Molecular Epidemiology of Inflammation

Link with Type 2 Diabetes and Coronary Heart Disease

Symen Ligthart

All of the studies described in this thesis involved the Rotterdam Study, which is supported by the Erasmus MC and the Erasmus University Rotterdam, the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw), the Dutch Heart Foundation, the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture, and Science, the Ministry of Health Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. The contribution of the inhabitants, general practitioners, and pharmacists of the Ommoord district to the Rotterdam Study are gratefully acknowledged.

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

Cover design: © Cas Oorthuys / Nederlands Fotomuseum. Werknemers van de RDM, Heijplaat.
Layout and printing: Optima Grafische Communicatie.
ISBN 978-94-6361-384-2

Molecular Epidemiology of Inflammation

Link with Type 2 Diabetes and Coronary Heart Disease

Moleculaire epidemiologie van inflammatie

Verband met type 2 diabetes en coronaire hartziekte

Proefschrift

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

Prof. dr. R.C.M.E. Engels

en volgens het besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

woensdag 19 februari 2020 om 15:30 uur

door

Symen Ligthart
geboren te Dordrecht

Erasmus University Rotterdam

Promotiecommissie

Promotor:

Prof. dr. M.A. Ikram

Overige leden:

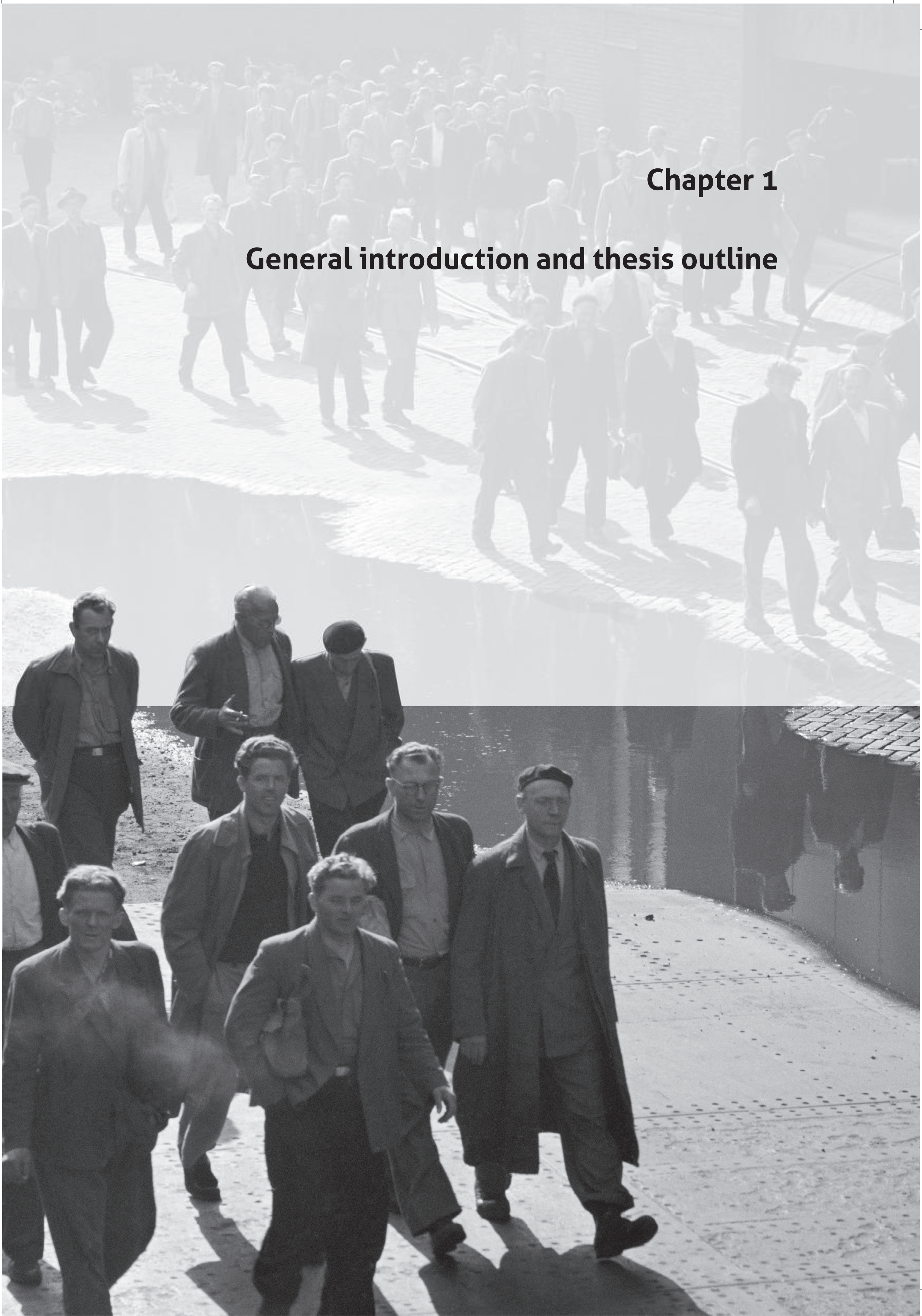
Prof. dr. E.J.G. Sijbrands
Prof. dr. H. Snieder
Prof. dr. A.G. Uitterlinden

Copromotor:

Dr. A. Dehghan

To my family

Table of Contents



Chapter 1

General introduction and thesis outline

Type 2 diabetes and cardiovascular disease (CVD) are complex diseases with a high disease burden worldwide. Although much effort has been done to reduce the occurrence of both diseases, their prevalence is rising.

Worldwide, it is estimated that 425 million adults live with diabetes[1]. In the United States of America, the lifetime risk to develop type 2 diabetes is approximately one in three[2]. The total health expenditure is 185 billion euros in Europe and 440 billion US dollars in North America, emphasizing the large burden on health cost.

Lifetime risk estimates for death from CVD are approximately one in three, with major differences according to the common risk factors for CVD[3]. Although incidence rates of CVD have reduced in developed countries in recent decades[4], CVD still remains the number one cause of death worldwide[5]. These data suggest that further understanding of the pathogenesis of diabetes and CVD is needed to reduce their burden on health.

Lifetime risk

Lifetime risks reflect the cumulative risk of developing a disease during an individual's remaining lifespan. Thus, the lifetime risk of diabetes at the age of 45 reflects the risk for an individual aged 45 to develop diabetes throughout the rest of his or her life. As patients and health care providers prefer absolute long-term risks over relative risks in disease risk communication[6], lifetime risks are an informative risk estimate to guide disease risk assessment and disease prevention. With the use of a modified version of survival analysis taking into account left- and right censoring, lifetime risks may be calculated using data from population-based cohort studies without lifelong follow-up[7]. It should be noted that when calculating lifetime risks, it is important to account for the competing risk of death from another cause to avoid overestimation of the lifetime risk of disease[8].

Chronic inflammation and C-reactive protein

Inflammation is the complex immune response of the body to a noxious stimulus. Cardinal signs of inflammation were first recorded by the Roman encyclopaedist Aulus Cornelius Celsus (25 BC-AD 50) in "De Medicina" as calor (warmth), dolor (pain), tumor (swelling), and rubor (redness). Where the ancient Romans referred to the more "acute" and "local" inflammatory response of the human body to a noxious stimulus, nowadays much interest has grown in the study of "chronic" and "systemic" inflammation. Chronic systemic inflammation is a common and highly complex response of the innate immune system involving a diversity of cytokines, interleukins, and other molecules that involves multiple organs. C-Reactive Protein (CRP), a sensitive acute phase reactant, has widely been used as an index for chronic systemic inflammation. It was first discovered in 1930 by William Tillet and Thomas Francis from the Rockefeller University[9].

Chronic inflammation and complex disorders

In recent decades, clinicians and researchers have been interested in causes and consequences of inflammation. Development of high-sensitivity CRP assays has enabled quantification of CRP in the low to very-low range of the pentameric protein and has revealed an association between low range CRP and complex diseases[10]. By doing so, the role of chronic inflammation in the pathogenesis of multiple complex disease including diabetes and CVD has been recognized[11,12], together with the development of interventions to treat chronic inflammation and illness[13,14,15,16]. Prospective observational research has established an association of circulating CRP levels with type 2 diabetes[17] and CVD[18]. Although the published data support a role of inflammation in the development of complex disorders, the causal role of the CRP protein itself has been disputed. Several well-conducted studies, randomized by the genetic variant for higher CRP (referred to as Mendelian randomization), have concluded that a causal role of CRP in diabetes and CVD is unlikely[19-22]. CRP thus seems to be an innocent bystander and this has led to the question which inflammatory processes and proteins upstream of CRP are causal to the disease.

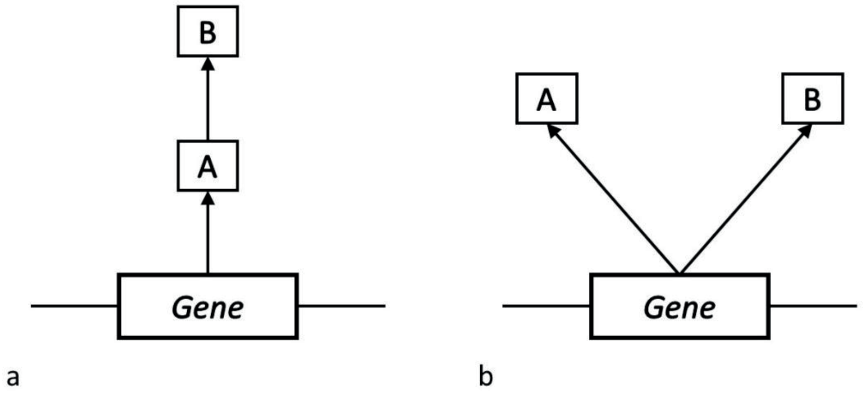
Genetic determinants of inflammation

The identification of genetic loci associated with inflammation may help in the search for molecular pathways underlying chronic systemic inflammation. Over the last decade, genome-wide association studies (GWAS) became a common approach to study genetic determinants of complex traits and diseases. GWAS investigate on a hypothesis-free basis the association between DNA sequence variants and phenotypes of interest, and have been successful in the identification of thousands of genetic loci for a wide range of phenotypes and diseases[23]. With an heritability of up to 50%[24], serum CRP levels has been one of the traits that researchers have searched for its genetic determinants. In candidate gene studies, researchers first identified polymorphisms in the *CRP* gene that related to serum CRP levels[25]. Later, genome-wide association analyses revealed genes outside the *CRP* gene that influence CRP levels[26]. In 2012, the largest GWAS on serum CRP levels identified 18 genetic loci explaining up to 5% of its heritability[27]. Thus, most of the heritability and molecular pathways underlying CRP levels remain to be determined. Further extending the sample size has been successful in GWAS to increase power and detect further genes for phenotypes of interest[28]. Also, the improvement of reference panels for the imputation of genetic variants, such as the 1000Genomes project[29] and the Haplotype Reference Consortium (HRC)[30], has the potential to study less frequent genetic variants, as well as insertions and deletions (INDELs). Considering these advantages, extending the sample size and the application of novel imputation panels may help to identify further genetic loci for serum CRP levels.

Genetic pleiotropy

DNA sequence variants may be associated with more than one phenotype, a phenomenon termed genetic pleiotropy[31]. The use of large GWAS meta-analyses has been successful in the identification of Single Nucleotide Polymorphisms (SNPs) that are associated with more than one phenotype. In general, genetic pleiotropy can be subdivided in two different types: vertical (also referred to as mediated) and horizontal (also referred to as biological) pleiotropy. Vertical pleiotropy refers to the scenario in which a gene causes phenotype A, and phenotype A causes phenotype B (Figure 1a). In horizontal pleiotropy, a gene is independently associated with phenotype A and phenotype B (Figure 1b). A better understanding of the shared genetic architecture of inflammation and associated phenotypes may contribute to a better understanding of how CRP levels link to those phenotypes, and may point to upstream mediators that are causal to clinical outcomes.

Figure 1. Vertical (a) and horizontal (b) genetic pleiotropy.



Epigenetics and inflammation

Complementary to the study of DNA sequence variants to identify determinants of phenotypes, recent studies have highlighted the importance of epigenetics in complex traits[32,33]. Epigenetics have the potential to change the function of the genome, without altering a person’s DNA sequence. DNA methylation is one of the most important and common epigenetic mechanism[34]. DNA methylation refers to the addition of a methyl-group to the DNA, which almost exclusively occurs at a cytosine nucleotide that is located next to a guanine (CpG sites). DNA methylation may change gene expression and genome stability, and is affected by both genetic and environmental factors[34]. Recently, techniques have been developed to quantify DNA methylation at thousands of CpG sites across the genome[35]. Whole blood is a readily available tissue in humans, and whole blood DNA is almost exclusively composed of white blood cell DNA. Hence, whole blood DNA methylation is highly suitable for studying DNA methylation in association with inflammation and related complex diseases. The selection of CpG sites to study with a phenotype of interest may be based on findings from prior research. Additionally, in the hypothesis-free epigenome-wide association studies (EWAS), DNA methylation at thousands of CpG sites are associated with phenotypes of interest.

Outline of this thesis

In part 1 of this thesis, the aim is to estimate the lifetime risk of diabetes for different subgroups of individuals. In chapter 2, the lifetime risk of prediabetes, diabetes, and insulin use is calculated in a community-dwelling European population stratified for body mass index. In chapter 3, I studied the lifetime risk of diabetes based on genetic background, and investigated whether adherence to a normal weight attenuates high genetic lifetime risk. In part 2, I aimed to identify novel inflammatory markers for diabetes and coronary heart disease (CHD). This to improve disease prediction and/or identify potential novel targets for future therapeutic interventions. In chapter 4, the aim is to identity novel inflammatory markers for diabetes, and in chapter 5 I seek to find novel markers for CHD. In part 3 the aim is to provide a better understanding of the genes and molecular pathways that regulate chronic inflammation and relate inflammation to cardiometabolic phenotypes. Therefore I sought to identify genetic determinants of CRP levels and studied the genetic overlap between CRP and related complex diseases. In chapter 6 I performed a GWAS of CRP levels, and estimated the causal inference of CRP levels on several clinical outcomes. Chapter 7 comments on the observation that genetically elevated CRP is associated with risk of schizophrenia. In chapter 8, shared genetic variants are studied between CRP and cardiometabolic diseases. In chapter 9, by applying a novel bivariate GWAS method, I aimed to identify novel genetic variants for CRP and lipid levels. In chapter 10, the causal effect of vitamin D on inflammation and vice versa was tested. In part 4, the role of DNA methylation in chronic inflammation and complex diseases was studied. In chapter 11, an EWAS was performed on serum CRP levels, and the link between inflammation related methylation sites and complex disease was investigated. Furthermore, in chapter 12 I performed an EWAS on tumor necrosis factor α (TNFα) levels, and tested the association between TNFα-associated methylation sites with incident CHD. Chapter 13 is devoted to the association between tobacco smoking and DNA methylation of diabetes susceptibility genes. In chapter 14, the association between tobacco smoking and DNA methylation at genes identified for coronary artery disease is studied. Finally, in part 5 (chapter 15), I summarize the main findings and discuss advantages and disadvantages of the methods used, and the implications for future research.

References

1. International Diabetes Federation. Diabetes Atlas, 7th edition. Brussels, Belgium: International Diabetes Federation, 2015.

2. Narayan KV, Boyle JP, Thompson TJ, Sorensen SW, Williamson DF. Lifetime risk for diabetes mellitus in the United States. *JAMA* 2003; 290(14): 1884-90.

3. Berry JD, Dyer A, Cai X, et al. Lifetime risks of cardiovascular disease. *N Engl J Med* 2012; 366(4): 321-9.

4. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. *Circulation* 2017; 135(10): e146-e603.

5. World Health Organization. Global status report on noncommunicable diseases 2010. 2011.

6. Fortin JM, Hirota LK, Bond BE, O'Connor AM, Col NF. Identifying patient preferences for communicating risk estimates: a descriptive pilot study. *BMC Med Inform Decis Mak* 2001; 1(1): 2.

7. Meister R, Schaefer C. Statistical methods for estimating the probability of spontaneous abortion in observational studies—analyzing pregnancies exposed to coumarin derivatives. *Reprod Toxicol* 2008; 26(1): 31-5.

8. Satagopan J, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach A. A note on competing risks in survival data analysis. *Br J Cancer* 2004; 91(7): 1229.

9. Tillett WS, Francis T. Serological reactions in pneumonia with a non-protein somatic fraction of pneumococcus. *J Exp Med* 1930; 52(4): 561-71.

10. Emerging Risk Factors Collaboration. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet* 2010; 375(9709): 132-40.

11. Pickup JC. Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes. *Diabetes Care* 2004; 27(3): 813-23.

12. Libby P. Inflammation in atherosclerosis. *Arterioscler Thromb Vasc Biol* 2012; 32(9): 2045-51.

13. Ridker PM, Danielson E, Fonseca F, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med* 2008; 359(21): 2195.

14. Ridker PM, Thuren T, Zalewski A, Libby P. Interleukin-1β inhibition and the prevention of recurrent cardiovascular events: rationale and design of the Canakinumab Anti-inflammatory Thrombosis Outcomes Study (CANTOS). *Am Heart J* 2011; 162(4): 597-605.

15. Ridker PM. Testing the inflammatory hypothesis of atherothrombosis: scientific rationale for the cardiovascular inflammation reduction trial (CIRT). *J Thromb Haemost* 2009; 7(s1): 332-9.

16. Larsen CM, Faulenbach M, Vaag A, et al. Interleukin-1–receptor antagonist in type 2 diabetes mellitus. *N Engl J Med* 2007; 356(15): 1517-26.

17. Pradhan AD, Manson JE, Rifai N, Buring JE, Ridker PM. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA* 2001; 286(3): 327-34.

18. Ridker PM, Buring JE, Shih J, Matias M, Hennekens CH. Prospective study of C-reactive protein and the risk of future cardiovascular events among apparently healthy women. *Circulation* 1998; 98(8): 731-3.

19. Brunner EJ, Kivimäki M, Witte DR, et al. Inflammation, insulin resistance, and diabetes—Mendelian randomization using CRP haplotypes points upstream. *PLoS Med* 2008; 5(8): e155.

20. Dehghan A, Kardys I, de Maat MP, et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. *Diabetes* 2007; 56(3): 872-8.

21. C-Reactive Protein Coronary Heart Disease Genetics Collaboration. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ* 2011; 342: d548.

22. Elliott P, Chambers JC, Zhang W, et al. Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA* 2009; 302(1): 37-48.

23. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 2017; 101(1): 5-22.

24. Pankow JS, Folsom AR, Cushman M, et al. Familial and genetic determinants of systemic markers of inflammation: the NHLBI family heart study. *Atherosclerosis* 2001; 154(3): 681-9.

25. Carlson CS, Aldred SF, Lee PK, et al. Polymorphisms within the C-reactive protein (CRP) promoter region are associated with plasma CRP levels. *Am J Hum Genet* 2005; 77(1): 64-77.

26. Ridker PM, Pare G, Parker A, et al. Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am J Hum Genet* 2008; 82(5): 1185-92.

27. Dehghan A, Dupuis J, Barbalic M, 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): 731-8.

28. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet* 2012; 90(1): 7-24.

29. Consortium GP. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; 491(7422): 56.

30. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; 48(10): 1279-83.

31. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 2013; 14(7): 483.

32. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* 2014; 383(9933): 1990-8.

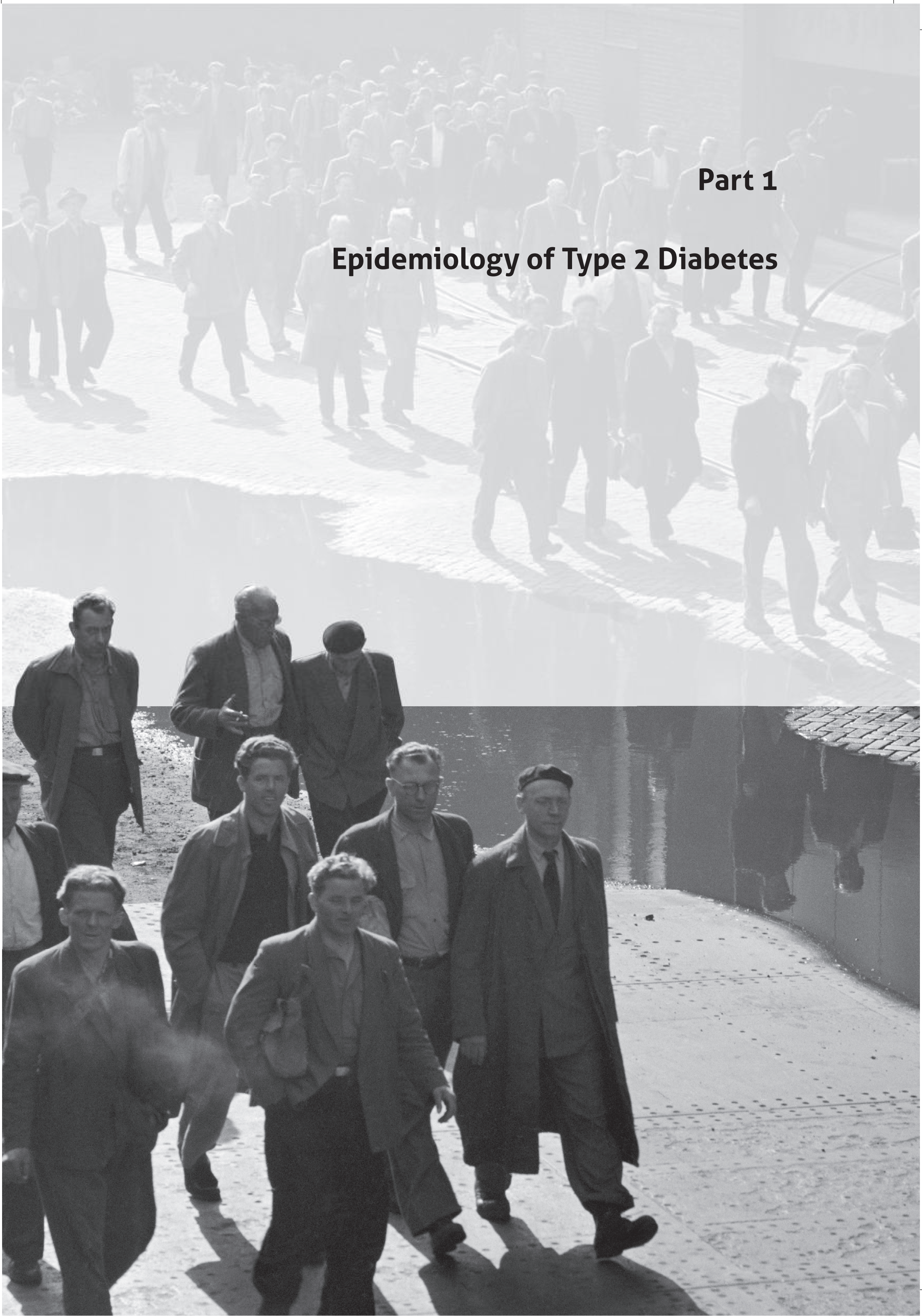
33. Irvin MR, Zhi D, Joehanes R, et al. Epigenome-wide association study of fasting blood lipids in the genetics of lipid lowering drugs and diet network study. *Circulation* 2014; 130(7): 565-72.

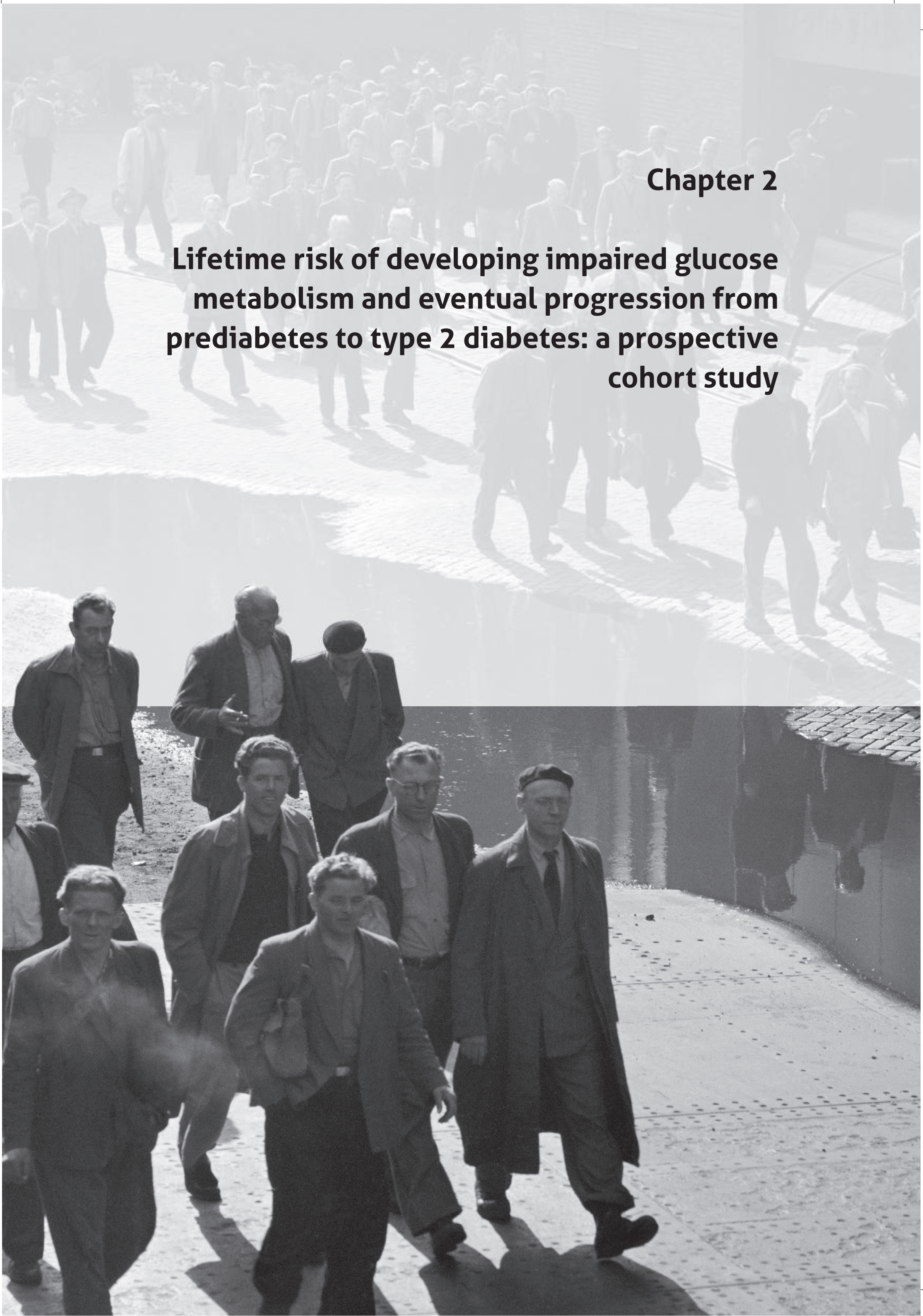
34. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012; 13(7): 484.

35. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011; 98(4): 288-95.

Part 1

Epidemiology of Type 2 Diabetes





Chapter 2

Lifetime risk of developing impaired glucose metabolism and eventual progression from prediabetes to type 2 diabetes: a prospective cohort study

Background: Data on lifetime risk of the full spectrum of impaired glucose metabolism including prediabetes and the risk to eventually progress to diabetes or start insulin therapy are scarce.

Methods: We used data from 10050 participants from the prospective population-based Rotterdam Study. Events were diagnosed by use of general practitioners records, hospital discharge letters, pharmacy dispensing data and serum fasting glucose measurements at the study center visits. Normoglycemia, prediabetes and diabetes were defined according to the WHO criteria for fasting glucose (normoglycemia: ≤6.0 mmol/L; prediabetes: >6.0mmol/L and <7.0mmol/L; diabetes ≥7.0 mmol/L or use of glucose lowering therapy). Lifetime risks were calculated using a modified version of survival analysis adjusted for the competing risk of death. In addition, we estimated the lifetime risk of progression from prediabetes to overt diabetes and from diabetes free of insulin therapy to insulin use. Further, we calculated years lived with healthy glucose metabolism.

Results: During a follow-up of up to 14.7 years, 1148 participants developed prediabetes, 828 diabetes and 237 started insulin therapy. At the age of 45, the remaining lifetime risk (95%CI) was 48.7% (46.2%-51.3%) for prediabetes, 31.3% (29.3%-33.3%) for diabetes and 9.1% (7.8%-10.3%) for insulin use. The lifetime risk to progress from prediabetes at the age of 45 to diabetes was 74.0% (67.6%-80.5%), and 49.1% (38.2%-60.0%) of the individuals with overt diabetes at the age of 45 started insulin therapy. The lifetime risks attenuated with advancing age but increased with increasing body mass index and waist circumference. On average, individuals with severe obesity lived 10 fewer years without glucose impairment compared to normal-weight individuals.

Conclusion: Our results highlight the public health burden posed by glycemic disturbances and demand further investigation into earlier and more effective prevention strategies.

Introduction

People with elevated blood glucose levels below the threshold of diabetes, a state referred to as prediabetes, have an excess risk of diabetes[1,2,3]. Today, more than 382 million people live with diabetes worldwide and due to the increasing prevalence of prediabetes and the rapid conversion of prediabetes to type 2 diabetes, the number is predicted to exceed half a billion by 2035[4]. Moreover, many diabetes patients are unable to achieve glycemic control goals through diet or oral medications only and ultimately require insulin treatment[5,6,7]. Estimates on the progression from prediabetes to diabetes and ultimately insulin therapy are scarce and have been limited to merely annual incidences and absolute risks within a restricted time period[2].

Lifetime risks provide estimation of the cumulative risk of developing a disease during an individual's remaining lifespan and comprise thus a clear message to patients, clinicians and policy makers[8,9,10]. A few reports have simulated the lifetime risk of type 2 diabetes in the US and Australia[10,11,12]. However, estimates using accurate and careful documentation of elevated blood glucose levels, diabetes diagnosis and diabetes drug use are lacking. Prospective population-based cohort studies with long-term follow-up and detailed data on the full spectrum of impaired glucose metabolism including prediabetes, diabetes and the eventual need for insulin therapy, permit estimation of the burden of elevated blood glucose levels in the context of overall survival.

Hence, we used mortality rates and incidences of the disease during every year of life taking into account the competing risk of death to assess the lifetime risks of prediabetes, diabetes and insulin use in a large prospective population-based cohort study of individuals aged 45 years and older. Additionally, we estimated the lifetime risk of individuals with prediabetes to eventually develop diabetes and for diabetes patients to ultimately use insulin.

Methods

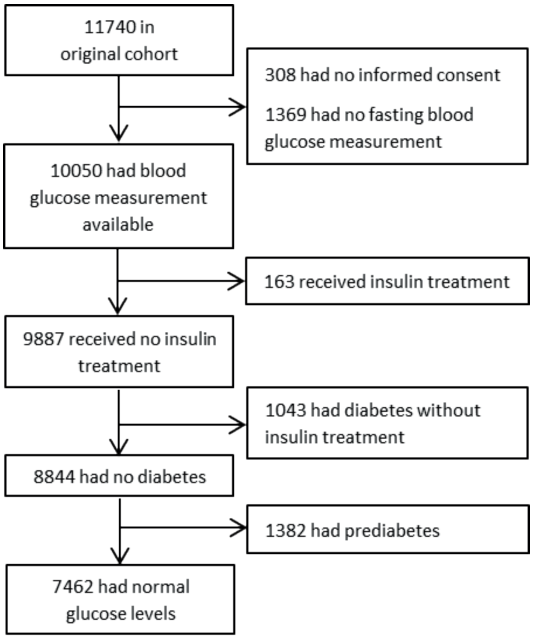
Study design and population

This study is embedded within the framework of the Rotterdam Study, a prospective cohort study among the community-dwelling population aged 45 years and older in the city of Rotterdam, the Netherlands. The study design of the Rotterdam Study has been described in detail previously[13]. Briefly, in 1990 all inhabitants of a well-defined district of Rotterdam were invited, of whom 7983 agreed to participate (78.1%). The study was extended in 2000 with a second cohort of individuals who had reached the age of 55 or moved into the study area after 1990 (n=3011). In 2006, a third cohort was enrolled including inhabitants aged 45 years and older (n=3932), bringing the total study size to 14926 individuals. There were no eligibility criteria to enter the Rotterdam study cohorts except the minimum age and residential area based on ZIP codes. We used the third center visit (1997–1999, n=4216) of the first cohort and the first visit of the second and third cohorts as baseline (2000–2001 and 2006–2008, respectively) for the current analysis. To ascertain the absence of prediabetes or diabetes by means of serum glucose measurement and use of blood glucose lowering medication, we excluded 1369 individuals without a valid baseline glucose measurement. Next, for the calculation of the lifetime risk of prediabetes, we only included individuals that were normoglycemic at study baseline (n=7462). To calculate the lifetime risk of diabetes, we only included individuals that were free of diabetes at study baseline (n=8844). Further, to calculate the lifetime risk of insulin use, we only included individuals that were free of insulin use at study baseline (n=9887). Selection of the individuals for the analyses can be found in Figure 1. The individuals with prediabetes (n=1382) were used to study progression from prediabetes to diabetes and individuals with diabetes without insulin treatment (n=1043) were used to study the progression from diabetes to insulin use. The Rotterdam Study has been approved by the medical ethics committee according to the Population Screening Act: Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. All participants in the present analysis provided written informed consent to participate and to obtain information from their treating physicians.

Ascertainment of prediabetes and type 2 diabetes

The participants were followed from the date of baseline center visit onwards. At baseline and during follow-up, cases of prediabetes and type 2 diabetes were ascertained through active follow-up using general practitioners' records (including laboratory glucose measurements), hospital discharge letters and serum glucose measurements from Rotterdam Study visits which take place approximately every four years[14]. Diabetes, prediabetes and normoglycemia were defined according to the recent WHO guidelines[15]. Normoglycemia was defined as a fasting blood glucose level ≤6.0 mmol/L; prediabetes was defined as a fasting blood glucose >6.0 mmol/L and <7.0 mmol/L or a non-fasting blood glucose >7.7 mmol/L and <11.1 mmol/L (when fasting samples were unavailable); type 2 diabetes was defined as a fasting blood glucose ≥7.0 mmol/L, a non-fasting blood glucose ≥11.1 mmol/L (when fasting samples were unavailable), or the use of blood glucose lowering medication. Information regarding the use of blood glucose lowering medication was derived from both structured home interviews and linkage to pharmacy dispensing records[14]. At baseline, more than 95% of the Rotterdam Study population was covered by the pharmacies in the study area. All potential events of prediabetes and type 2 diabetes were independently adjudicated by two study physicians. In case of disagreement, consensus was sought with an endocrinologist. Follow-up data was complete until January 1st 2012.

Figure 1. Participants Selection.


11740 in original cohort
308 had no informed consent
1369 had no fasting blood glucose measurement
10050 had blood glucose measurement available
163 received insulin treatment
9887 received no insulin treatment
1043 had diabetes without insulin treatment
8844 had no diabetes
1382 had prediabetes
7462 had normal glucose levels

Participants without insulin treatment were used for the lifetime risk of insulin dependency, participants without diabetes for the lifetime risk of diabetes and participants with normal glucose levels for the lifetime risk of prediabetes. Progression from prediabetes to diabetes was assessed in the individuals with prediabetes and the progression from diabetes to insulin dependency in the individuals with diabetes without use of insulin treatment.

Statistical analysis

Baseline characteristics were compared between normoglycemic individuals, individuals with prediabetes and individuals with type 2 diabetes using linear regression models, Kruskal-Wallis tests for continuous data, and χ^2 tests for categorical data.

Remaining lifetime risks at different ages were calculated for prediabetes, diabetes and insulin use. We used a modified version of survival analysis to take the competing event of death into account for the calculation of the absolute lifetime risk (see appendix page 3 for statistical details). Lifetime risk estimates were calculated at index ages 45, 55, 65, 75, and 85 years for men and women combined and separately. The lifetime risk estimates reflect the remaining risk at the index age to the age of last observation (107 years in our study). In addition, to compare lifetime risks at the index ages with absolute risks in a shorter time period, we also calculated 10-year risks for prediabetes, diabetes and insulin use at all index ages.

Next, we calculated the lifetime risk of diabetes only in individuals with prediabetes in order to obtain an estimate of the lifetime risk to progress from prediabetes to diabetes. Similarly, we calculated the lifetime risk of insulin use in individuals with diabetes free of insulin therapy to study what percentage of individuals with diabetes will eventually start insulin therapy.

In order to analyze the effect of anthropometric measures on the lifetime risk of prediabetes, diabetes and insulin use, we computed lifetime risks at the age of 45 stratified by BMI and waist circumference. The individuals were stratified into four categories of BMI (<25 kg/m^2, 25-30 kg/m^2, 30-35 kg/m^2, and >35 kg/m^2) and three categories of waist circumference based on the WHO classification scheme (for men: <94 cm, 94-101 cm and ≥102 cm; for women: <80 cm, 80-87 cm and ≥88 cm)[16].

To study the delay in onset of prediabetes, diabetes and insulin use we examined the difference in mean disease-free survival among BMI and waist circumference strata. As censoring precludes estimation of the mean survival time, we used Irwin's restricted mean survival to calculate the mean disease-free survival and overall mean survival[17]. Irwin's restricted mean survival is the mean of the survival time up to a point in time and mathematically is the area under the survival curve up to the selected point in time. As data from individuals aged older than 100 was limited, we set the restriction time point to 100 years of age.

All data were analyzed using the IBM SPSS Statistics version 21.0.0.1 (IBM Corp, Somers, NY, USA) and R version 2.1 with the 'etm' and 'survival' libraries[18,19].

Results

Baseline population characteristics

The mean (SD) age of the population was 65.2 (9.8) and women made up the majority of the study population (56.5%). Of the 10050 participants at baseline, 7462 (74.2%) had normoglycemia, 1382 (13.8%) had prediabetes, and 1206 (12.0%) had diabetes (Table 1). Prevalences of prediabetes and type 2 diabetes increased with advancing age in both men and women and were higher in men compared to women (appendix page 7). Individuals with prediabetes and diabetes had higher BMI and unfavorable lipid profile compared to normal glycemic individuals. Furthermore, people with diabetes had a higher prevalence of stroke, coronary heart disease and were more often smokers compared to normoglycemic individuals.

Lifetime risk of prediabetes, diabetes and insulin use

During 56230 person-years of follow-up in normoglycemic individuals, 1148 individuals developed prediabetes and 1343 died (incidence rate per 1000 person-years (IR): 20.4 (95%CI 19.3 to 21.6); mortality rate per 1000 person-years (MR): 23.9, 95%CI 22.7 to 25.2). We observed 828 cases of diabetes during 69639 person-years of follow-up in non-diabetic

Table 1: Baseline characteristics of participants by prevalent glycemic state.

Characteristics	Normal glucose (n=7462)	Prediabetes (n=1382)	Diabetes (n=1206)	P-value
Women (n, %)	4411 (59.1)	692 (51.0)	582 (47.7)	<0.001
Age (y)	64.4±9.8	66.6±9.4	67.5±9.6	<0.001
Waist circumference (cm)	90±11	96±12	101±12	<0.001
Body mass index (kg/m^2)	26.3±3.8	27.9±4.2	29.4±4.8	<0.001
Total cholesterol (mmol/L)	5.7±1.0	5.8±1.0	5.4±1.1	<0.001
HDL cholesterol (mmol/L)	1.4 (1.2-1.7)	1.3 (1.1-1.5)	1.2 (1.0-1.4)	<0.001
Triglycerides (mmol/L)[†]	1.3 (1.0-1.7)	1.5 (1.1-2.1)	1.7 (1.2-2.3)	<0.001
non HDL cholesterol (mmol/L)	4.3±1.0	4.4±1.0	4.2±1.1	<0.001
LDL cholesterol (mmol/L)	3.7±0.9	3.6±0.9	3.4±0.9	<0.001
Insulin (pmol/L)[†]	66 (47-93)	93 (64-133)	110 (73-177)	<0.001
Glucose (mmol/L)[†]	5.3 (5.0-5.6)	6.3 (6.1-6.5)	7.7 (7.0-9.5)	<0.001
eGFR (mL/min/1.73 m^2)	81±17	80±18	83±22	<0.001
C-reactive protein (mg/L)	1.4 (0.6-3.1)	2.0 (0.9-4.2)	2.5 (1.1-4.9)	<0.001
Systolic blood pressure (mmHg)	137±21	145±21	147±22	<0.001
Diastolic blood pressure (mmHg)	78±11	81±12	79±12	<0.001
Hypertension (n, %)	3448 (46.7)	873 (64.0)	875 (73.3)	<0.001
History of stroke (n, %)	178 (2.4)	35 (2.5)	77 (6.4)	<0.001
History of CHD (n, %)	422 (5.8)	110 (8.1)	164 (13.8)	<0.001
Use of blood pressure lowering drugs (n, %)	1408 (19.6)	437 (32.8)	474 (40.5)	<0.001
Use of lipid lowering agents (n, %)	1043 (14.4)	239 (17.8)	1178 (27.4)	<0.001
Current smoking (n, %)	707 (9.5)	144 (10.4)	147 (12.3)	0.001
Former smoking (n, %)	2769 (37.4)	553 (40.2)	480 (40.1)	0.001

Values are mean ± standard deviation or median (interquartile range) for characteristics with skewed distributions. eGFR denotes estimated glomerular filtration rate, HDL high-density-lipoprotein, and CHD coronary heart disease. [†]Only fasting samples.

Table 2: Remaining lifetime and 10-year risks of prediabetes, diabetes and insulin use.

Age, years		N	Lifetime risk prediabetes (95%CI)	N	Lifetime risk diabetes (95%CI)	N	Lifetime risk insulin use (95%CI)
45	Lifetime	7462	48.7% (46.2-51.3)	8844	31.3% (29.3-33.3)	9887	9.1% (7.8-10.3)
	10-year	1233	8.4% (5.4-11.4)	1344	3.4% (2.1-4.8)	1400	0.5% (0.0-1.0)
55	Lifetime	6939	44.5% (42.5-46.6)	8291	29.2% (2..4-31.0)	9329	8.8% (7.6-10.0)
	10-year	3788	13.2% (11.4-15.0)	4456	7.0% (5.8-8.3)	4914	2.2% (1.4-3.0)
65	Lifetime	5109	37.6% (35.6-39.5)	6257	24.8% (2..1-26.5)	7179	7.0% (6.0-7.9)
	10-year	3901	19.3% (17.7-20.9)	4754	11.2% (10.0-12.3)	5414	3.0% (2.4-3.6)
75	Lifetime	3073	25.8% (23.7-28.0)	3850	17.4% (15.7-19.1)	4547	4.7% (3.8-5.6)
	10-year	2885	19.1% (17.3-20.9)	3618	12.2% (10.9-13.6)	4273	2.6% (2.0-3.2)
85	Lifetime	1072	13.1% (10.4-15.7)	1405	8.9% (6.9-10.9)	1725	3.3% (2.2-4.4)
	10-year	1060	11.9% (9.5-14.4)	1390	8.2% (6.4-10.0)	1707	3.3% (2.2-4.4)

The lifetime risk of prediabetes is for individuals with normal glucose levels at the index age, lifetime risk of diabetes for individuals without diabetes at the index age and the lifetime risk of insulin use for individuals free of insulin use at the index age.

individuals and 1709 deaths (IR: 11.9, 95%CI 11.1 to 12.7; MR: 24.5, 95%CI 23.4 to 25.7). Among non-insulin users, 237 incident cases of insulin use were observed during 80832 person-years of follow-up (IR: 3.4, 95%CI 3.0 to 3.9) and 2183 individuals died (MR: 27.0, 95%CI 25.9 to 28.1). The remaining lifetime risk for a 45-year old individual to develop prediabetes was 48.7% (95%CI 46.2 to 51.3), whereas the lifetime risks of diabetes and insulin use were 31.3% (95%CI 29.3 to 33.3) and 9.1% (95%CI 7.8 to 10.3), respectively (Table 2). The cumulative incidences function subsequent to 45 years of age, are depicted in Figure 1. Lifetime risks of prediabetes, diabetes and insulin use subsequent to increasing ages attenuated. Compared to the lifetime risks, the 10-year risks of prediabetes, diabetes and insulin use were lower at all index ages. The remaining lifetime risks did not differ by gender irrespective of age (appendix page 8-9). With adjustment for the competing risk of death, the lifetime risks were lower as compared to the unadjusted risk derived from Kaplan-Meier estimates (appendix page 4).

Progression to diabetes and insulin use

In 1382 individuals with prediabetes we observed 425 incident cases of diabetes, whilst 257 died without diabetes (IR: 43.0, 95%CI 39.2 to 47.2; MR: 26.0, 95%CI 23.0 to 29.3). The lifetime risk for individuals that experience prediabetes at 45 years of age to progress to diabetes was 74.0% (95%CI 67.6 to 80.5). Further, among 1043 individuals with diabetes, we observed 183 incident cases of insulin use, whilst 302 died without ever using insulin treatment (IR: 24.0, 95%CI 20.8 to 27.7; MR: 39.7, 95%CI 35.5 to 44.3). The lifetime risk for individuals with diabetes at the age of 45 to start insulin therapy was 49.1% (95%CI 38.2 to 60.0).

Stratification by BMI and waist circumference

Stratification by BMI revealed that people with normal weight at the age of 45 have a significantly lower prediabetes lifetime risk compared to overweight and obese individuals (Table 3). Stratification by waist circumference revealed similar effects on the lifetime risks of prediabetes. In accordance with the lifetime risks for prediabetes, lifetime risks for diabetes and insulin use were higher with increasing BMI and waist circumference. The cumulative incidences by BMI strata as a function of age, for 45 year olds, are depicted in Figure 2. When we stratified individuals within the BMI strata by waist circumference categories, we observed an increasing risk of diabetes with increasing waist, except in the lowest BMI category (appendix page 10).

The lifetime risk to progress from prediabetes to diabetes was also substantially lower for individuals with a normal weight compared to overweight and obese individuals. However, the risk to start insulin therapy in individuals with diabetes did not differ substantially between strata of BMI (Table 4).

Table 3: Lifetime risk at the age of 45 for prediabetes, diabetes and insulin use by body mass index and waist circumference strata.

BMI (kg/m^2)	N	Lifetime risk prediabetes (95%CI)	P-value	N	Lifetime risk diabetes (95%CI)	P-value	N	Lifetime risk Insulin use (95%CI)	P-value
< 25	2686	36.9% (33.1-40.6)		2955	18.8% (15.8-21.7)		3124	4.6% (3.0-6.2)	
25-30	3452	52.1% (48.0-56.2)	<0.001	4132	33.1% (30.2-36.0)	<0.001	4598	9.6% (7.7-11.5)	<0.001
30-35	1013	60.2% (54.1-66.2)	0.02	1324	43.9% (38.6-49.2)	<0.001	1613	14.2% (10.5-17.9)	0.01
> 35	240	71.3% (60.5-82.0)	0.04	352	56.6% (46.7-66.6)	0.01	462	17.1% (9.9-24.2)	0.24
Waist circumference									
Small	2127	37.5% (33.2-41.7)		2329	19.5% (16.0-23.1)		2449	4.5% (2.8-6.3)	
Medium	2220	46.6% (41.1-52.2)	0.005	2589	25.1% (21.6-28.6)	0.01	2815	6.5% (4.4-8.5)	0.08
Large	2801	57.8% (54.2-61.5)	<0.001	3553	41.8% (38.5-45.0)	<0.001	4205	12.6% (10.5-14.8)	<0.001

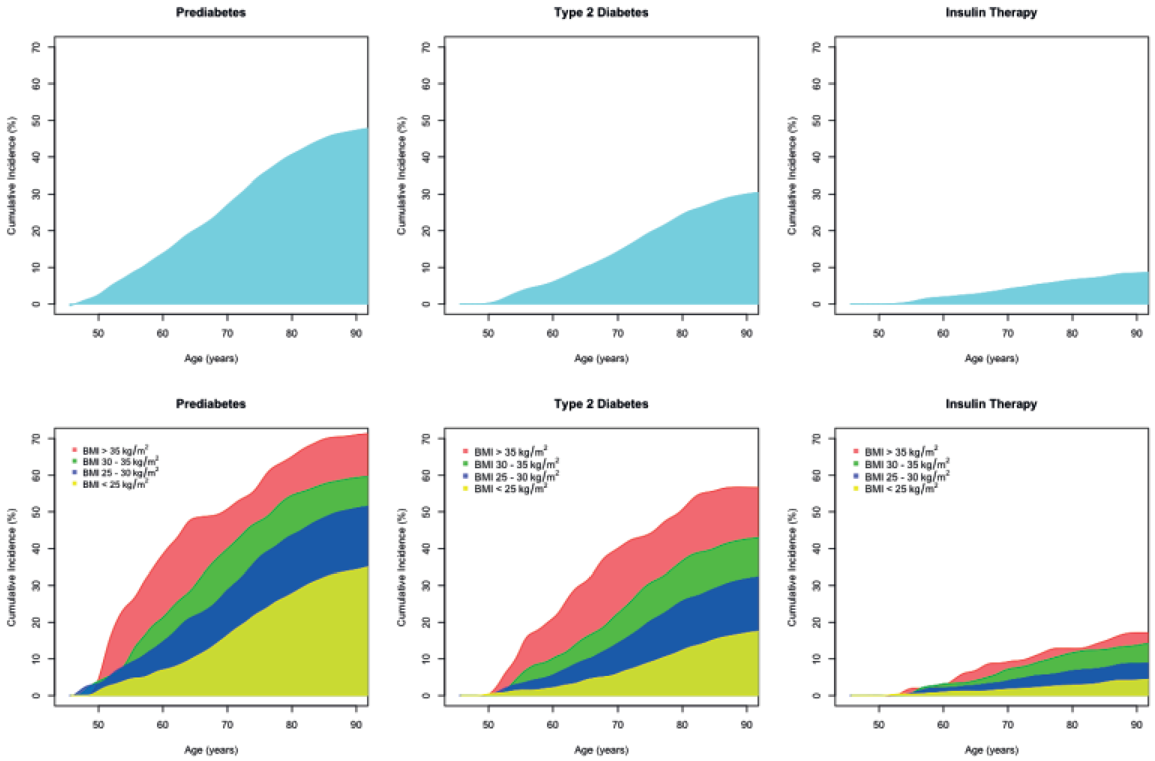
*Waist circumference categories small, medium and large represent the WHO classification scheme (for men: <94 cm, 94-102 cm and ≥102; for women: <80 cm, 80-88 cm and ≥88). BMI denotes body mass index. P-values are for the comparison with the lower BMI or waist category.
The lifetime risks are subsequent to the age of 45.

Table 4: Lifetime Risk to Develop Diabetes in Individuals with Prediabetes at the Age of 45 and the Lifetime Risk to Use Insulin in Individuals aged 45 with Diabetes but Free of Insulin Use, Stratified by Body Mass Index and Waist Circumference.

BMI (kg/m^2)	N	Prediabetes to diabetes (95%CI)	P-value	N	Non-insulin dependent diabetes to insulin use (95%CI)	P-value
< 25.0	269	35.9% (18.4-53.4)		169	58.6% (34.2-83.1)	
25 ⍰ 30	680	76.3% (68.9-83.7)	<0.0001	466	57.5% (45.2-69.9)	0.47
30 ⍰ 35	311	87.7% (79.4-96.0)	0.02	289	25.0% (0.0-60.1)	0.04
> 35	112	80.9% (65.9-95.9)	0.22	110	46.4% (28.6-64.1)	0.14
Waist circumference						
Small	202	49.8% (28.6-70.9)		120	43.0% (24.4-61.7)	
Medium	369	70.0% (56.4-83.7)	0.06	226	55.9% (38.6-73.2)	0.16
Large	752	78.8% (70.4-87.1)	0.14	652	45.0% (29.0-61.0)	0.18

*Waist circumference categories small, medium and large represent the WHO classification scheme (for men: <94 cm, 94-102 cm and ≥102; for women: <80 cm, 80-88 cm and ≥88). P-values are for the comparison with the lower BMI or waist category.

Figure 2. Lifetime risk of prediabetes, type 2 diabetes and insulin therapy among individuals at 45 years of age, adjusted for the competing risk of death.



The upper panel depicts the cumulative incidence of prediabetes (fasting glucose >6.0 mmol/L), type 2 diabetes (fasting glucose ≥7.0 mmol/L or use of glucose lowering medication) and insulin use among all individuals at 45 years of age, adjusted for the competing risk of death. The lower panel depicts the cumulative incidences of prediabetes, type 2 diabetes and insulin therapy among individuals at 45 years of age, according to body mass index and adjusted for the competing risk of death.

Diabetes-free survival

Remaining life years free of prediabetes, diabetes and insulin use from the age of 45 by sex, BMI and waist circumference strata are depicted in appendix page 5-6. Overall, years lived with normal glucose metabolism diminished with increasing levels of obesity. Also, individuals with higher BMI experienced more years lived with diabetes. For example, the average age of onset of prediabetes in men with normal weight was more than 10 years later compared to men with a BMI >35 kg/m^2. On average, insulin therapy is only used in a short time period at the end of life.

Discussion

The lifetime risk of prediabetes for an individual aged 45 is one in two, and one in three individuals aged 45 will develop diabetes. The vast majority of individuals that have prediabetes at age 45 will eventually progress to diabetes and one in two diabetes patients aged 45 will start insulin therapy. Furthermore, obesity substantially affects the risk to progress from prediabetes to diabetes and compresses the years lived with normal glucose metabolism.

Individuals with prediabetes have an increased risk of diabetes, cardiovascular disease, cancer and mortality[20,21,22]. Despite the high prevalence of prediabetes, estimates of what proportion of the population will eventually present with prediabetes have not been previously published. Evidence regarding the preventive effects of both lifestyle and pharmacological interventions on the progression of prediabetes to diabetes increases[23,24,25]. Lifetime risk estimates may indicate the proportion of individuals for whom early intervention would be applicable. We observed that half of our population will sooner or later present with prediabetes and may qualify for potential interventions during their lifespan.

In contrast to diabetes, prediabetes is a more fluctuating health state. The lifetime risk estimates of prediabetes in the current study should be interpreted as ever experiencing a serum glucose in the prediabetes range. However, individuals diagnosed with prediabetes could return to normoglycaemia. In the Diabetes Prevention Program (DPP), 19% of the placebo group returned to normoglycemia within 10 year[26]. In the pioglitazone for diabetes prevention study, 28% returned to normoglycemia in the placebo arm during a median follow-up of 2.4 years. These estimates are based on a limited time period and longer follow-up may result in different estimates. Our lifetime estimates show that 3 in 4 individuals with a glucose level in the prediabetes range at the age of 45 progress to diabetes. These estimates provide a better long-term perspective of individuals who ever meet prediabetes criteria, irrespective whether an individual remains prediabetes or returns to normoglycemia in the following years. This is in agreement with the American

Diabetes Association expert panel suggesting that 70% of the individuals with prediabetes progress to diabetes[2]. The higher prevalence of obesity in the US raises the concern of even higher progression rates compared to the estimates from our European population.

A previous report simulated the lifetime risk of diabetes in a US population based on questionnaire data for the adjudication of diabetes which does not comprise undiagnosed diabetes[11]. As the prevalence of undiagnosed diabetes is more than 25%[27], the risk estimates in the US study are likely to be underestimated. Furthermore, an Australian study estimated the lifetime risk of diabetes using two cross-sectional examinations (diabetes defined based on fasting plasma glucose (≥7.0 mmol/L) and 2hr plasma glucose (≥11.1 mmol/L)) with a short time interval (5 years) in a population with a large number of dropouts (39%) and without active follow[12]. Instead, we used active follow-up data and fasting glucose measurements as an objective and comprehensive assessment of diabetes diagnosis enabling us to provide accurate estimates of the entire spectrum of impaired glucose metabolism.

We observed a substantial impact of obesity on the remaining lifetime risk of prediabetes, diabetes and insulin use, which is in line with a previous report[28]. Also, obesity increased the risk to progress from prediabetes to diabetes. This is in agreement with the observation in the placebo group of the DPP in which obese individuals had a higher risk to progress to diabetes (9 vs 14 cases/100 person-years)[23]. Furthermore, obesity compressed the survival with normal glucose metabolism and influenced the time lived within each glycemic state underscoring the importance of weight management.

We estimate that one in two patients with diabetes eventually start insulin treatment. Together, the lifetime risk of diabetes and insulin use show the burden of pharmacological treatment of diabetes in our Western population. Despite the consensus statements from the ADA and the European Association for the Study of Diabetes in 2006[29] and 2009[30], physicians consider a variety of non-standardized factors to initiate glucose lowering treatment[31]. Therefore, our lifetime risks of insulin use may not reflect country and population-specific prescription behaviors. Furthermore, recent developments in diabetes care include the initiation of insulin therapy for beta-cell preservation, which has not been common practice in the calendar time period of our study. The lifetime risk of 9.1% in non-insulin users at the age of 45 may therefore be an underestimation of the risk of insulin use in current diabetes clinical care.

The strength of our study is the comprehensive assessment of incident diabetes diagnosis through use of blood glucose lowering treatment using medical records from hospitals and general practitioners, standardized blood glucose measurements at the repeated study center visits, and electronic linkage with the pharmacy dispensing records in the study area. Also, we used data from a prospective population-based cohort study with long-term follow-up and adjusted the lifetime risks for the competing risk of death to avoid overestimation. We need to address some limitations. First, we calculated remaining lifetime risks at the age of 45 because we did not have data for individuals younger than 45. Nevertheless, the cumulative incidence of type 2 diabetes before the age of 45 is low[10,11]. Also, for estimating the lifetime risk in BMI and waist circumference strata, we used anthropometric data at older ages than 45 as not all individuals entered the study at age 45. This could have led to the misclassification of individuals across the different categories as BMI and waist circumference could have changed with age. Third, we used data from an completely unselected sample of the general Dutch population with high participation rates (72.0%)[13]. However, all studies requiring active participation are to some extent subject to the “healthy volunteer effect” and this generally leads to slight underestimations of absolute risk estimates at short term follow-up. However, this underestimation attenuates at long-term follow-up[32]. Last, the vast majority of the Rotterdam Study participants are white (97%) and we therefore present lifetime risks for individuals from European ancestry. Half of the general population will sooner or later develop prediabetes defined as fasting glucose >6.0 mmol/L. Up to three in four of those with prediabetes aged 45 will progress to diabetes and one in two diabetics aged 45 eventually starts insulin therapy. These lifetime risks demonstrate the burden of impaired glucose metabolism on our society and demand earlier and more effective prevention strategies.

References

1. Gerstein HC, Santaguida P, Raina P, et al. Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies. *Diabetes Res Clin Pract* 2007; 78(3): 305-12.

2. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. *Lancet* 2012; 379(9833): 2279-90.

3. Yeboah J, Bertoni AG, Herrington DM, Post WS, Burke GL. Impaired fasting glucose and the risk of incident diabetes mellitus and cardiovascular events in an adult population: MESA (Multi-Ethnic Study of Atherosclerosis). *J Am Coll Cardiol* 2011; 58(2): 140-6.

4. International Diabetes Federation. IDF Diabetes Atlas, 6th edn. *Brussels, Belgium: International Diabetes Federation* 2013.

5. American Diabetes Association. 7. Approaches to Glycemic Treatment. *Diabetes Care* 2015; 38(Supplement 1): S41-S8.

6. Wallia A, Molitch ME. Insulin therapy for type 2 diabetes mellitus. *JAMA* 2014; 311(22): 2315-25.

7. Turner RC, Cull CA, Frighi V, Holman RR, Group UKPDS. Glycemic control with diet, sulfonylurea, metformin, or insulin in patients with type 2 diabetes mellitus: progressive requirement for multiple therapies (UKPDS 49). *JAMA* 1999; 281(21): 2005-12.

8. Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. *Lancet* 1999; 353(9147): 89-92.

9. Feuer EJ, Wun L-M, Boring CC, Flanders WD, Timmel MJ, Tong T. The lifetime risk of developing breast cancer. *J Natl Cancer Inst* 1993; 85(11): 892-7.

10. Narayan KMV, Boyle JP, Thompson TJ, Sorensen SW, Williamson DF. Lifetime risk for diabetes mellitus in the United States. *JAMA* 2003; 290(14): 1884-90.

11. Gregg EW, Zhuo X, Cheng YJ, Albright AL, Narayan KM, Thompson TJ. Trends in lifetime risk and years of life lost due to diabetes in the USA, 1985-2011: a modelling study. *Lancet Diabetes Endocrinol* 2014; 2(11): 867-74.

12. Magliano DJ, Shaw JE, Shortreed SM, et al. Lifetime risk and projected population prevalence of diabetes. *Diabetologia* 2008; 51(12): 2179-86.

13. Hofman A, Murad SD, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol* 2013; 28(11): 889-926.

14. Leening MJG, Kavousi M, Heeringa J, et al. Methods of data collection and definitions of cardiac outcomes in the Rotterdam Study. *Eur J Epidemiol* 2012; 27(3): 173-85.

15. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. *Geneva: World Health Organization* 2006: 1-50.

16. World Health Organization Expert Consultation. Waist circumference and waist-hip ratio. 2011.

17. Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *J Hyg* 1949; 47(2): 188-9.

18. Allignol A, Schumacher M, Beyersmann J. Empirical transition matrix of multistate models: the etm package. *J Stat Software* 2011; 38(4): 1-15.

19. Therneau TM. Modeling survival data: extending the Cox model: Springer; 2000.

20. Levitan EB, Song Y, Ford ES, Liu S. Is nondiabetic hyperglycemia a risk factor for cardiovascular disease?: A meta-analysis of prospective studies. *Arch Intern Med* 2004; 164(19): 2147-55.

21. Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* 2010; 375(9733): 2215-22.

22. Seshasai SRK, Kaptoge S, Thompson A, et al. Diabetes mellitus, fasting glucose, and risk of cause-specific death. *N Engl J Med* 2011; 364(9): 829.

23. Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002; 346(6): 393.

24. Tuomilehto J, Lindström J, Eriksson JG, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 2001; 344(18): 1343-50.

25. Dream Trial Investigators. Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. *Lancet* 2006; 368(9541): 1096-105.

26. Diabetes Prevention Program Research Group. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* 2009; 374(9702): 1677-86.

27. Centers for Disease Control and Prevention. National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. *Atlanta, GA: US Department of Health and Human Services* 2014.

28. Narayan KM, Boyle JP, Thompson TJ, Gregg EW, Williamson DF. Effect of BMI on lifetime risk for diabetes in the U.S. *Diabetes Care* 2007; 30(6): 1562-6.

29. Nathan DM, Buse JB, Davidson MB, et al. Management of hyperglycemia in type 2 diabetes: A consensus algorithm for the initiation and adjustment of therapy: a consensus statement from the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* 2006; 29(8): 1963-72.

30. Nathan DM, Buse JB, Davidson MB, et al. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* 2009; 32(1): 193-203.

31. Grant RW, Wexler DJ, Watson AJ, et al. How doctors choose medications to treat type 2 diabetes: a national survey of specialists and academic generalists. *Diabetes Care* 2007; 30(6): 1448-53.

32. Leening MJG, Heeringa J, Deckers JW, et al. Healthy volunteer effect and cardiovascular risk. *Epidemiology* 2014; 25(3): 470-1.

Supplementary material

Statistical analysis lifetime risk

We used data from individuals at each age during follow-up that they attained free of the disease (i.e. prediabetes, diabetes or insulin therapy)[1,2]. Individuals who reached to age j free of the disease at some point during follow-up constituted the population at risk for any age j (risk set, Rj). If an individual developed the disease, died or was censored at age j, he or she was removed from the risk set for age j+1 and older. If an individual entered the study at age j+1, he or she was added to the risk set for age j+1.[3] For the lifetime risk at 45, for instance, hazards (hj), age-specific incidences (fj), cumulative incidences (Fj), and survival probabilities (Sj) were calculated according to the standard Kaplan-Meier methods for each age j (assuming F44 = 0 and S44 = 1):

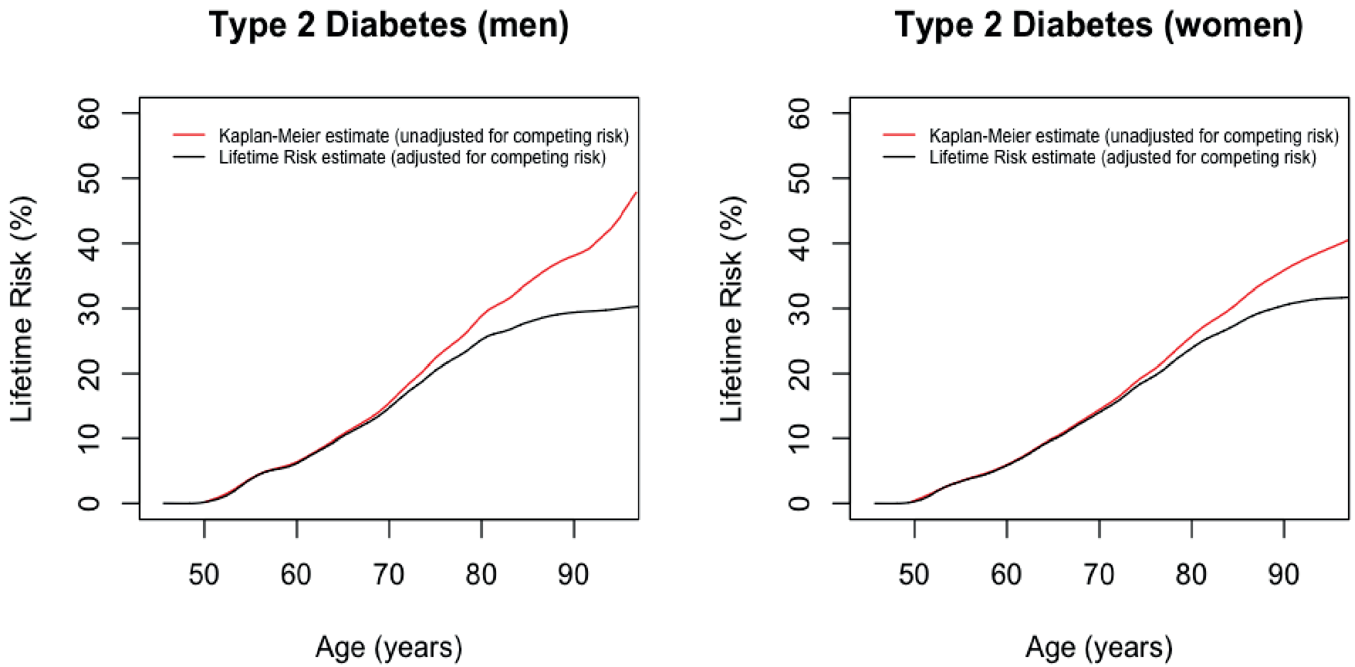
hj = ej / Rj (ej = # of events at age j)
f j = hj × Sj – 1
Fj = ji = 45fi
Sj = 1 – Fj

It should be noted that Fj is the cumulative incidence of the disease (prediabetes, diabetes or insulin therapy) which applies to individuals who survive through age j-1. This cumulative incidence does not take into account the competing risk of death from another cause. This means that individuals that decease count as withdrawals and are assumed to have the same risk of the disease compared to the individuals that are alive at censoring. However, individuals who die before age j have a zero future risk of the disease. This competing risk of death will result in overestimation of the lifetime risk[4]. Therefore, we used a separate survival function (Uj) with death included as an event alongside prediabetes, diabetes or insulin therapy to adjust for the competing risk of death. The adjusted incidence and true lifetime risk was calculated as follows:

fj* = hj × Uj – 1
Fj* = ji = 45fi*
Sj* = 1– Fj*

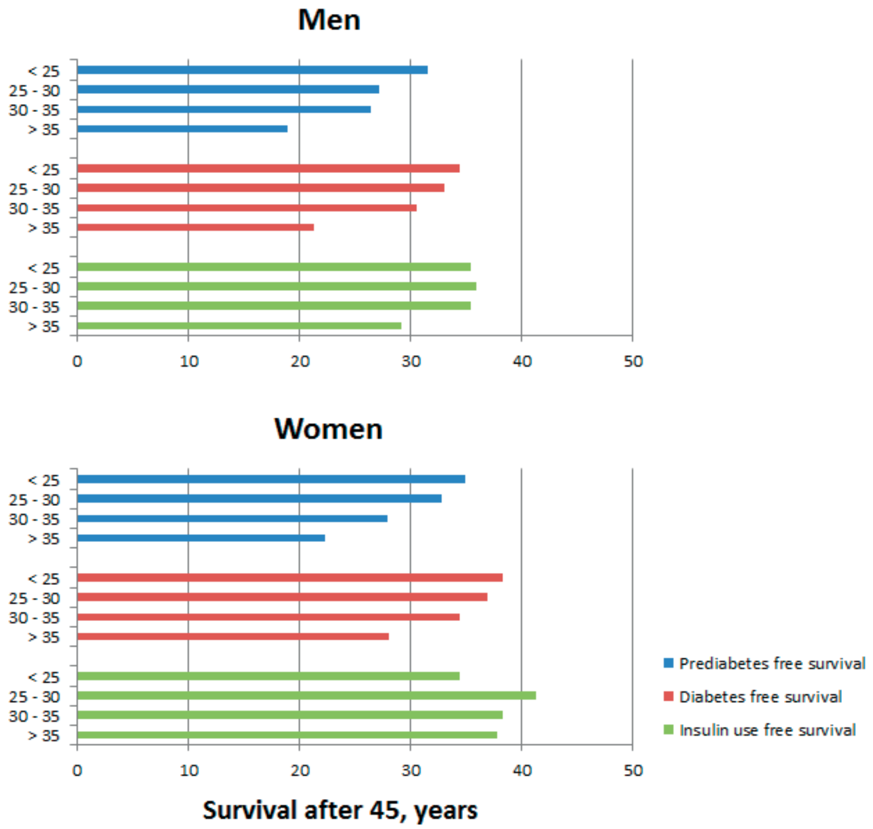
We used similar methods to calculate lifetime risks at the starting age 55, 65, 75 and 85. We set the FT-1 and UT-1 to 0 for every index age T and used the original hazard (hj) to calculate Uj for j T. The analysis methods for prediabetes, diabetes and insulin dependency were similar. Furthermore, we calculated the lifetime risks stratified by body mass index and waist circumference. Finally, we calculated the lifetime risk of diabetes conditional on the presence of prediabetes to study the progression from prediabetes to overt diabetes. Similarly, we studied the lifetime risk of insulin therapy conditional on the presence of diabetes (without insulin therapy) to study the progression from insulin-free diabetes to diabetes with insulin therapy.

Figure S1. Kaplan-Meier Estimate Compared to Competing-Risk Estimate of Lifetime Risk of Type 2 Diabetes.



