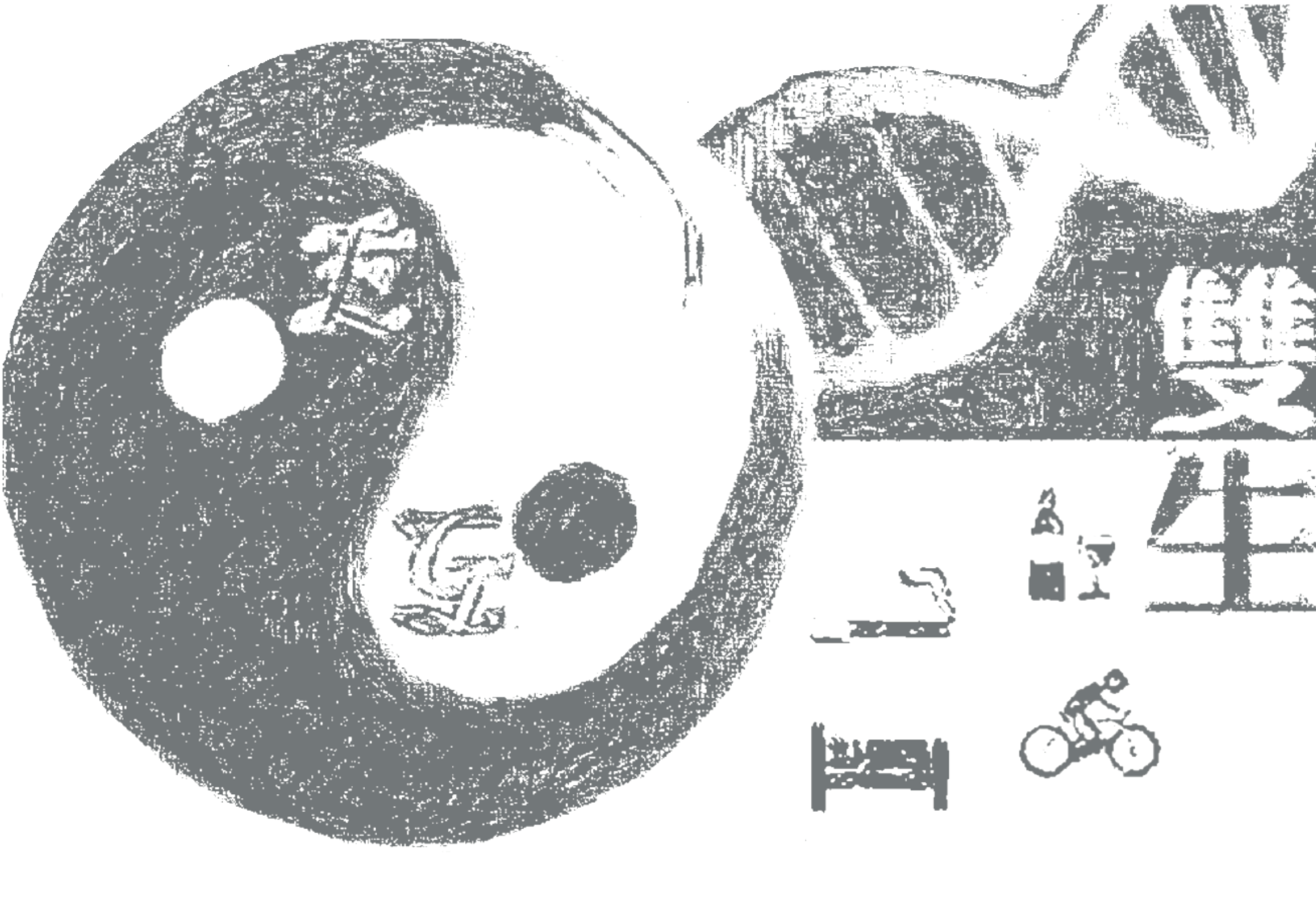
Thesis for doctoral degree (Ph.D.)
2018

# Genetic Epidemiology of Cardiometabolic Biomarkers: Twin Studies in the Genomic Era



Xu Chen

Karolinska Institutet

From the Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, Stockholm, Sweden

# GENETIC EPIDEMIOLOGY OF CARDIOMETABOLIC BIOMARKERS: TWIN STUDIES IN THE GENOMIC ERA

**Xu Chen**

陳旭

Stockholm 2018

All previously published papers were reproduced with permission from the publishers.

Front cover and all figures are illustrated by the author.

心血管代谢生物标志物的遗传流行病学: 基因组时代的双生子研究/陈旭, 2018.6.15

Published by Karolinska Institutet.

Printed by E-Print AB 2018.

ISBN 978-91-7831-037-1

9 789178 310371

# Genetic Epidemiology of Cardiometabolic Biomarkers: Twin Studies in the Genomic Era

THESIS FOR DOCTORAL DEGREE (Ph.D.)

Public defense at the lecture hall Petrén, Nobels väg 12B,
Solna campus, Karolinska Institutet, Stockholm, Sweden
**June 15th 2018, Friday, 09:00**

by

## Xu Chen

***Principal Supervisor:***

**Patrik K. E. Magnusson**
Associate Professor; PhD
*Karolinska Institutet*
*Department of Medical Epidemiology and Biostatistics*

***Co-supervisors:***

**Nancy L. Pedersen**
Professor; PhD
*Karolinska Institutet*
*Department of Medical Epidemiology and Biostatistics*

**Sara Hägg**
Associate Professor; PhD
*Karolinska Institutet*
*Department of Medical Epidemiology and Biostatistics*

**Johan Frostegård**
Professor; MD, PhD
*Karolinska Institutet*
*Institute of Environmental Medicine*

**Per Svensson**
Associate Professor; MD, PhD
*Karolinska Institutet*
*Department of Clinical Science and Education*
*Södersjukhuset, Department of Cardiology*

***Opponent:***

**Kerrin Small**
Senior Lecturer; PhD
*King's College London*
*Department of Twin Research and Genetic Epidemiology*

***Examination Board:***

**Liming Li**
Professor; MD, MPH
*Peking University Health Science Center*
*School of Public Health*
*Department of Epidemiology and Biostatistics*

**Lars Rönnegård**
Professor; PhD
*Dalarna University, Section of Statistics*
*The Swedish University of Agricultural Sciences*
*Department of Animal Breeding and Genetics*

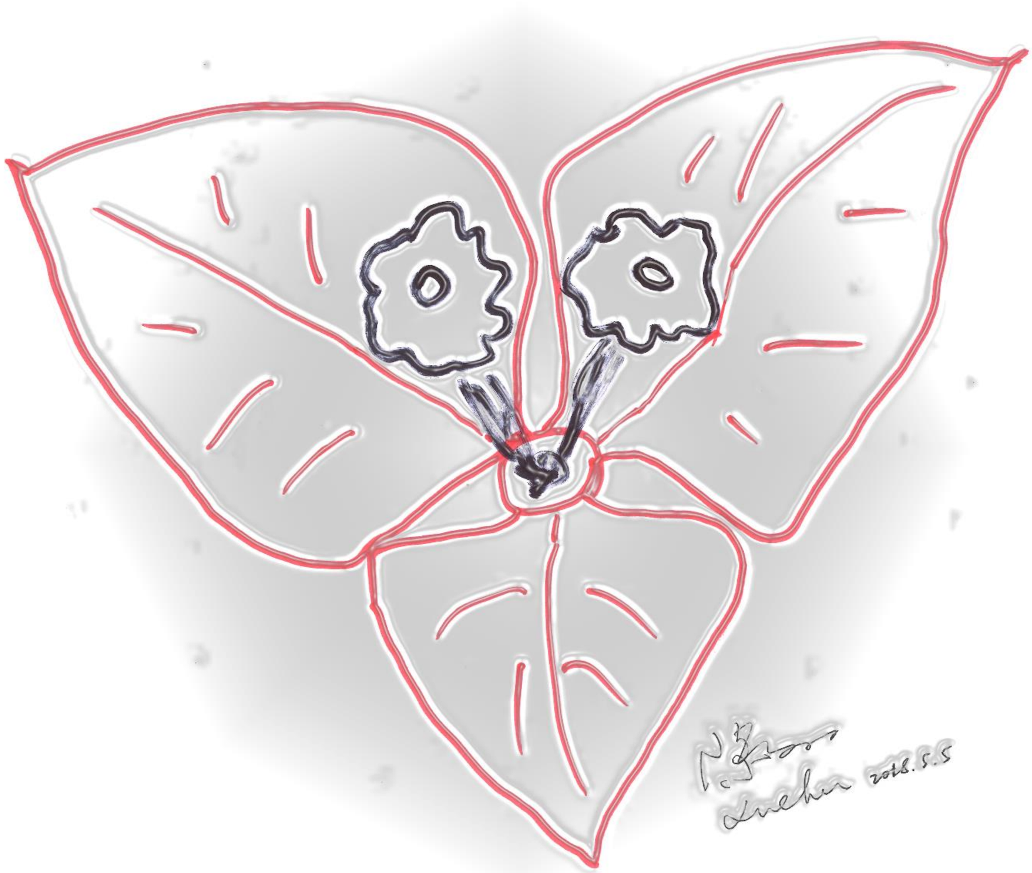
**Karin Broberg**
Professor; PhD
*Karolinska Institutet*
*Institute of Environmental Medicine*

*To all who appeared in the past thirty years of my life*

献给出现在我生命前三十年的所有人

*In memory of my dearest "waipo"*

仅此纪念未能等到我博士毕业归来的外婆



# ABSTRACT

Following the Human Genome Project, many genomic approaches have been developed in genetic epidemiology to investigate the genetic influences on human complex traits. This thesis aims to answer four genetic epidemiological questions for cardiometabolic biomarkers/ traits, by using classical twin studies and novel genomic methods.

*Whether the dominant genetic effects are important for "missing heritability"?* Heritability is a population specific estimate reflecting the relative importance of genes (versus environment) for human complex traits. "Missing heritability" is the proportion of heritability that remains unexplained by single nucleotide polymorphisms (SNPs). For 24 cardiometabolic traits, the univariate (**study I**) and bivariate (**study II**) heritabilities were estimated by using both twin and SNP models, within the same study base (10,682 twins in TwinGene). **Study I** supports that the main genetic influences on these traits are additive genetic effects (A), but significant contributions from dominant genetic effects (D) are also identified for certain traits. D effects are often masked by shared environment (C) in twin studies, thus D might have a more prominent role than what the estimates often suggest. It is difficult to distinguish D from A in too small twin studies, so the "missing heritability'' might be overestimated if all genetic influences (A and D) are erroneously attributed to the narrow-sense heritability.

*What's the pattern of genetic and environmental contributions to the covariation between cardiometabolic traits?* **Study II** demonstrates that the pattern varies by different clusters of cardiometabolic traits. Additive genetic effects (A) and non-shared environment (E) influence the covariation between blood pressure traits. Besides A and E, dominant genetic effects appear to be important for the covariation between obesity traits. However, shared environmental contributions seem generally to be weak between cardiometabolic traits in TwinGene samples.

*Which genetic variants are associated with the novel cardiometabolic biomarker — immunoglobulin M against phosphorylcholine (IgM anti-PC)?* By performing genome-wide association study (GWAS) in four Swedish cohorts (total n=3,648), **study III** identified a haplotype block at 11q24.1 close to the *GRAMD1B* gene to be the top locus shared between anti-PC and chronic lymphocytic leukemia (CLL). Prediction from bioinformatics suggests that the SNP rs35923643-G in this locus might be the functional variant by impeding the transcription factor binding. A small nested case-control study indicates a potential reverse causation between anti-PC and CLL.

*Whether the associations between blood lipids and amyotrophic lateral sclerosis (ALS) are causal?* By using summary GWAS results (~100,000 individuals for blood lipids and ~30,000 for ALS) in the polygenic risk score and Mendelian randomization settings, **study IV** tested the association and causality between blood lipids and ALS. It supports that high levels of low-density lipoprotein (LDL) and total cholesterol (TC) are risk factors for ALS. Based on current assumptions and evidence, it also suggests potential causal effects of LDL and TC on ALS.

In summary, this thesis quantified the proportion of genetic contributions to the variation (**study I**) and covariation (**study II**) for 24 traditional cardiometabolic biomarkers/traits; it identified the genetic variants (common SNPs) associated with novel biomarker IgM anti-PC (**study III**); it also tested whether polygenic evidence supports the association and causality between blood lipids and ALS (**study IV**). In general, the thesis suggests that twin studies have continuing important values for genetic epidemiology in the genomic era.

## LIST OF SCIENTIFIC PAPERS

This thesis is based on the following original articles, which have been referred in the text by their Roman numerals (I-IV).

I. **Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models.**
*Chen X, Kuja-Halkola R, Rahman I, Arpegård J, Viktorin A, Karlsson R, Hägg S, Svensson P, Pedersen NL, Magnusson PK.*
American Journal of Human Genetics. 2015, 97(5):708-714.

II. **Genetic and Environmental Contributions to the Covariation between Cardio-metabolic Traits.**
*Chen X, Kuja-Halkola R, Chang Z, Karlsson R, Hägg S, Svensson P, Pedersen NL, Magnusson PK.*
Journal of the American Heart Association. 2018, 7(9): e007806.

III. **A Genome-wide Association Study of IgM Antibody against Phosphorylcholine: Shared Genetics and Phenotypic Relationship to Chronic Lymphocytic Leukemia.**
*Chen X, Gustafsson S, Whitington T, Borné Y, Lorentzen E, Sun J, Almgren P, Su J, Karlsson R, Song J, Lu Y, Zhan Y, Hägg S, Svensson P, Smedby KE, Slager SL, Ingelsson E, Lindgren CM, Morris AP, Melander O, Karlsson T, de Faire U, Caidahl K, Engström G, Lind L, Karlsson MC, Pedersen NL, Frostegård J, Magnusson PK.*
Human Molecular Genetics. 2018, 27(10): 1809-1818.

IV. **Polygenic Link between Blood Lipids and Amyotrophic Lateral Sclerosis.**
*Chen X, Yazdani S, Piehl F, Magnusson PK, Fang F.*
Neurobiology of Aging. 2018, 67: 202.e1-202.e6.

## RELATED PAPERS

Not included in this thesis (*equal contribution, #joint supervision).

