not alter by omalizumab treatment alone. However, the combined effect of pOIT+O showed an up-and downregulation of several genes involved in T-cell functions and immune responses. Furthermore, pOIT+O treatment seems to alter the expression of genes affected during acute peanut allergic reactions.

4.5 METHODOLOGICAL CONSIDERATION

One strength of the conducted molecular epidemiological studies (**Study I and II**) in this thesis is that the material used is based on prospective cohorts, including BAMSE and other international birth cohorts. The data were collected in a prospective manner where the exposure measurements occurred prior to the outcome assessments. Detailed questionnaires were repeatedly filled out during the follow-up in all participating cohorts. In addition, the exposure and outcome definitions were harmonized. An additional strength is the large number of participants included in the cohorts and the analyses. Moreover, the omics integration and functional follow-up analyses used data from collaborating research groups and comprehensive bioinformatics online resources. However, molecular epidemiological studies are subject to limitations in the study design or applied method, including systemic error (bias) or random error that may raise the issue of reproducibility, validity, and stability of identified markers[239]; thus, one has to keep this in mind when interpreting the results of these studies.

4.5.1 Systematic error (bias)

The validity of a molecular epidemiological study becomes high by eliminating systematic error. Systematic error mainly depends on the definition or measurement of exposure and outcome and recruitment of a study population. The major types of bias are selection bias, information bias (misclassification), and confounding.

4.5.1.1 Selection bias

Selection bias can occur during the enrollment of subjects into the study population or during the follow-up of study subjects. This can be present if the selection of the subjects in the study is based on factors related to exposure or outcome[240], which may result in a different association between exposure and outcome for those included and not included in the study. In the BAMSE prospective population-based birth cohort, 75% of eligible children from Stockholm County were included in the study. The non-responders had a higher prevalence of parental smoking but no other significant differences in background characteristics were found compared with those who participated in BAMSE[133]. Dropout or loss of follow-up is another cause of section bias. The survey response rates in BAMSE at 4, 8, and 16 years were 91%, 84%, and 78%, respectively. Those participating in the clinical examinations and collected blood samples were 63%, 60%, and 62% at ages 4, 8, and 16, respectively, including 42% of children participating in all three clinical follow-ups. In the original cohort and the follow-up "sub-cohorts", including the clinical examinations at different time points, baseline characteristics follow more or less a similar distribution with no major selection bias problems identified except female participants more likely to stay in the follow-up than male[241,242]. One of the challenges in cohort studies (and other study types) is the choice of inclusion criteria (initially and for subgroup analyses) to obtain a representative sample for the population under study. In **Study I-II**, the inclusion criteria depended on the availability of blood samples and questionnaire data from the cohorts. The selection of participants in **Study III** included those who performed clinical examinations and were willing to undergo bronchoscopy given that they fulfilled the pre-determined criteria regarding lung function, annual FEV_1 decline, and smoking history. The controls for the COPD cases were both age- and sex-matched. In **Study IV**, the recruitment of patients was performed at the outpatient allergy clinic at Sachs' Children and Youth Hospital or referred by pediatric allergists in the Stockholm area. In addition, willing participants who received information about the study through the clinicaltrials.gov website were also included in the study. Thus, only very motivated adolescents with severe allergies were included to have a homogenous study population with any concurrent allergies to either pollen or pets.

4.5.1.2 Information bias (misclassification)

Information bias, or misclassification, is characterized by any systematic error in the exposure or outcome assessment. Information bias has two categories: differential misclassification, either the classification error depends on the exposure or outcome; otherwise, non-differential misclassification if occurs independently of the exposure or outcome [240,243]. In **Study I**, a potential bias (non-differential misclassification) could occur in relation to gestational age

because the participating cohorts estimated gestational age using ultrasound or according to the last menstrual period (or combined), and if one of these two methods was more correct than the other. However, a study shows a high correlation between gestational age estimated by ultrasound and the last menstrual period[244]. Similarly, the misclassification of true air pollution exposure could have influenced our findings, as we did not incorporate individual time-activity data and indoor exposure levels. However, validation studies had demonstrated good concordance of modeled outdoor levels of air pollution with measured outdoor levels, as well as personal exposure measurements[245,246]. Importantly, the air pollution exposure assessment was harmonized in each participating cohort.

COPD was defined using the GOLD criteria (stage 2-3) and was further characterized with respect to lung function decline and smoking history. The spirometry was assessed according to the ATS/ERS guidelines using a dry volume spirometer by well-trained and experienced staff. Any misclassification in COPD cases or controls is likely non-differential as the measurement should only be affected by a random error in the lung function measurement.

Peanut oral immunotherapy combined with omalizumab was given to peanut-allergic adolescents by well-trained and experienced physicians and nurses on each visit, while laboratory staff was blinded. During the treatment, misclassification is unlikely given the detailed clinical protocol, but cannot be 100% excluded.

In this study, we used two microarray-based DNA methylation Infinium 450K and EPIC chips. The EPIC chip includes >90 % of the CpGs from the 450K and an additional 413,743 CpGs, and these additional probes improve the coverage of regulatory regions[247]. But the EPIC array includes only 3% of the known 28.2 million CpG sites in the human genome. Several reasons could cause failure in microarray experiments, including low-quality DNA input, incomplete bisulfite conversion, or a failure in other Infinium assay steps[248]. However, this variation is likely to happen randomly, which may lead to a non-differential misclassification for exposure of interest. In addition, technical variations related to the bisulfite treatment and microarray chips are corrected using batch correction methods or adjusted in the model. In **Study I-II**, the participating cohorts followed their normalization and quality control (QC) protocols in DNA methylation, which may induce heterogeneity in the result. However, previous EWAS meta-analysis shows similar results between non-normalized methylation and normalized methylation using different protocols across the cohorts[154].

The gene expression data used in this study was based on microarray and RNA-seq; both have high reliability and reproducibility. The microarray was based on a hybridization-based technique, while the RNA seq refers to a sequencing-based technique[249]. In both cases, a technical variation occurs during RNA extraction, library preparation, and other production steps while running in different chips. This process might result in random technical variation with respect to exposure of interest that may be leading a non-differential misclassification unless the case and control are on different plates or different batches.

In **Study I** and **Study III**, the functional follow-up of CpGs on gene expression was done ±250-kb window. The co-localized within a ±500 kb window of one or more COPD-associated SNPs on CpGs were performed in **Study III**. However, there was no consensus on the optimal distance of cis-effects. The functional analyses in **Study I** was performed using a set of genes robustly affecting gestational age overlap with regions of Bonferroni significant at least three or more adjacent CpGs; these lead to not including potential important single CpGs in the functional analyses.

The results of this study might be influenced to some degree by the misclassification of exposure. However, the potential bias was likely non-differential as DNA methylation or gene expression levels and the exposures were assessed separately.

4.5.1.3 Confounders

A confounder is the third factor besides exposure and outcome (e.g., intervention, treatment, exposure) that influences both outcome and exposure[250]. Confounding can result in either under- or overestimation of the association[250]. In these studies, potential confounders or variables affecting the effect estimates along with known confounders were adjusted for in the models. The list of potential confounders included in this thesis is shown above in Table 4.

It is known that gene expression and DNA methylation are tissue- and cell-specific, and studying the right relevant tissues associated with exposure or disease is very important. In **Study I**, we assessed DNA methylation from the cord and whole blood associated with gestational age; in addition, fetal and brain tissues were compared with blood findings. However, in **Study II**, rather than nasal epithelium, where the first contact between air pollution and the respiratory tract occurs, blood samples were assessed for convenience reasons. The main target organ BAL cell DNA methylation profiles were evaluated in COPD cases and controls in **Study III**. However, the DNA methylation pattern may differ between BAL cells or other relevant lung tissues or blood. By the systemic nature of peanut allergy, blood RNA sequence was evaluated for pOIT in **Study IV**.

Cell types are important potential confounding factors in DNA methylation and gene expression, and adjusting for measured or estimated cell types is, therefore essential. In **Studies I-II**, a relative proportion of white blood cells from DNA methylation was calculated using a Bakulski reference for cord blood[169] and Reinius reference for whole blood[170,171]. In **Study IV**, the cell composition in RNA-seq was estimated via deconvolution through the CIBERSORT method[172]. Measured cell type counts were adjusted for BAL cell DNA methylation in **Study III**.

In **Study I**, we adjusted the model for known potential clinical/environmental confounders, including maternal smoking during pregnancy and birth weight. Still, some of the identified gestational age-related DNA methylation overlapped with those presented in the maternal smoking EWAS[154], which may reflect the under-reporting of smoking status by some pregnant women as smoking-related DNA methylation might capture quantitative smoking

history more accurately than self-report[251]. In addition, the adverse effect of shorter gestational age is linked with some biological pathways related to smoking CpGs[252,253]. Similarly, we observed an overlap of CpGs in birth weight[182] and gestational age in our study. The impact of DNA methylation may have been shared by the two correlated factors in newborns, which makes it difficult to uncouple the effect. Also, two potential confounders, maternal obesity and alcohol intake that may alter offspring DNA methylation, were not included in the adjustment. However, the impact of maternal obesity and alcohol intake is modest compared to maternal smoking[254,255].

Socioeconomic status is related to air pollution exposure, especially in urban areas where low socio-economy is often linked to high traffic-related air pollution exposure[256]. However, the opposite is observed in Stockholm, where a higher socio-economic status is more common in the inner-city areas that also have the highest traffic-related air pollution levels[70]. Therefore, in **Study II**, the results were adjusted for the study area (as well as for socio-economic status). One of the potential confounders for COPD is smoking; in **Study III**, the smoking history was retrieved from the longitudinal population-based cohort, and both the COPD cases and controls had a documented smoking history. Air pollution is known to increase the risk of having COPD and is associated with DNA methylation, but we did not have data on this confounding factor in **Study III**. The treatment outcome, either success or failure, was adjusted for in **Study IV.** As all the patients were required to have a concomitant respiratory allergy and many also had asthma, they might have used inhaled steroids while on pOIT. Such treatment was not controlled for in the analyses, although inhaled steroids are believed to leave few biological fingerprints when assessed systemically[257].

Thus, it is possible that residual or unmeasured confounders might have affected the associations presented in this thesis. In addition, cross-sectional analyses were performed in **Study I-IV**; therefore, the association cannot infer causation.