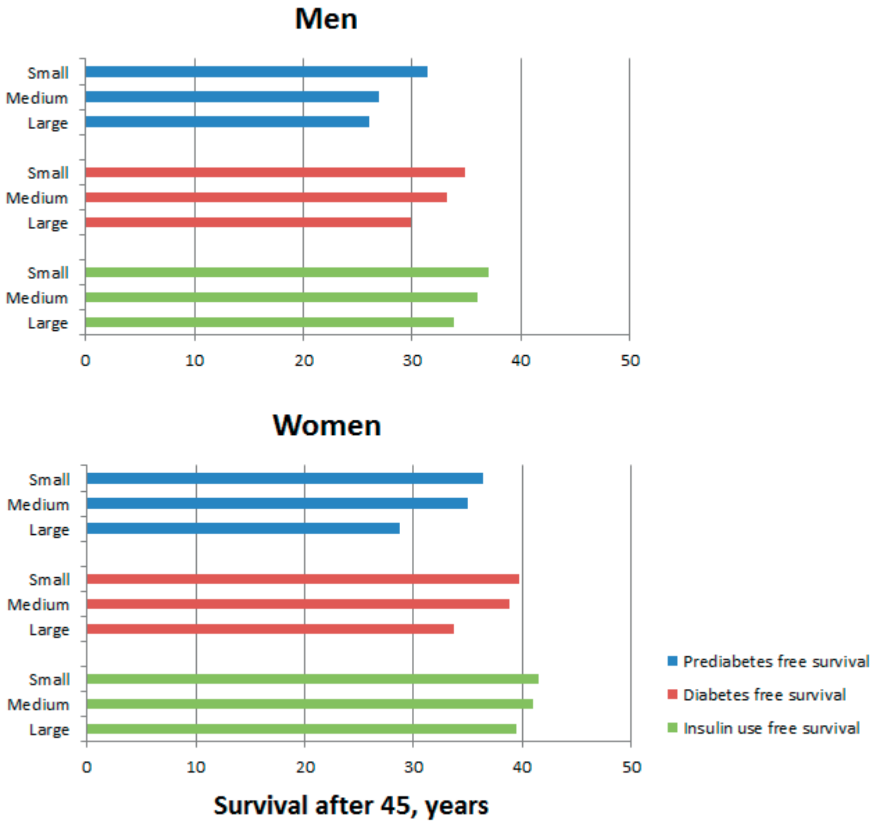
Comparison of the cumulative incidence of type 2 diabetes adjusted for the competing risk of death (Cumulative Incidence estimates) with the Kaplan-Meier estimate of the lifetime risk of type 2 diabetes (unadjusted for the competing risk of death) in men (n = 3741) and women (n = 5103) aged 45.

Figure S2. Survival free of prediabetes, diabetes and insulin use after the age of 45 by body mass index strata.



The blue bars represent the prediabetes-free survival in men and women with normal glucose levels at the age of 45. The red bars represents the diabetes-free survival in men and women without diabetes at the age of 45. Finally, the green bars represent the insulin-free survival in men and women that do not use insulin at the age of 45. All analyses are stratified by body mass index and adjusted for the competing risk of death.

Figure S3. Survival free of prediabetes, diabetes and insulin use after the age of 45 by waist circumference strata.



The blue bars represent the prediabetes-free survival in men and women with normal glucose levels at the age of 45. The red bars represents the diabetes-free survival in men and women without diabetes at the age of 45. Finally, the green bars represent the insulin-free survival in men and women that do not use insulin at the age of 45. All analyses are stratified by waist circumference and adjusted for the competing risk of death.

Table S1: Prevalence of Prediabetes and Diabetes at Baseline by Sex and Age (n=10050).

	Men (n = 4368)		Women (n = 5682)	
Age (years)	**Total N (cases)**	**Prevalence (95%CI)**	**Total N (cases)**	**Prevalence (95%CI)**
Prediabetes				
45-55	622 (58)	0.09 (0.07-0.12)	795 (43)	0.05 (0.04-0.07)
55-65	1824 (318)	0.17 (0.16-0.19)	2311 (279)	0.12 (0.11-0.13)
65-75	1243 (198)	0.16 (0.14-0.18)	1453 (191)	0.13 (0.11-0.15)
75-85	596 (105)	0.18 (0.15-0.21)	928 (146)	0.16 (0.13-0.18)
>85	83 (11)	0.13 (0.07-0.23)	195 (33)	0.17 (0.12-0.23)
Diabetes				
45-55	622 (43)	0.07 (0.05-0.10)	795 (40)	0.05 (0.04-0.07)
55-65	1824 (254)	0.14 (0.13-0.16)	2311 (206)	0.09 (0.08-0.10)
65-75	1243 (195)	0.16 (0.14-0.18)	1453 (169)	0.12 (0.10-0.13)
75-85	596 (118)	0.20 (0.17-0.23)	928 (135)	0.15 (0.12-0.17)
>85	83 (17)	0.20 (0.13-0.31)	195 (29)	0.15 (0.11-0.21)

CI denotes confidence interval.

Table S2: Remaining Lifetime Risk of Prediabetes, Type 2 Diabetes and Insulin Dependent Type 2 Diabetes in Men by Categories of Age, Body Mass Index, and Waist Circumference.

Age (years)	N	Lifetime risk prediabetes (95%CI)	N	Lifetime risk diabetes (95%CI)	N	Lifetime risk insulin use (95%CI)
45	3051	48.4% (44.1-52.6)	3741	30.5% (27.5-33.5)	4282	10.7% (8.7-12.7)
55	2835	42.6% (39.5-45.8)	3505	28.1% (25.4-30.7)	4045	10.3% (8.4-12.2)
65	2070	35.1% (32.2-38.1)	2619	24.1% (21.6-26.6)	3085	7.8% (6.3-9.3)
75	1228	24.5% (21.2-27.7)	1557	16.0% (13.4-18.6)	1887	5.3% (3.9-6.8)
85	346	12.9% (8.2-17.5)	463	8.3% (4.9-11.8)	584	4.0% (2.0-6.0)
BMI (kg/m^2)						
< 25	1047	37.5% (31.6-43.3)	1174	19.6% (15.3-23.9)	1273	6.9% (3.5-10.3)
25-30	1596	51.2% (43.6-58.8)	1979	31.5% (27.3-35.7)	2254	11.1% (8.3-13.8)
30-35	346	58.4% (49.1-67.8)	492	42.0% (33.7-50.3)	614	16.5% (10.3-22.7)
>35	41	72.4% (48.6-96.2)	69	59.7% (38.4-81.0)	110	16.8% (3.8-29.8)
Waist circumference						
Small	1129	40.2% (34.3-46.1)	1266	21.8% (17.0-26.6)	1357	5.0% (2.6-7.4)
Medium	1005	52.1% (41.6-62.6)	1237	29.0% (23.5-34.6)	1392	9.6% (6.2-12.9)
Large	797	53.4% (46.7-60.1)	1087	39.4% (33.8-44.9)	1359	15.3% (11.2-19.3)

*Waist circumference categories small, medium and large represent the WHO classification scheme (for men: <94 cm, 94-102 cm and ≥102 cm). BMI denotes body mass index. The lifetime risk of prediabetes is for individuals with normal glucose levels at the index age, lifetime risk of diabetes for individuals without diabetes at the index age and the lifetime risk of insulin use for individuals free of insulin use at the index age. For BMI and waist circumference, the lifetime risks are subsequent-the age of 45.

Table S3: Remaining Lifetime Risk of Prediabetes, Type 2 Diabetes and Insulin Dependent Type 2 Diabetes in Women by Categories of Age, Body Mass Index, and Waist Circumference.

Age (years)	N	Lifetime risk prediabetes (95%CI)	N	Lifetime risk diabetes (95%CI)	N	Lifetime risk insulin use (95%CI)
45	4411	49.1% (46.1-52.0)	5103	31.9% (29.3-34.6)	5605	7.8% (6.2-9.3)
55	4104	46.1% (43.4-48.8)	4786	30.1% (27.7-32.5)	5284	7.5% (6.1-9.0)
65	3039	39.4% (36.8-42.1)	3638	25.3% (23.1-27.6)	4094	6.3% (5.0-7.6)
75	1845	26.8% (24.0-29.6)	2293	18.5% (16.2-20.7)	2660	4.2% (3.0-5.3)
85	726	13.2% (10.0-16.4)	942	9.2% (6.8-11.6)	1141	2.9% (1.6-4.2)
BMI (kg/m^2)						
< 25	1639	36.6% (31.6-41.5)	1781	18.0% (14.0-22.0)	1851	3.2% (1.6-4.8)
25-30	1856	53.5% (49.3-57.7)	2153	34.8% (30.7-38.8)	2344	8.0% (5.4-10.7)
30-35	667	61.3% (42.8-69.8)	832	43.4% (36.7-50.1)	999	13.0% (8.3-17.6)
>35	199	71.0% (58.4-83.6)	283	55.9% (44.6-67.3)	352	16.1% (7.8-24.4)
Waist circumference						
Small	998	33.9% (27.8-39.9)	1063	16.0% (10.8-21.1)	1092	4.1% (1.3-7.0)
Medium	1215	42.4% (36.8-48.0)	1352	21.6% (17.1-26.1)	1423	2.9% (0.8-5.0)
Large	2004	59.7% (55.4-64.1)	2466	42.8% (38.8-46.8)	2846	11.4% (8.8-13.9)

*Waist circumference categories small, medium and large represent the WHO classification scheme (for women: <80 cm, 80-88 cm and ≥88 cm). BMI denotes body mass index. The lifetime risk of prediabetes is for individuals with normal glucose levels at the index age, lifetime risk of diabetes for individuals without diabetes at the index age and the lifetime risk of insulin use for individuals free of insulin use at the index age. For BMI and waist circumference, the lifetime risks are subsequent-the age of 45.

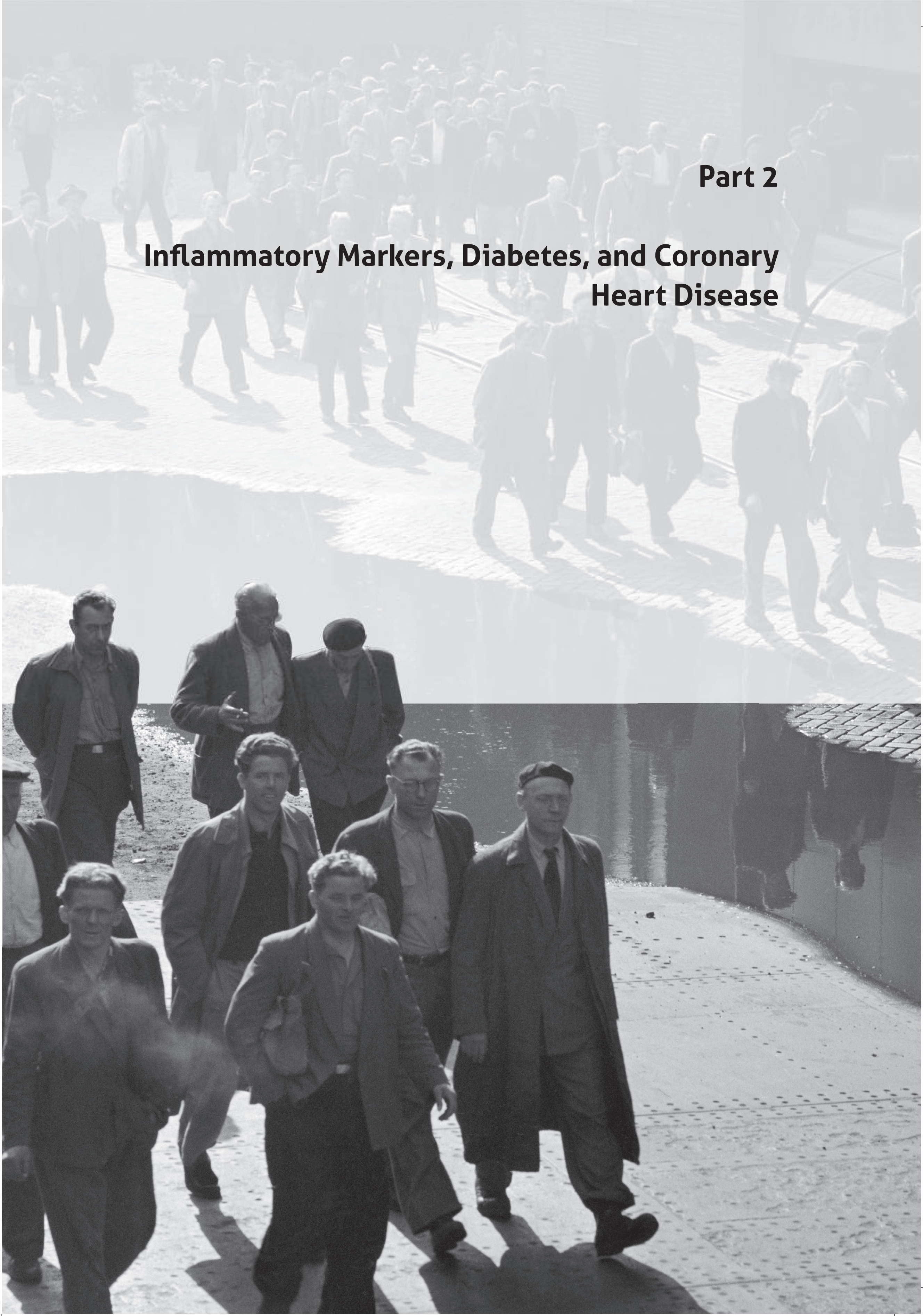
Table S4: Remaining Lifetime Risk of Diabetes at Age 45 by Body Mass Index and Waist Circumference.

	Body mass index (kg/m^2)			
Waist circumference	<25	25-30	30-35	>35
Small	18.4% (14.3-22.5)	22.8% (15.9-29.8)	N/A	N/A
Medium	18.5% (13.5-23.5)	28.5% (23.9-33.2)	27.6% (10.2-44.9)	N/A
Large	17.7% (10.5-24.9)	39.0% (34.6-43.4)	44.6% (39.0-50.1)	57.1% (46.8-67.3)

* Waist circumference categories small, medium and large represent the WHO classification scheme (for men: <94 cm, 94-102 cm and ≥102 cm; for women: <80 cm, 80-88 cm and ≥88 cm).

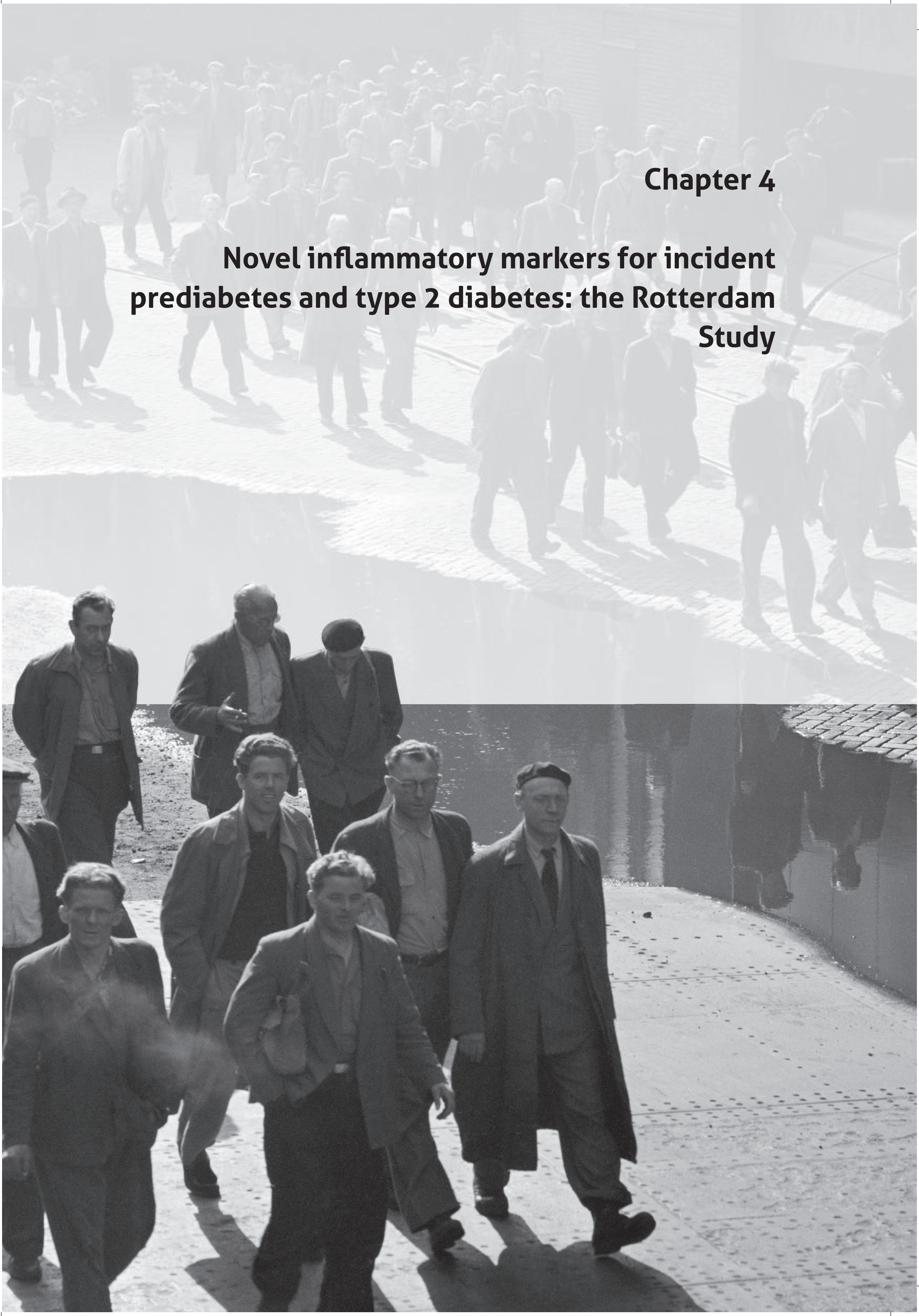
Part 2

Inflammatory Markers, Diabetes, and Coronary Heart Disease



Chapter 4

Novel inflammatory markers for incident prediabetes and type 2 diabetes: the Rotterdam Study



Background: The immune response involved in each phase of type 2 diabetes (T2D) development might be different. We aimed to identify novel inflammatory markers that predict progression from normoglycemia to prediabetes, incident T2D and insulin therapy.

Methods: We used plasma levels of 26 inflammatory markers in 971 subjects from the Rotterdam Study. Among them 17 are novel and 9 previously studied. Cox regression models were built to perform survival analysis.

Main Outcome Measures: During a follow-up of up to 14.7 years (between April 1, 1997, and Jan 1, 2012) 139 cases of prediabetes, 110 cases of T2D and 26 cases of insulin initiation were identified.

Results: In age and sex adjusted Cox models, IL13 (HR = 0.78), EN-RAGE (1.30), CFH (1.24), IL18 (1.22) and CRP (1.32) were associated with incident prediabetes. IL13 (0.62), IL17 (0.75), EN-RAGE (1.25), complement 3 (1.44), IL18 (1.35), TNFRII (1.27), IL1ra (1.24) and CRP (1.64) were associated with incident T2D. In multivariate models, IL13 (0.77), EN-RAGE (1.23) and CRP (1.26) remained associated with prediabetes. IL13 (0.67), IL17 (0.76) and CRP (1.32) remained associated with T2D. IL13 (0.55) was the only marker associated with initiation of insulin therapy in diabetics.

Conclusion: Various inflammatory markers are associated with progression from normoglycemia to prediabetes (IL13, EN-RAGE, CRP), T2D (IL13, IL17, CRP) or insulin therapy start (IL13). Among them, EN-RAGE is a novel inflammatory marker for prediabetes, IL17 for incident T2D and IL13 for prediabetes, incident T2D and insulin therapy start.

Introduction

There is increasing evidence that inflammation plays a role in the development of type 2 diabetes mellitus (DM)[1,2,3]. In this context, the identification of novel inflammatory markers associated with the risk of type 2 DM will shed light on the pathophysiology of the disease and might also help clinicians to target individuals at highest risk[4,5]. So far, a limited number of inflammatory markers have been investigated. Previous studies reported inflammatory markers including C-reactive protein (CRP), interleukin 6 (IL6) and adiponectin to associate with the risk of type 2 DM[6,7,8,9,10,11]. These studies merely investigated inflammatory markers that predict the conversion from normoglycemia to type 2 DM.

Healthy individuals are thought to experience a prediabetes phase before developing type 2 DM. Prediabetes is the presence of blood glucose levels higher than normal, but not yet high enough to be classified as diabetes[12]. Moreover, type 2 DM could further deteriorate to a stage, where glucose control is only possible by insulin therapy[12,13]. Progression from normoglycemia to prediabetes is thought to be driven by insulin resistance, while progression to type 2 DM and need for insulin therapy is further affected by beta cell dysfunction[14,15,16]. Therefore, the immune response involved in each of these phases might be different[17].

We hypothesized that inflammatory markers are phase-specific for conversion from normoglycemia to prediabetes, diabetes and need for insulin therapy. We agnostically studied the association of a set of inflammatory markers with progression from normoglycemia to prediabetes, type 2 DM and finally to insulin therapy.

Materials and Methods

Study population

The Rotterdam Study is a prospective population-based cohort study in Ommoord, a district of Rotterdam, the Netherlands. The design of the Rotterdam Study has been described in more detail elsewhere[18]. Briefly, in 1989 all residents within the well-defined study area aged 55 years or older were invited to participate of whom 78% (7983 out of 10275) agreed. There were no other eligibility criteria to enter the Rotterdam Study except minimum age and residency are based on ZIP code. The first examination took place from 1990 to 1993, after which follow-up examinations were conducted every 3-5 years. This study was based on data collected during the third visit(1997-1999). We used data from 971 individuals with available data on inflammatory markers, drawn as a random control sample in a case-cohort study of markers for dementia. The Rotterdam Study has been approved by the medical ethics committee according to the Population Screening Act: Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of Netherlands. All participants in the present analysis provided written informed consent to participate and to obtain information from their treating physicians.

Measurement of inflammatory markers

Fasting blood samples were collected at the research centre. Plasma was isolated and immediately put on ice and stored at -80°C. Citrate plasma (200Ul) was sent in July 2008 to Rules-Based Medicine, Austin, Texas (www.myriadrbm.com). The samples were thawed at room temperature, vortexed, spun at 4000 RPM for 5 minutes for clarification and volume was removed for MAP analysis into a master microtiter plate. Using automated pipetting, an aliquot of each sample was introduced into one of the capture microsphere multiplexes of the Multi Analyte Profile. The mixture of sample and capture microspheres were thoroughly mixed and incubated at room temperature for 1 hour. Multiplexed cocktails of biotinylated, reporter antibodies for each multiplex were then added robotically and after thorough mixing, were incubated for an additional hour at room temperature. Multiplexes were developed using an excess of streptavidin-phycoerythrin solution which was thoroughly mixed into each multiplex and incubated for 1 hour at room temperature. The volume of each multiplexed reaction was reduced by vacuum filtration and the volume increased by dilution into matrix buffer for analysis. Analysis was performed in a Luminex 100 instrument and the resulting data stream was interpreted using proprietary data analysis software developed at Rules-Based Medicine (https://myriadrbm.com/scientific-media/quality-control-systems-white-paper/). For each multiplex, both calibrators and controls were included on each microtiter plate. 8-point calibrators were run in the first and last column of each plate and 3-level controls were included in duplicate. Testing results were determined first for the high, medium and low controls for each multiplex to ensure proper assay performance. Unknown values for each of the analytes localized in a specific multiplex were determined using 4 and 5 parameter, weighted and non-weighted curve fitting algorithms included in the data analysis package.

Fifty inflammatory markers were quantified using multiplex immunoassay on a custom designed human multi-analyte profile. The intra-assay variability was less than 4% and the inter assay variability was less than 13%. Markers with more than 60% completeness of measurements were selected for analysis (26 from 50)[19].

Type 2 diabetes mellitus diagnosis

The participants were followed from the date of baseline center visit onwards. At baseline and during follow-up, cases of prediabetes and type 2 DM were ascertained through active follow-up using general practitioners' records, hospital discharge letters and glucose measurements from Rotterdam Study visits which take place approximately every 4 years[20]. Diabetes, prediabetes and normoglycemia were defined according to the current WHO guidelines. Normoglycemia was defined as a fasting blood glucose level < 6.0 mmol/L; prediabetes was defined as a fasting blood glucose between 6.0 mmol/L and 7.0 mmol/L or a non-fasting blood glucose between 7.7 mmol/L and 11.1 mmol/L (when fasting samples were unavailable); type 2 diabetes was defined as a fasting blood glucose ≥ 7.0 mmol/L, a non-fasting blood glucose ≥ 11.1 mmol/L (when fasting samples were unavailable), or the use of blood glucose lowering medication[20]. Information regarding the use of blood glucose lowering medication was derived from both structured home interviews and linkage to pharmacy dispensing records. At baseline, more than 95% of the Rotterdam Study population was covered by the pharmacies in the study area. All potential events of prediabetes and type 2 diabetes were independently adjudicated by two study physicians. In case of disagreement, consensus was sought with an endocrinologist. Follow-up data was complete until January 1st 2012, calculated as a separate variable for every outcome, taking in account the hierarchy of events as follows: prediabetes, type 2 diabetes, insulin therapy start[20].

Covariates

Height and weight were measured with the participants standing without shoes and heavy outer garments. Body mass index (BMI) was calculated as weight divided by height squared (kg/m^2). Waist circumference was measured at the level midway between the lower rib margin and the iliac crest with participants in standing position without heavy outer garments and with emptied pockets, breathing out gently. Blood pressure was measured at the right brachial artery with a random-zero sphygmomanometer with the participant in sitting position, and the mean of 2 consecutive measurements was used. Information on medication use, medical history and smoking behaviour was collected via computerized questionnaires during home visits. Smoking was classified as current versus non-current smokers. Participants were asked whether they were currently smoking cigarettes, cigars, or pipes. History of cardiovascular disease was defined as a history of coronary heart diseases (myocardial infarction, revascularization, coronary artery bypass graft surgery or percutaneous coronary intervention) and was verified from the medical records of the general practitioner. Alcohol intake was assessed in grams of ethanol per day. Insulin, glucose, total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG) were measured on the COBAS 8000 Modular Analyzer (Roche Diagnostics GmbH). The corresponding interassay coefficients of variations are the following: insulin <8%, glucose <1.4%, lipids <2.1%. HOMA-IR (the homeostatic model assessment to quantify insulin resistance) was calculated dividing the product of fasting glucose (in mmol/L) and fasting insulin (in mU/L) by 22.5. HOMA-B (the homeostatic model assessment of β-cell function) was calculated dividing the product of fasting insulin (in mU/L) and 20 by the difference of glucose (in mmol/L) with 3.5[21].

Statistical analyses

We used linear regression to investigate the association between each inflammatory marker and fasting glucose and fasting insulin in 851 subjects free of diabetes at baseline (excluding 120 prevalent diabetes cases from 971 subjects with available data) as presented at Figure 1.1, Figure 1.2, table S1, table S2. Also the associations between markers with HOMA-IR and HOMA-B were investigated using linear regression (table S3). Markers with a right-skewed distribution were transformed to the natural logarithmic scale (including fasting glucose and insulin). For a better comparison between the inflammatory markers, all markers were standardized by dividing the measured value by the standard deviation. We defined marker values as an outlier when the value was >4 standard deviations higher or lower than the mean of the normal variable (not natural log transformed). Participants were excluded from the analyses when the marker value for this person was an outlier. A multiple imputation procedure was used for missing covariates (n=5 imputations). The analyses with incident prediabetes, incident type 2 DM and need for insulin therapy were performed using Cox proportional hazard models to calculate hazard ratios (HRs) and 95% confidence intervals (CI). The first model with incident prediabetes and diabetes was adjusted for age and sex (table 2). Significant markers were further investigated in multivariable models (table 3). In the second model, we additionally adjusted for body mass index, waist circumference, total cholesterol, HDL-cholesterol, medication for hypertension, smoking, prevalent cardiovascular disease and lipid lowering medication. In the third model we additionally adjusted for C-reaction protein (CRP) levels (except for CRP marker). We sought to investigate the associations between the inflammatory markers and the need for insulin therapy in 115 prevalent diabetes cases with no prevalent use of insulin at baseline (from 120 prevalent cases in total). The inflammatory markers were not correlated to each other, representing 26 independent variables. As a sensitivity analysis, to identify the most robust findings in every analysis, we applied a Bonferroni corrected p-value of 1.9×10^{-3} (0.05/26 markers).The analyses were performed using IBM SPSS Statistics for Windows (IBM SPSS Statistics for Windows, Armonk, New York: IBM Corp) and R V.3.0.1 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Table 1 summarizes the baseline characteristics of 971 participants, including 120 prevalent diabetes cases. The mean (SD) age at baseline was 73.0 (7.5) years and 44.8% of our population sample were males. The mean BMI (SD) was 26.7 (3.9) kg/m^2 and 12.6% of the study population used statin. Baseline levels of inflammation markers are presented in table S4.

Table 1. **Baseline characteristics of study participants.**

Characteristic	P-value
Total population number	971
Age, years	73.0±7.5
Men, n (%)	435.0 (44.8)
Waist Circumference, m	0.9±0.1
Body mass index, kg/m^2	26.7±3.9
Systolic blood pressure, mmHg	144.0±21.7
Diastolic blood pressure, mmHg	75.0±11.0
Hypertension medication with indication, n(%)	744.0 (76.6)
Total cholesterol, mmol/L	5.8±1.0
HDL cholesterol, mmol/L	1.4±0.4
Fasting glucose, mmol/L	5.6 (3.54)
Fasting insulin, uIU/L	9.4 (19.87)
Current smokers, n (%)	137.0 (14.1)
Former smokers, n (%)	483.0 (49.7)
Prevalent CVD, n (%)	201.0 (20.7)
Alcohol intake in drinkers (76%), g/day	5.71 (42.73)
Lipid lowering medication, n (%)	122.0 (12.6)

Abbreviations: HDL, high density lipoproteins; CVD, cardiovascular disease.
[*]Plus-minus values are means ± standard deviation or median (inter-quartile range).

Cross-sectional analysis

Figure 1.1 and 1.2 present the multivariable adjusted associations between the inflammatory markers and fasting glucose, fasting insulin in 851 subjects free of diabetes at baseline. Three markers, EN-RAGE, IL13 and sRAGE were significantly associated with fasting glucose. CD40, EN-RAGE, FAS, HCC4, IL13, IL18, TRAILR3, CFH, complement 3, IL18 and IL1ra were significantly associated with fasting insulin.

Prospective analyses

During a median follow-up of 9.5 years in 698 subjects free of prediabetes at baseline, 139 cases of prediabetes were identified (21 prediabetes cases per 1000 person-years). table S1.1 presents baseline characteristics among prediabetes cases and non-cases. In age and sex adjusted model, EN-RAGE, IL13, CFH, IL18 and CRP were associated with incident prediabetes (table 2). In multivariate models, IL13 (HR=0.77), EN- RAGE (HR=1.23) and CRP (HR=1.26) remained associated with incident prediabetes (table 3).

Table 2. **Age and sex-adjusted associations between markers and incident prediabetes and incident overt diabetes mellitus.**

Marker	Incident prediabetes		Incident diabetes	
	HR(95%CI)	P-value	HR(95%CI)	P-value
CD40, ng/mL	0.93 (0.72-1.19)	0.5	1.18 (0.91-1.52)	0.2
CD40Ligand*, ng/mL	0.95 (0.79-1.16)	0.6	1.06 (0.85-1.32)	0.6
EN-RAGE*, ng/mL	1.30 (1.08-1.56)	5.0×10^{-3}	1.25 (1.01-1.54)	4.0×10^{-2}
Eotaxin*, pg/mL	0.95 (0.79-1.15)	0.6	0.98 (0.80-1.21)	0.8
FAS*, ng/mL	1.09 (0.88-1.35)	0.4	1.09 (0.87-1.38)	0.4
HCC4, ng/mL	1.11 (0.90-1.35)	0.3	1.24 (0.99-1.53)	5.4×10^{-2}
IL13*, pg/mL	0.78 (0.64-0.94)	8.0×10^{-3}	0.62 (0.50-0.76)	5.0×10^{-6}
IL16, pg/mL	1.07 (0.88-1.29)	0.4	1.17 (0.94-1.45)	0.1
IL17*, pg/mL	0.97 (0.81-1.16)	0.7	0.75 (0.62-0.91)	3.0×10^{-3}
IL8*, pg/mL	1.05 (0.87-1.27)	0.5	1.19 (0.97-1.47)	0.1
MDC, pg/mL	0.94 (0.75-1.17)	0.5	1.19 (0.94-1.50)	0.1
MIP1 alpha*, pg/mL	1.09 (0.90-1.32)	0.3	1.08 (0.87-1.34)	0.5
MIP1 beta*, pg/mL	1.05 (0.87-1.26)	0.6	1.00 (0.81-1.25)	0.9
PARC, ng/mL	1.08 (0.88-1.32)	0.4	0.94 (0.75-1.19)	0.6
sRAGE*, ng/mL	0.95 (0.79-1.14)	0.6	0.91 (0.75-1.11)	0.3
TRAILR3*, ng/mL	1.18 (0.98-1.41)	8.1×10^{-2}	1.19 (0.97-1.47)	9.1×10^{-2}
CFH*, ug/mL	1.24 (1.02-1.49)	2.8×10^{-2}	1.05 (0.87-1.28)	0.6
C3*, mg/mL	1.13 (0.94-1.36)	0.1	1.44 (1.17-1.77)	1.0×10^{-3}
IL18*, pg/mL	1.22 (1.02-1.47)	3.2×10^{-2}	1.35 (1.10-1.65)	4.0×10^{-3}
MCP1*, pg/mL	0.93 (0.76-1.14)	0.4	0.99 (0.79-1.25)	0.9
MIF*, ng/mL	0.97 (0.82-1.14)	0.6	1.11 (0.92-1.35)	0.2
RANTES*, ng/mL	0.89 (0.75-1.05)	0.1	1.05 (0.87-1.27)	0.6
Resistin*, ng/mL	1.02 (0.85-1.24)	0.8	0.96 (0.78-1.18)	0.7
TNFRII*, ng/mL	0.97 (0.79-1.18)	0.7	1.27 (1.03-1.58)	2.9×10^{-2}
Il1ra*, pg/mL	1.04 (0.87-1.25)	0.6	1.24 (1.02-1.51)	3.4×10^{-2}
CRP*, ug/mL	1.32 (1.10-1.58)	3.0×10^{-3}	1.64 (1.33-2.02)	4.0×10^{-6}

*Naturally log-transformed. CD40 denotes cluster of differentiation 40, CD40 ligand cluster of differentiation 40 ligand, EN-RAGE Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein, FAS Fas Cell Surface Death Receptor, HCC4 Human CC chemokine-4, IL13 interleukin 13, IL16 interleukin 16, IL17 interleukin 17, IL8 interleukin 8, MDC Monocyte Derived Chemokine, MIP1alpha Macrophage Inflammatory Protein 1 alpha, MIP1beta Macrophage Inflammatory Protein 1 beta, PARC Pulmonary and Activation-Regulated Chemokine, sRage Soluble Receptor of Advanced Glycation End-products, TRAILR3 Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor 3, CFH Complement Factor H, C3 complement 3, IL18 interleukin 18, MCP1 Monocyte Chemotactic Protein 1, RANTES Regulated Upon Activation, Normally T-Expressed, And Presumably Secreted, TNFR-II Tumor Necrosis Factor Receptor 2, IL1ra, Interleukin 1 Receptor Antagonist, CRP C-Reactive Protein.

Table 3. **Multivariable-adjusted associations between markers and incident prediabetes, incident type 2 diabetes mellitus.**

	Incident prediabetes		Incident type 2 diabetes	
Marker	**HR(95%CI)**	**P-value**	**HR(95%CI)**	**P-value**
Interleukin 13				
Model 1	0.78 (0.64-0.94)	8.0×10^{-3}	0.62 (0.50-0.76)	5.0×10^{-6}
Model 2	0.78 (0.63-0.98)	2.9×10^{-2}	0.68 (0.53-0.88)	4.0×10^{-3}
Model 3	0.77 (0.62-0.96)	2.2×10^{-2}	0.67 (0.52-0.86)	2.0×10^{-3}
Interleukin 17				
Model 1	0.97 (0.81-1.16)	0.7	0.75 (0.62-0.91)	3.0×10^{-3}
Model 2	0.97 (0.82-1.16)	0.7	0.75 (0.62-0.91)	4.0×10^{-3}
Model 3	0.98 (0.82-1.17)	0.8	0.76 (0.63-0.93)	7.0×10^{-3}
EN-RAGE				
Model 1	1.30 (1.08-1.56)	5.0×10^{-3}	1.25 (1.01-1.54)	4.0×10^{-2}
Model 2	1.28 (1.06-1.56)	1.2×10^{-2}	1.13 (0.89-1.41)	0.3
Model 3	1.23 (1.01-1.51)	4.1×10^{-2}	1.05 (0.83-1.32)	0.6
Complement 3				
Model 1	1.13 (0.94-1.36)	0.1	1.44 (1.17-1.77)	1.0×10^{-3}
Model 2	1.05 (0.86-1.27)	0.6	1.19 (0.96-1.49)	0.1
Model 3	0.99 (0.82-1.21)	0.9	1.10 (0.87-1.39)	0.4
Complement factor H				
Model 1	1.24 (1.02-1.49)	2.8×10^{-2}	1.05 (0.87-1.28)	0.6
Model 2	1.19 (0.99-1.45)	6.5×10^{-2}	0.98 (0.81-1.18)	0.8
Model 3	1.18 (0.97-1.42)	9.7×10^{-2}	0.98 (0.81-1.18)	0.8
Interleukin 18				
Model 1	1.22 (1.02-1.47)	3.2×10^{-2}	1.35 (1.10-1.65)	4.0×10^{-3}
Model 2	1.17 (0.97-1.41)	0.1	1.22 (0.98-1.50)	6.7×10^{-2}
Model 3	1.13 (0.94-1.36)	0.1	1.18 (0.96-1.46)	0.1
TNF-Receptor II				
Model 1	0.97 (0.79-1.18)	0.7	1.27 (1.03-1.58)	2.9×10^{-2}
Model 2	0.89 (0.73-1.09)	0.2	1.08 (0.86-1.37)	0.5
Model 3	0.81 (0.66-1.01)	6.1×10^{-2}	0.99 (0.78-1.28)	0.9
Interleukin 1ra				
Model 1	1.04 (0.87-1.25)	0.6	1.24 (1.02-1.51)	3.4×10^{-2}
Model 2	0.97 (0.80-1.17)	0.7	1.03 (0.83-1.27)	0.8
Model 3	0.94 (0.78-1.14)	0.5	0.98 (0.79-1.22)	0.8
C-reactive protein				
Model 1	1.32 (1.10-1.58)	3.0×10^{-3}	1.64 (1.33-2.02)	4.0×10^{-6}
Model 2	1.26 (1.04-1.53)	1.8×10^{-2}	1.32 (1.05-1.67)	1.7×10^{-2}
Model 3	NA	NA	NA	NA

Model 1 is adjusted for age and sex. Model 2 is additionally adjusted for BMI, waist circumference (WC), Total Cholesterol, HDL, medication for hypertension, smoking, prevalent CVD, and lipid lowering medication. Model 3 is additionally adjusted for CRP. The p-values are bold when they are less than or equal to the significance level cut-off of 0.05.

During a median follow-up of 12.1 years in 851 subjects free of diabetes at baseline, 110 cases of incident type 2 diabetes were identified (11 diabetes cases per 1000 person-years). table S1.2 presents baseline characteristics among diabetes cases and non-cases. In age and sex adjusted model, EN-RAGE, IL13, IL17, complement 3, IL18, TNFRII, IL1ra and CRP were associated with incident type 2 diabetes (table 2). In multivariate models, IL13 (HR=0.67), IL17 (HR=0.76) and CRP (HR=1.32) remained associated with incident type 2 diabetes (table 3).

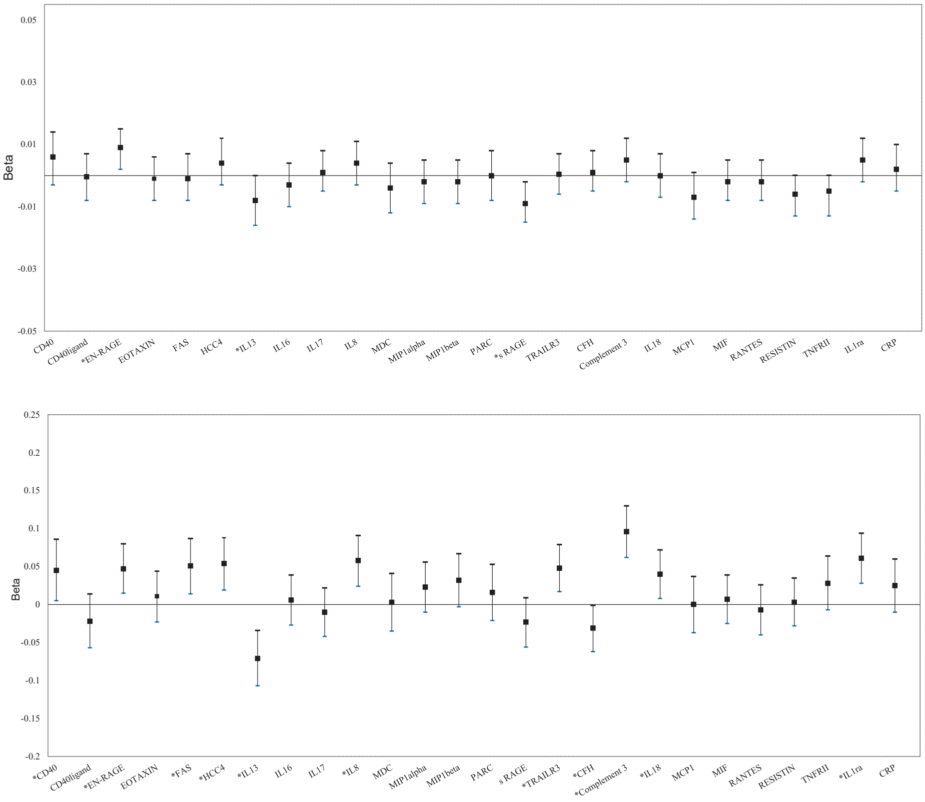
During a median follow-up of 7.5 years in 115 prevalent diabetics free of insulin at baseline, 26 started insulin therapy (30 insulin starters per 1000 person-years). Table S3 presents baseline characteristics among insulin starters and non-starters. The only marker associated with need for insulin therapy was IL13. In age and sex adjusted model, the risk for insulin therapy start was 45% lower per standard deviation increase in the natural log-transformed IL13 (HR=0.55, 95% CI: 0.34, 0.90), (Table S4). The association between 1L13 and initiation of insulin therapy remained significant after further adjustment for BMI, waist circumference, total cholesterol, HDL, medication for hypertension, smoking, prevalent CVD, lipid lowering medication (HR=0.49, 95% CI: 0.28, 0.91).

Discussion

Although a sizable number of studies have documented the association of inflammatory markers with type 2 DM, most of them investigated the risk to become diabetic, but not the risk of prediabetes and insulin therapy start[8]. In this study we investigated a wide range of inflammatory markers for phase-specific prediction of progression to type 2 DM and identified EN-RAGE, IL13 and IL17 as novel inflammatory markers. Higher EN-RAGE levels were associated with an increased risk of incident prediabetes, whereas higher IL13 levels were associated with a decreased risk of prediabetes, incident type 2 DM and need for insulin therapy. Higher IL17 levels were associated with a decreased risk of incident type 2 DM. In addition, this study reconfirm the previously found associations between high CRP levels and the increased risk for type 2 diabetes[6,7,8].

EN-RAGE, also known as S100A12 or Calgranulin C, is a calcium-binding pro inflammatory protein mainly secreted by granulocytes. The best known target protein of EN-RAGE are RAGE (Receptor for Advanced Glycation Endproducts)[22] and TLR4 (Toll-like receptor 4)[23]. Ligation of EN-RAGE with RAGE or TLR4, which are both gatekeepers of the innate immune system, activates inflammatory cascades, including the NF-κB pathway and JNK (c-Jun NH $_2$ –terminal kinase)[24]. NF-κB and JNK are both signaling pathways involved in the pathogenesis of insulin resistance and type 2 DM[25]. EN-RAGE is positively associated with chronic inflammatory disorders such as inflammatory bowel disease, chronic kidney disease,

Figure 1 Associations of inflammatory markers with fasting glucose (upper panel) or fasting insulin (lower panel).



CD40, cluster of differentiation 40; CD40 ligand, cluster of differentiation 40 ligand ; EN-RAGE, Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein; FAS, Fas Cell Surface Death Receptor; HCC4, Human CC chemokine-4; IL13, interleukin 13; IL16, interleukin 16; IL17, interleukin 17; IL8, interleukin 8; MDC, Monocyte Derived Chemokine; MIP1alpha, Macrophage Inflammatory Protein 1 alpha; MIP1beta, Macrophage Inflammatory Protein 1 beta; PARC, Pulmonary and Activation-Regulated Chemokine; sRage, Soluble Receptor of Advanced Glycation End-products; TRAILR3, Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor 3; CFH, Complement Factor H; IL18, interleukin 18; MCP1, Monocyte Chemotactic Protein 1; RANTES, Regulated Upon Activation, Normally T-Expressed, And Presumably Secreted; TNFR-II, Tumor Necrosis Factor Receptor 2; IL1ra, Interleukin 1 Receptor Antagonist; CRP, C-Reactive Protein. *Significant associations between the marker and fasting glucose or insulin. Adjusted for age, sex, BMI, waist circumference, total Cholesterol, HDL-cholesterol, medication for hypertension, smoking, prevalent CVD, lipid lowering medication.

subclinical atherosclerosis and coronary artery disease. A cross-sectional study in Italian population found that prediabetic patients exhibited lower RAGE plasma levels as well as increased levels of proinflammatory S100A12 in both prediabetic and diabetic patients[26]. In addition, we have previously reported the positive association between EN-RAGE and incident CHD in the Rotterdam Study[19]. Kosaki et al observed increased plasma EN-RAGE levels in patients with type 2 DM[27]. EN-RAGE was significantly associated with both HOMA-IR and HOMA-B, suggesting proinflammatory EN-RAGE leads to incident type 2 DM via inflammation –induced insulin resistance as well as via B-cell dysfunction (table S3).

Interleukin 13 (IL13) is a cytokine mainly produced by the T-helper (Th)-2 subset of lymphocytes, but also from non-T-cell populations such as mast cells, basophils, dendritic cells, keratinocytes and eosinophils.[28,29,30] IL13 is a regulator of inflammation and immune responses[31]. IL13 has a common receptor unit (α-chain) with interleukin 4 (IL4), which explains the similarities between IL13 and IL4[32]. Previous research has reported a preventive effect of IL4 on the onset of diabetes in non-obese diabetic mice (NOD mice)[33]. Furthermore, Zaccone et al found that IL13 prevents autoimmune diabetes in NOD mice, providing evidence that IL13 down-regulates the immune-inflammatory diabetogenic pathways[34], which is in agreement with our findings. Wong et al suggested the stimulation of IL13 receptors on T-cells, as a new pathway for tolerance induction in NOD mice[35]. In addition, IL13 is a B cell stimulating factor, which further supports our observation[36]. Stanya et al conclude that IL13 mitigates proinflammatory response in mice and regulates glucose homeostasis via the IL-13rα1–STAT3 signaling pathway in the liver, and that this pathway might provide a target for glycemic control in type 2 DM[37].

We also investigated the associations of IL13 with HOMA-IR and HOMA-B. IL13 was associated with both of them, suggesting a protective role against insulin resistance and B-cell dysfunction (table S3).

There are six members in the interleukin 17 (IL17) cytokine family, including IL17A, IL17B, IL17C, IL17D, IL17E (also known as IL25) and IL17F. Among all the members, the biological function and regulation of IL17A and IL17F are best understood. IL17A was produced mainly in T cells, whereas IL-17F was produced in T cells, innate immune cells, and epithelial cells. Functionally, both IL17A and IL17F mediate pro-inflammatory responses[38,39]. IL17 family cytokines have been linked to many autoimmune diseases, including multiple sclerosis, rheumatoid arthritis, inflammatory bowel disease and psoriasis[40]. The role of IL17 on the risk for type 2 DM remains unclear. Roohi A et al[41]. reported no association between serum IL17 levels and type 1 and 2 diabetes. Another study found that therapeutic improvement of glucoregulation in newly diagnosed type 2 DM patients is associated with a reduction of IL-17 levels[42]. Our study suggest a protective IL17 cytokine against the risk for type 2 DM (HR = 0.76), which is controversial and novel to the already known pro-inflammatory role of IL17 family cytokines. However, a cross-sectional case-control study has reported inverse associations of serum levels of IL17 with type 2 DM as well as with retinopathy, which is in line with our findings[43].

This study has certain strengths and limitations. To our knowledge, this is the first prospective population-based cohort study to investigate the association between a large set of novel inflammatory markers and the progression of type 2 DM with long-term follow-up. Furthermore, we performed sensitivity analysis, adjusting the type I error for multiple testing in order to highlight the most robust associations in every analyses. However, given the novelty of the markers, we reported significant findings at the level of 0.05 to avoid missing possibly important findings[44]. Beyond the identification of new novel inflammatory markers for type 2 diabetes, our findings relate them specifically to different stages of the disease. We are also aware of some limitations of the study. First, we had to exclude inflammatory markers with very low serum concentrations. However, the selected markers have >60% completeness of measurements, indicating acceptable quality of quantification. Second, our population is 55 years and older, thus generalization of the results to a younger age should be done with caution. Also, the Rotterdam Study mainly includes individuals from European Ancestry (98%). The effect estimates might differ between ethnicities.

A better prevention of type 2 DM requires the targeting of subjects at high risk in very early phases, such as prediabetes[12]. In this context, it is worth to investigate novel inflammatory markers that might be detectors of different stages of type 2 DM development[17].

In conclusion, our results show various inflammatory markers are associated with the progression from normoglycemia to type 2 DM and need for insulin therapy in a phase-specific manner. Among them, EN-RAGE is a novel inflammatory marker for prediabetes, IL17 for incident type 2 DM and IL13 for prediabetes, incident type 2 DM and insulin therapy start. This study only indicates new associations, emphasizing the need for further studies to establish the role of EN-RAGE, IL13 and IL17 in the development of type 2 diabetes.

References

1. Hotamisligil GS. Inflammation and metabolic disorders. *Nature* 2006; 444(7121): 860-7.
2. Bertoni AG, Burke GL, Owusu JA, et al. Inflammation and the incidence of type 2 diabetes: the Multi-Ethnic Study of Atherosclerosis (MESA). *Diabetes Care* 2010; 33(4): 804-10.
3. Barzilay JI, Abraham L, Heckbert SR, et al. The relation of markers of inflammation to the development of glucose disorders in the elderly: the Cardiovascular Health Study. *Diabetes* 2001; 50(10): 2384-9.
4. Galle J, Quaschning T, Seibold S, Wanner C. Endothelial dysfunction and inflammation: what is the link? *Kidney Int Suppl* 2003; (84): S45-9.
5. Meigs JB, Hu FB, Rifai N, Manson JE. Biomarkers of endothelial dysfunction and risk of type 2 diabetes mellitus. *Jama* 2004; 291(16): 1978-86.
6. Dehghan A, Kardys I, de Maat MP, et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. *Diabetes* 2007; 56(3): 872-8.
7. Wang X, Bao W, Liu J, et al. Inflammatory markers and risk of type 2 diabetes: a systematic review and meta-analysis. *Diabetes Care* 2013; 36(1): 166-75.
8. Pradhan AD, Manson JE, Rifai N, Buring JE, Ridker PM. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *Jama* 2001; 286(3): 327-34.
9. Dandona P, Aljada A, Bandyopadhyay A. Inflammation: the link between insulin resistance, obesity and diabetes. *Trends Immunol* 2004; 25(1): 4-7.
10. Krakoff J, Funahashi T, Stehouwer CD, et al. Inflammatory markers, adiponectin, and risk of type 2 diabetes in the Pima Indian. *Diabetes Care* 2003; 26(6): 1745-51.
11. Li S, Shin HJ, Ding EL, van Dam RM. Adiponectin levels and risk of type 2 diabetes: a systematic review and meta-analysis. *Jama* 2009; 302(2): 179-88.
12. Tabak AG, Herder C, Rathmann W, Brunner EJ, Kivimaki M. Prediabetes: a high-risk state for diabetes development. *Lancet* 2012; 379(9833): 2279-90.
13. Fonseca VA. Defining and characterizing the progression of type 2 diabetes. *Diabetes Care* 2009; 32 Suppl 2: S151-6.
14. Tabak AG, Jokela M, Akbaraly TN, Brunner EJ, Kivimaki M, Witte DR. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *Lancet* 2009; 373(9682): 2215-21.
15. Abdul-Ghani MA, Tripathy D, DeFronzo RA. Contributions of beta-cell dysfunction and insulin resistance to the pathogenesis of impaired glucose tolerance and impaired fasting glucose. *Diabetes Care* 2006; 29(5): 1130-9.
16. Gastaldelli A, Ferrannini E, Miyazaki Y, Matsuda M, DeFronzo RA, San Antonio metabolism s. Beta-cell dysfunction and glucose intolerance: results from the San Antonio metabolism (SAM) study. *Diabetologia* 2004; 47(1): 31-9.
17. Grossmann V, Schmitt VH, Zeller T, et al. Profile of the Immune and Inflammatory Response in Individuals With Prediabetes and Type 2 Diabetes. *Diabetes Care* 2015; 38(7): 1356-64.
18. Hofman A, van Duijn CM, Franco OH, et al. The Rotterdam Study: 2012 objectives and design update. *Eur J Epidemiol* 2011; 26(8): 657-86.

19. Ligthart S, Sedaghat S, Ikram MA, Hofman A, Franco OH, Dehghan A. EN-RAGE: a novel inflammatory marker for incident coronary heart disease. *Arterioscler Thromb Vasc Biol* 2014; 34(12): 2695-9.

20. Ligthart S, van Herpt TT, Leening MJ, et al. Lifetime risk of developing impaired glucose metabolism and eventual progression from prediabetes to type 2 diabetes: a prospective cohort study. *Lancet Diabetes Endocrinol* 2016; 4(1): 44-51.

21. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 1985; 28(7): 412-9.

22. Brett J, Schmidt AM, Yan SD, et al. Survey of the distribution of a newly characterized receptor for advanced glycation end products in tissues. *Am J Pathol* 1993; 143(6): 1699-712.

23. Foell D, Wittkowski H, Kessel C, et al. Proinflammatory S100A12 can activate human monocytes via Toll-like receptor 4. *Am J Respir Crit Care Med* 2013; 187(12): 1324-34.

24. Yang Z, Yan WX, Cai H, et al. S100A12 provokes mast cell activation: a potential amplification pathway in asthma and innate immunity. *J Allergy Clin Immunol* 2007; 119(1): 106-14.

25. Andreasen AS, Kelly M, Berg RM, Moller K, Pedersen BK. Type 2 diabetes is associated with altered NF-kappaB DNA binding activity, JNK phosphorylation, and AMPK phosphorylation in skeletal muscle after LPS. *PLoS One* 2011; 6(9): e23999.

26. Di Pino A, Currenti W, Urbano F, et al. Low advanced glycation end product diet improves the lipid and inflammatory profiles of prediabetic subjects. *J Clin Lipidol* 2016; 10(5): 1098-108.

27. Kosaki A, Hasegawa T, Kimura T, et al. Increased plasma S100A12 (EN-RAGE) levels in patients with type 2 diabetes. *J Clin Endocrinol Metab* 2004; 89(11): 5423-8.

28. Schmid-Grendelmeier P, Altznauer F, Fischer B, et al. Eosinophils express functional IL-13 in eosinophilic inflammatory diseases. *J Immunol* 2002; 169(2): 1021-7.

29. Kim J, Woods A, Becker-Dunn E, Bottomly K. Distinct functional phenotypes of cloned Ia-restricted helper T cells. *J Exp Med* 1985; 162(1): 188-201.

30. Mosmann TR, Cherwinski H, Bond MW, Giedlin MA, Coffman RL. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J Immunol* 1986; 136(7): 2348-57.

31. Minty A, Chalon P, Derocq JM, et al. Interleukin-13 is a new human lymphokine regulating inflammatory and immune responses. *Nature* 1993; 362(6417): 248-50.

32. Hilton DJ, Zhang JG, Metcalf D, Alexander WS, Nicola NA, Willson TA. Cloning and characterization of a binding subunit of the interleukin 13 receptor that is also a component of the interleukin 4 receptor. *Proc Natl Acad Sci U S A* 1996; 93(1): 497-501.

33. Rapoport MJ, Jaramillo A, Zipris D, et al. Interleukin 4 reverses T cell proliferative unresponsiveness and prevents the onset of diabetes in nonobese diabetic mice. *J Exp Med* 1993; 178(1): 87-99.

34. Zaccone P, Phillips J, Conget I, et al. Interleukin-13 prevents autoimmune diabetes in NOD mice. *Diabetes* 1999; 48(8): 1522-8.

35. Wong FS. Stimulating IL-13 receptors on T cells: a new pathway for tolerance induction in diabetes? *Diabetes* 2011; 60(6): 1657-9.

36. Defrance T, Carayon P, Billian G, et al. Interleukin 13 is a B cell stimulating factor. *J Exp Med* 1994; 179(1): 135-43.

37. Stanya KJ, Jacobi D, Liu S, et al. Direct control of hepatic glucose production by interleukin-13 in mice. *J Clin Invest* 2013; 123(1): 261-71.

38. Ishigame H, Kakuta S, Nagai T, et al. Differential roles of interleukin-17A and -17F in host defense against mucoepithelial bacterial infection and allergic responses. *Immunity* 2009; 30(1): 108-19.

39. Yang XO, Chang SH, Park H, et al. Regulation of inflammatory responses by IL-17F. *J Exp Med* 2008; 205(5): 1063-75.

40. Park H, Li Z, Yang XO, et al. A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nat Immunol* 2005; 6(11): 1133-41.

41. Roohi A, Tabrizi M, Abbasi F, et al. Serum IL-17, IL-23, and TGF-beta levels in type 1 and type 2 diabetic patients and age-matched healthy controls. *Biomed Res Int* 2014; 2014: 718946.

42. Sumarac-Dumanovic M, Jeremic D, Pantovic A, et al. Therapeutic improvement of glucoregulation in newly diagnosed type 2 diabetes patients is associated with a reduction of IL-17 levels. *Immunobiology* 2013; 218(8): 1113-8.

43. Nadeem A, Javaid K, Sami W, et al. Inverse relationship of serum IL-17 with type-II diabetes retinopathy. *Clin Lab* 2013; 59(11-12): 1311-7.

44. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; 1(1): 43-6.

Supplementary material

Table S1. Baseline characteristics among non prediabetes cases and incident prediabetes cases.

Characteristic	Non-prediabetes cases	Incident prediabetes cases	P-value
Total population number	559	139	
Age, years	73±7.6	71±6.1	< 0.001
Men, n (%)	238 (43)	60 (43.2)	0.9
Waist Circumference, m	0.9±0.1	0.9±0.1	0.3
Body mass index, kg/m^2	25.9±3.6	27.0±3.8	0.003
Systolic blood pressure, mmHg	143±21.7	142±20.1	0.4
Diastolic blood pressure, mmHg	74.6±10.9	74.9±10.9	0.7
Hypertension medication with indication, n(%)	107 (19.3)	30 (21.6)	0.052
Total cholesterol, mmol/L	5.8±1.0	5.8±0.9	0.9
HDL cholesterol, mmol/L	1.5±0.4	1.4±0.4	0.016
Fasting glucose, mmol/L	5.3 (4.7–5.9)	5.6 (4.9–6.0)	< 0.001
Fasting insulin, uIU/L	8.2 (3.7–18.4)	9.6 (3.7–22.4)	0.002
Current smokers, n (%)	85 (15.4)	19 (13.7)	0.6
Former smokers, n (%)	255 (46.1)	77 (55.4)	0.1
Prevalent CVD, n (%)	94 (17)	15 (10.8)	0.07
Alcohol intake in drinkers (76%), g/day	2.9 (0.0–40.2)	2.9 (0.0–40.0)	0.6
Lipid lowering medication, n (%)	57 (10.3)	26 (18.7)	0.007

Abbreviations: HDL, high density lipoproteins; CVD, cardiovascular disease.
Plus-minus values are means ± standard deviation or median (inter-quartile range).

Table S2. Baseline characteristics among non- diabetes cases/ diabetes cases.

Characteristic	Non- diabetes cases	Incident diabetes cases	P-value
Total population number	741	110	
Age, years	73±7.5	70±5.9	< 0.001
Men, n (%)	330 (44.9)	46 (41.8)	0.5
Waist Circumference, m	0.9±0.1	0.9±0.1	< 0.001
Body mass index, kg/m^2	26.2±3.7	28.4±4.2	< 0.001
Systolic blood pressure, mmHg	143.4±21.7	144.9±19.6	0.4
Diastolic blood pressure, mmHg	75.0±11.2	75.7±11.2	0.5
Hypertension medication with indication, n(%)	154 (21)	36 (32.7)	0.02
Total cholesterol, mmol/L	5.8±1.0	5.9±0.9	0.4
HDL cholesterol, mmol/L	1.4±0.4	1.3±0.4	< 0.001
Fasting glucose, mmol/L	5.4 (4.7–6.3)	6.1 (5.1–6.8)	< 0.001
Fasting insulin, uIU/L	8.5 (3.7–19.3)	12.9 (5.3–27.3)	< 0.001
Current smokers, n (%)	107 (14.6)	18 (16.4)	0.6
Former smokers, n (%)	355 (48.3)	57 (51.8)	0.6
Prevalent CVD, n (%)	124 (16.9)	9 (8.2)	0.02
Alcohol intake in drinkers (76%), g/day	2.9 (0.0–41.9)	2.9 (0.0–40.1)	0.5
Lipid lowering medication, n (%)	81 (11)	14 (12.7)	0.8

Abbreviations: HDL, high density lipoproteins; CVD, cardiovascular disease.

Plus-minus values are means ± standard deviation or median (inter-quartile range).

Table S3. Baseline characteristics among non-insulin starters and insulin starters.

Characteristic	Non-insulin starters	Insulin starters	P-value
Total population number	89	26	
Age, years	74.9±8.3	73.6±6.5	0.001
Men, n (%)	45 (50.6)	12 (46.2)	0.3
Waist Circumference, m	0.9±0.1	0.9±0.1	0.6
Body mass index, kg/m^2	28.3±4.4	28.7±4.7	0.09
Systolic blood pressure, mmHg	145.8±21.0	149.7±28.5	0.7
Diastolic blood pressure, mmHg	75.3±10.1	75.4±9.5	0.7
Hypertension medication with indication, n(%)	39 (43.8)	8 (30.8)	0.5
Total cholesterol, mmol/L	5.7±0.9	5.8±0.8	0.5
HDL cholesterol, mmol/L	1.2±0.4	1.1±0.2	0.2
Fasting glucose, mmol/L	7.5 (5.4–12.5)	8.8 (6.6–17.2)	0.02
Fasting insulin, uIU/L	12.2 (4.6–41.1)	12.7 (5.9–52.9)	0.7
Current smokers, n (%)	11 (12.4)	0 (0)	0.03
Former smokers, n (%)	49 (55.1)	17 (65.4)	0.1
Prevalent CVD, n (%)	22 (24.7)	8 (30.8)	0.005
Alcohol intake in drinkers (76%), g/day	1.4 (0.0–21.1)	1.4 (0.0–18.1)	0.4
Lipid lowering medication, n (%)	15 (16.9)	8 (30.8)	0.6

Abbreviations: HDL, high density lipoproteins; CVD, cardiovascular disease.
Plus-minus values are means ± standard deviation or median (inter-quartile range).

Table S4. Baseline characteristics of the inflammatory markers.

Marker	P-Value
Total population free of diabetes	851
CD40, ng/mL	0.73±0.27
CD40 ligand*, ng/mL	0.03 (0.01-0.06)
EN-RAGE*, ng/mL	10.70 (4.82-24.45)
Eotaxin*, pg/mL	161.00 (65.40-330.65)
FAS*, ng/mL	4.65 (2.94-8.18)
HCC4, ng/mL	4.87±1.95
IL13*, pg/mL	76.20 (48.70 - -123.00)
IL16, pg/mL	381.98±103.89
IL17*, pg/mL	12.90 (6.22-23.30)
IL8*, pg/mL	9.24 (4.25-20.92)
MDC, pg/mL	365.61±124.01
MIP1 alpha*, pg/mL	46.20 (27.22-73.46)
MIP1 beta*, pg/mL	122.00 (66.06-323.20)
PARC, ng/mL	29.93±11.12
sRAGE *, ng/mL	2.67 (1.27-5.66)
TRAILR3 *, ng/mL	6.55 (3.48-12.48)
CFH*, ug/mL	2520 (890.95-3700)
Complement 3*, mg/mL	0.82 (0.62-1.06)
IL18*, pg/mL	187 (100-381.40)
MCP1*, pg/mL	184 (113-309)
MIF*, ng/mL	0.06 (0.01-0.15)
RANTES*, ng/mL	0.51 (0.18-1.79)
Resistin*, ng/mL	0.42 (0.17-0.99)
TNFRII*, ng/mL	3.51 (2.25-6.21)
Il1ra*, pg/mL	66.80 (25.99-191)
CRP*, ug/mL	1.37 (0.23-8.90)

Plus-minus values are means ± standard deviation or median (inter-quartile range). *Naturally log-transformed markers. CD40, cluster of differentiation 40; CD40 ligand, cluster of differentiation 40 ligand ; EN-RAGE, Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein; FAS, Fas Cell Surface Death Receptor; HCC4, Human CC chemokine-4; IL13, interleukin 13; IL16, interleukin 16; IL17, interleukin 17; IL8, interleukin 8; MDC, Monocyte Derived Chemokine; MIP1alpha, Macrophage Inflammatory Protein 1 alpha; MIP1beta, Macrophage Inflammatory Protein 1 beta; PARC, Pulmonary and Activation-Regulated Chemokine; sRage, Soluble Receptor of Advanced Glycation End-products; TRAILR3, Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor 3; CFH, Complement Factor H; IL18, interleukin 18; MCP1, Monocyte Chemotactic Protein 1; RANTES, Regulated Upon Activation, Normally T-Expressed, And Presumably Secreted; TNFR-II, Tumor Necrosis Factor Receptor 2; IL1ra, Interleukin 1 Receptor Antagonist; CRP, C-Reactive Protein.

Table S5. Multivariable adjusted associations between markers of inflammation and fasting glucose.

Marker	N	Beta (95%CI)	P-value
CD40, ng/mL	848	0.006 (-0.003, 0.014)	0.1
CD40 ligand*, ng/mL	780	-0.0004 (-0.008, 0.007)	0.9
EN-RAGE*, ng/mL	843	0.009 (0.002, 0.015)	1.2×10^{-2}
Eotaxin*, pg/mL	841	-0.001 (-0.008, 0.006)	0.6
FAS*, ng/mL	837	-0.001 (-0.008, 0.007)	0.8
HCC4, ng/mL	850	0.004 (-0.003, 0.012)	0.2
IL13*, pg/mL	814	-0.008 (-0.016, 0.000)	4.7×10^{-2}
IL16, pg/mL	849	-0.003 (-0.010, 0.004)	0.3
IL17*, pg/mL	805	0.001 (-0.005, 0.008)	0.7
IL8*, pg/mL	824	0.004 (-0.003, 0.011)	0.2
MDC, pg/mL	846	-0.004 (-0.012, 0.004)	0.3
MIP1 alpha*, pg/mL	846	-0.002 (-0.009, 0.005)	0.5
MIP1 beta*, pg/mL	844	-0.002 (-0.009, 0.005)	0.5
PARC, ng/mL	845	-0.0001 (-0.008, 0.008)	0.9
sRAGE*, ng/mL	847	-0.009 (-0.015, -0.002)	1.4×10^{-2}
TRAILR3*, ng/mL	844	0.0004 (-0.006, 0.007)	0.8
CFH*, ug/mL	840	0.001 (-0.005, 0.008)	0.6
Complement 3*, mg/mL	851	0.005 (-0.002, 0.012)	0.1
IL18*, pg/mL	844	-0.0001 (-0.007, 0.007)	0.9
MCP1*, pg/mL	846	-0.007 (-0.014, 0.001)	7.5×10^{-2}
MIF*, ng/mL	831	-0.002 (-0.008, 0.005)	0.5
RANTES*, ng/mL	849	-0.002 (-0.008, 0.005)	0.6
Resistin*, ng/mL	844	-0.006 (-0.013, $1\text{x}10^{-4}$)	5.6×10^{-2}
TNFRII*, ng/mL	846	-0.005 (-0.013, 0.002)	0.1
Il1ra*, pg/mL	821	0.005 (-0.002, 0.012)	0.2
CRP*, ug/mL	837	0.002 (-0.005, 0.010)	0.4

* Naturally log-transformed. CD40, cluster of differentiation 40; CD40 ligand, cluster of differentiation 40 ligand ; EN-RAGE, Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein; FAS, Fas Cell Surface Death Receptor; HCC4, Human CC chemokine-4; IL13, interleukin 13; IL16, interleukin 16; IL17, interleukin 17; IL8, interleukin 8; MDC, Monocyte Derived Chemokine; MIP1alpha, Macrophage Inflammatory Protein 1 alpha; MIP1beta, Macrophage Inflammatory Protein 1 beta; PARC, Pulmonary and Activation-Regulated Chemokine; sRage, Soluble Receptor of Advanced Glycation End-products; TRAILR3, Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor 3; CFH, Complement Factor H; IL18, interleukin 18; MCP1, Monocyte Chemotactic Protein 1; RANTES, Regulated Upon Activation, Normally T-Expressed, And Presumably Secreted; TNFR-II, Tumor Necrosis Factor Receptor 2; IL1ra, Interleukin 1 Receptor Antagonist; CRP, C-Reactive Protein. Adjusted for age, sex, BMI, waist circumference (WC), Total Cholesterol, HDL, medication for hypertension, smoking, prevalent CVD, lipid lowering medication. [b]Sensitivity analysis: significant after Bonferroni correction (p-value=0.05/26 = 1.9×10^{-3}).

Table S6. Multivariable adjusted associations between markers of inflammation and fasting insulin.