V. **One CNV Discordance in _NRXN1_ Observed upon Genome-wide Screening in 38 Pairs of Adult Healthy Monozygotic Twins.**
_Magnusson PK, Lee D, Chen X, Szatkiewicz J, Pramana S, Teo S, Sullivan PF, Feuk L, Pawitan Y._
Twin Research and Human Genetics. 2016, 19(2):97-103.

VI. **Cystatin C Predicts Incident Cardiovascular Disease in Twins.**
_Arpegård J, Magnusson PK, Chen X, Ridefelt P, Pedersen NL, de Faire U, Svensson P._
Journal of the American Heart Association. 2016, 5(6): e003085.

VII. **A Large-scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure.**
_Sung YJ*, Winkler TW*, de Las Fuentes L*, Bentley AR*, Brown MR*, Kraja AT*, Schwander K*, Ntalla I*, …[32 authors]…, Chen X, …[248 authors]…, Caulfield MJ#, Elliott P#, Rice K#, Munroe PB#, Morrison AC#, Cupples LA#, Rao DC#, Chasman DI#._
American Journal of Human Genetics. 2018, 102(3):375-400.

VIII. **Novel Genetic Associations for Blood Pressure Identified via Gene-alcohol Interaction in up to 570K Individuals across Multiple Ancestries.**
_Feitosa MF*, Kraja AT*, Chasman DI*, Sung YJ*, Winkler TW*, Ntalla I*, …[31 authors]…, Chen X, …[228 authors]…, Rice K#, Morrison AC#, Elliott P#, Caulfield MJ#, Munroe PB#, Rao DC#, Province MA#, Levy D#._
PLoS One. 2018, in press.

IX. **Genetic and Environmental Influences on the Association between Metabolic Syndrome and Chronic Kidney Disease.**
_Bhuiyan I*, Chen X*#, Kuja-Halkola R, Magnusson PK, Svensson P#._
Manuscript.

X. **Apolipoprotein C3 Concentrations, Lipoprotein Concentrations and Risk of Incident Coronary Heart Disease.**
_Leander K*, Chen X*, Mannila M, Magnusson PK, Silveira A, van 't Hooft FM._
Manuscript.

# CONTENTS

**APPENDIX**

- **I:** Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models.
- **II:** Genetic and Environmental Contributions to the Covariation between Cardiometabolic Traits.
- **III:** A Genome-wide Association Study of IgM anti-PC: Shared Genetics and Phenotypic Relationship to Chronic Lymphocytic Leukemia.
- **IV:** Polygenic Link between Blood Lipids and Amyotrophic Lateral Sclerosis.

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **A** | Additive genetic effects |
| **$a^2$** | Additive genetic variance |
| **ACE** | Model including A, C, E components |
| **ADE** | Model including A, D, E components |
| **AE** | Model including A, E components |
| **AIC** | Akaike information criterion |
| **ALS** | Amyotrophic lateral sclerosis |
| **apoA1** | Apolipoprotein A1 |
| **apoB** | Apolipoprotein B |
| **ASCVD** | Atherosclerotic cardiovascular disease |
| **BMI** | Body mass index |
| **C** | Common/shared environment |
| **$c^2$** | Common/shared environmental variance |
| **CAD** | Coronary artery disease |
| **CLL** | Chronic lymphocytic leukemia |
| **Crea** | Creatinine |
| **CRP** | C-reactive protein |
| **CVD** | Cardiovascular disease |
| **CysC** | Cystatin C |
| **D** | Dominant genetic effects |
| **$d^2$** | Dominant genetic variance |
| **DBP** | Diastolic blood pressure |
| **DZ** | Dizygotic twin |
| **E** | Unique/non-shared environment |
| **$e^2$** | Unique/non-shared environmental variance |
| **EEA** | Equal environment assumption |
| **GCTA** | Genome-wide Complex Trait Analysis tool |
| ***GRAMD1B*** | Gram domain containing 1b gene |
| **GREML** | Genomic-relatedness-matrix restricted maximum likelihood |
| **GWAS** | Genome-wide association study |
| **$h^2$** | Narrow-sense heritability |
| **$H^2$** | Broad-sense heritability |
| **HbA1c** | Glycated hemoglobin, hemoglobin A1c |

| | |
|---|---|
| **HCY** | Homocysteine |
| **HDL** | High-density lipoprotein |
| **ICD** | International Classification of Diseases code |
| **IgM** | Immunoglobulin M |
| **IV** | Instrumental variable |
| **LDL** | Low-density lipoprotein |
| **LDSC** | Linkage disequilibrium score regression |
| **Lp(a)** | Lipoprotein (a) |
| **Lp-PLA2** | Lipoprotein-associated phospholipase A2 |
| **MAP** | Mean arterial pressure |
| **MDC** | Malmö Diet and Cancer cohort |
| **MI** | Myocardial infarction |
| **MR** | Mendelian randomization |
| **MZ** | Monozygotic twin |
| **PAD** | Peripheral artery disease |
| **PC** | Phosphorylcholine |
| **PIVUS** | Prospective Investigation of the Vasculature in Uppsala Seniors cohort |
| **PP** | Pulse pressure |
| **PRACSIS** | Prognosis and Risk in Acute Coronary Syndromes in Sweden cohort |
| **PRS** | Polygenic risk score |
| **$P_T$** | P-value threshold |
| **rDZ** | Intra-pair correlation in dizygotic twins |
| **rMZ** | Intra-pair correlation in monozygotic twins |
| **SBP** | Systolic blood pressure |
| **SEM** | Structural equation model |
| **SNP** | Single nucleotide polymorphism |
| **TC** | Total cholesterol |
| **TG** | Triglycerides |
| **WHR** | Waist-hip ratio |

# 1 BACKGROUND

## 1.1 Cardiometabolic traits: diseases and biomarkers

According to the latest fact sheets from the World Health Organization, cardiometabolic diseases have been the largest global mortality burden during 2000-2015 [1]. Cardiovascular diseases (CVDs) account for ~30% of deaths around the world, in which ischemic heart disease and stroke contribute most, with 15 million deaths in 2015 [2]. Diabetes, a common metabolic disease, is the sixth strongest killer accounting for 1 million deaths in 2000 and 1.6 million deaths in 2015 [1].

Cardiometabolic traits, including different types of cardiovascular and metabolic diseases, as well as a large number of related risk factors/biomarkers (e.g. high blood lipids, abdominal obesity) and complications (e.g. dyslipidemia, hyperglycemia, hypertension, insulin resistance, and declined kidney function) [3], display a lot of overlaps and interdependencies (*Figure 1.1*). During 1980-2010, high levels of four cardiometabolic biomarkers [blood pressure, fasting glucose, serum cholesterol and body mass index (BMI)] contributed to 65% of global mortality due to three major chronic/cardiometabolic diseases: CVD, chronic kidney disease and diabetes [4].

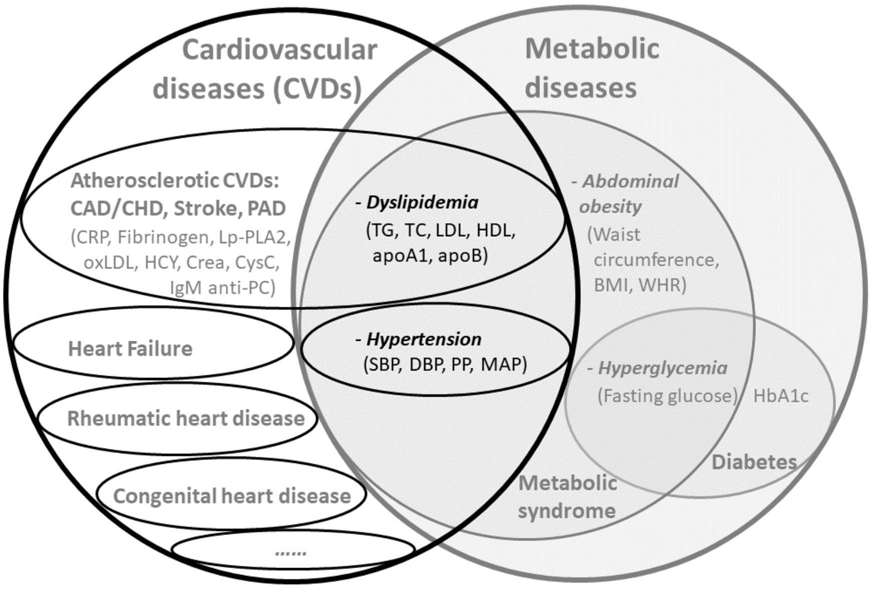


**Figure 1.1. Venn diagram over cardiometabolic traits**

Biomarkers are efficient indicators for the development of diseases [5], and some of them also constitute modifiable risk factors for the prevention and management of diseases [6]. To date, many cardiometabolic biomarkers are well established in current guidelines and widely used in clinical practice [7]. At the same time, novel biomarkers, reflecting different pathophysiological processes or displaying potential values in the clinical diagnosis and prevention, have been discovered for cardiometabolic diseases [8, 9].

### 1.1.1 Blood lipids, dyslipidemia, and ASCVD

In blood, lipids (mainly fatty acids and cholesterol) are bound by apolipoproteins (apo) and transported as lipoproteins. Dyslipidemia usually refers to the dysregulation of blood lipids, which is the primary modifiable risk factor for atherosclerotic CVD (ASCVD) such as myocardial infarction (MI) [10]. Several types of blood lipids have been used for the clinical management of CVD [7]: including triglycerides (TG), total cholesterol (TC), low-density lipoprotein (LDL), high-density lipoprotein (HDL), apoA1, apoB and lipoprotein (a) [Lp(a)]. Lp(a) is an LDL-like particle that also contains apo(a), high level of Lp(a) is suggested to be an independent risk factor for CVD [11].

### 1.1.2 Inflammatory biomarkers and atherosclerosis

Atherosclerosis is the predominant pathological process underlying ASCVD, in which plaques are mainly formed by the accumulation of lipids and immune competent cells [2]. Nowadays, atherosclerosis is regarded as a lipids-driven chronic inflammation process [12]. Atherosclerosis is initiated by the intracellular LDL accumulation, the LDL is susceptible to be oxidized into oxLDL by oxygen radicals or enzymes [13]. C-reactive protein (CRP) and fibrinogen are two inflammatory biomarkers recommended in current guidelines, but the specificity and sensitivity of them appear to be low for CVD diagnosis [7]. The lipoprotein-associated phospholipase A2 (Lp-PLA2) produced by inflammatory cells can bind to apoB on LDL, playing pro-inflammatory role in atherosclerosis [14]. The activity and mass of Lp-PLA2 are associated with coronary artery disease (CAD) and stroke [15].

### 1.1.3 Other metabolic biomarkers

Metabolic disorders occur in a wide range of metabolic processes (e.g. the biosynthesis and catabolism of carbohydrates, proteins and also lipids). Metabolic syndrome is a cluster of risk factors like abdominal obesity, hyperglycemia, hypertension, and dyslipidemia [16].

Waist circumference is the key measurement for abdominal obesity. Other obesity traits include weight, body mass index (BMI), hip circumference, waist-hip ratio (WHR) and so on. Most of them are important risk factors or predictors for cardiometabolic diseases [17].

Glycated hemoglobin (HbA1c) is not only the long-term biomarker for diabetes, but also a strong and independent risk factor for CAD [18].

Four blood pressure measurements are commonly used in clinics: systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), and pulse pressure

(PP). In the latest guideline issued in 2017 [19], blood pressure (in mmHg) has been categorized into normal (SBP<120 and DBP<80), elevated (SBP 120-129 and DBP<80), stage I hypertension (SBP 130-139 or DBP 80-89), and stage II hypertension (SBP≥140 or DBP>90).

Homocysteine (HCY) is involved in different processes of atherosclerosis, and high level of HCY indicates increased CAD risk [20]. Two blood biomarkers reflecting kidney function, cystatin C (Cys C) and creatinine (Crea), are also reported to be positively associated with the risk of MI and stroke [21].

## 1.2 Genetic epidemiology of cardiometabolic traits

Cardiometabolic traits, influenced by both genes and environment, are among the most commonly studied human complex phenotypes. Since 1980s, genetic epidemiology has been developed as an interdisciplinary subject to investigate the genetic influences on human complex traits; by using theories, designs and methodologies from the genetics, medical epidemiology and biostatistics. However, genetic epidemiology is also a special research field that mainly focus on genetic factors and family aggregation at the population level [22].

In the past decades, genetic epidemiological studies have been performed for many traditional cardiometabolic traits. The relative importance and identification of the genetic factors shed more lights on the genetic etiology and the molecular mechanisms linking biomarkers and diseases. The association and causality tested by genetic epidemiological methods also provide less biased genetic evidence for the observational findings [23, 24].

### 1.2.1 Whether genetic factors are important for a single trait?

This is the most basic "nature versus nurture" question in genetic epidemiology. *Heritability* is a concept to reflect the relative importance between genes and environment, which is defined as the proportion of the phenotypic variation attributed to genetic effects [25].

#### 1.2.1.1 Heritability estimation

Heritability can be estimated from several methods: either based on family designs to compare the phenotypic similarities among relatives [26], or based on genomic methods to compare the phenotypic and genotypic similarities in related or unrelated subjects [27, 28].

***The classical twin study*** is the most common family-based approach to estimate heritability [29]. Heritabilities of more than 17,800 human complex traits have been estimated among 2,800 twin studies during the past fifty years [30]. In the classical twin study, the observed resemblance between monozygotic (MZ) and dizygotic (DZ) twin pairs are compared. Interpretations of the results rely on three basic assumptions: co-twins within the MZ pair share 100% while co-twins within the DZ pair share 50% of their segregating genes; co-twins within MZ and DZ pair share their raising environment to the same extent (*equal environment assumption, EEA*) [29, 31]. Therefore, classical twin-based *structural equation model (SEM)* usually decompose the phenotypic variation of each trait into three components: *additive genetic effects (A)*, *common/shared environmental effects (C)* and *unique/non-shared environmental effects (E)*. The proportion of A to the sum of A, C and E is defined as the *narrow-sense heritability ($h^2$)*, but most often referred to as just "heritability". From the meta-analysis of twin studies in the past fifty years, the average estimates of heritability are ~40% for cardio-vascular traits and ~60% for metabolic traits; and the average estimates of *shared environmental variance ($c^2$)* are less than 20% for these traits (*Figure 1.2*).