4.5.2 Random error

The presence of random error in a selected sample can occur due to the population's variability, which affects the precision of an estimate. A way to minimize the random error and maximize precision is to have a larger sample size that still represents the source population. The selected subsample in **Studies I-II** from the original cohorts, depending on the availability of blood samples, questionnaire data, and measured exposure, did limit the sample size compared to the original cohorts but still reached very reasonable numbers. Meta-analysis of data within an international consortium with harmonized exposure and outcome variables increased the sample size of **Study I** and **Study II**; in total, 11,000 and 650 participants in **Study I** and **Study II** were included in the analysis, respectively. Even though we had a large sample size in **Study I,** rather few premature births were observed. The sample size in **Study III** (18 COPD cases and 15 controls) is limited because the sample collection depends on bronchoscopies to get the BAL fluid, and recruiting makes it difficult as the medical procedure is time-consuming and invasive to some extent. Similarly, in **Study IV**, the sample size was small due to challenging enrollment in pOIT from the peanut-allergic

participant, who feared allergic reactions, and disliked the taste of peanuts. Some participants also dropped out; at the end, a total of 17 participants' data were collected at four-time points. The small sample size in **Studies III-IV** may have introduced some statistical uncertainty, that we tried to overcome by rigorous methodologic, statistical approaches and the use of publicly available datasets for additional analyses.

4.5.3 Generalizability of results

The birth cohort BAMSE and other international studies included in **Study I-II** are population-based. The majority of cohorts contributing to **Study I** comprise participants of European ancestry, and few cohorts of Hispanic background. Only European ancestry was included in **Study II**.

The association between DNA methylation and gestational age in **Study I** should be generalizable to most populations of European and Hispanic origin.

The objective of **Study II** was to assess the impact of air pollution on transcriptomic and epigenetic profiles. The molecular mechanisms induced by air pollution could be generalizable to most populations of European origin but, since we only included data from three centers, applications to other areas with higher or lower exposure levels are uncertain.

Our results in **Study III** could be considered generalizable to an older Swedish population aged 62 years and above since participants from the longitudinal population-based OLIN COPD cohort were recruited. But the sample size was rather small and originated only from one Norrbotten-based study that prevents complete generalizability across regions or time periods etc. **Study IV** and the FASTX study could be regarded as fairly generalizable to an adolescent Swedish population with severe allergy although the patients were recruited in the Stockholm area. However, the **Study IV** analyses were aimed to elucidate mechanisms and detect biomarkers in relation to OIT, which is also feasible in a selected group of patients or participants as in this case.

5 CONCLUSIONS AND FUTURE PERSPECTIVES

The following conclusion can be drawn from the separate studies included in this thesis:

I: Using an epigenome-wide approach, we showed that gestational age was associated with methylome signatures at birth, which likely captures the epigenetic plasticity of fetal development across tissues. Methylation levels in the majority of identified CpGs changed over time and stabilized after school age. In addition, we presented functional links between identified CpGs and human diseases and enrichment of biological processes essential for fetal development, along with correlated gene activities.

II: $PM_{2.5}$ exposure at birth altered transcriptome profiles in children and adolescents. Integration of gene expression with matched DNA methylation data and protein-protein interactions revealed several epigenetic deregulation gene modules interactome hotspots in relation to $PM_{2.5}$ exposure both at birth and at the time of bio-sampling. We highlighted the added value of additional layers of omics information on $PM_{2.5}$ exposure to enhance the understanding of molecular mechanisms and biological responses induced by air pollution exposure.

III: COPD was associated with methylome signatures in BAL cells, primarily consisting of alveolar macrophages. These epigenetic effects likely reflected the local impact of the disease on the lung and airways. Additionally, we presented the functional pathways that are affected and altered in COPD and also have a link to gene expression. In addition, our results suggested that both genetic and epigenetic mechanisms play important roles at certain COPD-associated loci.

IV: Omalizumab treatment in patients with peanut allergy alone did not change peripheral gene expression levels. However, pOIT and combined treatment with omalizumab did alter transcriptome profiles. We demonstrated up- and downregulation of several genes involved in T-cell functions and immune responses. Furthermore, pOIT and combined treatment with omalizumab seemed to alter genes affected during acute peanut allergic reactions.

In an ideal setting, we could have addressed our study limitations and expanded the studies in this thesis in the following way: Collecting additional samples of premature individuals in **Study I**, especially among subjects borne extremely preterm. This would have allowed us to find methylation patterns on the whole spectrum of gestational ages. Including additional samples from heavily polluted or less polluted cities in **Study II** would allow us to compare the molecular pattern at different exposure levels. By increasing the sample size in **Study III**, we would have had better power to identify stable and replicable COPD markers. Adding a control group without omalizumab treatment and increasing the sample size in **Study IV** would have helped us identify transcriptomic changes associated with pOIT without concomitant treatment. Extensive use of publicly available bioinformatic databases and collaborating internationally in the consortium could have created a possibility of replicating

the research findings and investigating the functional role of early-life exposures, chronic diseases, and treatment response.

Going forward, studying relevant tissues or cell types other than surrogate tissues could produce relevant data to uncover the underlying molecular mechanisms induced by early environmental exposures and disease pathogenesis, and treatment response. In the future, in addition to total BAL cell investigations, a sorted BAL cell population from the lung in a genome-wide setting would give additional valuable information at the cellular level. Moreover, several studies have suggested that nasal epithelial cells could serve as a proxy to study the lower airways[258]. It would be very valuable to study how air pollution affects airway epithelium, as the nasal epithelium is a primary target for inhaled harmful substances. Epithelial cell analyses may help us understand the molecular changes that directly affect the inner airway system. Furthermore, adding additional layers of information using the single-cell technique in clinical settings would likely strengthen our understanding of, for example, treatment response.

Future research should focus on the integration of different omics layers of information to enhance understanding of molecular mechanisms and biological responses that arise due to early environmental exposures or in complex diseases like COPD. Moreover, having additional layers of omics information would be highly relevant for capturing the molecular treatment response from pOIT (and other treatments). However, handling the large-scale and complex data produced from the different layers of omics requires method development and user-friendly bioinformatics pipeline tools. But some of these are already underway[18].

The results in this thesis show that early-life exposure to air pollution affects DNA methylation as well as gene expression. In addition, DNA methylation in BAL cells is strongly associated with COPD. Furthermore, pOIT and combined treatment with omalizumab alter gene expression. However, the functional mechanisms induced by environmental exposure, the functional role of disease pathogenesis, and the functional effect of the treatment response still need additional clarification. It would be valuable to expand our analyses to identify the functional aspects of the identified key genes.

In summary, these findings may contribute valuable knowledge regarding the influence of environmental exposures and stressors *in utero* and early life on developmental processes, health, and disease. Our epigenetics results regarding COPD provided additional insights into the pathogenesis of disease, and analyses of gene signatures in relation to immunotherapy treatment response may open for novel approaches to study the biological effects of specific treatments. Overall, the findings in this thesis can contribute to translational efforts bridging epidemiology, experimental research, and clinical care.

6 POPULÄRVETENSKAPLIG SAMMANFATTING AV AVHANDLINGEN

Både innan och efter att en människa föds pågår en snabb utveckling av kroppen inkluderande celldifferentiering, bildning av organ och utveckling av organsystem. Exponering för miljöfaktorer eller andra stressorer under fosterlivet och de första levnadsåren skulle kunna påverka dessa utvecklingsprocesser, potentiellt genom molekylära och epigenetiska mekanismer, och därigenom ha en negativ inverkan på individens hälsa senare i livet. *Epigenetik* innebär förändringar i en gens aktivitet eller funktion, utan att själva DNA-sekvensen har ändrats, vilket kan resultera att genen ”slås på” eller ”stängs av”. Det kommer i sin tur påverka *genuttryck och kodning av protein*, det vill säga hur och när ett protein bildas utifrån instruktionen som genen innehåller. Epigenetiska förändringar kan vara antingen reversibla eller enkelriktade, och även ärftliga. Den mest utförligt studerade epigenetiska mekanismen är *DNA-metylering*, där cytosin, en av de fyra byggstenarna som DNA är uppbyggt av, ombildas till 5-methylcytosin. DNA-metylering har föreslagits vara en länk mellan genetik och miljöfaktorer. Epigenetiska mönster som uppkommit tidigt i livet (under graviditeten) kan påverka genuttrycket under hela livet och skulle därmed kunna påverka en individs benägenhet att utveckla kroniska sjukdomar.

Kroniskt obstruktiv lungsjukdom (KOL) är en komplex sjukdom och ses som ett globalt hälsoproblem. En huvudsaklig riskfaktor för KOL är tobaksrökning. Andra faktorer som genetik, luftföroreningar och upprepade luftvägsinfektioner har också visats vara kopplade till risken att utveckla KOL. Däremot har rollen som DNA-metylering spelar i utvecklingen av KOL inte studerats i detalj.

En annan komplex sjukdom, jordnötsallergi, är en av de vanligast förekommande födoämnesallergierna och den ledande orsaken till allvarlig allergisk reaktion (anafylaxi) hos barn. Oral immunoterapi vid jordnötsallergi (pOIT) innebär att en liten mängd jordnöt intas under kontrollerade former och kan leda till en minskad känslighet och utveckling av tolerans för födoämnet. Behandling med läkemedlet omalizumab, en anti-IgE-antikropp som minskar den allergiska reaktionen mot till exempel jordnöt, skulle kunna göra starten av oral immunoterapi vid jordnötsallergi säkrare. Mekanismerna bakom hur tolerans uppkommer vid oral immunoterapi är dock inte klarlagda.

Denna avhandling syftar till att identifiera de molekylära mönster som kan relateras till tidiga exponeringar, kronisk lungsjukdom, liksom svaret på allergibehandling.

I den första studien relaterades DNA-metyleringsmönstret vid födelsen till hur lång graviditeten hade varat (antal gestationsveckor). Denna storskaliga undersökning baserades på data från 26 olika studier som deltog i det stora PACE-konsortiet och kunde visa att det på flera platser i arvsmassan fanns ett samband mellan nivån av DNA-metylering och graviditetslängd. Genom att analysera blod från navelsträngen kunde vi visa på att dessa förändringar troligen kunde kopplas till fosterutveckling i flera olika vävnader.

I de flesta fall var nivån av DNA-metyleringar föränderlig under barndomen, d.v.s. om det fanns skillnader kopplade till gestationsålder vid födseln så ändrade sig nivåerna med åldern till jämförbara nivåer vid skolåldern, och därefter ligga kvar på samma nivå. Dock identifierades en mindre mängd metyleringsplatser (17%) där sambandet mellan lägre metyleringsnivå och lägre gestationsålder kvarstod upp i tonåren. Bioinformatikanalyser visade att de gener som kunde kopplas till de identifierade metylationsplatserna har i andra studier visats ha samband med olika sjukdomar och sannolikt är involverade i biologiska processer som är nödvändiga för fostrets utveckling. Flera av metylationsplatserna påverkade också uttrycket av närliggande gener.