Marker	N	Beta (95%CI)	P-value
CD40, ng/mL	848	0.045 (0.005, 0.086)	2.9×10^{-2}
CD40 ligand*, ng/mL	780	-0.022 (-0.057, 0.014)	0.2
EN-RAGE*, ng/mL	843	0.047 (0.015, 0.080)	5×10^{-3}
Eotaxin*, pg/mL	841	0.011 (-0.023, 0.044)	0.5
FAS*, ng/mL	837	0.051 (0.014, 0.087)	6×10^{-3}
HCC4, ng/mL	850	0.054 (0.019, 0.088)	3×10^{-3}
IL13*, pg/mL	814	-0.071 (-0.107, -0.034)	[b]1.6×10^{-4}
IL16, pg/mL	849	0.006 (-0.027, 0.039)	0.7
IL17*, pg/mL	805	-0.010 (-0.042, 0.022)	0.5
IL8*, pg/mL	824	0.058 (0.024, 0.091)	[b]1.0×10^{-3}
MDC, pg/mL	846	0.003 (-0.035, 0.041)	0.8
MIP1 alpha*, pg/mL	846	0.023 (-0.010, 0.056)	0.1
MIP1 beta*, pg/mL	844	0.032 (-0.003, 0.067)	6.9×10^{-2}
PARC, ng/mL	845	0.016 (-0.021, 0.053)	0.3
sRAGE*, ng/mL	847	-0.023 (-0.056, 0.009)	0.1
TRAILR3*, ng/mL	844	0.048 (0.017, 0.079)	3.0×10^{-3}
CFH*, ug/mL	840	-0.031 (-0.062, -0.001)	4.6×10^{-2}
Complement 3*, mg/mL	851	0.096 (0.062, 0.130)	[b]2.3×10^{-8}
IL18*, pg/mL	844	0.040 (0.008, 0.072)	1.5×10^{-2}
MCP1*, pg/mL	846	0.0003 (-0.037, 0.037)	0.9
MIF*, ng/mL	831	0.007 (-0.025, 0.039)	0.6
RANTES*, ng/mL	849	-0.007 (-0.040, 0.026)	0.6
Resistin*, ng/mL	844	0.003 (-0.028, 0.035)	0.8
TNFRII*, ng/mL	846	0.028 (-0.007, 0.064)	0.1
Il1ra*, pg/mL	821	0.061 (0.028, 0.094)	[b]2.7×10^{-4}
CRP*, ug/mL	837	0.025 (-0.010, 0.060)	0.1

* Naturally log-transformed. CD40, cluster of differentiation 40; CD40 ligand, cluster of differentiation 40 ligand ; EN-RAGE, Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein; FAS, Fas Cell Surface Death Receptor; HCC4, Human CC chemokine-4; IL13, interleukin 13; IL16, interleukin 16; IL17, interleukin 17; IL8, interleukin 8; MDC, Monocyte Derived Chemokine; MIP1alpha, Macrophage Inflammatory Protein 1 alpha; MIP1beta, Macrophage Inflammatory Protein 1 beta; PARC, Pulmonary and Activation-Regulated Chemokine; sRage, Soluble Receptor of Advanced Glycation End-products; TRAILR3, Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor 3; CFH, Complement Factor H; IL18, interleukin 18; MCP1, Monocyte Chemotactic Protein 1; RANTES, Regulated Upon Activation, Normally T-Expressed, And Presumably Secreted; TNFR-II, Tumor Necrosis Factor Receptor 2; IL1ra, Interleukin 1 Receptor Antagonist; CRP, C-Reactive Protein. Adjusted for age, sex, BMI, waist circumference (WC), Total Cholesterol, HDL, medication for hypertension, smoking, prevalent CVD, lipid lowering medication. [b]Sensitivity analysis: significant after Bonferroni correction (p-value=0.05/26=1.9×10^{-3}).

Table S7. Associations between markers of inflammation and HOMA-IR.

Marker	N	Beta (95%CI)	P-value
CD40, ng/mL	848	0.105 (0.056, 0.155)	$^{b}3.4\times10^{-5}$
CD40 ligand *, ng/mL	780	-0.020 (-0.065, 0.025)	0.3
EN-RAGE *, ng/mL	843	0.090 (0.050, 0.130)	$^{b}1.0\times10^{-5}$
Eotaxin *, pg/mL	841	-0.036 (-0.077, 0.005)	8.8×10^{-2}
FAS *, ng/mL	837	0.117 (0.073, 0.162)	$^{b}1.7\times10^{-7}$
HCC4, ng/mL	850	0.107 (0.064, 0.150)	$^{b}9.5\times10^{-7}$
IL13 *, pg/mL	814	-0.168 (-0.206, -0.130)	$^{b}5.0\times10^{-17}$
IL16, pg/mL	849	0.054 (0.013, 0.095)	1.0×10^{-2}
IL17 *, pg/mL	805	-0.027 (-0.067, 0.013)	0.1
IL8 *, pg/mL	824	0.060 (0.019, 0.102)	4.0×10^{-3}
MDC, pg/mL	846	0.049 (0.003, 0.095)	3.8×10^{-2}
MIP1 alpha *, pg/mL	846	0.051 (0.010, 0.092)	1.4×10^{-2}
MIP1 beta *, pg/mL	844	0.063 (0.020, 0.106)	4.0×10^{-3}
PARC, ng/mL	845	0.040 (-0.005, 0.086)	8.4×10^{-2}
sRAGE *, ng/mL	847	-0.057 (-0.097, -0.017)	6.0×10^{-3}
TRAILR3 *, ng/mL	844	0.085 (0.047, 0.124)	$^{b}1.5\times10^{-5}$
CFH *, ug/mL	840	0.003 (-0.036, 0.041)	0.8
Complement 3*, mg/mL	851	0.193 (0.156, 0.230)	$^{b}6.3\times10^{-23}$
IL18*, pg/mL	844	0.085 (0.046, 0.125)	$^{b}2.2\times10^{-5}$
MCP1*, pg/mL	846	-0.023 (-0.069, 0.023)	0.3
MIF*, ng/mL	831	0.040 (0.001, 0.079)	4.2×10^{-2}
RANTES*, ng/mL	849	-0.026 (-0.065, 0.013)	0.1
Resistin*, ng/mL	844	0.017 (-0.023, 0.057)	0.4
TNFRII*, ng/mL	846	0.106 (0.064, 0.148)	$^{b}6.9\times10^{-7}$
Il1ra*, pg/mL	821	0.150 (0.112, 0.189)	$^{b}1.9\times10^{-14}$
CRP*, ug/mL	837	0.132 (0.093, 0.170)	$^{b}3.6\times10^{-11}$

Age and sex adjusted. The number of subjects differs for each marker, after outliers exclusion. *Naturally log-transformed. CD40, cluster of differentiation 40; CD40 ligand, cluster of differentiation 40 ligand ; EN-RAGE, Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein; FAS, Fas Cell Surface Death Receptor; HCC4, Human CC chemokine-4; IL13, interleukin 13; IL16, interleukin 16; IL17, interleukin 17; IL8, interleukin 8; MDC, Monocyte Derived Chemokine; MIP1alpha, Macrophage Inflammatory Protein 1 alpha; MIP1beta, Macrophage Inflammatory Protein 1 beta; PARC, Pulmonary and Activation-Regulated Chemokine; sRage, Soluble Receptor of Advanced Glycation End-products; TRAILR3, Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor 3; CFH, Complement Factor H; IL18, interleukin 18; MCP1, Monocyte Chemotactic Protein 1; RANTES, Regulated Upon Activation, Normally T-Expressed, And Presumably Secreted; TNFR-II, Tumor Necrosis Factor Receptor 2; IL1ra, Interleukin 1 Receptor Antagonist; CRP, C-Reactive Protein. ^{b}Sensitivity analysis: significant after Bonferroni correction (p-value=0.05/26=1.9×10^{-3}).

Table S8. Associations between markers of inflammation and HOMA-B.

Marker	N	Beta (95%CI)	P-value
CD40, ng/mL	848	0.077 (0.033, 0.122)	$^{b}1.0\times10^{-3}$
CD40 ligand *, ng/mL	780	-0.012 (-0.052, 0.029)	0.5
EN-RAGE *, ng/mL	843	0.045 (0.009, 0.081)	1.4×10^{-2}
Eotaxin *, pg/mL	841	-0.019 (-0.056, 0.018)	0.3
FAS *, ng/mL	837	0.101 (0.062, 0.141)	$^{b}5.5\times10^{-7}$
HCC4, ng/mL	850	0.076 (0.037, 0.114)	$^{b}1.1\times10^{-4}$
IL13 *, pg/mL	814	-0.120 (-0.155, -0.085)	$^{b}2.0\times10^{-11}$
IL16, pg/mL	849	0.051 (0.014, 0.087)	7.0×10^{-3}
IL17 *, pg/mL	805	-0.024 (-0.060, 0.011)	0.1
IL8 *, pg/mL	824	0.051 (0.014, 0.088)	7.0×10^{-3}
MDC, pg/mL	846	0.048 (0.006, 0.089)	2.4×10^{-2}
MIP1 alpha *, pg/mL	846	0.054 (0.018, 0.091)	4.0×10^{-3}
MIP1 beta *, pg/mL	844	0.063 (0.025, 0.102)	$^{b}1.0\times10^{-3}$
PARC, ng/mL	845	0.039 (-0.002, 0.080)	5.9×10^{-2}
sRAGE *, ng/mL	847	-0.009 (-0.045, 0.027)	0.6
TRAILR3 *, ng/mL	844	0.073 (0.039, 0.108)	$^{b}3.0\times10^{-5}$
CFH *, ug/mL	840	-0.013 (-0.047, 0.022)	0.4
Complement 3*, mg/mL	851	0.140 (0.106, 0.174)	$^{b}8.8\times10^{-16}$
IL18*, pg/mL	844	0.077 (0.042, 0.112)	$^{b}1.8\times10^{-5}$
MCP1*, pg/mL	846	0.007 (-0.034, 0.048)	0.7
MIF*, ng/mL	831	0.040 (0.005, 0.075)	2.6×10^{-2}
RANTES*, ng/mL	849	-0.012 (-0.047, 0.022)	0.4
Resistin*, ng/mL	844	0.037 (0.002, 0.073)	3.9×10^{-2}
TNFRII*, ng/mL	846	0.103 (0.066, 0.141)	$^{b}5.0\times10^{-8}$
Il1ra*, pg/mL	821	0.110 (0.076, 0.145)	$^{b}4.4\times10^{-10}$
CRP*, ug/mL	837	0.086 (0.051, 0.122)	$^{b}2.0\times10^{-6}$

Age and sex adjusted. The number of subjects differs for each marker, after outliers exclusion. *Naturally log-transformed. CD40, cluster of differentiation 40; CD40 ligand, cluster of differentiation 40 ligand ; EN-RAGE, Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein; FAS, Fas Cell Surface Death Receptor; HCC4, Human CC chemokine-4; IL13, interleukin 13; IL16, interleukin 16; IL17, interleukin 17; IL8, interleukin 8; MDC, Monocyte Derived Chemokine; MIP1alpha, Macrophage Inflammatory Protein 1 alpha; MIP1beta, Macrophage Inflammatory Protein 1 beta; PARC, Pulmonary and Activation-Regulated Chemokine; sRage, Soluble Receptor of Advanced Glycation End-products; TRAILR3, Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor 3; CFH, Complement Factor H; IL18, interleukin 18; MCP1, Monocyte Chemotactic Protein 1; RANTES, Regulated Upon Activation, Normally T-Expressed, And Presumably Secreted; TNFR-II, Tumor Necrosis Factor Receptor 2; IL1ra, Interleukin 1 Receptor Antagonist; CRP, C-Reactive Protein. ^{b}Sensitivity analysis: significant after Bonferroni correction (p-value=0.05/26=1.9×10^{-3}).

Chapter 5

EN-RAGE: a Novel Inflammatory Marker for Incident Coronary Heart Disease

Background: Inflammation plays a key role in atherosclerosis. We hypothesized that novel inflammatory markers may predict the risk of coronary heart disease (CHD).

Methods: We investigated the association of 16 inflammatory biomarkers with the risk of CHD in a random subset of 839 CHD free individuals in a prospective population-based cohort study. A Bonferroni corrected p-value of 3.1×10^{-3} was used as a threshold of statistical significance.

Results: The mean age at baseline was 72.8 years. During a median follow-up of 10.6 years, 99 cases of incident CHD were observed. Among all inflammatory biomarkers, neutrophil derived human s100a12 (EN-RAGE) showed the strongest association with the risk of CHD (p-value 2.0×10^{-3}). After multivariable adjustment for established cardiovascular risk factors, each standard deviation increase in the natural log-transformed EN-RAGE was associated with 30% higher risk of incident CHD (Hazard ratio: 1.30; 95% confidence interval (CI) CI: 1.06–1.59). Further adjustment for previously studied inflammatory markers did not attenuate the association. Excluding individuals with prevalent type 2 diabetes, impaired kidney function or individuals using antihypertensive medication did not change the effect estimates. Cause-specific hazard ratios suggested a stronger association between EN-RAGE and CHD mortality compared to stable CHD.

Conclusion: Our results highlight EN-RAGE as an inflammatory marker for future CHD in a general population, beyond traditional CHD risk factors and inflammatory markers.

Introduction

With 7.3 million deaths per year globally, coronary heart disease (CHD) is still the world's leading cause of mortality[1]. Inflammation is thought to play a key role in the pathogenesis of atherosclerosis and CHD[2]. Accordingly, inflammatory markers have been investigated for predicting the risk of CHD, an effort that has led to the identification and validation of several inflammatory markers for CHD[3,4,5,6]. However, the inflammatory markers that have been investigated so far only represent a minor part of the diverse molecules that constitute the complex human immune response[7]. Exploring prospectively the association of relatively uninvestigated inflammatory markers with CHD may unravel novel inflammatory risk factors for CHD and may shed light on additional pathways that might be involved in the pathogenesis of atherosclerosis and CHD.

We hypothesized that indicators of inflammation which have not been studied previously with the incidence of CHD are associated with incident CHD beyond traditional risk factors and previously studied inflammatory markers. To this end, we studied the association of 16 biomarkers of inflammation with the risk of CHD in the Rotterdam Study, a prospective population-based cohort study.

Methods

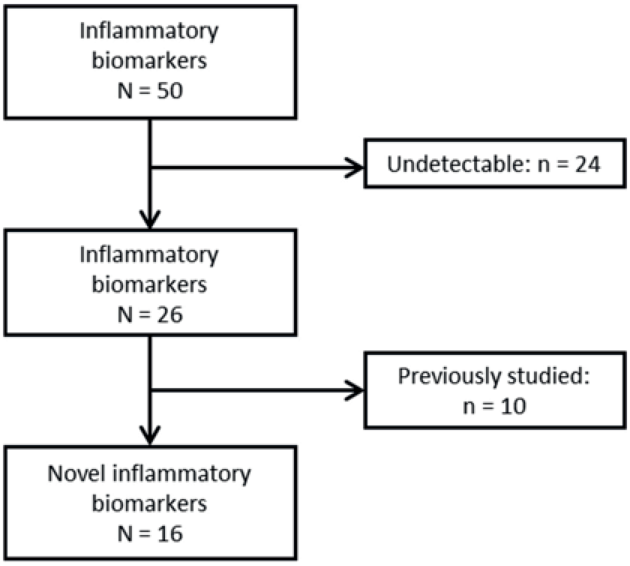
Study Population

The Rotterdam Study is a prospective population-based cohort study in Ommoord, a district of Rotterdam, the Netherlands. The design of the Rotterdam Study has been described in more detail elsewhere[8]. Briefly, in 1989 all residents of Ommoord aged 55 years or older were invited to participate of whom 78% (7,983 out of 10,275) agreed. The first examination round was completed between 1990 and 1993, after which follow-up examinations were conducted in 1993-1994, 1997-1999, 2002-2004 and 2009-2011. This study was based on data collected during the third visit (1997-1999). Among 5990 (80%) eligible individuals, 4797 individuals visited the research center. A random subset of 971 participants was selected as part of a separate case-cohort study to investigate biomarkers in association with dementia. Given the random sampling these persons can be considered representative of the source population. We excluded 132 participants with history of CHD (defined as clinically manifest myocardial infarction, coronary artery bypass grafting, or percutaneous trans luminal coronary angioplasty), resulting in 839 participants for analysis. The Rotterdam Study has been approved by the medical ethics committee according to the Population Study Act Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. A written informed consent was obtained from all participants.

Measurement of Inflammatory Biomarkers

In the third center visit, fasting blood samples were collected at the research center. Plasma was isolated and immediately put on ice and stored at -80°C. Citrate plasma (200 µL) was sent in July 2008 to Rules-Based Medicine, Austin, Texas (www.myriadrbm.com). Fifty inflammatory biomarkers were quantified using multiplex immunoassay on a custom designed human multianalyte profile. The intra-assay variability was less than 4% and the inter-assay variability was less than 13%. Biomarkers with more than 60% completeness of measurements were selected for imputation and further analysis (Figure 1). Among the 26 eligible biomarkers, 10 were excluded since they have previously been investigated prospectively with the incidence of CHD (table S1). This resulted in a final set of 16 novel inflammatory biomarkers that were selected to investigate with incidence of CHD (table S2). The inflammatory markers investigated in the current study have no standard international calibration reference, therefore interpretation of the absolute values should be with caution. Since the current study is conducted within one set of individuals, the use of relative measures does not affect the effect estimates.

Figure 1. Flow chart of inflammatory biomarker inclusion.


Inflammatory biomarkers N = 50
Undetectable: n = 24
Inflammatory biomarkers N = 26
Previously studied: n = 10
Novel inflammatory biomarkers N = 16

Coronary Heart Disease Diagnosis

Information on the incidence of CHD was obtained from general practitioners and from letters and discharge reports of medical specialists. Two independent study physicians coded all reported events and in case of disagreement, consensus was sought. Subsequently, a medical specialist validated all events. Incident CHD was defined as myocardial revascularization, fatal and non-fatal myocardial infarction and CHD mortality. Definition and coding of CHD events within the Rotterdam Study is described in more detail elsewhere[9]. Follow-up data until January 1st 2011 was used.

Covariates

Anthropometric measures were obtained during the visit to the research center. Body mass index (BMI) was defined as weight in kilogram divided by the square of height in meters. Blood pressure was measured during research center visit at the right brachial artery, with participants in sitting position. The mean of two consecutive measurements was used. Total and high-density lipoprotein cholesterol (HDL-cholesterol) levels, creatinine and white blood cell counts were measured in fasting blood samples with standard laboratory techniques. The glomerular filtration rate (GFR) was estimated by the abbreviated modification of diet in renal disease (MDRD) equation which is recommended by the National Kidney Foundation[10]. Chronic kidney disease was defined as an eGFR < 60 ml/min/1.73m^2 [11]. Prevalent diabetes mellitus was defined as a fasting plasma glucose level ≥ 7.0 mmol/L or use of anti-diabetic medication. Information on medication use, medical history and smoking behavior was collected via computerized questionnaires during home visits. Smoking was classified as current versus non-current smokers. The previously studied inflammatory markers were measured using the same multiplex immunoassay that was also used for the novel inflammatory biomarkers.

Statistical Analyses

In the first step, we used Cox proportional hazard models to investigate the age and sex adjusted association between each inflammatory biomarker and the incidence of CHD. All models met the proportional hazards assumption. Markers with a right-skewed distribution were transformed to the natural logarithmic scale (table S2). For a better comparison between the biomarkers, all markers were standardized by dividing the measured value by the standard deviation. We defined biomarker values as an outlier when the value was >4 standard deviations higher or lower than the mean. Participants were excluded from the analysis when the biomarker value for this person was an outlier. The maximum number of excluded individuals was 3 among all biomarkers. We selected the significant biomarkers from the first step to further assess their association with CHD in multivariable analyses. In this second step, we additionally adjusted the association for BMI, serum total cholesterol, HDL-cholesterol, systolic blood pressure, use of anti-hypertensive medication (defined as diuretics, anti-adrenergic agents, β blockers, calcium channel blockers and RAAS inhibitor), eGFR, prevalent type 2 diabetes and smoking. The hazard ratios were also calculated for the two upper tertiles with the first tertile as reference. In the third model, we additionally adjusted for the inflammatory markers that have previously been studied. In a sensitivity analysis, we excluded individuals with prevalent type 2 diabetes, chronic kidney disease and individuals using anti-hypertensive medication. Participants were censored at the time of occurrence of CHD, death, loss to follow-up or the end of the study period on January 1, 2011. We estimated 10-year risks for first-incident CHD for different tertiles of the identified biomarker(s). The cumulative incidence curves were created taking into account competing events[12,13].

In addition, we analyzed EN-RAGE with the different CHD outcomes separately (myocardial infarction, coronary revascularization and CHD mortality). To compare directly the effect estimates on these specific first CHD events using Cox regression, we applied the data augmentation method proposed by Lunn and McNeil[14]. This method estimates the difference in cause-specific hazard ratios of EN-RAGE on the specific CHD events when competing CHD events and non-CHD events are present[12]. We presented the results for the model in which we adjusted for the traditional CHD risk factors.

The measures of association are presented with 95% confidence intervals (CI). We hypothesized that inflammatory markers may predict the incidence of CHD. To this end, we tested the association between 16 markers of inflammation with the incidence of CHD. To avoid false positive findings, we applied a Bonferroni corrected p-value of 3.1×10^{-3} (0.05/16) as a robust threshold of significance. All other statistical tests were considered significant with a p-value < 0.05.

We compared the 10-year CHD risk prediction of the traditional Framingham risk score model to the new model that additionally included EN-RAGE using the c-statistic difference, continuous net reclassification improvement (NRI) and integrated discrimination improvement (IDI)[15,16,17]. The difference in c-statistic between the base model and the model with EN-RAGE was corrected for optimism using 100 bootstraps.

Approximately 5% of the participants lacked data on one or more of the cardiovascular covariates, except for the covariate “use of antihypertensive medication”, where 9% of the values were missing. Missing data for these covariates was imputed by multiple imputation where 5 datasets were pooled to obtain the risk estimates for the association between EN-RAGE and incident CHD[18,19]. Biomarkers with missing data due to values under the lower detection limit were imputed with the lower detection limit. Data were handled and analyzed using the IBM SPSS Statistics version 21.0.0.1 (IBM Corp., Somers, NY, USA) and R version 3.0.0[20].

Results

Table 1 summarizes the baseline characteristics of 839 participants (see table S2 for baseline characteristics in future CHD cases and non-cases). At the start of the

Table 1. Baseline characteristics of participants at risk for CHD.

Characteristics	Total (n=839)
Age, y	72.8±7.5
Men, n (%)	355 (42)
Body mass index, kg/m^2	27±4
Systolic blood pressure, mmHg	144±21
Diastolic blood pressure, mmHg	75±11
Antihypertensive medication use, n (%)	319 (38)
Total cholesterol, mmol/L	5.8±1.0
High-density lipoprotein cholesterol, mmol/L	1.4±0.4
Triglycerides, mmol/L	1.5±0.7
Current smokers, n (%)	144 (17.2)
Prevalent type 2 diabetes, n (%)	107 (12.8)
Estimated Glomerular Filtration Rate, ml/min/1.73m^2	74±15
CD40ligand*, ng/mL	0.028 (0.020–0.039)
Complement 3, mg/mL	0.84±0.14
C-reactive protein*, mg/L	1.43 (0.69–3.28)
Interleukin 8*, pg/mL	9.21 (7.20–12.40)
Interleukin 18*, pg/mL	190 (149–248)
Monocyte chemotactic protein*, pg/mL	183 (151–225)
Macrophage migration inhibitory factor*, ng/mL	0.056 (0.037–0.082)
Regulated on activation, normal T cell expressed and secreted*, ng/mL	0.50 (0.32–0.80)
Resistin*, ng/mL	0.42 (0.31–0.49)
Tumor necrosis factor receptor 2*, ng/mL	3.54 (2.93–4.34)
White blood cell count, x10^9/L	6.4±1.7

Plus-minus values are means ±SD. *Values are presented as median (inter-quartile range).

study, the mean (±SD) age was 72.8 (7.5) and 58% of the population were female. During a median follow-up of 10.6 years (interquartile range: 6.8–11.9), 2 were lost to follow-up, 353 individuals died (302 unrelated to CHD) and 99 developed CHD (incidence rate: 12.7 per 1000 person years). Out of the 16 inflammatory biomarkers, after Bonferroni correction, only EN-RAGE was significantly associated with CHD when adjusted for age and sex (table S3). The risk of CHD was nearly one third increased per standard deviation increase in the natural log-transformed EN-RAGE (Hazard Ratio (HR): 1.37; 95% confidence interval (CI): 1.12–1.67) (Table 2). Compared to the lowest tertile, participants in the highest tertile experienced approximately a 2.6 higher risk of developing CHD compared to participants in the lowest tertile (HR: 2.59; 95%CI: 1.52–4.40). When we further adjusted the association for traditional cardiovascular risk factors, the effect estimates attenuated slightly (HR: 1.30; 95%CI: 1.06–1.59). Additional adjustment for previously studied inflammatory markers yielded slightly increased effect estimates (Table 2, table S4).

Table 2. The association between EN-RAGE serum levels and incident CHD.

		*HR (95% CI)**	*HR (95% CI)**	*HR (95% CI)**
Tertile EN-RAGE	**n/N**	**Model 1**	**Model 2**	**Model 3**
First	20/277	1.00 (Reference)	1.00 (Reference)	1.00 (Reference)
Second	35/281	1.92 (1.11–3.33)	1.66 (0.95–2.90)	1.77 (1.00–3.15)
Third	44/279	2.59 (1.52–4.40)	2.15 (1.25–3.69)	2.47 (1.35–4.54)
P for trend		<0.001	0.006	0.006
Per SD Ln (EN-RAGE)	99/837	1.37 (1.12–1.67)	1.30 (1.06–1.59)	1.46 (1.11–1.90)

*Hazard ratios are represented per standard deviation increase in log-transformed EN-RAGE. Model 1: adjusted for age and sex. Model 2: adjusted for age, sex, BMI, systolic blood pressure, anti-hypertensive medication use, HDL- cholesterol, Total cholesterol, smoking status (current, non-current), prevalent type 2 diabetes, eGFR. Model 3: additionally adjusted for CD40ligand, Complement 3, C-reactive protein, interleukin 8, interleukin 18, monocyte chemotactic protein 1, macrophage migration inhibitory factor, RANTES, Resistin, TNF receptor 2 and white blood cells. HR=Hazard ratio, SD=Standard deviation, EN-RAGE=extracellular newly identified receptor for advanced glycation end-products binding protein.

Cumulative incidence curves for the tertiles of EN-RAGE adjusted for competing risks are depicted in Figure 2. The 10-year probability of first incident event of CHD was 0.05 (95%CI: 0.03 – 0.08) for the first tertile, 0.11 (95%CI: 0.07 – 0.14) for the second tertile and 0.14 (95%CI: 0.10 – 0.18) for the third tertile.

After excluding participants with chronic kidney disease at baseline, the association between EN-RAGE and incident CHD attenuated slightly (1.28 (95%CI: 1.03–1.59)) (Table 3). Excluding participants with type 2 diabetes, the effect estimates of the association between EN-RAGE and CHD did not change: hazard ratio 1.29 (95% CI:1.04–1.60) in the fully adjusted model. Finally, after excluding participants taking anti-hypertensive medication, the hazard ratio did not change (HR:1.40; 95% CI: 1.05–1.87).

Table 4 depicts the results for the associations between EN-RAGE and the different CHD manifestations separately. We observed the strongest association with CHD mortality (HR:1.56; 95% CI: 1.19–2.04) compared to myocardial infarction and revascularization which were not significant. Further adjustment for the traditional cardiovascular risk factors did not change the effect estimate for CHD mortality. Cause-specific HRs were not significantly lower for revascularization compared to myocardial infarction (Lunn and McNeil p-value=0.700), but they were borderline significant for CHD mortality compared to revascularization (Lunn and McNeil p-value=0.055).

Figure 2. Cumulative incidence curves for first, second and third tertile of serum EN-RAGE in relation to incidence of coronary artery disease adjusting for competing non-coronary heart disease death up to 10 years of follow-up.

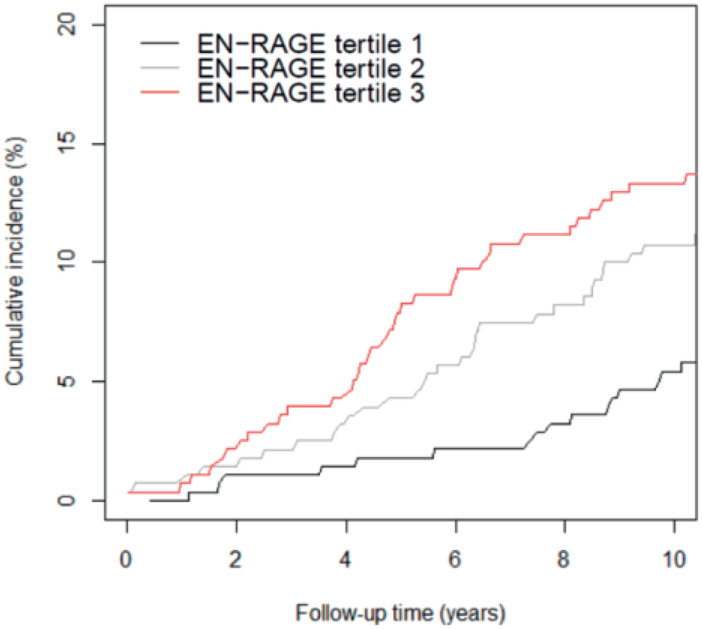


Table 3. The association of EN-RAGE with CHD in absence of patients with CKD, T2D or anti-hypertensive use.

		*HR (95% CI)**	*HR (95% CI)**
	n/N	**Model 1**	**Model 2**
eGFR < 60 excluded	81/720	1.34 (1.09–1.65)	1.28 (1.03–1.59)
Prevalent diabetes excluded	86/732	1.34 (1.09–1.64)	1.29 (1.04–1.60)
Anti-hypertensive use excluded	42/520	1.45 (1.10–1.92)	1.40 (1.05–1.87)

*Hazard ratios are represented per standard deviation increase in log-transformed EN-RAGE. Model 1: adjusted for age and sex. Model 2: adjusted for age, sex, BMI, systolic blood pressure, anti-hypertensive medication use, HDL- cholesterol, Total cholesterol, smoking status (current and non-current), prevalent type 2 diabetes, eGFR. HR=Hazard ratio, eGFR, estimated glomerular filtration rate, EN-RAGE=extracellular newly identified receptor for advanced glycation end-products binding protein.

Table 4. The association between EN-RAGE and incident myocardial infarction, coronary revascularization and CHD mortality.

		*HR (95% CI)**	*HR (95% CI)**
	n/N	**Model 1**	**Model 2**
Coronary revascularization	38/837	1.10 (0.77–1.55)	0.99 (0.68–1.44)
Myocardial infarction	51/837	1.32 (1.00–1.76)	1.30 (0.95–1.76)
CHD mortality	51/837	1.56 (1.19–2.04)	1.57 (1.17–2.10)

*Hazard ratios are represented per standard deviation increase in log-transformed EN-RAGE. Model 1: adjusted for age and sex. Model 2: adjusted for age, sex, BMI, systolic blood pressure, anti-hypertensive medication use, HDL- cholesterol, Total cholesterol, smoking status (current and non-current), prevalent type 2 diabetes, eGFR. HR=Hazard ratio, eGFR, estimated glomerular filtration rate, EN-RAGE=extracellular newly identified receptor for advanced glycation end-products binding protein.

The c-statistic of the traditional Framingham risk score model for 10-year CHD risk was 0.730 (95% CI: 0.672–0.788). When we added EN-RAGE to the model, the c-statistic improved to 0.741 (95% CI:0.683–0.799) with a difference of 0.011 (95% CI: -0.012–0.033, p-value=0.33). Adding EN-RAGE to the Framingham risk score model resulted in a significant continuous net reclassification index (NRI) of 0.36 (95% CI:0.05-0.67, p-value=0.02) and integrated discrimination improvement of 0.026 (95% CI:0.009–0.0437, p-value=3.3×10^{-3}).

Discussion

In this prospective, population-based cohort study, we found that higher EN-RAGE levels were associated with an increased risk of CHD beyond conventional risk factors. Further adjustments for inflammatory markers as well as excluding diseased individuals did not change the results. These findings suggest pro-inflammatory EN-RAGE as an new inflammatory risk marker for CHD that represents a distinct inflammatory pathway compared to other inflammatory markers.

Previous studies have observed increased levels of EN-RAGE in patients with chronic inflammatory disorders including inflammatory bowel disease (IBD)[21], type 2 diabetes (T2D)[22], chronic kidney disease (CKD), subclinical atherosclerosis[23,24] and coronary artery disease[25,26,27,28,29]. The design of the mentioned studies is mainly cross-sectional and conducted in patient populations. A positive association between EN-RAGE and (CHD) mortality was shown in a prospective study including patients on hemodialysis[30]. In addition, a recent study in Japanese CHD individuals observed a significant association between EN-RAGE and future major adverse cardiac events[31]. To our knowledge, we are the first to investigate the association between EN-RAGE and CHD in a prospective population-based cohort study with long-term follow-up.

To address the possibility of confounding, we adjusted in the multivariable model for the different traditional CHD risk factors and previously studied inflammatory markers. To address the question whether our results were driven by a certain subgroup, we analyzed the data excluding participants with chronic kidney disease, T2D and antihypertensive use in the sensitivity analyses. Across all these analyses, there was a consistent effect of EN-RAGE on the risk of CHD, even after adjusting for the established inflammatory markers. These results suggest an effect of EN-RAGE on the risk of CHD beyond well-established metabolic and inflammatory pathways.

We observed a stronger association between EN-RAGE and future myocardial infarction and CHD mortality compared to revascularization. This suggests that EN-RAGE is more a determinant of acute coronary events with plaque instability rather than stable coronary artery disease. This observation that EN-RAGE, a member of the S100 protein family, is a strong determinant of acute coronary events is in line with previous studies that reported higher levels of mRNA and plasma levels of family S100 proteins (S100A8/9) in patients with ST-elevated myocardial infarction compared to stable CAD cases[32]. Furthermore, a post-mortem study in people died from sudden cardiac death has found high expression levels of S100A12 in coronary artery smooth muscle in the ruptured plaques, especially in diabetics[33]. However, the cause-specific hazard ratio for the CHD events were not significantly different using the method proposed by Lunn and McNeil. We might have been underpowered to observe a significant difference due to the limited number of cases in this cause-specific analyses.

Studying the added value of EN-RAGE in 10-year CHD risk prediction, we found an improvement in risk prediction when we added EN-RAGE to the Framingham risk score. This suggests that EN-RAGE, as a non-invasive marker of future CHD, could be useful in predicting the risk of CHD in the general population. Although we corrected the change in c-statistic for optimism, we believe that further studies are needed to establish the potential role of EN-RAGE in CHD risk prediction.

EN-RAGE, a member of the S100 protein family of EF-hand calcium-binding proteins, is an endogenously produced inflammatory ligand of the Receptor of Advanced Glycation End products (RAGE)[34] and Toll-like receptor 4 (TLR4)[35]. RAGE is a member of the immunoglobulin superfamily of cell surface molecules and is expressed in multiple tissues including endothelium cells, vascular smooth muscle cells and monocyte derived macrophages[36]. The binding of RAGE by EN-RAGE activates inflammatory cascades, including the pro-inflammatory NF-κB signaling pathway, a well-known pathway of the innate immune system involved in the pathogenesis of CHD[34,37]. Moreover, intracellular signaling pathways triggered by EN-RAGE may alter gene expression and up-regulate the synthesis of vascular cell adhesion molecule-1 and intracellular adhesion molecule-1 synthesis[34]. Considering atherosclerosis as a chronic inflammatory disease, the engagement of RAGE by EN-RAGE may play an important role in the pathogenesis of atherosclerosis and subsequently CHD. In line with this evidence, the expression of EN-RAGE in vascular smooth muscle cells can modulate the remodeling of the aortic wall and stimulates cytokine production and increases oxidative stress[38]. Moreover, EN-RAGE accelerates atherosclerosis and vascular calcification in Apolipoprotein E-Null mice[39]. Recently EN-RAGE has been shown to accelerate the development of cardiac hypertrophy and diastolic dysfunction in mice with CKD[40]. The monocyte activation effect of EN-RAGE has also been observed to be TLR4 dependent[35]. It was demonstrated that EN-RAGE facilitates inflammatory monocyte activation by TLR4 and that this effect was modulated by RAGE. Toll-like receptors have been investigated extensively in the field of cardiovascular diseases as they are expressed in vascular and myocardial cells membranes[41]. The important role of EN-RAGE in the pathogenesis of atherosclerosis is further emphasized by a recent study where pharmacological inhibition of S100A12-mediated atherosclerosis improved atherosclerotic plaque features including smaller necrotic cores, diminished calcification and reduced number of inflammatory cells[42].

This study has certain strengths and limitations. The prospective population-based study design, the diversity of the available inflammatory biomarkers and the long-term follow-up of CHD can be marked as the main strengths of the current study. In addition, our findings are robust regarding the strict Bonferroni p-value we used as the threshold for significant associations in the first step. A number of limitations should also be acknowledged. First, although we adjusted our analysis for different potential confounders, we cannot exclude the effect of unknown or unmeasured confounders. However, since we adjusted for the traditional and commonly used risk factors for CHD and inflammatory pathways, we believe that EN-RAGE as a novel inflammatory marker for CHD is interesting since it might reflect other pathways that lead to CHD. Second, we had to exclude inflammatory biomarkers with very low serum concentrations. Nonetheless, the selected biomarkers have more than 60% completeness of measurements indicating acceptable quality of quantification. Third, our population is 55 years and older. Therefore, generalization of the results to a younger age category should be with caution. Our study only indicates an association, we think that further studies are needed to establish the causal role of EN-RAGE in the pathogenesis of CHD.

In conclusion, our study suggests that higher levels of serum EN-RAGE are associated with incidence of CHD beyond conventional cardiovascular risk factors and inflammatory markers. These results provide evidence for a role of EN-RAGE in the development of CHD and suggest this marker as a potential target for drug therapy and risk prediction.

References

1. Alwan A. Global status report on noncommunicable diseases 2010: World Health Organization; 2011.

2. Libby P. Inflammation in atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology* 2012; 32(9): 2045-51.

3. Danesh J, Wheeler JG, Hirschfield GM, et al. C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *New England Journal of Medicine* 2004; 350(14): 1387-97.

4. Danesh J, Kaptoge S, Mann AG, et al. Long-term interleukin-6 levels and subsequent risk of coronary heart disease: two new prospective studies and a systematic review. *PLoS medicine* 2008; 5(4): e78.

5. Pai JK, Pischon T, Ma J, et al. Inflammatory markers and the risk of coronary heart disease in men and women. *New England Journal of Medicine* 2004; 351(25): 2599-610.

6. Kaptoge S, Seshasai SRK, Gao P, et al. Inflammatory cytokines and risk of coronary heart disease: new prospective study and updated meta-analysis. *European heart journal* 2013: eht367.

7. Hansson GK, Hermansson A. The immune system in atherosclerosis. *Nature immunology* 2011; 12(3): 204-12.

8. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol* 2013; 28(11): 889-926.

9. Leening MJ, Kavousi M, Heeringa J, et al. Methods of data collection and definitions of cardiac outcomes in the Rotterdam Study. *Eur J Epidemiol* 2012; 27(3): 173-85.

10. Perrone RD, Madias NE, Levey AS. Serum creatinine as an index of renal function: new insights into old concepts. *Clin Chem* 1992; 38(10): 1933-53.

11. Bash LD, Coresh J, Kottgen A, et al. Defining incident chronic kidney disease in the research setting: The ARIC Study. *Am J Epidemiol* 2009; 170(4): 414-24.

12. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine* 2007; 26(11): 2389-430.

13. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. *British Journal of Cancer* 2004; 91(7): 1229-35.

14. Lunn M, McNeil D. Applying Cox regression to competing risks. *Biometrics* 1995: 524-32.

15. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis: Springer; 2001.

16. Steyerberg EW. Clinical prediction models: Springer; 2009.

17. Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. *Annals of internal medicine* 2010; 152(3): 195-6; author reply 6-7.

18. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; 91(434): 473-89.

19. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology* 1995; 142(12): 1255-64.

20. Team RC. R: A language and environment for statistical computing. *R foundation for Statistical Computing* 2005.

21. Foell D, Kucharzik T, Kraft M, et al. Neutrophil derived human S100A12 (EN-RAGE) is strongly expressed during chronic active inflammatory bowel disease. *Gut* 2003; 52(6): 847-53.

22. Kosaki A, Hasegawa T, Kimura T, et al. Increased plasma S100A12 (EN-RAGE) levels in patients with type 2 diabetes. *Journal of Clinical Endocrinology & Metabolism* 2004; 89(11): 5423-8.

23. Mori Y, Kosaki A, Kishimoto N, et al. Increased plasma S100A12 (EN-RAGE) levels in hemodialysis patients with atherosclerosis. *American journal of nephrology* 2008; 29(1): 18-24.

24. Shiotsu Y, Mori Y, Nishimura M, et al. Plasma S100A12 level is associated with cardiovascular disease in hemodialysis patients. *Clinical Journal of the American Society of Nephrology* 2011; 6(4): 718-23.

25. Tydén H, Lood C, Gullstrand B, et al. Increased serum levels of S100A8/A9 and S100A12 are associated with cardiovascular disease in patients with inactive systemic lupus erythematosus. *Rheumatology* 2013; 52(11): 2048-55.

26. Rosenberg S, Elashoff MR, Beineke P, et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. *Annals of internal medicine* 2010; 153(7): 425-34.

27. Zhao P, Wu M, Yu H, et al. Serum S100A12 levels are correlated with the presence and severity of coronary artery disease in patients with type 2 diabetes mellitus. *Journal of Investigative Medicine* 2013; 61(5): 861-6.

28. Mahajan N, Malik N, Bahl A, Dhawan V. Receptor for advanced glycation end products (RAGE) and its inflammatory ligand EN-RAGE in non-diabetic subjects with pre-mature coronary artery disease. *Atherosclerosis* 2009; 207(2): 597-602.

29. Liu J, Ren Y-G, Zhang L-H, Tong Y-W, Kang L. Serum S100A12 concentrations are correlated with angiographic coronary lesion complexity in patients with coronary artery disease. *Scandinavian Journal of Clinical & Laboratory Investigation* 2014; 74(2): 149-54.

30. Nakashima A, Carrero JJ, Qureshi AR, et al. Effect of circulating soluble receptor for advanced glycation end products (sRAGE) and the proinflammatory RAGE ligand (EN-RAGE, S100A12) on mortality in hemodialysis patients. *Clinical Journal of the American Society of Nephrology* 2010; 5(12): 2213-9.

31. Saito T, Hojo Y, Ogoyama Y, et al. S100A12 as a marker to predict cardiovascular events in patients with chronic coronary artery disease. *Circulation journal: official journal of the Japanese Circulation Society* 2011; 76(11): 2647-52.

32. Healy AM, Pickard MD, Pradhan AD, et al. Platelet expression profiling and clinical validation of myeloid-related protein-14 as a novel determinant of cardiovascular events. *Circulation* 2006; 113(19): 2278-84.

33. Burke AP, Kolodgie FD, Zieske A, et al. Morphologic findings of coronary atherosclerotic plaques in diabetics a postmortem study. *Arteriosclerosis, thrombosis, and vascular biology* 2004; 24(7): 1266-71.

34. Hofmann MA, Drury S, Fu C, et al. RAGE mediates a novel proinflammatory axis: a central cell surface receptor for S100/calgranulin polypeptides. *Cell* 1999; 97(7): 889-901.

35. Foell D, Wittkowski H, Kessel C, et al. Pro-inflammatory S100A12 can Activate Human Monocytes via Toll-like Receptor 4. *American journal of respiratory and critical care medicine* 2013; (ja).

36. Brett J, Schmidt AM, Du Yan S, et al. Survey of the distribution of a newly characterized receptor for advanced glycation end products in tissues. *The American journal of pathology* 1993; 143(6): 1699.

37. Hansson GK, Libby P, Schönbeck U, Yan Z-Q. Innate and adaptive immunity in the pathogenesis of atherosclerosis. *Circulation research* 2002; 91(4): 281-91.

38. Bowman MH, Wilk J, Heydemann A, et al. S100A12 mediates aortic wall remodeling and aortic aneurysm. *Circulation research* 2010; 106(1): 145-54.

39. Bowman MAH, Gawdzik J, Bukhari U, et al. S100A12 in vascular smooth muscle accelerates vascular calcification in apolipoprotein E–null mice by activating an osteogenic gene regulatory program. *Arteriosclerosis, thrombosis, and vascular biology* 2011; 31(2): 337-44.

40. Yan L, Mathew L, Chellan B, et al. S100/Calgranulin-Mediated Inflammation Accelerates Left Ventricular Hypertrophy and Aortic Valve Sclerosis in Chronic Kidney Disease in a Receptor for Advanced Glycation End Products–Dependent Manner. *Arteriosclerosis, thrombosis, and vascular biology* 2014: ATVBAHA. 114.303508.

41. Frantz S, Ertl G, Bauersachs J. Mechanisms of disease: Toll-like receptors in cardiovascular disease. *Nature clinical practice cardiovascular medicine* 2007; 4(8): 444-54.

42. Yan L, Bjork P, Butuc R, et al. Beneficial effects of quinoline-3-carboxamide (ABR-215757) on atherosclerotic plaque morphology in S100A12 transgenic ApoE null mice. *Atherosclerosis* 2013; 228(1): 69-79.

Supplementary material

Table S1. List of previously investigated inflammatory markers with incident Coronary Heart Disease.

Biomarker
- C-reactive protein[1]
- CD40 Ligand[2,3]
- Complement factor 3[4]
- Interleukin 18[5]
- Monocyte chemotactic protein-1[6]
- Regulated on activation, normal T cell expressed and secreted[7]
- Resistin[8]
- TNF receptor II[9]
- Interleukin 8[10]
- Macrophage inhibitory factor[11]

Table S2. Differences in baseline characteristics between future CHD cases and CHD non-cases.

Characteristics	CHD cases (n=99)	CHD non-cases (n=740)	P-value
Age, y	74.0±7.9	72.6±7.4	0.09
Men, n (%)	48(49)	307(42)	0.19
Body mass index, kg/m^2	27±4	27±4	0.06
Systolic blood pressure, mmHg	148±20	144±22	0.06
Diastolic blood pressure, mmHg	75±12	75±11	0.66
Antihypertensive medication use, n (%)	56 (57)	260 (35)	<0.001
Total cholesterol, mmol/L	5.9±1.2	5.8±0.9	0.48
HDL-cholesterol, mmol/L	1.3±0.3	1.4±0.4	0.002
Triglycerides, mmol/L	1.6±0.7	1.5±0.8	0.18
Current smokers, n (%)	20 (20.2)	124 (16.8)	0.35
Prevalent type 2 diabetes, n (%)	13 (13.1)	94 (12.7)	0.90
eGFR, ml/min/1.73m^2	74±14	74±15	0.63
CD40ligand*, ng/mL	0.028 (0.020–0.038)	0.028(020–0.038)	0.87
Complement 3, mg/mL	0.85±0.14	0.84±0.14	0.38
C-reactive protein*, mg/L	1.64 (0.78–3.73)	1.42 (0.68–3.13)	0.36
Interleukin 8*, pg/mL	10.40 (7.36–13.40)	9.15 (7.01–12.40)	0.06
Interleukin 18*, pg/mL	206 (150–270)	188 (149–245)	0.12
MCP1*, pg/mL	189 (151–236)	183 (151–225)	0.50
Macrophage migration inhibitory factor*, ng/mL	0.059 (0.032–0.091)	0.055 (0.037–0.082)	0.79
RANTES*, ng/mL	0.46 (0.33–0.75)	0.51 (0.32–0.81)	0.51
Resistin*, ng/mL	0.40 (0.27–0.64)	0.43 (0.31–0.59)	0.68
TNFR-II*, ng/mL	3.64 (2.97–4.45)	3.52 (2.91–4.33)	0.24

Plus-minus values are means ±SD. *Values are presented as median (inter-quartile range). CD40ligand, cluster of differentiation 40 ligand; eGFR, estimated glomerular filtration rate; HDL-cholesterol, high-density lipoprotein cholesterol; MCP1, Monocyte chemotactic protein; RANTES, regulated on activation, normal T cell expressed and secreted; TNFR-II, Tumor necrosis factor receptor 2.

Table S3. Biomarkers and their age and sex adjusted association results with incident Coronary Heart Disease.

Marker[1]	Median (SD/IQR)[1]	N[2]	Beta (95% CI)	P-value	nr. under LOD
CD40, ng/mL	0.70 (0.58–0.83)	836	1.11 (0.89–1.39)	0.38	0
CFH, µg/mL	2455.6 (838.9)	839	1.00 (1.00–1.00)	0.27	90[3]
EN-RAGE, ng/mL	10.80 (7.66–14.70)	837	1.37 (1.12–1.67)	0.002	0
Eotaxin, pg/mL	161 (116–217)	838	1.11 (0.90–1.37)	0.34	3
FASLR, ng/mL	4.69 (3.97–5.54)	832	1.06 (0.85–1.31)	0.62	0
HCC4, ng/mL	4.90 (1.96)	838	1.17 (0.97–1.41)	0.11	0
IL13, pg/mL	4.32 (4.09–4.52)	838	0.97 (0.79–1.19)	0.76	30
IL16, pg/mL	5.94 (5.74–6.08)	837	1.03 (0.83–1.27)	0.79	0
IL17, pg/mL	13.67 (5.20)	838	0.88 (0.72–1.08)	0.23	47
IL1ra, pg/mL	68.5 (47.75–102.00)	838	1.19 (0.97–1.46)	0.10	20
MDC, pg/mL	352 (294 -419)	836	1.09 (0.89–1.33)	0.42	0
MIP1alpha, pg/mL	45 (38–56)	835	1.21 (0.98–1.49)	0.07	4
MIP1beta, pg/mL	122 (95–153)	828	0.92 (0.73–1.17)	0.51	0
PARC, ng/mL	3.38 (3.18–3.56)	834	0.97 (0.78–1.20)	0.75	0
sRAGE, ng/mL	2.66 (1.94–3.63)	839	0.99 (0.81–1.21)	0.92	0
TRAILR3, mg/mL	6.62 (5.16–8.41)	837	0.90 (0.73–1.10)	0.28	0

[1]Markers that were not following a normal distribution were log transformed and presented as median and interquartile range. Measures are presented based on non-imputed values. [2]Samples included in analysis, outliers excluded. [3]CFH values are missing due to insufficient quantity of serum. CD40, cluster of differentiation 40; CFH, Complement Factor H; EN-RAGE, Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein; FASLR, Fas Ligand Receptor; HCC4, Human CC chemokine-4; IL13, interleukin 13; IL17, interleukin 16; IL17, interleukin 17; IL1ra, interleukin 1 receptor antagonist; LOD, limit of detection; MDC, Monocyte Derived Chemokine; MIP1alpha, Macrophage Inflammatory Protein 1 alpha; MIP1beta, Macrophage Inflammatory Protein 1 beta; PARC, Pulmonary and Activation-Regulated Chemokine; sRAGE, soluble Receptor of Advanced Glycation End-products; TRAILR3, TNF-Related Apoptosis-Inducing Ligand Receptor.

Table S4. Effect estimates for all covariates included in model 3 per standard deviation.

Covariate	HR (95% CI)	P-value
Age	1.05 (1.01–1.08)	0.004
Gender	0.66 (0.43–1.01)	0.05
EN-RAGE*	1.41 (1.13–1.76)	0.002
CD40*	1.06 (0.80–1.40)	0.68
Complement 3	1.08 (0.86–1.37)	0.51
C-reactive protein*	0.93 (0.73–1.19)	0.57
Interleukin 18*	1.11 (0.90–1.37)	0.33
Interleukin 8*	1.17 (0.94–1.46)	0.15
Monocyte chemotactic protein 1*	0.93 (0.75–1.16)	0.52
Macrophage migration inhibitory factor*	0.97 (0.77–1.22)	0.79
RANTES*	0.84 (0.68–1.04)	0.11
Resistin*	0.85 (0.68–1.06)	0.14
Tumor necrosis factor receptor II*	1.08 (0.81–1.43)	0.61

EN-RAGE, extracellular newly identified receptor for advanced glycation end products binding protein; RANTES, regulated on activation, normal T cell expressed and secreted. *Markers were natural log-transformed.

Supplementary references

1. Koenig W, Sund M, Fröhlich M, et al. C-reactive protein, a sensitive marker of inflammation, predicts future risk of coronary heart disease in initially healthy middle-aged men results from the MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) Augsburg Cohort Study, 1984 to 1992. *Circulation* 1999; 99(2): 237-42.

2. Jefferis BJ, Whincup PH, Welsh P, et al. Prospective study of circulating soluble CD40 ligand concentrations and the incidence of cardiovascular disease in a nested prospective case–control study of older men and women. *Journal of Thrombosis and Haemostasis* 2011; 9(8): 1452-9.

3. Kaptoge S, Seshasai SRK, Gao P, et al. Inflammatory cytokines and risk of coronary heart disease: new prospective study and updated meta-analysis. *European heart journal* 2013: eht367.

4. Muscari A, Bozzoli C, Puddu GM, et al. Association of serum C3 levels with the risk of myocardial infarction. *The American journal of medicine* 1995; 98(4): 357-64.

5. Blankenberg S, Luc G, Ducimetière P, et al. Interleukin-18 and the risk of coronary heart disease in European men the Prospective Epidemiological Study of Myocardial Infarction (PRIME). *Circulation* 2003; 108(20): 2453-9.

6. Hoogeveen RC, Morrison A, Boerwinkle E, et al. Plasma MCP-1 level and risk for peripheral arterial disease and incident coronary heart disease: Atherosclerosis Risk in Communities study. *Atherosclerosis* 2005; 183(2): 301-7.

7. Herder C, Peeters W, Illig T, et al. RANTES/CCL5 and risk for coronary events: results from the MONICA/KORA Augsburg case-cohort, Athero-Express and CARDIoGRAM studies. *PloS one* 2011; 6(12): e25734.

8. Weikert C, Westphal S, Berger K, et al. Plasma resistin levels and risk of myocardial infarction and ischemic stroke. *Journal of Clinical Endocrinology & Metabolism* 2008; 93(7): 2647-53.

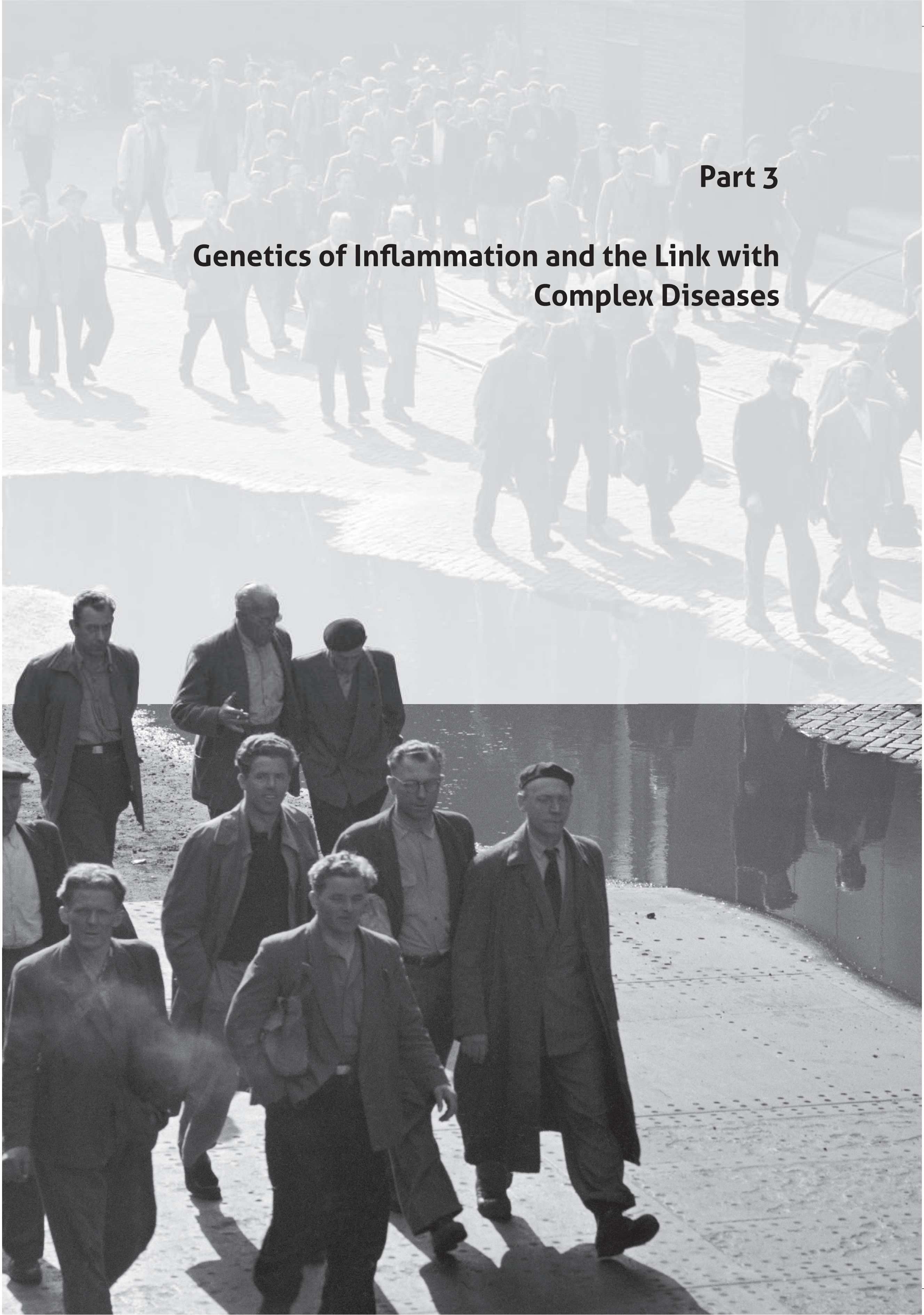
9. Pai JK, Pischon T, Ma J, et al. Inflammatory markers and the risk of coronary heart disease in men and women. *New England Journal of Medicine* 2004; 351(25): 2599-610.

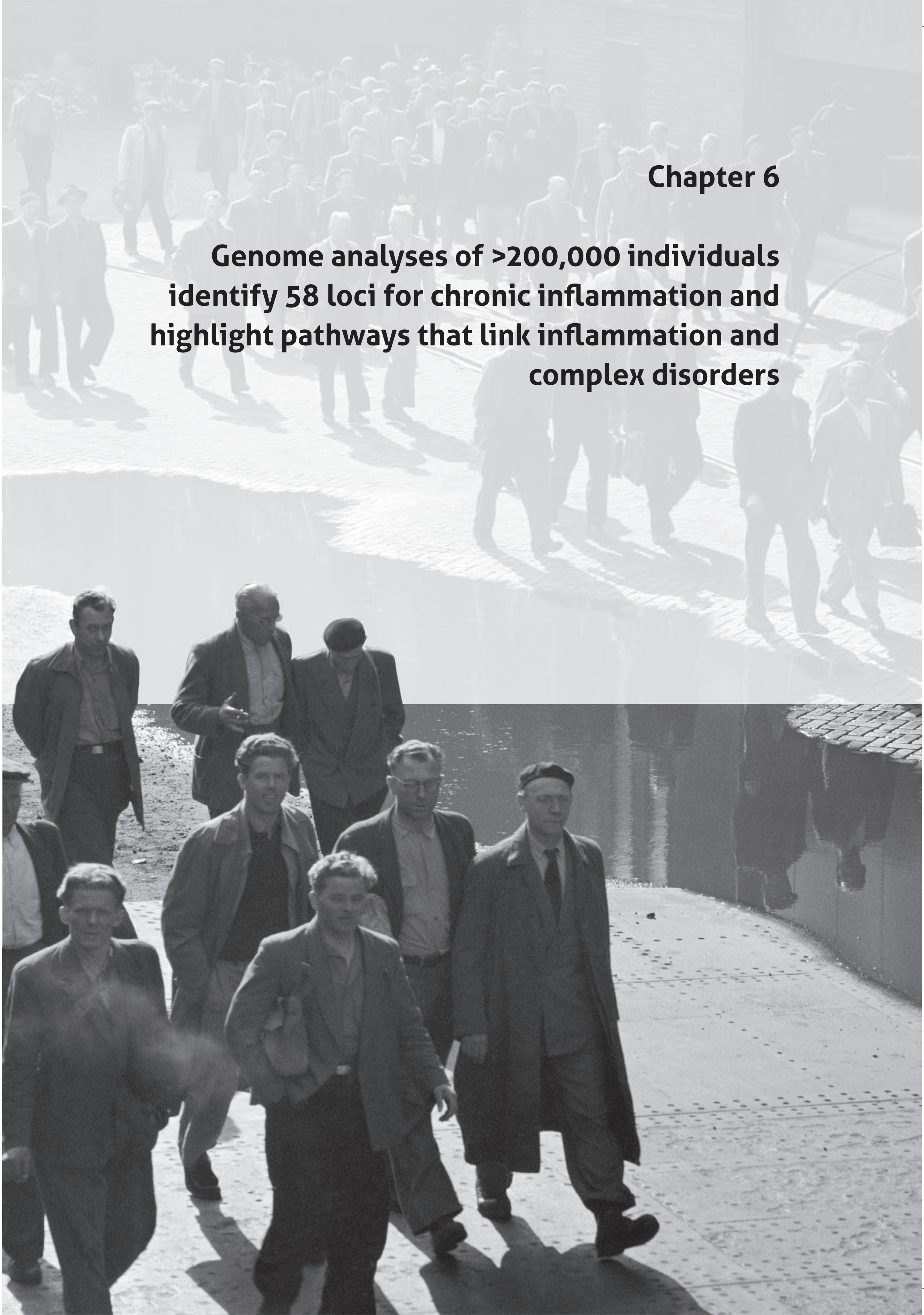
10. Boekholdt SM, Peters RJG, Hack CE, et al. IL-8 Plasma Concentrations and the Risk of Future Coronary Artery Disease in Apparently Healthy Men and Women The EPIC-Norfolk Prospective Population Study. *Arteriosclerosis, thrombosis, and vascular biology* 2004; 24(8): 1503-8.

11. Herder C, Illig T, Baumert J, et al. Macrophage migration inhibitory factor (MIF) and risk for coronary heart disease: results from the MONICA/KORA Augsburg case-cohort study, 1984–2002. *Atherosclerosis* 2008; 200(2): 380-8.

Part 3

Genetics of Inflammation and the Link with Complex Diseases





Chapter 6

Genome analyses of >200,000 individuals identify 58 loci for chronic inflammation and highlight pathways that link inflammation and complex disorders

Background: C-reactive protein (CRP) is a sensitive biomarker of chronic low-grade inflammation that is associated with multiple complex diseases. The genetic determinants of chronic inflammation remain largely unknown, and the causal role of CRP in several clinical outcomes is debated.

Methods: We performed genome-wide association analyses of circulating CRP levels in 204,402 European individuals. Additionally, we performed *in silico* functional analyses and Mendelian randomization analyses with several clinical outcomes.

Results: We identify 42 novel distinct CRP-associated loci (P-value<5×10^{-8}). The lead variants at the distinct loci explained up to 7.0% of the variance in circulating CRP levels. We identified 66 gene sets that were organized in two substantially correlated clusters, one mainly comprised of immune pathways, and the other characterized by metabolic pathways in the liver. Mendelian randomization analyses revealed a causal protective effect of CRP on schizophrenia and a risk increasing effect on bipolar disorder.

Conclusion: Our findings provide further insights in the biology of inflammation that may lead to novel interventions to treat inflammation and its clinical consequences.

Introduction

Inflammation plays a key role in the development of complex diseases such as cardiovascular disease[1], type 2 diabetes[2], Alzheimer's disease[3], and schizophrenia[4]. C-reactive protein (CRP) is a sensitive marker of chronic low-grade inflammation[5], and elevated serum levels of CRP have been associated with a wide range of diseases[6,7,8]. Unraveling the genetics of inflammation may provide further insights into the underlying biology of inflammation, and may identify novel therapeutic targets to attenuate inflammation.

The genetic determinants of CRP have only been partly characterized. In 2011, our group published a HapMap-based meta-analysis of genome-wide association studies (GWAS) including a discovery panel of up to 65,000 individuals and found 18 loci that were associated with CRP levels[9]. Increasing study sample size in GWAS and denser mapping of the genome with further advanced imputation panels may help to identify further genes associated with the phenotypes of interest[10,11]. Furthermore, by using genetic instrumental variables (i.e. a genetic score), Mendelian randomization (MR) allows investigation of the potential causal effect of an exposure on clinical outcomes, and may help to understand the causal pathways that link the exposure with the outcome[12]. The causal role of CRP in the development of diseases is still controversial[13], and the causal pathways that link inflammation to complex disorders are only partly understood.

We applied two large-scale GWAS on circulatory levels of CRP using HapMap and 1000Genomes (1KG) imputed data to identify genetic determinants of chronic inflammation. Because body mass index (BMI) is a major determinant of CRP levels, we additionally conducted GWAS adjusted for BMI to identify associated loci independent of BMI. To identify any sex differences in genetic determinants of chronic inflammation, we further conducted GWAS in men and women separately. We applied *in silico* functional analyses on the identified loci to obtain better insights into the biological processes potentially regulating chronic inflammation. Finally, MR analyses were conducted to provide an improved understanding of the causal relation between CRP and several related clinical outcomes.

Methods

GWAS for circulating CRP levels

We conducted a meta-analysis of GWAS including individuals of European ancestry within the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium Inflammation Working Group of the (CIWG)[14]. The CIWG invited cohorts for participation in the HapMap GWAS meta-analysis of CRP levels in 2012. In 2014 and in the light of our assessment which showed complementary value of HapMap and 1KG imputed GWAS[10], studies were further invited to participate in the 1KG GWAS meta-analysis. The 1KG GWAS may help to identify loci that were not covered in the HapMap GWAS and fine map loci found in the HapMap GWAS. Cohorts were allowed to participate in either the HapMap or 1KG GWAS, or both. Here we present both a HapMap (204,402 individuals from 78 studies) and 1KG (148,164 individuals from 49 studies) imputed genotypes GWAS meta-analysis. All participating cohorts implemented a pre-specified study plan comprising study design, data quality check, data analysis, and data sharing. Serum CRP was measured in mg/L using standard laboratory techniques (Supplemental Methods), and values were natural log-transformed. Individuals with auto-immune diseases, individuals taking immune-modulating agents (if this information was available), and individuals with CRP levels 4SD or more away from the mean were excluded from all analyses. The characteristics of the participants are presented in Table S1. Individuals and genetic variants were filtered based on study-specific quality control criteria (Table S2). At each individual study site, genetic variants were tested for association with CRP levels using an additive linear regression model adjusted for age, sex, and population substructure, and accounting for relatedness, if relevant. Before meta-analysis, variants were filtered based on imputation quality at R^2 index of >0.4. To avoid type-I error inflation, study-specific GWAS were corrected for genomic inflation. For the HapMap study, fixed effect meta-analyses were conducted for each genetic variant, using the inverse variance-weighted method implemented in GWAMA[15]. For the 1KG imputed GWAS, METAL[16] was used to perform a fixed effect meta-analysis. We removed variants that were only available in <50% of the samples. The HapMap meta-analysis included 2,254,727 variants, and the 1KG GWAS included 10,019,203 variants. Associations with P-value<5×10^{-8} were considered genome-wide significant. We used a stringent distance criterion, 500kb minimum between two significant variants, to identify distinct loci. In each locus, the variant with the smallest p-value was called the “lead variant”. Additionally, sex-stratified analyses were performed among HapMap imputed studies, and we tested for heterogeneity between sex-specific effect estimates as described previously[17]. The false-discovery rate of Benjamini-Hochberg was used to assess significance of the P-value for sex difference (<0.05). BMI adjusted analyses were conducted in the 1KG meta-analysis to determine the role of BMI in mediating the genetic associations with CRP, and to increase power to detect associations not mediated by BMI.

LD Score regression

Because population stratification is a major concern in GWAS and may lead to false-positive associations, we applied Linkage Disequilibrium Score regression (LDSC) to distinguish whether the inflation of test statistics observed in the CRP GWAS is due to the polygenic architecture of CRP or reflects confounding bias due to cryptic relatedness or population stratification. The LD Score measures collective genetic variation acquired from all genetic variants in LD with the index tagging (causal) variant[18]. A higher LD score of an index variant implicates more nearby genetic variants in high LD with the index variant, which makes it more likely that the index variant tags causal variant(s). More genetic variants in LD with the index genetic variant (i.e. a higher LD score due to polygenicity) may yield higher (i.e. inflated) test statistics. In contrast, higher test statistics caused by cryptic population stratification will not be related to LD score. LD Score regression analysis performs regression of the summary statistics from the GWAS meta-analysis (χ^2 statistics from the GWAS) on the LD scores across the genome. An intercept of the LD Score regression that equals one suggests no confounding bias, whereas an inflated intercept (larger than one) suggests contribution of confounding due to relatedness to the test statistics. We used the LDHub web interface to perform LD Score regression[19]. Variants were filtered to the subset of HapMap 3 variants, and variants with duplicated rs numbers, ambiguous variants, minor allele frequency (MAF)<0.01, and reported sample size <66.7% of total sample size were excluded. The default European LD Score file based on the European 1KG reference panel was used.

Furthermore, we applied cross-trait LD score regression to estimate genetic correlation of chronic inflammation (using the HapMap GWAS meta-analysis) with other phenotypes using published GWAS summary statistics[20]. In brief, the cross-product of two GWAS test statistics is calculated at each genetic variant, and this cross-product is regressed on the LD Score. The slope of the regression is used to estimate the genetic covariance between two phenotypes.

Identification of additional distinct variants in associated loci

To identify additional distinct variants in the associated loci, we performed joint approximate conditional analysis using the 1KG meta-analysis summary statistics and the linkage disequilibrium (LD) matrix derived from the first cohort of the Rotterdam study (RS-I) (n=5,974). We used the Genome-wide Complex Trait Analysis (GCTA) tool, which performs a genome-wide step-wise procedure to identify variants according to their distinct association with CRP (i.e. conditional P value)[21,22]. Only variants with an imputation quality of R^2>0.8 in the reference set (RS-I) were used. This approximate conditional analysis may reveal different lead signals in a locus where multiple associated variants are in the final joint association model. The distinct variants identified in the *CRP* gene were tested jointly for an association with CRP using individual level data from the second and third cohort of the Rotterdam Study (RS-II and RS-III, totaling 5,024 subjects), and the Women's Genome Health Study (WGHS) of 16,299 individuals.

Proportion of CRP variance explained

The variance explained in serum CRP levels was estimated using the formula *(2*MAF(1-MAF)beta^2)/var(CRP)*, where beta is the estimated effect of the individual variants on CRP[23] and var(CRP) is the variance in natural log-transformed CRP estimated in the RS-I cohort. We calculated the variance explained for four combinations of associated variants: 1. the lead variant at just the *CRP* locus; 2. the distinct variants at the CRP locus derived from the 1KG joint conditional analysis; 3. all lead variants in the distinct loci; 4. all lead variants in the distinct loci and, when applicable, the distinct variants at associated loci derived from the approximate joint conditional analysis.

Pathway analysis and gene expression

We used Data-Driven Expression-Prioritized Integration for Complex Traits (DEPICT v.1 rel173 beta)[24] to systematically prioritize the most likely causal genes, highlight the pathways that are enriched by the likely causal genes and identify tissues and cell types in which genes from associated loci are highly expressed. DEPICT requires summary statistics from the GWAS meta-analysis. First, genome-wide associated variants from both GWAS meta-analyses were filtered by MAF>0.01, and variants with low correlation with other variants were selected by PLINK (version 1.90) using a clumping distance of 500 kb apart and/or index of LD r^2 threshold <0.1. The settings for the analysis involved the usage of 1KG pilot phase data[25] (phase 1 integrated release, version 3, CEU, GBR, TSI unrelated individuals; 2010.11.23) with r^2>0.5 LD threshold for locus definition, 10,000 permutations for bias correction, and 500 repetitions for FDR calculation. To summarize and visualize the results, pairwise Pearson correlation coefficients were calculated between all gene-specific Z-scores for every pair of reconstituted DEPICT gene sets. Affinity Propagation Clustering (apcluster command; *APCluster* R package[26]) was used to identify clusters and representative examples of the clusters, and Cytoscape v3.2.1 was used for visualization of the results. The results of the pathway and gene prioritization results were summarized as a heatmap (R. v2.3.3, *pheatmap* v1.0.8 package[27]). The gene-specific Z-score describes the likelihood that a given gene is part of the corresponding GO term, KEGG pathway, REACTOME pathway, Mouse Phenotype, or protein-protein interaction network.

Also, we performed Multi-marker Analysis of GenoMic Annotation (MAGMA)[28]. MAGMA performs gene and gene-set analysis and requires the association results of all variants, therefore we chose the larger HapMap GWAS for MAGMA. We used the Functional Mapping and Annotation (FUMA)[29] tool to perform MAGMA, and applied standard settings for running MAGMA.