***Single nucleotide polymorphisms (SNPs)-based methods*** have been developed to estimate the heritability since 2010. Because SNPs are genotyped by using gene chip (DNA microarray), the SNP-based estimate of genetic variance is called "*chip heritability*" [24]. So far, there are two common SNP-based methods to estimate chip heritability: *genomic-relatedness-matrix restricted maximum likelihood (GREML)* and *linkage disequilibrium score (LDSC) regression*.

By comparing the genotypic and phenotypic similarities within the unrelated individuals, GREML can exclude C and just estimate the A and E [28]. LDSC regression can estimate the SNP-based heritability by using the LD scores from the reference population and mean $\chi^2$ statistics from the *genome-wide association study (GWAS)* in target population [32]. However, the heritability estimated from either GREML or LDSC just represents the phenotypic variation explained by SNPs, which is lower than the twin-based estimate that represents all genetic factors. The gap between twin- and SNP-based estimates of heritability is the main topic in the "*missing heritability*" debate.

#### 1.2.1.2 Missing heritability

The term missing heritability was coined in 2008, originally referring to the observation that genome-wide significant SNPs identified from GWAS explained an extremely small proportion (~5%) of the variation of human complex traits [33]. In 2010, GREML, which includes contributions from all common SNPs was developed [28]. The method captures much larger parts of the total genetic variation and provides larger estimates of SNP-based heritability (~30% to 50%) [34]. However, the gap between twin- and SNP-based heritability still remains large (the proportion of $h^2_{SNP}/h^2_{Twin}$ is usually less than 50%). So far, two major explanations have been proposed and further investigated: 1) the numerator $h^2_{SNP}$ is underestimated, because current SNP-based methods haven't included the rare SNPs or other types of genetic variants (e.g. copy number variations, insertions and deletions), nor accounting for *gene-environment interactions* [35]; 2) the denominator $h^2_{Twin}$ is overestimated in classical twin studies, due to potential violation of EEA, or falsely ascribing true *non-additive effects* to A (and thereby the narrow-sense heritability is overestimated) [36-38].

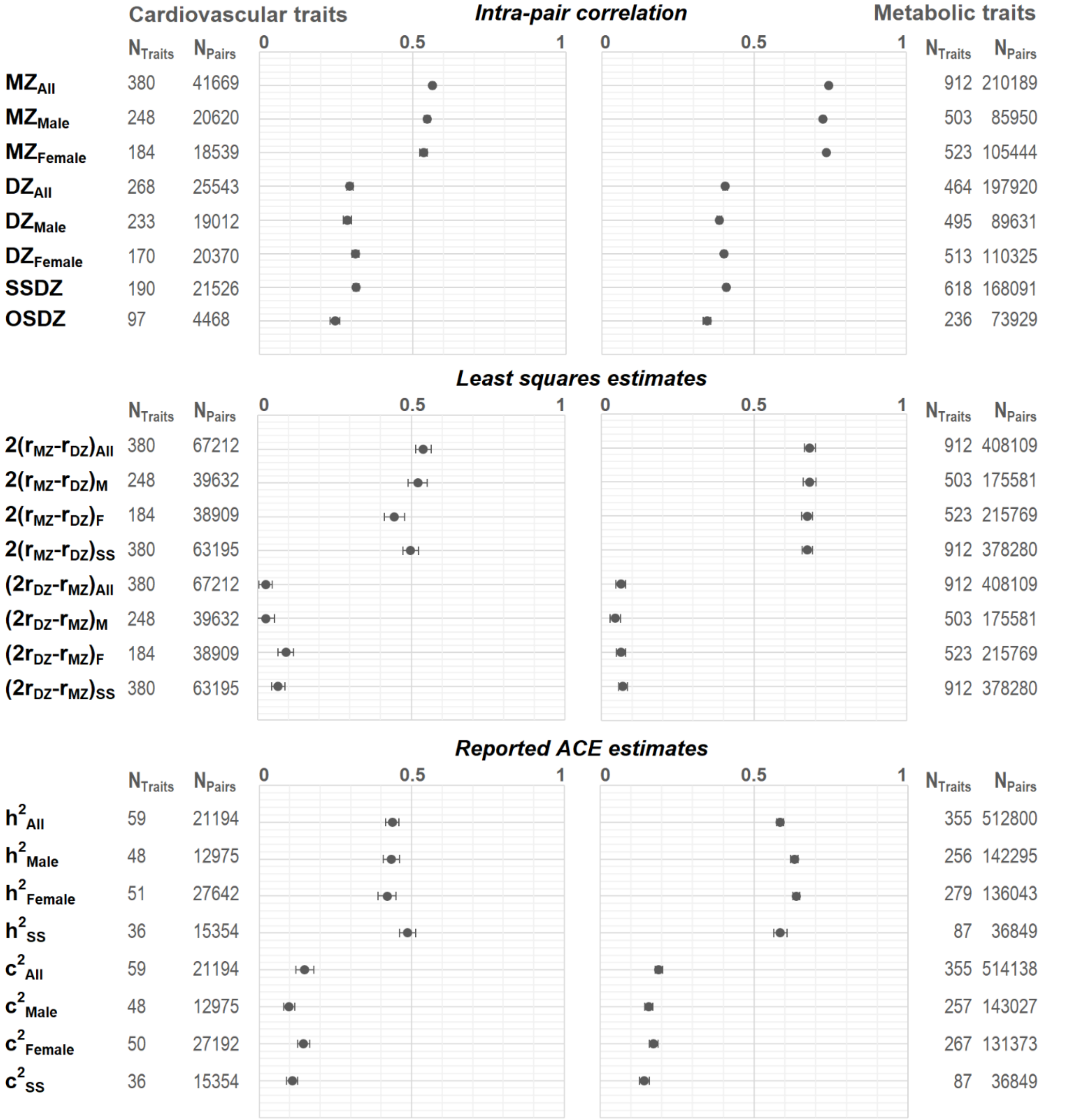


**Figure 1.2. Twin-based estimates for cardiometabolic traits in the past 50 years**

Data are from the MaTCH (*Polderman TJ, et al. Nature Genetics, 2015* [27]), re-plotted by the author.

SSDZ: same-sex dizygotic twins; OSDZ: opposite-sex dizygotic twins; rMZ, rDZ: intra-pair correlation coefficients in MZ and DZ, respectively; M: male; F: female; SS: same-sex; $h^2$: heritability; $c^2$: shared environmental variance.

### 1.2.1.3 Non-additive genetic effects

Non-additive genetic effects are the genetic interactions between alleles. Two types of non-additive genetic effects are usually defined: 1) *dominance* or *dominant genetic effects (D)*, representing the interactions between two alleles within the same locus; 2) *epistasis*, representing the interactions between alleles from different loci [39]. Because the classical twin study only can estimate a maximum of three variance components and D is dependent on additive genetic effect (A) of each allele, it is not possible to estimate C and D within the same model. Therefore, the phenotypic variation of each trait can be decomposed in either ACE, or ADE, or AE model in classical twin study. The proportion, (A+D)/(A+D+E) is defined as the *broad-sense heritability ($H^2$).*

Most classical twin designs and GWASs assume that the individual effect of each allele is additive. The latest meta-analysis of 50 years' twin studies suggests that the twin resemblance for 69% of 17,804 traits are only from additive genetic effects [30]. The newly developed GREML(dominant, d) method also find that SNP-based estimates of *dominant genetic variance ($d^2$)* are too small (~3%) to be able to explain the "missing heritability" [40]. However, certain classical- (with larger sample size) and extended- (including more family members) twin studies have identified significant and considerable $d^2$ (~30%) for many cardiometabolic traits [41-43].

**Therefore, two questions arose and motivated our study I:**

— *Why is there such a big difference between twin- and SNP-based estimates of $d^2$?*

— *Whether D is really not important for the "missing heritability"?*

## 1.2.2 Which proportion of genetic factors is shared between traits?

The genetic and environmental contributions to the covariation between two traits can also be estimated from the twin- and SNP-based bivariate models [44].

### 1.2.2.1 Bivariate heritability and genetic correlation

*Bivariate heritability* is defined as the proportion of two traits' phenotypic correlation explained by genetic factors; *genetic correlation* reflects the overlap of genetic factors between two traits [26]. Similar with the *univariate heritability*, the SNP-based estimates of bivariate heritability are notably lower than twin-based estimates (before our study II, comparisons were however only available between estimates obtained from different populations).

#### 1.2.2.2 Heritability is a population specific estimate

Heritability often varies by age, sex and other factors of the population samples. To the best of our knowledge, our study I might have been the first study to compare the twin- and SNP-based univariate heritabilities within the same population.

Moreover, our study I and other studies have indicated that small sample size might hamper accurate quantification of A, C or D and E contributions to the variation for certain traits. For example, no matter in large or small classical twin studies, ACE is usually the best-fitted model for human height ($a^2$≈80%, $c^2$≈10% and $e^2$≈10%); and the SNP-based estimate of $a^2$ is from 45% (using directed genotyped SNPs, [34]) to 56% (including more imputed SNPs, [45]). Therefore, the gap between twin- and SNP-based $h^2$ for height is just 30%-40%. While for BMI, AE or ACE is the most frequently reported model in small twin studies ($a^2$≈70% and $e^2$≈30%) [46]; but ADE model is reported in certain studies with larger samples ($a^2$≈30%, $d^2$≈40% and $e^2$≈30%) [42]. The particular pattern including different types of genetic and environmental contributions to the covariation between cardiometabolic traits haven't been comprehensively identified.

**Then, two questions were further illuminated in our study II:**

— *How A, C or D, E contribute to the covariation between cardiometabolic traits?*

— *Whether the bivariate twin- and SNP-based estimates also differ a lot?*

### 1.2.3 Where are the important genetic factors?

After quantifying the relative importance of genetic factors to the phenotypic variation and covariation, the natural next step in genetic epidemiology is to try to find the particular genetic factors.

#### 1.2.3.1 GWASs for traditional cardiometabolic traits

GWAS is a hypothesis-free and efficient design to identify the genetic factors (from genome-wide common SNPs) associated with human complex traits [47]. Since the first GWAS on age-related macular degeneration in 2005 [48], more than 68,000 SNP-trait associations have been identified from ~5,000 GWASs [47, 49]. Numerous genome-wide significance loci (association P-value<$5\times10^{-8}$) have also been identified for traditional cardiometabolic traits. Current results show: ~100 loci for blood lipids [50] and CAD [51, 52]; ~20 loci for CRP [53] and fibrinogen [54, 55]; ~10 loci for Lp-PLA2 [56-58]; ~60 loci for HbA1c [59];

~30 loci for fasting insulin and glucose [60]; ~20 loci for type I [61] and type II diabetes [62]; ~10 loci for HCY [63]; ~15 loci for Cys C and CKD [64]; ~80 loci for blood pressure [65]; hundreds of loci have also been identified for human anthropometric traits (e.g. height, BMI and so on) from large-scale studies [66].

#### 1.2.3.2 IgM anti-PC, a potential novel cardiometabolic biomarker

Phosphorylcholine (PC) is an exposed antigen on apoptotic cells, oxLDL and *Streptococcus pneumoniae* [67]. As shown in *Figure 1.3*, PC links to immunity, apoptosis, atherosclerosis, pathogens and chronic lymphocytic leukemia (CLL). Immunoglobulin M against PC (IgM anti-PC) induced by PC immunization can inhibit the uptake of oxLDL through macrophages, thus preventing the development of atherosclerosis [68]. In recent years, several studies have reported that IgM anti-PC is inversely associated with ASCVD risk and displays potential value for the prevention, diagnosis and therapy of atherosclerosis [69, 70]. However, the association between IgM anti-PC level and CLL risk has never been tested.

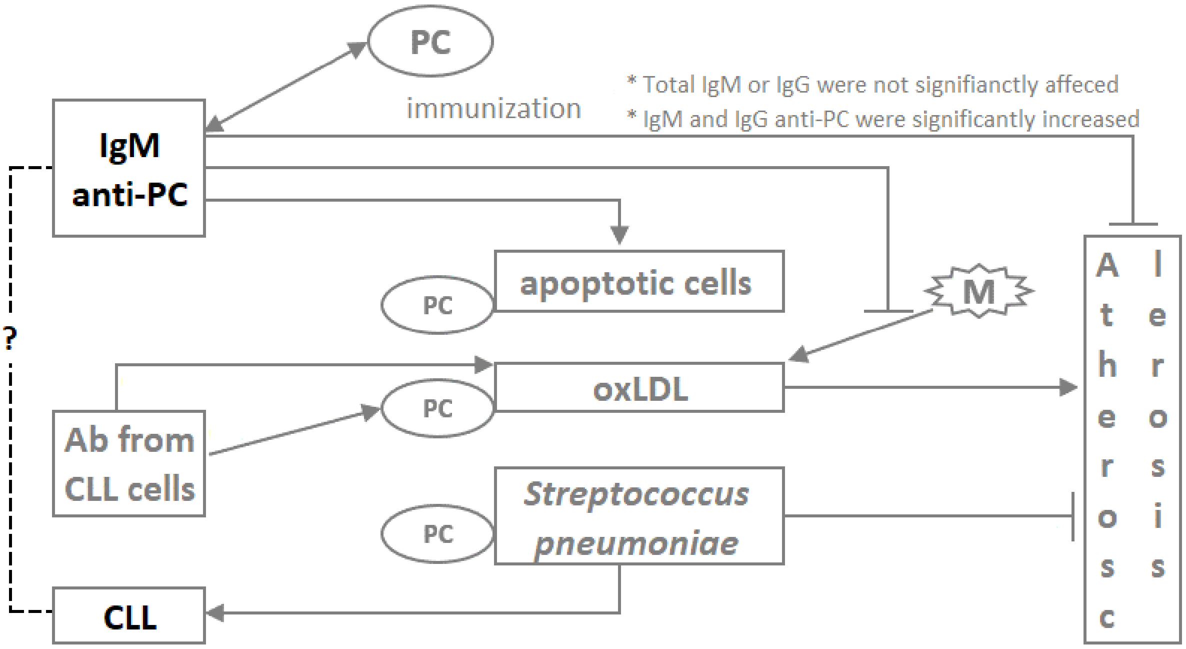


**Figure 1.3. Schematic overview of previous literature about IgM anti-PC**
Line with arrow means positive association/effect; Line with —| means inhibition. Ab: antibodies; M: macrophage.

A study including 1,018 complete twin pairs has estimated the heritability for serum level of IgM anti-PC and found that about 40% of its phenotypic variation is explained by genetic effects [71]. However, no studies (before our study III) had identified specific genetic variants associated with IgM anti-PC.