I nästa studie undersöktes sambandet mellan den beräknade utomhusexponeringen för partiklar med storlek mindre än 2.5 mikrometer ($PM_{2.5}$) vid bostadsadressen vid födseln samt under tonåren, och effekten på DNA-metylering. Genyttryck undersöktes under barndomen och tonåren. Vi fann att genuttryck hos barn och tonåringar var kopplat till den exponering för $PM_{2.5}$ som fanns vid födelseadressen. När vi kombinerade metylering och genuttryck och relaterade detta till exponering för $PM_{2.5}$ fann vi flera exempel på interactome hotspots, det vill säga tillfällen då både metylering och genuttryck påverkas. Några av de identifierade generna kunde kopplas till sjukdomar som orsakas eller förvärras av luftföroreningsexponering.

I den tredje studien studerades sambandet mellan metylering i celler från de nedre luftvägarna, i huvudsak makrofager, KOL-status och rökning hos vuxna individer, med avsikten att öka förståelsen om sjukdomsuppkomsten vid KOL. Vi fann flera samband mellan KOL och DNA-metyleringsnivåer i dessa celler, med en stark funktionell koppling till nivåerna för genuttryck. Vår analys pekar också mot att både genetiska och epigenetiska mekanismer spelar viktiga roller för KOL-sjukdomen.

I det fjärde delarbetet studerades genuttrycket före, under och efter pOIT hos tonåringar med allvarlig jordnötsallergi. Här återfanns både upp- och nedreglering av immunrelaterade gener i relation till pOIT och omalizumab-behandling. Dessa resultat kan hjälpa oss att förstå mekanismerna bakom toleransutveckling vid svår allergi.

Sammanfattningsvis har de presenterade resultaten ökat vår kunskap avseende den roll DNA-metylering och genuttryck spelar i människans utveckling, gällande effekter av föroreningsexponering, för sjukdomsmekanismer och vid svar på behandling. Dessa fynd kan bidra till att överbrygga epidemiologisk och experimentell forskning med klinisk vård.

7 ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation to the following people who contributed immensely in the process of making this thesis.

First and foremost, I would like to thank all the participants in the BAMSE, FASTX, and KOLIN studies and the cohorts in the PACE and MeDALL consortia; without your participation, this thesis would not have been possible.

Erik Melén, Special thanks to my main supervisor Erik; thank you for the professional and academic guidance and comments you have provided me throughout this project. Your continuous support, guidance, and encouragement helped me to strive for excellence. It is my pleasure and an honor to be in your team. This journey would not have been possible if it was not for your excellent leadership style.

Olena Gruzieva, my co-supervisor, I would like to extend my sincere thanks for your inspiration, willingness to share your knowledge, and always keeping your doors open whenever I have questions. And thank you for introducing me to the field of environmental epidemiology. Your ability to always create a great working atmosphere and fika time are very much appreciated.

Åsa Wheelock, my co-supervisor, for your valuable and encouraging advice. Thank you for sharing your incredible expertise in bioinformatics.

Gerard Koppelman, my co-supervisor, for your extensive knowledge and invaluable feedback. Thank you for sharing your enormous knowledge and expertise in the field of respiratory genetics.

Petter Ljungman, my mentor. Thank you for giving me your time and sound advice.

Per Tornvall, present head of the Department of Clinical Science and Education, Södersjukhuset, for providing the possibilities to perform research in an excellent environment.

I would like to thank and acknowledge all the co-authors for sharing your invaluable knowledge. It has been a great honor to work with many research groups from different disciplines.

The BAMSE co-authors: **Göran Pershagen**, thank you for the warm welcome in the fantastic IMM group; you were a true inspiration and gave incredible leadership and asked critical questions in the research meetings; **Anna Bergström**, a supportive and kind person with outstanding leadership; **Inger Kull**, a positive, smiling, and warm personality with a fantastic BAMSE-support; **Ashish Kumar** shared your knowledge and expertise in biostatistics and bioinformatics, and **Sophia Björkander** introducing me in the field of immunology and kind, supportive that has an eye for detail. Thank you all for your unending guidance, help, and support.

All the collaborators and co-authors in PACE consortium. Special thanks to **Stephanie J London** and **Janine F Felix** for organizing, creating a fantastic atmosphere, sharing your vast knowledge, and consistent, timely follow-up of the project.

The MeDALL collaborator and co-authors, thank you for providing me with valuable input and sharing your knowledge.

All the co-authors in KOLIN project, your contribution was very crucial. Furthermore, special thanks to **Jonas Eriksson Ström** for sharing your knowledge in COPD, providing valuable clinical inputs in the project, and asking practical analytical questions. I have learned so much from our discussion and email correspondence; it is a pleasure working with you. Most importantly, the PI of the KOLIN study **Annelie F Behndig**, thanks for your contribution and for sharing your vast knowledge.

Special thanks to the FASTX teams' physicians and nurses that made the project possible. Many thanks to FASTX PI **Caroline Nilsson** and **Anna Nopp** for your valuable knowledge and contribution. Moreover, all the co-authors in FASTX project, for your invaluable contribution.

A supportive and smiling BAMSE co-PI **Antonios Georgelis,** thank you. The BAMSE secretariat **Alexandra Lövquist**, **Sandra Ekström**, and **André Lauber**, for your tremendous and appreciated work. Thanks to former BAMSE crew **Eva Hallner** and **Sara Nilsson**.

To my former and current colleagues at the Institute of Environmental Medicine: **Ulla Stenius** (present chairmen of the Institute of Environmental Medicine that creates an excellent research environment), **Niklas Andersson** (thank you for having a good discussion about statistical issues), **Tom Bellander**, **Magnus Wickman**, **Andrei Pyko**, **Alva Wallas**, **Anna Gref**, **Erika Schultz**, **Jessica Magnusson**, **Marcus Dahlquist**, **Elin Dahlén**, **Jennifer Protudjer**, **Michal Korek**, **Charlotta Eriksson**, **Auriba Raza**, **Natalia Ballardini**, **Emma Johansson**, **Lara Stucki**, **Tomas Lind**, **Ann-Sophie Merritt, Mare Löhmus Sundström, Getahun Bero Bedada**, **Marina Jonsson**, **Åsa Persson**, **Zhebin Yu**, **Shizhen He**, **Jeroen De Bon**, **Massimo Stafoggia**, **Emmanouela Sdona**, **Anna Zettergren;** it has been a privilege working with you all. The *fantasy football* crew **Jesse Thacher** and **Ayman Alhamdow,** your discussion at the lunch table, and coming up with ideas to convince me to join the "*Bitcoin*" community were hilarious.

All colleges at the Department of Clinical Science and Education, Södersjukhuset whom I had shared the journey. Thank you; **Petra Um Bergström** for being very kind, always smiling and bringing Ester candy baskets to our office, **Jenny Hallberg** for being friendly with a positive personality; tusen tack för den svenska sammanfattande översättningen, **Susanna Klevebro** for your encouragement and valuable research discussion, **Natalia Hernandez-Pacheco** for your invaluable Ph.D. studies guidance and inspiration, your input was very helpful, **Maura Kere** for your inspiration through and valuable commitments, **Gang Wang** for having a great statistical talk and always willing to help, **Björn Lundberg**

for the excellent research discussion and fantastic time during ERS Barcelona, **Maria Ödling** for giving me your advice and share your journey of writing kappa, **Thomas Olsson** for the excellent lunchtime discussions, and **Louise Betton** for being cheerful and ready to answer all the question. **Hans-Jacob Koefoed** for bringing a new vibe to the office. All members of Melén group, **Susanne Lundin** and **Ida Mogensen**; it has been a privilege working with you all. **Dimitra Karampatsi** and **Ellen Vercalsteren,** thanks for having a great talk at the lunch table.

Thank you, all colleges at Forskningscenter, Södersjukhuset: **Carina Wallén**, **Ulrika Hellberg**, **Ann-Charlotte Sundqvist**, **Margareta Eriksson**, **Anna Castel,** and **Zekiye Cansu**. Special thanks to **Fuad Bahram** for making great illustrations (FB scientific art design).

To all my friends, specially **Frezer Eshetu** and **Anteneh Moges,** thank you for your thoughtful support and encouragement in this life journey.

To my family, my mother **Sisay,** you are the best. I am truly blessed. አንቺ ለእኔ አርአያ እና መነሳሻ ሆነሽ ሁል ጊዜም ከጎኔ ሳትለይ ስለምትሰጭኝ ድጋፍ እወድሻለው::. My father **Kebede**; Dad, I wish you were here to see this achievement; የአንተ ማበረታቻ፣ ምክር እና የህይወት ተመኩሮ ሁል ጊዜ በአእምሮዬ ውስጥ ናቸው።. My big brothers **Andualem** and **Kidus**, thanks for your guidance in life. Especially, Kidus, thank you for the fantastic cover design. ቅዱስ ስለ አስደናቂው የሽፋን ስዕል አመሰግናለሁ::. The eldest sister, **Amarech,** you are super. My lovely sister **Yodit**, thank you for your support. My beloves two little sisters, **Eden** and **Rediet**; love you all.

Last but not least **Hiwot,** my beautiful wife and the love of my life, you are simply the best and perfect wife and mother to our daughter **Deborah**. You two always inspire me to work hard, and I am lucky to have you in my life. I love you so much!!

8 REFERENCES

1 Watson, C. T. *et al.* Integrative transcriptomic analysis reveals key drivers of acute peanut allergic reactions. *Nature communications* **8**, 1943, doi:10.1038/s41467-017-02188-7 (2017).

2 Gannon, F. Molecular biology--what's in a name? *EMBO reports* **3**, 101, doi:10.1093/embo-reports/kvf039 (2002).

3 Alberts B, J. A., Lewis J, Morgan D, Raff M, Roberts K, Walter P Molecular Biology of the Cell. 1-10 (2014).

4 Pukkila, P. J. in *eLS*.

5 Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **13**, 484-492, doi:10.1038/nrg3230 (2012).

6 Holliday, R. Epigenetics: an overview. *Developmental genetics* **15**, 453-457, doi:10.1002/dvg.1020150602 (1994).

7 Gibney, E. R. & Nolan, C. M. Epigenetics and gene expression. *Heredity* **105**, 4-13, doi:10.1038/hdy.2010.54 (2010).

8 Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571**, 489-499, doi:10.1038/s41586-019-1411-0 (2019).

9 Feinberg, A. P. The Key Role of Epigenetics in Human Disease Prevention and Mitigation. **378**, 1323-1334, doi:10.1056/NEJMra1402513 (2018).

10 Sharp, G. C. & Relton, C. L. Epigenetics and noncommunicable diseases. **9**, 789-791, doi:10.2217/epi-2017-0045 (2017).