To prioritize the most likely trait-relevant gene for each GWAS locus, we run colocalization analysis using the "coloc" R package v3.1[30] separately for the HapMap and 1KG GWAS. We used publicly available genome-wide eQTL data from 5,311 whole blood samples[31], and from the Genome Tissue Expression (GTEx) V6p portal incorporating eQTL data from 44 post-mortem tissues[32]. “Coloc” uses approximate Bayes factors to estimate the posterior probability that GWAS and eQTL effects share a single causal variant. All significant cis-eGenes or cis-eProbes ($q<0.05$ in GTEx; lowest cis-eQTL FDR<0.05 in Westra et al.[31]) were extracted ±1Mb from the lead SNP of each locus. The HapMap SNP positions were converted to hg19 with the liftOver command from the rtracklayer v1.38.3 package. We used the SNPs present in both the GWAS and eQTL datasets. For the HapMap GWAS, the 1KG GWAS and the GTEx eQTL datasets, we performed the test using association beta, standard error of beta, and minor allele frequency (MAF). For the data from Westra et al.[31], we used association P-value, MAF, and sample size, and included only the subset of cis-eQTLs which are publicly available (up to significance FDR<0.5). We used default priors supplied by the coloc package (P1=1e-4, P2=1e-4, P12=1e-5; prior probabilities for association in GWAS, eQTL, and both datasets). Full MAF data were not available for the eQTL datasets, therefore we used the GIANT 1KG p1v3 EUR reference panel instead. We visualized the results as a heatmap using the *pheatmap* v1.0.8 R package[27].

Mendelian randomization analyses

To assess the effect of CRP on complex disorders, we performed a two-sample Mendelian randomization (MR) study on nine clinical outcomes (Alzheimer’s disease (AD), bipolar disorder (BD), coronary artery disease (CAD), Crohn’s disease (CD), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), schizophrenia, and diastolic (DBP) and systolic blood pressure (SBP)) to which CRP showed a potentially causal association at a $P<0.1$ in a previous MR study[13]. We used the effect estimates of the 48 lead SNPs found to be associated with CRP in the HapMap GWAS, and the effect estimates of the four SNPs that were additionally found to be associated with CRP in the 1KG GWAS in a multiple instrument approach for the MR analyses (n=52 SNPs). Additionally, we separately studied the effect of rs2794520 at the *CRP* locus to minimize the probability of horizontal pleiotropy that may be introduced in a multiple instrument approach. We tested the statistical significance of the association between the instrument and CRP using the formula:

$$F = \frac{R^2(n-1-k)}{(1-R^2)\times k}$$

R^2 is the variance explained of CRP by the genetic instrument (0.014 for the rs2794520 SNP and 0.065 for the 52-SNP score), n is the number of individuals included in the CRP GWAS, and k the number of variants included in the genetic score. The *F* statistic for the 52-SNP score was 273, and for the rs2794520 SNP 2,902, indicating that both instruments were strong.

For the clinical outcomes, we used summary statistics from the most recent meta-analysis of GWA studies. For diastolic and systolic blood pressure, we used data from the UK

Biobank. The details of the outcome studies are summarized in Table S12. We implemented four different methods of MR analyses: Inverse-variance weighted method (IVW), MR-Egger, Weighted median (WM), and Penalized weighted median (PWM). We used the "*TwoSampleMR"* package in R for the MR analyses[33]. Further, we applied the Bonferroni method to correct for multiple testing (0.05/9 phenotypes = 5.6×10^{-3}). When the Q-statistic of the IVW analyses provided evidence for heterogeneity, the weighted median estimates were used for significance. The MR methods are described briefly below.

Inverse-variance weighted (IVW): The causal estimate is obtained by regressing the SNP associations with the outcome on the SNP associations with the risk factor, with the intercept set to zero and weights being the inverse-variances of the SNP associations with the outcome. With a single genetic variant, the estimate is the ratio of coefficients betaY/betaX and the standard error is the first term of the delta method approximation betaYse/betaX. When all CRP-SNPs are valid IVs, the IVW estimates converge to the true causal effect. When one or more invalids IVs are present, (ie. one SNP has effect on outcome through a different pathway than CRP), the IVW estimate deviates from the true causal effect.

MR-Egger: We used MR-Egger to account for potential unbalanced pleiotropy in the multiple variant instrument[34]. When unbalanced pleiotropy is present, an alternative effect (positive or negative) is present between the SNP and the outcome that may bias the estimate of the causal association. The MR-Egger method is similar to the IVW analysis, but does not force the intercept to pass through the origin. The slope of the MR-Egger regression provides the estimate of the causal association between CRP and the clinical outcome. An MR-Egger intercept that is significantly different from zero suggests directional pleiotropic effects that may bias uncorrected estimates of the causal effect. MR-Egger regression depends on the InSIDE (Instrument Strength Independent of Direct Effect) assumption, that states that the strengths of the effect of the SNP on the outcome is uncorrelated with the direct pleiotropic effect of the SNP on the outcome.

Weighted median (WM) and penalized Weighted Median (PWM): We applied the median based method to provide robust estimates of causal association even in the presence of horizontal pleiotropy when up to 50% of the information contributed by the genetic variants is invalid[35]. In PWM analysis the effect of each variants is weighted by a factor that corresponds to the Q statistics (heterogeneity test) of the SNP; this means that most variants will not be affected by this correction, but the causal effect of the outlying variants, which are most likely to be invalid IVs, will be down-weighted.

We displayed the individual SNP causal effect estimates and corresponding 95% confidence intervals in a forest plot. To assess whether one of the variants used in the genetic score had disproportionate effects, we performed "leave-one-out" analyses where one SNP at a time is removed from the score. We depicted the relationship between the SNP effect on

CRP and the SNP effect on the clinical outcomes in a scatter plot, and plotted the individual SNP effect against the inverse of their standard error in a funnel plot. When unbalanced pleiotropy is absent, the causal effect estimates of the individual should center around the meta-analysis estimate in the funnel plot.

We used the proportion of variance in CRP explained by the genetic instruments (0.014 for the rs2794520 SNP and 0.065 for the 52-SNP score) to perform power calculations for each outcome using the online tool mRnd[36]. We calculated the power to detect a relative 5%, 10%, 15%, and 20% difference in outcome risk. For example, a 10% difference refers to an OR of at least 0.90 or 1.10 in outcome risk (Table S13).

Results

HapMap GWAS meta-analysis for CRP levels

The HapMap meta-analysis identified 3,977 genome-wide significant variants at P-value$<5\times10^{-8}$ (QQ-plot Supplementary Fig. 1; Manhattan plot Supplementary Fig. 2), which mapped to 48 distinct loci (Table 1). Of the previously reported 18 variants for CRP, 16 remained associated. Compared to the previous GWAS, the rs6901250 variant at the *GPRC6A* locus (P-value=0.09) and the rs4705952 variants at the *IRF1* locus (P-value=2.7×10^{-3}) were not significant. The beta estimates for natural log-transformed CRP for each of the associated loci ranged from 0.020 to 0.229. The strongest association was observed for rs2794520 at the *CRP* gene (β=0.182 in the natural log-transformed CRP (mg/L) per copy increment in the coded allele, P-value=4.17×10^{-523}), followed by rs4420638 at the *APOC1/E* gene (β=0.229, P-value=1.23×10^{-305}). Similarly to previous GWAS meta-analysis, the lead variant within the interleukin-6 receptor gene (*IL6R*) was rs4129267 (β=0.088, P-value=1.2×10^{-129}). Related to the interleukin-6 pathway, we identified rs1880241 upstream of the *IL6* gene (β= 0.028, P-value=8.4×10^{-14}). In addition to the previously described interleukin-1 signaling, the *IL1RN-IL1F10* locus (interleukin-1 receptor antagonist and interleukin-1 family member 10), we found rs9284725 within the interleukin-1 receptor 1 gene (*IL1R1*, β=0.02, P-value=7.3×10^{-11}, Table 1**)**. The sex-specific meta-analyses did not identify additional loci for CRP compared to the overall meta-analysis including both sexes, but at four genetic variants we found evidence for heterogeneity in effect estimates between sexes (Supplementary table 3), though the directions of associations were consistent.

In the 1KG meta-analysis, 8,002 variants were associated with CRP at P-value$<5\times10^{-8}$ (QQ-plot Supplementary Fig. 3; Manhattan plot Supplementary Fig. 4). This resulted in 40 distinct loci, of which 36 overlapped with the HapMap meta-analysis (Table 1). The lead variant at the *CRP* locus in the 1KG GWAS was rs4287174 (β=-0.185, P-value=$1.95\times^{-398}$), which is in high LD with rs2794520 (r^2=0.98), the lead variant at the *CRP* locus in the HapMap GWAS.

Table 1. Novel identified loci associated with C-reactive protein.

Variant	Chr	Position	Coded allele	Coded allele freq	Beta	SE	P-value	Closest Gene	1KG lead variant
Loci found in the HapMap GWAS									
rs469772	1	91530305	T	0.19	-0.031	0.005	5.54×10^{-12}	*ZNF644*	rs469882
rs12995480	2	629881	T	0.17	-0.031	0.005	1.24×10^{-10}	*TMEM18*	rs62105327
rs4246598	2	88438050	A	0.46	0.022	0.004	5.11×10^{-10}	*FABP1*	-
rs9284725	2	102744854	C	0.24	0.027	0.004	7.34×10^{-11}	*IL1R1*	rs1115282
rs1441169	2	214033530	G	0.53	-0.025	0.004	2.27×10^{-11}	*IKZF2*	-
rs2352975	3	49891885	C	0.30	0.025	0.004	6.43×10^{-10}	*TRAIP*	rs10049413
rs17658229	5	172191052	C	0.05	0.056	0.010	5.50×10^{-09}	*DUSP1*	rs34471628
rs9271608	6	32591588	G	0.22	0.042	0.005	2.33×10^{-17}	*HLA-DQA1*	rs2647062
rs12202641	6	116314634	T	0.39	-0.023	0.004	3.00×10^{-10}	*FRK*	-
rs1490384	6	126851160	T	0.51	-0.025	0.004	2.65×10^{-12}	*C6orf173*	rs1490384
rs9385532	6	130371227	T	0.33	-0.026	0.004	1.90×10^{-11}	*L3MBTL3*	-
rs1880241	7	22759469	G	0.48	-0.028	0.004	8.41×10^{-14}	*IL6*	rs13241897
rs2710804	7	36084529	C	0.37	0.021	0.004	1.30×10^{-08}	*KIAA1706*	-
rs2064009	8	117007850	C	0.42	-0.027	0.004	2.28×10^{-14}	*TRPS1*	rs6987444
rs2891677	8	126344208	C	0.46	-0.020	0.004	1.59×10^{-08}	*NSMCE2*	rs10956251
rs643434	9	136142355	A	0.37	0.023	0.004	1.02×10^{-09}	*ABO*	9:136146061
rs1051338	10	91007360	G	0.31	0.024	0.004	2.27×10^{-09}	*LIPA*	-
rs10832027	11	13357183	G	0.33	-0.026	0.004	4.43×10^{-12}	*ARNTL*	rs10832027
rs10838687	11	47312892	G	0.22	-0.031	0.004	9.12×10^{-13}	*MADD*	rs7125468
rs1582763	11	60021948	A	0.37	-0.022	0.004	2.37×10^{-09}	*MS4A4A*	rs1582763
rs7121935	11	72496148	A	0.38	-0.022	0.004	5.28×10^{-09}	*STARD10*	-
rs11108056	12	95855385	G	0.42	-0.028	0.004	5.42×10^{-14}	*METAP2*	rs12813389
rs2239222	14	73011885	G	0.36	0.035	0.004	9.87×10^{-20}	*RGS6*	rs2239222

rs4774590	15	51745277	A	0.35	-0.022	0.004	2.71×10^{-08}	*DMXL2*	rs1189402
rs1558902	16	53803574	A	0.41	0.034	0.004	5.20×10^{-20}	*FTO*	rs55872725
rs178810	17	16097430	T	0.56	0.020	0.004	2.95×10^{-08}	*NCOR1*	-
rs10512597	17	72699833	T	0.18	-0.037	0.005	4.44×10^{-14}	*CD300LF,RAB37*	rs2384955
rs4092465	18	55080437	A	0.35	-0.027	0.004	3.11×10^{-10}	*ONECUT2*	-
rs12960928	18	57897803	C	0.27	0.024	0.004	1.91×10^{-09}	*MC4R*	-
rs2315008	20	62343956	T	0.31	-0.023	0.004	5.36×10^{-10}	*ZGPAT*	-
rs2836878	21	40465534	G	0.27	0.043	0.004	7.71×10^{-26}	*DSCR2*	rs4817984
rs6001193	22	39074737	G	0.35	-0.028	0.004	6.53×10^{-14}	*TOMM22*	rs4821816
Additional loci found in the 1KG GWAS									
rs75460349	1	27180088	A	0.97	0.086	0.014	4.50×10^{-10}	*ZDHHC18*	
rs1514895	3	170705693	A	0.71	-0.027	0.004	2.70×10^{-09}	*EIF5A2*	
rs112635299	14	94838142	T	0.02	-0.107	0.017	2.10×10^{-10}	*SERPINA1/2*	
rs1189402	15	53728154	A	0.62	0.025	0.004	3.90×10^{-09}	*ONECUT1*	
Additional loci found in the BMI adjusted 1KG GWAS									
3:47431869	3	47431869	D	0.59	0.024	0.004	1.10×10^{-08}	*PTPN23*	
rs687339	3	135932359	T	0.78	-0.030	0.005	2.80×10^{-10}	*MSL2*	
rs7795281	7	74122854	A	0.76	0.028	0.005	3.10×10^{-08}	*GTF2I*	
rs1736060	8	11664738	T	0.60	0.029	0.004	2.60×10^{-13}	*FDFT1*	
17:58001690	17	58001690	D	0.44	-0.026	0.004	9.50×10^{-10}	*RPS6KB1*	
rs9611441	22	41339367	C	0.49	-0.022	0.004	1.40×10^{-08}	*XPNPEP3*	

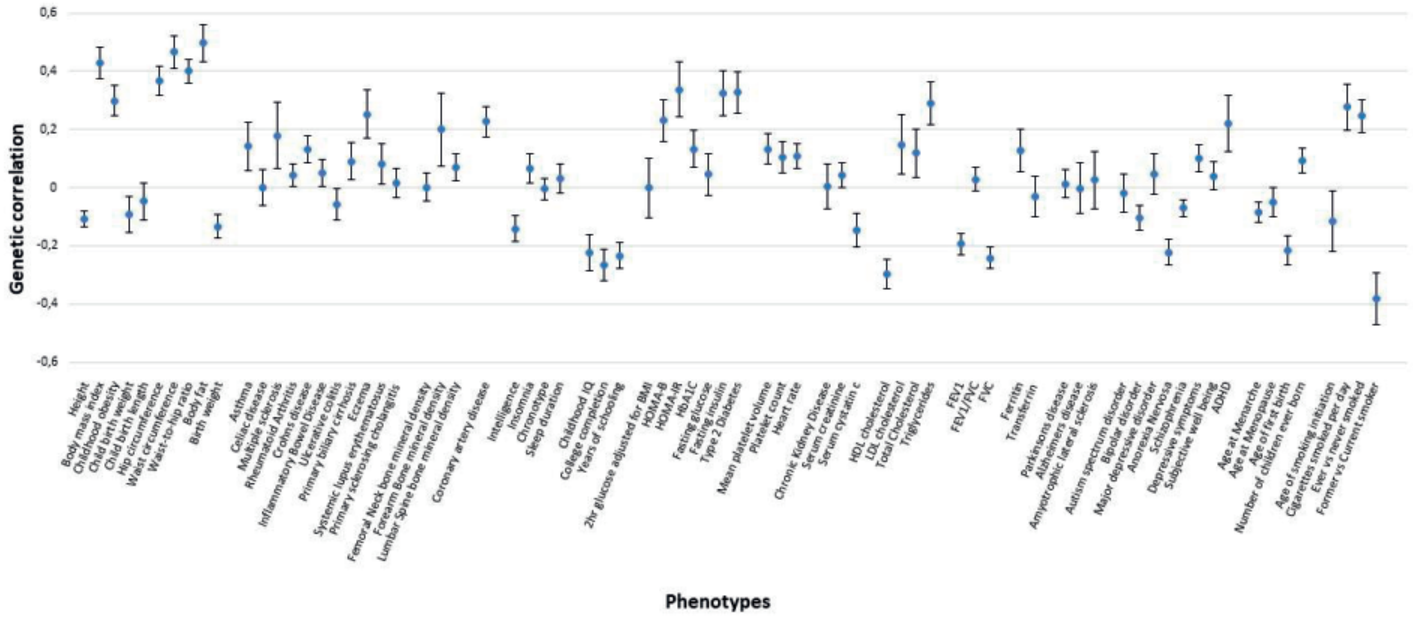
β coefficient represents 1-unit change in the natural log-transformed CRP (mg/L) per copy increment in the coded allele. Position is according to Hg19. When a variant is located within a gene, that gene is reported in the closest gene column, otherwise the closest gene. The HapMap variants are presented, except for the 1KG additional findings. For the HapMap loci, the lead variant from the 1KG GWAS is presented when the locus was also found in the 1KG GWAS.1KG GWAS meta-analysis for CRP levels.

Among eight of the overlapping loci, the lead variant was at the same position in both GWAS (rs1260326, rs1490384, rs10832027, rs1582763, rs7310409, rs2239222, rs340005, and rs1800961). Compared with HapMap, the four additional variants identified in 1KG were rs75460349 (near *ZDHHC18*), rs1514895 (near *EIF5A2*), rs112635299 (near *SERPINA1/2*), and rs1189402 (near *ONECUT1*). The variants rs1514895 and rs1189402 were available in the HapMap GWAS, but were not associated at the genome-wide threshold (respectively P-value=1.2×10^{-7} and P-value=8.1×10^{-3}). The two variants rs75460349 and rs112635299 were not available in the HapMap GWAS, nor were variants in high LD ($r^2<0.8$). The rs75460349 is a low frequency variant with a coded allele frequency of 0.97 (β=0.086, P-value=4.5×10^{-10}). Also rs112635299 near the *SERPINA1/2* gene is a low frequency variant with a MAF of 0.02 (β=0.107, P-value=2.1×10^{-10}). Adjustment for BMI in the 1KG GWAS (n=147,827) revealed six additional loci that were not associated with CRP in the HapMap and 1KG primary analyses (Table 1; Supplementary table 4, QQ-plot Supplementary Fig. 5; Manhattan plot Supplementary Fig. 6). The associations at three lead variants were much reduced after adjustment for BMI (rs1558902 (*FTO*), P-value$_{adjusted}$=0.40; rs12995480 (*TMEM18*), P-value$_{adjusted}$=0.02; rs64343 (*ABO*), P-value$_{adjusted}$ =1.0×10^{-7}). Both the *FTO* and *TMEM18* gene are well-known obesity genes. Except for the *FTO*, *TMEM18*, and *ABO* loci, all distinct loci identified in the primary 1KG analysis were also associated with CRP in the BMI adjusted 1KG analysis. No genome-wide significant association was observed on the X-chromosome in the 1KG GWAS including 102,086 individuals.

LD score regression

The HapMap GWAS LD Score regression intercept was 1.03 (standard error: 0.013), and the 1KG intercept was 1.02 (standard error 0.011). This suggests that a small proportion of the inflation is attributable to confounding bias (~12% for the HapMap GWAS and ~13% for the 1KG GWAS). Hence, the vast majority of inflation is due to the polygenic architecture of circulating CRP levels. As depicted in Figure 1, CRP showed strong positive genetic correlations with anthropometric traits (e.g. BMI: R_g=0.43, P-value=5.4×10^{-15}), glycemic phenotypes (e.g. type 2 diabetes R_g=0.33, P-value=3.1×10^{-6}), lipid phenotypes (e.g. triglycerides R_g=0.29, P-value=7.9×10^{-5}), and coronary artery disease (R_g=0.23, P-value=2.4×10^{-5}) (Supplementary table 5). By comparison, CRP showed inverse genetic correlations with educational attainment (e.g. college completion R_g=-0.27, P-value=9.2×10^{-7}), lung function (e.g. forced vital capacity R_g=-0.24, P-value=4.6×10^{-12}), and HDL-cholesterol (R_g=-0.30, P-value=4.8×10^{-9}).

Figure 1. Genome-wide genetic correlation between serum CRP levels and different phenotypes and clinical diseases.



The genetic correlation and its standard error are estimated with linkage disequilibrium score regression analysis. ADHD, attention deficit and hyperactivity disorder; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; HOMA-B, homeostatic model assessment β-cell function; HOMA-IR, homeostatic model assessment insulin resistance; HbA1C, Hemoglobin A1c.

Additional signals at distinct loci

Approximate conditional analyses in the 1KG GWAS revealed additional signals at nine loci (Supplementary table 6). Five loci showed one secondary signal (*IL6R, NLRP3, HNF1A, CD300LF,* and *APOE/APOC1*), the *PPP1R3B* locus had two additional signals, the *LEPR* locus had three additional signals, and the *SALL1* locus had four additional signals, whereas the *CRP* locus showed a total of 13 distinct associated variants. Interestingly, the rs149520992 rare variant (MAF=0.01) mapping to the *CRP* locus showed an association at P-value$_{\text{conditional}}$=3.7×10^{-15} with β=-0.272 for the T-allele. The GCTA effect estimates for the ten distinct variants in the vicinity of the *CRP* gene identified in the 1KG conditional analysis are in high correlation with the effect estimates of these variants obtained from the RS-I and WGHS individual level data (r_{RS}=0.97, and r_{WGHS}=0.84), confirming the reliability of the GCTA estimates.

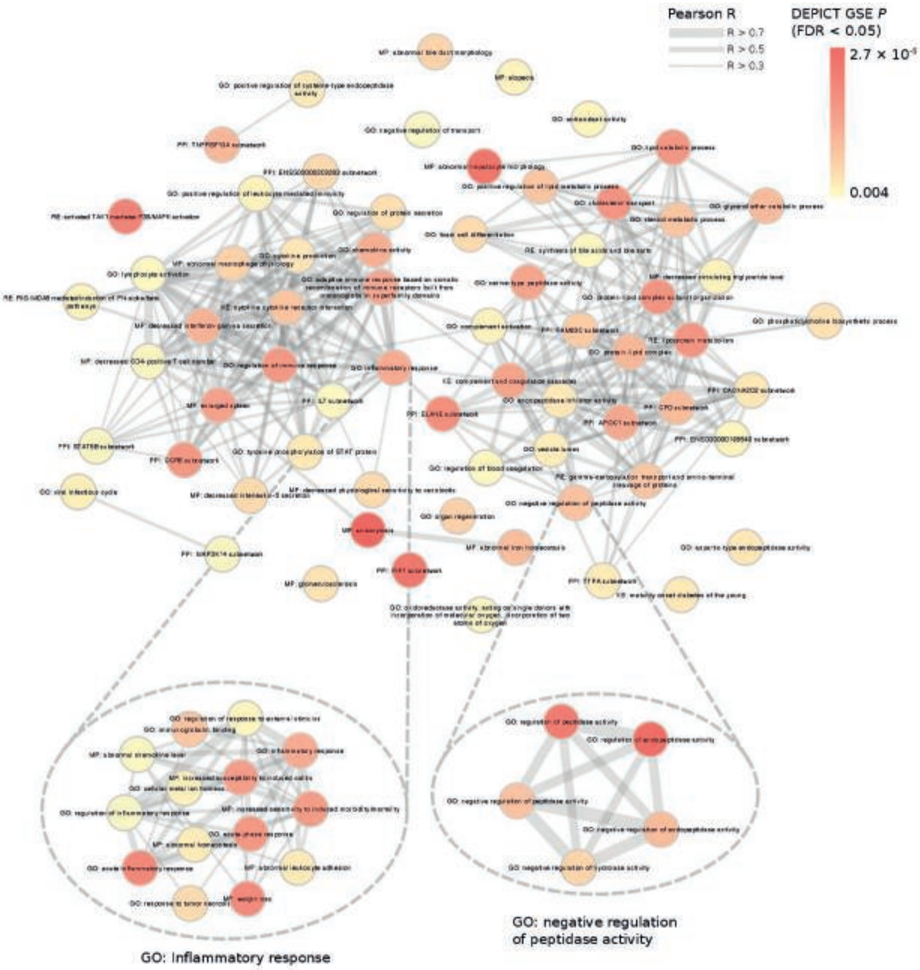
Variance explained of CRP

The lead variant at the *CRP* locus in both the HapMap (rs2794520) and 1KG (rs4287174) GWAS explained 1.4% of the variance in natural log-transformed CRP levels. The distinct variants at the *CRP* locus derived from the joint conditional analysis in the 1KG GWAS explained 4.3% of the variance. The lead variants at all distinct loci together explained 6.2% of the CRP variance in the HapMap GWAS, and 6.5% in the 1KG GWAS. When we added the distinct variants at associated loci derived from the conditional analysis, the variance explained by all associated loci was 11.0% in the 1KG GWAS.

Functional annotation

We applied DEPICT and MAGMA analyses for functional annotation and biological interpretation of the findings. The DEPICT analysis included 9,497 genome-wide significant variants, covering 283 genes, and prioritized 55 candidate genes across 29 regions (FDR<0.05, Table S8). The prioritized genes included *IL6R* mapping to the 1q21.3 locus (represented by rs4129267) and *APCS* to the 1q32.2 locus. Investigating 10,968 reconstituted gene sets for enrichment, DEPICT highlighted 583 (5.3%) gene sets to be significantly enriched among CRP-associated loci at FDR<0.05 (Table S9). Using further clustering, we identified 66 groups of gene sets that substantially correlated and clustered in two sets, one mainly comprised of immune pathways, and the other enriched for metabolic pathways (Figure 2). In Figure 3, we present the prioritized genes and the most significant gene sets. We found synovial fluid, liver tissue, and monocytes to be enriched for the expression of the prioritized genes (FDR<0.05). The MAGMA analysis was applied on the HapMap GWAS, identifying five significantly enriched gene sets (Bonferroni-corrected P<0.05, Table S10). Results included consequences of gene EGF induction, positive regulation of gene expression, and IL-6 signaling pathway, in line with the most strongly

Figure 2. Results of the DEPICT functional annotation analysis.


Pearson R
R > 0.7
R > 0.5
R > 0.3
DEPICT GSE P (FDR < 0.05)
2.7×10^{-5}
0.004
MP: abnormal bile duct morphology
MP: alopecia
GO: positive regulation of cysteine-type endopeptidase activity
GO: antioxidant activity
GO: negative regulation of transport
PPI: TNFRSF13A subnetwork
GO: lipid catabolic process
MP: abnormal hepatocyte morphology
PPI: ENSG00000203093 subnetwork
GO: positive regulation of lipid metabolic process
GO: positive regulation of leukocyte mediated immunity
GO: cholesterol transport
GO: glycerol ether metabolic process
RE: activated TAK1 mediates P38/MAPK activation
GO: regulation of protein secretion
GO: steroid metabolic process
GO: foam cell differentiation
GO: cytokine production
GO: chemokine activity
RE: synthesis of bile acids and bile salts
MP: abnormal macrophage physiology
GO: lymphocyte activation
GO: adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
GO: serine-type peptidase activity
MP: decreased circulating triglyceride level
RE: RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways
GO: protein-lipid complex subunit organization
RE: cytokine-cytokine receptor interaction
MP: decreased interferon-gamma secretion
GO: complement activation
PPI: SAMD9L subnetwork
GO: phosphatidylcholine biosynthetic process
RE: lipoprotein metabolism
MP: decreased CD4-positive T cell number
GO: regulation of immune response
GO: protein-lipid complex
GO: inflammatory response
KE: complement and coagulation cascades
PPI: CACNA2D2 subnetwork
GO: endopeptidase inhibitor activity
MP: enlarged spleen
PPI: IL7 subnetwork
PPI: ELANE subnetwork
PPI: CPD subnetwork
PPI: APOC1 subnetwork
PPI: ENSG00000109648 subnetwork
PPI: STAT5B subnetwork
GO: vesicle lumen
PPI: CCR6 subnetwork
GO: tyrosine phosphorylation of STAT protein
GO: regulation of blood coagulation
RE: gamma-carboxylation, transport and amino-terminal cleavage of proteins
GO: viral infectious cycle
MP: decreased interleukin-5 secretion
MP: decreased physiological sensitivity to xenobiotic
GO: negative regulation of peptidase activity
GO: organ regeneration
MP: anisocytosis
MP: abnormal iron homeostasis
GO: aspartic-type endopeptidase activity
PPI: MAP3K14 subnetwork
PPI: FHT subnetwork
PPI: TTFR subnetwork
KE: maturity onset diabetes of the young
MP: glomerulosclerosis
GO: oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen
GO: regulation of response to external stimulus
GO: immunoglobulin binding
MP: abnormal chemokine level
GO: inflammatory response
MP: increased susceptibility to induced colitis
GO: cellular metal ion homeostasis
GO: regulation of inflammatory response
MP: increased sensitivity to induced morbidity/mortality
GO: acute-phase response
MP: abnormal homeostasis
GO: acute inflammatory response
MP: abnormal leukocyte adhesion
GO: response to tumor necrosis
MP: weight loss
GO: Inflammatory response
GO: regulation of peptidase activity
GO: regulation of endopeptidase activity
GO: negative regulation of peptidase activity
GO: negative regulation of endopeptidase activity
GO: negative regulation of hydrolase activity
GO: negative regulation of peptidase activity

Each node represents exemplar gene set from Affinity Propagation clustering and links represent corresponding Pearson correlation coefficients between individual enriched gene sets (only the links with r>0.3 are shown). As an example, outlined are the individual gene sets inside two clusters ("Inflammatory response" and "negative regulation of peptidase activity").

prioritized gene from DEPICT gene prioritization. MAGMA analysis prioritized liver as a sole enriched tissue (P-value=0.048).

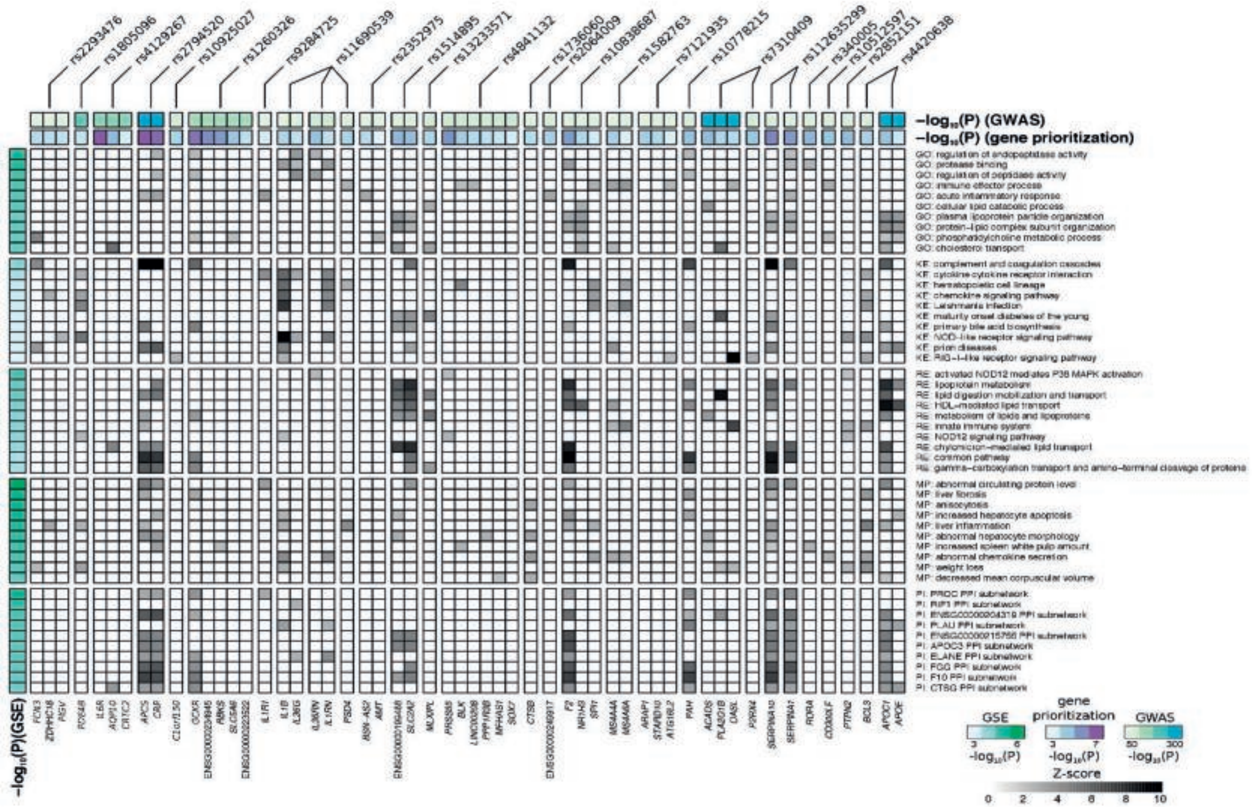
To prioritize the most likely trait-relevant gene for each GWAS locus, we interrogated the GWAS data with *cis*-eQTL data identified from 44 post-mortem tissues and a large whole blood eQTL meta-analysis using colocalization analysis (Table S11). Figure S7 presents the GWAS loci that colocalize with *cis*-eQTLs with the corresponding tissue, the colocalizing gene, and the posterior probability of one shared underlying variant driving both associations. Out of the 58 lead gSNPs, 25 SNPs (43%) showed evidence of colocalization with one or more local eQTL effects (posterior probability>0.9). For example, the rs2293476 locus colocalizes with several *cis*-eQTL effects for *PABC4*, and pseudogenes *OXCT2P1*, *RP11−69E11.4*, and *RP11−69E11.8*. The rs10925027 locus shows colocalization with cis-eQTL effect for *NLRP3*, exclusively in the highly powered blood meta-analysis. Out of 25 loci, for nine loci there was only one colocalizing gene. Altogether, gSNP-associated cis-eQTL effects were present in up to 14 different tissues, with whole blood, esophagus mucosa, skin, and tibial nerve being the most frequent.

Mendelian randomization analyses

We observed a protective effect of genetically determined variance in CRP with schizophrenia with an IVW odds ratio (OR) of the 52-SNP score of 0.89 (95%CI: 0.81-0.97, P-value=6.6×10^{-3}) (Tables S14-S15, Figure S8-S11). The MR-Egger intercept was compatible with no unbalanced pleiotropy (P-value=0.48). The estimate of the rs2794520 variant was comparable to the 52-SNP score estimate (OR 0.89, 95% CI 0.84-0.94, P-value=0.046). The WM and PWM estimates were comparable to the IVW estimate (OR_{WM} 0.89, P-$value_{WM}$=5.1×10^{-3}; OR_{PWM}=0.89, P-$value_{PWM}$=4.4×10^{-3}). The “leave-one-out” analysis provided evidence that no single variant was driving the IVW point estimate (Figure S10). The causal OR between the rs2794520 variant and BD was 1.33 (95% CI 1.03-1.73, P-value=0.032). For the 52-SNP score, the IVW OR was 1.16 (95% CI 1.00-1.35, P-value=0.054). The MR-Egger intercept was compatible with unbalanced pleiotropy (P-value=0.049). The MR-Egger estimate OR of the 52-SNP score was comparable to the rs2794520 estimate (OR=1.36, 95%CI 1.1-1.69, P-value=6.7×10^{-3}), as were the WM and PWM estimates (OR_{WM}=1.33, P-$value_{WM}$=3.4×10^{-3}; OR_{PWM}=1.32, P-$value_{PWM}$=4.3×10^{-3}).

We observed evidence against a causal association between either *CRP* rs2794520 (OR=1.01, 95%CI 0.91-1.12, P-value=0.88), or the 52-SNP instrument (OR=0.96, 95%CI 0.84-1.09, P-value=0.51) and CAD. An Egger intercept of 0.014 suggested presence of unbalanced pleiotropy (P-value=5.8×10^{-3}), with an MR-Egger causal estimate of OR 0.79 (95%CI 0.67-0.94, P-value=0.012). However, the WM and PWM showed no association between CRP and CAD. For AD, there was evidence against an association with rs2794520 (P-value=0.592), though the IVW OR showed a protective effect (OR=0.51, 95%CI 0.30-0.88, P-value=0.015).

Figure 3. Heatmap representing the results of DEPICT functional annotation analysis.



Each row represents enriched (FDR < 0.05) gene sets and each column represents prioritized (FDR<0.05) genes. Colors on the heatmap represent each gene´s contribution to gene set enrichment (depicted as Z-score, only top 10 highest Z-scores per gene set are visualized). Sidebars represent p-values for GWAS, gene set enrichment (GSE), and gene prioritization (nominal P-value on log_{10} scale). Top 10 gene sets per annotation category are visualized. GO, Gene Ontology; KE, Kyoto Encyclopedia of Gene and Genomes; RE, REACTOME pathways; MP, Mouse Phenotypes; PI, protein-protein interactions.

The Egger intercept of 0.046 suggested unbalanced pleiotropy (P-value=0.042), and the MR-Egger OR was 0.27 (95%CI 0.12-0.60). However, the association was null for the WM and PWM analyses (OR_{WM}=1.04, P-$value_{WM}$=0.61; OR_{PWM}=1.05, P-$value_{PWM}$=0.53). We observed evidence against an effect for CD, DBP, IBD, RA, and SBP for the rs2794520 variant and the IVW, MR-Egger, WM, and PWM analyses.

Discussion

Using genomic data from >200,000 individuals, we have identified 42 novel distinct signals for circulating CRP levels, and confirmed 16 previously identified CRP loci totaling 58 genetic loci associated to CRP levels. BMI-adjusted GWAS suggested that the vast majority of genetic risk variants affect CRP levels independent of its main determinant (BMI). The genome-wide *in silico* functional annotation analysis highlights 55 genes which are likely to explain the association of 29 signals to CRP levels. The data identified gene sets involved in the biology of immune system and liver as main regulators of serum CRP levels. Mendelian randomization analyses supported causal associations of genetically increased CRP with a protective effect on schizophrenia, and increased risk of bipolar disorder.

Obesity is one of the main determinants of chronic low-grade inflammation in the general population[37,38]. Adjustment for BMI in the CRP GWAS abolished the association at only three lead variants, suggesting that the genetic regulation of chronic low-grade inflammation is largely independent from BMI. Notably, BMI adjustment resulted in the identification of six variants that were not associated with CRP in the BMI-unadjusted GWAS. This supports the notion that adjustment for covariates that explain phenotypic variance may improve the statistical power in linear model analyses of quantitative traits[39]. Although adjustment for heritable correlated traits in GWAS may bias effect estimates (collider bias)[40], there is consistent evidence in the literature that BMI has a causal direct effect on CRP levels[41], and therefore, collider bias in CRP GWAS adjusted for BMI is less likely.

The sex-stratified analyses revealed significant heterogeneity in effect estimates between men and women at only four lead variants, which represent less than 10% of all CRP loci. Even among these four loci the effect directions were similar, thus the heterogeneity was limited to effect sizes. The data suggest that the difference between men and women in CRP levels is less likely to be explained by genetic factors.

The top variant at the *CRP* locus in both the HapMap and 1KG GWAS explained 1.4% of the variance in circulating CRP levels. The approximate conditional analysis resulted in 13 variants jointly associated within the *CRP* locus in the 1KG GWAS. With respect to locus definition, we used a conservative distance criterion compared to other GWA studies that often use ±500kb surrounding the GWAS peak[42]. Here, we used the criterion that the minimum distance between the boundaries of loci is 500kb. In order to identify further variants associated with CRP levels, we performed approximate conditional analyses resulting in multiple putative additional variants, also inside and near genes that were not identified in the primary GWAS. As an example, the *CRP* locus spanned >2MB according to our criterion. Approximate conditional analysis revealed that two variants, namely rs3027012 near *DARC* and rs56288844 near *FCER1A*, both downstream of the *CRP* gene, were associated with CRP levels. Furthermore, upstream of *CRP*, we identified a variant near *FCGR2A* (Immunoglobulin G Fc Receptor II). These results show that for a given lead variant, potentially multiple causal loci, here *DARC*, *FCER1A*, and *FCGRA2,* alongside *CRP* contribute to chronic low-grade inflammation and variation in circulating CRP levels.

DEPICT analysis provided further evidence that the genes annotated to the associated CRP variants mainly cluster in the immune and liver biological systems. Notably, the gene set "inflammatory response", which captures both immune response and liver metabolism, was the main connector network between the two networks. This is in line with the observation that CRP is mainly produced by liver cells in response to inflammatory cytokines during acute and chronic inflammation[43]. Interestingly, the analysis highlighted iron homeostasis as an enriched gene set. In agreement, the conditional analysis highlighted a distinct genetic association at the hemochromatosis gene *HFE,* a transmembrane protein of the major histocompatibility complex (MHC) class I family. Previous studies show that iron metabolism plays a pivotal role in inflammation[44,45]. However, genetic pleiotropy may highlight co-regulated networks in pathway analysis that are not causal to inflammation per se. It is also important to note that the results of DEPICT analyses apply to reconstituted gene sets which may sometimes have slightly different overlaying biological theme than the original gene set annotation.

The MR analyses validate previous evidence that genetically-elevated CRP is protective for the risk of schizophrenia[13,46], although observational data suggest a positive association between CRP and risk of schizophrenia[47]. For bipolar disorder we observed a positive causal effect, which is in line with previous MR and observational studies[13,48]. Although the causal underlying mechanisms remain to be elucidated, a hypothesis for the schizophrenia observation might be the immune response to infections early in life. Levels of acute-phase response proteins in dry blood spots collected at birth are lower for patients with non-affective psychosis, which includes schizophrenia, compared to controls, suggesting a weaker immune response at birth[49]. Also, neonates that have been exposed to a maternal infection and have low levels of acute-phase response proteins, have a higher risk of schizophrenia[50]. Altogether, the evidence suggests that a deficient immune response may contribute to chronic infection in children and the development of schizophrenia. For AD and CHD, the Egger intercept showed evidence of unbalanced pleiotropy and the Egger estimate showed a protective effect of CRP on the risk of AD and CHD. However, for both outcomes, the effects of the WM and PWM analyses, as well as analyses using the single rs2794520 variant (which is least likely to be affected by pleiotropy) were null. The MR-Egger estimate relies on the InSIDE assumption which states that the strength of the association between the genetic variants and CRP is independent from the strength of the direct pleiotropic effects of the genetic variants on the outcome. This assumption may be violated when the genetic variants are associated with a confounder of the CRP-outcome association. Such a scenario may occur when the genetic variants are associated with an exposure that is causally upstream of the exposure under study. In the context of the association of CRP with AD and CHD, this could be lipids or glycemic phenotypes. Several genetic variants used in the CRP_{GWAS} instrument are associated with metabolic phenotypes that may affect CRP levels. In agreement, the WM and PWM, in which the InSIDE assumption is relaxed, and the single variant analysis showed no association. Furthermore, the observation that CRP is not causally related to CAD in the MR analyses is in comparison to previous published studies[51]. Power calculation showed that we had 100% power to detect a 10% difference in CAD risk, thus the probability of a false negative finding is small. Also, CRP is associated with future CAD in observational studies, and randomized trials have shown a beneficial effect of lowering inflammation using statins[52] and canakinumab[53] on CAD risk, but this effect is unlikely to be attributable to CRP.

The strengths of our study are the use of the largest sample size for CRP to date and the use of both HapMap and 1KG imputed data. Furthermore, we conducted sex-specific and BMI-adjusted analyses to study the effect of sex and body mass on the associations between genetic variants and CRP. To maximize power and to efficiently use the data, we meta-analyzed all available samples in a discovery setting without replication. The consistent association of the variants in >50 studies at a strict Bonferroni corrected threshold provide confidence that our findings represent true associations. We used both HapMap and 1KG imputed data to identify novel genetic variants for circulating CRP levels. At the start of the project, more studies had HapMap imputed data available. Hence, the sample size and thus power in the HapMap GWAS was higher compared to the 1KG. Also, HapMap may identify variants that are not identified in 1KG GWAS[54]. Nevertheless, 1KG offers better coverage of uncommon variants and includes INDELs, which are not included in the HapMap reference panel. Including both reference panels, we used all available samples and maximized the possibility to identify novel genetic variants for CRP, both common and uncommon.

However, we note limitations to our study. GWAS merely identify loci associated with complex phenotypes and the identification of causal genes remains challenging. We only included individuals of European ancestry; the generalizability of our findings to other races/ethnicities is uncertain. In addition, although our analyses provided support for causal associations, we acknowledge that we may not have identified the causal variants and we may not have eliminated residual confounding.

In conclusion, we performed the largest GWAS meta-analysis to identify novel loci associated with circulating CRP levels, a sensitive marker of chronic low-grade inflammation, and found support for a causal role of CRP with a decreased risk of schizophrenia and higher risk of bipolar disorder. As inflammation is implicated in the pathogenesis of multiple complex diseases, the new insights into the biology of inflammation obtained in the current study may contribute to future therapies and interventions.

References

1. Libby P. Inflammation in atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology* 2012; 32(9): 2045-51.

2. Pickup JC. Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes. *Diabetes care* 2004; 27(3): 813-23.

3. Akiyama H, Barger S, Barnum S, et al. Inflammation and Alzheimer’s disease. *Neurobiology of aging* 2000; 21(3): 383-421.

4. Khandaker GM, Cousins L, Deakin J, Lennox BR, Yolken R, Jones PB. Inflammation and immunity in schizophrenia: implications for pathophysiology and treatment. *The Lancet Psychiatry* 2015; 2(3): 258-70.

5. Pepys MB. The acute phase response and C-reactive protein. *Oxford textbook of medicine* 1995; 2: 1527-33.

6. Danesh J, Wheeler JG, Hirschfield GM, et al. C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *N Engl J Med* 2004; 350(14): 1387-97.

7. Dehghan A, Kardys I, de Maat MP, et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. *Diabetes* 2007; 56(3): 872-8.

8. Wium-Andersen MK, Ørsted DD, Nordestgaard BG. Elevated C-reactive protein associated with late-and very-late-onset schizophrenia in the general population: a prospective study. *Schizophrenia bulletin* 2013: sbt120.

9. Dehghan A, Dupuis J, Barbalic M, 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): 731-8.

10. De Vries PS, Sabater-Lleal M, Chasman DI, et al. Comparison of HapMap and 1000 genomes reference panels in a large-scale genome-wide association study. *PloS one* 2017; 12(1): e0167742.

11. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* 2017; 101(1): 5-22.

12. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine* 2008; 27(8): 1133-63.

13. Prins BP, Abbasi A, Wong A, et al. Investigating the causal relationship of C-reactive protein with 32 complex somatic and psychiatric outcomes: a large-scale cross-consortium Mendelian randomization study. *PLoS Med* 2016; 13(6): e1001976.

14. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation: Cardiovascular Genetics* 2009; 2(1): 73-80.

15. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC bioinformatics* 2010; 11(1): 288.

16. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; 26(17): 2190-1.

17. Randall JC, Winkler TW, Kutalik Z, et al. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet* 2013; 9(6): e1003500.

18. Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* 2015; 47(3): 291-5.

19. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 2016: btw613.

20. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics* 2015.

21. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 2011; 88(1): 76-82.

22. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* 2012; 44(4): 369-75.

23. Park J-H, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics* 2010; 42(7): 570-5.

24. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications* 2015; 6.

25. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; 491(7422): 56-65.

26. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* 2011; 27(17): 2463-4.

27. Kolde R. Pheatmap: pretty heatmaps. *R package version* 2012; 61.

28. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 11, e1004219.

29. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat commun 8, 1826.

30. Giambartolomei, C., Vukcevic, D., Schadt, E.E.,et al. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 10, e1004383.

31. Westra H-J, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics* 2013; 45(10): 1238-43.

32. Consortium G. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 2015; 348(6235): 648-60.

33. Hemani, G., Zheng, J., Elsworth, B., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. eLife 7, e34408.

34. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology* 2015; 44(2): 512-25.

35. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology* 2016; 40(4): 304-14.

36. Brion, M.-J.A., Shakhbazov, K., and Visscher, P.M. (2012). Calculating statistical power in Mendelian randomization studies. Int J Epidemiol 42, 1497-1501.

37. Visser, M., Bouter, L.M., McQuillan, G.M., Wener, M.H., and Harris, T.B. (1999). Elevated C-reactive protein levels in overweight and obese adults. JAMA 282, 2131-2135.

38. Wellen, K.E., and Hotamisligil, G.S. (2003). Obesity-induced inflammatory changes in adipose tissue. J Clin Invest 112, 1785.

39. Robinson, L.D., and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. International Statistical Review/Revue Internationale de Statistique, 227-240.

40. Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am J Hum Genet 96, 329-339.

41. Timpson, N.J., Nordestgaard, B.G., Harbord, R.M., et al. (2011). C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. Int J Obes 35, 300-308.

42. Locke, A.E., Kahali, B., Berndt, S.I., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197-206.

43. Moshage, H.J., Roelofs, H.M.J., Van Pelt, J.F., et al. (1988). The effect of interleukin-1, interleukin-6 and its interrelationship on the synthesis of serum amyloid A and C-reactive protein in primary cultures of adult human hepatocytes. Biochem Biophys Res Commun 155, 112-117.

44. Ganz, T., and Nemeth, E. (2015). Iron homeostasis in host defence and inflammation. Nat Rev Immunol 15, 500-510.

45. Alizadeh, B., Njajou, O., Hazes, J.,et al. (2007). The H63D variant in the HFE gene predisposes to arthralgia, chondrocalcinosis and osteoarthritis. Ann Rheum Dis 66, 1436-1442.

46. Hartwig, F.P., Borges, M.C., Horta, B.L., Bowden, J., and Smith, G.D. (2017). Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study. JAMA psychiatry.

47. Fernandes, B.S., Steiner, J., Bernstein, H.-G., et al. (2016). C-reactive protein is increased in schizophrenia but is not altered by antipsychotics: meta-analysis and implications. Mol Psychiatry 21, 554-565.

48. Fernandes, B.S., Steiner, J., Molendijk, et al. (2016). C-reactive protein concentrations across the mood spectrum in bipolar disorder: a systematic review and meta-analysis. Lancet Psychiatry 3, 1147-1156.

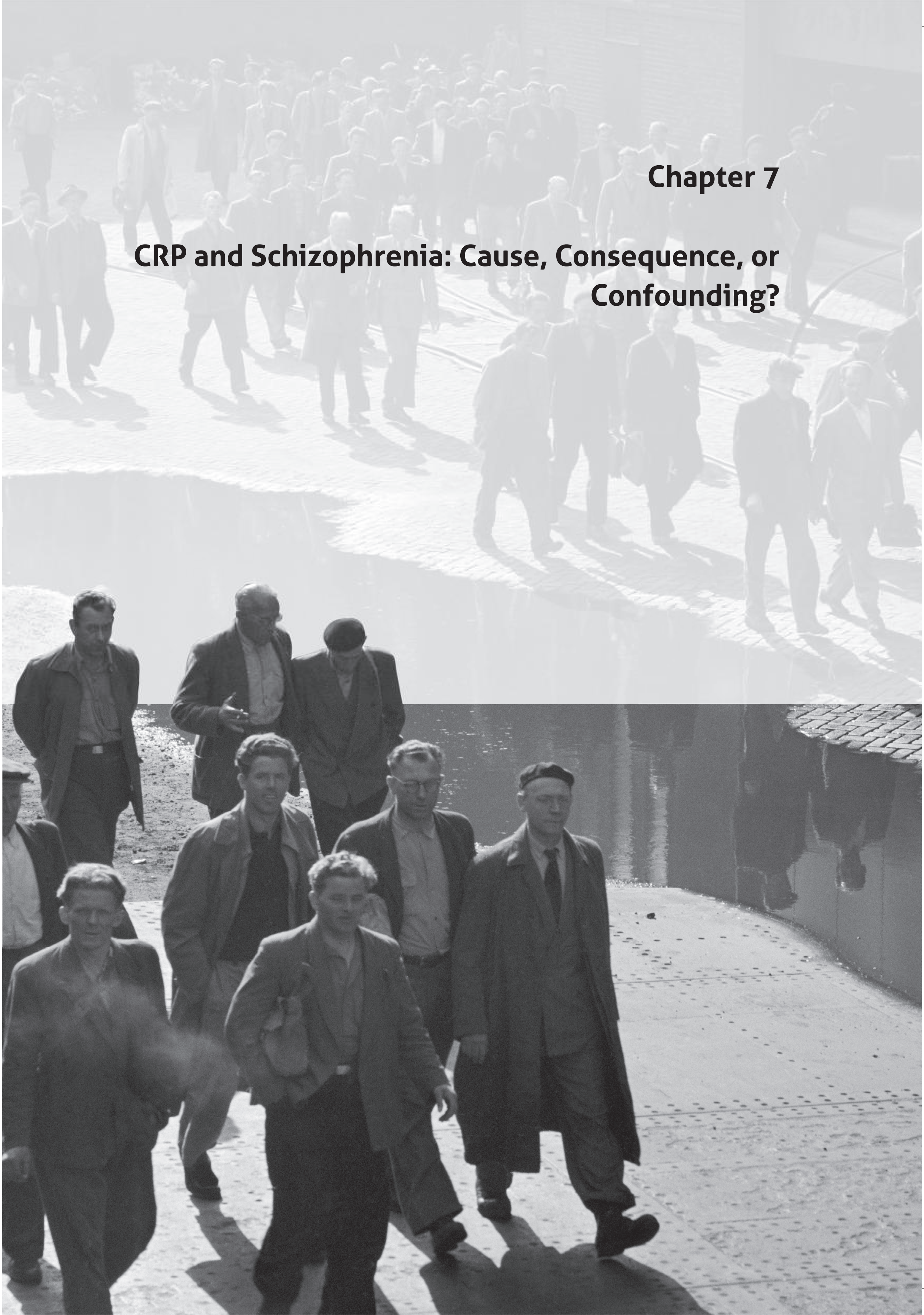
49. Gardner, R., Dalman, C., Wicks, S., Lee, B., and Karlsson, H. (2013). Neonatal levels of acute phase proteins and later risk of non-affective psychosis. Transl psychiatry 3, e228.

50. Blomström, Å., Gardner, R., Dalman, C., Yolken, R., and Karlsson, H. (2015). Influence of maternal infections on neonatal acute phase proteins and their interaction in the development of non-affective psychosis. Transl psychiatry 5, e502.

51. Elliott, P., Chambers, J.C., Zhang, et al. (2009). Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. JAMA 302, 37-48.
52. Ridker, P.M., Danielson, E., Fonseca, et al. (2008). Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. N Engl J Med 359, 2195-2207.
53. Ridker, P.M., Everett, B.M., Thuren, T., et al. (2017). Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. N Engl J Med 377, 1119-1131.
54. Wood, A.R., Perry, J.R., Tanaka, T., et al. (2013). Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. PLoS One 8, e64343.

Supplementary material

The supplementary material of this manuscript can found at the following webpage: https://www.cell.com/ajhg/fulltext/S0002-9297(18)30320-3.



Chapter 7

CRP and Schizophrenia: Cause, Consequence, or Confounding?

Schizophrenia is a debilitating psychiatric disorder affecting millions of people worldwide. Both genetic and environmental factors contribute to disease development, but the pathophysiology of schizophrenia is poorly understood and treatment mainly consists of psychosocial interventions and antipsychotic medication. Mendelian randomization (MR) analysis has the potential to identify causal factors for an outcome of interest, providing insights in the biological pathways that cause disease and could aid in detecting novel therapeutic targets. Genome-wide association studies (GWAS) have identified more than one hundred loci for schizophrenia that can be used in MR analysis for the identification of causal factors for schizophrenia.

In 2016, Prins and colleagues described for the first time an association between genetically determined CRP and schizophrenia in an MR study[1]. In contrast to prior published observational association studies in which higher CRP levels were observed in cases of schizophrenia patients compared to controls[2], Prins et al. found a protective causal effect of CRP on schizophrenia. Using similar CRP and schizophrenia GWAS data, this finding was confirmed in a subsequent MR study that incorporated robust MR sensitivity analyses[3]. Additionally, Hartwig et al. investigated the role of IL-6 which is the major upstream regulator of CRP and findings in the MR of IL-6 were in agreement with the protective effect of CRP. Recently, in a novel CRP GWAS effort, up to 52 genetic variants were included in MR analyses of CRP and schizophrenia and a similar protective effect of CRP on schizophrenia risk was found[4]. Furthermore, the MR analysis that included only the genetic variant within the *CRP* gene, which reduces the chance of horizontal pleiotropy in which the genetic variant is independently associated with multiple phenotypes, showed similar results.

The study by Bochao et al[5], published in this issue, examined the causal association between CRP and several blood metabolites with schizophrenia in an MR study applying different MR methods. The study is important and novel for several reasons. First, the authors used the most recent GWAS data sets available for both CRP and schizophrenia to perform the first bidirectional MR analysis. Second, they are the first to apply Generalized Summary data-based Mendelian Randomization (GSMR) analyses which has the advantage of more statistical power compared to MR Egger and includes the possibility to detect putative pleiotropic effects through the Heterogeneity in Dependent Instruments (HEIDI) test. Third, extensive MR sensitivity analyses were applied to exclude weak instrument bias, horizontal pleiotropy, and heterogeneity in the instrumental variables. Finally, with regards to blood metabolites, a novel MR technique developed to handle high-throughput data performing multiple multivariable MR models was applied. The authors observed in all MR analyses a protective effect of CRP on schizophrenia, and no effect of genetic liability to schizophrenia on CRP levels, confirming published MR work. In sensitivity analyses, no evidence was found for weak instrument bias or horizontal pleiotropy, and selection or survivor bias is unlikely to explain the association between CRP and schizophrenia.

Although robust MR analyses suggest a causal protective effect of CRP on schizophrenia, could the results still be confounded? MR analyses rely on several assumptions, one of them being that the instrument affects the outcome only through the risk factor. This assumption may be violated in the CRP-schizophrenia association when the CRP variants affect schizophrenia directly, not through CRP. Even the single risk variant at the *CRP* gene or a variant in high linkage disequilibrium may have an effect on schizophrenia that is not through serum CRP levels. Another assumption of MR analysis is that the genetic variants are not associated with confounders of the association between the exposure and the outcome. We may be unaware of confounding factors of the CRP-schizophrenia association, and the association of genetic variants with these confounding factors.

If we assume that CRP truly does have a causal effect on schizophrenia, what is the biological explanation? C-reactive protein is a pentameric protein first discovered by William Tillet and Thomas Francis in 1930 and named after the C-polysaccharide of the *pneumococcus* bacteria[6]. CRP has a notable role in the immune system as an activator of the classic complement cascade, among other things. Therefore, CRP is important for antimicrobial defense. Prior research has indeed shown that CRP may protect against bacterial infections[7], and infections have been hypothesized as a cause for schizophrenia[8]. Considering the role of CRP in antimicrobial defense, CRP may thus lower schizophrenia risk by reducing infection risk. However, there is no strong evidence yet to support this hypothesis. Furthermore, CRP may have a biological effect on neurocognitive function that is yet unknown. Since MR studies estimate the lifetime effect of the exposure on the outcome, Bochao et al. speculate that other CRP risk variants affect CRP levels in children and that possibly childhood infections attributable to environmental factors increase schizophrenia risk. This hypothesis does not explain the protective effect of CRP observed in the MR analyses, and there is no data to support the hypothesis that the genetic background of CRP levels in children is different from adults.

In order to get a better understanding of the association between CRP and schizophrenia, it would be of interest to examine the association between CRP and infection risk in well-powered MR studies. Genetic data on infection risk is scarce, but GWAS have been published for specific pathogens[9]. Also, an assessment of the causal association between infections with specific pathogens and schizophrenia may elucidate if, and which, pathogens may contribute to the risk of schizophrenia. Furthermore, thinking outside the field of epidemiology, possibly wet lab experiments designed to assess the effect of CRP on neural cells identify an effect for CRP on the brain. The results of Bochao et al. provide further evidence for a causal protective effect of CRP on schizophrenia, and future studies will hopefully shed light on the biological mechanism behind this remarkable observation.

References

1. Prins BP, Abbasi A, Wong A, et al. Investigating the Causal Relationship of C-Reactive Protein with 32 Complex Somatic and Psychiatric Outcomes: A Large-Scale Cross-Consortium Mendelian Randomization Study. 2016; 13(6): e1001976.

2. Fernandes BS, Steiner J, Bernstein HG, et al. C-reactive protein is increased in schizophrenia but is not altered by antipsychotics: meta-analysis and implications. *Molecular psychiatry* 2016; 21(4): 554-64.

3. Hartwig FP, Borges MC, Horta BL, Bowden J, Davey Smith G. Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study. *JAMA psychiatry* 2017; 74(12): 1226-33.

4. Ligthart S, Vaez A, Vosa U, et al. Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. *American journal of human genetics* 2018; 103(5): 691-706.

5. Lin B, Alkema A, Peters T, et al. Assesing the causal links between metabolic traits, inflammation and schizophrenia: a univariable and multivariable bidirectional Mendelian Randomization study. *Int J Epidemiol* 2019; in press.

6. Tillett WS, Francis T. SEROLOGICAL REACTIONS IN PNEUMONIA WITH A NON-PROTEIN SOMATIC FRACTION OF PNEUMOCOCCUS. *The Journal of experimental medicine* 1930; 52(4): 561-71.

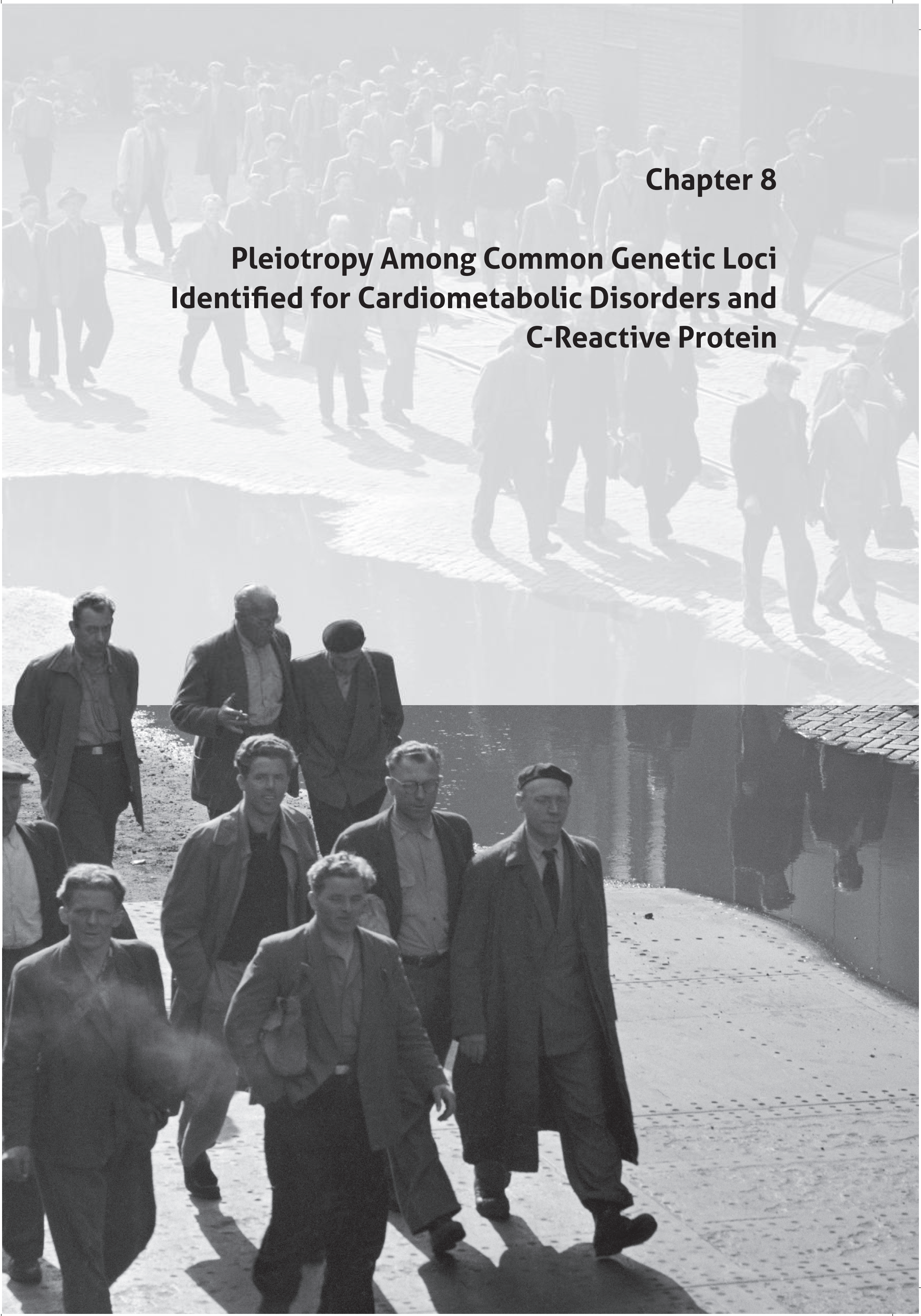
7. Sproston NR, Ashworth JJ. Role of C-Reactive Protein at Sites of Inflammation and Infection. *Frontiers in immunology* 2018; 9: 754.

8. Blomstrom A, Karlsson H, Wicks S, Yang S, Yolken RH, Dalman C. Maternal antibodies to infectious agents and risk for non-affective psychoses in the offspring--a matched case-control study. *Schizophrenia research* 2012; 140(1-3): 25-30.

9. Rautanen A, Pirinen M, Mills TC, et al. Polymorphism in a lincRNA Associates with a Doubled Risk of Pneumococcal Bacteremia in Kenyan Children. *American journal of human genetics* 2016; 98(6): 1092-100.

Chapter 8

Pleiotropy Among Common Genetic Loci Identified for Cardiometabolic Disorders and C-Reactive Protein



Background: Pleiotropic genetic variants have independent effects on different phenotypes. C-reactive protein (CRP) is associated with several cardiometabolic phenotypes. Shared genetic backgrounds may partially underlie these associations.

Methods: We conducted a genome-wide analysis to identify the shared genetic background of inflammation and cardiometabolic phenotypes using published genome-wide association studies (GWAS). We also evaluated whether the pleiotropic effects of such loci were biological or mediated in nature. First, we examined whether 283 common variants identified for 10 cardiometabolic phenotypes in GWAS are associated with CRP level. Second, we tested whether 18 variants identified for serum CRP are associated with 10 cardiometabolic phenotypes. We used a Bonferroni corrected p-value of 1.1×10^{-04} (0.05/463) as a threshold of significance. We evaluated the independent pleiotropic effect on both phenotypes using individual level data from the Women Genome Health Study.

Results: Evaluating the genetic overlap between inflammation and cardiometabolic phenotypes, we found 13 pleiotropic regions. Additional analyses showed that 6 regions (*APOC1, HNF1A, IL6R, PPP1R3B, HNF4A* and *IL1F10*) appeared to have a pleiotropic effect on CRP independent of the effects on the cardiometabolic phenotypes. These included loci where individuals carrying the risk allele for CRP encounter higher lipid levels and risk of type 2 diabetes. In addition, 5 regions (*GCKR, PABPC4, BCL7B, FTO* and *TMEM18)* had an effect on CRP largely mediated through the cardiometabolic phenotypes.

Conclusion: The results show genetic pleiotropy among inflammation and cardiometabolic phenotypes. In addition to reverse causation, the data suggest that pleiotropic genetic variants partially underlie the association between CRP and cardiometabolic phenotypes.

Introduction

The risk of cardiometabolic diseases, the world's leading cause of mortality, is higher in people with elevated levels of systemic inflammation, independent of traditional cardiometabolic risk factors[1]. Elevated levels of C-reactive protein (CRP), as a measurement of systemic inflammation, are associated with hypertension[2], type 2 diabetes (T2D)[3,4], coronary artery disease (CAD)[1,5,6], stroke[7,8], peripheral artery disease[9], and mortality[10]. Although observational data suggest a link between CRP and cardiometabolic phenotypes, Mendelian randomization approaches have provided evidence against a causal link between CRP and these cardiometabolic phenotypes[11,12,13,14].

Genome-wide association studies (GWAS) have discovered multiple single-nucleotide polymorphisms (SNPs) associated with inflammatory markers including CRP and different cardiometabolic phenotypes including T2D, coronary artery disease (CAD), lipids and hypertension [15-21]. From these GWAS we already learned that several genes, such as *IL6R*, *APOC1*, *GCKR* and *HNF1A*, are associated both with systemic inflammation and cardiometabolic phenotypes such as CAD, lipids and diabetes[15,17,21,22]. This phenomenon of one genetic locus affecting more than one phenotype is called genetic "pleiotropy". In general, two types of pleiotropy can be defined. As previously defined by Solovieff et al., "biological pleiotropy" refers to a gene that has independent biological effects on more than one phenotype, and "mediated pleiotropy" refers to the situation where the genetic effect on phenotype B is mediated by phenotype A that is causally related to phenotype B[23]. Although both types of pleiotropy are interesting, only biological pleiotropy refers to the genuine pleiotropy where the effect of the genetic variant on two or more phenotypes is independent.

We hypothesize that in addition to reverse causation, genetic loci with pleiotropic effects may underlie the association between CRP and cardiometabolic phenotypes. To this end, we applied a simple and robust approach to point out these pleiotropic genetic variants[24]. First, we examined whether common variants identified for cardiometabolic phenotypes are associated with serum CRP levels as a measure of systemic inflammation. Second, we conversely examined whether variants so far identified for serum CRP associate with cardiometabolic phenotypes. In addition, we adjusted the association between the SNP and CRP for the cardiometabolic phenotypes and vice versa to distinguish a genuine biological pleiotropic effect from mediated pleiotropy.

Methods

Study design

To examine the overlap between genes for inflammation and cardio-metabolic disorders we collected GWAS meta-analyses data from published GWAS on cardiometabolic phenotypes and CRP[15,16,17,19,22]. These GWAS are mainly conducted in individuals from European ancestry (Table 1). We tested the genetic association of cardiometabolic SNPs with systemic inflammation using the largest published GWAS meta-analysis on CRP levels from the CHARGE (the Cohorts for Heart and Aging Research in Genomic Epidemiology) inflammation working group[22]. Testing the genetic association of the CRP SNPs with 10 cardiometabolic phenotypes we used the recent GWAS data from the following consortia: Coronary Artery Disease Genome-wide Replication and Meta-analysis plus the Coronary Artery Disease, CARDIoGRAMplusC4D[15], International Consortium for Blood Pressure, ICBP[16], the Meta-Analyses of Glucose and Insulin-related traits Consortium, MAGIC[17,18], DIAbetes Genetics Replication And Meta-analysis, DIAGRAM[19], The Genetic Investigation of Anthropometric Traits, GIANT[20] and Global Lipids Genetic Consortium, GLGC[21]. Additionally, we carried out analyses in a population based cohort study to explore the type of pleiotropy of the overlapping SNPs.

Table 1. Genome-wide association studies of cardiometabolic phenotypes and inflammation.

Consortium	Phenotype	Sample size	No. of Studies
GIANT[20]	BMI	249,796	62
GLGC[21]	HDLC, LDLC, TG, TC	99,900	46
ICBP[16]	SBP, DBP	69,395	29
MAGIC[19]	FG, FI	133,010	32
DIAGRAM[18]	T2D	149,821	38
CARDIoGRAMplusC4D[15]	CAD	194,427	49
CHARGE inflammation[22]	CRP	82,725	25

Abbreviations: BMI, body mass index; CAD, coronary artery disease; CARDIoGRAMplusC4D, Coronary Artery Disease Genome-wide Replication and Meta-Analysis plus Coronary Artery Disease Genetics Consortium; CHARGE, Cohorts for Heart and Aging Research in Genomic Epidemiology; CRP, c-reactive protein; DBP, diastolic blood pressure; DIAGRAM, DIAbetes Genetics Replication And Meta-analysis; FG, fasting glucose; FI, fasting insulin; GIANT, Genetic Investigation of ANthropometric Traits; GLGC, Global Lipids Genetic Consortium; HDLC, HDL-cholesterol; ICBP, International Consortium for Blood Pressure; LDLC, LDL-cholesterol; MAGIC, Meta-Analyses of Glucose and Insulin-related traits Consortium; SBP, systolic blood pressure; T2D, type 2 diabetes; TC, total cholesterol; TG, triglycerides.