**Thus, our study III addressed two questions about IgM anti-PC:**

— *Which genetic variants are associated with serum level of IgM anti-PC?*

— *What is the association between IgM anti-PC and CLL?*

## 1.2.4 What can we learn from the important genetic factors?

The genetic variants identified from GWASs can be used to increase the understanding about the genetic etiology of human complex traits and molecular mechanisms of diseases [47]. Moreover, thanks to the continuous development of polygenic methods, these genetic variants can now also be used to test the association and causality previously reported from traditional epidemiological studies [24].

### 1.2.4.1 Association tested by polygenic risk score analysis

GWASs and the "missing heritability" phenomenon indicate that most human complex traits are highly polygenic, which means that they are influenced by numerous genetic variants with small effects [24]. *Polygenic risk scores (PRSs)* are calculated in target samples by weighting the risk alleles identified from GWASs. The number of alleles to be used depends on a flexible threshold of GWAS P-value [72]. In recent years, as a complement to traditional biomarkers, PRSs based on risk alleles of biomarkers have been investigated as predictors of diseases [73].

### 1.2.4.2 Causality of the association tested by Mendelian randomization study

Because alleles are randomly assigned during meiosis and generally unchanged throughout human life, they can be used as *instrumental variables (IVs)* to test causality in the *Mendelian randomization (MR) study* [74]. However, the causal inference of MR study is based on three core assumptions: 1) IVs are only specifically associated with the exposure; 2) IVs are independent of any measured or unmeasured confounders; 3) the influences of IVs on the outcome only go through the exposure [75].

### 1.2.4.3 PRS and MR studies between cardiometabolic biomarkers and diseases

Many associations between the risk factors/biomarkers and cardiometabolic diseases (suggested from previous observational studies) have been tested by PRS studies [76]. In the MR studies, LDL and TG are supported to be causal for CAD [77], while HDL is not [78, 79]. It is well in line with the clinical treatment outcomes: statins are still the most effective drug to prevent ASCVD, because they inhibit LDL biosynthesis by blocking the 3-hydroxy-

3-methylglutaryl-coenzyme A reductase [80]; but niacin fails to prevent ASCVD by increasing HDL levels. By using IVs of WHR adjusted for BMI, PRS and MR analyses support that abdominal adiposity is causal for type II diabetes and CAD [81].

#### 1.2.4.4 Controversial associations between blood lipids and ALS

Whether dyslipidemia is a risk or protective factor for *amyotrophic lateral sclerosis (ALS)* has been debated for more than 10 years [82-89], perhaps because less than 500 ALS cases were included in previous observational studies. Currently, the summary statistics from large-scale GWASs on blood lipids (in ~100,000 individuals, [90]) and ALS (including 12,577 ALS cases and 23,475 controls, [91]), are publicly available.

**Thereby, two questions about blood lipids and ALS were addressed in study IV:**

— *Whether polygenic evidence support the association between blood lipids and ALS?*

— *If so, which direction and whether the association is causal?*

# 2 AIMS

**The general aim** of this thesis is to investigate the genetic epidemiology of cardiometabolic biomarkers, by using classical twin studies and current SNP-based genomic methods.

**The specific aim of each study:**

***Study I*** aims to illuminate the role of dominant genetic effects in the “missing heritability”.

***Study II*** aims to quantify the genetic and environmental contributions to the covariation between cardiometabolic traits.

***Study III*** aims to identify the genetic variants associated with serum level of IgM anti-PC.

***Study IV*** aims to test the association and causality between blood lipids and ALS.

# 3 STUDY DESIGN

The study design for each study is summarized in the *Table* and outlined in this chapter. More details about materials and methods used in each study can be found in the published papers I-IV [92-95].

**Table. Study design for each study**

| Study | Materials | Methods |
|---|---|---|
| I | *Same study base:* 10,682 twins in TwinGene (3,870 complete twin pairs; 5,779 unrelated individuals) | Univariate twin-based SEM and SNP-based GREML(d) |
|  | *Phenotypes:* 24 cardiometabolic biomarkers |  |
| II | *Genotypes:* directly genotyped 700K SNPs<br>*Covariates:* age, sex, 10 principle components | Bivariate twin-based SEM and SNP-based GREML(d) |
| III | *Subjects* in four Swedish cohorts (total n=3,648)<br>*Phenotype:* serum level of IgM anti-PC<br>*Genotypes:* ~8 million SNPs after imputation<br>*Covariates:* age, sex, 2-4 principle components | GWAS, Meta-analysis<br>PRS<br>Bioinformatics' prediction<br>Nested-case control study |
| IV | Summarized GWAS results of lipids and ALS: $n_{LDL}$=95,454; $n_{TC}$=100,184; $n_{TG}$=96,598; $n_{HDL}$=99,900; $n_{ALS}$=36,052 (12,577 cases) | PRS<br>MR |

## 3.1 Materials

### 3.1.1 Phenotypes and genotypes in cohorts

**TwinGene from the Swedish Twin Registry**

TwinGene is a Swedish population-based cohort including ~12,000 twins born between 1911 and 1958, the medical records of TwinGene participants are accessed from the national registers in Sweden [96]. Blood samples and health check-up information were collected during 2004-2008, blood biomarkers and genotypes were measured by methods described in our paper I and II [92, 93].

Study I and II used the same materials from TwinGene: 10,682 twins with both genotypes (644,556 directly genotyped autosomal SNPs passed quality control) and 24 traditional cardiometabolic biomarkers available. From the same study base, 3,870 complete twin pairs were used for twin-based SEM and 5,779 unrelated individuals were used for SNP-based GREML(d) .

IgM anti-PC was measured in 1,018 complete twin pairs (2,036 twins) randomly selected from TwinGene to estimate the heritability [71]. After quality control (QC), 1,175 twins with both IgM anti-PC measurements and genotypes (~8 million autosomal SNPs after imputation and QC) were used for GWAS in study III. For the nested case-control study in study III, CLL cases were identified from TwinGene Biobank (serum and DNA samples from ~12,000 twins) by using the International Classification of Diseases (ICD) code (ICD7/8/9: 204.1; ICD10: C91.1). For each CLL case, three age- and sex- matched controls were also randomly selected from the biobank.

**PIVUS**

Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) cohort was established in 2001, including 1,106 seventy-years-old individuals who lived in Uppsala community [97]. IgM anti-PC was measured and genomic DNA was genotyped for all PIVUS participants. After QC and phenotype-genotype matching, 945 individuals were used for IgM anti-PC GWAS in study III.

**MDC**

Malmö Diet and Cancer (MDC) cohort includes ~30,000 individuals living in Malmö city [98, 99]. Within a nested case-control study for CAD, IgM anti-PC was measured in 1,042 individuals [100], from which 882 individuals with both IgM anti-PC and genotypes available were used for GWAS in our study III.

**PRACSIS**

During 1995-2001, Prognosis and Risk in Acute Coronary Syndromes in Sweden (PRACSIS) cohort was established to recruit acute coronary syndromes patients [101]. IgM anti-PC was measured for 1,185 patients and genomic DNA was genotyped for 1,268 patients. Finally, 646 subjects with both IgM anti-PC and genotypes were used for GWAS replication in study III.

### 3.1.2 Summary statistics from GWASs

In study III and IV, summary GWAS results (from the European-ancestry populations) for CLL, immunoglobulins, blood lipids, CAD and ALS were used for PRS or MR analyses.

**CLL and immunoglobulins in study III**

Full summary GWAS results of CLL (from 3,100 unrelated cases and 7,677 controls) were accessed from the InterLymph Consortium [102]. Public GWAS results of immunoglobulins (~19,000 individuals) only are ~5,000 SNPs with association P-value<$1\times10^{-6}$ [103].

**Blood lipids, CAD and ALS in study IV**

Summary GWAS results of blood lipids (TG, TC, LDL and HDL) were based on 2.69 million SNPs among ~100,000 Europeans [90]. Because the association and causality between blood lipids and CAD have been clearly tested, CAD was used as a "reference outcome" in MR study. GWAS results of CAD (including 22,233 cases and 64,762 controls, with 2.42 million SNPs) were accessed from the CARDIoGRAMplusC4D Consortium [51]. Summary statistics for ALS were from the latest GWAS in 2016, including 8.71 million SNPs for 12,577 cases and 23,475 controls [91].

# 3.2 Methods

## 3.2.1 Classical twin design

Based on human consanguinity (degree of kinship or biological relationship), the classical twin study compares the phenotypic similarities between MZ and DZ twins [29]. *Falconer's formula* can roughly quantify the additive genetic effects (A), common/shared (C) and unique/non-shared (E) environmental effects, by using the monozygotic and dizygotic intra-pair correlations (rMZ, rDZ) in the following equations:

① Similarity between MZ twin is due to the shared A (100%) and C (100%), rMZ=A+C;

② Similarity between DZ twin is from the shared A (50%) and C (100%), rDZ=0.5A+C;

③ Dissimilarity between MZ twin is because of non-shared environment, E=1-rMZ.

Therefore, A=2(rMZ-rDZ), and C=2rDZ-rMZ.

Falconer's formula assumes that all genetic effects are additive and that the phenotypic variance is only due to contributions of A, C and E. Thus, the heritability can be simply calculated as A/(A+C+E).

### 3.2.2 Twin-based SEM

Although Falconer's formula provides an easy way to obtain point estimates, more sophisticated model fitting approaches are needed to evaluate statistical significance, to obtain confidence intervals and to test more complex models.

By using the OpenMx package (version 2.8.3) in R (version 3.4.1) [104], the observed variance-covariance matrices were constructed for MZ and DZ pairs. We constructed ACE model, ADE model and AE model for each trait, respectively. The model fitting was evaluated by the Akaike information criterion (AIC), considering the model with the lowest AIC value as best-fitted [105].

### 3.2.3 SNP-based GREML(d)

In the tool of Genome-wide Complex Trait Analysis (GCTA), GREML(d) fit all SNPs as random effects within a mixed linear model, in which the empirical genetic resemblance between "unrelated individuals" (to exclude the shared environmental effects) were compared [28].

In this thesis, the "unrelated individuals" were selected from the same study base in the following steps: 1) one twin within each MZ pair and both twins in DZ pairs were genotyped; 2) one twin within each DZ pair was randomly removed; 3) among the remaining individuals, related individuals were further removed based on the genetic-related-matrix (cut-off value for relatedness was 0.025).

For the 24 traditional cardiometabolic biomarkers in our study I and II, the univariate and bivariate twin-SEM and SNP-GREML(d) were performed and compared within the same study base (10,682 twins from TwinGene), respectively.

### 3.2.4 A direct test for effects from shared environment

The self-reported contact frequency (in four levels: 1-contact less than once per year, 2-yearly contact, 3-monthly contact, 4-weekly contact), and separation age were used to test the existence of shared environmental effects. The t-test on their mean levels between MZ and DZ twins was used to test the EEA (that co-twin within MZ and DZ pairs share environment to the same extent). The potential relation between the degree of shared environment and the intra-pair trait difference was also investigated by estimating their correlation in MZ pairs for each trait.

### 3.2.5 Genome-wide association study

All four cohorts in the IgM anti-PC GWAS used the same analysis procedure, as below:

**Phenotype:** IgM anti-PC raw values were adjusted for age at blood sampling and sex in the linear regression model, outliers [individuals with the residuals beyond ±4 standard deviations (SDs) from the mean] were removed, then residuals were rank order normalized (to achieve standard normal distribution) and used as the phenotype in GWAS.

**Genotype:** directly genotyped SNPs and imputed SNPs by using the 1000 Genome reference panel (GRCh 37/hg 19, Phase 1, Version 3); QC details can be found in Paper III.

**Model:** linear regression model.

**Covariates:** the first 4 genetic principal components (the first 2 genetic principal components in PIVUS). The relatedness in TwinGene participants was handled by using the "--within" option in PLINK.

### 3.2.6 Polygenic risk score analysis

In study III and IV, PRS analyses were performed by using summary GWAS data in the PRSice tool [106]. Independent SNPs were kept in the base data by LD clumping (reference panel: HapMap_ceu_all, release 22), with the following settings: clumping threshold p1=p2= 0.5, LD threshold $r^2$=0.05 and distance threshold=300Kb. Then independent SNPs were grouped into quantiles with gradually increasing P-value threshold ($P_T$). The quantile explaining the largest trait variance in the target sample is denoted the best-fitted, and the corresponding $P_T$ is defined as the best-fitted $P_T$.

### 3.2.7 Mendelian randomization study

In study IV, MR was performed if PRS analyses identified significant polygenic association between blood lipids and ALS. Independent SNPs that were only associated with the exposure (P-value<$5\times10^{-8}$) but not with any other traits (P-value>$5\times10^{-8}$) in the PhenoScanner database were used as IVs [107]. By using "gtx v0.0.8" package in R 3.2.5, the causal effect of the exposure on the outcome was tested by the *inverse-variance weighted method* [74]. CAD was used as a reference outcome.

# 4 RESULTS AND DISCUSSION

The main results and interpretations of study I-IV are briefly presented here, more details and supplemental information can be found in the published paper I-IV [92-95].

## 4.1 Study I

### 4.1.1 Intra-pair correlation and model fitting

The intra-pair correlation coefficients in MZ and DZ (rMZ, rDZ) are plotted in the figure below. For height and apoA1, rMZ<2rDZ, indicating contributions from C; while rMZ>2rDZ for all other 22 biomarkers, which indicates some potential dominant deviations from the pure additive model. From the model fitting according to AIC, ACE was the best-fitted model for height; AE was the best-fitted model for apoA1, HDL, PP, DBP and MAP; while ADE was the best-fitted model for all other biomarkers (*Figure 4.1.1*).