11 Boyce, W. T., Sokolowski, M. B. & Robinson, G. E. Genes and environments, development and time. **117**, 23235-23241, doi:doi:10.1073/pnas.2016710117 (2020).

12 Jawaid, A., Jehle, K.-L. & Mansuy, I. M. Impact of Parental Exposure on Offspring Health in Humans. *Trends in Genetics* **37**, 373-388, doi:https://doi.org/10.1016/j.tig.2020.10.006 (2021).

13 Ryan, C. P. "Epigenetic clocks": Theory and applications in human biology. **33**, e23488, doi:https://doi.org/10.1002/ajhb.23488 (2021).

14 Campagna, M. P. *et al.* Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clinical epigenetics* **13**, 214, doi:10.1186/s13148-021-01200-8 (2021).

15 Li, D. in *Computational Biology* (ed H. Husi) (Codon Publications

Copyright: The Authors., 2019).

16 Koh, E. J. & Hwang, S. Y. Multi-omics approaches for understanding environmental exposure and human health. *Molecular & Cellular Toxicology* **15**, 1-7, doi:10.1007/s13273-019-0001-4 (2019).

17 Kaur, P., Singh, A. & Chana, I. Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions. *Archives of Computational Methods in Engineering* **28**, 4595-4631, doi:10.1007/s11831-021-09547-0 (2021).

18 Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and biology insights* **14**, 1177932219899051, doi:10.1177/1177932219899051 (2020).

19 Lu Shi, J. *et al.* A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data. *Current Proteomics* **12**, 14-27, doi:http://dx.doi.org/10.2174/1570164612011505062009 27 (2015).

20 Bader, G. D., Cary, M. P. & Sander, C. Pathguide: a pathway resource list. *Nucleic acids research* **34**, D504-506, doi:10.1093/nar/gkj126 (2006).

21 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).

22 Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic acids research* **47**, D590-d595, doi:10.1093/nar/gky962 (2019).

23 Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39**, D691-697, doi:10.1093/nar/gkq1018 (2011).

24 Pascual-Montano, A. & Carazo, J. M. Efficient functional bioinformatics tools: towards understanding biological processes. *EMBnet.journal; Vol 16, No 1*, doi:10.14806/ej.16.1.184 (2010).

25 Lynch, C. D. & Zhang, J. The research implications of the selection of a gestational age estimation method. *Paediatric and perinatal epidemiology* **21 Suppl 2**, 86-96, doi:10.1111/j.1365-3016.2007.00865.x (2007).

26 Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. *Lancet (London, England)* **371**, 75-84, doi:10.1016/s0140-6736(08)60074-4 (2008).

27 Engle, W. A. Morbidity and mortality in late preterm and early term newborns: a continuum. *Clinics in perinatology* **38**, 493-516, doi:10.1016/j.clp.2011.06.009 (2011).

28 Been, J. V. *et al.* Preterm birth and childhood wheezing disorders: a systematic review and meta-analysis. *PLoS medicine* **11**, e1001596, doi:10.1371/journal.pmed.1001596 (2014).

29 den Dekker, H. T. *et al.* Early growth characteristics and the risk of reduced lung function and asthma: A meta-analysis of 25,000 children. *The Journal of allergy and clinical immunology* **137**, 1026-1035, doi:10.1016/j.jaci.2015.08.050 (2016).

30 Leung, J. Y., Lam, H. S., Leung, G. M. & Schooling, C. M. Gestational Age, Birthweight for Gestational Age, and Childhood Hospitalisations for Asthma and Other Wheezing Disorders. *Paediatric and perinatal epidemiology* **30**, 149-159, doi:10.1111/ppe.12273 (2016).

31 Raby, B. A. *et al.* Low-normal gestational age as a predictor of asthma at 6 years of age. *Pediatrics* **114**, e327-332, doi:10.1542/peds.2003-0838-L (2004).

32 Aarnoudse-Moens, C. S., Weisglas-Kuperus, N., van Goudoever, J. B. & Oosterlaan, J. Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. *Pediatrics* **124**, 717-728, doi:10.1542/peds.2008-2816 (2009).

33 Kerkhof, G. F., Breukhoven, P. E., Leunissen, R. W., Willemsen, R. H. & Hokken-Koelega, A. C. Does preterm birth influence cardiovascular risk in early adulthood? *The Journal of pediatrics* **161**, 390-396.e391, doi:10.1016/j.jpeds.2012.03.048 (2012).

34 Geldof, C. J., van Wassenaer, A. G., de Kieviet, J. F., Kok, J. H. & Oosterlaan, J. Visual perception and visual-motor integration in very preterm and/or very low birth weight children: a meta-analysis. *Research in developmental disabilities* **33**, 726-736, doi:10.1016/j.ridd.2011.08.025 (2012).

35 Hille, E. T. *et al.* Functional outcomes and participation in young adulthood for very preterm and very low birth weight infants: the Dutch Project on Preterm and Small for Gestational Age Infants at 19 years of age. *Pediatrics* **120**, e587-595, doi:10.1542/peds.2006-2407 (2007).

36 Kwinta, P. & Pietrzyk, J. J. Preterm birth and respiratory disease in later life. *Expert review of respiratory medicine* **4**, 593-604, doi:10.1586/ers.10.59 (2010).

37 Thunqvist, P. *et al.* Lung Function at 8 and 16 Years After Moderate-to-Late Preterm Birth: A Prospective Cohort Study. *Pediatrics* **137**, doi:10.1542/peds.2015-2056 (2016).

38 Institute of Medicine Committee on Understanding Premature, B. & Assuring Healthy, O. in *Preterm Birth: Causes, Consequences, and Prevention* (eds R. E. Behrman & A. S. Butler) (National Academies Press (US)

National Academy of Sciences., 2007).

39 Koning, S. M. & Ehrenthal, D. B. Stressor landscapes, birth weight, and prematurity at the intersection of race and income: Elucidating birth contexts through patterned life events. *SSM - population health* **8**, 100460, doi:10.1016/j.ssmph.2019.100460 (2019).

40 Kondracki, A. J. & Hofferth, S. L. A gestational vulnerability window for smoking exposure and the increased risk of preterm birth: how timing and intensity of maternal smoking matter. *Reproductive Health* **16**, 43, doi:10.1186/s12978-019-0705-x (2019).

41 Fuchs, F., Monet, B., Ducruet, T., Chaillet, N. & Audibert, F. Effect of maternal age on the risk of preterm birth: A large cohort study. *PloS one* **13**, e0191002, doi:10.1371/journal.pone.0191002 (2018).

42 Zhang, G. *et al.* Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. **377**, 1156-1167, doi:10.1056/NEJMoa1612665 (2017).

43 Kemp, M. W. Preterm birth, intrauterine infection, and fetal inflammation. *Frontiers in immunology* **5**, 574, doi:10.3389/fimmu.2014.00574 (2014).

44 Dolgun, Z. N., Inan, C., Altintas, A. S., Okten, S. B. & Sayin, N. C. Preterm birth in twin pregnancies: Clinical outcomes and predictive parameters. *Pakistan journal of medical sciences* **32**, 922-926, doi:10.12669/pjms.324.10409 (2016).

45 Gruzieva, O., Merid, S. K. & Melen, E. An update on epigenetics and childhood respiratory diseases. *Paediatric respiratory reviews* **15**, 348-354, doi:10.1016/j.prrv.2014.07.003 (2014).

46 Robins, J. C., Marsit, C. J., Padbury, J. F. & Sharma, S. S. Endocrine disruptors, environmental oxygen, epigenetics and pregnancy. *Frontiers in bioscience (Elite edition)* **3**, 690-700 (2011).

47 Cruickshank, M. N. *et al.* Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. *Genome medicine* **5**, 96, doi:10.1186/gm500 (2013).

48 Cutfield, W. S., Hofman, P. L., Mitchell, M. & Morison, I. M. Could epigenetics play a role in the developmental origins of health and disease? *Pediatric research* **61**, 68r-75r, doi:10.1203/pdr.0b013e318045764c (2007).

49 Parets, S. E., Bedient, C. E., Menon, R. & Smith, A. K. Preterm birth and its long-term effects: methylation to mechanisms. *Biology* **3**, 498-513, doi:10.3390/biology3030498 (2014).

50 Lee, H. *et al.* DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth. *International journal of epidemiology* **41**, 188-199, doi:10.1093/ije/dyr237 (2012).

51 Schroeder, J. W. *et al.* Neonatal DNA methylation patterns associate with gestational age. *Epigenetics* **6**, 1498-1504, doi:10.4161/epi.6.12.18296 (2011).

52 Parets, S. E. *et al.* Fetal DNA Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age. *PloS one* **8**, e67489, doi:10.1371/journal.pone.0067489 (2013).

53 Knight, A. K. *et al.* An epigenetic clock for gestational age at birth based on blood methylation data. *Genome biology* **17**, 206, doi:10.1186/s13059-016-1068-z (2016).

54 Simpkin, A. J. *et al.* Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Human molecular genetics* **24**, 3752-3763, doi:10.1093/hmg/ddv119 (2015).

55 Bohlin, J. *et al.* Prediction of gestational age based on genome-wide differentially methylated regions. *Genome biology* **17**, 207, doi:10.1186/s13059-016-1063-4 (2016).

56 Relton, C. L. *et al.* DNA methylation patterns in cord blood DNA and body size in childhood. *PloS one* **7**, e31821, doi:10.1371/journal.pone.0031821 (2012).

57 Perez, L., Rapp, R. & Kunzli, N. The Year of the Lung: outdoor air pollution and lung health. *Swiss medical weekly* **140**, w13129, doi:10.4414/smw.2010.13129 (2010).

58 Brook, R. D. *et al.* Hemodynamic, autonomic, and vascular effects of exposure to coarse particulate matter air pollution from a rural location. *Environmental health perspectives* **122**, 624-630, doi:10.1289/ehp.1306595 (2014).

59 World Health Organization. Fact Sheet: Ambient (Outdoor) Air Quality and Health. 2018. Available online: https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health ((01 May 2019)).

60 Pedersen, M. *et al.* Ambient air pollution and low birthweight: a European cohort study (ESCAPE). *The Lancet. Respiratory medicine* **1**, 695-704, doi:10.1016/s2213-2600(13)70192-9 (2013).

61 Minelli, C. *et al.* Interactive effects of antioxidant genes and air pollution on respiratory function and airway disease: a HuGE review. *American journal of epidemiology* **173**, 603-620, doi:10.1093/aje/kwq403 (2011).

62 Duan, R. R., Hao, K. & Yang, T. Air pollution and chronic obstructive pulmonary disease. *Chronic diseases and translational medicine* **6**, 260-269, doi:10.1016/j.cdtm.2020.05.004 (2020).