Cardiometabolic SNPs and association with CRP

We first compiled a list of genome-wide significant SNPs (p-value<5×10^{-8}) previously identified in large GWAS on cardiometabolic traits to test the genetic association in the CRP

GWAS. The following cardiometabolic traits were included to generate the SNP list: coronary artery disease (51 SNPs in CARDIOGRAMplusC4D, n=130,681 with 63,746 cases)[15]; blood pressure (29 SNPs in ICBP, n=69,395)[16]; fasting glucose, fasting insulin (53 SNPs in MAGIC, n=133,010)[17,18]; type 2 diabetes (55 SNPs in DIAGRAM, n=149,821 with 34,840 cases)[19]; body-mass index (38 SNPs in GIANT, n=123,865)[20]; LDL cholesterol, HDL cholesterol, triglycerides and total cholesterol (102 loci in GLGC, n=100,184)[21]. When the SNP was not available in the CRP GWAS, we searched for a proxy with an r^2>0.8. For 6 SNPs, this was not possible. LD-based pruning was performed (r^2 threshold of 0.3) using HapMap LD information to make sure that independent SNPs were included in the analysis[25]. The SNP with the lowest p-value in one of the cardiometabolic GWAS was chosen. The final list included 283 independent SNPs that are genome-wide significantly associated with one or more cardiometabolic phenotypes.

CRP SNPs and association with cardiometabolic phenotypes

We used the publicly available GWAS meta-analyses data to test whether any of the 18 independent genome-wide significant SNPs identified in the CRP GWAS were associated with the following cardiometabolic phenotypes: LDL cholesterol, HDL cholesterol, triglycerides and total cholesterol (GLGC); body mass index (GIANT); systolic blood pressure (ICBP); coronary artery disease (CARDIoGRAMplusC4D consortium); fasting glucose and fasting insulin (MAGIC); type 2 diabetes (DIAGRAM). All available GWASs provided p-values for all 18 CRP SNPs, except the GWAS on CAD and the glycemic phenotypes which were based on a custom chip array (Metabochip array[26]) containing 79,000 SNPs and this array did not include 8 of the CRP SNPs. For the SNPs that were not on the Metabochip, we used for fasting glucose and fasting insulin the previous GWAS published by Dupuis et al.[17], for type 2 diabetes only the stage 1 GWAS including all HapMap SNPs[19] and for CAD we used the summary data from the CARDIoGRAM meta-analysis only[27].

CRP and cardiometabolic measures

Coronary artery diseases was defined in the CARDIoGRAMplusC4D consortium using standard criteria for myocardial infarction or coronary artery disease namely symptoms of angina pectoris, previous myocardial infarction or cardiac intervention[15]. Hypertension was defined in the ICBP consortium as systolic blood pressure ≥140 mmHg or diastolic blood pressure ≤90 mmHg[16]. Fasting glucose and fasting insulin were measured in MAGIC using standard laboratory techniques[17]. Type 2 diabetes was in the DIAGRAM consortium defined as fasting plasma glucose level ≥7.0 mmol/l or non-fasting glucose plasma level ≥11.0 mmol/l and/or treatment with oral antidiabetic medication or insulin[19]. LDL cholesterol, HDL cholesterol, triglycerides and total cholesterol were measured in the GLGC using standard laboratory techniques[21].

We used the discovery panel of the recently published GWAS meta-analysis on serum CRP (CHARGE Inflammation)[22]. The meta-analysis included 15 studies in the discovery panel (n=65,000). CRP was natural log-transformed (lnCRP) and effects represented a 1-unit change in lnCRP per copy increase in risk allele.

Statistical methods

In this study we evaluated 463 possible SNP-phenotype associations including 283 independent cardiometabolic SNPs in the CRP GWAS and 18 independent CRP SNPs in 10 different cardiometabolic GWAS. To address the issue of multiple testing we used a Bonferroni corrected alpha of 1.1×10^{-4} (0.05/463 tests) as a robust threshold for a significant association between the SNP and the phenotype in our study[28].

In a quantile-quantile (Q-Q) plot, a nominal probability distribution is compared against an empirical distribution. In the scenario that the nominal p-values form a straight line on a Q-Q plot when they are plotted against the empirical distribution, all relations are null. When the observed distribution is deflected to the left from the uniform null distribution, lower p-values are observed compared to that expected by chance (enrichment). We used QQ-plots to evaluate whether SNPs that are genome-wide significant associated with the cardiometabolic phenotype, were in the CRP GWAS distributed differently from what is expected under the null-hypothesis. Vice versa, we evaluated whether genes identified for CRP were in the cardiometabolic GWAS distributed differently from what is expected under the null-hypothesis. We used Fisher's combined probability test to test for significant enrichment in the QQ-plots[29].

Evaluation of the type of pleiotropy

To clarify the type of genetic pleiotropy (biological or mediated), we performed additional analyses in the Women's Genome Health Study (WGHS) including 23,294 women[30]. In the first model, we analyzed the age-adjusted association between CRP (dependent variable) and the lead SNP for CRP in the pleiotropic regions. To examine whether the association is independent of cardiometabolic traits we further adjusted this association for BMI, lipid levels (HDL-cholesterol, LDL-cholesterol, triglycerides and total cholesterol) and HbA1C. We used HbA1C as a proxy for glycemic metabolism given the fact that glycated hemoglobin is an acceptable marker of average blood glucose level in the last 2-3 months[31]. In addition, we adjusted the association for age and in a stepwise manner we added lipids, BMI and HbA1C to the model to evaluate the different effects of the phenotypes on the association. Last, we analyzed the association between the pleiotropic SNP and the associated cardiometabolic phenotypes unadjusted and adjusted for CRP. As we tested 43 SNP-phenotype associations in the WGHS, we used a Bonferroni corrected alpha of 1.2×10^{-03} as a threshold of study-wide significance. All regression analyses were carried out in the statistical software R version 2.15.3[32].

Women's Genome Health Study (WGHS)

The WGHS is a prospective cohort of female North American health care professionals representing participants in the Women's Health Study who provided a blood sample at baseline and consent for blood-based analyses. Participants were 45 or older at enrollment and free of cardiovascular disease, cancer or other major chronic illness. The current data are derived from 23,294 WGHS participants with whole genome genetic data and verified self-reported, European ancestry. The study protocol was approved by the institutional review board of the Brigham and Women's Hospital (Boston, MA, USA). All participants provided written informed consent to participate in the study.

Covariates WGHS

BMI (weight in kilograms divided by height in meters squared) was calculated from responses to the baseline questionnaire. All baseline blood samples underwent measurement for high-sensitivity C-reactive protein (hsCRP) via a validated immunoturbidometric method (Denka Seiken, Tokyo, Japan). Concentrations of total cholesterol (TC) and HDL-C were measured enzymatically on a Hitachi 911 autoanalyzer (Roche Diagnostics) with day-to-day reproducibility of 1.36% and 1.07% for TC concentrations of 129.8 and 277.2 mg/dL, respectively, (throughout this report, concentrations and units given are those reported in the referenced sources) and of 1.98% and 2.68% for HDL-C concentrations of 35 and 55 mg/dL, respectively. LDL-C was determined directly (Genzyme) with reproducibility of 2.16% and 1.98% for concentrations of 76.2 and 148.7 mg/dL, respectively. Triglycerides were measured enzymatically, with correction for endogenous glycerol, using a Hitachi 917 analyzer and reagents and calibrators from Roche Diagnostics; reproducibility was 1.52% and 1.49% for triglyceride concentrations of 82.5 and 178.8 mg/dL, respectively. Hemoglobin A1c was measured using turbidimetric immunoinhibition directly from packed red blood cells (Roche Diagnostics) with reproducibility of 3.63% and 3.77% at levels of 5.2% and 8.8%, respectively. A total of 22,773 participants with genotyped and covariates available were included in this study.

Genotyping WGHS

DNA extracted from the baseline blood samples underwent SNP genotyping via the Illumina Infinium II assay for querying of a genome-wide set of SNPs from the Illumina HumanHap300 Duo "+" platform. This panel including the standard content of approximately 318,237 SNPs covering the entire genome from the HumanHap300 panel with an additional focused panel of 45,571 SNPs selected to enhance coverage of cardiovascular candidate genes and SNPs with suspected functional consequences. For the current analysis, all samples had successful genotype information for >98% of the SNPs, while all SNPs had successful genotype information for >90% of the samples. SNPs with significance p-value$<10^{-6}$ for deviations from Hardy-Weinberg equilibrium were excluded from analysis. Self-reported European ancestry was confirmed in the 23,294 samples on the basis of a principal component analysis using PLINK among 1,443 ancestry informative SNPs selected for Fst>0.4 in the HapMap2. In total, 339,875 genotyped SNPs passing the criteria for inclusion also had minor allele frequency at least 1 percent. On the basis of linkage disequilibrium relationships in the HapMap (release 22), genotypes for a total of 2,621,896 SNPs were imputed from the 23,294 samples passing the quality criteria using Mach v. 1.0.16.

Pathway analysis

Pathway analysis was performed on the pleiotropic loci that we identified using Ingenuity Pathway Analysis software tool (IPA Ingenuity Systems). The Ingenuity Knowledge Base (including only genes) was used as a reference set and we considered molecules and/or direct and indirect relationships. The confidence filter was set to experimentally observed or high (predicted). Pathways were generated with a maximum size of 35 genes and we allowed up to 25 pathways. The significance p-value associated with enrichment of pathways was calculated using the right-tailed Fisher's exact test, considering the number of query molecules that participate in that pathway and the total number of molecules that are known to be associated with that pathway in the reference set. A False Discovery Rate of five percent was used as a threshold of significance using the Benjamini-Hochberg method.

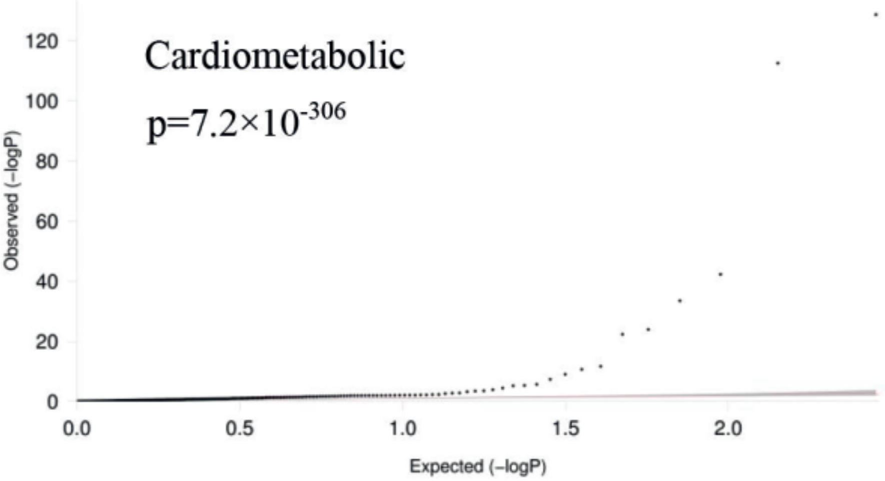
Results

Cardiometabolic SNPs in CRP GWAS

First, we used QQ-plots to evaluate whether the p-values for the associations of the 283 cardiometabolic SNPs with serum CRP are distributed differently from what is expected under the null hypothesis in each trait group. As depicted in Figure 1, the leftward deviation of the dotted lines in the QQ-plots shows that the 283 SNPs known for cardiometabolic phenotypes to have p-values smaller than expected under the null hypothesis in the CRP GWAS (p-value: 7.2×10^{-306}).

A total of 19 SNPs out of 283 independent cardiometabolic SNPs (6.7%) were associated with CRP after correction for multiple testing (p-value threshold 1.1×10^{-4}). These 19 SNPs were located within or close to 12 different genes *APOC1, HNF1A, GCKR, IL6R, PPP1R3B, HNF4A, PABPC4, BCL7B, FTO, TMEM18, PLTP* and *MC4R.* Table 2 shows the SNPs with the

Figure 1. Quantile-quantile plot of cardiometabolic SNPs in CRP GWAS.



QQ-plot was used to evaluate whether SNPs that are genome-wide significant associated with the cardiometabolic phenotypes, were in the CRP GWAS distributed differently from what is expected under the null-hypothesis. The observed p-values (dotted line) for the phenotypes deviated significantly leftwards indicating that these p-values are smaller than expected under null hypothesis.

lowest p-values in the 12 pleiotropic loci based on the CRP GWAS, i.e. the lowest p-value in that genomic locus. The eight SNPs in Table 2 with the lowest p-value were already known to be associated with CRP based on the recent CRP GWAS[22]. The next four SNPs were not identified in the genome-wide association study of CRP. The first novel association was rs1558902 with a p-value of 2.2×10^{-6}. This SNP is located in the first intron of the *FTO* gene on chromosome 16. The second novel signal was the SNP rs2867125 which is located on chromosome 2, near 46kb downstream of *TMEM18*. This SNP had a p-value of 5.0×10^{-6} in the CRP meta-analysis. The third association was with rs6065906 which is located on chromosome 20, near the *PLTP* and *PCIF1* gene (p-value=6.7×10^{-6}). The last finding was rs571312 which is located 2 Mb upstream of the *MC4R* gene on chromosome 18 (p-value=3.8×10^{-5}).

Among the associated SNPs, we observed many SNPs with different directions of effect on the cardiometabolic phenotypes and CRP than one would expect based on the association of CRP and these phenotypes in observational data. As an example, the A-allele of the SNP rs4420638 in the *APOC1* locus increases serum CRP levels. However, this allele is associated with a decrease in the level of total cholesterol, LDL-cholesterol and triglycerides. We also observed such effects for the G-allele of the SNP rs1183910 in the *HNF1A* locus. This allele increases serum CRP levels and is associated with a decline in total cholesterol and LDL-cholesterol.

Table 2. The association of known loci for cardiometabolic traits with serum CRP.

SNP	Band	A1/A2[a]	Effect[b](SE)	P-value	Gene	Phenotypes (effect direction)	Top-SNP[c] (r^2;P-value)
rs4420638	19q13.32	A/G	0.240(0.010)	2.1×10^{-129}	*APOC1*	TG(-),TC(-),HDLC(+),LDLC(-)	The same
rs1169288	12q24.31	A/C	0.152(0.007)	3.3×10^{-113}	*HNF1A*	TC(-),LDLC(-),T2D(+)	rs1183910 (0.96;3.3×10^{-113})
rs1260326	2p23.3	T/C	0.089(0.007)	5.5×10^{-43}	*GCKR*	TC(+),TG(+),FG(-),FI(-)	The same
rs4845625	1q21.3	T/C	0.062(0.006)	4.8×10^{-23}	*IL6R*	CAD(+)	rs4129267 (0.52;1.1×10^{-47})
rs9987289	8p23.1	G/A	0.079(0.011)	2.3×10^{-12}	*PPP1R3B*	TC(+),HDLC(+),LDLC(+),FI(-),FG(-)	The same
rs1800961	20q13.12	C/T	0.120(0.018)	2.3×10^{-11}	*HNF4A*	TC(+),HDLC(+),T2D(+)	The same
rs4660293	1p32.4	G/A	0.044(0.007)	9.9×10^{-10}	*PABPC4*	HDLC(-)	rs12037222 (0.96;4.5×10^{-10})
rs17145738	7q11.23	C/T	0.054(0.010)	4.7×10^{-8}	*BCL7B*	HDLC(-),TG(+)	rs13233571 (1.00;2.8×10^{-8})
rs1558902	16q12.2	A/T	0.032(0.007)	2.2×10^{-6}	*FTO*	BMI(+),T2D(+)	The same
rs2867125	2p25.3	C/T	0.038(0.008)	5.0×10^{-6}	*TMEM18*	BMI(+)	rs10189761 (0.93;1.2×10^{-6})
rs6065906	20q13.12	T/C	0.036(0.008)	6.7×10^{-6}	*PLTP*	HDLC(+),TG(-)	rs6073972 (1.00;2.9×10^{-6})
rs571312	18q22	A/C	0.033(0.008)	3.8×10^{-5}	*MC4R*	BMI(+)	The same

[a] Effect represents 1-unit change in the natural log-transformed CRP (mg/L) per copy increase in the risk allele. SE, standard error.

[b] A1 and A2 represent respectively the risk allele and non-risk allele.

[c] Top-SNP represents the SNP with the lowest p-value in the genomic region in the CRP meta-analysis. If the top-SNP is “The same”, the top SNP for the cardiometabolic traits is the same as the SNP with the lowest p-value in the CRP meta-analysis.

Note: p-value ≤ 1.1×10^{-4} is considered as study-wide significant (0.05/463).

Abbreviations: BMI, body mass index; CAD, coronary artery disease; FG, fasting glucose; FI, fasting insulin; HDLC, high-density lipoprotein cholesterol; LDL, low-density lipoprotein cholesterol; T2D, type 2 diabetes; TC, total cholesterol; TG, triglycerides.

Out of the 12 pleiotropic loci, 6 loci had the same lead SNP in both the CRP and one or more of the cardiometabolic GWAS. In the other 6 loci the lead SNPs were different between the CRP GWAS and the cardiometabolic GWAS. However, in the majority of these loci the lead SNPs of the cardiometabolic GWAS were in high LD (r^2>0.8) with the lead SNP in the CRP GWAS. In the *IL6R* locus we observed the lowest LD between the top hit in the CRP GWAS and the CAD GWAS (r^2=0.52).

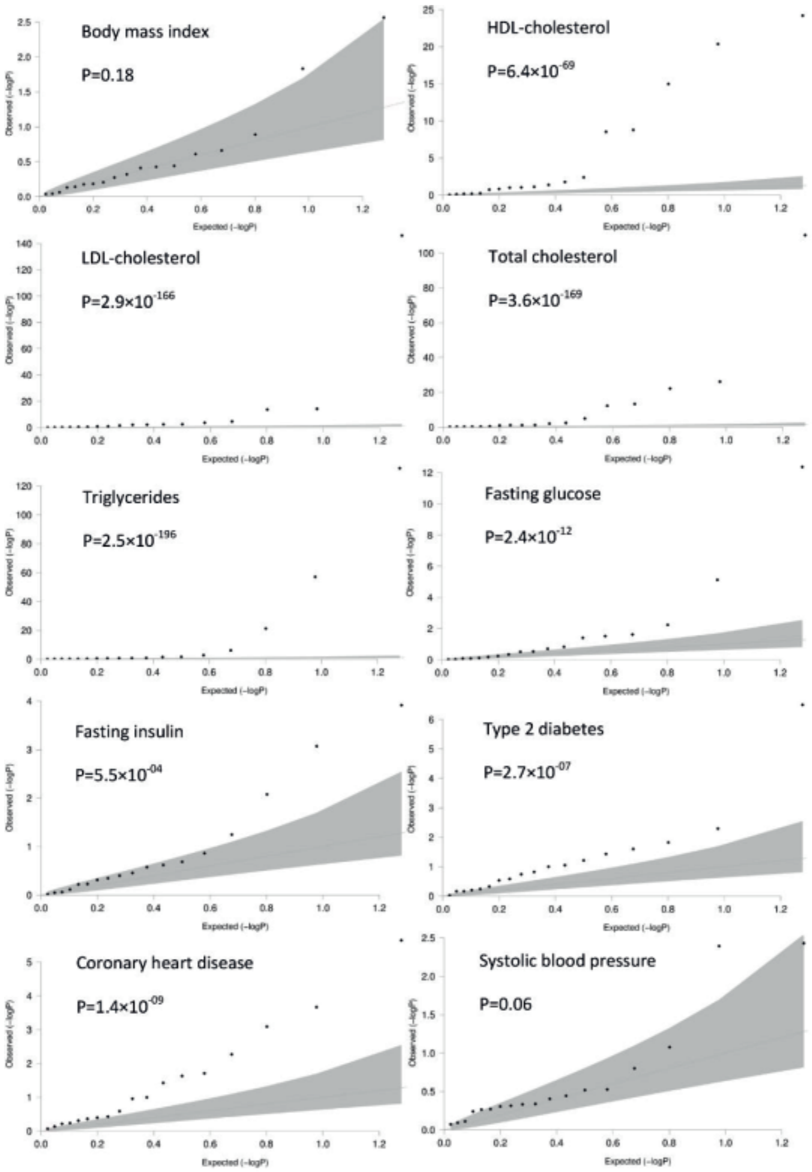
CRP SNPs in cardiometabolic GWAS

We used the same QQ-plots as described previously to evaluate whether the association p-values for the 18 CRP SNPs are distributed differently from what is expected under the null hypothesis in the different cardiometabolic GWAS. As depicted by the leftward deviation of the dotted lines in the QQ-plots for CAD (P-value=1.4×10^{-9}), the cholesterol phenotypes (HDL-cholesterol, P-value=6.4×10^{-69}; LDL-cholesterol, P-value=2.9×10^{-166}; total cholesterol, P-value=3.6×10^{-169} and triglycerides, P-value=2.5×10^{-196}) and the glycemic phenotypes (fasting glucose, P-value=2.4×10^{-12} and fasting insulin 5.5×10^{-4}), the p-values for the association between the 18 CRP SNPs and these phenotypes are significantly smaller than expected under the null hypothesis (Figure 2). We did not observe such a significant deviation in the QQ-plots of BMI (P-value=0.18) and SBP (P-value=0.06).

Results of the association of the 18 genome-wide significant associations with CRP-level are depicted in Figure 3 (Tables S2 and S3). We observed 9 associations with one or more of the 10 cardiometabolic phenotypes close to or within the genes *APOC1, HNF1A, IL6R, GCKR, IL1F10, PPP1R3B, HNF4A, PABPC* and *BCL7B* (p-value<1.1×10^{-4}). Only 1 gene (*IL1F10*) was not identified in the previous analysis where we tested the association between the cardiometabolic SNPs and CRP. Among all 9 associations, we found three associations that are not reported in the GWAS for that specific phenotype. The first was rs1183910 with CAD (p-value 5.6×10^{-6}). This SNP is located in the first intron of the *HNF1A* gene on chromosome 12. The second was rs6734238 with total cholesterol (p-value 1.16×10^{-5}). This SNP is located on chromosome 2, nearby the *IL1F10* gene and other interleukin 1 family genes. The third was rs4420638 with T2D (p-value 4.0x10^{-6}) nearby the *APOC1* gene on chromosome 19.

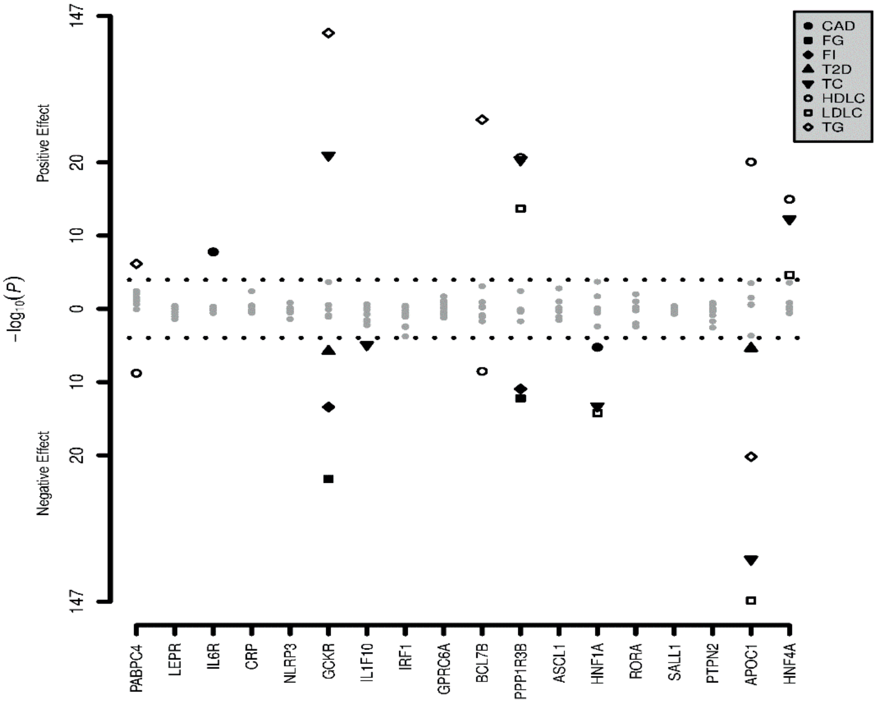
Comparable with the previous associations results, we observed many different direction of effects. For instance, the A-allele of the SNP rs4420638 in the *APOC1* locus increases serum CRP levels and is associated with a lower risk of type 2 diabetes. Furthermore, the G-allele of the SNP rs1183910 in the *HNF1A* locus increases serum CRP levels and is associated with a lower risk of coronary artery disease.

Figure 2. Quantile-quantile plots of CRP SNPs in cardiometabolic GWAS.



QQ-plots were used to evaluate whether SNPs that are genome-wide significant associated with CRP, were in the cardiometabolic GWAS distributed differently from what is expected under the null-hypothesis. The observed p-values (dotted line) for the phenotypes HDL-cholesterol, fasting glucose, type 2 diabetes and coronary artery disease deviated significantly leftwards indicating that these p-values are smaller than expected under null hypothesis.

Figure 3. P-values for the associations of the 18 CRP SNPs with different cardiometabolic phenotypes.



P-values for the associations between the 18 CRP SNPs and BMI, lipids, glycemic phenotypes, SBP and coronary artery disease. The genes on the x-axis represent the genes in which the CRP SNPs are located or closest by. The numbers on the y-axis indicate the p-values of the association between the SNPs and the cardiometabolic phenotypes. Significant associations are colored as depicted in the figure legend. For BMI and SBP, no significant associations were observed. CAD, coronary artery disease; FG, fasting glucose; FI, fasting insulin; HDLC, HDL-cholesterol; LDLC, LDL-cholesterol; T2D, type 2 diabetes; TC, total cholesterol; TG, Triglycerides.

Exploring the type of pleiotropy

We observed a total number of 13 genetic regions with pleiotropic effects on CRP and cardiometabolic phenotypes: 12 regions identified in the first step testing the cardiometabolic SNPs with CRP and 1 additional region identified in the second step testing the associations of the CRP SNPs with the cardiometabolic phenotypes. Table 3 shows the unadjusted and adjusted associations between the 13 overlapping SNPs and CRP-level using individual level data from the WGHS. There was no significant association in the WGHS between the SNPs located near *PLTP* and *MC4R* and CRP after correction for multiple testing. The effect sizes of the genetic loci in or near the genes *APOC1, HNF1A, IL6R, PPP1R3B, HNF4A* and *IL1F10* did not diminish substantially after adjustment for BMI,

Table 3. Pleiotropic SNPs and their association with CRP.

		MODEL 1[a]		*MODEL 2*[b]			
SNP	**Chr**	**Effect(se)**	**P-value**	**Effect(se)**	**P-value**	**Gene**	**pleiotropy**[c]
rs4420638	19	0.269(0.019)	4.4×10^{-47}	0.272(0.016)	1.7×10^{-65}	*APOC1*	B
rs1169288	12	0.165(0.012)	2.3×10^{-43}	0.160(0.010)	4.0×10^{-56}	*HNF1A*	B
rs1260326	2	0.110(0.011)	1.6×10^{-22}	0.073(0.010)	3.4×10^{-14}	*GCKR*	M
rs4845625	1	0.067(0.011)	2.0×10^{-9}	0.065(0.009)	8.8×10^{-12}	*IL6R*	B
rs9987289	8	0.076(0.019)	4.5×10^{-5}	0.086(0.016)	1.5×10^{-7}	*PPP1R3B*	B
rs1800961	20	0.146(0.033)	8.4×10^{-6}	0.141(0.028)	4.7×10^{-7}	*HNF4A*	B
rs4660293	1	0.048(0.013)	1.9×10^{-4}	0.036(0.011)	1.2×10^{-3}	*PABPC4*	M
rs17145738	7	0.075(0.017)	1.3×10^{-5}	0.019(0.015)	1.8×10^{-1}	*BCL7B*	M
rs1558902	16	0.041(0.012)	6.0×10^{-4}	0.012(0.010)	2.3×10^{-1}	*FTO*	M
rs7561317	2	0.055(0.015)	1.5×10^{-4}	0.013(0.012)	2.9×10^{-1}	*TMEM18*	M
rs6065906	20	0.026(0.014)	6.6×10^{-2}	0.039(0.012)	1.2×10^{-3}	*PLTP*	B
rs571312	18	0.038(0.013)	3.5×10^{-3}	0.006(0.011)	6.0×10^{-1}	*MC4R*	M
rs6734238	2	0.040(0.011)	3.9×10^{-4}	0.051(0.010)	1.3×10^{-7}	*IL1F10*	B

[a]Model 1: adjusted for age

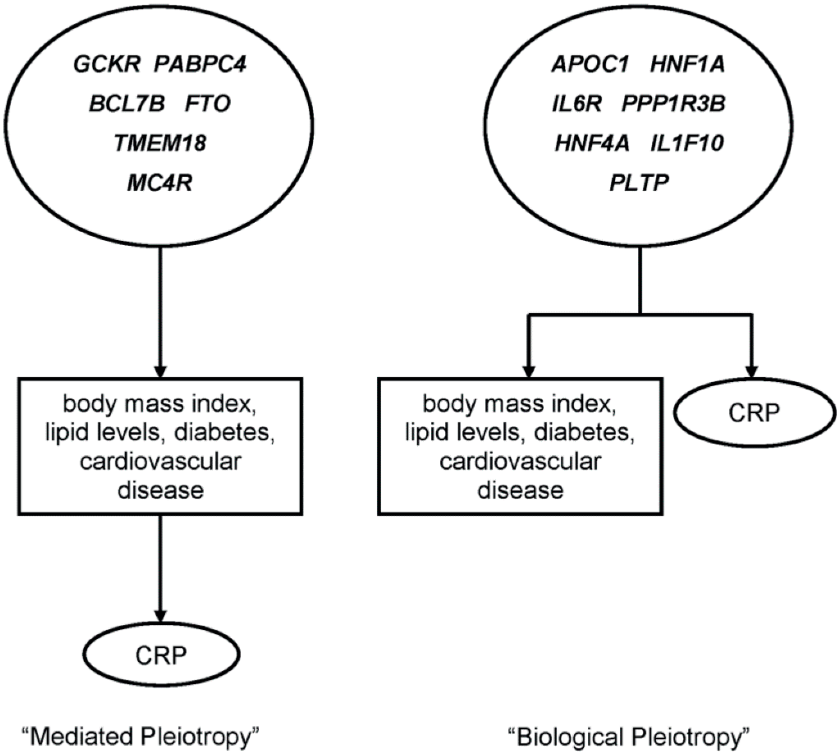
[b]Model 2: additionally adjusted for BMI, HDL-cholesterol, LDL-cholesterol, triglycerides, total cholesterol and HbA1C

[c]B: biological pleiotropy; M: mediated pleiotropy.

cholesterol levels and HbA1C suggesting biological pleiotropy. For *BCL7B, FTO* and *TMEM18* the effect sizes decreased considerably implying mediated pleiotropy. The estimate of the association between rs1260326 (*GCKR*) and CRP decreased substantially after adjustment but was still strongly associated. We observed the same scenario for the association between rs4660293 (*PABPC4*) and CRP. When we added the phenotypes in a stepwise manner to the model, we observed for the mediated pleiotropic loci *FTO* and *TMEM18* that the effect was mainly mediated through BMI (Table S3). For *BCL7B* and *PABPC4*, lipids appeared to be the most important mediators. Figure 4 shows graphically the biological and mediated pleiotropic effects.

The results for the associations between the pleiotropic SNPs and the associated cardiometabolic phenotypes are presented in Table S4. Eight SNPs were not significantly associated with the cardiometabolic phenotype in the WGHS after adjustment for multiple testing. The majority of the estimates in- or decreased slightly after adjustment for CRP. However, the estimates between *APOC1* and HbA1C, *PABPC4* and triglycerides and *BCL7B* and HDL-cholesterol decreased considerably after the adjustment for CRP.

Figure 4. Biological and mediated pleiotropy of overlapping loci among inflammation and cardiometabolic phenotypes.


GCKR PABPC4
BCL7B FTO
TMEM18
MC4R
body mass index, lipid levels, diabetes, cardiovascular disease
CRP
"Mediated Pleiotropy"
APOC1 HNF1A
IL6R PPP1R3B
HNF4A IL1F10
PLTP
body mass index, lipid levels, diabetes, cardiovascular disease
CRP
"Biological Pleiotropy"

Overlapping loci among inflammation and cardiometabolic phenotypes and type of pleiotropy according to the additional analyses. We identified six overlapping loci with mediated pleiotropic effects on CRP (left) and seven with a biological pleiotropic effect (right).

Pathway analysis

The results from the pathway analysis including all 13 pleiotropic genes are listed in the Table S5. A total number of 13 canonical pathways were significantly enriched using an FDR of five percent. The top pathways included the FXR/RXR activation (P-value=7.4×10^{-9}) , LXR/RXR activation (P-value=4.6×10^{-5}), Maturity Onset Diabetes of the Young (MODY) signaling (P-value=7.6×10^{-5}), hepatic cholestasis (P-value=1.1×10^{-4}) and acute phase response signaling (P-value=1.3×10^{-4}).

We observed several overlapping common genetic risk factors for cardiometabolic phenotypes and systemic inflammation. The additional analyses provided evidence for six biological pleiotropic loci with independent effects on both CRP and the cardiometabolic phenotype. These pleiotropic loci suggest a shared genetic background for CRP and cardiometabolic phenotypes. In addition, 5 pleiotropic loci appeared to have an effect on CRP mediated through the cardiometabolic phenotypes. Taken together, our results highlight the complex shared genetic architecture of cardiometabolic phenotypes and chronic inflammation.

Several of the identified biological pleiotropic loci suggest that the association between CRP and cardiometabolic phenotypes is not only reverse causation, but also shared independent genetic effects. Both the *HNF1A* and *HNF4A* loci were associated with CRP after adjustment for the cardiometabolic phenotypes. The effect directions were the same for type 2 diabetes and CRP, implying people carrying the risk allele for type 2 diabetes also have higher CRP values. We observed this also for the *PPP1R3B* locus where people carrying the risk allele for higher cholesterol also experience higher CRP levels. In both cases the effect on CRP is independent of the effect on the corresponding cardiometabolic trait.

Three of the associations that were not reported in the GWAS on CRP-level (*FTO, TMEM18* and *MC4R*) are associations with SNPs discovered in the GWAS on BMI by Speliotes et al[20]. Moreover, these SNPs were also the leading findings in this large BMI GWAS meta-analysis. Our additional analyses clearly showed that after adjustment for BMI, the effects of *FTO* and *TMEM18* decreased substantially, resulting in a non-significant association, which suggests that their effect on inflammation is indeed mediated by BMI. This is in line with previous research that already provided evidence for a causal role of BMI in inflammation[14]. Conversely, none of the SNPs identified in the CRP GWAS were associated with BMI when we tested these SNPs in the BMI GWAS.

Our results also suggest a role for lipids in systemic inflammation. When we adjusted the association between *BCL7B* loci and CRP for the cardiometabolic phenotypes including lipids, the association was not present anymore. This locus appears to increase systemic inflammation through their effect on lipids. Also the association between *PABPC4* and CRP decreased after adjustment for CRP, but there was a significant residuals effect suggesting partly mediated effects through lipids. The observation that lipids may cause inflammation is in line with previous studies that have shown an important role for oxidized LDL-cholesterol molecules and free fatty acids in the development of systemic inflammation[33]. However, in addition to the mediated pleiotropic loci among lipids and CRP, we also observed loci with independent effects (biological pleiotropy) on lipids and CRP including *APOC1, HNF1A* and *HNF4A*, highlighting the complex interrelationship of lipids and inflammation. Moreover, the pathway analysis confirmed the role of the pleiotropic genes in both inflammation and lipid metabolism.

We observed little overlap between risk loci for CAD and CRP. Apart from the *IL6R* gene, our results suggest an association with CAD at the *HNF1A* locus. The *HNF1A* gene is an important hepatic nuclear transcription factor that has been associated in GWAS with lipids and diabetes[19,21]. This gene is known to regulate many target genes involved in lipid metabolism and transport[34]. A previous study has associated this locus with different cardiovascular phenotypes including coronary artery calcification and incident CHD[35]. Unfortunately we were not able to look-up 9 CRP SNPs in the larger CAD Metabochip GWAS because these variants were not on the Metabochip and no appropriate proxies were available. This might partly explain the little overlap between CRP and CAD genetic risk variants.

In the additional analyses we used glycated hemoglobin (HbA1C) to adjust for fasting glucose, fasting insulin, T2D and other components of the glucose homeostasis. HbA1C represents the average glucose level in the last 3 months, implying that this is only a proxy for the complex glucose homeostasis rather than a covariate that reflects its entire biological metabolism. Therefore, there may still be residual confounding from other biological pathways that have an effect on glucose and insulin levels. This could explain the observed residual effect of *GCKR* on CRP after adjustment for the cardiometabolic phenotypes.

In the evaluation of the type of pleiotropy, we adjusted the association between the pleiotropic SNP and CRP for the cardiometabolic phenotypes. For some variants we observed a convincing attenuation in the effect estimates (*BCL7B*, *FTO* and *TMEM18*). For other variants, the attenuation was less pronounced (*GCKR* and *PABPC4*). From these results we cannot conclude whether the residual effect is residual confounding or a true residual effect. Additionally, for several variants the effect estimates were the same or even increased after adjustment suggesting biological pleiotropy. The latter increase in estimate might be due to negative confounding where the SNP has an opposite direction of effect on the cross-associated phenotype compared to CRP and the effect of this phenotype on CRP is in the same direction (negative confounding). We also analyzed the association between the pleiotropic SNP and the cardiometabolic phenotypes unadjusted and adjusted for CRP. Although there is ample of evidence against a causal role for CRP in the development of cardiometabolic phenotypes, for some associations the effect estimates attenuated considerably[11,12,13,14]. This might be explained by the fact that CRP is correlated with many intermediate phenotypes that mediate the association between the SNP and the cardiometabolic phenotype.

Among some pleiotropic SNPs we observed opposite direction of effects on the phenotypes than one would expect based on their effects on the health of the possessor and the association of CRP and these cardiometabolic phenotypes in observational data. This phenomenon is known as “antagonistic pleiotropy”[36]. For instance, the SNP in the *HNF1A* locus increases serum CRP level according to the G allele and decreases LDL-cholesterol level. For biological pleiotropic loci we can substantiate this antagonistic effect. A genetic locus may have a deleterious effect on one phenotype, but an independent beneficial effect on a second phenotype. An explanation for these findings might be the fact that the effect sizes and variances explained by the genetic variants are small and therefore they only play a minor role in the phenotypical correlations. Moreover, the high frequency of seemingly detrimental alleles in human populations may partly be the effect of antagonistic pleiotropy[37]. As expected, among the loci where no independent effect was observed (mediated pleiotropy), we did not observe antagonistic pleiotropy.

Our study has certain strengths. We used the largest available GWAS data on lipids, blood pressure, BMI, CAD, glycemic traits, T2D and CRP from the GLGC, ICBP, GIANT, CARDIoGRAMplusC4D, MAGIC, DIAGRAM and CHARGE Inflammation consortia to attain as much power as possible. By including only genome-wide significant findings, we restricted the analysis to the most robust genetic associations. Moreover, we used a conservative method to correct for multiple testing, lowering the probability of false positive findings. Nonetheless, some limitations should be acknowledged. Although we used the largest available GWAS sample sizes, the identified common genetic variants for above mentioned phenotypes only explain a modest fraction of the genetic variance of these phenotypes (ranging from 5 to 12 percent). Therefore, the effects of the cardiometabolic SNPs on CRP and vice-versa may still be too small to detect cross-phenotype associations, resulting in an underestimation of the amount of genetic overlap. Moreover, we only focused on common SNPs and it might be that also rare variants underlie the shared genetic associations. The method we applied to distinguish “biological” from “mediated” pleiotropy is a classical and widely used approach in the field of epidemiology. However, we cannot completely rule out reverse causation or unknown confounders as potential drivers of the association between the genetic variant and CRP. Furthermore, we only studied GWAS including participants of European ancestry. We are aware of differences in haplotype structures between different ethnicities; however, the results are likely to be generalizable given the biological pathways.

In conclusion, we observed several genetic loci with independent effects on both CRP and one or more cardiometabolic phenotypes. These results suggest that the association between CRP and cardiometabolic phenotypes is partly explained by a shared genetic background.

References

1. Danesh J, Wheeler JG, Hirschfield GM, et al. C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *N Engl J Med* 2004; 350(14): 1387-97.

2. Sesso HD, Buring JE, Rifai N, Blake GJ, Gaziano JM, Ridker PM. C-reactive protein and the risk of developing hypertension. *JAMA* 2003; 290(22): 2945-51.

3. Dehghan A, Kardys I, de Maat MP, et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. *Diabetes* 2007; 56(3): 872-8.

4. Pradhan AD, Manson JE, Rifai N, Buring JE, Ridker PM. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA* 2001; 286(3): 327-34.

5. Cushman M, Arnold AM, Psaty BM, et al. C-reactive protein and the 10-year incidence of coronary heart disease in older men and women: the cardiovascular health study. *Circulation* 2005; 112(1): 25-31.

6. Blake GJ, Rifai N, Buring JE, Ridker PM. Blood pressure, C-reactive protein, and risk of future cardiovascular events. *Circulation* 2003; 108(24): 2993-9.

7. Ridker PM, Cushman M, Stampfer MJ, Tracy RP, Hennekens CH. Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *N Engl J Med* 1997; 336(14): 973-9.

8. Rost NS, Wolf PA, Kase CS, et al. Plasma concentration of C-reactive protein and risk of ischemic stroke and transient ischemic attack: the Framingham study. *Stroke* 2001; 32(11): 2575-9.

9. Ridker PM, Cushman M, Stampfer MJ, Tracy RP, Hennekens CH. Plasma concentration of C-reactive protein and risk of developing peripheral vascular disease. *Circulation* 1998; 97(5): 425-8.

10. Hindorff LA, Rice KM, Lange LA, et al. Common variants in the CRP gene in relation to longevity and cause-specific mortality in older adults: the Cardiovascular Health Study. *Atherosclerosis* 2008; 197(2): 922-30.

11. Elliott P, Chambers JC, Zhang W, et al. Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA* 2009; 302(1): 37-48.

12. Collaboration CRPCHDG. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ: British Medical Journal* 2011; 342.

13. Smith GD, Lawlor DA, Harbord R, et al. Association of C-reactive protein with blood pressure and hypertension life course confounding and Mendelian randomization tests of causality. *Arteriosclerosis, thrombosis, and vascular biology* 2005; 25(5): 1051-6.

14. Timpson NJ, Nordestgaard BG, Harbord RM, et al. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *International Journal of Obesity* 2011; 35(2): 300-8.

15. The CDC, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2012; 45(1): 25-33.

16. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011; 478(7367): 103-9.

17. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010; 42(2): 105-16.

18. Scott RA, Lagou V, Welch RP, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* 2012; 44(9): 991-1005.

19. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012; 44(9): 981-90.

20. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010; 42(11): 937-48.

21. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; 466(7307): 707-13.

22. Dehghan A, Dupuis J, Barbalic M, 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): 731-8.

23. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 2013; 14(7): 483-95.

24. Olden M, Teumer A, Bochud M, et al. Overlap between common genetic polymorphisms underpinning kidney traits and cardiovascular disease phenotypes: the CKDGen Consortium. *American Journal of Kidney Diseases* 2013; 61(6): 889-98.

25. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008; 24(24): 2938-9.

26. Voight BF, Kang HM, Ding J, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 2012; 8(8): e1002793.

27. Schunkert H, Konig IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 2011; 43(4): 333-8.

28. McIntyre LM, Martin ER, Simonsen KL, Kaplan NL. Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. *Genet Epidemiol* 2000; 19(1): 18-29.

29. Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics* 2010; 18(1): 111-7.

30. Ridker PM, Chasman DI, Zee RY, et al. Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000 initially healthy american women. *Clin Chem* 2008; 54(2): 249-55.

31. Nathan DM, Singer DE, Hurxthal K, Goodson JD. The clinical information value of the glycosylated hemoglobin assay. *N Engl J Med* 1984; 310(6): 341-6.

32. Team RC. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria Available from: http://wwwR-projectorg/* 2012.

33. Rocha VZ, Libby P. Obesity, inflammation, and atherosclerosis. *Nat Rev Cardiol* 2009; 6(6): 399-409.

34. Odom DT, Zizlsperger N, Gordon DB, et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science* 2004; 303(5662): 1378-81.

35. Reiner AP, Gross MD, Carlson CS, et al. Common Coding Variants of the HNF1A Gene Are Associated With Multiple Cardiovascular Risk Phenotypes in Community-Based Samples of Younger and Older European-American Adults The Coronary Artery Risk Development in Young Adults Study and The Cardiovascular Health Study. *Circulation: Cardiovascular Genetics* 2009; 2(3): 244-54.

36. Williams GC. Pleiotropy, natural selection, and the evolution of senescence. *Science's SAGE KE* 2001; 2001(1): 13.

37. Carter AJ, Nguyen AQ. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC Med Genet* 2011; 12: 160.

Supplementary material

Table S1. The associations of CRP SNPs with body mass index and cholesterol levels.

SNP	Gene	A1*	CRP	BMI	TC	HDLC	LDLC	TG
rs2794520	*CRP*	C	9.5×10^{-189}(+)	0.53(-)	0.59(+)	0.77(-)	0.35(+)	0.66(-)
rs4420638	*APOC1*	A	2.1×10^{-129}(+)	0.25(+)	5.2×10^{-111}(-)	4.4×10^{-21}(+)	8.7×10^{-147}(-)	5.4×10^{-22}(-)
rs1183910	*HNF1A*	G	3.3×10^{-113}(+)	0.86(+)	5.2×10^{-14}(-)	4.0×10^{-3}(-)	5.8×10^{-15}(-)	0.58(-)
rs4420065	*LEPR*	C	3.2×10^{-64}(+)	0.36(-)	0.08(-)	0.09(-)	0.38(-)	0.18(-)
rs4129267	*IL6R*	C	1.1×10^{-47}(+)	0.62(-)	0.68(-)	0.90(-)	0.80(-)	0.84(-)
rs1260326	*GCKR*	T	5.4×10^{-43}(+)	0.13(-)	7.3×10^{-27}(+)	0.08(-)	2.2×10^{-4}(+)	5.7×10^{-133}(+)
rs12239046	*NLRP3*	C	1.6×10^{-13}(+)	0.92(-)	0.37(-)	0.04(-)	0.95(+)	0.30(-)
rs6734238	*IL1F10*	G	3.4×10^{-13}(+)	0.74(-)	1.2×10^{-5}(-)	0.18(-)	6.5×10^{-3}(-)	0.03(-)
rs9987289	*PPP1R3B*	G	2.3×10^{-12}(+)	0.39(-)	7.1×10^{-23}(+)	6.4×10^{-25}(+)	2.0×10^{-14}(+)	0.02(-)
rs10745954	*ASCL1*	A	1.6×10^{-11}(+)	0.48(-)	0.66(-)	0.10(+)	0.03(-)	1.6×10^{-3}(+)
rs1800961	*HNF4A*	C	2.3×10^{-11}(+)	0.91(-)	5.7×10^{-13}(+)	1.1×10^{-15}(+)	2.4×10^{-5}(+)	0.59(+)
rs340029	*RORA*	T	2.6×10^{-11}(+)	0.01(+)	0.01(-)	0.67(+)	3.8×10^{-3}(-)	0.74(+)
rs10521222	*SALL1*	C	1.3×10^{-10}(+)	0.72(+)	0.57(-)	0.65(+)	0.58(-)	0.20(-)
rs12037222	*PABPC4*	A	4.5×10^{-10}(+)	0.22(+)	0.07(+)	1.6×10^{-9}(-)	8.7×10^{-3}(+)	7.0×10^{-7}(+)
rs4705952	*IRF1*	G	1.3×10^{-8}(+)	0.37(+)	$4.1x10^{-3}$(-)	0.56(-)	$3.2x10^{-3}$(-)	0.52(+)
rs2847281	*PTPN2*	A	2.2×10^{-8}(+)	$2.73x10^{-3}$(-)	0.77(+)	0.15(+)	0.64(-)	0.13(-)
rs13233571	*BCL7B*	C	2.8×10^{-8}(+)	0.66(+)	0.12(+)	2.9×10^{-9}(-)	0.14(-)	9.3×10^{-58}(+)
rs6901250	*GPRC6A*	A	4.8×10^{-8}(+)	0.66(+)	0.08(+)	0.02(+)	0.13(+)	0.25(-)

*A1 represents the risk allele according to the CRP GWAS.

Note: p-value$\leq1.1\times10^{-4}$ is considered as study-wide significant (0.05/463).

Abbreviations: BMI, body mass index; CRP, c-reactive protein; HDLC, HDL-cholesterol; LDLC, LDL-cholesterol; SNP, single-nucleotide polymorphism; TC, total cholesterol; TG, triglycerides.

Table S2. The associations of CRP SNPs with coronary artery disease, glycaemic phenotypes and blood pressure.

SNP	Gene	A1*	CRP	CAD	FG	FI	T2D	SBP
rs2794520	*CRP*	C	9.5×10^{-189}(+)	0.66(-)	0.39(-)	0.32(-)	0.29 (-)	4.0×10^{-3}(+)
rs4420638	*APOC1*	A	2.1×10^{-129}(+)	2.1×10^{-4}(-)	3.1×10^{-4} (+)	0.03(+)	4.0×10^{-6} (-)	0.30(+)
rs1183910	*HNF1A*	G	3.3×10^{-113}(+)	5.6×10^{-6}(-)	0.29(-)	0.02(+)	2.0×10^{-4} (+)	0.46(-)
rs4420065	*LEPR*	C	3.2×10^{-64}(+)	0.04(-)	0.57(+)	0.77(-)	0.94(+)	0.39(+)
rs4129267	*IL6R*	C	1.1×10^{-47}(+)	1.7×10^{-8}(+)	0.70(-)	0.88(-)	0.26 (-)	0.50(+)
rs1260326	*GCKR*	T	5.4×10^{-43}(+)	0.84(-)	2.17×10^{-41}(-)	3.8×10^{-14}(-)	1.6×10^{-6} (-)	0.30(+)
rs12239046	*NLRP3*	C	1.6×10^{-13}(+)	0.73(+)	0.15(+)	0.97(-)	0.47(-)	0.54(-)
rs6734238	*IL1F10*	G	3.4×10^{-13}(+)	0.02(-)	5.8×10^{-3}(-)	0.24(+)	0.62(+)	0.77(-)
rs9987289	*PPP1R3B*	G	2.3×10^{-12}(+)	0.81(-)	6.1×10^{-13}(-)	1.1×10^{-11}(-)	3.7×10^{-3} (+)	0.58(-)
rs10745954	*ASCL1*	A	1.6×10^{-11}(+)	0.08(-)	0.71(+)	0.62(-)	0.79 (+)	0.81(+)
rs1800961	*HNF4A*	C	2.3×10^{-11}(+)	0.27(-)	0.15(+)	0.25(-)	2.7×10^{-4} (+)	0.54(+)
rs340029	*RORA*	T	2.6×10^{-11}(+)	0.61(-)	0.67(-)	0.59(+)	0.10(+)	0.85(+)
rs10521222	*SALL1*	C	1.3×10^{-10}(+)	0.58(+)	0.83(+)	0.40(+)	0.68(-)	0.36(-)
rs12037222	*PABPC4*	A	4.5×10^{-10}(+)	0.86(-)	0.03(+)	0.10(+)	4.9×10^{-3} (+)	3.7×10^{-3}(+)
rs4705952	*IRF1*	G	1.3×10^{-8}(+)	0.26(-)	1.7×10^{-4}(-)	0.14(-)	0.68(+)	0.08(-)
rs2847281	*PTPN2*	A	2.2×10^{-8}(+)	0.43(-)	0.02(-)	0.90(-)	0.29(+)	0.16(+)
rs13233571	*BCL7B*	C	2.8×10^{-8}(+)	0.08(-)	0.09(-)	0.02(-)	8.3×10^{-4} (+)	0.49(+)
rs6901250	*GPRC6A*	A	4.8×10^{-8}(+)	0.85(+)	0.30(+)	0.06(-)	0.15(-)	0.47(-)

*A1 represents the risk allele according to the CRP GWAS.

Note: p-value≤1.1×10^{-4} is considered as study-wide significant (0.05/463).

Abbreviations: CRP, C-reactive protein; CAD, coronary artery disease; FG, fasting glucose; FI, fasting insulin; SBP, systolic blood pressure; SNP, single-nucleotide polymorphism; T2D, type 2 diabetes.

Table S3. Pleiotropic SNPs and their association with CRP stepwise adjusted for cardiometabolic phenotypes.

		Model 1[a]			*Model 2*[b]			*Model 3*[c]			*Model 4*[d]		
SNP	**Gene**	**beta**	**se**	**P-value**	**beta**	**se**	**P-value**	**beta**	**se**	**P-value**	**beta**	**se**	**P-value**
rs4420638	*APOC1*	0.269	0.019	4.4×10^{-47}	0.310	0.017	1.3×10^{-72}	0.276	0.016	8.2×10^{-68}	0.272	0.016	1.7×10^{-65}
rs1169288	*HNF1A*	0.165	0.012	2.3×10^{-43}	0.164	0.011	1.0×10^{-50}	0.161	0.010	5.9×10^{-57}	0.160	0.010	4.0×10^{-56}
rs1260326	*GCKR*	0.110	0.011	1.6×10^{-22}	0.052	0.010	7.2×10^{-07}	0.073	0.010	3.5×10^{-14}	0.073	0.010	3.4×10^{-14}
rs4845625	*IL6R*	0.067	0.011	2.0×10^{-9}	0.065	0.010	1.9×10^{-10}	0.064	0.009	1.7×10^{-11}	0.065	0.009	8.8×10^{-12}
rs9987289	*PPP1R3B*	0.076	0.019	4.5×10^{-5}	0.099	0.018	2.0×10^{-08}	0.085	0.016	2.3×10^{-07}	0.086	0.016	1.5×10^{-7}
rs1800961	*HNF4A*	0.146	0.033	8.4×10^{-6}	0.154	0.030	3.1×10^{-07}	0.143	0.028	3.1×10^{-07}	0.141	0.028	4.7×10^{-7}
rs4660293	*PABPC4*	0.048	0.013	1.9×10^{-4}	0.031	0.012	1.0×10^{-02}	0.037	0.011	8.2×10^{-04}	0.036	0.011	1.2×10^{-3}
rs17145738	*BCL7B*	0.075	0.017	1.3×10^{-5}	0.012	0.016	4.3×10^{-01}	0.020	0.015	1.7×10^{-01}	0.019	0.015	1.8×10^{-1}
rs1558902	*FTO*	0.041	0.012	6.0×10^{-4}	0.034	0.011	1.5×10^{-03}	0.011	0.010	2.7×10^{-01}	0.012	0.010	2.3×10^{-1}
rs7561317	*TMEM18*	0.055	0.015	1.5×10^{-4}	0.044	0.013	9.4×10^{-04}	0.014	0.012	2.7×10^{-01}	0.013	0.012	2.9×10^{-1}
rs6065906	*PLTP*	0.026	0.014	6.6×10^{-2}	0.061	0.013	2.0×10^{-06}	0.038	0.012	1.7×10^{-03}	0.039	0.012	1.2×10^{-3}
rs571312	*MC4R*	0.038	0.013	3.5×10^{-3}	0.024	0.012	4.6×10^{-02}	0.004	0.011	6.9×10^{-01}	0.006	0.011	6.0×10^{-1}
rs6734238	*IL1F10*	0.040	0.011	3.9×10^{-4}	0.047	0.010	5.0×10^{-06}	0.051	0.010	1.0×10^{-07}	0.051	0.010	1.3×10^{-7}

[a] Model 1: adjusted for age

[b] Model 2: adjusted for age and lipids (HDL-cholesterol, LDL-cholesterol, triglycerides and total cholesterol)

[c] Model 3: adjusted for age, lipids and BMI

[d] Model 4: adjusted for age, lipids, BMI and HbA1C

Table S4. Pleiotropic SNPs and their association with cardiometabolic phenotype.

		Age adjusted			*Age + CRP adjusted*				
SNP	**chr**	**beta**	**se**	**P-value**	**beta**	**se**	**P-value**	**Gene**	**Phenotype**
rs4420638	19	7.657	0.642	9.5×10^{-33}	9.047	0.637	1.4×10^{-45}	*APOC1*	TC
rs4420638	19	0.053	0.008	2.5×10^{-10}	0.103	0.008	1.4×10^{-39}	*APOC1*	TG
rs4420638	19	-1.545	0.236	6.0×10^{-11}	-2.131	0.233	7.6×10^{-20}	*APOC1*	HDLC
rs4420638	19	8.287	0.526	1.6×10^{-55}	8.875	0.527	3.4×10^{-63}	*APOC1*	LDLC
rs4420638	19	-0.026	0.009	4.2×10^{-03}	-0.006	0.009	5.0×10^{-01}	*APOC1*	HbA1C
rs1169288	12	2.471	0.411	1.9×10^{-09}	3.299	0.409	7.2×10^{-16}	*HNF1A*	TC
rs1169288	12	2.261	0.338	2.4×10^{-11}	2.594	0.339	2.0×10^{-14}	*HNF1A*	LDLC
rs1169288	12	0.003	0.006	6.4×10^{-01}	0.017	0.006	3.5×10^{-03}	*HNF1A*	HbA1C
rs1260326	2	3.595	0.388	2.2×10^{-20}	3.070	0.385	1.7×10^{-15}	*GCKR*	TC
rs1260326	2	0.070	0.005	8.5×10^{-43}	0.050	0.005	7.5×10^{-27}	*GCKR*	TG
rs1260326	2	-0.013	0.006	1.7×10^{-02}	-0.023	0.005	2.5×10^{-05}	*GCKR*	HbA1C
rs9987289	8	-2.773	0.664	3.0×10^{-05}	-2.405	0.658	2.6×10^{-04}	*PPP1R3B*	TC
rs9987289	8	-1.085	0.244	8.5×10^{-06}	-1.246	0.240	2.2×10^{-07}	*PPP1R3B*	HDLC
rs9987289	8	-2.210	0.546	5.2×10^{-05}	-2.068	0.545	1.5×10^{-04}	*PPP1R3B*	LDLC
rs9987289	8	0.015	0.010	1.1×10^{-01}	0.021	0.009	2.3×10^{-02}	*PPP1R3B*	HbA1C
rs1800961	20	-4.927	1.130	1.3×10^{-05}	-4.217	1.119	1.6×10^{-04}	*HNF4A*	TC
rs1800961	20	-2.163	0.415	1.8×10^{-07}	-2.474	0.409	1.5×10^{-09}	*HNF4A*	HDLC
rs1800961	20	-2.960	0.929	1.4×10^{-03}	-2.685	0.928	3.8×10^{-03}	*HNF4A*	LDLC
rs1800961	20	0.022	0.016	1.7×10^{-01}	0.030	0.016	6.3×10^{-02}	*HNF4A*	HbA1C
rs4660293	1	-0.575	0.164	4.6×10^{-04}	-0.473	0.162	3.5×10^{-03}	*PABPC4*	HDLC
rs4660293	1	0.020	0.006	7.5×10^{-04}	0.011	0.005	3.9×10^{-02}	*PABPC4*	TG
rs17145738	7	0.458	0.218	3.5×10^{-02}	0.299	0.215	1.6×10^{-01}	*BCL7B*	HDLC
rs17145738	7	-0.074	0.008	3.7×10^{-21}	-0.060	0.007	4.6×10^{-17}	*BCL7B*	TG
rs1558902	16	0.541	0.049	5.0×10^{-28}	0.470	0.044	9.7×10^{-27}	*FTO*	BMI
rs1558902	16	0.028	0.006	2.2×10^{-06}	0.025	0.006	1.9×10^{-05}	*FTO*	HbA1C
rs7561317	2	-0.344	0.060	1.3×10^{-08}	-0.243	0.054	6.6×10^{-06}	*TMEM18*	BMI
rs6065906	20	-1.154	0.178	9.0×10^{-11}	-1.209	0.175	5.7×10^{-12}	*PLTP*	HDLC
rs6065906	20	0.039	0.006	1.4×10^{-09}	0.043	0.006	1.5×10^{-13}	*PLTP*	TG
rs571312	18	0.281	0.055	2.8×10^{-07}	0.209	0.049	1.7×10^{-05}	*MC4R*	BMI
rs6734238	2	-1.164	0.388	2.7×10^{-03}	-1.358	0.385	4.1×10^{-04}	*IL1F10*	TC

Abbreviations: BMI, body mass index; chr, chromosome; CRP, C-reactive protein; HbA1C, haemoglobin A1C; HDLC, HDL-cholesterol; LDLC, LDL-cholesterol; se, standard error; SNP, single-nucleotide polymorphism; TC, total cholesterol; TG, triglycerides

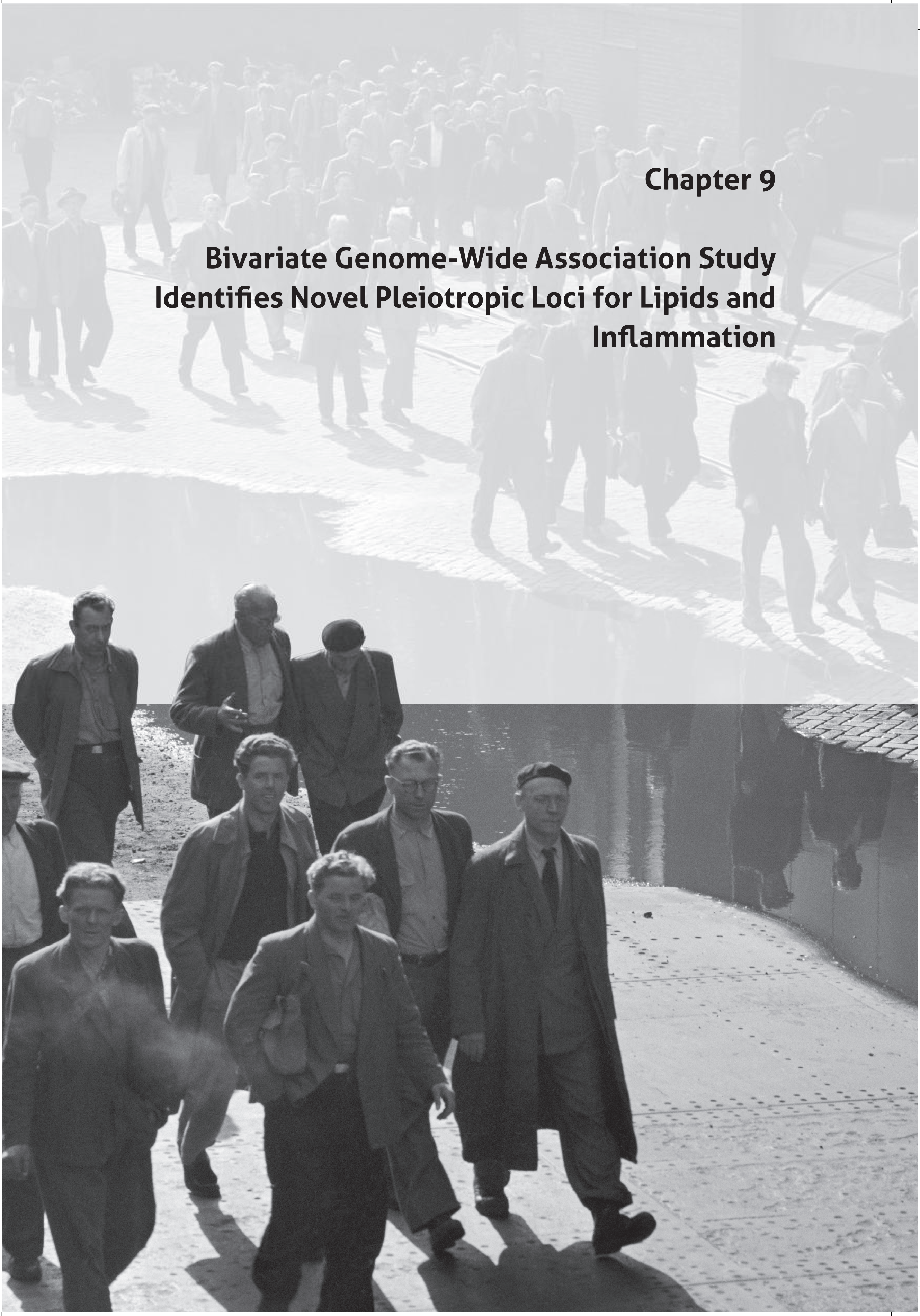
Table S5. Pathway analysis results from the 13 pleiotropic genes.

Caninocal Pathway	**p-value**[a]
FXR/RXR Activation	7.4×10^{-9}
LXR/RXR Activation	4.6×10^{-5}
Maturity Onset Diabetes of the Young (MODY) signaling	7.6×10^{-5}
Hepatic Cholestasis	1.1×10^{-4}
Acute Phase Response signaling	1.3×10^{-4}
LPS/IL-1 Mediated Inhibition of RXR function	2.6×10^{-4}
Role of Macrophages, Ficorblasts and Endothelial Cells in Rheumatoid Arthritis	6.6×10^{-4}
IL-6 signaling	2.1×10^{-3}
Atherosclerosis Signaling	2.4×10^{-3}
Acyl-CoA Hydrolysis	7.2×10^{-3}
Role of Osetoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	7.4×10^{-3}
Systemic Lupus Erythematosus Signaling	7.5×10^{-3}
Colorectal Cancer Metastasis Signaling	8.5×10^{-3}

[a]Significant at False Discovery Rate of 5 percent.

Chapter 9

Bivariate Genome-Wide Association Study Identifies Novel Pleiotropic Loci for Lipids and Inflammation



Background: Genome-wide association studies (GWAS) have identified multiple genetic loci for C-reactive protein (CRP) and lipids, of which some overlap. We aimed to identify genetic pleiotropy among CRP and lipids in order to better understand the shared biology of chronic inflammation and lipid metabolism.

Methods: In a bivariate GWAS, we combined summary statistics of published GWAS on CRP (n=66,185) and lipids, including LDL-cholesterol, HDL-cholesterol, triglycerides, and total cholesterol (n=100,184), using an empirical weighted linear-combined test statistic. We sought replication for novel CRP associations in an independent sample of 17,743 genotyped individuals, and performed *in silico* replication of novel lipid variants in 93,982 individuals.

Results: Fifty potentially pleiotropic SNPs were identified among CRP and lipids: 21 for LDL-cholesterol and CRP, 20 for HDL-cholesterol and CRP, 21 for triglycerides, and CRP and 20 for total cholesterol and CRP. We identified and significantly replicated three novel SNPs for CRP in or near *CTSB/FDFT1* (rs10435719, P-value$_{replication}$: 2.6×10^{-5}), *STAG1/PCCB* (rs7621025, P-value$_{replication}$: 1.4×10^{-3}) and *FTO* (rs1558902, P-value$_{replication}$: 2.7×10^{-5}). Seven pleiotropic lipid loci were replicated in the independent set of MetaboChip samples of the Global Lipids Genetics Consortium. Annotating the effect of replicated CRP SNPs to the expression of nearby genes, we observed an effect of rs10435719 on gene expression of *FDFT1*, and an effect of rs7621025 on *PCCB*.

Conclusion: Our large scale combined GWAS analysis identified numerous pleiotropic loci for CRP and lipids providing further insight in the genetic interrelation between lipids and inflammation. In addition, we provide evidence for *FDFT1, PCCB* and *FTO* to be associated with CRP levels.

Introduction

Genome-wide association studies (GWAS) have identified hundreds of genetic loci for cardiovascular disease and it's risk factors, including chronic inflammation and lipids[1,2,3]. Some of the identified genetic variants are associated with more than one phenotype, termed genetic pleiotropy[4]. Examples are *APOC1(rs4420638)* and *HNF1A (rs1183910),* which are associated both with lipids and C-reactive protein (CRP)[2,3]. As randomized clinical trials have shown a coextending effect of statin treatment on the lowering of LDL-cholesterol and CRP, we do expect inflammation and lipids to share certain biological pathways[5,6]. Moreover, there is accumulating evidence that the pleiotropic effects are partially independent, although the biological mechanisms are not fully understood[7]. The identification of further pleiotropic genes could provide insight into the biological mechanisms that link chronic inflammation to lipids.

Therefore, we aimed to identify further shared genes for lipids and CRP. In order to enhance the statistical power of genetic studies to find pleiotropic genes for the correlated phenotypes of interest, we applied a method that combines GWAS meta-analysis summary statistics allowing for mixed directions of effect, a common observed phenomenon in genetic pleitropy[8]. In a second step we sought to replicate novel associations with lipids and CRP in an independent sample of 93,982 genotyped individuals for lipids and 17,743 genotyped individuals for CRP. We identified multiple overlapping genetic variants between CRP and lipids and confirmed novel genes implicated in the biology of chronic inflammation.

Methods

The present study includes three stages. First, we performed a bivariate GWAS combining published GWAS data on CRP and lipids to identify pleiotropic variants for CRP and lipids. In a second step, we sought replication of novel associations in independent samples of genotyped individuals. Finally, we carried out functional analyses in a third step to point out potential underlying transcriptional mechanisms.

We used the data from the largest published GWAS on CRP as well as the publicly available GWAS on lipids from GLGC to explore the genetic pleiotropy between inflammation and lipids[2,3]. We combined summary association test statistics from the CRP GWAS separately with the GWAS on HDL-cholesterol, LDL-cholesterol, triglycerides and total cholesterol. The CRP GWAS meta-analysis included 65,000 individuals from 15 different studies in the discovery panel and after replication, 18 loci were genome-wide significantly associated with serum CRP level[3]. The lipids GWAS comprised 100,184 individuals for total cholesterol, 95,454 for LDL-cholesterol, 99,900 for HDL-cholesterol and 96,598 for triglycerides across 46 studies. The lipid GWAS identified a total of 95 lipid loci (52 for total cholesterol, 37 for

LDL-cholesterol, 47 for HDL-cholesterol and 32 for triglycerides)[2]. The CRP and lipids GWAS used HapMap imputed data (build 36). All studies that contributed genotype data to the CRP GWAS also contributed data to the lipids GWAS. We ensured that effect alleles were harmonized across the two GWAS before applying the bivariate GWAS method. Overall, 2,501,549 common Single Nucleotide Polymorphisms (SNPs) were tested for their association with CRP and total cholesterol, 2,501,711 with CRP and triglycerides, 2,501,543 with CRP and HDL-cholesterol and 2,501,749 with CRP and LDL-cholesterol. An aggregated p-value was calculated using the method described below.