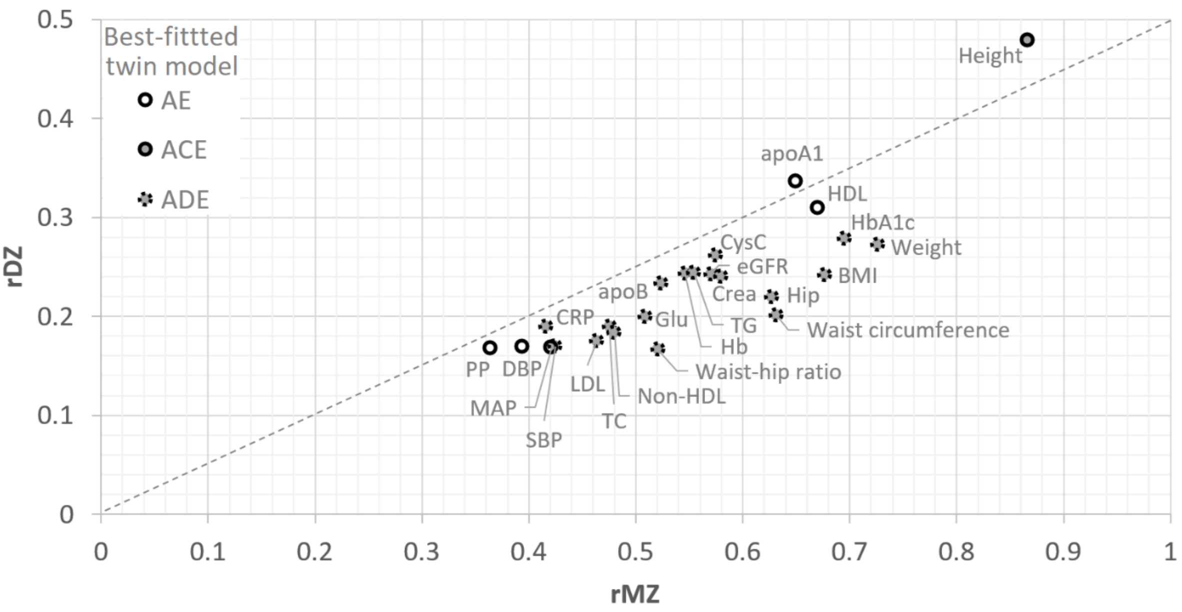


**Figure 4.1.1. Intra-pair correlation and model fitting**

Among the 24 cardiometabolic biomarkers, intra-pair MZ and DZ twin correlations (rMZ, rDZ) indicate dominant deviation from additive model for 22 biomarkers, ADE model is best-fitted for 18 of them.

### 4.1.2 Twin- versus SNP-based univariate heritability

The decomposing of phenotypic variation of each trait by twin and SNP model is presented in *Figure 4.1.2*. Twin-based estimate of shared environmental variance ($c^2$) was 9% for height. The SNP-based estimates of additive genetic variance ($a^2$) for SBP, DBP and MAP were not significant, but all twin- and SNP-based $a^2$ and unique environmental variance ($e^2$) were significantly estimated for other 21 traits. Significant contributions from the dominant genetic effects were identified for 13 traits in twin model, while SNP-based estimates of the dominant genetic variance ($d^2$) were significant just for TG (28%, 95%CI 10%-46%) and waist circumference (19%, 95%CI 1%-37%).

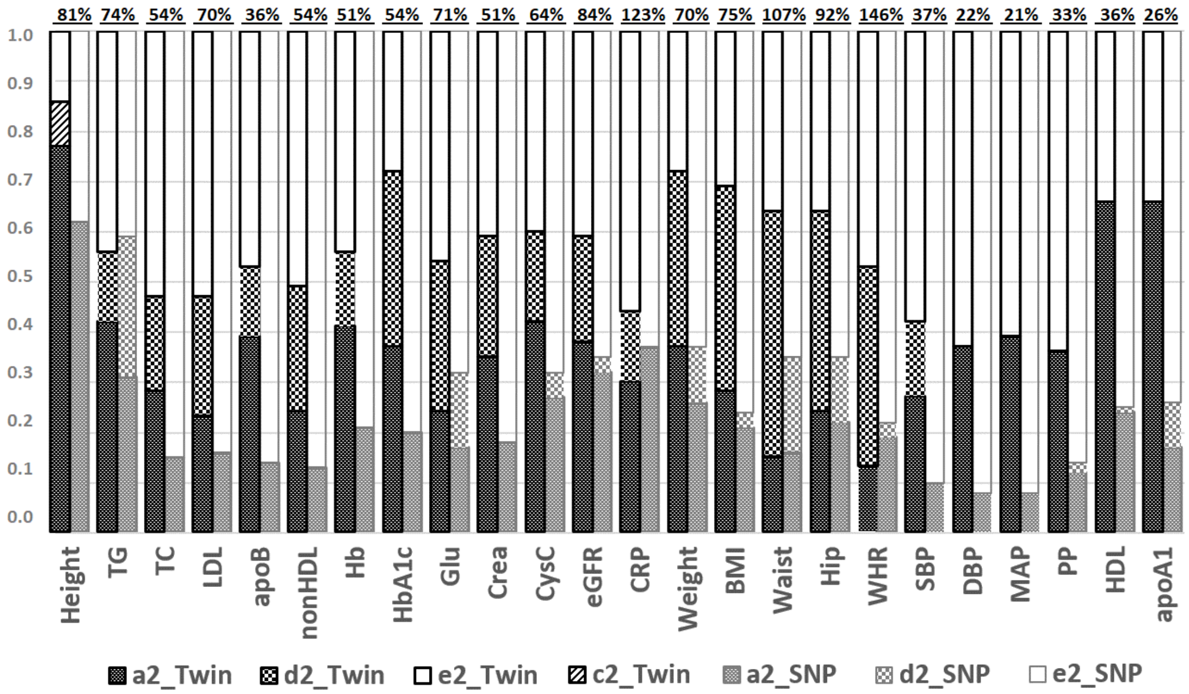


**Figure 4.1.2. Phenotypic variation partitioned by twin and SNP models**

Twin-based SEM identifies significant dominant genetic influences (D) on the phenotypic variation of 13 biomarkers, while SNP-based GREML just identifies significant D for 2 biomarkers. Statistically significant estimates (P-value<0.05) are labeled in solid line, the percentage values on the top of bars represent $h^2_{SNP}/h^2_{Twin}$.

For the 13 traits with significant estimates of $d^2$, the average value of $h^2_{SNP}/h^2_{Twin}$ was 76%; while for the 5 AE best-fitted traits, the average value of $h^2_{SNP}/h^2_{Twin}$ was 28%.

## 4.1.3 Test for shared environment

The mean values of contact frequency (in four levels: 1, 2, 3, 4) and separation age (years spent together in raising household) were significantly higher in MZ pairs (3.03±0.82, 19.80±3.43 years) than SSDZ (2.71±0.82, 18.55±3.59 years) and OSDZ (2.45±0.69, 18.25±3.75 years) pairs, which indicate potential violation of the equal environment assumption for co-twins within MZ and DZ pairs.

However, the correlations between trait difference and shared environment (contact frequency and separation age) were weak within MZ pairs, the absolute values of correlation coefficient were less than 0.1 (*Figure 4.1.3*); from which significant correlations were found between these two shared environmental factors and 8 cardiometabolic traits (all five obesity traits: weight, BMI, waist circumference, hip circumference, WHR; SBP, PP and HDL).

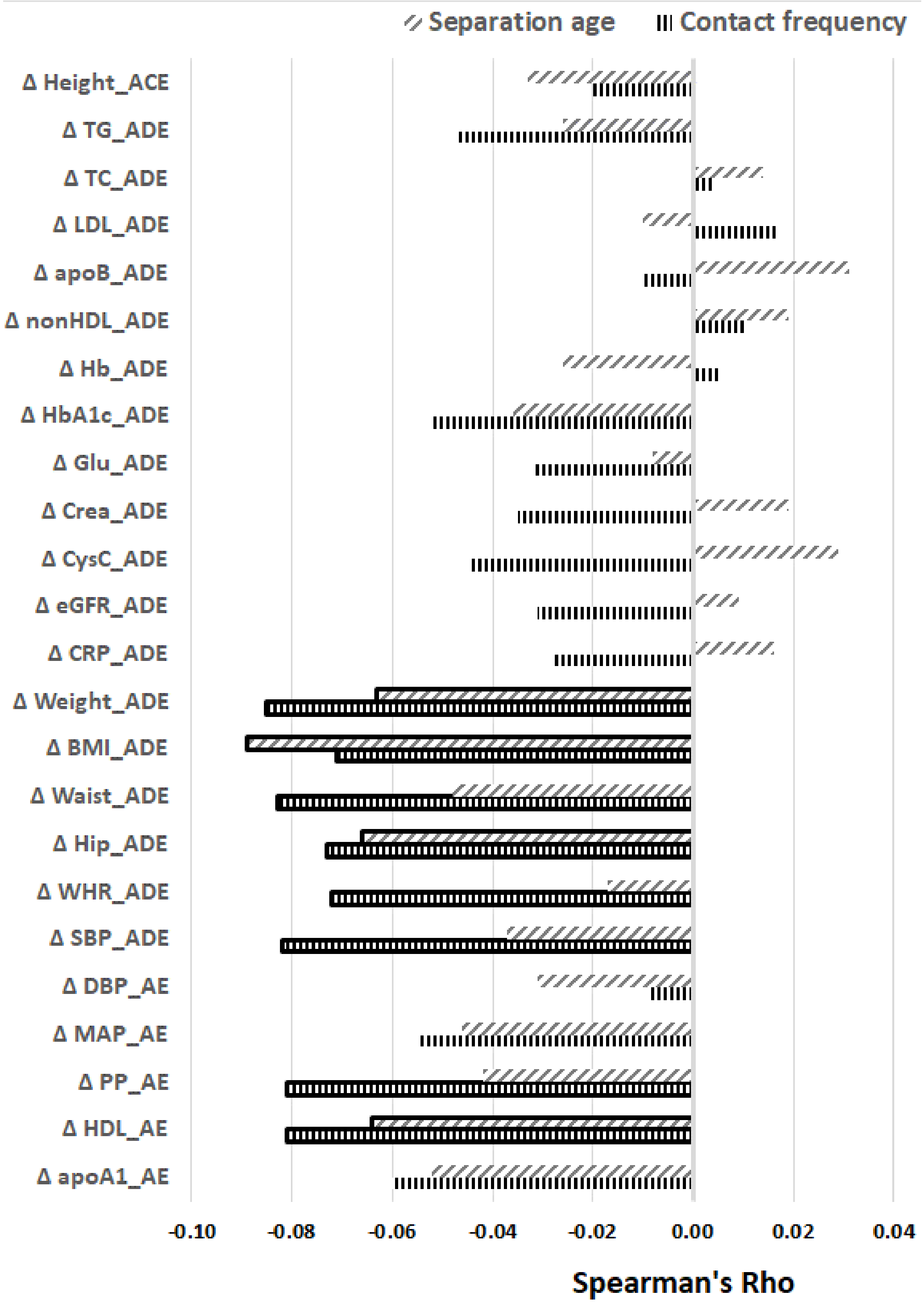


**Figure 4.1.3. Correlations between intra-pair trait difference and degree of shared environment (separation age and contact frequency) in MZ pairs**

Statistically significant estimates (P-value<0.05) are labeled in solid lines. ACE, ADE and AE represent the best-fitted model for each trait in the univariate twin-based structural equation model.

# 4.2 Study II

## 4.2.1 Phenotypic correlations

Among the 276 pairs of correlations between the 24 cardiometabolic biomarkers/traits, 27 pairs with the absolute phenotypic correlation coefficient larger than 0.40 were further investigated in study II. In line with the biological knowledge, the genetic and environmental contributions to their phenotypic covariation can be illuminated in four clusters: blood lipids, metabolic biomarkers, obesity traits and blood pressure (*Figure 4.2*).

## 4.2.2 Covariation decomposition by twin and SNP model

Among the bivariate twin-SEM for these 27 correlated pairs of cardiometabolic traits, the AE model was best-fitted for 7 pairs (TG-HDL, TC-apoB and all 5 pairs in the blood pressure cluster); ACE was the best-fitted model for 4 pairs (HDL-apoA1, LDL-apoB, apoB-nonHDL, CysC-eGFR), but estimates of $c^2$ were close to zero; ADE was the best-fitted model for the remaining 16 pairs, in which significant bivariate $d^2$ were identified for 13 pairs (including all the 9 pairs in the obesity cluster).

The SNP-based estimates of bivariate $a^2$ were non-significant for weight-WHR, BMI-WHR and 4 blood pressure pairs (SBP-DBP, SBP-MAP, DBP-MAP, MAP-PP), and SNP-based estimates of bivariate $d^2$ were neither significant for any pairs.

In general, the SNP-based bivariate $a^2$ (~19% on average) were lower than twin-based bivariate $a^2$ (~36% on average); the SNP- and twin-based estimates of additive genetic correlation (rA) were highly similar (both were 0.67 on average). The estimates of phenotypic correlation (rP) and environmental correlation (rE) showed only small differences between twin and SNP models.