63 Calderon-Garciduenas, L., Torres-Jardon, R., Kulesza, R. J., Park, S. B. & D'Angiulli, A. Air pollution and detrimental effects on children's brain. The need for a multidisciplinary approach to the issue complexity and challenges. *Frontiers in human neuroscience* **8**, 613, doi:10.3389/fnhum.2014.00613 (2014).

64 Rivas, I. *et al.* Association between Early Life Exposure to Air Pollution and Working Memory and Attention. *Environmental health perspectives* **127**, 57002, doi:10.1289/ehp3169 (2019).

65 Chen, G. *et al.* Early life exposure to particulate matter air pollution (PM1, PM2.5 and PM10) and autism in Shanghai, China: A case-control study. *Environment international* **121**, 1121-1127, doi:10.1016/j.envint.2018.10.026 (2018).

66 Wang, G. *et al.* Early-life risk factors for reversible and irreversible airflow limitation in young adults: findings from the BAMSE birth cohort. *Thorax* **76**, 503-507, doi:10.1136/thoraxjnl-2020-215884 (2021).

67 Wang, G. *et al.* Assessment of chronic bronchitis and risk factors in young adults: results from BAMSE. *The European respiratory journal* **57**, doi:10.1183/13993003.02120-2020 (2021).

68 Gruzieva, O. *et al.* Exposure to air pollution from traffic and childhood asthma until 12 years of age. *Epidemiology (Cambridge, Mass.)* **24**, 54-61, doi:10.1097/EDE.0b013e318276c1ea (2013).

69 Schultz, E. S. *et al.* Traffic-related air pollution and lung function in children at 8 years of age: a birth cohort study. *American journal of respiratory and critical care medicine* **186**, 1286-1291, doi:10.1164/rccm.201206-1045OC (2012).

70 Schultz, E. S. *et al.* Early-Life Exposure to Traffic-related Air Pollution and Lung Function in Adolescence. *American journal of respiratory and critical care medicine* **193**, 171-177, doi:10.1164/rccm.201505-0928OC (2016).

71 Schultz, E. S. *et al.* Early life exposure to traffic-related air pollution and lung function in adolescence assessed with impulse oscillometry. *The Journal of allergy and clinical immunology* **138**, 930-932.e935, doi:10.1016/j.jaci.2016.04.014 (2016).

72 Lodovici, M. & Bigagli, E. Oxidative stress and air pollution exposure. *J Toxicol* **2011**, 487074-487074, doi:10.1155/2011/487074 (2011).

73 Ghio, A. J., Carraway, M. S. & Madden, M. C. Composition of air pollution particles and oxidative stress in cells, tissues, and living systems. *Journal of toxicology and environmental health. Part B, Critical reviews* **15**, 1-21, doi:10.1080/10937404.2012.632359 (2012).

74 Neophytou, A. M. *et al.* Traffic-related exposures and biomarkers of systemic inflammation, endothelial activation and oxidative stress: a panel study in the US trucking industry. *Environmental Health* **12**, 105, doi:10.1186/1476-069X-12-105 (2013).

75 Gruzieva, O. *et al.* Prenatal Particulate Air Pollution and DNA Methylation in Newborns: An Epigenome-Wide Meta-Analysis. *Environmental health perspectives* **127**, 57012, doi:10.1289/ehp4522 (2019).

76 Gruzieva, O. *et al.* Epigenome-Wide Meta-Analysis of Methylation in Children Related to Prenatal NO2 Air Pollution Exposure. *Environmental health perspectives* **125**, 104-110, doi:10.1289/ehp36 (2017).

77 Chi, G. C. *et al.* Long-term outdoor air pollution and DNA methylation in circulating monocytes: results from the Multi-Ethnic Study of Atherosclerosis (MESA). *Environmental Health* **15**, 119, doi:10.1186/s12940-016-0202-4 (2016).

78 Plusquin, M. *et al.* DNA methylation and exposure to ambient air pollution in two prospective cohorts. *Environment international* **108**, 127-136, doi:https://doi.org/10.1016/j.envint.2017.08.006 (2017).

79 de, F. C. L. A. J. *et al.* Long-term Air Pollution Exposure, Genome-wide DNA Methylation and Lung Function in the LifeLines Cohort Study. *Environmental health perspectives* **126**, 027004, doi:10.1289/ehp2045 (2018).

80 Lee, M. K. *et al.* Genome-wide DNA methylation and long-term ambient air pollution exposure in Korean adults. *Clinical epigenetics* **11**, 37, doi:10.1186/s13148-019-0635-z (2019).

81 Mostafavi, N. *et al.* Associations Between Genome-wide Gene Expression and Ambient Nitrogen Oxides. *Epidemiology (Cambridge, Mass.)* **28**, 320-328, doi:10.1097/ede.0000000000000628 (2017).

82 Chu, J. H. *et al.* Gene expression network analyses in response to air pollution exposures in the trucking industry. *Environmental health : a global access science source* **15**, 101, doi:10.1186/s12940-016-0187-z (2016).

83 van Leeuwen, D. M. *et al.* Genome-wide differential gene expression in children exposed to air pollution in the Czech Republic. *Mutation research* **600**, 12-22, doi:10.1016/j.mrfmmm.2006.05.032 (2006).

84 Honkova, K. *et al.* Gene expression profiling in healthy newborns from diverse localities of the Czech Republic. *Environmental and molecular mutagenesis* **59**, 401-415, doi:10.1002/em.22184 (2018).

85 Vlaanderen, J. *et al.* Impact of long-term exposure to PM2.5 on peripheral blood gene expression pathways involved in cell signaling and immune response. *Environment international* **168**, 107491, doi:https://doi.org/10.1016/j.envint.2022.107491 (2022).

86 Tian, Z. *et al.* Constructing an integrated gene similarity network for the identification of disease genes. *J Biomed Semantics* **8**, 32, doi:10.1186/s13326-017-0141-1 (2017).

87 Jiao, Y., Widschwendter, M. & Teschendorff, A. E. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics (Oxford, England)* **30**, 2360-2366, doi:10.1093/bioinformatics/btu316 (2014).

88 Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11**, 333-337, doi:10.1038/nmeth.2810 (2014).

89 Li, C. X., Wheelock, C. E., Skold, C. M. & Wheelock, A. M. Integration of multi-omics datasets enables molecular classification of COPD. *The European respiratory journal* **51**, doi:10.1183/13993003.01930-2017 (2018).

90 Yang, X. *et al.* Integrative analysis of methylome and transcriptome variation of identified cardiac disease-specific genes in human cardiomyocytes after PM2.5 exposure. *Chemosphere* **212**, 915-926, doi:10.1016/j.chemosphere.2018.09.010 (2018).

91 Vargas, J. E. *et al.* A systemic approach to identify signaling pathways activated during short-term exposure to traffic-related urban air pollution from human blood.

Environmental science and pollution research international **25**, 29572-29583, doi:10.1007/s11356-018-3009-8 (2018).

92 WHO. The top 10 causes of death 2017 [updated Jan 2017; cited 2017 Nov 17]. Available from: www.who.int/mediacentre/factsheets/fs310/en/.

93 GOLD. 2022 GOLD Reports, Global Strategy for Prevention, Diagnosis and Management of COPD 2022 [updated 2022; cited 2022 Mar 4]. https://goldcopd.org/2022-gold-reports-2/.

94 Diaz-Guzman, E. & Mannino, D. M. Epidemiology and prevalence of chronic obstructive pulmonary disease. *Clinics in chest medicine* **35**, 7-16, doi:10.1016/j.ccm.2013.10.002 (2014).

95 Lindberg, A. *et al.* Prevalence of chronic obstructive pulmonary disease according to BTS, ERS, GOLD and ATS criteria in relation to doctor's diagnosis, symptoms, age, gender, and smoking habits. *Respiration; international review of thoracic diseases* **72**, 471-479, doi:10.1159/000087670 (2005).

96 Lundbäck, B. *et al.* Not 15 but 50% of smokers develop COPD?--Report from the Obstructive Lung Disease in Northern Sweden Studies. *Respiratory medicine* **97**, 115-122, doi:10.1053/rmed.2003.1446 (2003).

97 Agustí, A. & Faner, R. COPD beyond smoking: new paradigm, novel opportunities. *The Lancet. Respiratory medicine* **6**, 324-326, doi:10.1016/s2213-2600(18)30060-2 (2018).

98 Melén, E., Guerra, S., Hallberg, J., Jarvis, D. & Stanojevic, S. Linking COPD epidemiology with pediatric asthma care: Implications for the patient and the physician. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* **30**, 589-597, doi:10.1111/pai.13054 (2019).

99 Doyle, L. W. *et al.* Expiratory airflow in late adolescence and early adulthood in individuals born very preterm or with very low birthweight compared with controls born at term or with normal birthweight: a meta-analysis of individual participant data. *The Lancet. Respiratory medicine* **7**, 677-686, doi:10.1016/s2213-2600(18)30530-7 (2019).

100 Simpson, S. J. *et al.* Lung function trajectories throughout childhood in survivors of very preterm birth: a longitudinal cohort study. *The Lancet. Child & adolescent health* **2**, 350-359, doi:10.1016/s2352-4642(18)30064-6 (2018).

101 Bui, D. S. *et al.* Association between very to moderate preterm births, lung function deficits, and COPD at age 53 years: analysis of a prospective cohort study. *The Lancet. Respiratory medicine* **10**, 478-484, doi:10.1016/s2213-2600(21)00508-7 (2022).

102 Sakornsakolpat, P. *et al.* Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nature genetics* **51**, 494-505, doi:10.1038/s41588-018-0342-2 (2019).

103 Ingebrigtsen, T. *et al.* Genetic influences on Chronic Obstructive Pulmonary Disease - a twin study. *Respiratory medicine* **104**, 1890-1895, doi:10.1016/j.rmed.2010.05.004 (2010).

104 Meteran, H., Backer, V., Kyvik, K. O., Skytthe, A. & Thomsen, S. F. Comorbidity between chronic obstructive pulmonary disease and type 2 diabetes: A nation-wide

cohort twin study. *Respiratory medicine* **109**, 1026-1030, doi:10.1016/j.rmed.2015.05.015 (2015).

105 Agustí, A., Melén, E., DeMeo, D. L., Breyer-Kohansal, R. & Faner, R. Pathogenesis of chronic obstructive pulmonary disease: understanding the contributions of gene–environment interactions across the lifespan. *The Lancet Respiratory Medicine* **10**, 512-524, doi:https://doi.org/10.1016/S2213-2600(21)00555-5 (2022).

106 Barnes, P. J. & Celli, B. R. Systemic manifestations and comorbidities of COPD. *The European respiratory journal* **33**, 1165-1185, doi:10.1183/09031936.00128008 (2009).