Bivariate Genome-Wide Association Study

To better understand the shared biology of CRP and lipids by further identifying shared genes between CRP and lipids, we aimed to increase power by combining the summary statistics from the CRP and lipid GWAS. We chose to use a recently introduced method that performs bivariate GWAS allowing for mixed directions of effect. The method combines summary statistics (Z test statistics) from univariate GWAS of CRP pairing with the summary statistics of each univariate GWAS meta-analysis of lipid phenotypes, using an empirical-weighted linear-combined test statistics (eLC), implemented in a C++ eLX package. We have recently used this method in the identification of pleiotropic genes for menopause and menarche and the details of the method are presented elsewhere[8,9]. eLC allows having opposite direction of effect on the combined phenotypes, which is common between CRP and cholesterol phenotypes[2,3]. Briefly, eLC directly combines correlated Z test statistics (calculated as *β*/*SE* derived from the original GWAS) obtained from univariate GWAS meta-analyses with a weighted sum of univariate test statistics to empirically maximize the overall association signals and also to account for the phenotypical correlations among CRP and lipids. Our eLC approach is expressed as

$$S_{eLC} = \sum_{1}^{k} [\max(|T_k|, c)^* \, |T_k|]$$

where T_k is a matrix of K statistics for K phenotypes (for bivariate, K is equal to 2) and c is a given non-negative constant. The optimal weighting is estimated empirically using the Monte Carlo Simulation[10] and the bona-fide p-values for eLC test statistics are calculated through permutation. The sample covariance matrix of the test statistics of all SNPs from the univariate GWAS analyses is used as an approximation of the variance-covariance matrix **Σ** of univariate test statistics. **Σ**:

$$\begin{bmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) \\ \mathrm{Cov}(Z_1, Z_2) & \mathrm{Var}(Z_2) \end{bmatrix}$$

where Z_1 and Z_2 consist of unbiased univariate test statistics of all the SNPs for the two traits on genome-wide scale for the first (Z_1) and second (Z_2) trait. The null hypothesis in the bivariate analysis is β_1=0 AND β_2=0; the H1 is β_1 not equal to 0 or β_2 not equal to 0.

The results were considered genome-wide significant when (1) the bivariate p-values were $< 5\times10^{-8}$ and (2) the bivariate p-value was at least one order of magnitude lower than both individual trait p-values and (3) when the individual trait p-values were at least nominally significant (p-value<0.05). When multiple SNPs were significant in a locus, the SNP with the lowest p-value was chosen for replication. The eLC method is implemented in eLX package using C++ (see Weblinks).

Replication Study

The bivariate GWAS resulted in three possible scenarios. First, the pleiotropic variant or the locus harboring the pleiotropic variant (defined as ±500MB of the pleiotropic SNP) was genome-wide significant in both the primary univariate GWAS of CRP and the lipid trait. Second, the pleiotropic signal was significant in either the CRP or the lipid univariate GWAS. Third, the pleiotropic signal was neither genome-wide significant in the CRP nor in the lipid GWAS. Per definition, a variant is considered pleiotropic when there is robust evidence for an association with two or more phenotypes. Therefore, we only selected the variants that were not genome-wide significant in the primary univariate GWAS for replication in an independent sample of genotyped samples. We intended to replicate the novel associations with CRP levels in three cohort studies that did not contribute to the original CRP GWAS. The independent cohorts were the second (n=1,943) and third (n=2,962) cohort of the Rotterdam Study and the LifeLines cohort study (n=12,838; supplementary method) [11,12]. The total sample size for the replication of potentially novel CRP variants comprised 17,743 individuals. In an attempt to replicate the potential novel lipid variants, we performed an in silico replication in the publicly available association results from the participants of the GLGC that did not contribute to the original lipids GWAS we used for the pleiotropy analysis. This replication set comprises 93,982 individuals genotyped using the Metabochip array[13, 14]. For the SNPs that were not available on the Metabochip, we selected the best available proxy SNP on the Metabochip for replication (r^2>0.5). We used a Bonferroni corrected p-value of 0.05 divided by the number of SNPs tested for replication as a threshold of significance in the replication study.

Expression Quantitative Trait Loci (eQTL)

In an attempt to annotate the pleiotropic variants to a pleiotropic gene, we searched in tissues related to lipids and inflammation for eQTL effects of the pleiotropic variants or reasonable proxy variants (r^2>0.80).

The eQTL analyses in whole blood comprised 5,311 individuals from seven studies in the discovery setting with both genetic and gene expression data available[15]. The discovery meta-analysis including the seven studies (EGCUT, InCHIANTI, Rotterdam Study, Fehrmann, HVH, SHIP-TREND and DILGOM). Results are publicly available (access URL:

http://genenetwork.nl/bloodeqtlbrowser/). eQTLs were deemed cis when the distance between the SNP and the midpoint of the RNA probe was <250kb. We only considered a significant eQTL effect of the pleiotropic SNP when the p-value exceeded the FDR corrected threshold for multiple testing.

We searched for liver eQTL effects by use of the eQTL browser provided by the university of Chicago (access URL: http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/). The liver tissue dataset by Schadt et al. comprised 427 individuals from European ancestry with liver specific gene expression and genotyping data available[16]. An eQTL was deemed cis when the SNP was within 1Mb of the annotated start or stop site of the corresponding structural gene. The authors used an FDR correction of 10% for a significant association. The dataset by Innocenti et al. comprised 266 individuals from 2 different studies. Cis eQTL was defined as <250kb from the gene transcription start site and the FDR for significant association was set to 5%[17].

We used the GTEx adipose tissue dataset (access URL: http://www.gtexportal.org/home/eqtls/tissue?tissueName=Adipose_Subcutaneous) to search for potential eQTLs in adipose tissue. The dataset consisted of 111 individuals with both gene expression and genotype data available[18] Cis radius was defined as +/- 1mb from transcription start site. An eQTL was deemed significant when the FDR q-value<=5%.

Results

Bivariate Genome-Wide Association Analysis

Manhattan plots for the bivariate GWAS are depicted in Figure 1. Table 1 indicates the results from the bivariate analysis combining CRP and LDL-cholesterol genetic association data. The bivariate analysis resulted in 21 potentially pleiotropic loci. We identified fourteen loci associated with CRP levels which had no genome-wide significant SNP in the original GWAS of CRP. These potential novel associations were located in or near *CELSR2*, *IRF2BP2*, *ABCG8*, *GCNT4*, *HLA-DQB1*, *FRK*, *TRIB1*, *FADS2*, *ST3GAL4*, *BRAP*, *C12orf51*, *CARM1/LDLR*, *NCAN* and *RASIP1*. The potential novel associations for LDL-cholesterol were located in or near *GCKR*, *IL1F10*, *RORA*, *RASIP1* and in *HNF4A*. The SNPs identified in the bivariate GWAS near *HLA-DQB1*, *FRK*, *BRAP*, *c12orf51* and *CARM1/LDLR* were not genome-wide significant in the original univariate GWAS on LDL-cholesterol, however other SNPs in their vicinity were significant in the original GWAS on LDL-cholesterol and the loci have thus been reported previously. The variants in and near *PPP1R3B, HNF1A* and *APOC1* were already genome-wide significant in both GWAS of CRP and LDL-cholesterol.

Figure 1. Manhattan plots of the bivariate genome-wide association studies combining C-reactive protein with LDL-cholesterol, HDL-cholesterol, triglycerides, and total cholesterol.

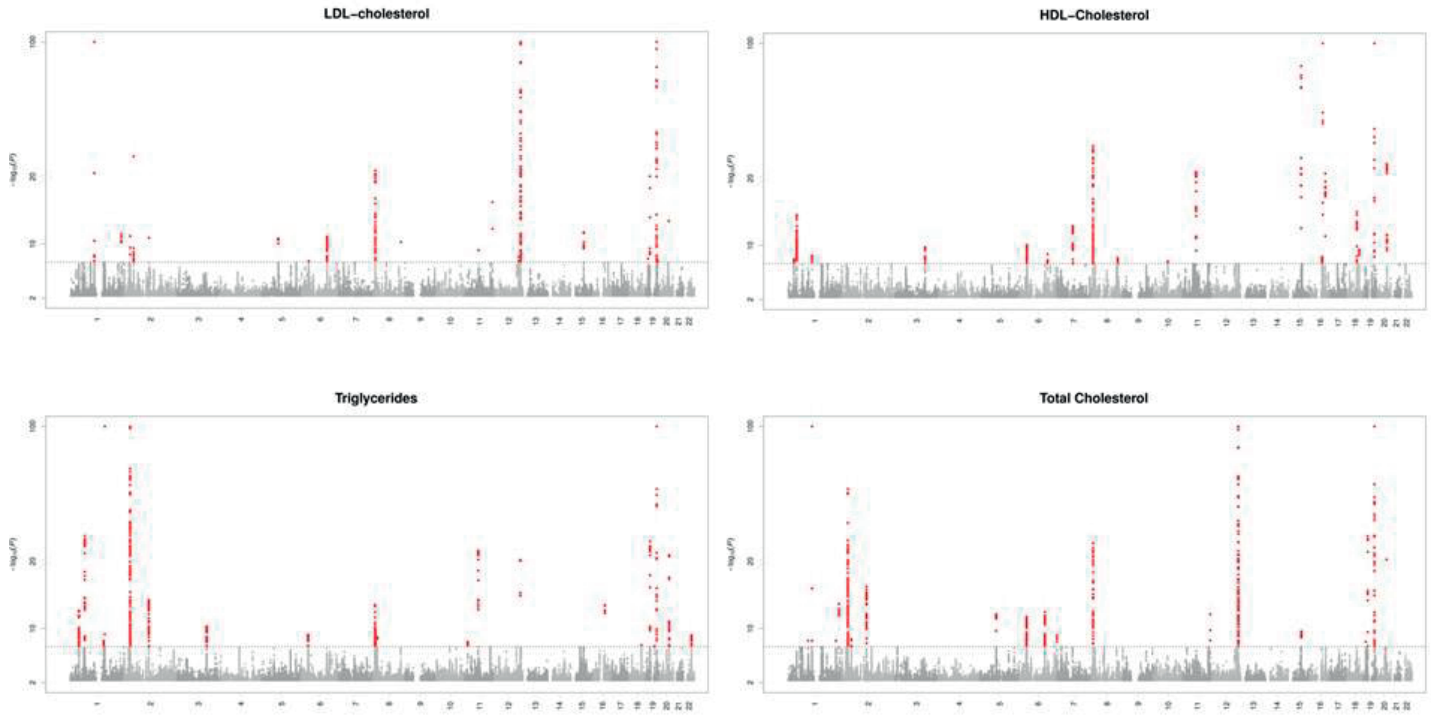


Table 1. Results of Bivariate GWAS for C-Reactive Protein and LDL-Cholesterol Levels.

SNP	Chr	Position	Effect Allele	C-reactive protein		LDL-cholesterol		Pleiotropy significance	Gene
				Beta	P-value	Beta	P-value		
rs646776	1	109620053	T	-0.018	0.02	0.171	4.5×10^{-169}	4.3×10^{-170}	*CELSR2*
rs661955	1	232909479	C	-0.021	1.7×10^{-3}	0.034	1.2×10^{-10}	3.2×10^{-12}	*IRF2BP2*
rs3817588	2	27584716	T	0.053	1.8×10^{-10}	0.024	4.2×10^{-4}	6.4×10^{-12}	*GCKR*
rs11887534	2	43919751	C	-0.049	2.5×10^{-4}	-0.134	1.1×10^{-31}	9.0×10^{-33}	*ABCG8*
rs12711751	2	113554236	T	-0.044	1.6×10^{-10}	0.014	4.8×10^{-3}	1.2×10^{-11}	*IL1F10*
rs4703642	5	74297918	A	0.018	3.0×10^{-3}	-0.031	3.1×10^{-10}	1.5×10^{-11}	*GCNT4*
rs9275292	6	32771267	A	0.022	3.6×10^{-4}	0.023	1.1×10^{-5}	3.3×10^{-8}	*HLA-DQB1*
rs3822857	6	116420624	C	-0.032	2.7×10^{-6}	-0.030	2.3×10^{-7}	7.6×10^{-12}	*FRK*
rs9987289	8	9220768	A	-0.079	2.1×10^{-12}	-0.071	2.0×10^{-14}	2.3×10^{-24}	*PPP1R3B*
rs8180991	8	126569532	C	-0.026	9.0×10^{-4}	-0.041	8.0×10^{-10}	5.1×10^{-11}	*TRIB1*
rs174574	11	61356918	A	-0.027	1.7×10^{-3}	-0.050	1.1×10^{-8}	7.8×10^{-10}	*FADS2*
rs11220463	11	125753421	A	0.032	2.8×10^{-3}	-0.070	1.3×10^{-15}	5.8×10^{-17}	*ST3GAL4*
rs10744775	12	110580598	T	0.021	4.0×10^{-3}	-0.030	5.3×10^{-7}	3.1×10^{-8}	*BRAP*
rs2285810	12	111183923	T	0.019	6.8×10^{-3}	-0.030	8.3×10^{-8}	8.3×10^{-9}	*C12orf51*
rs1183910	12	119905190	A	-0.151	4.6×10^{-113}	0.042	5.8×10^{-15}	5.6×10^{-128}	*HNF1A*
rs340005	15	58665322	A	0.044	3.2×10^{-11}	-0.015	3.4×10^{-3}	1.7×10^{-12}	*RORA*
rs1529711	19	10884434	T	0.030	8.4×10^{-4}	0.037	1.5×10^{-6}	1.5×10^{-8}	*CARM1/LDLR*
rs2228603	19	19190924	T	0.036	2.9×10^{-3}	-0.089	1.4×10^{-19}	6.5×10^{-21}	*NCAN*
rs4420638	19	50114786	A	0.240	1.0×10^{-129}	-0.215	8.7×10^{-147}	1.2×10^{-283}	*APOC1*
rs2287921	19	53920084	T	-0.019	3.6×10^{-3}	-0.026	3.4×10^{-7}	2.8×10^{-8}	*RASIP1*
rs1800961	20	42475778	T	-0.120	2.4×10^{-11}	-0.070	2.4×10^{-5}	3.8×10^{-14}	*HNF4A*

For both CRP and the lipid phenotype, the effect estimates are according to the original GWAS. Chromosome and position are in NCBI genome build 36.

Beta coefficient for CRP represents 1-unit change in the natural log–transformed CRP (mg/L) per copy increment in the coded allele.

Beta coefficient for LDL-cholesterol represents 1-unit change in the standardized LDL-cholesterol levels per copy increment in the coded allele.

Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphism.

We identified 20 potential pleiotropic SNPs (Table 2). The variants near *CELSR2, STAG1, HLA-DRA, JMJD1C, FADS1, LIPC, CETP, LYPLA3, LIPG* and *MC4R* were not genome-wide significant in the original CRP meta-GWAS analysis. Seven SNPs were potentially novel for both CRP and HDL-cholesterol: the SNP rs12742376 located in *C1orf172* on chromosome 1 ($P_{\text{bivariate}}$ = 1.4×10^{-8}) , rs7621025 in *STAG1* on chromosome 3 ($P_{bivariate}$=1.2×10^{-9}), rs9378212 near *HLA-DRA* ($P_{bivariate}$=6.7×10^{-10}), rs10761731 in *JMJD1C* ($P_{bivariate}$=2.2×10^{-8}), rs1936797 in *RSPO3* on chromosome 6 ($P_{\text{bivariate}}$=6.7×10^{-9}), rs4871137 near *SNTB1* ($P_{\text{bivariate}}$=3.3×10^{-8}) on chromosome 8 and the *FTO* SNP rs1558902 ($P_{\text{bivariate}}$=5.0×10^{-9}) on chromosome 16. The variants near *CELSR2* and *PLTP* were not significant in the original GWAS on HDL-cholesterol, but these loci were identified in the original GWAS. The variants in or near *PABPC4, BAZ1B, PPP1R3B, APOC1* and *HNF4A* were already genome-wide significant in both the CRP and HDL-cholesterol univariate GWAS.

Table 3 lists the 21 potentially pleiotropic SNPs that were identified combining the GWAS results of triglycerides and CRP. For triglycerides, we identified eleven potential novel associations compared to the original GWAS located in or near *PABPC4, LEPR, ADAR, CRP, IL1F10, PPP1R3B, CTSB/FDFT1, ARNTL, CABP1, MC4R* and *HPN*. The variant near *PLA2G6* was not genome-wide significant in the original GWAS, but this locus was identified in the original GWAS. The variants in and near *ADAR, MSL2L1, HLA-C, CTSB/FDFT1, LPL, ARNTL, FADS1, CETP, MC4R, SF4, HPN, ZNF335/PLTP* and *PLA2G6* were potential novel associations with CRP level. Five loci were not genome-wide significant in either the original GWAS on CRP or triglycerides: the SNP rs1127311 within *ADAR* on chromosome 1 ($P_{bivariate}$=6.4×10^{-9}), rs10435719 located 77Kb upstream of *CTSB* on chromosome 8 ($P_{bivariate}$=2.0×10^{-10}), rs10832027 located in the second intron of *ARNTL* on chromosome 11 ($P_{bivariate}$=9.4×10^{-9}), rs571312 on chromosome 18 near *MC4R* ($P_{bivariate}$=2.8×10^{-8})*,* and the chromosome 19 rs1688043 in the fifth intron of *HPN* ($P_{bivariate}$=4.1×10^{-8}). In both the original GWAS of CRP and triglycerides, *GCKR* and *APOC1* were already genome-wide significant.

Twenty potentially pleiotropic SNPs were identified combining CRP and total cholesterol (Table 4). The SNPs in or near *ZNF644*, *SLC44A4*, *C7orf50* and *RORA* were potentially novel for total cholesterol. The variants near *HLX, ABCG5, IL1F10, C7orf60* and *CARM1* were not genome-wide significant in the GWAS on total cholesterol, but the loci were identified in this original GWAS. For CRP, *ZNF664*, *CELSR2*, *HLX*, *IRF2BP2*, *ABCG5*, *GCNT4*, *SLC44A4*, *HLA-DQB1*, *FRK*, *ST3GAL4*, *CARM1* and *NCAN* were potentially novel compared to the univariate GWAS. The SNPs near *ZNF644* and *C7orf50* were novel pleiotropic loci for both CRP and total cholesterol.

Table 2. Results of Bivariate GWAS Analyses for C-Reactive Protein and HDL-Cholesterol Levels.

SNP	Chr	Position	Effect	C-reactive protein		HDL-cholesterol		Pleiotropy	Gene
			Allele	Beta	P-value	Beta	P-value	significance	
rs12742376	1	27157782	T	-0.027	1.7×10^{-2}	-0.046	2.8×10^{-7}	1.4×10^{-8}	*C1orf172*
rs4660293	1	39800767	A	-0.044	1.2×10^{-9}	0.034	4.0×10^{-10}	3.1×10^{-15}	*PABPC4*
rs646776	1	109620053	T	-0.018	1.8×10^{-2}	-0.033	6.4×10^{-8}	3.2×10^{-9}	*CELSR2*
rs7621025	3	137754936	T	0.028	1.7×10^{-4}	0.026	4.1×10^{-6}	1.2×10^{-9}	*STAG1*
rs9378212	6	32553669	T	0.027	4.9×10^{-5}	0.021	8.1×10^{-6}	6.7×10^{-10}	*HLA-DRA*
rs1936797	6	127474350	A	0.022	2.8×10^{-3}	0.022	9.9×10^{-7}	6.7×10^{-9}	*RSPO3*
rs13244268	7	72549779	T	0.054	2.6×10^{-8}	-0.045	1.3×10^{-9}	1.2×10^{-13}	*BAZ1B*
rs9987289	8	9220768	A	-0.079	2.1×10^{-12}	-0.083	6.4×10^{-25}	1.2×10^{-39}	*PPP1R3B*
rs4871137	8	121937732	T	-0.021	2.2×10^{-3}	-0.026	5.6×10^{-6}	3.3×10^{-8}	*SNTB1*
rs10761731	10	64697616	A	0.023	2.7×10^{-4}	-0.025	2.5×10^{-7}	2.2×10^{-8}	*JMJD1C*
rs174546	11	61326406	T	-0.017	1.2×10^{-2}	-0.048	2.6×10^{-22}	1.6×10^{-24}	*FADS1*
rs1077834	15	56510771	T	-0.016	4.0×10^{-2}	-0.114	9.6×10^{-84}	2.5×10^{-87}	*LIPC*
rs1558902	16	52361075	A	0.032	2.0×10^{-6}	-0.021	4.6×10^{-6}	5.0×10^{-9}	*FTO*
rs711752	16	55553712	A	0.016	1.8×10^{-2}	0.192	2.1×10^{-297}	4.3×10^{-308}	*CETP*
rs17688076	16	66843928	A	0.019	4.9×10^{-2}	0.070	3.9×10^{-22}	1.8×10^{-23}	*LYPLA3*
rs11874381	18	45457406	A	0.013	4.9×10^{-2}	0.038	1.2×10^{-14}	1.0×10^{-15}	*LIPG*
rs12967135	18	56000003	A	0.029	1.2×10^{-4}	-0.036	6.6×10^{-9}	4.3×10^{-10}	*MC4R*
rs4420638	19	50114786	A	0.240	1.0×10^{-129}	0.071	4.4×10^{-21}	2×10^{-164}	*APOC1*
rs1800961	20	42475778	T	-0.120	2.4×10^{-11}	-0.129	1.1×10^{-15}	3.9×10^{-28}	*HNF4A*
rs6065906	20	43987422	T	0.036	5.9×10^{-6}	0.058	1.9×10^{-22}	5.1×10^{-29}	*PLTP*

For both CRP and the lipid phenotype, the effect estimates are according to the original GWAS. Chromosome and position are in NCBI genome build 36.

β coefficient for CRP represents 1-unit change in the natural log–transformed CRP (mg/L) per copy increment in the coded allele.

Beta coefficient for HDL-cholesterol represents 1-unit change in the standardized HDL-cholesterol levels per copy increment in the coded allele.

Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphism.

Table 3. Results of Bivariate GWAS Analyses for C-Reactive Protein and Triglycerides Levels.

SNP	Chr	Position	Effect Allele	C-reactive protein		Triglycerides		Pleiotropy significance	Gene
				Beta	P-value	Beta	P-value		
rs4660808	1	39791096	T	0.046	8.6×10^{-10}	0.028	3.1×10^{-7}	2.2×10^{-13}	*PABPC4*
rs11208722	1	65943589	A	-0.083	1.2×10^{-32}	0.012	0.02	8.1×10^{-36}	*LEPR*
rs1127311	1	152823287	A	-0.031	9.3×10^{-7}	0.012	5.5×10^{-3}	6.4×10^{-9}	*ADAR*
rs12755606	1	157936960	C	-0.153	3.0×10^{-112}	0.012	0.01	4.0×10^{-120}	*CRP*
rs1260326	2	27584444	T	0.089	1.7×10^{-42}	0.116	5.7×10^{-133}	4.4×10^{-151}	*GCKR*
rs13409360	2	113554573	A	0.048	1.3×10^{-12}	-0.013	8.6×10^{-3}	5.3×10^{-15}	*IL1F10*
rs645040	3	137409312	T	-0.023	2.5×10^{-3}	0.030	2.5×10^{-8}	4.6×10^{-11}	*MSL2L1*
rs2524163	6	31367558	T	0.025	1.5×10^{-4}	0.027	1.7×10^{-8}	7.9×10^{-10}	*HLA-C*
rs9987289	8	9220768	A	-0.079	2.1×10^{-12}	0.020	0.02	2.9×10^{-14}	*PPP1R3B*
rs10435719	8	11814313	T	0.026	7.6×10^{-5}	-0.022	4.1×10^{-6}	2.0×10^{-10}	*CTSB*
rs1441759	8	19909843	C	0.11	3.3×10^{-4}	0.125	2.1×10^{-8}	2.0×10^{-9}	*LPL*
rs10832027	11	13313759	A	0.032	8.5×10^{-7}	0.020	1.1×10^{-4}	9.4×10^{-9}	*ARNTL*
rs174546	11	61326406	T	-0.017	0.01	0.048	5.4×10^{-24}	5.2×10^{-27}	*FADS1*
rs2686555	12	119579555	A	-0.059	1.7×10^{-19}	0.010	0.03	1.6×10^{-21}	*CABP1*
rs11508026	16	55556829	T	0.014	0.03	-0.038	1.3×10^{-12}	3.1×10^{-14}	*CETP*
rs571312	18	55990749	A	0.033	3.5×10^{-5}	0.026	1.2×10^{-5}	2.8×10^{-8}	*MC4R*
rs10401969	19	19268718	T	-0.031	0.02	0.112	1.6×10^{-29}	1.6×10^{-32}	*SF4*
rs1688043	19	40245181	T	-0.038	2.4×10^{-3}	0.037	1.2×10^{-5}	4.1×10^{-8}	*HPN*
rs4420638	19	50114786	A	0.24	1.0×10^{-129}	-0.068	5.4×10^{-22}	1.7×10^{-171}	*APOC1*
rs4465830	20	44018827	A	0.036	7.0×10^{-6}	-0.050	2.0×10^{-17}	2.0×10^{-24}	*ZNF335/PLTP*
rs2277844	22	36907461	A	-0.018	5.7×10^{-3}	0.025	1.5×10^{-7}	9.2×10^{-10}	*PLA2G6*

For both CRP and the lipid phenotype, the effect estimates are according to the original GWAS.

Chromosome and position are in NCBI genome build 36.

β coefficient for CRP represents 1-unit change in the natural log–transformed CRP (mg/L) per copy increment in the coded allele.

Beta coefficient for triglycerides represents 1-unit change in the standardized triglyceride levels per copy increment in the coded allele.

Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphism.

Table 4. Results of Bivariate GWAS Analyses for C-Reactive Protein and Total Cholesterol Levels.

SNP	Chr	Position	Effect Allele	C-reactive protein		Total cholesterol		Pleiotropy significance	Gene
				Beta	P-value	Beta	P-value		
rs469772	1	91302893	T	-0.042	1.6×10^{-7}	-0.020	1.5×10^{-3}	1.5×10^{-8}	*ZNF644*
rs629301	1	109619829	T	-0.017	2.8×10^{-2}	0.149	5.8×10^{-131}	5.7×10^{-132}	*CELSR2*
rs17597773	1	219121384	C	0.020	7.5×10^{-3}	-0.031	7.1×10^{-8}	6.6×10^{-9}	*HLX*
rs661955	1	232909479	C	-0.021	1.7×10^{-3}	0.036	1.0×10^{-12}	2.2×10^{-14}	*IRF2BP2*
rs1260326	2	27584444	T	0.089	1.7×10^{-42}	0.055	7.3×10^{-27}	2.6×10^{-63}	*GCKR*
rs4148191	2	43896408	A	-0.050	2.5×10^{-4}	-0.054	1.1×10^{-6}	3.7×10^{-09}	*ABCG5*
rs6734238	2	113557501	A	-0.047	4.8×10^{-13}	0.023	1.2×10^{-5}	5.8×10^{-17}	*IL1F10*
rs4703642	5	74297918	A	0.018	3.0×10^{-3}	-0.033	2.0×10^{-11}	7.3×10^{-13}	*GCNT4*
rs577272	6	31945942	A	0.020	1.1×10^{-3}	0.026	2.3×10^{-7}	1.6×10^{-8}	*SLC44A4*
rs2858310	6	32776301	A	0.026	8.7×10^{-5}	0.033	3.3×10^{-10}	3.8×10^{-12}	*HLA-DQB1*
rs3822857	6	116420624	C	-0.032	2.7×10^{-6}	-0.033	4.7×10^{-9}	2.1×10^{-12}	*FRK*
rs6951245	7	1024719	A	0.03	5.5×10^{-4}	0.037	6.1×10^{-8}	2.6×10^{-9}	*C7orf50*
rs2126259	8	9222556	T	-0.072	5.7×10^{-12}	-0.085	9.0×10^{-24}	1.4×10^{-31}	*PPP1R3B*
rs11220463	11	125753421	A	0.032	2.8×10^{-3}	-0.057	2.1×10^{-11}	7.3×10^{-13}	*ST3GAL4*
rs1183910	12	119905190	A	-0.151	4.6×10^{-113}	0.040	5.2×10^{-14}	8.2×10^{-128}	*HNF1A*
rs340025	15	58695599	T	-0.036	8.3×10^{-9}	0.015	2.4×10^{-3}	2.5×10^{-10}	*RORA*
rs1529711	19	10884434	T	0.030	8.4×10^{-4}	0.038	6.3×10^{-7}	3.4×10^{-8}	*CARM1*
rs2228603	19	19190924	T	0.036	2.9×10^{-3}	-0.118	4.3×10^{-34}	1.1×10^{-35}	*NCAN*
rs4420638	19	50114786	A	0.240	1.0×10^{-129}	-0.184	5.2×10^{-111}	3.8×10^{-249}	*APOC1*
rs1800961	20	42475778	T	-0.120	2.4×10^{-11}	-0.118	5.7×10^{-13}	1.0×10^{-20}	*HNF4A*

For both CRP and the lipid phenotype, the effect estimates are according to the original GWAS.

Chromosome and position are in NCBI genome build 36.

Beta coefficient for total cholesterol represents 1-unit change in the standardized total cholesterol levels per copy increment in the coded allele.

β coefficient for CRP represents 1-unit change in the natural log–transformed CRP (mg/L) per copy increment in the coded allele.

Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphism.

Replication of the Novel Pleiotropic Loci

In total, we sought replication for 36 potential novel SNPs for CRP in 17,743 genotyped individuals from three independent cohort studies. Using a Bonferroni corrected threshold for multiple testing (0.05/36=1.4×10^{-3}), three SNPs remained significantly associated with CRP levels when we performed replication analysis (Supplementary table I). These variants included the SNPs rs10435719 in *CTSB/FDFT1* ($P_{replication}$=2.6×10^{-5})*,* rs1558902 near FTO ($P_{replication}$=2.7×10^{-5}) and rs7621025 near *STAG1* ($P_{replication}$=1.4×10^{-3}).

We aimed replication for 23 potential novel SNPs for lipids (4 for LDL-cholesterol, 7 for HDL-cholesterol, 9 for triglycerides and 3 for total cholesterol) in an *in silico* analysis including 93,982 individuals. We could significantly replicate 2 variants for LDL-cholesterol (*HNF4A* and *RASIP1*), three for HDL-cholesterol (*C1orf172, RSPO3* and *STAG1),* one for triglycerides (*CTSB*) and one for total cholesterol (*C7orf50*) (Supplementary table II).

Expression Quantitative Trait Loci (eQTL)

To annotate the effect of the replicated pleiotropic variants to the expression level of nearby genes, we investigated the association between the pleiotropic variants and gene expression levels in three different tissues relevant to CRP and lipids by use of large publicly available datasets: whole blood (n=5,311)[15], liver (n=427[16] and 266[17]) and adipose tissue(n=111)[18]. For the replicated pleiotropic variant rs10435719 near *CTSB* and *FDFT1*, we observed significant associations in whole blood with expression levels of two genes: *CTSB* itself (P-value=1.67×10^{-6}), and *FDFT1* (P-value=1.10×10^{-96}). In addition, the SNP rs7621025 near *STAG1* and *PCCB* was strongly associated with expression of the gene *PCCB* in whole blood (P-value=1.1×10^{-40}). No eQTL effect was observed in the liver and adipose tissue.

Discussion

We identified fifty potential pleiotropic SNPs which affect both CRP and lipid levels, of which we replicated three novel CRP variants: rs10435719 (*CTSB/FDFT1*), rs7621025 (*STAG1/PCCB*) and rs1558902 (*FTO*). In silico expression analyses suggested a role for rs10435719 in the gene expression of both *CTSB* and *FDFT1* and rs7621025 appeared to have an effect on the gene expression of *PCCB*.

The locus harboring rs10435719 near *CTSB* and *FDFT1* that was identified for CRP in our study has previously been identified for triglycerides in the joint analysis of the Global Lipids Genetics Consortium combining GWAS data with Metabochip association results[14]*.* We observed a significant effect of rs10435719 on the expression of both *CTSB* and *FDFT1*. The effect of the CRP increasing allele (T) was weakly associated with a decrease in the expression of *CTSB,* whilst we observed a strong association of the T-allele with an increase of *FDFT1* gene expression*. FDFT1* encodes the enzyme squalene synthase which is involved in the cholesterol biosynthesis[19]. Apart from lipids, *FDFT1* has been identified in a GWAS on fatty liver disease[20]. Squalene Synthase Inhibitors (SSI) have been developed and are successful in the reduction of cholesterol levels as well as CRP levels[21]. This pleiotropic effect of cholesterol synthesis blockers on both lipid levels and inflammation is thought to be the consequence of altered isoprenoids levels that may activate pro-inflammatory pathways[22]. The observation that the CRP increasing allele is associated with an increase in *FDFT1* gene expression suggests an effect of rs10435719 on serum CRP through *FDFT1.* However, we searched in large databases to identify robust eQTL effects of the novel variants. Therefore, we were unable to test the association between the expression and CRP and we cannot draw a firm conclusion on the causal effect of the gene expression in the association between the genetic variant and CRP.

We identified the SNP rs7621025 (*STAG1/PCCB*) as a pleiotropic variant for HDL-cholesterol and CRP. We confirmed the effect of rs7621025 on serum CRP in an independent set of individuals and this genomic region has been identified in a GWAS of lipids[14]. The SNP rs7621025 is located within *STAG1*, but has a strong effect on the expression of *PCCB*, located ±300kb downstream of rs7621025 on chromosome 3. *PCCB* has been identified in a GWAS of the protein fibrinogen, an acute phase response protein sharing many genes with CRP[23]. Our results provide further evidence that the *PCCB* gene is involved in inflammation.

We identified the *FTO* gene as a pleiotropic locus for CRP and HDL-cholesterol. The A allele of rs1558902 was associated with an increase of CRP and a decrease in HDL cholesterol. In several GWAS on BMI, the A allele of rs1558902 was also associated with an increase in BMI[24,25]. Previous studies have highlighted the causal effect of obesity on inflammation[26], and the effect directions are consistent with mediation of both the association with CRP and HDL-cholesterol by BMI. We have previously shown that the effect of FTO on CRP is indeed mediated through BMI[27]. Further research is needed to demonstrate whether this is also true for HDL-cholesterol. Our results provide further evidence for the role of obesity in inflammation and highlight the pleiotropic effects of the *FTO* locus on both chronic inflammation and lipid metabolism.

Genetic pleiotropy can be divided in biological and mediated pleiotropy[4]. In biological pleiotropy, the effect of the pleiotropic variant on two or more phenotypes is independent. In mediated pleiotropy, one phenotype mediates the association between the genetic variant and the second phenotype. Both biological and mediated pleiotropic effects may occur for CRP and lipids[28]. In the current study, we did not disentangle the different subtypes of pleiotropy. Moreover, we observed pleiotropic variants with an opposite direction of effect than expected based on the phenotypical correlation in observational epidemiological studies. In biological pleiotropy, opposite directions of effect may occur. As an example, although CRP and LDL-cholesterol are positively associated in observational epidemiological studies, the A-allele of the SNP rs1183910 (*HNF1A*) is associated with lower

CRP levels but higher LDL-cholesterol. Opposite direction of effects are often seen in genetic studies and highlight the complex interplay between correlated phenotypes, in our study CRP and lipids[25]. We did not disentangle the different subtypes of pleiotropy, which is a limitation of the current study.

Our study has certain strengths. We add to previous studies showing that the multivariate method we applied can be effectively utilized to identify potential novel and pleiotropic loci. This method only requires GWAS summary data instead of individual level data from all participating cohorts. Thanks to close collaboration between studies across the world, researchers have performed large GWAS meta-analyses for a vast amount of phenotypes and this data is available for further research. Second, we used the largest GWAS meta-analyses that have so far been done on CRP and lipid levels to identify pleiotropic genetic loci. By doing so, we enhanced the statistical power to detect these loci considerably. Third, we provided robust evidence for three novel CRP loci by replication in an independent sample of genotyped individuals. A limitation of the bivariate meta-analysis is that very strong signals in one of the individual traits may overshadow the weak association with the other phenotype. We set a criterion for the univariate p-values <0.05 to minimize the chance of false positive findings. In many instances the effect of the pleiotropic loci on CRP or lipids is very small. We did not replicate all our pleiotropic loci. This could be due to lack of power in the replication. In concordance, we replicated a larger proportion of the lipid variants in the larger lipid replication sample compared to CRP. Also, variants closer to significance did replicate in the replication study of both CRP and lipids. Also, several variants had substantial heterogeneity I^2 in the replication which lowers the power for replication. Furthermore, the replication sample size was for some variants smaller than 17,743 due to absence of the variants in one or more of the replication studies. However, we cannot rule out the possibility that bivariate p-values are driven by strong associations with one of the phenotypes and produce false positive results. In addition, for the replication of the lipid variants, we used the Metabochip results from the GLGC. Several variants selected for replication were not present on the Metabochip. Although we selected the best available proxy SNP for replication, variants in moderate LD may have limited power for replication. The method used in the current manuscript to prioritize variants with pleiotropic effects among inflammation and cholesterol are hypothesis generating and further functional work regarding the role of the identified variants in cholesterol metabolism and inflammation is necessary.

In conclusion, our results provide evidence for substantial overlap in genetic susceptibility for chronic inflammation and lipid metabolism. In addition, through bivariate genome-wide association studies and replication in an independent sample of individuals we could identify novel genes for CRP.

References

1. Deloukas P, Kanoni S, Willenborg C, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature genetics* 2012.

2. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; 466(7307): 707-13.

3. Dehghan A, Dupuis J, Barbalic M, 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): 731-8.

4. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* 2013.

5. Albert MA, Danielson E, Rifai N, Ridker PM. Effect of statin therapy on C-reactive protein levels. *JAMA* 2001; 286(1): 64-70.

6. Ridker PM, Danielson E, Fonseca FAH, et al. Reduction in C-reactive protein and LDL cholesterol and cardiovascular event rates after initiation of rosuvastatin: a prospective study of the JUPITER trial. *The Lancet* 2009; 373(9670): 1175-82.

7. Jain MK, Ridker PM. Anti-inflammatory effects of statins: clinical evidence and basic mechanisms. *Nature reviews Drug discovery* 2005; 4: 977-87.

8. Perry JRB, Hsu Y-H, Chasman DI, et al. DNA mismatch repair gene MSH6 implicated in determining age at natural menopause. *Human molecular genetics* 2013: ddt620.

9. Hsu YH, Chen X. Identifying Pleiotropic Genetic Effects: A Two-Stage Approach Using Genome-Wide Association Meta-Analysis Data. https://sites.google.com/site/multivariateyihsianghsu/.

10. Hsu Y. The Multi-Phenotype GWAS Analysis For Pleiotropic Genetic Effects. 2013. https://sites.google.com/site/multivariateyihsianghsu/ (accessed 26-11-2013 2013).

11. Hofman A, Murad SD, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *European journal of epidemiology* 2013; 28(11): 889-926.

12. Almqvist C, Adami H-O, Franks PW, et al. LifeGene—a large prospective population-based study of global relevance. *European journal of epidemiology* 2011; 26(1): 67-77.

13. Voight BF, Kang HM, Ding J, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics* 2012; 8(8): e1002793.

14. Global Lipids Genetics C. Discovery and refinement of loci associated with lipid levels. *Nature genetics* 2013.

15. Westra H-J, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics* 2013; 45(10): 1238-43.

16. Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS biology* 2008; 6(5): e107.

17. Innocenti F, Cooper GM, Stanaway IB, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS genetics* 2011; 7(5): e1002078.

18. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* 2013; 45(6): 580-5.

19. Pandit J, Danley DE, Schulte GK, et al. Crystal Structure of Human Squalene Synthase A KEY ENZYME IN CHOLESTEROL BIOSYNTHESIS. *Journal of Biological Chemistry* 2000; 275(39): 30610-7.

20. Chalasani N, Guo X, Loomba R, et al. Genome-wide association study identifies variants associated with histologic features of nonalcoholic fatty liver disease. *Gastroenterology* 2010; 139(5): 1567-76. e6.

21. Stein EA, Bays H, O'Brien D, Pedicano J, Piper E, Spezzi A. Lapaquistat Acetate Development of a Squalene Synthase Inhibitor for the Treatment of Hypercholesterolemia. *Circulation* 2011; 123(18): 1974-85.

22. Schönbeck U, Libby P. Inflammation, Immunity, and HMG-CoA Reductase Inhibitors Statins as Antiinflammatory Agents? *Circulation* 2004; 109(21 suppl 1): II-18-II-26.

23. Dehghan A, Yang Q, Peters A, et al. Association of Novel Genetic Loci With Circulating Fibrinogen Levels A Genome-Wide Association Study in 6 Population-Based Cohorts. *Circulation: Cardiovascular Genetics* 2009; 2(2): 125-33.

24. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics* 2010; 42(11): 937-48.

25. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; 518(7538): 197-206.

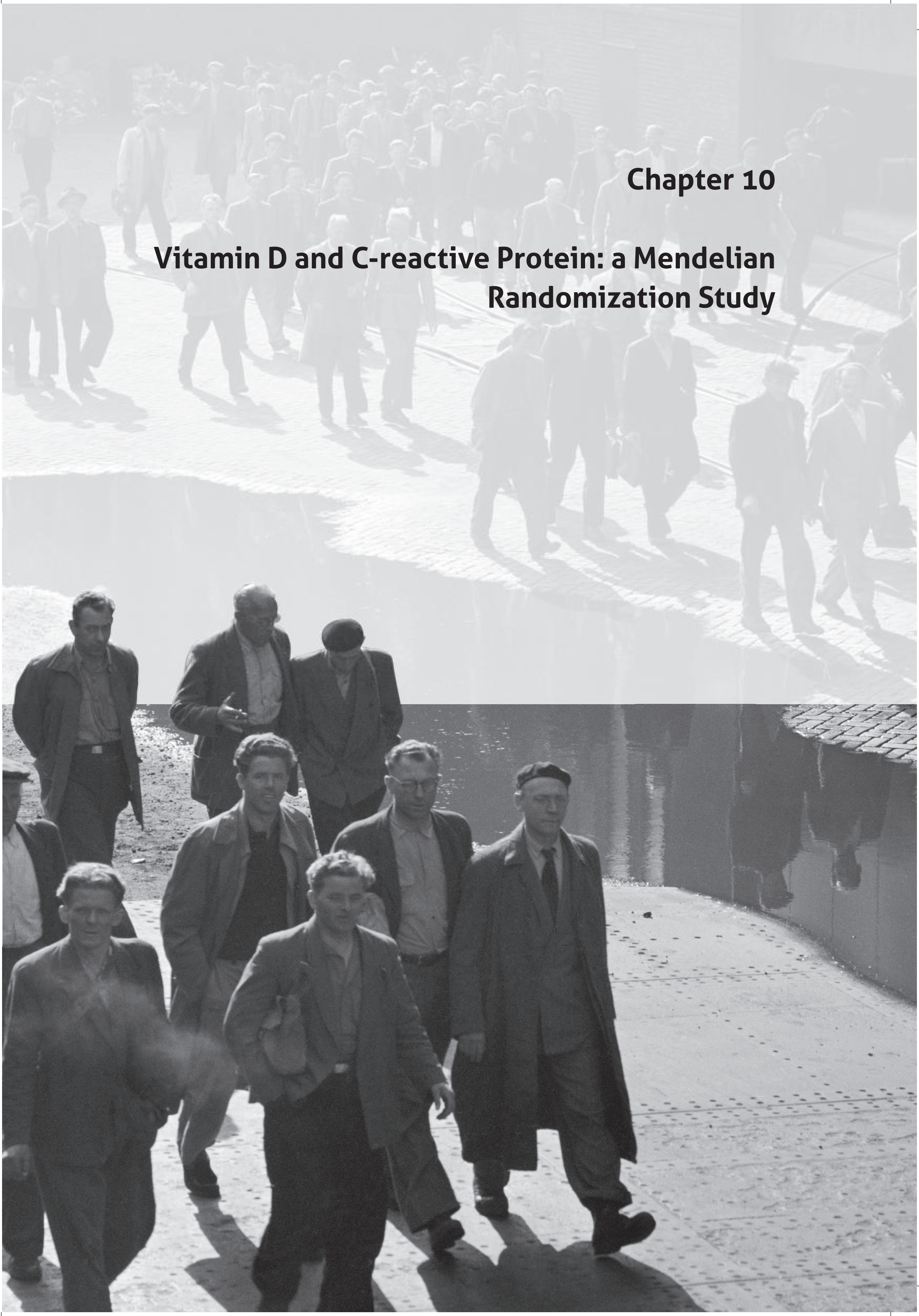
26. Timpson NJ, Nordestgaard BG, Harbord RM, et al. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *International Journal of Obesity* 2011; 35(2): 300-8.

27. Ligthart S, de Vries PS, Uitterlinden AG, et al. Pleiotropy among Common Genetic Loci Identified for Cardiometabolic Disorders and C-Reactive Protein. *PloS one* 2015; 10(3): e0118859.

28. van Diepen JA, Berbée JFP, Havekes LM, Rensen PCN. Interactions between inflammation and lipid metabolism: relevance for efficacy of anti-inflammatory drugs in the treatment of atherosclerosis. *Atherosclerosis* 2013; 228(2): 306-15.

Chapter 10

Vitamin D and C-reactive Protein: a Mendelian Randomization Study



Background: Vitamin D deficiency is widely prevalent and has been associated with many diseases. It has been suggested that vitamin D has effects on the immune system and inhibits inflammation. The aim of our study was to investigate whether vitamin D has an inhibitory effect on systemic inflammation by assessing the association between serum levels of vitamin D and C-reactive protein.

Methods: We studied the association between serum 25-hydroxyvitamin D and C-reactive protein through linear regression in 9,649 participants of the Rotterdam Study, an observational, prospective population-based cohort study. We used genetic variants related to vitamin D and CRP to compute a genetic risk score and perform bi-directional Mendelian randomization analysis.

Results: In linear regression adjusted for age, sex, cohort and other confounders, natural log-transformed CRP decreased with 0.06 (95% CI: -0.08, -0.03) unit per standard deviation increase in 25-hydroxyvitamin D. Bi-directional Mendelian randomization analyses showed no association between the vitamin D genetic risk score and lnCRP (Beta per SD=-0.018; p-value=0.082) or the CRP genetic risk score and 25-hydroxyvitamin D (Beta per SD=0.001; p-value=0.998).

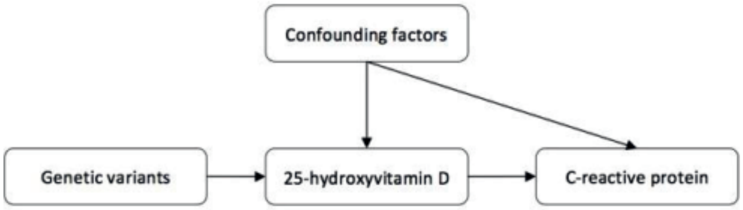
Conclusion: Higher levels of Vitamin D are associated with lower levels of C-reactive protein. In this study we did not find evidence for this to be the result of a causal relationship.

Introduction

Low vitamin D levels are present in up to 50% of the adult population in developed countries[1]. The most important causes for low vitamin D are lack of sun exposure, which leads to inadequate production of the precursor of vitamin D in the skin, and insufficient nutritional intake. The vitamin D receptor is present on immune cells, such as monocytes and T-helper cells. Therefore it is speculated that vitamin D could have effect on immune response and chronic inflammation[2,3,4]. Inflammation is known to be involved in several complex disorders, potentially through its influence on cell growth, tissue damage, pancreatic beta-cell failure and the development of atherosclerosis[5]. Previous studies investigating the association between vitamin D and inflammation have shown inconsistent results[6,7,8,9,10,11,12,13,14,15]. Some studies found inverse associations between serum vitamin D and inflammatory markers, yet due to the observational nature of these studies the question of causality remains unanswered[8,9].

Conclusions about causality cannot be drawn merely based on the presence of an association in an observational design. A complementary alternative is to apply the Mendelian randomization approach, in which the relationship between a genetic determinant of a predictor variable and a specific outcome is studied (Figure 1)[16,17]. If there is indeed a causal effect of vitamin D on inflammation as measured with C-reactive protein (CRP), genetic determinants related to vitamin D should be associated with CRP levels In turn, if inflammation would lower vitamin D levels, genetic determinants of CRP would be expected to be associated with vitamin D levels. These associations are less prone to confounding, since the genetic variants are inherited randomly and do not associate with any other factors. Moreover, reverse causation is unlikely, due to the constant nature of genetic variants over their life course[16,17].

Figure 1. Concept of Mendelian randomization



We investigated the association between serum 25-hydroxyvitamin D and CRP in the Rotterdam Study, a prospective population–based cohort. Furthermore, we evaluated a potential causal effect by using genetic variants in bi-directional Mendelian randomization analysis.

Methods

Study population

This study was conducted among participants of the first (RSI), second (RSII) and third (RSIII) cohort of the Rotterdam Study, a prospective population-based cohort study that has been ongoing since 1989 in the district of Ommoord in the city of Rotterdam, The Netherlands. The design of this study has been described previously[18,19]. In brief, residents aged 55 and over living in the district of Ommoord in Rotterdam, the Netherlands, were invited to participate. Seventy-eight percent of the invitees agreed to participate and were included in the first study cohort (n=7,983). In 1999 the study was extended with a second cohort, comprising 3,011 subjects that had reached the age of 55 years and over. Finally, a third cohort consisting of 3,932 subjects aged 45 and over was included in 2006, after which the study population totals 14,926 subjects. The study was approved by the medical ethics committee at Erasmus University Rotterdam. All participants gave written informed consent.

25-hydroxyvitamin D

Plasma levels of 25-hydroxyvitamin D were measured in non-fasting samples of 1,428 subjects at the first visit of RSI (RSI-1) and 3,799 samples at the third visit of RSI (RSI-3), of which 1,323 were overlapping. Plasma 25-hydroxyvitamin D was measured in fasting samples of 2,464 and 3,420 subjects at the first visits of RSII (RSII-1) and RSIII (RSIII-1) respectively.

In RSI-1, 25-hydroxy vitamin D (25OHD) serum levels were measured using a radioimmunoassay (IDS Ltd, Boldon, UK, available at www.idsltd.com). This test detects levels within a range of 4 to 400 nmol/l, with a sensitivity of 3 nmol/l, a within-run precision <8% and a total precision <12%. Measurements in RSI-3, RSII-1 and RSIII-1 were done using an electrochemiluminescense-based assay (Elecsys Vitamin D Total, Roche Diagnostics, Mannheim, Germany). This test detects levels within a range of 7.50 - 175 nmol/l, with a sensitivity of 10 nmol/l, a within-run precision <6.5% and a total precision <11.5%.

C-reactive protein

At RSI-1, plasma levels of CRP were measured in non-fasting samples of 6,569 subjects, and at RSI-3 in 3,986 subjects, of which 3,694 were overlapping. The samples were put on ice immediately and were processed within 30 minutes. Samples were kept frozen at -20°C until CRP was measured. High-sensitivity CRP was measured using a rate near-infrared particle immunoassay (IMMAGE Immunochemistry System, Beckman Coulter, Fullerton, CA). This system detects concentrations from 0.2 to 1,440 mg/l, with a within-run precision <5.0%, a total precision <7.5%, and a reliability coefficient of 0.995.

In RSII-1 and RSIII-1, plasma levels of CRP were measured in fasting samples of 2,512 and 3,440 subjects respectively. CRP was measured using a particle enhanced immunoturbidimetric assay (Roche Diagnostics, Mannheim, Germany), which detects concentrations from 0.3-350 mg/l, with a sensitivity of 0.6 mg/l.

Genotyping

Genotyping was done using genomic DNA extracted from peripheral venous blood samples according to standard procedures. Genotyping was performed with the version 3 Illumina Infinium HumanHap 550K chip RSI and RSII and the Illumina Infinium HumanHap 610 Quad chip in RSIII. SNPs with allele frequency ≤1%, Hardy–Weinberg equilibrium P-value$<10^{-6}$, or SNP call rate <98% were excluded. Imputation was performed with 1000 Genome phase I, version 3 as the reference panel using the maximum likelihood method implemented in MACH[20,21]. We selected four vitamin D related SNPs based on a genome-wide association study (GWAS) on serum 25-hydroxyvitamin D[22]. For C-reactive protein, we selected 18 SNPs from the latest available GWAS on serum C-reactive protein[23]. The selected SNPs are listed in Table 1.

Table 1. SNPs associated with 25-hydroxyvitamin D or C-reactive protein.

SNP	Associated with	Risk Allele	Nearest Gene
rs12785878	25-hydroxyvitamin D	G	DHCR7
rs10741657	25-hydroxyvitamin D	G	CYP2R1
rs2282679	25-hydroxyvitamin D	G	GC
rs6013897	25-hydroxyvitamin D	A	CYP24A1
rs2794520	C-reactive protein	C	CRP
rs4420638	C-reactive protein	A	APOC1
rs1183910	C-reactive protein	G	HNF1A
rs4420065	C-reactive protein	C	LEPR
rs4129267	C-reactive protein	C	IL6R
rs1260326	C-reactive protein	T	GCKR
rs12239046	C-reactive protein	C	NLRP3
rs6734238	C-reactive protein	G	IL1F10
rs9987289	C-reactive protein	A	PPP1R3B
rs10745954	C-reactive protein	A	ASCL1
rs1800961	C-reactive protein	C	HNF4A
rs340029	C-reactive protein	T	RORA
rs10521222	C-reactive protein	C	SALL1
rs12037222	C-reactive protein	A	PABPC4
rs13233571	C-reactive protein	C	BCL7B
rs2847281	C-reactive protein	A	PTPN2
rs6901250	C-reactive protein	A	GPRC6A
rs4705952	C-reactive protein	G	IRF1

Covariates

Body Mass Index (BMI) was calculated as weight in kilogram divided by the square height in meters. Height and body weight were measured while the participants wore indoor clothing and no shoes. Blood pressure was defined as the mean of two consecutive measurements, which were obtained by trained research assistants from the right brachial artery, with the patient in a sitting position.

Total cholesterol and high-density lipoprotein were measured with standard laboratory techniques, after which the TC/HDL ratio was calculated. Prevalent diabetes mellitus was defined as a fasting serum glucose≥7.0 mmol/l, a non-fasting serum glucose≥11.1 mmol/l and/or use of anti-diabetic medication. The abbreviated modification of diet in renal disease (MDRD) equation was used to estimate glomerular filtration rate[24]. Smoking habits were divided in three categories: former smoker, current smoker and never smoker. Information on current health status, medical history, medication use, alcohol use, smoking behavior and education was obtained by trained research assistants during home visits. Level of education was categorized according to the International Standard Classification of Education.[25] Bone mineral density measurement of the femoral neck was performed by dual energy X-ray absorptiometry (DXA) (Lunar DPX-L densitometer, Madison, WI, USA)[26]. From these measurements, sex-specific T-scores were calculated using the NHANES reference population of Caucasian males and females aged 20 to 29 years[27].

Statistical analysis

To assess the relation between 25-hydroxyvitamin D and CRP we performed linear regression analysis. Due to its right skewed distribution, CRP levels were natural log-transformed prior to analysis. Participants with values larger than 4 standard deviations from the mean in natural log-transformed CRP (lnCRP) and/or 25-hydroxyvitamin D were excluded from the analyses.

In the first model, we assessed the association between lnCRP and 25-hydroxyvitamin D in samples taken from RSI-3, RSII-1 and RSIII-1, adjusting for age, sex and cohort. In the second model, additional adjustments were made for variables including body mass index (BMI), total cholesterol to high-density lipoprotein ratio (TC/HDL ratio), systolic blood pressure (SBP), smoking status, alcohol intake, estimated glomerular filtration rate (eGFR), prevalent type 2 diabetes mellitus (DM), season of blood drawing and level of education. We also performed stratified linear regression analysis for deficient (<50 nmol/l), insufficient (50 – 75 nmol/l) and sufficient (>75 nmol/l) plasma levels of vitamin D, in accordance with the guidelines of the Endocrine Society[28]. Additionally, we repeated these analyses in a quadratic model, in which we added a variable for squared 25-hydroxyvitamin D to assess whether the relation between 25-hydroxyvitamin D and CRP was non-linear. To account for potential confounding by use of vitamin D supplements, we repeated our analyses in a subset of RSI-3 (n=2,746), which we adjusted for prevalent osteoporosis as a proxy for supplement use.

We constructed a genetic risk score (GRS) by adding the 25-hydroxyvitamin D lowering alleles (coded 0–2) from each selected SNP for each individual[22]. For C-reactive protein, we created a similar genetic risk score from 18 CRP related SNPs, with the effect allele being the CRP raising allele[23]. We performed linear regression analysis to confirm the association between the genetic risk scores and their respective phenotypes. We then performed bi-directional Mendelian randomization analyses. First, we tested the associations between individual 25-hydroxyvitamin D related SNPs and lnCRP and corrected them using Bonferroni correction[29]. We used age, sex and cohort adjusted linear regression to examine the effect of the GRS for 25-hydroxyvitamin D on lnCRP and the effect of the GRS for CRP on 25-hydroxyvitamin D. Furthermore, we used a method proposed by Dastani et al. to approximate the effect of the GRS for 25-hydroxyvitamin D on lnCRP using data of a CRP GWAS with a sample size of 66,185 so we would be able to achieve greater power[23,30].

For all but one variable, less than 2% of participants had missing data. For alcohol intake the percentage missing was 6.7%. We used multiple imputation, creating 5 datasets, to complete cases with missing values for the variables included in our analysis. We did not impute 25-hydroxyvitamin D or C-reactive protein levels, but we did enter them as predictor variables in our imputation model. An overview of missing data is given in Table S1.

Tests were considered statistically significant at p-values lower than 0.05. Analyses were performed with IBM SPSS Statistics version 21.0.

Results

Characteristics of the population under study are shown in Table 2, categorized according to vitamin D status. The mean age of the participants was 64.9 years and 43.2 % were male. The mean plasma 25-hydroxyvitamin D level was 55.9 nmol/l (SD 27.6) and median CRP was 1.6 mg/l (IQR: 0.70–3.55). Study participants that had data on 25-hydroxyvitamin D available (n=9,649) were divided in groups of sufficient vitamin D levels (n=2,294), insufficient levels (n=2,784) or deficient levels (n=4,571). Participants from the population eligible for analysis were younger, had lower blood pressure, a lower prevalence of diabetes and a higher education than those from the non-eligible population (Table S2). After correcting for age, the differences in systolic blood pressure and alcohol intake disappeared.

Table 3 shows the results of the linear regression analysis of lnCRP on 25-hydroxyvitamin D. In the age, sex and cohort adjusted linear regression, lnCRP decreased with 0.13 unit (95% CI: -0.15, -0.11) per standard deviation increase in 25-hydroxyvitamin D. There was a consistent trend across the three different categories of vitamin D levels (p-value=4.98×10^{-25}). After further adjustment for BMI, SBP, eGFR, TC/HDL ratio, alcohol intake, smoking,

Table 2. Characteristics of study participants.

	<50 nmol/l	50–75 nmol/l	>75 nmol/l
Number of subjects	4,571	2,784	2,294
Age, years	70.9 (10.7)	63.5 (8.7)	62.1 (7.9)
Sex, male	1,725 (37.7)	1,303 (46.8)	1,139 (49.7)
Body mass index, kg/m^2	28 (5)	27 (4)	26 (4)
25-hydroxyvitamin D, nmol/l	32.6 (10.6)	61.8 (7.1)	95.0 (16.5)
C-reactive protein, mg/l	2.0 (0.8-4.1)	1.4 (0.6-3.1)	1.2 (0.5-2.7)
Systolic blood pressure, mmHg	141 (22)	138 (20)	136 (20)
eGFR, ml/min/1,73m^2	81 (19)	82 (17)	82 (16)
TC/HDL ratio	4.5 (1.4)	4.3 (1.3)	4.2 (1.3)
Alcohol intake, gram/day	5.7 (0.3-15.0)	15.0 (1.4-16.3)	15.0 (2.9-24.3)
Smoking			
Never	1,504 (32.9)	799 (28.7)	623 (27.2)
Former	1,931 (42.2)	1,388 (49.9)	1,156 (50.4)
Current	1,064 (23.3)	566 (21.0)	499 (21.8)
Prevalent DM	701 (15.3)	272 (9.8)	148 (6.5)
Level of education			
ISCED 0	692 (15.1)	286 (10.3)	225 (9.8)
ISCED 1	1,838 (40.2)	1,130 (40.6)	904 (39.4)
ISCED 2	1,275 (27.5)	806 (29.0)	714 (31.1)
ISCED 3	742 (16.2)	548 (19.7)	424 (18.5)

Numbers show mean (SD) for age, body mass index, 25-hydroxyvitamin D, systolic blood pressure, eGFR and TC/HDL ratio, median (IQR) for C-reactive protein and alcohol intake, and frequency (%) for sex, smoking, prevalent DM and level of education.

Abbreviations: eGFR, estimated glomerular filtration rate; TC/HDL ratio, total cholesterol to high-density lipoprotein ratio; DM, diabetes mellitus; ISCED, International Standard Classification of Education.

Table 3. Association between serum 25-hydroxyvitamin D and C-reactive protein.

	N	Model 1	Model 2
		Beta (95% CI)	Beta (95% CI)
<50 nmol/l	4,571	Reference	Reference
50 – 75 nmol/l	2,784	-0.23 (-0.28, -0.18)	-0.12 (-0.17, -0.07)
>75 nmol/l	2,294	-0.28 (-0.34, -0.22)	-0.12 (-0.18, -0.07)
P-value for trend		4.98×10^{-25}	4.48×10^{-6}
Per SD 25OHD*	9,649	-0.13 (-0.15, -0.11)	-0.06 (-0.08, -0.03)
P-value		2.31×10^{-27}	1.70×10^{-6}

Model 1: adjusted for age, sex and cohort.

Model 2: adjusted for age, sex, cohort, body mass index, total cholesterol to high-density lipoprotein ratio, systolic blood pressure, prevalent diabetes mellitus, estimated glomerular filtration rate, smoking, alcohol intake, season and level of education.

*25OHD denotes 25-hydroxyvitamin D.

prevalent diabetes, season of blood drawing, income and level of education, the effect estimates attenuated substantially (B=-0.06, 95% CI: -0.08, -0.03, p-value for trend=$4.48 \cdot 10^{-6}$). We repeated these analyses with a quadratic term for vitamin D added to the regression model. Squared vitamin D was significantly associated with log-transformed CRP in both the first (p-value=$8.55 \cdot 10^{-9}$) and the second model (p-value=$3.21 \cdot 10^{-6}$) (Table S3). Moreover, in a subset of RSI-3 in which we additionally adjusted for osteoporosis, we found similar results in the first and second model as in the previous analyses comprising the larger study population (Table 4). Our quadratic model was not significant in this subset (Table S4).

Table 4. Association between serum 25-hydroxyvitamin D and C-reactive protein in subjects with data on osteoporosis available.

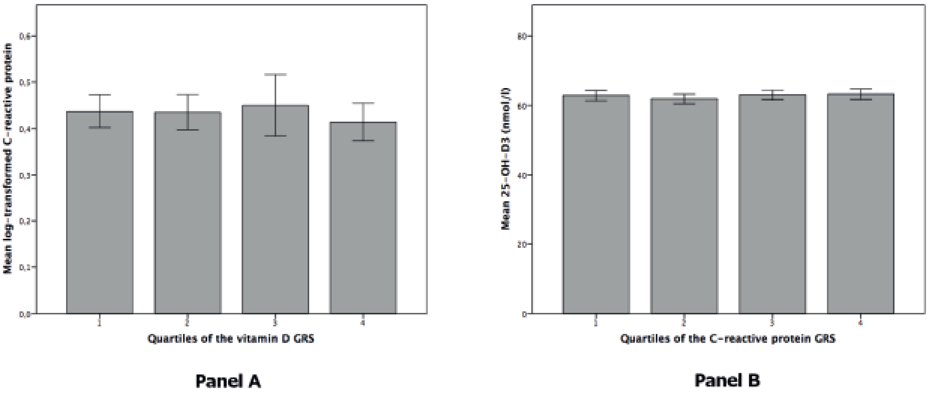
		Model 1	*Model 2*	*Model 3*
	N	**Beta (95% CI)**	**Beta (95% CI)**	**Beta (95% CI)**
<50 nmol/l	1,579	Reference	Reference	Reference
50–75 nmol/l	749	-0.22 (-0.31, -0.12)	-0.12 (-0.21, -0.03)	-0.12 (-0.21, -0.03)
>75 nmol/l	418	-0.26 (-0.37, -0.14)	-0.15 (-0.26, -0.04)	-0.15 (-0.26, -0.04)
P-value for trend		6.15×10^{-7}	0.003	0.003
Per SD 25OHD*	2,746	-0.12 (-0.17, -0.08)	-0.07 (-0.12, -0.03)	-0.07 (-0.11, -0.02)
P-value		5.48×10^{-7}	0.004	0.004

Model 1: adjusted for age and sex.
Model 2: adjusted for age, sex, body mass index, total cholesterol to high-density lipoprotein ratio, systolic blood pressure, prevalent diabetes mellitus, estimated glomerular filtration rate, smoking, alcohol intake, season and level of education.
Model 3: additionally adjusted for osteoporosis.
* 25OHD denotes 25-hydroxyvitamin D.

Mendelian randomization analyses

The genetic risk scores for vitamin D and CRP were robustly associated with their respective phenotypes (S1 and S2 Figs). The 25-hydroxyvitamin D GRS explained 5.1% of the variation in serum 25-hydroxyvitamin D. The 25-hydroxyvitamin D GRS was not associated with lnCRP (n=10,788, β=-0.018 per SD, p-value=0.082). Moreover, there was no significant trend across the GRS quartiles (Figure 2). Associations of individual SNPs with lnCRP are shown in S5 Table. Among all, rs2282679 (*GC*: Vitamin D binding protein) was significantly associated with lnCRP (p-value=0.027), however, after correcting for multiple testing this was no longer significant. The additional analysis that estimated the effect of the GRS for 25-hydroxyvitamin D on lnCRP in data of a CRP GWAS did not provide a significant result (p-value=0.23). The CRP GRS explained 5.5% of the variation in lnCRP. We did not observe a significant association between the CRP GRS and serum 25-hydroxyvitamin D (n=6,267, β=0.001 per SD, p-value=0.998). Similarly, after dividing the GRS in quartiles, there was no significant trend (Figure 2).

Figure 2. Results of Mendelian randomization analyses with the genetic risk scores in quartiles.



Panel A: quartiles of the 25-hydroxyvitamin D genetic risk score in relation to C-reactive protein. P-value for trend=0.056.
Panel B: quartiles of the C-reactive protein genetic risk score in relation to 25-hydroxyvitamin D. P-value for trend=0.374
Error bars represent 95% confidence intervals.

Discussion

Our observational data suggest an inverse association between serum 25-hydroxyvitamin D and C-reactive protein. However, since genetic determinants of serum vitamin D were not associated with serum CRP in the Mendelian randomization approach, our study does not provide evidence for a causal relationship between vitamin D and inflammation.

There are several ways in which vitamin D is able to affect the immune system that could explain the observed association with CRP. It has been shown that immune cells, such as macrophages and dendritic cells, express 1-a-hydroxylase, and thus are able to locally convert 25-hydroxyvitamin D into the active form of vitamin D, 1.25-dihydroxyvitamin D[31,32]. Moreover, the vitamin D receptor is present on leukocytes, T-helper cells and monocytes. 1.25-dihydroxyvitamin D has been shown to inhibit production of inflammatory markers such as IFN-γ, IL-2, and IL-5 by T-helper 1 lymphocytes[33,34]. Vitamin D also inhibits synthesis of IL-6 by monocytes, which is the primary stimulant of CRP production in the liver[35,36].

Previous observational studies that investigated the relationship between vitamin D and inflammatory markers such as CRP have shown mixed results. Shea et al. studied the relation of vitamin D with several inflammatory markers cross-sectionally in 1,381 subjects from the Framingham Offspring Study cohort and did not find a significant association for most of the markers, including CRP[6]. Another, smaller study by Michos et al. did also not find a significant association between vitamin D and CRP[7]. Patel et al. observed an inverse relation between vitamin D and CRP in patients with polyarthritis[8]. Amer et al. found a significant inverse association between 25-hydroxyvitamin D and CRP in a cross-sectional setting in a population of 15,167 adults with a mean age of 46 years from the United States. However, for vitamin D levels above the population median of 21 ng/ml, this relationship reversed, leading the authors to conclude that above this level, vitamin D may actually be pro-inflammatory[9]. In our study, we found that a quadratic model fit the data better than a linear model, suggesting that the relation between vitamin D and CRP may indeed not be linear. The analyses by Amer et al. were done in a younger population and were not adjusted for season of blood drawing or geographical location, which may explain the difference compared to our results.

Several randomized controlled trials have been performed to investigate the effect of vitamin D supplementation on CRP. Coussens et al. found that 95 patients who were treated for tuberculosis and received additional vitamin D supplementation had a faster drop in CRP levels than those who received placebo[10]. In a small study of 54 subjects by Timms et al. there was a decrease in CRP after one year of vitamin D supplementation, but the study was unblinded and included severely vitamin D deficient subjects (25-hydroxyvitamin D <11 ng/ml or <27 nmol/l) only[11]. Chen et al. performed a meta-analysis of randomized controlled trials that investigated the effect of vitamin D on high-sensitive C-reactive protein. They analyzed data of 10 studies, totaling 924 subjects, and found that vitamin D had a significant effect on C-reactive protein. Since there was evidence of heterogeneity these results should be interpreted with caution[12]. However, other randomized trials have not been able to confirm these effects. Schleithoff et al. investigated cytokine profiles in 93 heart failure patients who received vitamin D supplementation or placebo. After 9 months of follow-up there was no effect on CRP[13]. In a study of 314 subjects, Pittas et al. found that after 3 years of vitamin D supplementation there was no significant difference in the decrease of CRP between the placebo and treatment group[14]. Bjorkman et al. did not find an effect of vitamin D supplementation versus placebo in a 6-month trial in 218 older patients[15].

High vitamin D levels may be the result of oral supplementation. Subjects that have an indication to use vitamin D supplements are generally people with decreased bone mineral density[28]. These subjects are more likely to have comorbidities, and thus increased CRP levels. Therefore, use of supplements is a possible confounder of the association between vitamin D and CRP. Since no reliable data were available for vitamin D supplementation, we used prevalent osteoporosis as a proxy for use of vitamin D supplements and adjusted for this in a sensitivity analysis. This did not influence our effect estimate. The quadratic model was not significant in this subset, possibly due to a small sample size and limited power.

Mendelian randomization analyses did not provide significant results. The association between the vitamin D GRS and lnCRP is not consistent with the observational association that we found between serum vitamin D and lnCRP, since the direction of effect is opposite. The result was mainly driven by one SNP, rs2282679, which is located in the gene that encodes the vitamin D binding protein that has no other known functions.

The major strengths of this study are the large sample size for measurements of both CRP and vitamin D, and a comprehensive assessment of this association using both observational and genetic data. By using analytic methods proposed by Dastani et al., we were able to greatly increase the number of subjects for Mendelian randomization analysis. We are the first study to investigate the causal relationship between vitamin D and inflammation through the Mendelian randomization approach. Some limitations should be acknowledged. The 25-hydroxyvitamin D GRS explained only 5.1% of the variation in serum 25-hydroxyvitamin D and the CRP GRS only explained 5.5 of the variation in serum CRP, which could mean that our study is underpowered to find a significant association in Mendelian randomization analyses. We only studied one inflammatory marker to assess the association between vitamin D and inflammation. However, CRP is a widely used marker for chronic inflammation that comprises different aspects of the complex immune system. We aimed to adjust for vitamin D supplement intake, but we did not have a representative variable and had to use a proxy on which information was only available for a small number of people. Our population consisted of elderly individuals, who have more co-morbidities than younger people and are more likely to be sun deprived, which could have had impact on our results. Furthermore, the results may not be valid for all ethnic groups, since our population consisted of Caucasian individuals.