# 4.3 Study III

## 4.3.1 GWAS meta-analysis, PRS and functional prediction

The meta-analysis of three individual discovery GWASs found two SNPs in 1p31.3 and six SNPs in 11q24.1 that achieved the genome-wide significance. The six SNPs close to *GRAMD1B* gene in 11q24.1 were successfully replicated in the fourth cohort. Based on the meta-analysis of four cohorts, rs35923643-G was the top allele, with the combined beta =0.19 rank order normalized SD of IgM anti-PC per allele (P-value=$4.34\times10^{-11}$, *Figure 4.3*).

| Pearson correlation coefficient ( r ) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | -1.0 | -0.8 | -0.6 | -0.4 | -0.2 | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |

**Blood lipids**

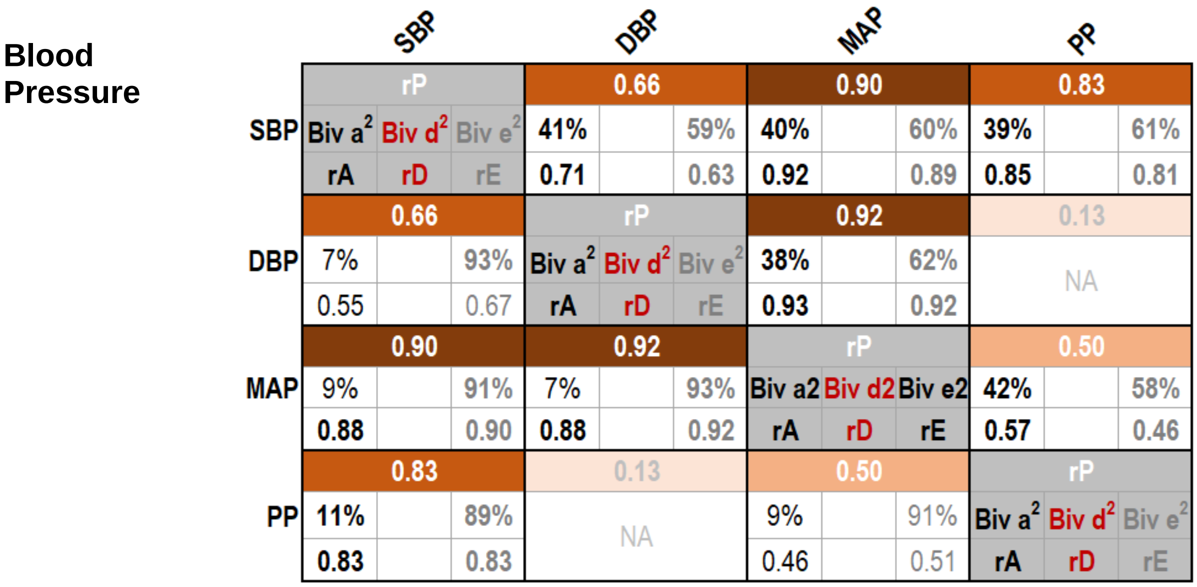
| | TG | | | TC | | | LDL | | | apoB | | | nonHDL | | | HDL | | | apoA1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TG | rP | | | 0.24 | | | 0.15 | | | 0.38 | | | 0.41 | | | -0.46 | | | -0.21 | | |
| | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | NA | | | NA | | | NA | | | 33% | 23% | 44% | 66% | | 34% | NA | | |
| | rA | rD | rE | | | | | | | | | | 0.43 | 0.49 | 0.38 | -0.51 | | -0.39 | | | |
| TC | 0.23 | | | rP | | | 0.94 | | | 0.87 | | | 0.93 | | | 0.27 | | | 0.29 | | |
| | NA | | | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 25% | 21% | 54% | 46% | | 54% | 25% | 22% | 53% | NA | | | NA | | |
| | | | | rA | rD | rE | 0.92 | 0.94 | 0.95 | 0.83 | | 0.91 | 0.88 | -0.96 | 0.96 | | | | | | |
| LDL | 0.14 | | | 0.94 | | | rP | | | 0.90 | | | 0.96 | | | 0.06 | | | 0.06 | | |
| | NA | | | 15% | | 85% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 47% | | 53% | 23% | 25% | 52% | NA | | | NA | | |
| | | | | 0.92 | | 0.94 | rA | rD | rE | 0.91 | | 0.92 | 0.94 | 0.98 | 0.96 | | | | | | |
| apoB | 0.37 | | | 0.87 | | | 0.90 | | | rP | | | 0.94 | | | -0.10 | | | -0.02 | | |
| | NA | | | 14% | | 86% | 14% | | 86% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 47% | | 53% | NA | | | NA | | |
| | | | | 0.82 | | 0.87 | 0.89 | | 0.91 | rA | rD | rE | 0.95 | | 0.95 | | | | | | |
| nonHDL | 0.40 | | | 0.93 | | | 0.96 | | | 0.94 | | | rP | | | -0.08 | | | -0.01 | | |
| | 11% | | 89% | 13% | | 87% | 14% | | 86% | 14% | | 86% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | NA | | | NA | | |
| | 0.23 | | 0.46 | 0.89 | | 0.94 | 0.95 | | 0.96 | 0.96 | | 0.93 | rA | rD | rE | | | | | | |
| HDL | -0.45 | | | 0.27 | | | 0.05 | | | -0.10 | | | -0.09 | | | rP | | | 0.84 | | |
| | 29% | | 71% | NA | | | NA | | | NA | | | NA | | | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 66% | | 34% |
| | -0.48 | | -0.45 | | | | | | | | | | | | | rA | rD | rE | 0.89 | | 0.85 |
| apoA1 | -0.21 | | | 0.28 | | | 0.05 | | | -0.03 | | | -0.02 | | | 0.84 | | | rP | | |
| | NA | | | NA | | | NA | | | NA | | | NA | | | 21% | | 79% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ |
| | | | | | | | | | | | | | | | | 0.88 | | 0.84 | rA | rD | rE |

**Metabolic Biomarkers**

| | HbA1c | | | Glu | | | Crea | | | CysC | | | eGFR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HbA1c | rP | | | 0.51 | | | 0.00 | | | 0.10 | | | -0.09 | | |
| | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 29% | 49% | 22% | NA | | | NA | | | NA | | |
| | rA | rD | rE | 0.51 | 0.76 | 0.32 | | | | | | | | | |
| Glu | 0.50 | | | rP | | | -0.06 | | | 0.03 | | | -0.02 | | |
| | 23% | | 77% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | NA | | | NA | | | NA | | |
| | 0.61 | | 0.48 | rA | rD | rE | | | | | | | | | |
| Crea | 0.01 | | | -0.05 | | | rP | | | 0.56 | | | -0.53 | | |
| | NA | | | NA | | | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 44% | 18% | 38% | 39% | 22% | 39% |
| | | | | | | | rA | rD | rE | 0.65 | 0.48 | 0.53 | -0.57 | 0.54 | -0.50 |
| CysC | 0.09 | | | 0.01 | | | 0.58 | | | rP | | | -0.97 | | |
| | NA | | | NA | | | 22% | | 78% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 58% | | 42% |
| | | | | | | | 0.59 | | 0.59 | rA | rD | rE | -0.98 | | -0.97 |
| eGFR | -0.09 | | | -0.01 | | | -0.55 | | | -0.96 | | | rP | | |
| | NA | | | NA | | | 25% | | 75% | NA | | | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ |
| | | | | | | | -0.57 | | -0.55 | | | | rA | rD | rE |

**Obesity Traits**

| | Weight | | | BMI | | | WC | | | Hip | | | WHR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Weight | rP | | | 0.86 | | | 0.82 | | | 0.83 | | | 0.43 | | |
| | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 25% | 45% | 30% | 25% | 47% | 28% | 33% | 40% | 27% | 14% | 54% | 32% |
| | rA | rD | rE | 0.68 | 1.00 | 0.89 | 0.83 | 0.93 | 0.74 | 0.91 | 0.90 | 0.70 | 0.26 | 0.62 | 0.38 |
| BMI | 0.86 | | | rP | | | 0.81 | | | 0.78 | | | 0.47 | | |
| | 18% | | 82% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 21% | 50% | 29% | 26% | 46% | 28% | 15% | 56% | 29% |
| | 0.66 | | 0.93 | rA | rD | rE | 0.80 | 0.93 | 0.70 | 0.76 | 0.89 | 0.66 | 0.36 | 0.66 | 0.37 |
| WC | 0.82 | | | 0.81 | | | rP | | | 0.77 | | | 0.74 | | |
| | 22% | | 78% | 17% | | 83% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | 23% | 46% | 31% | 12% | 48% | 40% |
| | 0.87 | | 0.82 | 0.76 | | 0.83 | rA | rD | rE | 0.94 | 0.80 | 0.66 | 0.64 | 0.80 | 0.72 |
| Hip | 0.83 | | | 0.78 | | | 0.76 | | | rP | | | 0.16 | | |
| | 28% | | 72% | 21% | | 79% | 22% | | 78% | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ | NA | | |
| | 0.99 | | 0.78 | 0.78 | | 0.78 | 0.90 | | 0.73 | rA | rD | rE | | | |
| WHR | 0.41 | | | 0.46 | | | 0.74 | | | 0.14 | | | rP | | |
| | 7% | | 93% | 10% | | 90% | 10% | | 90% | NA | | | Biv a$^2$ | Biv d$^2$ | Biv e$^2$ |
| | 0.14 | | 0.49 | 0.24 | | 0.52 | 0.46 | | 0.79 | | | | rA | rD | rE |



Blood Pressure

| | SBP | | | DBP | | | MAP | | | PP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | rP | | | 0.66 | | | 0.90 | | | 0.83 | | |
| SBP | Biv $a^2$ | Biv $d^2$ | Biv $e^2$ | **41%** | | 59% | **40%** | | 60% | **39%** | | 61% |
| | rA | rD | rE | **0.71** | | 0.63 | **0.92** | | 0.89 | **0.85** | | 0.81 |
| | 0.66 | | | rP | | | 0.92 | | | 0.13 | | |
| DBP | 7% | | 93% | Biv $a^2$ | Biv $d^2$ | Biv $e^2$ | **38%** | | 62% | NA | | |
| | 0.55 | | 0.67 | rA | rD | rE | **0.93** | | 0.92 | | | |
| | 0.90 | | | 0.92 | | | rP | | | 0.50 | | |
| MAP | 9% | | 91% | 7% | | 93% | Biv a2 | Biv d2 | Biv e2 | **42%** | | 58% |
| | **0.88** | | 0.90 | **0.88** | | 0.92 | rA | rD | rE | **0.57** | | 0.46 |
| | 0.83 | | | 0.13 | | | 0.50 | | | rP | | |
| PP | **11%** | | 89% | NA | | | 9% | | 91% | Biv $a^2$ | Biv $d^2$ | Biv $e^2$ |
| | **0.83** | | 0.83 | | | | 0.46 | | 0.51 | rA | rD | rE |

**Figure 4.2. Genetic and environmental contributions to cardiometabolic pairs**

Bivariate twin-SEM and SNP-GREML(d) are performed for 27 highly correlated pairs (absolute phenotypic correlation coefficient |rP|≥0.4). Twin-based estimates are in the upper triangle, and SNP-based estimates are in the lower triangle. Statistically significant estimates (P-value<0.05) are in bold. NA: not available because of the weak phenotypic correlation.

The SNP rs35923643-G and its proxy variant rs735665-A are also the top risk alleles for CLL. In the PRS analysis, the top variant in 11q24.1 explained the largest variance of CLL (Nagelkerke $r^2$=0.006, P-value=1.2×$10^{-15}$). Based on bioinformatics tools and databases, our functional predictions suggested that rs35923643-G might be the functional variant affecting the transcription factors binding, especially impeding the binding of tumor suppressor RUNX3.

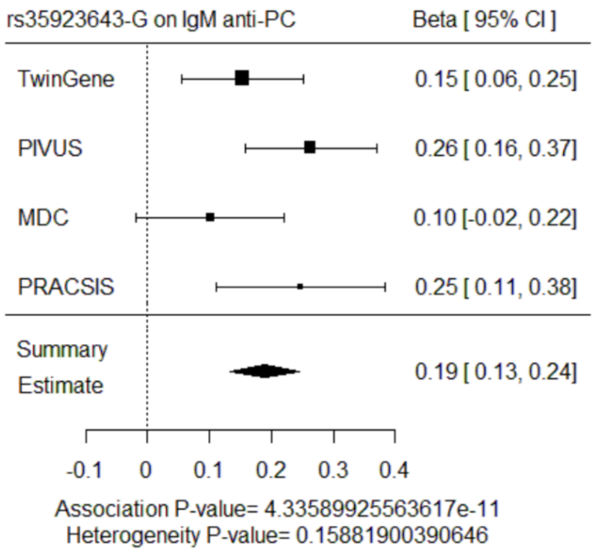


**Figure 4.3. Association of the top allele rs35923643-G with IgM anti-PC in four Swedish cohorts**

### 4.3.2 Nested case-control study

The small nested case-control study found that IgM anti-PC level was significantly lower in 7 prevalent CLL cases than in 21 matched controls (P-value=0.006); IgM anti-PC was also lower in the 23 incident CLL cases than in 69 matched controls, but the difference was not statistically significant (P=0.227). The hazard ratio from the stratified Cox proportional hazards model indicated an inverse association between IgM anti-PC and incident risk of CLL, hazard ratio estimate was 0.75 (95% CI 0.40-1.39) but not significant (P-value=0.354).

## 4.4 Study IV

### 4.4.1 Bi-directional PRS analyses

When using blood lipids as the base and ALS as the target in the PRS analyses, PRSs based on the increasing alleles of LDL or TC ($PRS_{LDL}$ or $PRS_{TC}$) were significantly associated with ALS risk. The estimates and predictions for ALS were very similar between $PRS_{LDL}$ and $PRS_{TC}$, likely reflecting the strong phenotypic correlation between them (LDL is the major type of TC). For the best-fitted $PRS_{LDL}$ and $PRS_{TC}$, the $P_T$ was the same ($=5\times10^{-5}$), and effect sizes were also quite similar (log OR=0.15 for $PRS_{LDL}$ calculated from 233 independent risk alleles; log OR=0.14 for $PRS_{TC}$ calculated from 270 independent risk alleles). However, no significant association with ALS risk was identified for PRSs based on TG increasing alleles or PRSs based on HDL increasing/decreasing alleles.

In the reverse PRS analysis, no significant association was identified between PRSs based on ALS risk alleles and any of the studied lipids. Perhaps because the sample size of ALS GWAS was a bit small (12,577 cases and 23,475 controls) compare with blood lipids (~100,000 individuals).

As a reference comparison, the PRSs based on large-scale CAD GWAS (22,233 cases and 64,762 controls) was significantly but also weakly associated with blood lipids ($|\log OR| \leq 0.01$, P-value$<2\times10^{-25}$).

### 4.4.2 MR study

The association between LDL, TC and ALS was suggested to be causal ($\beta$=0.23, P=0.03), by using 13 independent SNPs that are specially associated with both LDL and TC (but not associated with any other traits) as instrumental variables in the MR study.

# 5 STRENGTHS AND LIMITATIONS

**Study I and II** have the possibility to compare the twin- and SNP-based estimates within the same study base. This provides a straightforward way to control for population differences (extra variances or "noises") arising from age, sex, ethnicity, life-style and other factors. TwinGene is a population-based cohort of the Swedish Twin Registry, in which elderly Swedish born twins living all over Sweden were invited without selections besides willingness to participate. Thus, the geographic and demographic distribution provides a homogenous genetic background of the sample, which is a valuable feature in genetic studies. All the blood samples are collected, extracted, and stored by the same biobank using the same procedures. The same laboratory, using the same methods, measured all the clinical biomarkers and measurements in the same procedure. These features are of vital importance in order to diminish the risk of biases due to batch effects.

**Study III** is the first GWAS for IgM anti-PC and it is also the first study to investigate the shared genetics and phenotypic relationship between IgM anti-PC and CLL. All individuals from the four cohorts used in study III were European-ancestry and born in Sweden, providing low heterogeneity (population or genetic stratification). The TwinGene cohort is also linked to several national health registers in Sweden, which enabled us to identify diseases (e.g. CLL) and test the associations with many biomarkers/factors.