107 Rider, C. F. & Carlsten, C. Air pollution and DNA methylation: effects of exposure in humans. *Clinical epigenetics* **11**, 131, doi:10.1186/s13148-019-0713-2 (2019).

108 Zong, D., Liu, X., Li, J., Ouyang, R. & Chen, P. The role of cigarette smoke-induced epigenetic alterations in inflammation. *Epigenetics & chromatin* **12**, 65, doi:10.1186/s13072-019-0311-8 (2019).

109 Kachroo, P. *et al.* Co-methylation analysis in lung tissue identifies pathways for fetal origins of COPD. *The European respiratory journal* **56**, doi:10.1183/13993003.02347-2019 (2020).

110 Machin, M. *et al.* Systematic review of lung function and COPD with peripheral blood DNA methylation in population based studies. *BMC pulmonary medicine* **17**, 54, doi:10.1186/s12890-017-0397-3 (2017).

111 Morrow, J. D. *et al.* DNA methylation profiling in human lung tissue identifies genes associated with COPD. *Epigenetics* **11**, 730-739, doi:10.1080/15592294.2016.1226451 (2016).

112 Vucic, E. A. *et al.* DNA methylation is globally disrupted and associated with expression changes in chronic obstructive pulmonary disease small airways. *American journal of respiratory cell and molecular biology* **50**, 912-922, doi:10.1165/rcmb.2013-0304OC (2014).

113 Faner, R. *et al.* Do sputum or circulating blood samples reflect the pulmonary transcriptomic differences of COPD patients? A multi-tissue transcriptomic network META-analysis. *Respir Res* **20**, 5, doi:10.1186/s12931-018-0965-y (2019).

114 Lieberman, J. A. *et al.* The global burden of illness of peanut allergy: A comprehensive literature review. *Allergy* **76**, 1367-1384, doi:10.1111/all.14666 (2021).

115 Sicherer, S. H. & Sampson, H. A. Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *The Journal of allergy and clinical immunology* **141**, 41-58, doi:10.1016/j.jaci.2017.11.003 (2018).

116 Peters, R. L. *et al.* Natural history of peanut allergy and predictors of resolution in the first 4 years of life: A population-based assessment. *The Journal of allergy and clinical immunology* **135**, 1257-1266.e1251-1252, doi:10.1016/j.jaci.2015.01.002 (2015).

117 Sicherer, S. H., Muñoz-Furlong, A., Godbold, J. H. & Sampson, H. A. US prevalence of self-reported peanut, tree nut, and sesame allergy: 11-year follow-up. *The Journal of allergy and clinical immunology* **125**, 1322-1326, doi:10.1016/j.jaci.2010.03.029 (2010).

118 Kotz, D., Simpson, C. R. & Sheikh, A. Incidence, prevalence, and trends of general practitioner-recorded diagnosis of peanut allergy in England, 2001 to 2005. *The Journal of allergy and clinical immunology* **127**, 623-630.e621, doi:10.1016/j.jaci.2010.11.021 (2011).

119 Asarnoj, A. *et al.* Reported symptoms to peanut between 4 and 8 years among children sensitized to peanut and birch pollen - results from the BAMSE birth cohort. *Allergy* **65**, 213-219, doi:10.1111/j.1398-9995.2009.02138.x (2010).

120 Vetander, M. *et al.* Anaphylaxis and reactions to foods in children--a population-based case study of emergency department visits. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **42**, 568-577, doi:10.1111/j.1365-2222.2011.03954.x (2012).

121 Krogulska, A. & Wood, R. A. Peanut allergy diagnosis: Moving from basic to more elegant testing. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* **31**, 346-357, doi:10.1111/pai.13215 (2020).

122 Lanser, B. J., Wright, B. L., Orgel, K. A., Vickery, B. P. & Fleischer, D. M. Current Options for the Treatment of Food Allergy. *Pediatric clinics of North America* **62**, 1531-1549, doi:10.1016/j.pcl.2015.07.015 (2015).

123 Mäntylä, J. *et al.* The effect of oral immunotherapy treatment in severe IgE mediated milk, peanut, and egg allergy in adults. *Immunity, inflammation and disease* **6**, 307-311, doi:10.1002/iid3.218 (2018).

124 Kauppila, T. K. *et al.* Outcome of oral immunotherapy for persistent cow's milk allergy from 11 years of experience in Finland. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* **30**, 356-362, doi:10.1111/pai.13025 (2019).

125 Chu, D. K. *et al.* Oral immunotherapy for peanut allergy (PACE): a systematic review and meta-analysis of efficacy and safety. *Lancet (London, England)* **393**, 2222-2232, doi:10.1016/s0140-6736(19)30420-9 (2019).

126 Fiocchi, A. *et al.* Oral immunotherapy for peanut allergy: The con argument. *The World Allergy Organization journal* **13**, 100445, doi:10.1016/j.waojou.2020.100445 (2020).

127 Ramsey, N. & Berin, M. C. Pathogenesis of IgE-mediated food allergy and implications for future immunotherapeutics. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* **32**, 1416-1425, doi:10.1111/pai.13501 (2021).

128 Barshow, S. M., Kulis, M. D., Burks, A. W. & Kim, E. H. Mechanisms of oral immunotherapy. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **51**, 527-535, doi:10.1111/cea.13824 (2021).

129 Yu, W., Freeland, D. M. H. & Nadeau, K. C. Food allergy: immune mechanisms, diagnosis and immunotherapy. *Nature Reviews Immunology* **16**, 751-765, doi:10.1038/nri.2016.111 (2016).

130 Dantzer, J. A. & Wood, R. A. Omalizumab as an adjuvant in food allergen immunotherapy. *Current opinion in allergy and clinical immunology* **21**, 278-285, doi:10.1097/aci.0000000000000736 (2021).

131 Wang, W. *et al.* Transcriptional changes in peanut-specific CD4+ T cells over the course of oral immunotherapy. *Clinical immunology (Orlando, Fla.)* **219**, 108568, doi:10.1016/j.clim.2020.108568 (2020).

132 Anvari, S. *et al.* Memory and naïve gamma delta regulatory T-cell gene expression in the first 24-weeks of peanut oral immunotherapy. *Clinical immunology (Orlando, Fla.)* **230**, 108820, doi:10.1016/j.clim.2021.108820 (2021).

133 Wickman, M., Kull, I., Pershagen, G. & Nordvall, S. L. The BAMSE project: presentation of a prospective longitudinal birth cohort study. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* **13 Suppl 15**, 11-13 (2002).

134 Thacher, J. D. *et al.* Pre- and postnatal exposure to parental smoking and allergic disease through adolescence. *Pediatrics* **134**, 428-434, doi:10.1542/peds.2014-0427 (2014).

135 Yu, Z. *et al.* Association of Short-term Air Pollution Exposure With SARS-CoV-2 Infection Among Young Adults in Sweden. *JAMA network open* **5**, e228109, doi:10.1001/jamanetworkopen.2022.8109 (2022).

136 Felix, J. F. *et al.* Cohort Profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. *International journal of epidemiology* **47**, 22-23u, doi:10.1093/ije/dyx190 (2018).

137 Bousquet, J. *et al.* MeDALL (Mechanisms of the Development of ALLergy): an integrated approach from phenotypes to systems medicine. **66**, 596-604, doi:10.1111/j.1398-9995.2010.02534.x (2011).

138 Heinrich, J. *et al.* GINIplus and LISAplus - Design and selected results of two German birth cohorts about natural course of atopic diseases and their determinants. *Allergologie select* **1**, 85-95, doi:10.5414/alx01455e (2017).

139 Guxens, M. *et al.* Cohort Profile: The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project. *International journal of epidemiology* **41**, 930-940, doi:10.1093/ije/dyr054 %J International Journal of Epidemiology (2011).

140 Lindberg, A. *et al.* From COPD epidemiology to studies of pathophysiological disease mechanisms: challenges with regard to study design and recruitment process: Respiratory and Cardiovascular Effects in COPD (KOLIN). *European clinical respiratory journal* **4**, 1415095, doi:10.1080/20018525.2017.1415095 (2017).

141 Lindberg, A. & Lundbäck, B. The Obstructive Lung Disease in Northern Sweden Chronic Obstructive Pulmonary Disease Study: design, the first year participation and mortality. *The clinical respiratory journal* **2 Suppl 1**, 64-71, doi:10.1111/j.1752-699X.2008.00086.x (2008).

142 Glaumann, S. *et al.* Basophil allergen threshold sensitivity, CD-sens, IgE-sensitization and DBPCFC in peanut-sensitized children. *Allergy* **67**, 242-247, doi:10.1111/j.1398-9995.2011.02754.x (2012).

143 Santos, A. F. *et al.* Basophil activation test discriminates between allergy and tolerance in peanut-sensitized children. *The Journal of allergy and clinical immunology* **134**, 645-652, doi:10.1016/j.jaci.2014.04.039 (2014).

144 Vetander, M. *et al.* Anaphylaxis to foods in a population of adolescents: incidence, characteristics and associated risks. *Clinical and experimental allergy : journal of the*

British Society for Allergy and Clinical Immunology **46**, 1575-1587, doi:10.1111/cea.12842 (2016).

145 Brandström, J. *et al.* Individually dosed omalizumab: an effective treatment for severe peanut allergy. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **47**, 540-550, doi:10.1111/cea.12862 (2017).

146 Brandström, J. *et al.* Individually dosed omalizumab facilitates peanut oral immunotherapy in peanut allergic adolescents. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **49**, 1328-1341, doi:10.1111/cea.13469 (2019).

147 Eeftens, M. *et al.* Development of Land Use Regression models for PM(2.5), PM(2.5) absorbance, PM(10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. *Environmental science & technology* **46**, 11195-11205, doi:10.1021/es301948k (2012).

148 Vogelmeier, C. F. *et al.* Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. *The European respiratory journal* **49**, doi:10.1183/13993003.00214-2017 (2017).

149 Standardization of Spirometry, 1994 Update. American Thoracic Society. *American journal of respiratory and critical care medicine* **152**, 1107-1136, doi:10.1164/ajrccm.152.3.7663792 (1995).

150 Berglund, E. *et al.* Forced expirograms in subjects between 7 and 70 years of age. **173**, 185-192 (1963).

151 Rojas, D. *et al.* Prenatal arsenic exposure and the epigenome: identifying sites of 5-methylcytosine alterations that predict functional changes in gene expression in newborn cord blood and subsequent birth outcomes. *Toxicological sciences : an official journal of the Society of Toxicology* **143**, 97-106, doi:10.1093/toxsci/kfu210 (2015).