In conclusion, serum vitamin D was inversely associated with CRP, but results of Mendelian randomization analyses do not provide evidence for a causal association. The observed association between vitamin D and CRP is possibly due to residual confounding, but a causal relationship cannot be ruled out yet. Further studies are necessary to understand the role and mechanisms of vitamin D on non-communicable disease prevention and the potential effect of vitamin D supplementation on inflammation.

References

1. Holick MF. Vitamin D deficiency. *The New England journal of medicine* 2007; 357(3): 266-81.

2. Bhalla AK, Amento EP, Clemens TL, Holick MF, Krane SM. Specific high-affinity receptors for 1,25-dihydroxyvitamin D3 in human peripheral blood mononuclear cells: presence in monocytes and induction in T lymphocytes following activation. *J Clin Endocrinol Metab* 1983; 57(6): 1308-10.

3. Provvedini DM, Tsoukas CD, Deftos LJ, Manolagas SC. 1,25-dihydroxyvitamin D3 receptors in human leukocytes. *Science* 1983; 221(4616): 1181-3.

4. Guillot X, Semerano L, Saidenberg-Kermanac'h N, Falgarone G, Boissier MC. Vitamin D and inflammation. *Joint Bone Spine* 2010; 77(6): 552-7.

5. Donath MY, Ehses JA, Maedler K, et al. Mechanisms of beta-cell death in type 2 diabetes. *Diabetes* 2005; 54 Suppl 2: S108-13.

6. Shea MK, Booth SL, Massaro JM, et al. Vitamin K and vitamin D status: associations with inflammatory markers in the Framingham Offspring Study. *American journal of epidemiology* 2008; 167(3): 313-20.

7. Michos ED, Streeten EA, Ryan KA, et al. Serum 25-hydroxyvitamin d levels are not associated with subclinical vascular disease or C-reactive protein in the old order amish. *Calcif Tissue Int* 2009; 84(3): 195-202.

8. Patel S, Farragher T, Berry J, Bunn D, Silman A, Symmons D. Association between serum vitamin D metabolite levels and disease activity in patients with early inflammatory polyarthritis. *Arthritis Rheum* 2007; 56(7): 2143-9.

9. Amer M, Qayyum R. Relation between serum 25-hydroxyvitamin D and C-reactive protein in asymptomatic adults (from the continuous National Health and Nutrition Examination Survey 2001 to 2006). *The American journal of cardiology* 2012; 109(2): 226-30.

10. Coussens AK, Wilkinson RJ, Hanifa Y, et al. Vitamin D accelerates resolution of inflammatory responses during tuberculosis treatment. *Proceedings of the National Academy of Sciences of the United States of America* 2012; 109(38): 15449-54.

11. Timms PM, Mannan N, Hitman GA, et al. Circulating MMP9, vitamin D and variation in the TIMP-1 response with VDR genotype: mechanisms for inflammatory damage in chronic disorders? *QJM : monthly journal of the Association of Physicians* 2002; 95(12): 787-96.

12. Chen N, Wan Z, Han SF, Li BY, Zhang ZL, Qin LQ. Effect of vitamin D supplementation on the level of circulating high-sensitivity C-reactive protein: a meta-analysis of randomized controlled trials. *Nutrients* 2014; 6(6): 2206-16.

13. Schleithoff SS, Zittermann A, Tenderich G, Berthold HK, Stehle P, Koerfer R. Vitamin D supplementation improves cytokine profiles in patients with congestive heart failure: a double-blind, randomized, placebo-controlled trial. *Am J Clin Nutr* 2006; 83(4): 754-9.

14. Pittas AG, Harris SS, Stark PC, Dawson-Hughes B. The effects of calcium and vitamin D supplementation on blood glucose and markers of inflammation in nondiabetic adults. *Diabetes Care* 2007; 30(4): 980-6.

15. Bjorkman MP, Sorva AJ, Tilvis RS. C-reactive protein and fibrinogen of bedridden older patients in a six-month vitamin D supplementation trial. *J Nutr Health Aging* 2009; 13(5): 435-9.

16. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology* 2003; 32(1): 1-22.

17. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine* 2008; 27(8): 1133-63.

18. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *European journal of epidemiology* 2013; 28(11): 889-926.

19. Hofman A, van Duijn CM, Franco OH, et al. The Rotterdam Study: 2012 objectives and design update. *European journal of epidemiology* 2011; 26(8): 657-86.

20. Genomes Project C, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; 491(7422): 56-65.

21. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* 2010; 34(8): 816-34.

22. Wang TJ, Zhang F, Richards JB, et al. Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* 2010; 376(9736): 180-8.

23. Dehghan A, Dupuis J, Barbalic M, 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): 731-8.

24. Levey AS, Coresh J, Greene T, et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Ann Intern Med* 2006; 145(4): 247-54.

25. United Nations Educational SaCOU. International Standard Classification of Education (ISCED). 1976. http://unesdoc.unesco.org/images/0002/000209/020992eb.pdf.

26. Burger H, van Daele PL, Algra D, et al. The association between age and bone mineral density in men and women aged 55 years and over: the Rotterdam Study. *Bone Miner* 1994; 25(1): 1-13.

27. Looker AC, Wahner HW, Dunn WL, et al. Updated data on proximal femur bone mineral levels of US adults. *Osteoporos Int* 1998; 8(5): 468-89.

28. Holick MF, Binkley NC, Bischoff-Ferrari HA, et al. Evaluation, treatment, and prevention of vitamin D deficiency: an Endocrine Society clinical practice guideline. *J Clin Endocrinol Metab* 2011; 96(7): 1911-30.

29. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *Bmj* 1995; 310(6973): 170.

30. Dastani Z, Hivert MF, Timpson N, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS genetics* 2012; 8(3): e1002607.

31. Holick MF. Vitamin D: the underappreciated D-lightful hormone that is important for skeletal and cellular health. *Current Opinion in Endocrinology, Diabetes and Obesity* 2002; 9(1): 87-98.

32. Hewison M, Zehnder D, Chakraverty R, Adams JS. Vitamin D and barrier function: a novel role for extra-renal 1 alpha-hydroxylase. *Mol Cell Endocrinol* 2004; 215(1-2): 31-8.

33. Cantorna MT, Zhu Y, Froicu M, Wittke A. Vitamin D status, 1,25-dihydroxyvitamin D3, and the immune system. *Am J Clin Nutr* 2004; 80(6 Suppl): 1717S-20S.

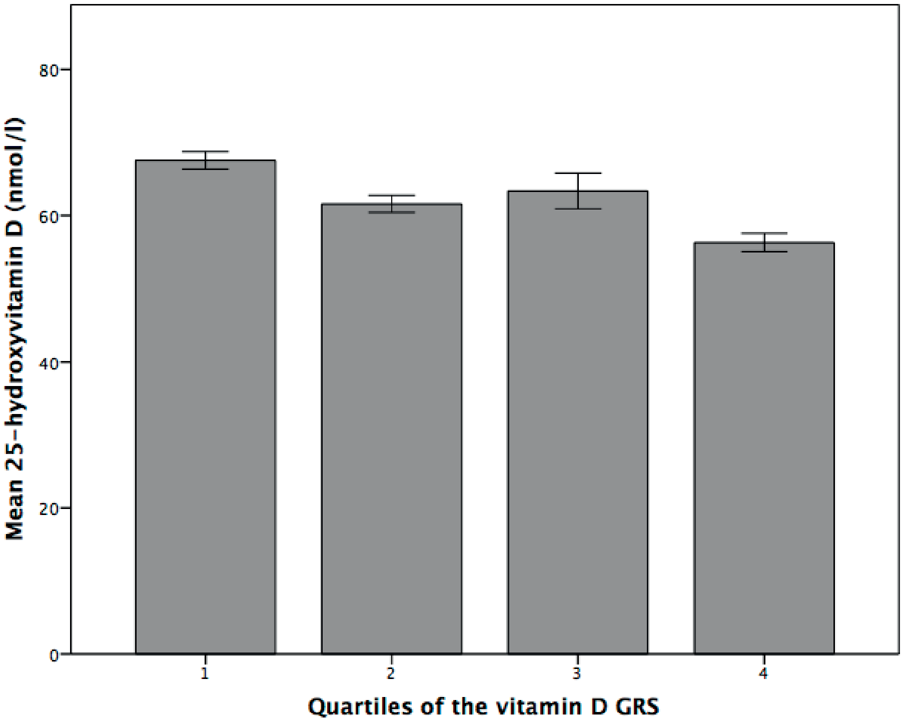
34. Mahon BD, Wittke A, Weaver V, Cantorna MT. The targets of vitamin D depend on the differentiation and activation status of CD4 positive T cells. *J Cell Biochem* 2003; 89(5): 922-32.

35. Zhang Y, Leung DY, Richers BN, et al. Vitamin D inhibits monocyte/macrophage proinflammatory cytokine production by targeting MAPK phosphatase-1. *J Immunol* 2012; 188(5): 2127-35.

36. Dickie LJ, Church LD, Coulthard LR, Mathews RJ, Emery P, McDermott MF. Vitamin D3 down-regulates intracellular Toll-like receptor 9 expression and Toll-like receptor 9-induced IL-6 production in human monocytes. *Rheumatology (Oxford)* 2010; 49(8): 1466-71.

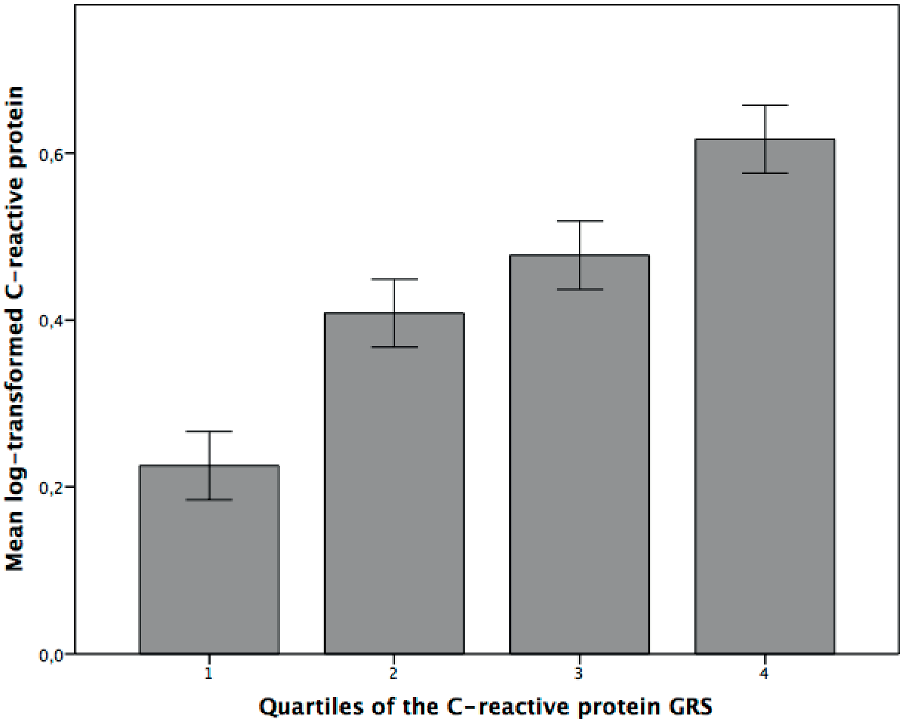
Supplementary material

Figure S1. Quartiles of the 25-hydroxyvitamin D genetic risk score in relation to 25-hydroxyvitamin D.



Error bars represent 95% confidence intervals.
P-value for trend=1.07×10^{-35}.

Figure S2. Quartiles of the C-reactive protein genetic risk score in relation to C-reactive protein.



Error bars represent 95% confidence intervals.
P-value for trend=7.99×10^{-40}.

Table S1. Overview of missing data.

Variable	Percentage missing	Imputed yes/no
Cohort	0	No
Age	0	No
Sex	0	No
25-hydroxyvitamin D	0	No
lnCRP	0	No
Body Mass Index	1.4	Yes
Systolic blood pressure	0.6	Yes
TC/HDL	1.4	Yes
Diabetes Mellitus	0.4	Yes
eGFR	1.6	Yes
Season	0.3	Yes
Alcohol intake	6.7	Yes
Smoking	1.0	Yes
Level of education	0.9	Yes

Abbreviations: lnCRP=natural log-transformed C-reactive protein; TC/HDL ratio=total cholesterol/high density lipoprotein ratio; eGFR=estimated glomerular filtration rate.

Table S2. Comparison of the population under study with the population not under study.

	Population for analysis	Population not eligible for analysis	P-value
Number	9,649	4,977	
Age, years	64.9 (9.8)	73.5 (10.8)	<0.001
Sex, male	4,167 (43.2)	1,860 (37.4)	<0.001
Body mass index, kg/m^2	27 (5)	27 (4)	0.927
Systolic blood pressure, mmHg	140 (21)	142 (23)	0.001
eGFR, ml/min/1,73m^2	81.2 (17.9)	80.4 (18.6)	0.392
TC/HDL ratio	4.4 (1.4)	4.4 (1.3)	0.841
Alcohol Intake, gram/day	12.1 (0.7-15.0)	2.9 (0.0-15.0)	<0.001
Smoking			<0.001
Never	2,926 (30.3)	641 (12.9)	
Former	4,475 (46.4)	796 (16.0)	
Current	2,129 (22.5)	484 (9.7)	
Prevalent DM	1,121 (11.6)	695 (14.0)	<0.001
Level of education			<0.001
ISCED 0	1,203 (12.5)	1,458 (29.9)	
ISCED 1	3,872 (40.1)	1,841 (37.0)	
ISCED 2	2,777 (28.8)	1,064 (21.4)	
ISCED 3	1,714 (17.8)	402 (8.1)	

Numbers show mean (SD) for age, body mass index, systolic blood pressure, eGFR and TC/HDL ratio, median (IQR) for alcohol intake, and frequency (%) for sex, smoking, prevalent DM and level of education.

Abbreviations: eGFR=estimated glomerular filtration rate; TC/HDL ratio=total cholesterol to high-density lipoprotein ratio; DM=diabetes mellitus; ISCED=International Standard Classification of Education.

Table S3. P-values for the association between serum 25-hydroxyvitamin D and C-reactive protein in a quadratic model.

	N	Model 1	Model 2
Squared 25OHD*	9,649	p-value=8.55×10^{-9}	p-value=3.21×10^{-6}

Model 1: adjusted for age, sex and cohort.
Model 2: adjusted for age, sex, cohort, body mass index, total cholesterol to high-density lipoprotein ratio, systolic blood pressure, prevalent diabetes mellitus, estimated glomerular filtration rate, smoking, alcohol intake, season and level of education.
*25OHD denotes 25-hydroxyvitamin D.

Table S4. P-values for the association between serum 25-hydroxyvitamin D and C-reactive protein in a quadratic model in subjects with data on osteoporosis available.

	N	Model 1	Model 2	Model 3
Squared 25OHD*	2,746	p-value=0.153	p-value=0.333	p-value=0.336

Model 1: adjusted for age and sex.
Model 2: adjusted for age, sex, body mass index, total cholesterol to high-density lipoprotein ratio, systolic blood pressure, prevalent diabetes mellitus, estimated glomerular filtration rate, smoking, alcohol intake, season and level of education.
Model 3: additionally adjusted for osteoporosis.
*25OHD denotes 25-hydroxyvitamin D.

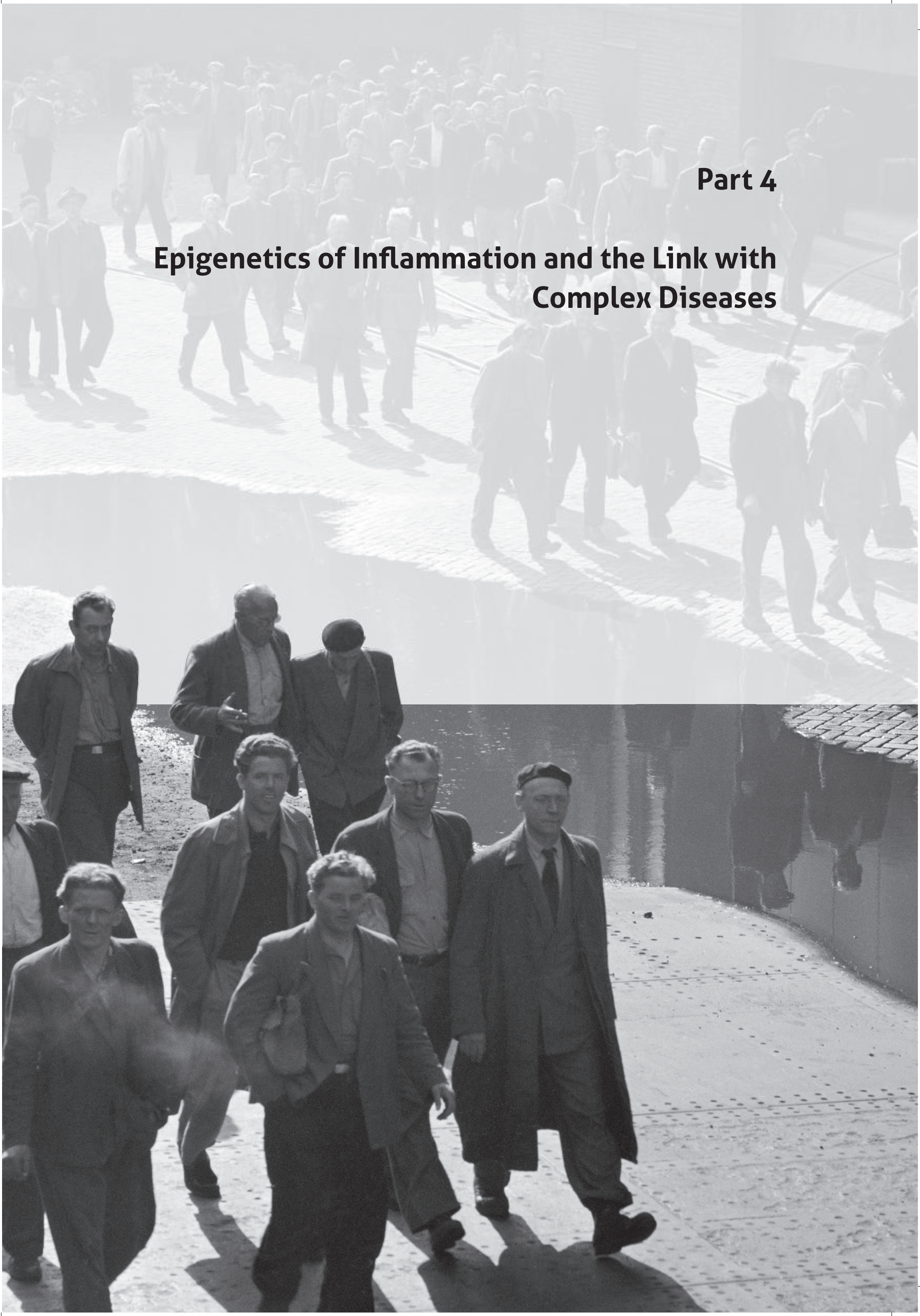
Table S5. Individual associations of vitamin D related SNPs with C-reactive protein.

SNP	Beta	p-value
rs12785878	-0.003	0.897
rs10741657	-0.007	0.659
rs2282679	-0.036	0.027
rs6013897	-0.012	0.493

After Bonferroni correction the threshold for significance lies at P-value=0.0125.

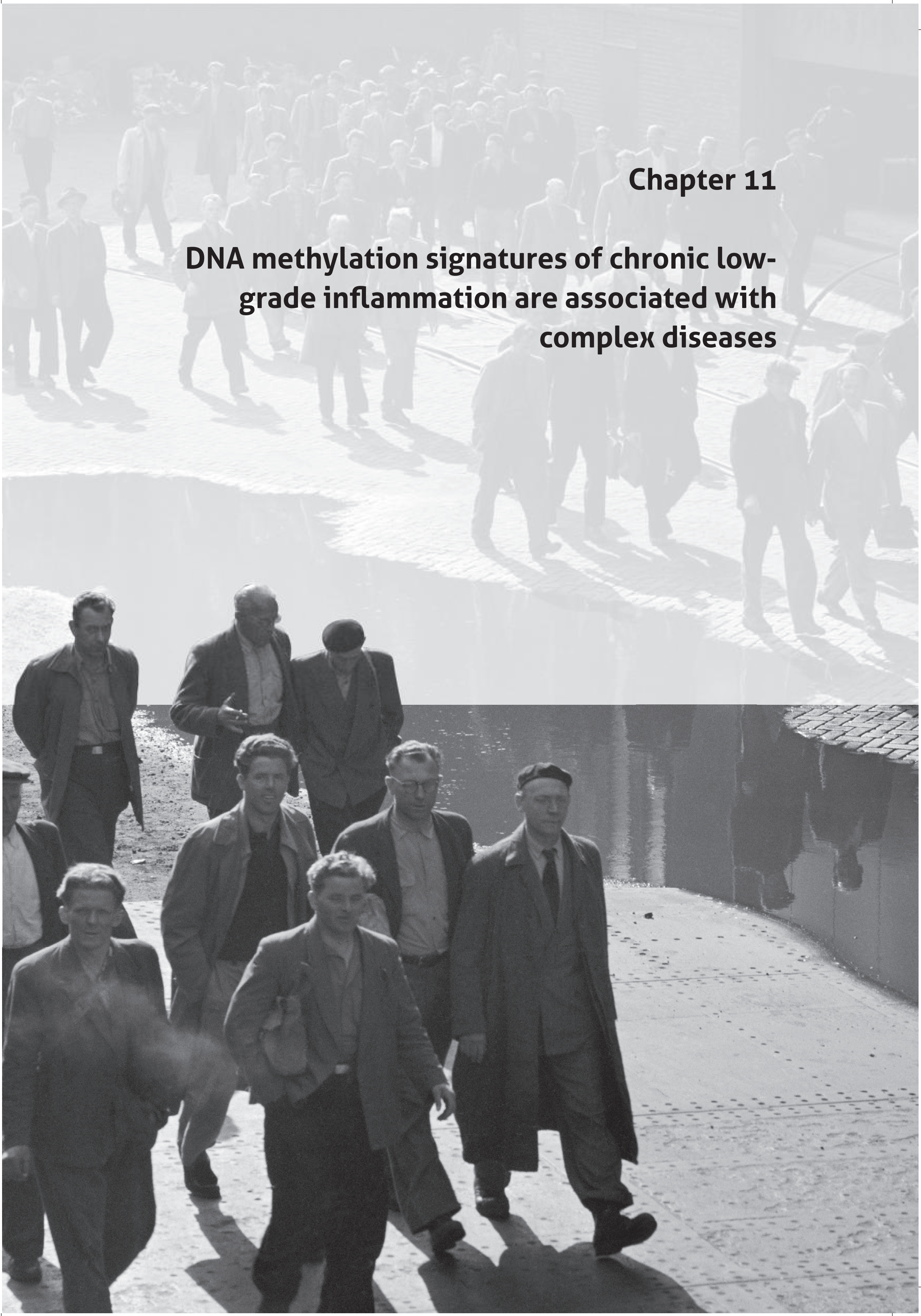
Part 4

Epigenetics of Inflammation and the Link with Complex Diseases



Chapter 11

DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases



Background: Chronic low-grade inflammation reflects a subclinical immune response implicated in the pathogenesis of complex diseases. Identifying genetic loci where DNA methylation is associated with chronic low-grade inflammation may reveal novel pathways or therapeutic targets for inflammation.

Methods: We performed a meta-analysis of epigenome-wide association studies of serum C-reactive protein (CRP), which is a sensitive marker of low-grade inflammation, in a large European population (n=8,863) and trans-ethnic replication in African-Americans (n=4,111).

Results: We found differential methylation at 218 CpG sites to be associated with CRP (P-value<1.15×10^{-7}) in the discovery panel of European ancestry, and replicated (P-value<2.29×10^{-4}) 58 CpG sites (45 unique loci) among African-Americans. To further characterize the molecular and clinical relevance of the findings, we examined the association with gene expression, genetic sequence variants, and clinical outcomes. DNA methylation at 9 (16%) CpG sites was associated with whole blood gene expression in cis (P-value<8.47×10^{-5}), 10 (17%) CpG sites were associated with a nearby genetic variant (P<2.50×10-3), and 51 (88%) also were associated with at least one related cardiometabolic entity (P-value<9.58×10^{-5}). An additive weighted score of replicated CpG sites accounted for up to 6% inter-individuals variation (R2) of age- and sex-adjusted CRP, independent of known CRP-related genetic variants.

Conclusion: We have completed an epigenome-wide association study of chronic low-grade inflammation and identified many novel genetic loci underlying inflammation that may serve as targets for the development of novel therapeutic interventions for inflammation.

Introduction

Chronic low-grade inflammation is a complex immune response that plays an important role in the pathogenesis of multiple chronic diseases, including diabetes and cardiovascular disease[1,2]. C-reactive protein (CRP) is a sensitive marker of chronic low-grade inflammation in community-dwelling adults[3], and is associated in population-based studies with an increased risk of incident coronary heart disease, stroke, and nonvascular mortality[4]. Several pathways have been identified for chronic low-grade inflammation[1,5], and genetic studies have found candidate loci through discovery of genetic sequence determinants of circulating CRP levels[6]. However, most of the molecular mechanisms underlying inter-individual variation in inflammation in the general population and the inter-relation with complex diseases remain to be elucidated.

Epigenetic modifications comprise biochemical alterations to the genome that leave the underlying nucleic acid sequence unchanged but can affect phenotypic expression. DNA methylation is a pivotal and stable epigenetic mechanism whereby a methyl group is attached to the DNA sequence, most often a cytosine nucleotide that neighbors a guanine nucleotide. DNA methylation is affected by both genetic and environmental factors, and regulates gene expression and chromosome stability[7]. Investigating DNA methylation in chronic low-grade inflammation may point to functional epigenetic changes that occur in the context of inflammation.

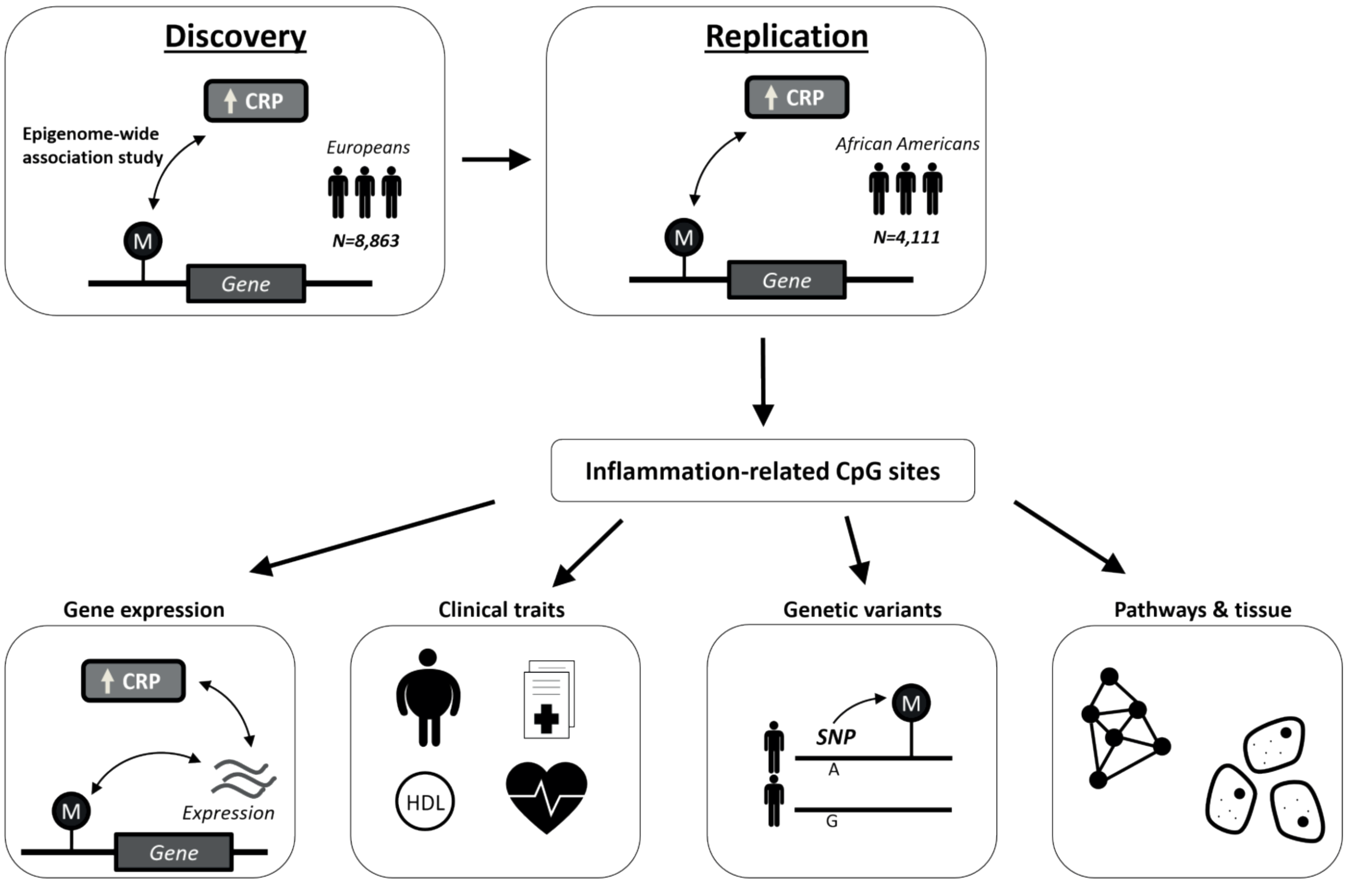
We performed the first meta-analysis of epigenome-wide association studies of methylation of DNA on chronic low-grade inflammation using CRP as a sensitive inflammatory biomarker (Figure 1). We first conducted a discovery meta-analysis, comprising 8,863 participants of European ancestry. Since race or ethnicity may affect epigenetic associations[8], we conducted trans-ethnic replication in 4,111 individuals of African-American ancestry. We further investigated the association between replicated DNA methylation sites and both *cis*- gene expression and genetic variants. Finally, differentially methylated CpG sites were examined for association with cardiometabolic phenotypes to study potential epigenetic links between inflammation and cardiometabolic diseases.

Methods

Discovery and replication study population

Our study was conducted within the framework of the Epigenetics working group of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium[9]. The discovery study population was comprised of 8,863 individuals from the following eleven cohort studies (listed in alphabetical order): the Cardiovascular Health Study (CHS),

Figure 1: Illustration of overall study flow.



the European Prospective Investigation into Cancer and Nutrition (EPIC) Norfolk study, the Framingham Heart Study (FHS), the Invecchiare in Chianti study (InCHIANTI), the Kooperative Gesundheitsforschung in der Region Augsburg (KORA) study, the Lothian Birth Cohorts 1921 and 1936 (LBC1921/1936), the Normative Aging Study (NAS), the Rotterdam Study (RS), and the Women's Health Initiative (WHI). All individuals in the discovery cohorts were of European descent. The trans-ethnic replication population consisted of 4,111 African American individuals from the Atherosclerosis Risk in Communities (ARIC) study, the CHS, the Genetic Epidemiology Network of Arteriopathy (GENOA) study, the Grady Trauma Project (GTP), and the WHI. The studies are described in detail in the supplemental methods (Additional file 13: Supplemental methods). Individuals with autoimmune diseases (rheumatoid arthritis, lupus erythematosus, Crohn's disease, type 1 diabetes) and individuals receiving immune-modulating agents were excluded from all analyses, when disease status and medication data were available. Individuals without such data were assumed to be disease-free and non-users. All participants gave written informed consent and protocols were approved by local institutional review boards and ethic committees.

C-reactive protein measurements

Serum CRP was measured in mg/L using high-sensitivity assays in all studies except the Lothian Birth Cohorts (LBC), in which CRP was measured with the use of a normal sensitivity assay. CRP was measured in blood samples drawn at the same time and center visit as blood was drawn for DNA methylation quantification. CRP values were natural log-transformed (lnCRP). Study-specific methods on the quantification of CRP are described in the Additional file 13: Supplemental methods. Distributions of the natural log transformed serum CRP levels per study are depicted in Additional file 6: Figure S1.

DNA methylation quantification

For the quantification of the DNA methylation, DNA was extracted from whole blood in all studies. All studies used the Illumina Infinium Human Methylation450K BeadChip (Illumina Inc, San Diego, CA, USA) for DNA methylation measurement except GENOA, which used the Illumina Infinium HumanMethylation27K BeadChip (Illumina Inc, San Diego, CA, USA). The 450K Beadchip assays methylation of >480,000 CpGs and is enriched for gene regions and covers 99% of all genes. DNA methylation data pre-processing was conducted independently in different studies and β values were normalized using study-specific methods. We used methylation β values to represent the proportion of the total signal intensity, which ranges from 0 to 1. Further study-specific methods and filtering criteria can be found in Additional file 13: Supplemental methods and Additional file 2: Table S2. A CpG site was deemed polymorphic when a SNP in the 1000 Genomes Project (Phase 1) with a minor allele frequency ≥0.01 resided at the position of the cytosine or guanine on either strand, or within 10 basepairs from the CpG within the probe binding site[8]. Polymorphic CpG sites were excluded from all analyses. Also, cross-reactive probes were excluded from all analyses[10]. In total, 434,253 probes were available for analysis.

Epigenome-wide association study

The epigenome-wide association study was performed at each center separately. Individuals with CRP values >4 standard deviations (SD) from the respective cohort mean lnCRP were excluded from all analyses. In the primary model, we used linear mixed effect regression models to study the methylation β-values, specified as the dependent variable, as a function of lnCRP adjusting for age, sex, white blood cell proportions, technical covariates (array number and position on array), smoking (current, former and never), and body mass index (BMI). Technical covariates were modeled as random effects. Measured or estimated (Houseman method implemented in the *minfi* package in R[11,12]) leukocyte proportions were included to account for cell type admixture (Additional file 2: Table S2). When applicable, models were additionally adjusted for study specific covariates such as study site (fixed effect) and family structure (random effect). Regression models and adjustments were comparable in the discovery and replication analyses. The effect size represents the change in DNA methylation per 1-unit increase in lnCRP.

Meta-analysis

Fixed effects meta-analyses were conducted using the inverse-variance weighted method implemented in METAL, corrected for double lambda control (individual studies and meta-analysis).[13] In the discovery phase, a Bonferroni correction was applied to correct for multiple testing with a significance threshold of $0.05/434{,}253=1.15\times10^{-7}$. We then examined the significant CpG sites for trans-ethnic replication in 4,111 individuals of African-American ancestry using a Bonferroni-corrected significance threshold for the number of CpG sites taken forward for replication. Between-study heterogeneity was examined with Cochran's Q statistic with a Bonferroni-corrected significance threshold for the number of replicated CpG sites. We performed a power calculation for the replication analysis using the GPower 3.1 tool (Additional file 6: Figure S2).[14] Additionally, the European and African-American samples were combined in one meta-analysis.

Sensitivity analyses

In a subset of the discovery cohorts that had further confounders available (CHS, FHS, InCHIANTI, KORA, NAS, RS, and WHI), the replicated CpG sites were additionally adjusted for other potential confounders. These covariates were selected based on strong associations with CRP in observational research[15]. In addition to the variables of the primary model, the sensitivity model included waist circumference, total/high-density lipoprotein

(HDL)-cholesterol ratio, prevalent diabetes (defined as fasting glucose ≥7.0 mmol/L, non-fasting glucose ≥11.1 mmol/L, or the use of diabetes medication), hypertension treatment (use of diuretics, anti-adrenergic agents, β-blockers, calcium channel blockers, and RAAS inhibitors), lipid treatment (use of statins, ezetimibe, and colestyramine), hormone replacement therapy, and prevalent coronary heart disease. Since the population for analysis in the second model was expected to be slightly smaller compared to the primary model due to missing data for certain covariates, we repeated the primary model to include only individuals present in the second model.

To investigate the association between the replicated CpG sites and serum CRP levels in CD4+ cells, we tested the association in the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study which quantified DNA methylation in CD4+ cells. Associations with a consistent effect direction and P-value<0.05 were considered significant.

Annotation of CpG sites

We used the genome coordinates provided by Illumina (GRCh37/hg19) to identify independent loci. A distance criterion of 500kb on either side of each epigenome-wide significant signal was used to define independent loci. In addition to the gene annotation provided by Illumina based on RefSeq database, the UCSC database was explored to further annotate the CpG sites to potential genes (nearest gene).

Methylation and genetic score

To calculate the variance explained by the replicated CpGs, we first selected independent CpGs based on pairwise Pearson correlation R^2. To this end, we first ranked the significant CpGs by discovery p-value in ascending order. We then iteratively excluded CpGs correlated with the top CpG site (r^2>0.1) until we reached a list of independent CpGs (n=8). The eight CpGs were used to construct a methylation score weighted by the effect estimates from regression in the FHS with lnCRP as the dependent variable, and residuals of the DNA methylation (after regressing out age, sex, batch effect, cell counts, smoking, and BMI) as the independent variable. Using a linear regression model, we calculated the CRP variance explained by the methylation score (multiple R^2, adjusting for age and sex) in ARIC, KORA, NAS, and RS. Furthermore, an additive effect-size weighted genetic score for CRP was constructed in RS to include 18 SNPs identified in the largest GWAS of CRP (genotyping information RS in Additional file 13: Supplemental methods)[6]. We calculated weighted dosages by multiplying the dosage of each risk allele (0, 1 or 2) with the published effect estimate. We calculated the CRP variance explained by the genetic score, and both the methylation and genetic score combined[6]. Additionally, the interaction between the methylation and genetic score on CRP was studied using a multiplicative interaction term. Finally, we assessed the association between the genetic and methylation scores.

Association with cardiometabolic phenotypes

The association between the significant CpGs and BMI, total cholesterol, HDL-cholesterol, triglycerides, fasting glucose, fasting insulin, prevalent diabetes, prevalent coronary heart disease (CHD) and incident CHD was explored in CHS, FHS, InCHIANTI, KORA, NAS, RS, and WHI. The analyses on fasting glucose and fasting insulin only included non-diabetic individuals. Diabetes was defined as fasting glucose ≥7.0 mmol/L, non-fasting glucose ≥11.1 mmol/L or the use of glucose-lowering medication. The lipid traits and fasting glucose were analyzed in mmol/L, whilst fasting insulin was analyzed in pmol/L. Fasting insulin and triglycerides were natural log-transformed. CHD (available in ARIC, CHS, EPICOR, FHS, KORA, NAS, RS, and WHI) was defined as fatal or non-fatal myocardial infarction, coronary revascularization, and unstable angina. The statistical models for the cross-phenotype analyses were similar to the basic CRP model (including age, sex, white blood cell counts, technical covariates and smoking) with DNA methylation as the dependent variable. The associations were also adjusted for BMI, except the association with BMI itself. We conducted fixed effect meta-analyses using inverse-variance method for total cholesterol, HDL-cholesterol, fasting glucose, fasting insulin and prevalent diabetes. For incident CHD, associations were analyzed using (penalized) Cox regression models. Results of the cross-phenotype associations with BMI and triglycerides were meta-analyzed combining p-values taking into account the study sample size and direction of effect. Both methods are implemented in METAL. We used a Bonferroni corrected p-value of 0.05 divided by the number of significant CpGs multiplied by nine phenotypes as a threshold of significant cross-phenotype association.

Gene expression analyses

To assess the relations of replicated CpGs with gene expression, we examined the association between replicated CpGs and whole blood gene expression of *cis*-genes (250kb up- and downstream of the CpG). The methylation-expression analyses were conducted in 3,699 individuals from the FHS, KORA, and RS with both DNA methylation and gene expression available from the same blood samples. In RS and KORA, we first created residuals for both DNA methylation and mRNA expression after regressing out age, sex, blood cell counts (fixed effect) and technical covariates (random effect). We then examined the association between the residuals of DNA methylation (independent variable) and mRNA expression (dependent variable) using a linear regression model. In FHS, we removed 25 surrogate variables (SVs)[16] from the gene expression, along with sex, age, and imputed blood cell fractions as fixed effects, and technical covariates, such as batch effects and lab effects as random effects. We also removed 25 separately computed SVs from the methylation data, along with sex, age, and imputed blood cell fractions as fixed effects, and technical covariates, such as batch effects and lab effects as random effects. We then associated the two data using simple linear model. Expression probes were aligned to genes, and unique methylation-gene expression results from FHS (n=2,262), KORA (n=707), and RS (n=730) were meta-analysed using the sample size weighted method implemented in METAL, based on p-values and direction of the effects. To reduce the type 1 error, results for the methylation-expression associations were adjusted for multiple testing using the Bonferroni correction (0.05/590 tests: $P<8.47\times10^{-5}$). Furthermore, for the significant methylation-expression associations, we tested the association between the gene expression and serum CRP levels. We examined the association between gene expression (dependent variable) and CRP levels (independent variable) in a linear model adjusted for age, sex, blood cell counts, technical covariates (plate ID and RNA quality score), tobacco smoking and body mass index. Results from GTP (n=114), FHS (n=5,328), InCHIANTI (n=590), KORA (n=724), and RS (n=870) were meta-analysed using the sample size weighted method implemented in METAL (P-value<0.05 was considered significant)[13]. Information on gene expression quantification in the specific studies can be found in the Additional file 13: Supplemental methods.

Genetic correlates of DNA methylation

We studied genetic variants in the proximity (±250kb) of the inflammation-related CpGs for a methylation quantitative trait effect on the percentage of methylation of the CpG site (*cis*-mQTL). The discovery analyses were conducted in the RS in which 730 participants were available with both genetic and epigenetic data. Genotyping information for the RS is described in Additional file 13: Supplemental methods. We used the expression quantitative trait loci (eQTL) mapping pipeline to study associations between genetic variants in a 500kb window around the CpG site and the percentage of methylation at this CpG site[17]. This pipeline has been applied previously to study expression quantitative trait loci (eQTLs). Instead of analyzing gene expression, we modeled the correlation between genetic variants and DNA methylation and adjusted for 20 principal components derived from the DNA methylation data to account for potential unrelated variation in the DNA methylation caused by environmental or technical effects (batch effects). The threshold of significance for *cis*-mQTLs was defined according to the pipeline specifications by FDR of 5%. When multiple *cis*-mQTLS were identified for the same CpG site, only the SNP with the lowest p-value was reported. Next, significant *cis*-mQTLs were replicated in FHS. The *cis*-mQTL analysis in FHS was performed on 2,408 individuals having both genotype and methylation data. Genotyping information for FHS is described in Additional file 13: Supplemental methods. We removed 50 principal components from the epigenomics data, along with sex, age, and imputed blood cell fractions as fixed effects, and technical covariates, such as batch effects and lab effects as random effects. We then associate the epigenomic residual data with the genotypic data accounting for 10 principal components computed using the

Eigenstrat software using fixed effect linear model. We collected effect value, T statistics, and p-value. We used a Bonferroni corrected p-value of $0.05/20=2.5\times10^{-3}$ (based on 20 findings in the discovery) for significant replication in FHS. Subsequently, replicated *cis*-mQTLs were tested for association with serum CRP in the largest published CRP GWAS (n=66,185) to strengthen the causal inference from our findings[6].

GWAS catalog, pathway analysis, and tissue enrichment

We used the National Human Genome Research Institute (NHGRI) GWAS catalog to query whether genes annotated to replicated CpGs were enriched for genes identified in published GWAS[18]. Altogether, 7,600 SNPs, annotated to 4,498 genes, associated with 988 phenotypes at GWAS p-value $\leq 5\times10^{-8}$, were retrieved on August 25, 2016 from the NHGRI GWAS catalog. Methylation CpGs were matched by gene symbols with the reported genes in the GWAS catalog. CpGs not annotated to a gene were discarded. Enrichment statistics were performed using one-sided Fisher's test. Next, enrichment of canonical pathways was explored using Ingenuity® Pathway Analysis software tool (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity). Replicated CpGs which mapped to a UCSC Refseq gene were included in pathway analyses. Pathway analyses were performed using the Ingenuity Pathway Analyses (IPA) software tool (IPA build version 338830M, content version: 23814503, release date 2016-10-04, analysis date 2015-08-03; http://www.ingenuity.com/). Gene enrichment in canonical pathways was assessed in the core analysis module using Fisher's exact test right tailed. Furthermore, we used experimentally-derived Functional element Overlap analysis of ReGions from EWAS (eFORGE) to identify tissue specific or cell-type specific signals[19]. eFORGE analyzes a set of differentially methylated CpGs for enrichment of overlap with DNase 1 hypersensitivity sites in different cell types of the ENCODE project. All 58 replicated CpGs were entered as the input of the eFORGE analysis. The set of 58 CpGs were tested for enrichment for overlap with putative functional elements compared to matched background CpGs. The functional elements considered are DNase I hotpsots fromthe ENCODE project. The matched background is a set of the same number of CpGs as the test set, matched for gene relationship and CpG island relationship annotation. Thousand matched background sets were applied. The enrichment analysis was performed for different tissues, since functional elements may differ across tissues. Enrichment outside the 99.9th percentile (-log10 binomial p-value: ≥3.38) was considered statistically significant (red).

Results

Clinical characteristics

The nine participating discovery (n=8,863) and four replication cohorts (n=4,111), and the clinical characteristics of the participants are presented in Table 1 (further details, Additional file 1: Table S1). The mean age in the participating studies ranged from 41 years in the GTP cohort to 87 years in LBC1921. The majority (54%) of the samples were from women. Some of the cohorts differed based on selection criteria for entry into the study. The NAS only included men, while the WHI only included women. Mean serum CRP levels (SD) ranged from 2.3 (3.7) mg/L in the KORA study to 7.2 (8.4) mg/L in the African American coronary heart disease cases of WHI.

Discovery meta-analysis

We identified 218 CpG sites significantly associated (P-value<1.15×10^{-7}) with CRP in the meta-analysis of European participants, adjusted for age, sex, white blood cell proportions,

Table 1. Characteristics of the discovery (n=8,863) and replication (n=4,111) studies.

Study	N	Country	Age (years)	Women (%)	CRP (mg/L)	BMI (kg/m^2)
Discovery (European)						
CHS	187	USA	76 (5)	56	6.6 (11.0)	31 (6)
EPIC-Norfolk	1,287	UK	60 (9)	54	3.3 (5.4)	27 (4)
FHS	2,427	USA	66 (9)	52	3.1 (6.7)	28 (5)
InCHIANTI	498	Italy	63 (16)	55	3.2 (3.5)	27 (4)
KORA	1,700	Germany	61 (9)	51	2.3 (3.7)	28 (5)
LBC 1921	169	UK	87 (0)	54	3.7 (8.4)	26 (4)
LBC 1936	296	UK	70 (1)	50	5.3 (6.8)	28 (4)
NAS	648	USA	73 (7)	0	3.3 (6.1)	28 (4)
Rotterdam	702	Netherlands	60 (8)	54	2.7 (4.7)	28 (5)
WHI controls	471	USA	68 (6)	100	3.8 (5.5)	28 (6)
WHI cases	478	USA	69 (6)	100	4.9 (6.4)	29 (6)
Replication (African American)						
ARIC	2,264	USA	56 (6)	64	5.9 (7.8)	30 (6)
CHS	193	USA	73 (5)	65	5.2 (5.6)	29 (5)
GENOA	939	USA	66 (8)	71	6.7 (12.3)	31 (6)
GTP	112	USA	41 (13)	70	5.9 (8.1)	33 (8)
WHI controls	309	USA	62 (6)	100	6.1 (7.5)	31 (7)
WHI cases	294	USA	64 (7)	100	7.2 (8.4)	32 (6)

Characteristics are mean (SD), unless otherwise specified. BMI denotes body mass index, CRP C-reactive protein, UK United Kingdom and USA United States of America.

technical covariates, smoking, and BMI (Manhattan and QQ-plot, Figure 2, and Additional file 2: Table S2, and Additional file 3: Table S3). Serum CRP was positively associated with 125 CpG sites and negatively associated with 93. The top CpG site was cg10636246 at 1q23.1 located within 1,500 base pairs of the transcription start site of *Absent in melanoma 2* (*AIM2*) (effect size=-0.0069, P-value=2.53×10^{-27}), an interferon-gamma induced protein involved in the innate immune response by inducing caspase-1-activating inflammasome formation in macrophages.

Replication meta-analysis

Of the 218 CpG sites significantly associated with CRP in our discovery meta-analysis, 58 replicated (P-value<2.29×10^{-4}) in a trans-ethnic replication meta-analysis of 4,111 individuals of African-American ancestry (Table 2). The replicated CpG sites annotated to 45 separate loci. The most significant CpG site in the discovery panel (cg10636246; *AIM2*) was also strongly related to serum CRP in individuals of African-American ancestry (effect size=-0.0081, P-value=6.31×10^{-9}). Effect estimates of the 58 replicated CpG sites assessed in the European and African-American panel were highly correlated (r=0.97). Cochrane's Q statistics displayed homogeneity for >95% of the 58 replicated loci in both the European discovery panel and the African-American replication panel (study specific effect estimates Additional file 4). In addition, we conducted a meta-analysis combining the European and African-American whole blood samples resulting in 258 significant CpGs (Additional file 5).

Sensitivity analyses

Further adjustment of the replicated CpG sites for additional potential confounders (waist circumference, total/HDL-cholesterol ratio, prevalent diabetes, hypertension treatment, lipid treatment, hormone replacement therapy, and prevalent coronary heart disease) did not substantially change the effect estimates and P-values. Additional file 6: Figure S3 depicts the correlation between the effect estimates and –log10 P-values in the primary model compared to the multivariable adjusted model, respectively. Furthermore, 18 CpGs were found to be associated with serum CRP levels in CD4+ cells in the GOLDN study (P-value<0.05) (Additional file 7: Table S6).

Methylation and genetic scores

Additive weighted methylation and genetic scores were constructed to calculate percentage of total CRP variance explained. A methylation score including eight independent CpGs (cg10636246, cg17501210, cg18608055, cg03957124, cg04987734, cg04523589, cg17980786, and cg02341197) explained 5.8% of the variance of CRP in ARIC, 2.3% in KORA, 5.0% in NAS, and 4.6% in RS. A genetic score including 18 independent CRP

Figure 2: Manhattan and QQ-plot depicting the –log$_{10}$(P-values) of the associations between all CpG sites and C-reactive protein.

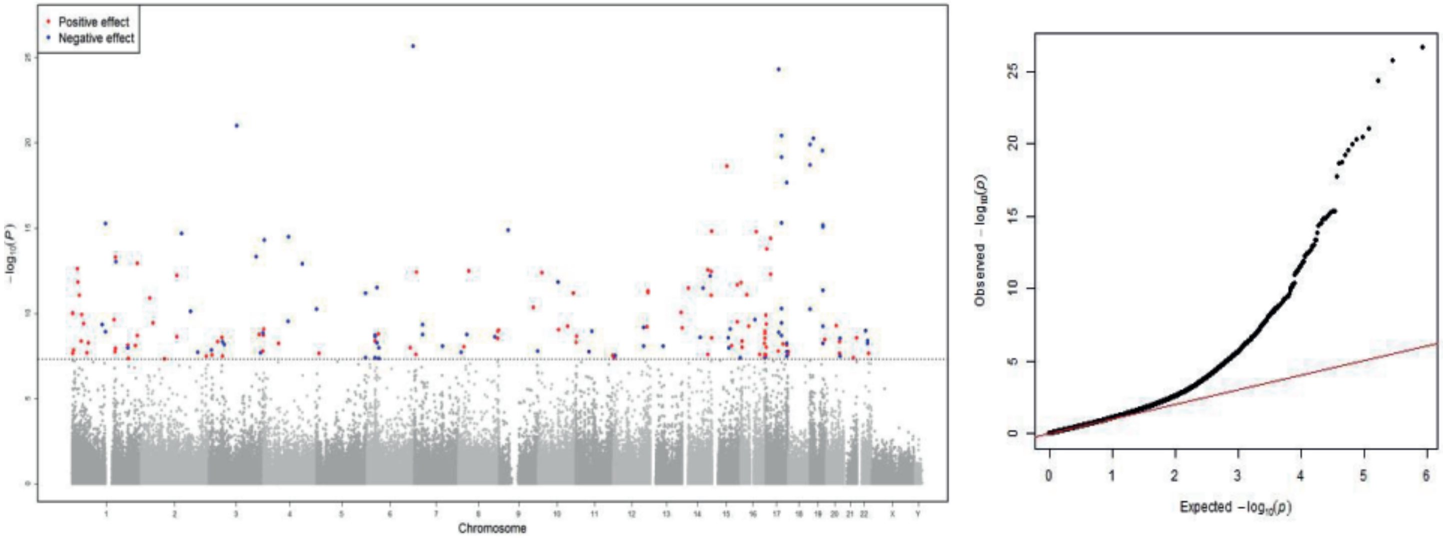


Table 2. DNA methylation sites associated with serum CRP levels.

CpG sites	Chr	Position	Effect size EA	P-value EA	Effect size AA	P-value AA	Gene
cg10636246	1	159046973	-0.0069	2.53×10^{-27}	-0.0081	6.31×10^{-09}	*AIM2*
cg17501210	6	166970252	-0.0065	2.06×10^{-26}	-0.0076	9.45×10^{-05}	*RPS6KA2*
cg02650017	17	47301614	-0.0021	4.87×10^{-25}	-0.0011	7.71×10^{-06}	*PHOSPHO1*
cg12992827	3	101901234	-0.0057	9.73×10^{-22}	-0.0086	4.42×10^{-14}	*NFKBIZ*
cg16936953	17	57915665	-0.0077	3.74×10^{-21}	-0.0125	1.13×10^{-13}	*TMEM49*
cg19821297	19	12890029	-0.0051	5.19×10^{-21}	-0.0055	6.58×10^{-06}	*GCDH*
cg07573872	19	1126342	-0.0052	1.24×10^{-20}	-0.0068	2.98×10^{-09}	*SBNO2*
cg26470501	19	45252955	-0.0045	2.85×10^{-20}	-0.0051	4.08×10^{-07}	*BCL3*
cg12054453	17	57915717	-0.0082	6.96×10^{-20}	-0.0117	4.25×10^{-12}	*TMEM49*
cg18608055	19	1130866	-0.0043	1.94×10^{-19}	-0.0078	2.96×10^{-11}	*SBNO2*
cg06192883	15	52554171	0.0045	2.29×10^{-19}	0.0073	8.29×10^{-12}	*MYO5C*
cg18181703	17	76354621	-0.0053	2.13×10^{-18}	-0.0091	7.08×10^{-13}	*SOCS3*
cg18942579	17	57915773	-0.0056	4.77×10^{-16}	-0.0098	8.70×10^{-12}	*TMEM49*
cg19769147	14	105860954	0.0029	1.51×10^{-15}	0.0029	6.60×10^{-05}	*PACS2*
cg20995564	2	145172035	-0.0051	2.04×10^{-15}	-0.0089	2.69×10^{-10}	*ZEB2*
cg02734358	4	90227074	-0.0048	3.09×10^{-15}	-0.0051	5.51×10^{-05}	*GPRIN3*
cg07094298	4	2748026	-0.0056	4.76×10^{-15}	-0.0058	5.32×10^{-06}	*TNIP2*
cg01059398	3	172235808	-0.0042	4.51×10^{-14}	-0.0068	2.27×10^{-05}	*TNFSF10*
cg06690548	4	139162808	-0.0048	1.21×10^{-13}	-0.0029	1.52×10^{-07}	*SLC7A11*
cg02003183	14	103415882	0.0047	3.59×10^{-13}	0.0051	4.36×10^{-05}	*CDC42BPB*
cg26804423	7	8201134	0.0027	3.87×10^{-13}	0.0038	4.82×10^{-07}	*ICA1*
cg13585930	10	72027357	-0.0037	1.42×10^{-12}	-0.0046	7.95×10^{-05}	*NPFFR1*
cg03957124	6	37016869	-0.0030	3.13×10^{-12}	-0.0039	1.39×10^{-05}	*FGD2*
cg12053291	12	125282342	0.0029	5.99×10^{-12}	0.0038	9.80×10^{-05}	*SCARB1*
cg02481950	16	21665002	0.0022	7.84×10^{-12}	0.0034	2.92×10^{-06}	*METTL9*
cg04987734	14	103415873	0.0041	8.40×10^{-12}	0.0051	1.40×10^{-04}	*CDC42BPB*
cg15551881	9	123688715	0.0039	4.62×10^{-11}	0.0049	3.99×10^{-07}	*TRAF1*
cg27023597	17	57918262	-0.0050	5.02×10^{-11}	-0.0070	5.96×10^{-06}	*MIR21*
cg05575921	5	373378	-0.0059	5.44×10^{-11}	-0.0063	1.17×10^{-04}	*AHRR*
cg27469606	19	1154485	-0.0020	5.62×10^{-11}	-0.0023	1.96×10^{-06}	*SBNO2*
cg01409343	17	57915740	-0.0037	3.56×10^{-10}	-0.0081	6.12×10^{-10}	*TMEM49*
cg21429551	7	30635762	-0.0069	4.42×10^{-10}	-0.0080	1.68×10^{-05}	*GARS*
cg23761815	10	73083123	0.0022	8.86×10^{-10}	0.0029	6.85×10^{-05}	*SLC29A3*
cg08548559	22	31686097	-0.0038	9.94×10^{-10}	-0.0049	9.88×10^{-05}	*PIK3IP1*
cg26610247	8	142297175	0.0029	1.07×10^{-09}	0.0041	4.59×10^{-06}	*TSNARE1*
cg27050612	17	46133198	-0.0019	1.30×10^{-09}	-0.0029	8.23×10^{-05}	*NFE2L1*
cg15721584	3	181326755	0.0055	1.71×10^{-09}	0.0072	1.14×10^{-05}	*SOX2OT*
cg06126421	6	30720080	-0.0052	1.80×10^{-09}	-0.0059	1.53×10^{-04}	*TUBB*
cg00851028	1	234905772	0.0023	1.95×10^{-09}	0.0042	1.46×10^{-05}	-
cg24174557	17	57903544	-0.0038	1.97×10^{-09}	-0.0051	1.65×10^{-04}	*TMEM49*

cg05316065	8	130799007	-0.0027	2.26×10^{-09}	-0.0051	2.28×10^{-07}	*GSDMC*
cg04523589	3	48265146	0.0022	2.49×10^{-09}	0.0031	4.47×10^{-05}	*CAMP*
cg17980786	3	32933637	0.0026	4.58×10^{-09}	0.0055	1.47×10^{-09}	*TRIM71*
cg25325512	6	37142220	-0.0031	5.31×10^{-09}	-0.0052	4.94×10^{-05}	*PIM1*
cg00812761	4	53799391	0.0025	5.60×10^{-09}	0.0036	1.36×10^{-04}	*SCFD2*
cg27637521	17	76355202	-0.0016	5.69×10^{-09}	-0.0017	3.69×10^{-05}	*SOCS3*
cg26846781	17	61620942	0.0018	5.99×10^{-09}	0.0033	3.03×10^{-05}	*KCNH6*
cg00159243	12	109023799	-0.0026	8.22×10^{-09}	-0.0036	1.38×10^{-04}	*SELPLG*
cg15310871	8	20077936	0.0022	8.63×10^{-09}	0.0027	2.96×10^{-05}	*ATP6V1B2*
cg15020801	17	46022809	0.0024	1.67×10^{-08}	0.0033	9.47×10^{-05}	*PNPO*
cg03128029	2	203143288	-0.0027	1.90×10^{-08}	-0.0036	2.03×10^{-04}	*NOP58*
cg22749855	17	76353952	-0.0024	3.22×10^{-08}	-0.0035	5.15×10^{-05}	*SOCS3*
cg02341197	21	34185927	0.0030	3.92×10^{-08}	0.0045	2.54×10^{-05}	*C21orf62*
cg12269535	6	43142014	-0.0028	4.39×10^{-08}	-0.0046	1.57×10^{-04}	*SRF*
cg25392060	8	142297121	0.0025	5.60×10^{-08}	0.0036	2.15×10^{-04}	*TSNARE1*
cg27184903	15	29285727	0.0024	5.84×10^{-08}	0.0052	4.91×10^{-07}	*APBA2*
cg18663307	21	46341389	0.0029	6.98×10^{-08}	0.0048	1.04×10^{-04}	*ITGB2*
cg09182678	22	50328711	-0.0016	9.02×10^{-08}	-0.0019	1.26×10^{-04}	*DENND6B*

Effect sizes represent the changes in normalized DNA methylation Beta-values per 1-unit increase in natural log-transformed CRP (mg/L). Chr and Position are in GRCh37/hg19. AA denotes African American and EA European Ancestry.

SNPs explained 4.9% of the CRP variance in RS, and the methylation and genetic scores together explained 9.0%. Notably, no significant interaction or association was observed between the genetic and methylation scores, suggesting that they independently explain variance in CRP.

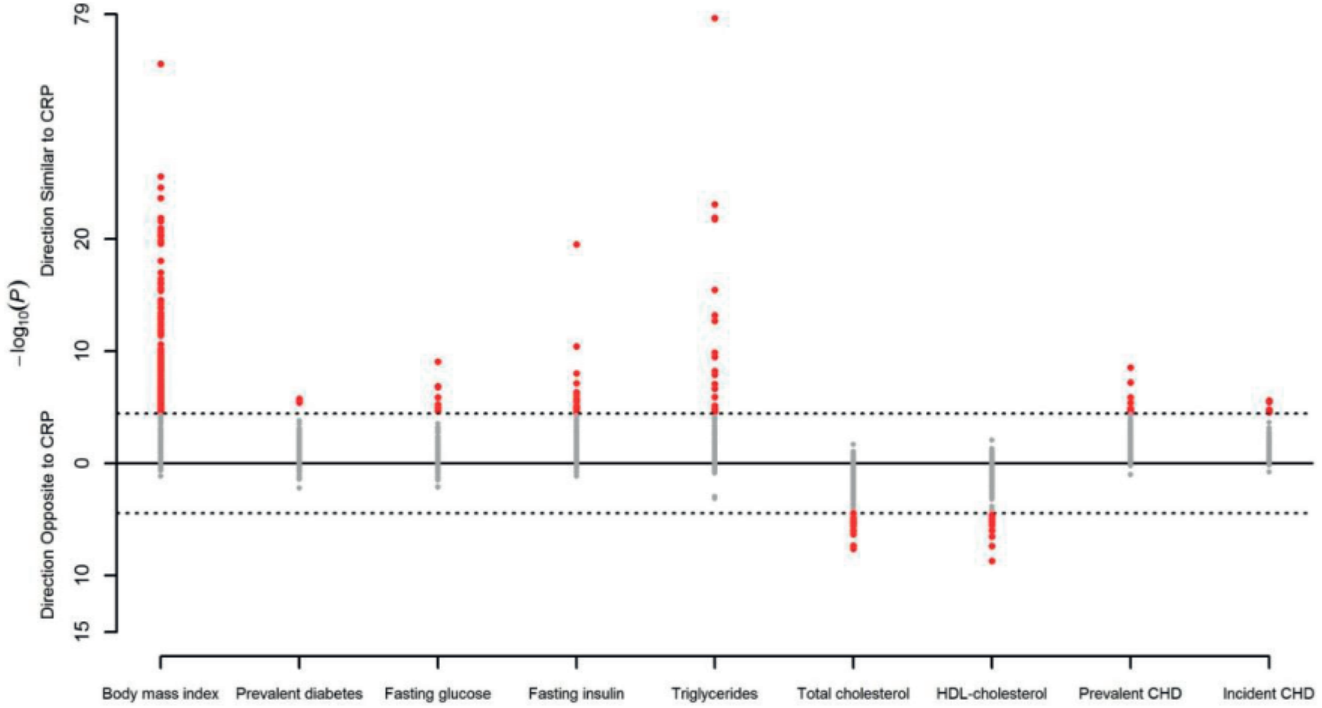
Association with cardiometabolic phenotypes

We examined the associations between the 58 replicated CRP-related CpG sites and nine cardiometabolic traits and diseases (BMI, lipids, glycemic phenotypes, prevalent coronary heart disease, and incident coronary heart disease). After Bonferroni correction for multiple testing based on 58 CpG sites and nine phenotypes (P-value<0.05/522=9.58×10^{-5}), we observed 89 significant associations with 51 unique CpG sites (Additional file 8: Table S7). There was major overlap with BMI (46 CpGs). CpGs that were significantly associated with higher BMI, fasting glucose, fasting insulin, risk of diabetes, triglycerides, and risk of CHD were also associated with higher CRP levels. For HDL-cholesterol and total cholesterol, CpGs were associated with lower CRP levels (Figure 3).

Gene expression analyses

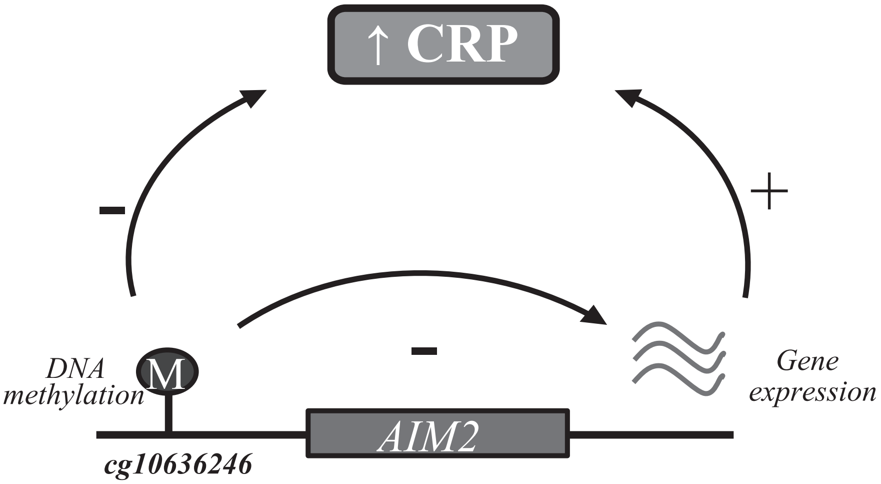
Of the 58 replicated CpG sites, 9 (16%) were significantly associated with expression of 9 unique genes in *cis* (P-value<8.47×10^{-5}) (Additional file 9: Table S8). Furthermore, of those 9 genes, the expression levels of 4 genes were associated with serum CRP levels (P-

Figure 3. Manhattan plot depicting the –log₁₀(P-values) and effect direction (respectively to CRP) of the associations between the 58 replicated CpG sites and each cardiometabolic phenotype. The dotted lines indicate the Bonferroni threshold of 9.58×10^{-5} for significance.



value<0.05). In these 4 cases we could show corresponding triangular relationships between DNA methylation, gene expression, and serum CRP levels. For example, increased methylation at cg10636246 was associated with lower serum CRP levels and lower expression of *AIM2*, and lower expression of *AIM2* was associated with lower CRP levels (Figure 4).

Figure 4. Illustration of the methylation-CRP, methylation-expression, and expression-CRP association for cg10636246 (*AIM2*).



Genetic correlates of DNA methylation in cis

In the RS, we identified 20 *cis*-mQTL pairs (19 unique SNPs and 20 unique CpG sites) for the replicated CpG sites, 10 of these cis-mQTL pairs could be replicated in the FHS (P-value<2.5×10^{-3}) (Additional file 10: Table S9). For example, the strongest correlation was observed between rs12677618 and cg25392060 (located 4,903 base pairs away from each other; β=-0.011; P-value=2.73×10^{-126}). None of the 10 replicated *cis*-mQTL variants was significantly associated with serum CRP levels after Bonferroni correction for multiple testing (P-value>0.005) in the largest published GWAS to date of 66,185 individuals[6].

GWAS catalog, pathway analysis and tissue enrichment

The 58 CpG sites were annotated to 47 genes, which are associated in GWAS with 18 phenotypes (Additional file 11: Table S10). We found enrichment in GWAS of epilepsy, renal cell carcinoma, and lipoprotein-associated phospholipase A2 (Lp-PLA2) activity and mass.

Pathway enrichment analyses were carried out in 47 unique genes that were annotated to the 58 replicated CpG sites in the Ingenuity Pathway Analysis database. The top pathways included growth hormone signaling, IL-9 signaling, atherosclerosis and IL-6 signaling (Additional file 12: Table S11).

Analysis of tissue specific DNase I hotspots yielded enrichment predominantly in epithelium, blood vessel, and various blood cells (especially CD14+ macrophages) (Additional file 6: Table S4).

Discussion

This meta-analysis of epigenome-wide association studies of CRP, a sensitive marker of chronic low-grade inflammation, identified and validated 58 CpG sites in or near 45 unique loci in leukocytes of individuals of European and African descent. The associations were robust to adjustment for potential confounders and explained more than 6% of the variation in circulating CRP concentrations. We demonstrated that several inflammation-related CpG sites were associated with expression of nearby genes, and many CpG sites showed pleiotropic associations with cardiometabolic phenotypes as well as the clinical disease coronary heart disease.

DNA methylation may differ by race or ethnicity[8], challenging replication across individuals of varying descent in epigenetic studies. We were able to replicate up to 27% of our findings with comparable effect estimates, demonstrating that our results are generalizable across Europeans and African-Americans. The trans-ethnic replication approach of our study strengthens the confidence of true-positive findings and supports the notion that despite differing baseline epigenetic profiles, different ethnicities may have consistent epigenetic associations with respect to inflammation.

Increased DNA methylation at the top signal cg10636246 near *AIM2*, was associated with lower expression of *AIM2* and lower CRP levels. In agreement, lower *AIM2* expression was associated with lower serum CRP levels. As an inflammasome receptor for double stranded DNA activating inflammatory cascades, *AIM2* is implicated in host defense mechanisms against bacterial and viral pathogens, and thus is key in the human innate immune response[20,21]. The data suggest that methylation near *AIM2* play a role in low-grade inflammation in the general population. Nevertheless, the results from the current study do not infer causal directionality.

Several of our hits were associated with future clinical events. For example, three inflammation-related CpG sites were also associated with incident CHD. Hypomethylation at cg18181703 (*SOCS3)*, cg06126421 (*TUBB*), and cg05575921 (*AHRR*) were associated with higher CRP levels and increased risk of future CHD. The gene product of *SOCS3,* suppressor of cytokine signaling 3, plays a pivotal role in the innate immune system as a regulator of cytokine signaling[22]. The role of *SOCS3* in atherosclerosis has been established[23]. We observed that lower DNA methylation was associated with increased expression of *SOCS3* and increased serum CRP. Differential methylation at the *AHRR* loci has been robustly demonstrated to be associated with cigarette smoking[24]. The association of *AHRR* methylation with CRP and incident CHD may highlight a connection between CRP and cardiovascular disease that is shared between cigarette smoking and independent mechanisms. Furthermore, we found two CpG sites that have recently been identified in an EWAS of incident type 2 diabetes[25]. We hypothesize that inflammation-related epigenetic features may explain at least part of the observed associations between CRP, a sensitive marker of chronic low-grade inflammation, and related clinical events including CHD and diabetes.

Many replicated CpG sites demonstrated associations with cardiometabolic phenotypes, emphasizing the substantial epigenetic overlap with those phenotypes. Taken together, these pleiotropic epigenetic associations across various phenotypes may provide novel insights into shared epigenetic mechanisms and provide opportunities to link chronic low-grade inflammation and cardiometabolic phenotypes. Our findings may help to focus on genomic regulation of pertinent loci that may be attractive targets for perturbation or therapeutic intervention.