**Study IV** is the first polygenic analysis between blood lipids and ALS, which also provides polygenic evidence to support the causal effects of LDL and TC on ALS risk.

However, there are also some limitations needed to be noted:

***Sample size*** is more likely a limitation than strength from the overview of this thesis.

— Although 3,870 complete twin pairs were used in our study I and II, which are larger than the average sample size of previous twin studies (≤2,104 pairs per study in the past fifty years [30]), the power to identify significant bivariate estimates in study II is still not adequate. Such situation is also the same for SNP-based GREML, 5,779 unrelated individuals are too small to get enough power for estimating additive genetic variance for SBP, DBP and MAP (and even harder for estimating dominant genetic variance) from genome-wide common SNPs.

— In study III, sample size of IgM anti-PC GWAS is also small (3,002 individuals in the discovery phase and 646 individuals for replication), thus only one locus 11q24.1 is successfully identified. In the PRS analysis between anti-PC and CLL, the quantile including the single top variant explains larger (actually the largest) variance of CLL than all other quantiles including more SNPs across the genome, perhaps also because of the small base data (IgM anti-PC GWAS).

— Similarly, the nested case-control study in study III is also small (7 prevalent and 23 incident CLL cases were identified among all ~12,000 TwinGene participants). There is also a lack of detailed information about the CLL stage, therapy and stereotyped B-cell status for the CLL cases; further impeding the possibility to draw firm conclusion about the association between IgM anti-PC and CLL.

— In the reverse PRS analyses in study IV, PRSs based on ALS risk alleles are not significantly associated with any studied lipids (and explain 0% of the variance), one potential explanation is also the relatively small base data (ALS GWAS).

***Generalizability*** is also a potential concern for study I and II. Because heritability is population specific, the reported contribution for cardiometabolic traits might only represent the baseline measurements in TwinGene (old Swedish twins born between 1911 and 1958). Since the relative importance of genes and environment might vary by different factors like age at blood sampling/measuring, more efforts are needed to further assess the generalizability of our conclusions.

# 6 ETHICAL CONSIDERATIONS

The first three studies in this thesis were mainly performed in TwinGene cohort from the Swedish Twin Registry. Ethical permits for TwinGene project had been approved in 2007 (Dnr: 2007/644-3) and amended in 2012 (Dnr: 2012/257-32). Study IV is based on the public summary GWAS results from consortia, so ethical permit is not required.

**Privacy**

By using questionnaires and interviews, the participants were asked to answer questions about certain personal information. They have also received some physical examinations, donated their blood for research purposes to biobank. Biomarkers in their serum/plasma have been measured; genomic DNA which carries all the genetic information has been isolated and genotyped. From certain analyses, we can identify participants that carry genotypes associated with increased risk of certain diseases. That means the researchers can even know more private information of the participant than themselves. How to handle these informative and sensitive data is also a crucial aspect we should consider.

**Right to know**

Since most of our studies can get certain predictive information, telling such information to the participants might bring both harmful and beneficial effects, which also can raise ethical dilemmas. There are three examples: 1) the zygosity was previously identified by the answers from the questionnaires, then a more accurate genotype-based method was used. Although these two methods were mostly consistent (95-98% match), some twins have their zygosity re-classified; 2) genotypes can be used to estimate polygenic risk for several complex traits/diseases. Such estimate has a market value and is offered by commercial companies. However, the clinical value is very limited and it is very difficult to communicate such information in an adequate way; 3) we might occasionally observe participants who carry risk alleles for certain diseases (e.g. breast or ovarian cancers). Should we inform and suggest them to adopt some prevention strategies?

**Causality**

As mentioned in the third point above, genetic variants are only associated with higher risk of disease, but mostly we do not have enough evidence to declare the causality between them. This is also a problem raised in our study III: because IgM anti-PC has been regarded as an atheroprotective biomarker, monoclonal antibodies targeting PC have been produced and planned to be evaluated as therapy to decrease CVD risk. Our GWAS results indicated that the haplotype which were significantly associated with increased anti-PC is also strongly associated with CLL risk. If this genetic sharing would reflect a causal relation between anti-PC and CLL, the suggested therapeutic treatment by increasing anti-PC levels might come with harmful side-effects. However, our nested case-control study gave no support for a positive association between anti-PC and CLL. Instead a weak negative association was detected.

Furthermore, associations between anti-PC and CLL may be due to confounding. One possible confounder is *S. pneumoniae* infection, which can increase anti-PC level due to the human innate immune response to the PC antigen on *S. pneumoniae*; pneumonia is also associated with increased risk of CLL [108]. If association is due to confounding only, artificially increasing anti-PC would not affect the CLL risk.

Nevertheless, it is very important to further investigate the causality among anti-PC, CVD and CLL, to provide useful scientific evidence for the clinical translation of IgM anti-PC.

# 7 CONCLUSIONS

**Study I** finds that significant contribution of dominant genetic effects (D) to the variation of most cardiometabolic biomarkers in TwinGene samples; it also indicates that the missing heritability ($1\text{-}h^2_{\text{SNP}}/h^2_{\text{twin}}$) becomes smaller when the twin model has enough power to distinguish true D from additive genetic effects (A).

**Study II** suggests that D also contribute to the covariation between certain blood lipids, metabolic biomarkers and all obesity traits in TwinGene samples.

**Study III** identifies that SNP rs35923643-G is the top genetic variant shared between IgM anti-PC and CLL risk, and it is also the potential functional variant affecting the binding of transcription factors.

**Study IV** provides polygenic evidence to support the positive associations between LDL, TC and ALS risk, such positive associations are also suggested to be causal based on current assumptions and evidence.

# 8 FUTURE PERSPECTIVES

For future studies to quantify the genetic and environmental effects on human complex traits, the following aspects are worth to be considered:

1) **Larger sample size** is quite vital for both twin and SNP model to get enough power to quantify and also distinguish the additive and non-additive genetic effects;
2) **Extended model** including more family members/relationships enables estimation of more components within the same model;
3) **Multivariate model** can be considered to capture the common A, C, D, E components shared among more than two correlated traits within clusters (like LDL-TC-apoB-nonHDL, weight-BMI-waist circumference-hip circumference-WHR);
4) **Longitudinal model** is useful to investigate the continuous changes of A, C, D, E contributions across the lifespan;
5) **More types of genetic variants as well as gene-environment interactions and correlations** can also be considered in future molecular/genomic methods.

**More samples** are needed for IgM anti-PC GWAS to identify more associated genetic variants. In order to further illustrate the relationship between IgM anti-PC and CLL, the nested case-control study also need more CLL cases and longer follow-up time. **Confounders** (e.g. Pneumonia or other infections) between IgM anti-PC and CLL are worth to be investigated. **More functional experiments** following the GWAS findings also need to be done, such as allele specific chromatin immunoprecipitation-sequencing and *in silico* approaches to validate the binding affinity of transcription factors.

Similarly, **larger samples** of GWAS on common SNPs are also critical for the MR causal inference between cardiometabolic biomarkers and ALS. Besides inverse-variance weighted method, **other types of MR methods** (e.g. MR-Egger) can be used to test or validate the causality between blood lipids and ALS. Since ALS is a rare disease, **deeper sequencing for rare variants** might also be essential for "missing heritability" and genetic mechanisms between dyslipidemia and ALS. In parallel, **functional experiments** are also needed to validate the pathogenic mechanisms related to the pathways suggested by polygenic evidence.

# 9 ACKNOWLEGEMENTS

Five years ago, when I was doing moleaular cloning experiments in Beijing Hospital, two questions came to my mind: Why do we repeatedly focus on genes? Whether genes are more important than enviroment? After reading some literatures, I found twin study is the natrual design to answer my questions. Then I tried to do my PhD in twin studies, finally I came to Sweden. Looking back to my 56 months in Sweden, I still believe this quick "match/deal" (within 2 weeks since I knew Karolincka) is the most correct/lucky choice/decision in the past 30 years. Because I met many nice persons and spent most happy time here. This acknowledgement will be short, not only because the name list is too long to write, but also because any beautiful words cannot express my gratitude to you all.

Patrik Magnusson, my main supervisor. Thanks for your comprehensive cares along this whole journey, not only from the academics (you always tried your best to help improve the studies, languages and settle my salaries), but also from the daily life (like helping repair my bike, staying with me in the emergency room when I got food poisoning). I also want to express my deep appreciations to your core and big families (I have met 12 of them at least), they are so nice and even learning to cook rice inorder to welcome my family in your house.

Nancy Pedersen, my co-supervisor, was also the first Swedish PI I contacted in 2013. Thanks for your generous financial support and role model as a successful but also humble scientist.

Sara Hägg, my co-supervisor. Thankss for your guidence and help from differenent ways. I am very appreciated to your attitude about equality to the scientists and your students.

Johan Frostegård, my co-supervisor, Thanks for your discoveries about anti-PC. I believe it is an useful biomarker and the anti-PC GWAS is the most interecting study in my thesis.

Per Svensson, my co-supervisor. Thanks for your help in collaboration, student supervision and finance. I am on the way to be a physician-scientist-teacher/supervisor like you.

Paul Lichtenstein, former head of MEB and the second Swedish PI I contacted in 2013. Thanks for your quick recommendation to Patrik and financial support. Fang Fang, you are a so nice friend for everyone. Thanks for your brilliant ideas and financial support. Olof Nyrén, it's my great honor to have you as my external mentor. We rarely met because it's lucky that no any conflicts occured between my supervisors and me. ☺

All the collaborators, thanks for your generous contributions and help for the studies. All members in the aging-molecular epidemiology group and twin group thanks for your enthusiasm, optimism and kindness shared within these two big families of academics.

All Chinese and other friends I met around the world since my first time to travel abroad in 2012, thanks for the time for eating, drinking, travelling, having fun and also complaining.

All colleagues, ITs and TAs in MEB. thanks for making such a wonderful department.

All previous teachers, supervisors, classmates and friends, thanks for your encouragement.

My dear parents and relatives, thanks for your constant love and understanding.

I also want to thank all participants and staffs in the cohorts used in this thesis, as well as financial support from the Swedish Research Council (grant no. 2015-03170 and 2017-00641); the Swedish Heart-Lung Foundation (grant no. 20070481) and the China Scholarship Council (grant no. 201306210065).

# 博士感懷

## 打油詩

丙戌複印紅寶書，開啟留學寄託路。
北京向北至斯京，北緯六十差一度。
一謝吾師帕特克，亦師亦友悉照顧。
二謝副導師南希，雙生研究老師傳。
三謝副導師薩拉，年輕獨美勤幫助。
四謝副導師約翰，磷酸臘驗抗體父。
五謝副導師派得，臨床科研均兼顧。
六謝系主任保羅，當初郵件秒回復。
七謝各位合作著，慷慨傳授肯幫助。
八謝中國小夥伴，人員眾多難細數。
九謝全系同事們，兩三百人士氣鼓。
十謝既往師與長，激勵陪伴求學路。
雖說科學無國界，科研人員有故土。
打鐵還需自身硬，願其國強民也富。
自打出生到如今，萬事攜心是父母。
寒窗苦讀廿四年，而立終把博士東。

不料癸巳二月初，兩週決定來瑞土。
回首四年攻博路，甜辣勝過酸鹹苦。
日常接觸如親屬，願其全家永幸福。
當初首位聊摯汝，最後仍是大金主。
衰老表觀女英雄，碩果結滿科研路。
論文實驗多勞心，科研小說兩不誤。
悉心指導勤合作，鄭伴東回不言苦。
伯榮之恩記心著，同謝方芳烏洛夫。
基因關聯結構樸，孿生環境遺傳度。
異國他鄉同胞情，吃喝玩樂也許苦。
精彩文章常常有，工作環境甚幸福。
感謝話兒說不完，尤其祖國和親屬。
異國求學四年多，國際形勢常反复。
無論漂泊在何處，黑髮紅心黃皮膚。
遠隔萬裡少照顧，遊子最念眾親屬。
醫道求索十二載，心懷感恩續征途。

戊戌年小滿九日於瑞典斯德哥爾摩卡羅林斯卡醫學院

# 10 REFERENCES

1. World Health Organization. *The top 10 causes of death*. 2017.

2. Mendis, S., P. Puska, and B. Norrving, *Global Atlas on cardiovascular disease prevention and control*. 2011, Geneva: World Health Organization.

3. Ndisang, J.F. and S. Rastogi, *Cardiometabolic diseases and related complications: current status and future perspective.* Biomed Res Int, 2013.

4. Global Burden of Metabolic Risk Factors for Chronic Diseases Collaboration, *Cardiovascular disease, chronic kidney disease, and diabetes mortality burden of cardiometabolic risk factors from 1980 to 2010: a comparative risk assessment.* Lancet Diabetes Endocrinol, 2014. **2**(8): p. 634-47.

5. Vasan, R.S., *Biomarkers of cardiovascular disease: molecular basis and practical considerations.* Circulation, 2006. **113**(19): p. 2335-62.

6. Cannon, C.P., *Cardiovascular disease and modifiable cardiometabolic risk factors.* Clin Cornerstone, 2008. **9**(2): p. 24-38; discussion 39-41.

7. Perk, J., et al., *European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts).* Eur Heart J, 2012. **33**(13): p. 1635-701.

8. Jensen, M.K., et al., *Novel metabolic biomarkers of cardiovascular disease.* Nat Rev Endocrinol, 2014. **10**(11): p. 659-72.

9. Roberts, L.D. and R.E. Gerszten, *Toward new biomarkers of cardiometabolic diseases.* Cell Metab, 2013. **18**(1): p. 43-50.

10. Yusuf, S., et al., *Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study.* Lancet, 2004. **364**(9438): p. 937-52.

11. Nordestgaard, B.G., et al., *Lipoprotein(a) as a cardiovascular risk factor: current status.* Eur Heart J, 2010. **31**(23): p. 2844-53.

12. Hansson, G.K., *Inflammation, atherosclerosis, and coronary artery disease.* N Engl J Med, 2005. **352**(16): p. 1685-95.

13. Hansson, G.K. and A. Hermansson, *The immune system in atherosclerosis.* Nat Immunol, 2011. **12**(3): p. 204-12.

14. Zalewski, A. and C. Macphee, *Role of lipoprotein-associated phospholipase A2 in atherosclerosis: biology, epidemiology, and possible therapeutic target.* Arterioscler Thromb Vasc Biol, 2005. **25**(5): p. 923-31.