152 Ringh, M. V. *et al.* Tobacco smoking induces changes in true DNA methylation, hydroxymethylation and gene expression in bronchoalveolar lavage cells. *EBioMedicine* **46**, 290-304, doi:10.1016/j.ebiom.2019.07.006 (2019).

153 Rager, J. E. *et al.* Prenatal arsenic exposure and the epigenome: altered microRNAs associated with innate and adaptive immune signaling in newborn cord blood. *Environmental and molecular mutagenesis* **55**, 196-208, doi:10.1002/em.21842 (2014).

154 Joubert, B. R. *et al.* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *American journal of human genetics* **98**, 680-696, doi:10.1016/j.ajhg.2016.02.019 (2016).

155 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

156 Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology* **6**, 92, doi:10.1186/1752-0509-6-92 (2012).

157 The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research* **49**, D325-d334, doi:10.1093/nar/gkaa1113 (2021).

158 Heng, T. S. & Painter, M. W. The Immunological Genome Project: networks of gene expression in immune cells. *Nature immunology* **9**, 1091-1094, doi:10.1038/ni1008-1091 (2008).

159 R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

160 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

161 Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic acids research* **44**, e20, doi:10.1093/nar/gkv907 (2016).

162 Venables, W. R., B. . Modern Applied Statistics with S. *Springer* (2002).

163 Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics (Oxford, England)* **28**, 2986-2988, doi:10.1093/bioinformatics/bts545 (2012).

164 Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics & chromatin* **8**, 6, doi:10.1186/1756-8935-8-6 (2015).

165 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47-e47, doi:10.1093/nar/gkv007 %J Nucleic Acids Research (2015).

166 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

167 Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**, 86, doi:10.1186/1471-2105-13-86 (2012).

168 Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics (Oxford, England)* **30**, 1363-1369, doi:10.1093/bioinformatics/btu049 (2014).

169 Bakulski, K. M. *et al.* DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics* **11**, 354-362, doi:10.1080/15592294.2016.1161875 (2016).

170 Reinius, L. E. *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one* **7**, e41361, doi:10.1371/journal.pone.0041361 (2012).

171 Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology* **15**, R31, doi:10.1186/gb-2014-15-2-r31 (2014).

172 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).

173 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).

174 Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American journal of human genetics* **88**, 586-598, doi:10.1016/j.ajhg.2011.04.014 (2011).

175 Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine* **21**, 1539-1558, doi:10.1002/sim.1186 (2002).

176 Horvath, S. DNA methylation age of human tissues and cell types. *Genome biology* **14**, 3156, doi:10.1186/gb-2013-14-10-r115 (2013).

177 Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic acids research* **39**, D712-D717, doi:10.1093/nar/gkq1156 (2011).

178 Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic acids research* **37**, D623-D628, doi:10.1093/nar/gkn698 (2009).

179 Ogris, C., Guala, D., Helleday, T. & Sonnhammer, Erik L L. A novel method for crosstalk analysis of biological networks: improving accuracy of pathway annotation. *Nucleic acids research* **45**, e8-e8, doi:10.1093/nar/gkw849 (2017).

180 Ogris, C., Helleday, T. & Sonnhammer, E. L. PathwAX: a web server for network crosstalk based pathway annotation. *Nucleic acids research* **44**, W105-109, doi:10.1093/nar/gkw356 (2016).

181 Yu, G. enrichplot: Visualization of Functional Enrichment Result. R package version 1.10.2. *https://yulab-smu.top/biomedical-knowledge-mining-book/* (2012).

182 Kupers, L. K. *et al.* Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nature communications* **10**, 1893, doi:10.1038/s41467-019-09671-3 (2019).

183 Šidák, Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association* **62**, 626-633, doi:10.1080/01621459.1967.10482935 (1967).

184 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **57**, 289-300, doi:10.1111/j.2517-6161.1995.tb02031.x (1995).

185 Hannon, E. *et al.* Variable DNA methylation in neonates mediates the association between prenatal smoking and birth weight. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **374**, 20180120, doi:10.1098/rstb.2018.0120 (2019).

186 Chhabra, D. *et al.* Fetal lung and placental methylation is associated with in utero nicotine exposure. *Epigenetics* **9**, 1473-1484, doi:10.4161/15592294.2014.971593 (2014).

187 Spiers, H. *et al.* Methylomic trajectories across human fetal brain development. *Genome research* **25**, 338-352, doi:10.1101/gr.180273.114 (2015).

188 Minelli, C. *et al.* Association of Forced Vital Capacity with the Developmental Gene NCOR2. *PloS one* **11**, e0147388, doi:10.1371/journal.pone.0147388 (2016).

189 Garg, B. D., Bansal, A. & Kabra, N. S. Role of vitamin A supplementation in prevention of bronchopulmonary dysplasia in extremely low birth weight neonates: a systematic review of randomized trials. *The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the*

Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstet, 1-8, doi:10.1080/14767058.2018.1441282 (2018).

190 Gasparoni, G. *et al.* DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. *Epigenetics & chromatin* **11**, 41, doi:10.1186/s13072-018-0211-3 (2018).

191 Cuajungco, M. P. *et al.* Abnormal accumulation of human transmembrane (TMEM)-176A and 176B proteins is associated with cancer pathology. *Acta histochemica* **114**, 705-712, doi:10.1016/j.acthis.2011.12.006 (2012).

192 Goyal, N. K., Fiks, A. G. & Lorch, S. A. Association of late-preterm birth with asthma in young children: practice-based study. *Pediatrics* **128**, e830-838, doi:10.1542/peds.2011-0809 (2011).

193 Wang, Y. F., Wu, L. Q., Liu, Y. N., Bi, Y. Y. & Wang, H. Gestational age and childhood leukemia: A meta-analysis of epidemiologic studies. *Hematology (Amsterdam, Netherlands)* **23**, 253-262, doi:10.1080/10245332.2017.1396056 (2018).

194 Sonntag, B. *et al.* Preterm birth but not mode of delivery is associated with an increased risk of developing inflammatory bowel disease later in life. *Inflammatory bowel diseases* **13**, 1385-1390, doi:10.1002/ibd.20206 (2007).

195 Li, S. *et al.* Preterm birth and risk of type 1 and type 2 diabetes: systematic review and meta-analysis. *Obesity reviews : an official journal of the International Association for the Study of Obesity* **15**, 804-811, doi:10.1111/obr.12214 (2014).

196 Deng, H., Xie, C., Ye, Y. & Du, Z. MicroRNA-1296 expression is associated with prognosis and inhibits cell proliferation and invasion by Wnt signaling in non-small cell lung cancer. *Oncology letters* **19**, 623-630, doi:10.3892/ol.2019.11154 (2020).

197 Majid, S. *et al.* Regulation of Minichromosome Maintenance Gene Family by MicroRNA-1296 and Genistein in Prostate Cancer. **70**, 2809-2818, doi:10.1158/0008-5472.CAN-09-4176 %J Cancer Research (2010).

198 Phan, B. *et al.* Tumor suppressor role of microRNA-1296 in triple-negative breast cancer. *Oncotarget* **7**, 19519-19530, doi:10.18632/oncotarget.6961 (2016).

199 Tao, Y. *et al.* MicroRNA-1296 Facilitates Proliferation, Migration And Invasion Of Colorectal Cancer Cells By Targeting SFPQ. *Journal of Cancer* **9**, 2317-2326, doi:10.7150/jca.25427 (2018).

200 Xu, Q. *et al.* MicroRNA-1296 inhibits metastasis and epithelial-mesenchymal transition of hepatocellular carcinoma by targeting SRPK1-mediated PI3K/AKT pathway. *Molecular cancer* **16**, 103, doi:10.1186/s12943-017-0675-y (2017).

201 Cakmak, H. A. *et al.* The prognostic value of circulating microRNAs in heart failure: preliminary results from a genome-wide expression study. *Journal of cardiovascular medicine (Hagerstown, Md.)* **16**, 431-437, doi:10.2459/jcm.0000000000000233 (2015).

202 Friedman, J. R. & Kaestner, K. H. The Foxa family of transcription factors in development and metabolism. *Cellular and Molecular Life Sciences CMLS* **63**, 2317-2328, doi:10.1007/s00018-006-6095-6 (2006).

203 Lee, C. S., Friedman, J. R., Fulmer, J. T. & Kaestner, K. H. The initiation of liver development is dependent on Foxa transcription factors. *Nature* **435**, 944-947, doi:10.1038/nature03649 (2005).

204 Wan, H. *et al.* Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *The Journal of biological chemistry* **280**, 13809-13816, doi:10.1074/jbc.M414122200 (2005).

205 Wan, H. *et al.* Foxa2 regulates alveolarization and goblet cell hyperplasia. *Development (Cambridge, England)* **131**, 953-964, doi:10.1242/dev.00966 (2004).

206 Wang, Y. *et al.* Discovery of potential asthma targets based on the clinical efficacy of Traditional Chinese Medicine formulas. *J Ethnopharmacol* **252**, 112635, doi:10.1016/j.jep.2020.112635 (2020).

207 Carmona, J. J. *et al.* Short-term airborne particulate matter exposure alters the epigenetic landscape of human genes associated with the mitogen-activated protein kinase network: a cross-sectional study. *Environmental health : a global access science source* **13**, 94, doi:10.1186/1476-069x-13-94 (2014).

208 Alvarez, A. & Woolf, P. J. RegNetB: predicting relevant regulator-gene relationships in localized prostate tumor samples. *BMC bioinformatics* **12**, 243, doi:10.1186/1471-2105-12-243 (2011).

209 Ye, J., Liu, H., Xu, Z.-L., Zheng, L. & Liu, R.-Y. Identification of a multidimensional transcriptome prognostic signature for lung adenocarcinoma. **33**, e22990, doi:10.1002/jcla.22990 (2019).

210 Brown, C. O. *et al.* Scavenger receptor class A member 3 (SCARA3) in disease progression and therapy resistance in multiple myeloma. *Leukemia research* **37**, 963-969, doi:10.1016/j.leukres.2013.03.004 (2013).

211 Karachanak-Yankova, S. *et al.* Epigenetic alterations in patients with type 2 diabetes mellitus. *Balkan J Med Genet* **18**, 15-24, doi:10.1515/bjmg-2015-0081 (2016).

212 Li, J.-W. *et al.* Interactome-transcriptome analysis discovers signatures complementary to GWAS Loci of Type 2 Diabetes. *Scientific reports* **6**, 35228-35228, doi:10.1038/srep35228 (2016).

213 Lv, D. *et al.* Genetic variations in SEC16B, MC4R, MAP2K5 and KCTD15 were associated with childhood obesity and interacted with dietary behaviors in Chinese school-age population. *Gene* **560**, 149-155, doi:10.1016/j.gene.2015.01.054 (2015).