15. Lp-PLA Studies Collaboration. *Lipoprotein-associated phospholipase A(2) and risk of coronary disease, stroke, and mortality: collaborative analysis of 32 prospective studies.* Lancet, 2010. **375**(9725): p. 1536-44.

16. Alberti, K.G., et al., *The metabolic syndrome--a new worldwide definition.* Lancet, 2005. **366**(9491): p. 1059-62.

17. Calabro, P. and E.T. Yeh, *Intra-abdominal adiposity, inflammation, and cardiovascular risk: new insight into global cardiometabolic risk.* Curr Hypertens Rep, 2008. **10**(1): p. 32-8.

18. Pai, J.K., et al., *Hemoglobin a1c is associated with increased risk of incident coronary heart disease among apparently healthy, nondiabetic men and women.* J Am Heart Assoc, 2013. **2**(2): p. e000077.

19. Whelton, P.K., et al., *2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines.* Hypertension, 2017.

20. Splaver, A., G.A. Lamas, and C.H. Hennekens, *Homocysteine and cardiovascular disease: biological mechanisms, observational epidemiology, and the need for randomized trials.* Am Heart J, 2004. **148**(1): p. 34-40.

21. Shlipak, M.G., et al., *Cystatin C and the risk of death and cardiovascular events among elderly persons.* N Engl J Med, 2005. **352**(20): p. 2049-60.

22. Burton, P.R., M.D. Tobin, and J.L. Hopper, *Key concepts in genetic epidemiology.* Lancet, 2005. **366**(9489): p. 941-51.

23. Benke, K.S. and M.D. Fallin, *Methods: genetic epidemiology.* Clin Lab Med, 2010. **30**(4): p. 795-814.

24. Dudbridge, F., *Polygenic Epidemiology.* Genet Epidemiol, 2016. **40**(4): p. 268-72.

25. Tenesa, A. and C.S. Haley, *The heritability of human disease: estimation, uses and abuses.* Nat Rev Genet, 2013. **14**(2): p. 139-49.

26. Neale, M.C. and H.H.M. Maes, *Methodology for Genetic Studies of Twins and Families*. 2004, Dordrecht, Netherlands: Kluwer Academic Publishers B.V.

27. Visscher, P.M., et al., *Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings.* PLoS Genet, 2006. **2**(3): p. e41.

28. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.

29. Boomsma, D., A. Busjahn, and L. Peltonen, *Classical twin studies and beyond.* Nat Rev Genet, 2002. **3**(11): p. 872-82.

30. Polderman, T.J., et al., *Meta-analysis of the heritability of human traits based on fifty years of twin studies.* Nat Genet, 2015. **47**(7): p. 702-9.

31. van Dongen, J., et al., *The continuing value of twin studies in the omics era.* Nat Rev Genet, 2012. **13**(9): p. 640-53.

32. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nat Genet, 2015. **47**(3): p. 291-5.

33. Maher, B., *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.

34. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.

35. Kaprio, J., *Twins and the mystery of missing heritability: the contribution of gene-environment interactions.* J Intern Med, 2012. **272**(5): p. 440-8.

36. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proc Natl Acad Sci U S A, 2012. **109**(4): p. 1193-8.

37. Zuk, O., et al., *Searching for missing heritability: designing rare variant association studies.* Proc Natl Acad Sci U S A, 2014. **111**(4): p. E455-64.

38. Ritchie, M.D., *Finding the epistasis needles in the genome-wide haystack.* Methods Mol Biol, 2015. **1253**: p. 19-33.

39. Omholt, S.W., et al., *Gene regulatory networks generating the phenomena of additivity, dominance and epistasis.* Genetics, 2000. **155**(2): p. 969-80.

40. Zhu, Z., et al., *Dominance genetic variation contributes little to the missing heritability for human complex traits.* Am J Hum Genet, 2015. **96**(3): p. 377-85.

41. Rahman, I., et al., *Genetic dominance influences blood biomarker levels in a sample of 12,000 Swedish elderly twins.* Twin Res Hum Genet, 2009. **12**(3): p. 286-94.

42. van Dongen, J., et al., *Heritability of metabolic syndrome traits in a large population-based sample.* J Lipid Res, 2013. **54**(10): p. 2914-23.

43. Keller, M.C., et al., *Widespread evidence for non-additive genetic variation in Cloninger's and Eysenck's personality dimensions using a twin plus sibling design.* Behav Genet, 2005. **35**(6): p. 707-21.

44. Lee, S.H., et al., *Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood.* Bioinformatics, 2012. **28**(19): p. 2540-2.

45. Yang, J., et al., *Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index.* Nat Genet, 2015. **47**(10): p. 1114-20.

46. Elks, C.E., et al., *Variability in the heritability of body mass index: a systematic review and meta-regression.* Front Endocrinol (Lausanne), 2012. **3**: p. 29.

47. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.

48. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.

49. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.

50. Global Lipids Genetics Consortium. *Discovery and refinement of loci associated with lipid levels.* Nat Genet, 2013. **45**(11): p. 1274-83.

51. Schunkert, H., et al., *Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease.* Nat Genet, 2011. **43**(4): p. 333-8.

52. van der Harst, P. and N. Verweij, *Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease.* Circ Res, 2018. **122**(3): p. 433-443.

53. Dehghan, A., et al., *Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels.* Circulation, 2011. **123**(7): p. 731-8.

54. Sabater-Lleal, M., et al., *Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease.* Circulation, 2013. **128**(12): p. 1310-24.

55. de Vries, P.S., et al., *A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration.* Hum Mol Genet, 2016. **25**(2): p. 358-70.

56. Suchindran, S., et al., *Genome-wide association study of Lp-PLA(2) activity and mass in the Framingham Heart Study.* PLoS Genet, 2010. **6**(4): p. e1000928.

57. Chu, A.Y., et al., *Genome-wide association study evaluating lipoprotein-associated phospholipase A2 mass and activity at baseline and after rosuvastatin therapy.* Circ Cardiovasc Genet, 2012. **5**(6): p. 676-85.

58. Grallert, H., et al., *Eight genetic loci associated with variation in lipoprotein-associated phospholipase A2 mass and activity and coronary heart disease: meta-analysis of genome-wide association studies from five community-based studies.* Eur Heart J, 2012. **33**(2): p. 238-51.

59. Wheeler, E., et al., *Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis.* PLoS Med, 2017. **14**(9): p. e1002383.

60. Manning, A.K., et al., *A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance.* Nat Genet, 2012. **44**(6): p. 659-69.

61. Plagnol, V., et al., *Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases.* PLoS Genet, 2011. **7**(8): p. e1002216.

62. Bonas-Guarch, S., et al., *Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes.* Nat Commun, 2018. **9**(1): p. 321.

63. van Meurs, J.B., et al., *Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease.* Am J Clin Nutr, 2013. **98**(3): p. 668-76.

64. Kottgen, A., et al., *New loci associated with kidney function and chronic kidney disease.* Nat Genet, 2010. **42**(5): p. 376-84.

65. Sung, Y.J., et al., *A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure.* Am J Hum Genet, 2018. **102**(3): p. 375-400.

66. Yengo, L., et al., *Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry.* bioRxiv, 2018.

67. Binder, C.J., et al., *Innate and acquired immunity in atherogenesis.* Nat Med, 2002. **8**(11): p. 1218-26.

68. Frostegard, J., *Immunity, atherosclerosis and cardiovascular disease.* BMC Med, 2013. **11**: p. 117.

69. Silverman, G.J., *Protective natural autoantibodies to apoptotic cells: evidence of convergent selection of recurrent innate-like clones.* Ann N Y Acad Sci, 2015. **1362**: p. 164-75.

70. Rahman, M., et al., *IgM antibodies against malondialdehyde and phosphorylcholine are together strong protection markers for atherosclerosis in systemic lupus erythematosus: Regulation and underlying mechanisms.* Clin Immunol, 2016. **166-167**: p. 27-37.

71. Rahman, I., et al., *Genetic and environmental regulation of inflammatory CVD biomarkers Lp-PLA2 and IgM anti-PC.* Atherosclerosis, 2011. **218**(1): p. 117-22.

72. Lewis, C.M. and E. Vassos, *Prospects for using risk scores in polygenic medicine.* Genome Med, 2017. **9**(1): p. 96.

73. Smith, J.A., et al., *Current Applications of Genetic Risk Scores to Cardiovascular Outcomes and Subclinical Phenotypes.* Curr Epidemiol Rep, 2015. **2**(3): p. 180-190.

74. Lawlor, D.A., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Stat Med, 2008. **27**(8): p. 1133-63.

75. Burgess, S., et al., *Mendelian randomization: where are we now and where are we going?* Int J Epidemiol, 2015. **44**(2): p. 379-88.

76. Holmes, M.V., M. Ala-Korpela, and G.D. Smith, *Mendelian randomization in cardiometabolic disease: challenges in evaluating causality.* Nat Rev Cardiol, 2017. **14**(10): p. 577-590.

77. Holmes, M.V., et al., *Mendelian randomization of blood lipids for coronary heart disease.* Eur Heart J, 2015. **36**(9): p. 539-50.

78. Voight, B.F., et al., *Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study.* Lancet, 2012. **380**(9841): p. 572-80.

79. Burgess, S. and E. Harshfield, *Mendelian randomization to assess causal effects of blood lipids on coronary heart disease: lessons from the past and applications to the future.* Curr Opin Endocrinol Diabetes Obes, 2016. **23**(2): p. 124-30.

80. Istvan, E.S. and J. Deisenhofer, *Structural mechanism for statin inhibition of HMG-CoA reductase.* Science, 2001. **292**(5519): p. 1160-4.

81. Emdin, C.A., et al., *Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease.* JAMA, 2017. **317**(6): p. 626-634.

82. Albert, S.M., *Dyslipidemia in ALS: good, bad, or unclear?* Neurology, 2008. **70**(13): p. 988-9.

83. Chio, A., et al., *Lower serum lipid levels are related to respiratory impairment in patients with ALS.* Neurology, 2009. **73**(20): p. 1681-5.

84. Dorst, J., et al., *Patients with elevated triglyceride and cholesterol serum levels have a prolonged survival in amyotrophic lateral sclerosis.* J Neurol, 2011. **258**(4): p. 613-7.

85. Dupuis, L., et al., *Dyslipidemia is a protective factor in amyotrophic lateral sclerosis.* Neurology, 2008. **70**(13): p. 1004-9.

86. Ikeda, K., et al., *Relationships between disease progression and serum levels of lipid, urate, creatinine and ferritin in Japanese patients with amyotrophic lateral sclerosis: a cross-sectional study.* Intern Med, 2012. **51**(12): p. 1501-8.

87. Sutedja, N.A., et al., *Beneficial vascular risk profile is associated with amyotrophic lateral sclerosis.* J Neurol Neurosurg Psychiatry, 2011. **82**(6): p. 638-42.

88. Yang, J.W., et al., *Hypolipidemia in patients with amyotrophic lateral sclerosis: a possible gender difference?* J Clin Neurol, 2013. **9**(2): p. 125-9.

89. Mariosa, D., et al., *Blood biomarkers of carbohydrate, lipid, and apolipoprotein metabolisms and risk of amyotrophic lateral sclerosis: A more than 20-year follow-up of the Swedish AMORIS cohort.* Ann Neurol, 2017. **81**(5): p. 718-728.

90. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.

91. van Rheenen, W., et al., *Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis.* Nat Genet, 2016. **48**(9): p. 1043-8.

92. Chen, X., et al., *Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models.* Am J Hum Genet, 2015. **97**(5): p. 708-14.

93. Chen, X., et al., *Genetic and Environmental Contributions to the Covariation Between Cardiometabolic Traits.* J Am Heart Assoc, 2018. **7**(9): p. e007806.

94. Chen, X., et al., *A genome-wide association study of IgM antibody against phosphorylcholine: shared genetics and phenotypic relationship to chronic lymphocytic leukemia.* Hum Mol Genet, 2018. **27**(10): p. 1809-1818.

95. Chen, X., et al., *Polygenic link between blood lipids and amyotrophic lateral sclerosis.* Neurobiol Aging, 2018. 67: p. 202.e1-e6.

96. Magnusson, P.K., et al., *The Swedish Twin Registry: establishment of a biobank and other recent developments.* Twin Res Hum Genet, 2013. **16**(1): p. 317-29.

97. Lind, L., et al., *A comparison of three different methods to evaluate endothelium-dependent vasodilation in the elderly: the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) study.* Arterioscler Thromb Vasc Biol, 2005. **25**(11): p. 2368-75.

98. Manjer, J., et al., *The Malmo Diet and Cancer Study: representativity, cancer incidence and mortality in participants and non-participants.* Eur J Cancer Prev, 2001. **10**(6): p. 489-99.

99. Hedblad, B., et al., *Relation between insulin resistance and carotid intima-media thickness and stenosis in non-diabetic subjects. Results from a cross-sectional study in Malmo, Sweden.* Diabet Med, 2000. **17**(4): p. 299-307.

100. Sjoberg, B.G., et al., *Low levels of IgM antibodies against phosphorylcholine-A potential risk marker for ischemic stroke in men.* Atherosclerosis, 2009. **203**(2): p. 528-32.

101. Caidahl, K., et al., *IgM-phosphorylcholine autoantibodies and outcome in acute coronary syndromes.* Int J Cardiol, 2013. **167**(2): p. 464-9.

102. Berndt, S.I., et al., *Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia.* Nat Commun, 2016. **7**: p. 10933.

103. Jonsson, S., et al., *Identification of sequence variants influencing immunoglobulin levels.* Nat Genet, 2017. **49**(8): p. 1182-1191.

104. Neale, M.C., et al., *OpenMx 2.0: Extended Structural Equation and Statistical Modeling.* Psychometrika, 2016. **81**(2): p. 535-49.

105. deLeeuw, J., *Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle*. 1992, New York: Springer. 11.

106. Euesden, J., C.M. Lewis, and P.F. O'Reilly, *PRSice: Polygenic Risk Score software.* Bioinformatics, 2015. **31**(9): p. 1466-8.

107. Staley, J.R., et al., *PhenoScanner: a database of human genotype-phenotype associations.* Bioinformatics, 2016. **32**(20): p. 3207-3209.

108. Landgren, O., et al., *Respiratory tract infections and subsequent risk of chronic lymphocytic leukemia.* Blood, 2007. **109**(5): p. 2198-201.

Karolinska
Institutet