214 Ng, M. C. *et al.* Implication of genetic variants near NEGR1, SEC16B, TMEM18, ETV5/DGKG, GNPDA2, LIN7C/BDNF, MTCH2, BCDIN3D/FAIM2, SH2B1, FTO, MC4R, and KCTD15 with obesity and type 2 diabetes in 7705 Chinese. *The Journal of clinical endocrinology and metabolism* **95**, 2418-2425, doi:10.1210/jc.2009-2077 (2010).

215 Andersson, C. *et al.* An Integrated Multi-Omics Approach to Identify Genetic Underpinnings of Heart Failure and its Echocardiographic Precursors: The Framingham Heart Study. *Circ Genom Precis Med*, doi:10.1161/circgen.118.002489 (2019).

216 Estimates, trends, and drivers of the global burden of type 2 diabetes attributable to PM(2·5) air pollution, 1990-2019: an analysis of data from the Global Burden of Disease Study 2019. *The Lancet. Planetary health* **6**, e586-e600, doi:10.1016/s2542-5196(22)00122-x (2022).

217 Lin, L. *et al.* Global association between atmospheric particulate matter and obesity: A systematic review and meta-analysis. *Environmental research* **209**, 112785, doi:https://doi.org/10.1016/j.envres.2022.112785 (2022).

218 Du, Y., Xu, X., Chu, M., Guo, Y. & Wang, J. Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence. *J Thorac Dis* **8**, E8-e19, doi:10.3978/j.issn.2072-1439.2015.11.37 (2016).

219 Yang, W., Ni, H., Wang, H. & Gu, H. NLRP3 inflammasome is essential for the development of chronic obstructive pulmonary disease. *International journal of clinical and experimental pathology* **8**, 13209-13216 (2015).

220 Birrell, M. A. & Eltom, S. The role of the NLRP3 inflammasome in the pathogenesis of airway disease. *Pharmacology & therapeutics* **130**, 364-370, doi:10.1016/j.pharmthera.2011.03.007 (2011).

221 Hu, W. P., Zeng, Y. Y., Zuo, Y. H. & Zhang, J. Identification of novel candidate genes involved in the progression of emphysema by bioinformatic methods. *International journal of chronic obstructive pulmonary disease* **13**, 3733-3747, doi:10.2147/copd.S183100 (2018).

222 Almansa, R. *et al.* Critical COPD respiratory illness is linked to increased transcriptomic activity of neutrophil proteases genes. *BMC research notes* **5**, 401, doi:10.1186/1756-0500-5-401 (2012).

223 Barnawi, J., Jersmann, H., Haberberger, R., Hodge, S. & Meech, R. Reduced DNA methylation of sphingosine-1 phosphate receptor 5 in alveolar macrophages in COPD: A potential link to failed efferocytosis. *Respirology (Carlton, Vic.)* **22**, 315-321, doi:10.1111/resp.12949 (2017).

224 Ahn, S. V. *et al.* Cancer development in patients with COPD: a retrospective analysis of the National Health Insurance Service-National Sample Cohort in Korea. *BMC pulmonary medicine* **20**, 170, doi:10.1186/s12890-020-01194-8 (2020).

225 Proskocil, B. J. & Fryer, A. D. Beta2-agonist and anticholinergic drugs in the treatment of lung disease. *Proceedings of the American Thoracic Society* **2**, 305-310; discussion 311-302, doi:10.1513/pats.200504-038SR (2005).

226 Oldenburger, A., Maarsingh, H. & Schmidt, M. Multiple facets of cAMP signalling and physiological impact: cAMP compartmentalization in the lung. *Pharmaceuticals (Basel, Switzerland)* **5**, 1291-1331, doi:10.3390/ph5121291 (2012).

227 Belmonte, K. E. Cholinergic pathways in the lungs and anticholinergic therapy for chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society* **2**, 297-304; discussion 311-292, doi:10.1513/pats.200504-043SR (2005).

228 Nuñez, F. J. *et al.* Glucocorticoids rapidly activate cAMP production via G(αs) to initiate non-genomic signaling that contributes to one-third of their canonical genomic effects. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **34**, 2882-2895, doi:10.1096/fj.201902521R (2020).

229 Dunne, A. E. *et al.* Direct Inhibitory Effect of the PDE4 Inhibitor Roflumilast on Neutrophil Migration in Chronic Obstructive Pulmonary Disease. *American journal of respiratory cell and molecular biology* **60**, 445-453, doi:10.1165/rcmb.2018-0065OC (2019).

230 Mercado, N., Ito, K. & Barnes, P. J. Accelerated ageing of the lung in COPD: new concepts. *Thorax* **70**, 482-489, doi:10.1136/thoraxjnl-2014-206084 (2015).

231 MacNee, W. Is Chronic Obstructive Pulmonary Disease an Accelerated Aging Disease? *Annals of the American Thoracic Society* **13 Suppl 5**, S429-s437, doi:10.1513/AnnalsATS.201602-124AW (2016).

232 Li, L. *et al.* FLI1 Exonic Circular RNAs as a Novel Oncogenic Driver to Promote Tumor Metastasis in Small Cell Lung Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **25**, 1302-1317, doi:10.1158/1078-0432.Ccr-18-1447 (2019).

233 Lin, S. F., Wu, C. C. & Chai, C. Y. Increased FLI-1 Expression is Associated With Poor Prognosis in Non-Small Cell Lung Cancers. *Applied immunohistochemistry & molecular morphology : AIMM* **24**, 556-561, doi:10.1097/pai.0000000000000227 (2016).

234 Zhao, X. *et al.* microRNA-214 Governs Lung Cancer Growth and Metastasis by Targeting Carboxypeptidase-D. *DNA and cell biology* **35**, 715-721, doi:10.1089/dna.2016.3398 (2016).

235 Hobbs, B. D. *et al.* Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nature genetics* **49**, 426-432, doi:10.1038/ng.3752 (2017).

236 Hobbs, B. D. & Cho, M. H. Dissecting respiratory disease heterogeneity through the genetics of diffusing capacity. *The European respiratory journal* **52**, doi:10.1183/13993003.01468-2018 (2018).

237 Wikenheiser, D. J. & Stumhofer, J. S. ICOS Co-Stimulation: Friend or Foe? **7**, doi:10.3389/fimmu.2016.00304 (2016).

238 Li, D. Y. & Xiong, X. Z. ICOS(+) Tregs: A Functional Subset of Tregs in Immune Diseases. *Frontiers in immunology* **11**, 2104, doi:10.3389/fimmu.2020.02104 (2020).

239 Honardoost, M., Rajabpour, A. & Vakili, L. Molecular epidemiology; New but impressive. *Medical journal of the Islamic Republic of Iran* **32**, 53, doi:10.14196/mjiri.32.53 (2018).

240 Vineis, P. & McMichael, A. J. Bias and confounding in molecular epidemiological studies: special considerations. *Carcinogenesis* **19**, 2063-2067, doi:10.1093/carcin/19.12.2063 (1998).

241 Sdona, E. *et al.* Dietary fibre in relation to asthma, allergic rhinitis and sensitization from childhood up to adulthood. *Clinical and translational allergy* **12**, e12188, doi:10.1002/clt2.12188 (2022).

242 Melén, E. *et al.* Male sex is strongly associated with IgE-sensitization to airborne but not food allergens: results up to age 24 years from the BAMSE birth cohort. *Clinical and translational allergy* **10**, 15, doi:10.1186/s13601-020-00319-w (2020).

243 Rothman, K. J. & Greenland, S. Modern epidemiology. 2nd ed . Philadelphia, PA: LippincottRaven. (1998).

244 Hoffman, C. S. *et al.* Comparison of gestational age at birth based on last menstrual period and ultrasound during the first trimester. *Paediatric and perinatal epidemiology* **22**, 587-596, doi:10.1111/j.1365-3016.2008.00965.x (2008).

245 de Hoogh, K. *et al.* Comparing land use regression and dispersion modelling to assess residential exposure to ambient air pollution for epidemiological studies. *Environment international* **73**, 382-392, doi:10.1016/j.envint.2014.08.011 (2014).

246 Bellander, T., Wichmann, J. & Lind, T. Individual exposure to NO2 in relation to spatial and temporal exposure indices in Stockholm, Sweden: the INDEX study. *PloS one* **7**, e39536, doi:10.1371/journal.pone.0039536 (2012).

247 Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome biology* **17**, 208, doi:10.1186/s13059-016-1066-1 (2016).

248 Heiss, J. A. & Just, A. C. Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clinical epigenetics* **10**, 73, doi:10.1186/s13148-018-0504-1 (2018).

249 Liu, F. *et al.* Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. *BMC genomics* **8**, 153, doi:10.1186/1471-2164-8-153 (2007).

250 KJ, R. Epidemiology: an introduction. New York, USA: Oxford university press Inc. (2002).

251 Zhang, Y., Florath, I., Saum, K. U. & Brenner, H. Self-reported smoking, serum cotinine, and blood DNA methylation. *Environmental research* **146**, 395-403, doi:10.1016/j.envres.2016.01.026 (2016).

252 Maas, S. C. E. *et al.* Smoking-related changes in DNA methylation and gene expression are associated with cardio-metabolic traits. *Clinical epigenetics* **12**, 157, doi:10.1186/s13148-020-00951-0 (2020).

253 Kodal, J. B., Kobylecki, C. J., Vedel-Krogh, S., Nordestgaard, B. G. & Bojesen, S. E. AHRR hypomethylation, lung function, lung function decline and respiratory symptoms. *The European respiratory journal* **51**, doi:10.1183/13993003.01512-2017 (2018).

254 Sharp, G. C. *et al.* Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. *Human molecular genetics* **26**, 4067-4085, doi:10.1093/hmg/ddx290 (2017).

255 Sharp, G. C. *et al.* Maternal alcohol consumption and offspring DNA methylation: findings from six general population-based birth cohorts. *Epigenomics* **10**, 27-42, doi:10.2217/epi-2017-0095 (2018).

256 Pratt, G. C., Vadali, M. L., Kvale, D. L. & Ellickson, K. M. Traffic, air pollution, minority and socio-economic status: addressing inequities in exposure and risk. *International journal of environmental research and public health* **12**, 5355-5372, doi:10.3390/ijerph120505355 (2015).

257 Kere, M. *et al.* Effects of inhaled corticosteroids on DNA methylation in peripheral blood cells in children with asthma. *Allergy* **75**, 688-691, doi:10.1111/all.14043 (2020).

258 McDougall, C. M. *et al.* Nasal epithelial cells as surrogates for bronchial epithelial cells in airway inflammation studies. *American journal of respiratory cell and molecular biology* **39**, 560-568, doi:10.1165/rcmb.2007-0325OC (2008).