CRP is affected by both genetic and environmental factors[15]. Although we may have slightly overestimated the variance explained since the testing cohorts participated in the discovery and replication meta-analysis, the CRP methylation score augmented the explained variance beyond that accounted for by the CRP genetic score. This suggests that the methylation score harbors information that may be independent from the genetic factors underlying CRP. In agreement with a previous report on the added value of a methylation score in explaining variance in BMI, we further add that methylation may explain further variation of complex traits that have substantial environmental components[26].

In the present study, we were able to present stringent triangular relationships between DNA methylation, gene expression, and serum CRP levels at four loci. However, firm conclusions regarding causal directionality are challenging in epigenetic studies. Although ten (17%) of the replicated methylation sites had *cis*-mQTLs, we were not able to detect a significant association between these mQTLs and CRP levels in the largest published CRP GWAS, which may be due to the limited power, or the findings represent methylation changes downstream of CRP. However, our findings were biologically plausible and consistent with previous observations. For example, GWAS enrichment analysis suggested enrichment in genes identified for renal cell carcinoma. CRP is commonly elevated in renal cell carcinoma patients[27]. Furthermore, pathway analyses identified regulatory mechanisms related to inflammatory processes such as STAT3 and IL-6 signaling pathway, the pro-inflammatory upstream regulator of serum CRP levels[28]. Taken together, these results suggest that DNA methylation plays a role in establishing or maintaining CRP levels in the general population.

The major strengths of the present study are its large sample size and multi-ethnic nature, allowing a valid interpretation of results for both European and African-American populations. Furthermore, careful and comprehensive adjusting models reduced the chance of confounding. In addition, DNA methylation was quantified in whole blood, which is primarily composed of leukocytes, a key component of the human immune system and therefore highly relevant to systemic inflammation. The combination of epigenomics with genomics and transcriptomics data as well as enrichment analyses allowed the exploration of functional properties of our findings.

The study has limitations. The 450K array captures approximately 2-4% of the total human DNA methylation, mainly in genic regions, thus limits the discovery of potentially important CpG sites that are not measured on the array. Furthermore, although we adjusted the analyses for measured or estimated cell type proportions, we cannot completely rule out the presence of residual confounding by white blood cell distributions. Residual confounding from differences in unmeasured cell count heterogeneity introduced by correlation between CRP and unknown cell subtypes may bias our results. Also, the annotation of CpGs and SNPs to genes is challenging in genomic studies. We annotated primarily based on distances, which may have incorrectly annotated genes. Further, we replicated our findings from the European discovery in African-Americans. The differences in ethnicities and the African-American sample size may have limited replication of the findings. Our study was limited to blood samples and while this has been demonstrated to be a good surrogate tissue[29], we would not be able to infer tissue specific methylation changes. Specifically, as CRP is synthesized in the liver, our current study design would not allow us to detect hepatic methylation changes. We did not observe associations with nearby gene expression for all CpGs we identified. However, the limited sample size for methylation-expression analyses, failure for expression probes to pass quality control, tissue-specificity, and long-distance effects may explain this observation. Furthermore, DNA methylation may also affect chromosome stability and alternative splicing, two functional consequences of DNA methylation which we have not investigated in the present study. Finally, we cannot exclude residual confounding, and cannot determine causal directionality.

We performed the first meta-analysis of epigenome-wide association studies of CRP, a sensitive marker of low-grade inflammation. We identified 58 DNA methylation sites that are significantly associated with CRP levels in individuals of both European and African-American ancestry. Since inflammation is implicated in the development of multiple complex diseases, the discoveries from the current study may contribute to the identification of novel therapies and interventions for treatment of inflammation and its clinical consequences.

References

1. Hansson GK, Hermansson A. The immune system in atherosclerosis. *Nature immunology* 2011; 12(3): 204-12.

2. Donath MY, Shoelson SE. Type 2 diabetes as an inflammatory disease. *Nature Reviews Immunology* 2011; 11(2): 98-107.

3. Pepys MB. The acute phase response and C-reactive protein. *Oxford textbook of medicine* 1995; 2: 1527-33.

4. Emerging Risk Factors Collaboration. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *The Lancet* 2010; 375(9709): 132-40.

5. Xu H, Barnes GT, Yang Q, et al. Chronic inflammation in fat plays a crucial role in the development of obesity-related insulin resistance. *Journal of Clinical Investigation* 2003; 112(12): 1821.

6. Dehghan A, Dupuis J, Barbalic M, 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): 731-8.

7. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* 2012; 13(7): 484-92.

8. Barfield RT, Almli LM, Kilaru V, et al. Accounting for population stratification in DNA methylation studies. *Genetic epidemiology* 2014; 38(3): 231-41.

9. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation: Cardiovascular Genetics* 2009; 2(1): 73-80.

10. Chen Y-a, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpG sites in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013; 8(2): 203-9.

11. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* 2012; 13(1): 86.

12. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014; 30(10): 1363-9.

13. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; 26(17): 2190-1.

14. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 2009; 41(4): 1149-60.

15. Schnabel RB, Lunetta KL, Larson MG, et al. The relation of genetic and environmental factors to systemic inflammatory biomarker concentrations. *Circulation: Cardiovascular Genetics* 2009; 2(3): 229-37.

16. Parker HS, Leek JT, Favorov AV, et al. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* 2014; 30(19): 2757-63.

17. Westra H-J, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics* 2013; 45(10): 1238-43.

18. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 2014; 42(D1): D1001-D6.

19. Breeze C, Paul D, Butcher L, et al. eFORGE: A tool for identifying tissue-specific signal in epigenomic data. *Manuscript in preparation.*

20. Hornung V, Ablasser A, Charrel-Dennis M, et al. AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. *Nature* 2009; 458(7237): 514-8.

21. Martinon F, Tschopp J. Inflammatory caspases and inflammasomes: master switches of inflammation. *Cell Death & Differentiation* 2007; 14(1): 10-22.

22. Carow B, Rottenberg ME. SOCS3, a major regulator of infection and inflammation. *Frontiers in immunology* 2014; 5.

23. Ortiz-Munoz G, Martin-Ventura JL, Hernandez-Vargas P, et al. Suppressors of cytokine signaling modulate JAK/STAT-mediated cell responses during atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology* 2009; 29(4): 525-31.

24. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clinical epigenetics* 2015; 7(1): 1-10.

25. Chambers JC, Loh M, Lehne B, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *The Lancet Diabetes & Endocrinology* 2015; 3(7): 526-34.

26. Shah S, Bonder MJ, Marioni RE, et al. Improving phenotypic prediction by combining genetic and epigenetic associations. *The American Journal of Human Genetics* 2015; 97(1): 75-85.

27. Jabs WJ, Busse M, Krüger S, Jocham D, Steinhoff J, Doehn C. Expression of C-reactive protein by renal cell carcinomas and unaffected surrounding renal tissue. *Kidney international* 2005; 68(5): 2103-10.

28. Kishimoto T, Akira S, Narazaki M, Taga T. Interleukin-6 family of cytokines and gp130. *Blood* 1995; 86(4): 1243-54.

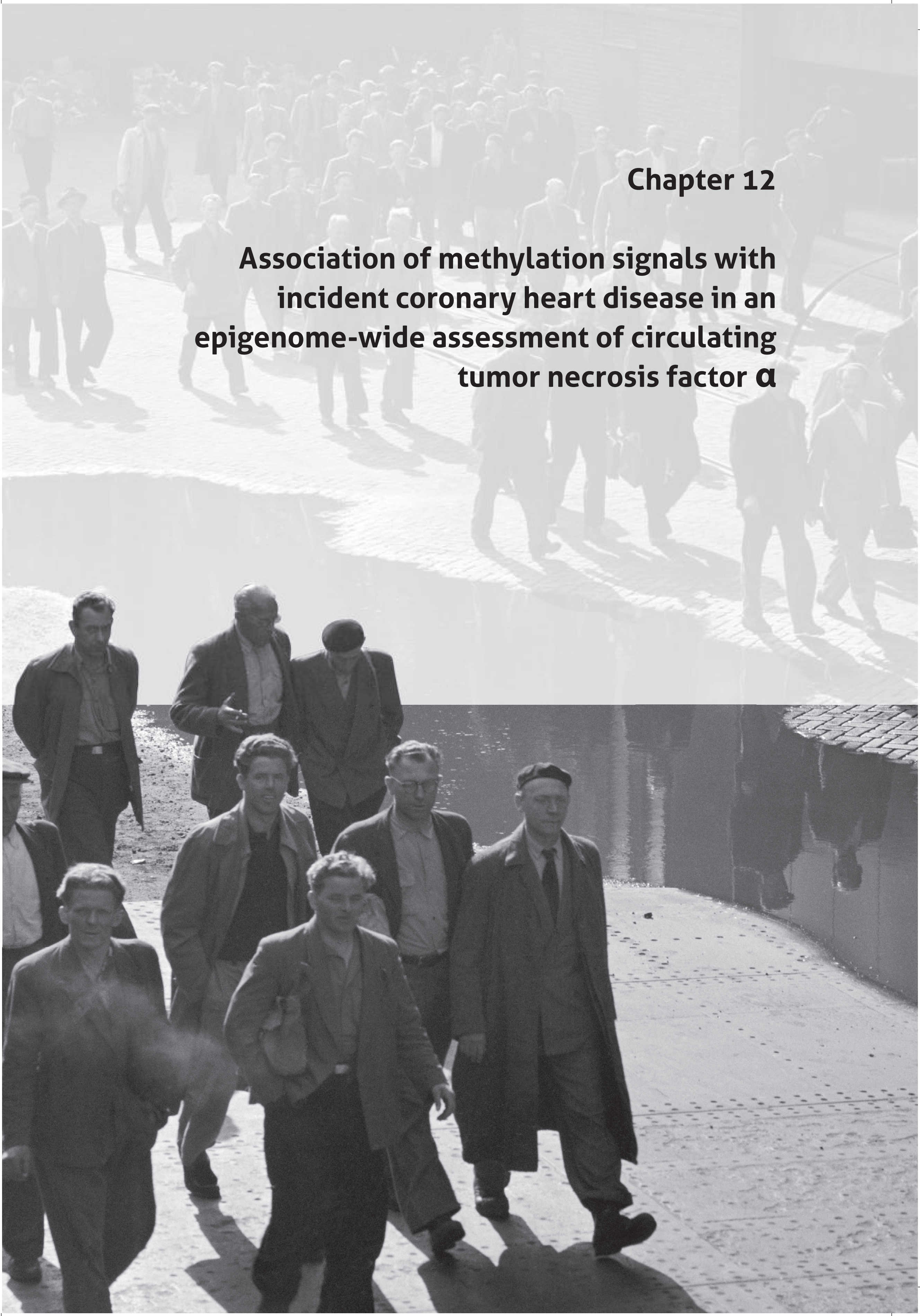
29. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. *The Lancet* 2014; 383(9933): 1990-8.

Supplementary material

The supplementary material of this manuscript can found at the following webpage: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1119-5.

Chapter 12

Association of methylation signals with incident coronary heart disease in an epigenome-wide assessment of circulating tumor necrosis factor α



Background: Tumor necrosis factor alpha (TNFα) is a proinflammatory cytokine with manifold consequences for mammalian pathophysiology, including cardiovascular disease. Strategies for therapeutic inhibition of TNFα have produced mixed results, necessitating a deeper understanding of TNFα biology to enhance treatment precision.

Methods: We performed a discovery meta-analysis (n=4,163) of epigenome-wide associations with cross-sectional circulating TNFα from five studies with external replication (n=667). Follow-up analyses investigated associations of identified methylation loci with gene expression and incident coronary heart disease (n=11,461 with 1,895 events).

Results: In the discovery stage, circulating TNFα levels were associated with methylation of seven cytosine-phosphate-guanine (CpG) sites (P-value≤$2.24\times10^{-7)}$, located in or near *DTX3L-PARP9* (cg00959259, cg08122652, cg22930808), *NLRC5* (cg16411857, cg07839457), and *ABO* (cg13683939, cg24267699) after accounting for multiple testing. Of those, negative associations between TNFα and the methylation of two loci in *NLRC5* (cg16411857 and cg07839457) and one in *DTX3L-PARP9* (cg08122652) externally replicated (P≤0.003). Methylation at the replicated TNFα loci was negatively associated with neighboring gene expression in two of the three participating cohorts; in turn, expression of *NLRC5*, *DTX3L*, and *PARP9* was strongly (P-value≤0.003) positively associated with TNFα. Methylation of cg07839457 in *NLRC5* was weakly associated with neighboring sequence variant on chromosome 16 (rs17369768), located at a transcriptionally active region in multiple tissues, and nominally associated with metabolic traits (visceral adipose tissue volume, waist circumference, weight) and inflammatory conditions (psoriasis, rheumatoid arthritis) in external databases. All replicated TNFα-related CpGs were associated with a significant reduction in the risk of incident CHD (9-19% decreased risk per 10% higher methylation per CpG, P-value≤0.003).

Conclusion: We identified and replicated novel epigenetic correlates of circulating TNFα in blood samples and linked these loci to CHD risk, opening opportunities for validation and therapeutic applications.

Introduction

Tumor necrosis factor alpha (TNFα) is a proinflammatory cytokine with pleiotropic effects in human health and disease. In addition to its well-characterized pathogenic contributions to inflammatory and autoimmune diseases, atherosclerosis, type 2 diabetes, and cancer, TNFα also plays a key homeostatic role in pathogen defense, tissue repair and regeneration, and organ development (reviewed in Kalliolias, et al.[1]). Therapeutic inhibition of TNFα is used in clinical settings with both successes (e.g. in various forms of autoimmune diseases) and failures (e.g. in multiple sclerosis[2]). Furthermore, treatment with TNF inhibitors has long been known to lower the risk of cardiovascular disease among autoimmune disease patients[3], and currently several trials (e.g. NCT01893996) are assessing cardioprotective effects of inhibiting inflammatory cytokines. Recently, the randomized placebo-controlled CANTOS trial[4] (NCT01327846) reported significant reductions in recurrent cardiovascular risk due to interleukin-1β inhibition, achieved independently of changes in lipid levels. While such findings highlight the clinical promise of targeting systemic inflammation in the setting of cardiovascular disease, the underlying mechanisms of action remain elusive.

Circulating levels of TNFα have a moderate genetic determinant, with heritability estimates ranging from 17%[5] to 39%[6] in large-scale European twin studies to 68% in a Ugandan community with a high prevalence of tuberculosis[7]. Known common mutations account for a small fraction of that heritable component, explaining <4% of TNFα variance in a recent meta-analysis of genome-wide association studies (personal correspondence). Emerging evidence suggests that epigenetic processes like DNA methylation, which reflect changes in gene expression that occur without sequence mutations, may offer promising clues in the search for missing TNFα heritability. For example, methylation of cytosine-phosphate-guanine (CpG) loci in the *TNF* promoter was associated with *TNFα* expression and plasma TNFα levels in several population-based studies[8,9]. *In vitro*, experimental manipulation of DNA methylation has been shown to alter the cells' ability to produce TNFα[10], offering causal support for the association observed in population studies. To date, however, no study has comprehensively examined the DNA methylation across the entire genome in relation to circulating levels of TNFα in large human populations, or has interrogated TNFα epigenetics with regards to cardiovascular risk.

Therefore, we conducted the first epigenome-wide meta-analysis of associations between circulating TNFα levels and DNA methylation in whole blood samples or isolated lymphocytes from 4,163 individuals in the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium. We subsequently achieved replication of the top CpG loci in an independent population, evaluated the associations between DNA methylation and *cis*-gene expression, and assessed genotype contributions to the observed CpG methylation variation in the regions of interest. Finally, we investigated the association of the top epigenetic correlates of circulating TNFα with incident coronary heart disease (CHD) in a meta-analysis comprising 11,461 participants with 1,895 CHD events.

Methods

Discovery and Replication Populations

In the discovery phase, the epigenome-wide study included individuals of European descent from six studies participating in the CHARGE consortium[11]: Framingham Heart Study (FHS), Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study, The Invecchiare in Chianti Study (InCHIANTI), Kooperative Gesundheitsforschung in der Region Augsburg (KORA), Lothian Birth Cohort 1921 (LBC1921), and Normative Aging Study (NAS). Two Finnish cohorts, the Northern Finland Birth Cohort 1966 (NFBC66) and the Helsinki Birth Cohort Study (HBCS), were designated for replication. Individuals who reported an autoimmune diagnosis or taking immune-modulating agents (e.g. TNFα blockers) were not included in the analyses. Further details about each study are included in Table 1 and eMethods 1. Notably, NAS and HBCS were excluded from the main analysis based on the extreme variability in TNFα measurements (Table 1), which could be due to previously reported poor performance of multiplexed assays compared to ELISA.[12] All study protocols were approved by Institutional Review Boards of the participating study sites, and all participants provided written informed consent.

Laboratory Measurements

Circulating TNFα levels were measured in pg/ml using the approaches listed in Table 1. In all but one cohort (FHS), TNFα was measured at the same time and center visit as blood was drawn for the quantification of DNA methylation. In FHS, TNFα was measured in the same individuals approximately seven years prior to the DNA methylation assay. Circulating TNFα was natural log-transformed (lnTNFα) to reduce skewness of the distribution. Individuals whose lnTNFα measurements were more than four standard deviations away from the cohort mean were excluded from subsequent analyses.

DNA Methylation Measurements, Normalization, and Quality Control

All studies used the Illumina Infinium Human Methylation450 Beadchip (Illumina Inc, San Diego, CA) to quantify epigenome-wide DNA methylation. In all studies but one, these measurements were performed on DNA extracted from whole blood samples; the GOLDN study isolated and quantified DNA methylation on CD4+ T-cells. Study-specific approaches to methylation data processing are summarized in eTable 1.

Statistical Analyses

In the discovery phase, each cohort fit three linear mixed effect regression models to assess associations between lnTNFα (predictor) and normalized methylation β scores (outcomes). The base model adjusted for age and sex as fixed effects and technical covariates (array, row, and/or column number) as a random effect. The second model additionally adjusted for white blood cells (WBC) subtypes for studies reporting methylation in whole blood samples. The third model adjusted for the same covariates as the second model plus smoking (current, former, or never) and body mass index (BMI) in kg/m^2. All covariates were selected based on their known associations with DNA methylation. Cohorts additionally adjusted for relatedness or other study-specific covariates as necessary (eTable 2). Results from the five cohorts participating in the main discovery analysis were meta-analyzed using a fixed effects, inverse-variance weighted approach in METASOFT[13]. Because GOLDN was the only cohort that used CD4+ T-cells and not whole blood samples, we ran a sensitivity meta-analysis excluding GOLDN. We performed additional sensitivity analyses including NAS and HBCS, cohorts with extremely variable Milliplex-measured TNFα.

To maximize statistical power of discovery, we carried CpGs forward to the replication phase if the false discovery rate (FDR) for the specific CpG was below 0.05. Models used in the replication analysis were identical to those implemented in the discovery phase. To minimize the chance of false positives, we implemented the more stringent Bonferroni correction in the replication phase: 0.05/number of statistically significant hits from the discovery meta-analysis.

Gene Expression Measurements and Analysis

The CpG sites that significantly replicated in the independent replication sample were further tested for association with *cis*-gene expression in whole blood in 3,738 participants with available gene expression measurements: FHS, KORA, and RS. Methods for the expression measurements and analysis are described in eMethods 2. mRNA transcripts that achieved statistical significance in at least two cohorts were further evaluated for association with circulating TNFα in FHS, using regression models adjusted for age, sex, imputed WBC counts, smoking, BMI, and technical covariates.

Integrating DNA Methylation and Sequence Data

To establish genetic contributions to the observed methylation of the top loci, we studied genetic associations with DNA methylation in *cis* (±20kb) using the GOLDN study as the discovery cohort and RS as the replication population. Genotyping, imputation procedures, and statistical analysis for both cohorts are described in eMethods 3. The variants that achieved nominal significance in the replication phase were tested for association with circulating TNFα in GOLDN using linear mixed models adjusting for age, sex, study site (fixed effects) and family (random effect).

In addition to meQTL analyses, we searched for overlap with the genomic regions containing the replicated sites in two genome-wide association study (GWAS) catalogs (http://www.ebi.ac.uk/gwas/search, accessed 25-4-2017, and http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner, accessed 7-8-2017) to assess previously reported associations of sequence variants in the regions of interest and disease traits.

Associations with Incident CHD

We tested associations between the replicated epigenetic correlates of circulating TNFα and incident CHD in a CHARGE consortium fixed effects meta-analysis that included 470,346 CpGs, 1,895 disease events, and 11,641 participants from the following cohorts: Atherosclerosis Risk in Communities, Cardiovascular Health Study, Long-term Follow-up of Antithrombotic Management Patterns In Acute Coronary Syndrome Patients, FHS, InCHIANTI, KORA, NAS, and Women's Health Initiative. The definition of CHD events included coronary insufficiency, coronary revascularization, recognized MI (hospitalization with diagnostic ECG changes and/or biomarkers of MI), and coronary death. Participants with prevalent CHD at enrollment were excluded. Each cohort study obtained written informed consent from participants and ethics approval from its respective institutional review boards and ethics committees. In each cohort, associations were adjusted for age, sex, smoking status, education, BMI, differential WBC counts (either directly measured or imputed), and technical covariates. Data were meta-analyzed using an inverse-variance weighted fixed effects method. This lookup was restricted to the top four CpG sites from the circulating TNFα meta-analysis; findings were considered statistically significant if P-value<0.05/4=0.0125. Further details about the incident CHD meta-analysis[14] are available in eMethods 4.

Functional Annotation

We used Hudson Alpha Institute for Biotechnology - ENCODE project custom methylation tracks implemented in the UCSC genome browser as well as the Illumina annotation file to visualize and annotate the functional potential of the top CpGs, including such indicators of regulatory activity as H3K27Ac marks, DNAseI hypersensitivity elements in relevant cell types, and genomic location of the CpGs (promoter vs gene body, exon vs intron, etc).

Table 1. Characteristics of the participating cohorts.

Study	N	Country	Mean Age, years±SD	Female, %	Mean TNFα pg/ml±SD	TNFα Assay	Coefficient of variation of TNFα measurements
Discovery Phase							
FHS	1730	USA	67.0±9.0	52	1.4±1.2	ELISA (R&D Systems)	Intra 6-8%, inter 5-11%
GOLDN	970	USA	47.8±16.3	52	3.2±1.7	ELISA (R&D Systems)	Intra 5%, inter 10%
InCHIANTI	498	Italy	62.8±15.8	55	4.3±2.2	ELISA (R&D Systems)	Intra 7%, inter <21%
KORA	800	Germany	69.0±4.3	49	2.5±2.2	ELISA (R&D Systems)	Intra 6%, inter 14%
LBC1921	165	UK	86.7±0.4	54	1.5±1.6	Immunonepheometry (Dade-Behring)	Not available
NAS[a]	631	USA	74.6±6.8	0	55.0±155.8	Milliplex Human Cytokine/ Chemokine Panel (EMD Millipore)	Not available
Replication Phase							
NFBC66	667	Finland	31.0±0.3	56	7.8±9.0	ELISA (Merck)	Intra 3%, inter 7%
HBCS[a]	149	Finland	63.3±2.7	0	16.7±47.6	Milliplex Map Human Metabolic Hormone Panel Kit (HMH-34K)	Not available

Abbreviations: FHS, Framingham Heart Study; GOLDN, Genetics of Lipid Lowering Drugs and Diet Network; HBCS, Helsinki Birth Cohort Study; InCHIANTI, Invecchiare in Chianti Study; KORA, Kooperative Gesundheitsforschung in der Region Augsburg Study; LBC1921, Lothian Birth Cohort 1921; NAS, Normative Aging Study; NFBC66, Northern Finland Birth Cohort 1966; TNFα, tumor necrosis factor α.

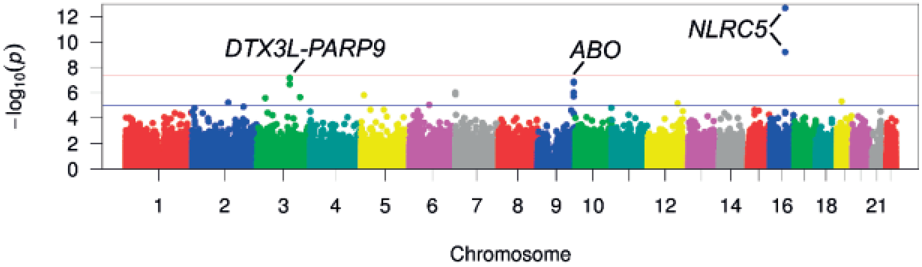
[a] Excluded from the analyses due to concerns over reliability of TNFα measurements.

Results

General Characteristics

Demographic and TNFα characteristics of all participating cohorts are summarized in Table 1. The cohorts varied in age, ranging from the mean age of 31(±0.3) years in NFBC66 to 87(±0.4) years in the LBC1921 cohort. The cohorts were approximately half female with the exception of NAS and HBCS, which only recruited men. NAS and HBCS were also unique in reporting exceptionally high and variable concentrations of TNFα compared to all other cohorts. The observed discrepancy may be due to technical differences between approaches to TNFα quantification (Milliplex vs. ELISA), as documented in Table 1.

Figure 1. Epigenome-wide associations between DNA methylation and circulating TNFa in the discovery stage.



This graph includes data of 4,163 participants in the discovery analyses. The red horizontal line denotes the FDR=0.05 threshold for statistical significance.

Meta-analysis, Replication, and Sensitivity Analyses

The three models fit to test associations between epigenome-wide methylation and TNFα levels yielded similar results (Table 2, eTables 3 and 4); namely, the cg16411857 site in NLRC5 emerged as the top hit with all approaches. Based prior evidence in support of adjusting epigenetic models for smoking and BMI[15], all subsequent analyses focused on the third model.

The results of the epigenome-wide analyses are summarized in Table 2 and visualized in Figure 1 and eFigure 1. Seven CpG sites located in three genomic regions emerged as the top hits in the discovery stage (FDR<0.05). Of those, three CpG sites—two located in the NLR family CARD domain containing 5 gene (*NLRC5*) and one in the deltex E3 ubiquitin ligase 3L/ poly(ADP-ribose) polymerase family member 9 gene complex (*DTX3L/PARP9*)—replicated in 677 individuals from the NFBC66 cohort. All replicated associations were in the same direction and had comparable effect sizes. Another locus in *DTX3L/PARP9* was not able to undergo replication due to failing quality control procedures in NFBC66; however, due to its proximity and similarity of associations to the replicated cg08122652, we included it in subsequent analyses.

Inclusion of the two cohorts that measured TNFα using the Milliplex method, namely NAS in discovery (eTable 5) and HBCS in replication, increased genomic inflation by 19% (λ=0.94 in the main analysis vs. λ=1.13 with NAS included). While the *NLRC5* and *DTX3L/PARP9* hits still emerged as significant in the discovery stage, none replicated in HBCS. Excluding GOLDN from the discovery stage due to differences in methylation measurements (CD4+ T-cells vs. whole blood, eTable 6) yielded five of the seven significant hits observed in the main analysis, including the *NLRC5* and *DTX3L/PARP9* loci.

Table 2. Associations of methylation sites and circulating TNFα.[a]

		Discovery			*Replication*	
CpG site	**Gene**	**β±SE**	**P-value**	**I^2**	**β±SE**	**P-value**
cg16411857	*NLRC5*	-0.01±0.002	2.14×10^{-13}	15%	-0.009±0.003	0.003
cg07839457	*NLRC5*	-0.02±0.003	6.31×10^{-10}	70%	-0.01±0.004	0.0003
cg00959259	*DTX3L; PARP9*	-0.01±0.003	7.36×10^{-8}	56%	NA	NA
cg22930808	*DTX3L; PARP9*	-0.01±0.002	6.92×10^{-8}	58%	-0.008±0.004	0.04
cg13683939	Intergenic; proximal to *ABO*	0.04±0.008	1.42×10^{-7}	71%	-0.02±0.01	0.18
cg24267699	*ABO*	-0.009±0.002	1.67×10^{-7}	89%	0.002±0.002	0.47
cg08122652	*DTX3L; PARP9*	-0.008±0.002	2.24×10^{-7}	78%	-0.007±0.002	0.003

Abbreviations: β, regression coefficient; CpG, cytosine-phosphate-guanine; I^2, heterogeneity statistic; standard error; TNFα, tumor necrosis factor α.

[a] Model adjusted for age, sex, white blood cell proportions, technical covariates, smoking, and BMI with FDR < 0.05 in the discovery stage. Results that met the Bonferroni threshold in the replication stage (P-value=0.05/7=0.007) are in bold. [b] hg19.

Methylation vs. Expression vs. Circulating TNFα

We observed nine cis-eQTMs (methylation-expression pairs) between methylation at the four TNFα-associated loci and cis-gene expression in FHS. All were robust in RS, while none reached significance in KORA. CpG-transcript pairs that satisfied the Bonferroni threshold in at least two cohorts are presented in Table 3. Direction (negative for all transcripts except the methylation with karyopherin subunit alpha 1 gene (*KPNA1*) pair) and magnitude of associations were consistent between FHS and RS. Of the five transcripts that were significantly associated with TNFα-linked CpGs in both FHS and RS (*NLRC5*, *DTX3L*, *KPNA1*, *PARP9*, and the poly(ADP-ribose) polymerase family member 14 gene (*PARP14*)), three were positively associated with circulating TNFα in FHS, with respective P of 5.47×10^{-5} (*NLRC5*), 0.003 (*DTX3L*), and 0.003 (*PARP14*) meeting the Bonferroni threshold.

Table 3. Associations between methylation status of top TNFα CpG sites and neighboring gene expression.

CpG site	Transcript	P-value in RS (n=750)	P-value in FHS (n=2,262)	P-value in KORA (n=726)	Direction of Association
cg16411857	*NLRC5*	0.0002	2.56×10^{-8}	0.16	---
cg07839457	*NLRC5*	2.10×10^{-7}	1.85×10^{-8}	0.44	---
cg00959259	*DTX3L*	1.07×10^{-6}	1.80×10^{-9}	0.80	--+
	PARP9	2.86×10^{-22}	2.58×10^{-13}	0.81/0.07[a]	----
	PARP14	9.21×10^{-23}	6.51×10^{-17}	n/a	--?
	DTX3L	2.91×10^{-7}	2.13×10^{-6}	0.61	---
	KPNA1	0.0003	0.00004	n/a	++?
cg08122652	*PARP9*	1.04×10^{-25}	1.09×10^{-9}	0.12/0.26[a]	----
	PARP14	1.25×10^{-24}	8.48×10^{-15}	n/a	--?

Abbreviations: CpG, cytosine-phosphate-guanine; FHS, Framingham Heart Study; KORA, Kooperative Gesundheitsforschung in der Region Augsburg Study; TNFα, tumor necrosis factor α. [a]Two probes (ILMN_1731224 and ILMN_2053527) corresponded to *PARP9* in KORA.

Methylation vs. Sequence Variation vs. Circulating TNFα

Of all significant methylation correlates of TNFα, only cg07839457 showed nominally significant (P-value<0.05) replication of associations with two neighboring *NLRC5* sequence variants: rs17369768 and a deletion at the 57042641 position on chromosome 16 (eTable7). Neither loci were significantly associated with circulating TNFα in GOLDN (P-value=0.60 and P-value=0.61, respectively). However, rs17369768 was nominally associated with visceral adipose tissue volume, waist circumference, weight, psoriasis, and rheumatoid arthritis in public databases (http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner).

Associations with Incident CHD

Methylation at all four TNFα-associated loci was robustly negatively associated with the risk of incident CHD in the meta-analysis of CHARGE cohorts (Table 4, eFigure 2). Adjusted for the appropriate covariates, each 10% increase in methylation of a given TNFα-associated locus was associated with a 9% to 19% decrease in the risk of an adverse CHD event.

GWAS Catalog Look-Up and Functional Annotation

Of the three common SNPs in or near *NLRC5* that were reported in the GWAS catalog, two (rs821470 and rs17290922) were associated with schizophrenia-related phenotypes[16,17]. The closest reported variant to the *DTX3L/PARP9* locus (rs2173763) was associated with major depressive disorder[18]. Conversely, a search for SNPs previously reported to be associated with circulating TNFα yielded no results located in the regions harboring the replicated epigenetic hits, although an *ABO* polymorphism (from a region that emerged as a top hit yet did not replicate) was identified as a TNFα protein quantitative trait locus in an earlier analysis of KORA data[19].

Bioinformatic regulatory annotations for the *DTX3L/PARP9* and *NLRC5* regions are presented in eFigures 3 and 4, respectively. Both sets of loci are adjacent to or overlap regulatory elements, supporting observed associations with gene expression.

Table 4. Associations between incident CHD and methylation status of top TNFα CpG sites in a meta-analysis of 8 cohorts with 1,895 disease events and 11,641 participants.

CpG site	Chr	Position[a]	Gene	HR (95% CI)	P-value
cg16411857	16	57023191	*NLRC5*	0.86 (0.78, 0.95)	0.003
cg07839457	16	57023022	*NLRC5*	0.89 (0.80, 0.94)	3.1×10^{-5}
cg00959259	3	122281975	*DTX3L;PARP9*	0.91 (0.84, 0.97)	0.002
cg08122652	3	122281939	*DTX3L;PARP9*	0.81 (0.74, 0.89)	2.0×10^{-5}

Abbreviations: β, regression coefficient; Chr, chromosome; CI, confidence interval; CpG, cytosine-phosphate-guanine; HR, hazard ratio per 10% increase in methylation; TNFα, tumor necrosis factor α. [a]hg19.

Discussion

Using epigenome-wide data from adult participants of European descent, we have identified and replicated novel associations between leukocyte DNA methylation loci in two genomic regions mapping to *NLRC5* and *DTX3L/PARP9*, the expression of corresponding genes, and circulating TNFα. Most notably, DNA methylation at the same loci that were correlated with lower plasma TNFα levels was also associated with a substantial reduction in the risk of incident CHD in a multi-ethnic, well-powered meta-analysis.

Both genomic regions that were discovered and validated in our analysis encode proteins that play a pivotal role in the immune response. *NLRC5* is a specific transactivator of major histocompatibility complex (MHC) class I genes[20], which encode human leukocyte antigens (HLA) proteins that set off the adaptive immune reaction[21], These processes are induced chiefly by interferon-gamma (IFNγ) stimulation, although also by toll-like receptor ligands, other interferons, and viral infections[22]. By activating CD8+ T-cells via MHC class I proteins, NLRC5 has also been shown to upregulate IFNγ, creating a positive feedback loop that ensures an effective response to intracellular pathogens[23].

The role of *NLRC5* as a master regulator of the immune response, combined with its remarkable specificity, has positioned it as a promising therapeutic target in multiple clinical settings. The specific methylation loci that emerged as our top findings, cg16411857 and cg07839457, have been shown to be significantly hypomethylated in blood from immune-suppressed HIV-infected individuals, also correlating negatively with viral load[24]. In another whole blood DNA methylation study, both CpGs were linked to circulating IL-18, offering a

12

possible mechanism for the association we observed with CHD incidence[24]. Of clinical interest, the *NLRC5* promoter (and specifically the cg16411857 locus) was shown to be hypermethylated in 13 distinct cancer types, with a corresponding reduction in expression of not only *NLRC5* but also of other genes in the MHC class I family—providing a mechanism for evasion of CD8+ T-lymphocyte antitumor activity[25]. Therefore, our study adds to the robust body of evidence in support of *NLRC5* involvement in a wide range of pathophysiologic conditions.

Similarly to *NLRC5*, increased expression of *DTX3L-PARP9* has been shown to enhance IFNγ signaling and therefore host immune response[26]. Recent evidence suggests that DTX3L-PARP9 may also play a key role in vascular inflammation and atherosclerosis. In macrophage-like cell lines stimulated with IFNγ, experimental silencing of *PARP9* has suppressed the induction of TNFα (consistently with the directions of association observed in our analyses) while silencing of *PARP14* has had opposite effects (in contrast with our observations); additionally, PARP14 deficiency was shown to promote atherogenesis in mice[27]. Possible explanations for the discrepancy in the direction of association may include cell type (macrophages vs. T-lymphocytes or whole blood), tightly controlled experimental conditions in cell culture/murine models vs. observational data from free-living humans, chance, or other factors. Therapeutic inhibition of other PARP enzymes—specifically PARP1—has also been shown to confer cardioprotective effects[28] as well as to reduce circulating TNFα *in vivo*[29]. Although the inconsistency of the PARP14 finding across studies merits close attention in future investigations, our analysis contributes to growing evidence linking PARP enzymes with systemic inflammation and CHD[27].

In follow-up analyses, we found only limited evidence of genotype contributions to the methylation of the CpG sites of interest, suggesting the importance of environmental determinants. A prior analysis of the GOLDN study reported moderate heritabilities for the top loci associated with TNFα in our analysis, with some of them (e.g. cg07839457) likely to be enriched in the genomic regions that evade erasure during embryogenesis[30]. It is therefore possible that the methylation of loci like cg07839457 in *NLRC5* could be programmed by environmental exposures (notably pathogens) and transmitted across generations, although further targeted studies are needed to rigorously test this hypothesis.

To date, the presented analysis is the largest epigenetic study of circulating TNFα, both in sample size and scope. Previously, a number of studies have interrogated relationships between methylation in the tumor necrosis factor gene (*TNF*) gene promoter, corresponding gene expression (where available), and circulating TNFα levels in various disease contexts, e.g. rheumatoid arthritis, chronic periodontitis[31,32], type 1 diabetes[33], or obesity[8]. Interestingly, *TNF* was not among the top regions associated with circulating TNFα in our meta-analysis or in published GWAS of TNFα.[19] Furthermore, there was little overlap between findings of our epigenome-wide meta-analysis and previous GWAS of TNFα. The only exception concerns our observed but unreplicated association between TNFα and methylation loci in ABO, alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase gene (*ABO*) that were also observed in a previous protein quantitative trait loci GWAS[19], which presented evidence that the effect was assay-specific and may be driven by cross-reactivity with ABO antigens. Finally, to the best of our knowledge, the *NLRC5* and *DTX3L-PARP9* findings have not been reported in epigenetic studies of other proinflammatory cytokines, although a recent meta-analysis of C-reactive protein reported multiple associations with methylation loci in other interferon pathway genes[34], illustrating distinct yet related epigenetic determinants of the human immune response.

Given the inflammatory relevance of the TNFα phenotype, the use of leukocyte-derived DNA for methylation measurements constitutes a clear strength of the study. Furthermore, the accessibility of blood facilitates future translational applications of our findings (e.g. development of risk stratification tools or other personalized approaches). The second strength of our study stems from restricting our main analyses to cohorts that measured TNFα using ELISA, considered the 'gold standard' for clinical use[35], thus reducing spurious variation. Third, we achieved independent replication of our top hits in NFBC66, increasing confidence in the validity of our findings. Finally, DNA methylation measurements were available in multiple cohorts that also offered genotype and expression data, enabling an integrative approach to identify the mechanisms linking methylation and circulating TNFα.

However, several limitations of our integrative analyses must be noted. First, the expression findings replicated robustly between FHS and RS, but not in KORA. Possible reasons include discrepancies in population characteristics, gene expression measurements, or chance. Second, FHS measurements of methylation and TNFα were taken several years apart, while all other cohorts performed them contemporaneously. However, the FHS findings were similar to those derived from cross-sectional studies, and the difference in time between the measurements would bias the effect estimates towards the null, further reassuring our results. Third, the reported associations may not be interpreted as causal because they were established in observational data that do not preclude bias, e.g. due to residual confounding. Causal inference methods such as Mendelian randomization, used widely to corroborate findings of epigenome-wide studies, are not optimal for our study because strong genetic instruments for either 1) the methylation at the top loci, which we showed to be only weakly related to the genotype and 2) TNFα itself are not currently available. Future studies may consider directly interrogating the relationship between DNA methylation in *NLRC5* and *PARP9-DTX3L* and systemic inflammation in experimental models.

In summary, we report novel evidence linking DNA methylation in two immune response-related regions—*NLRC5* and *PARP9-DTX3L*—with corresponding gene expression, circulating TNFα, and incident CHD in a population-based meta-analysis, highlighting the potential of these regions as translational targets. Further, our findings illustrate the utility of agnostic methylome-wide studies in identifying physiologically meaningful phenomena. In concert with evidence from *in vitro* and *in vivo* functional studies, our findings yield valuable insights into immunopathology of CHD.

References

1. Kalliolias GD, Ivashkiv LB. TNF biology, pathogenic mechanisms and emerging therapeutic strategies. *Nat Rev Rheumatol.* 2016;12(1):49-62.
2. Brenner D, Blaser H, Mak TW. Regulation of tumour necrosis factor signalling: live or let die. *Nat Rev Immunol.* 2015;15(6):362-374.
3. Pope JE. Rheumatoid arthritis: TNF inhibitors and cardiovascular risk management in RA. *Nat Rev Rheumatol.* 2016;12(6):317-318.
4. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. *N Engl J Med.* 2017;377(12):1119-1131.
5. Sas AA, Jamshidi Y, Zheng D, et al. The age-dependency of genetic and environmental influences on serum cytokine levels: a twin study. *Cytokine.* 2012;60(1):108-113.
6. Neijts M, van Dongen J, Kluft C, Boomsma DI, Willemsen G, de Geus EJ. Genetic architecture of the pro-inflammatory state in an extended twin-family design. *Twin Res Hum Genet.* 2013;16(5):931-940.
7. Stein CM, Guwatudde D, Nakakeeto M, et al. Heritability analysis of cytokines as intermediate phenotypes of tuberculosis. *J Infect Dis.* 2003;187(11):1679-1685.
8. Hermsdorff HH, Mansego ML, Campion J, Milagro FI, Zulet MA, Martinez JA. TNF-alpha promoter methylation in peripheral white blood cells: relationship with circulating TNFalpha, truncal fat and n-6 PUFA intake in young women. *Cytokine.* 2013;64(1):265-271.
9. Marques-Rocha JL, Milagro FI, Mansego ML, Mourao DM, Martinez JA, Bressan J. LINE-1 methylation is positively associated with healthier lifestyle but inversely related to body fat mass in healthy young individuals. *Epigenetics.* 2016;11(1):49-60.
10. Sullivan KE, Reddy AB, Dietzmann K, et al. Epigenetic regulation of tumor necrosis factor alpha. *Mol Cell Biol.* 2007;27(14):5147-5160.
11. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet.* 2009;2(1):73-80.
12. Liu MY, Xydakis AM, Hoogeveen RC, et al. Multiplexed analysis of biomarkers related to obesity and the metabolic syndrome in human plasma, using the Luminex-100 system. *Clin Chem.* 2005;51(7):1102-1109.
13. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet.* 2011;88(5):586-598.
14. Agha G, Group IHDCEW. DNA methylation is associated with incident cardiovascular disease. *Gerontologist.* 2016;56((Supp 3)):34-35.
15. Mendelson MM, Marioni RE, Joehanes R, et al. Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med.* 2017;14(1):e1002215.
16. Bergen SE, O'Dushlaine CT, Ripke S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry.* 2012;17(9):880-886.
17. Fanous AH, Zhou B, Aggen SH, et al. Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. *Am J Psychiatry.* 2012;169(12):1309-1317.

18. Ripke S, Wray NR, Lewis CM, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry.* 2013;18(4):497-511.

19. Melzer D, Perry JR, Hernandez D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* 2008;4(5):e1000072.

20. Meissner TB, Li A, Biswas A, et al. NLR family member NLRC5 is a transcriptional regulator of MHC class I genes. *Proc Natl Acad Sci USA.* 2010;107(31):13794-13799.

21. Braud VM, Allan DS, McMichael AJ. Functions of nonclassical MHC and non-MHC-encoded class I molecules. *Curr Opin Immunol.* 1999;11(1):100-108.

22. Zhao Y, Shao F. NLRC5: a NOD-like receptor protein with many faces in immune regulation. *Cell Res.* 2012;22(7):1099-1101.

23. Yao Y, Wang Y, Chen F, et al. NLRC5 regulates MHC class I antigen presentation in host defense against intracellular pathogens. *Cell Res.* 2012;22(5):836-847.

24. Zhang X, Justice AC, Hu Y, et al. Epigenome-wide differential DNA methylation between HIV-infected and uninfected individuals. *Epigenetics.* 2016:1-11.

25. Yoshihama S, Roszik J, Downs I, et al. NLRC5/MHC class I transactivator is a target for immune evasion in cancer. *Proc Natl Acad Sci USA.* 2016;113(21):5999-6004.

26. Zhang Y, Mao D, Roswit WT, et al. PARP9-DTX3L ubiquitin ligase targets host histone H2BJ and viral 3C protease to enhance interferon signaling and control viral infection. *Nature Immunol.* 2015;16(12):1215-1227.

27. Iwata H, Goettsch C, Sharma A, et al. PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. *Nature Commun.* 2016;7:12849.

28. Pacher P, Szabo C. Role of poly(ADP-ribose) polymerase 1 (PARP-1) in cardiovascular diseases: the therapeutic potential of PARP inhibitors. *Cardiovasc Drug Rev.* 2007;25(3):235-260.

29. Zakaria EM, El-Bassossy HM, El-Maraghy NN, Ahmed AF, Ali AA. PARP-1 inhibition alleviates diabetic cardiac complications in experimental animals. *Eur J Pharmacol.* 2016;791:444-454.

30. Day K, Waite LL, Alonso A, et al. Heritable DNA Methylation in CD4+ Cells among Complex Families Displays Genetic and Non-Genetic Effects. *PLoS One.* 2016;11(10):e0165488.

31. Zhang S, Barros SP, Moretti AJ, et al. Epigenetic regulation of TNFA expression in periodontal disease. *J Periodontol.* 2013;84(11):1606-1616.

32. Kojima A, Kobayashi T, Ito S, Murasawa A, Nakazono K, Yoshie H. Tumor necrosis factor-alpha gene promoter methylation in Japanese adults with chronic periodontitis and rheumatoid arthritis. *J Periodontal Res.* 2016;51(3):350-358.

33. Arroyo-Jousse V, Garcia-Diaz DF, Codner E, Perez-Bravo F. Epigenetics in type 1 diabetes: TNFa gene promoter methylation status in Chilean patients with type 1 diabetes mellitus. *Br J Nutr.* 2016;116(11):1861-1868.

34. Ligthart S, Marzi C, Aslibekyan S, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* 2016;17(1):255.

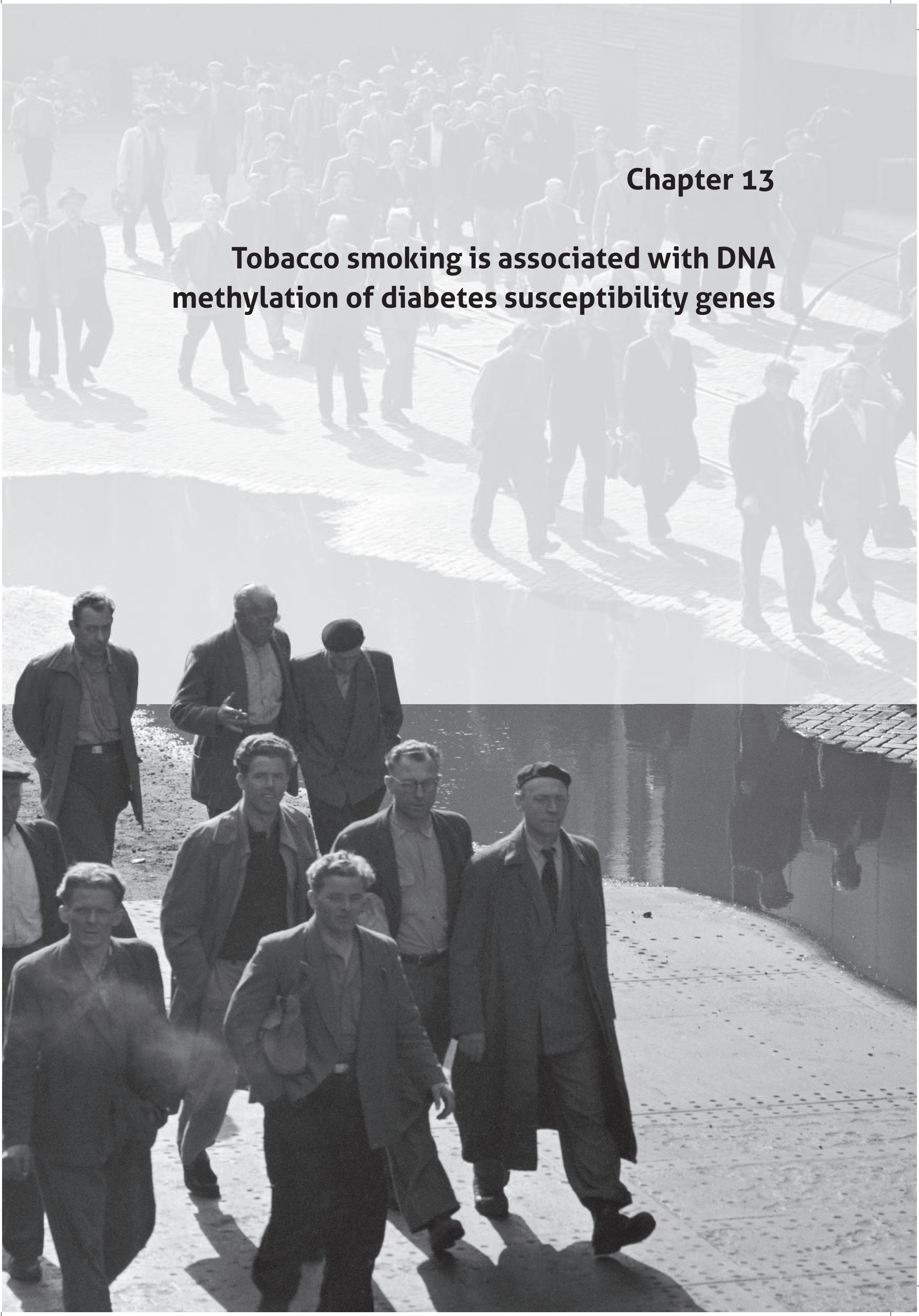
35. Elshal MF, McCoy JP. Multiplex bead array assays: performance evaluation and comparison of sensitivity to ELISA. *Methods.* 2006;38(4):317-323.

Supplementary material

The supplementary material of this manuscript can found at the following webpage: https://jamanetwork.com/journals/jamacardiology/fullarticle/2677631.

Chapter 13

Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes



Background: Tobacco smoking, a risk factor for diabetes, is an established modifier of DNA methylation. We hypothesized that tobacco smoking modifies DNA methylation of genes previously identified for diabetes.

Methods: We annotated CpG sites available on the Illumina450K array to diabetes genes previously identified by genome-wide association studies (GWAS), and investigated them for an association with smoking comparing current to never smokers. The discovery study consisted of 630 individuals (Bonferroni corrected P-value=1.4×10^{-5}), and we sought replication in an independent sample of 674 individuals. The replicated sites were tested for association with nearby genetic variants and gene expression and fasting glucose and insulin levels.

Results: We annotated 3,620 CpG sites to the genes identified in the GWAS on type 2 diabetes. Comparing current to never smokers, we found 12 differentially methylated CpG sites, of which five replicated: cg23161492 within *ANPEP* (P-value=1.3×10^{-12}); cg26963277 (P-value=1.2×10^{-9}), cg01744331 (P-value=8.0×10^{-6}) and cg16556677 (P-value=1.2×10^{-5}) within *KCNQ1;* cg03450842 (P-value=3.1×10^{-8}) within *ZMIZ1.* The effect of smoking on DNA methylation at the replicated CpG sites attenuated after smoking cessation. Increased DNA methylation at cg23161492 was associated with decreased gene expression levels of *ANPEP* (P-value=8.9×10^{-5}). rs231356-T, which was associated with hypomethylation of cg26963277 (KCNQ1), was associated with an higher odds of diabetes (OR=1.06, P-value=1.3×10^{-5}). Additionally, hypomethylation of cg26963277 was associated with lower fasting insulin levels (P-value=0.04).

Conclusion: Tobacco smoking is associated with differential DNA methylation of the diabetes risk genes *ANPEP, KCNQ1* and *ZMIZ1*. Our study highlights potential biological mechanisms connecting tobacco smoking to excess risk of type 2 diabetes.

Introduction

In the last decade, genome-wide association studies (GWAS) have been conducted in order to identify DNA sequence variants for a wide range of diseases including type 2 diabetes[1,2,3]. These GWAS have successfully identified numerous single-nucleotide polymorphisms (SNPs) located in and near genes that may be key in the development of type 2 diabetes. Up to now, a total number of 88 genetic loci have been identified for type 2 diabetes[4].

Tobacco smoking is associated with an increased risk of type 2 diabetes[5]. Several biological mechanisms have been proposed through which smoking may have an effect on the development of diabetes, including inflammation and the effect of nicotine on insulin resistance[6]. However, the exact molecular mechanisms connecting smoking to an increased risk of diabetes remain largely unknown. Previous research has established an important role of tobacco smoking on DNA methylation, the epigenetic mechanism of attachment of a methyl-group to a nucleotide[7,8,9]. DNA methylation has several functions on the human genome including the regulation of gene expression and maintaining genome stability[10]. In line with this, previous studies have suggested a role for DNA methylation as a potential pathway in the association between tobacco smoking and an increased risk of diabetes[11].

We hypothesized that tobacco smoking changes DNA methylation of susceptibility loci identified in GWAS for type 2 diabetes. We therefore investigated the association between DNA methylation in whole blood at loci identified for type 2 diabetes through GWAS and current tobacco smoking in a Dutch population-based cohort study. Furthermore, we investigated the potential effect of DNA methylation on the expression of genes nearby the identified methylation sites.

Methods

Study population

The study was conducted using data from the Rotterdam Study. The design of the Rotterdam Study has been described elsewhere[12]. In brief, 1990 all inhabitants living in the neighborhood Ommoord in Rotterdam, the Netherlands, aged 55 years and over were invited to participate (RS-1). In 2000, the cohort was extended with 3,011 participants aged 55 years and over that had reached the age of 55 or had moved into the research area (RS-2). In 2006, a third cohort of 3,934 participants aged 45 years and older was initiated (RS-3). The discovery panel consisted of 630 non-diabetic participants in the first visit of RS-3 (diabetes was defined as a serum glucose level ≥7.0 mmol/L or the use of glucose-lowering medication) of a random subset of 747 Caucasian subjects with DNA methylation data available. We sought replication of the identified CpG sites in a set of 674 non-diabetic participants from the third visit of RS-2 and the second visit of RS-3. The individuals in the replication study did not participate in the discovery study. The Rotterdam Study has been approved by the medical ethics committee according to the Population Screening Act: Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. All participants in the present analysis provided written informed consent to participate and to obtain information from their treating physicians.

Data collection

Data on tobacco smoking was collected during home interviews. Participants were asked about past and present cigarette, cigar and pipe smoking behavior and where then categorized into current, former and never tobacco smokers. We asked current smokers the start age of smoking and amount of cigarettes per day. Former smokers were asked for the age of smoking cessation. Five of the participants had missing smoking status and were therefore excluded from any analysis. During the center visit, weight and height were measured in standing position wearing normal cloths. Body mass index (BMI) was calculated as height in meters by weight in kilograms squared. All participants had blood samples taken during the visit to quantify DNA methylation, messenger RNA (mRNA) expression levels, DNA sequence variants and other blood measurements.

DNA methylation data

DNA was extracted from whole peripheral blood (stored in EDTA tubes) by standardized salting out methods. Genome-wide DNA methylation levels were measured using the Illumina Human Methylation 450K array[13]. In short, samples (500ng of DNA per sample) were first bisulfite treated using the Zymo EZ-96 DNA-methylation kit (Zymo Research, Irvine, CA, USA). Next, samples were hybridized to the arrays according to the manufacturers' protocol. The methylation percentage of a CpG site was reported as a beta-value ranging between 0 (no methylation) and 1 (full methylation). Processing of the Rotterdam Study DNA methylation samples was performed at the Genetic Laboratory of Internal Medicine, Erasmus University Medical Centre Rotterdam.

Quality control of the samples was done with Genome Studio (v2011.1, methylation module version 1.9.0). In the discovery panel, a total number of 16 samples were removed: 7 had a sample call rate below 99%; 5 had incomplete bisulfite conversion and 4 had gender swaps. In the replication set, all samples passed the quality control based on the first two principal components obtained using principal component analysis (PCA), and no gender swaps were detected. Further quality control of the probes was done based on the detection p-value calculated with Genome Studio. Probes with a detection p-value of more than 0.01 in more than 1% of the samples were excluded. Additionally, sample level QC was performed using MethylAid[14]. This resulted in a total set of 474,528 probes which were normalized using the Dasen option of the WateRmelon R-package[15].

mRNA expression data

Whole-blood was collected (PAXGene Tubes – Becton Dickinson) and total RNA was isolated (PAXGene Blood RNA kits - Qiagen). To ensure a constant high quality of the RNA preparations, all RNA samples were analyzed using the Labchip GX (Calliper) according to the manufacturer's instructions. Samples with an RNA Quality Score more than 7 were amplified and labelled (Ambion TotalPrep RNA), and hybridized to the Illumina HumanHT12v4 Expression Beadchips as described by the manufacturer's protocol. Processing of the Rotterdam Study RNA samples was performed at the Genetic Laboratory of Internal Medicine, Erasmus University Medical Centre Rotterdam. The RS-III expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE33828: 881 samples are available for analysis.

Illumina gene expression data was quantile-normalized to the median distribution and subsequently log2-transformed. The probe and sample means were centered to zero. Genes were declared significantly expressed when the detection p-values calculated by GenomeStudio were less than 0.05 in more than 10% of all discovery samples, which added to a total number of 21,238 probes. Quality control was done using the eQTL-mapping pipeline[16]. We only analyzed probes that uniquely mapped to the human genome build 37 and represented gene mRNA expression[17].

Selection of methylation sites

A recent review summarizing all diabetes GWAS findings was used to compile a list of variants significantly associated with diabetes (88 variants)[4]. Next, the list of 88 variants was extended with polymorphisms in linkage disequilibrium ($R^2 > 0.8$) in the HapMap panel and within 500kb using the SNAP Proxy Search tool (https://www.broadinstitute.org/mpg/snap/ldsearch.php). The final list included 890 SNPs and those SNPs were tested for in-gene variants and effects on expression of a gene within 1Mb as found in a large publically available blood cis-expression-quantitative trait loci (*cis*-eQTL) database (FDR<0.05)[16]. We identified 525 SNPs that were in-gene (mapping to 72 unique genes) and 316 SNPs with an eQTL effect (mapping to 50 unique genes). The final number of unique genes was 111. The methylation probes within and near these diabetes-related genes as provided by Illumina were included in the analysis. We excluded probes from the Infinium HD methylation SNP list with a minor allele frequency above 1% as provided by Illumina, since variations in these SNPs can cause bias in the methylation measurement[18]. We further excluded known cross-reactive probes, since they can introduce bias in the results[19]. In total, we included 3,620 CpG sites in the analyses.

Statistical analysis

The characteristics of the discovery and replication population were compared between current and never smokers using IBM SPSS Statistics version 21.0.0.1 (IBM Corp.). The p-values were calculated using independent sample T-tests for continuous variables and Chi-square tests for dichotomous variables.

The 3,620 methylation probes were tested for association with tobacco smoking using a linear mixed model with the LME4 package in R version 3.1.0 with Dasen normalized beta-values of the CpG sites as outcome measure[20]. Extreme outliers (>4SD from the mean and >4SD from the before last) in the DNA methylation values were excluded. We first compared current to never smokers and then performed a sensitivity analysis on the identified CpG sites comparing former to never smokers. Covariates were selected based on known association with DNA methylation. The selected covariates with fixed effects were age, sex and BMI[21-24]. Houseman estimated white blood cell proportions were used as fixed effects to correct for cell mixture distribution[25]. Array number and position on array were added in the model as covariates with random effects to correct for batch effects. We corrected for multiple testing using a robust Bonferroni corrected p-value of 1.4×10^{-5} as the threshold for significance (0.05/3,620 probes).

The probes identified in the discovery analysis were tested for replication in the independent samples from the Rotterdam study. We used identical models with the addition of cohort (RS-2 or RS-3) as a variable in the model to adjust for a potential cohort effect. A Bonferroni corrected p-value of 0.05 divided by the number of significant findings in the discovery study was used as a threshold of significant replication.

The replicated probes were further tested with total pack years in the current smokers to test the association between tobacco smoking and cumulative exposure of smoking. We further investigated the association between the replicated probes and time since cessation in former smokers to study the change in methylation after smoking cessation. To decrease the possibility of confounding in our association, we further adjusted the model in a second analysis for other possible confounders and mediators. This analysis included total cholesterol, HDL-cholesterol, triacylglycerol levels (natural log-transformed), systolic blood pressure, daily alcohol intake, and C-reactive protein levels (natural log-transformed).

Functional analysis

Since DNA methylation may have an effect on gene expression, we tested the association between DNA methylation and mRNA expression levels of nearby genes (*cis*) within 500kb of the replicated CpG sites (250kb up and downstream of the CpG location). First, residuals for mRNA expression were created after regressing out the measured cell counts (granulocytes, lymphocytes, monocytes, platelets and erythrocytes), fasting state, RNA quality score, plate number, age and sex on the mRNA expression levels using a linear mixed model. We then created residuals for DNA methylation regressing out the measured white blood cells, age, sex, array number and position on array on the Dasen normalized beta-values of the CpG sites using a linear mixed model. The residuals of the mRNA expression levels and the residuals of the Dasen normalized beta-values of the CpG sites were tested for association using a linear regression model.

We also studied the association between the replicated CpG sites and serum measures of fasting glucose and insulin combining both the discovery and replication samples. Serum glucose and insulin were measured using standard laboratory techniques. The models were adjusted for the same covariates as in the main analyses, with the addition of smoking category. Serum insulin was natural log-transformed. A Bonferroni corrected p-value for five tests was used. Furthermore, we searched for genetic variants (methylation Quantitative Trait Loci (metQTLs)) associated with the replicated methylation sites in the publicly available data from the paper by Grundberg et al[26]. Significant met-QTLs were then tested for an association with type 2 diabetes in the publicly available data from the DIAGRAM consortium, using a Bonferroni corrected p-value of 0.01 (0.05/5 met-QTLs)[3].

Results

A total of 630 participants were included in the discovery study. Clinical characteristics of the study population by smoking category are listed in Table 1. The participants were on average 59.5±8.0 years old and 45% was male. The samples consisted of 175 current smokers and 184 never smokers, whereas 271 participants were former smokers. On average, current smokers had lower HDL-cholesterol, higher triacylglycerol and serum C-reactive protein compared to never smokers. Also alcohol consumption was higher in current smokers compared to former and never smokers. In the replication panel, 68 individuals were current smoker and 238 never smokers, whereas 368 individuals were former smokers. Clinical characteristics of the replication population can be found in Supplementary table 1.

After correction for multiple testing (P-value=1.4×10^{-5}), we identified 12 differentially methylated CpG sites when comparing current smokers to never smokers in the discovery study (Table 2). The 12 differentially methylated CpG sites were located within eight genes. The most significant finding was cg23161492 located within the gene ANPEP on chromosome 15 (P-value= 1.3×10^{-12}). On chromosome 11, four CpG sites located within the gene KCNQ1 were significantly associated with current tobacco smoking (cg26963277, P-value=1.2×10^{-9}; cg13428066, 5.8×10^{-6}; cg01744331, 8.0×10^{-6}; cg16556677, 1.2×10^{-5}). Within the gene ZMIZ1 on chromosome 10, two CpG sites were significant differentially methylated between current and never smokers (cg03450842, 3.1×10^{-8}; cg21344746, 6.6×10^{-6}). In addition, we identified CpG sites in and near *INPP5E*, *NDUFS5*, *FCHSD2*, *PBX4* and *TCF19* to be differentially methylated in current smokers compared to never smokers.

Table 1. Baseline characteristics of the study population.

		Smoking category			
	Total	**Current**	**Former**	**Never**	**P-value[a]**
N	630	175	271	184	
Age, years	59.5 (8.0)	57.9 (6.6)	60.9 (8.5)	59.0 (8.1)	0.16
Sex, male (%)	283 (45%)	85 (49%)	126 (47%)	72 (39%)	0.07
Body mass index, kg/m^2	27.4 (4.5)	26.7 (4.4)	27.6 (4.3)	27.6 (4.8)	0.07
Fasting glucose, mmol/l	5.35 (0.55)	5.33 (0.58)	5.40 (0.55)	5.30 (0.52)	0.65
Systolic blood pressure, mmHg	138.5 (63.0)	136.4 (60.4)	139.7 (67.4)	138.7 (58.8)	0.71
Diastolic blood pressure, mmHg	88.0 (65.0)	86.0 (62.1)	89.0 (69.8)	88.4 (60.3)	0.71
Total cholesterol, mmol/l	5.60 (1.03)	5.60 (1.07)	5.62 (1.01)	5.56 (1.02)	0.72
HDL-cholesterol, mmol/l	1.41 (0.40)	1.34 (0.39)	1.44 (0.41)	1.44 (0.37)	0.01
Triglycerides, mmol/l	1.45 (0.81)	1.62 (1.02)	1.39 (0.62)	1.40 (0.81)	0.02
C-reactive protein	2.55 (4.74)	3.17 (7.03)	2.52 (3.54)	2.03 (3.31)	0.05
Alcohol, g/day	18.3 (11.0)	19.4 (12.7)	19.0 (10.9)	16.1 (9.3)	0.006
Fasting[b], yes (%)	628 (100%)	173 (99%)	271 (100%)	184 (100%)	0.15

Data are mean (SD) or n (%). HDL-cholesterol denotes high density lipoprotein.

[a] Current versus never smokers.

[b] The subjects who provided blood after an overnight fast.

Table 2. Significant associations between current versus never tobacco smoking and methylation of diabetes genes.

			Discovery		*Replication*		
CpG site	**Chr**	**Position**	**Beta(SE)**	**P-value**	**Beta(SE)**	**P-value**	**Gene**
cg23161492	15	90357202	-0.044(0. 006)	1.3×10^{-12}	-0.045(0.006)	3.4×10^{-11}	*ANPEP*
cg26963277	11	2722407	-0.026(0.004)	1.2×10^{-9}	-0.034(0.004)	3.3×10^{-14}	*KCNQ1*
cg03450842	10	80834947	-0.017(0.003)	3.1×10^{-8}	-0.030(0.004)	2.2×10^{-12}	*ZMIZ1*
cg14024579	9	139332845	-0.022(0.004)	1.1×10^{-7}	-0.015(0.006)	0.01	*INPP5E*
cg14656441	1	39500070	0.026(0.005)	1.5×10^{-6}	0.016(0.008)	0.05	*NDUFS5*
cg13912027	11	72759293	0.022(0.005)	2.1×10^{-6}	-0.001(0.006)	0.89	*FCHSD2*
cg00591868	19	19729048	-0.015(0.003)	4.6×10^{-6}	-0.003(0.005)	0.51	*PBX4*
cg13428066	11	2677768	0.015(0.003)	5.8×10^{-6}	0.007(0.006)	0.28	*KCNQ1*
cg21344746	10	80831230	0.016(0.004)	6.6×10^{-6}	0.001(0.005)	0.82	*ZMIZ1*
cg16095155	6	31127863	-0.013(0.003)	7.2×10^{-6}	-0.007(0.004)	0.12	*TCF19*
cg01744331	11	2722358	-0.013(0.003)	8.0×10^{-6}	-0.025(0.003)	7.4×10^{-12}	*KCNQ1*
cg16556677	11	2722401	-0.015(0.003)	1.2×10^{-5}	-0.027(0.004)	3.9×10^{-10}	*KCNQ1*

Adjusted for age, sex, body mass index, houseman estimated white blood cell proportions and batch effects. Chr denotes chromosome. Position according to Hg19. Bonferroni corrected threshold for significance: 0.05/3,620=1.4×10^{-5}.

We attempted replication of the 12 differentially methylated CpG sites from the discovery study in 674 independent participants of the second and third cohort of the Rotterdam Study. We used a p-value of 4.2×10^{-3} (0.05/12) as a threshold of significant replication. We significantly replicated the five CpG sites cg23161492 (*ANPEP*), cg26963277 (*KCNQ1*), cg03450842 (*ZMIZ1*), cg01744331 (*KCNQ1*) and cg16556677 (*KCNQ1*) (Table 2). Furthermore, the replicated associations were robust to further adjustment for possible confounders including systolic blood pressure, total cholesterol, HDL-cholesterol, triacylglycerol, alcohol consumption and C-reactive protein (Table S2). Boxplots of replicated probe beta-values per smoking category are presented in Figure 1.

When we adjusted the effect of the top signal within the *KCNQ1* gene (cg26963277) for the second (cg01744331) or third (cg16556677) signal within *KCNQ1*, cg26963277 was associated with current smoking, whereas cg01744331 and cg16556677 did not show an association (*P* 0.84 and 0.35, respectively).

To study the effect of smoking cessation on the replicated CpG sites, we compared former to never smokers and tested the association between time since smoking cessation and DNA methylation. DNA methylation at the five CpG sites were not differentially methylated comparing former to never smokers (Table 3). Methylation at cg23161492 (P-value=2.6×10^{-6}), cg26963277 (P-value=2.1×10^{-4}), cg01744331 (P-value=5.1×10^{-5}) and cg16556677 (P-value=1.2×10^{-3}) was associated with time since smoking cessation. Additionally, methylation at the CpG sites cg23161492, cg26963277, cg03450842 and cg01744331 was associated with cumulative exposure of tobacco smoking.

In the 630 individuals from the discovery panel, six genes out of 20 candidates were found to be significantly expressed in the analyzed whole blood samples. The 12 methylation-

Table 3. Association between CpG sites and former compared to never smokers, time since smoking cessation and cumulative smoking exposure in pack years.

	Former versus never smokers		*Cessation time*		*Packyears*	
CpG site	**Beta(SE)**	**P-value**	**Beta[a](SE)**	**P-value**	**Beta[a](SE)**	**P-value**
cg23161492	-0.007(0.006)	0.24	0.014(0.003)	2.6×10^{-6}	-0.007(0.002)	2.8×10^{-3}
cg26963277	-0.006(0.003)	0.05	0.006(0.002)	2.1×10^{-4}	-0.006(0.002)	9.0×10^{-4}
cg03450842	-0.005(0.002)	0.06	0.002(0.001)	0.20	-0.003(0.001)	1.6×10^{-3}
cg01744331	-0.003(0.002)	0.21	0.005(0.001)	5.1×10^{-5}	-0.004(0.001)	1.1×10^{-4}
cg16556677	-0.007(0.003)	7.1×10^{-3}	0.005(0.001)	1.2×10^{-3}	-0.003(0.001)	0.05

[a]Beta represent change in methylation per 10 years since smoking cessation and per 10 packyears.
Adjusted for age, sex, body mass index, white blood cell counts and batch effects. Bonferroni corrected threshold for significance: 0.05/15 =3.3×10^{-3}.

expression combinations can be found in Table S3. The p-value threshold for association was 4.2×10^{-3} (0.05/12 tests). Increased methylation at cg23161492 was negatively associated with gene expression levels of *ANPEP* (P-value=8.9×10^{-5}) (Figure S1).

We observed a putative effect of the CpG site cg26963277 with fasting serum insulin (effect: 0.004, P-value=0.04). Results for the associations between all replicated CpG sites and serum fasting glucose and insulin are presented in Table S4.

We identified a significant met-QTL for all replicated CpG sites, except cg0345084 (Table S5). The T-allele of the SNP rs231356 is associated with lower methylation of both cg26963277 and cg01744331 (*KCNQ1*). Also, the T-allele of the SNP rs231356 is associated with an increased odds of type 2 diabetes (Odds Ratio=1.06, P-value=1.3×10^{-5}).

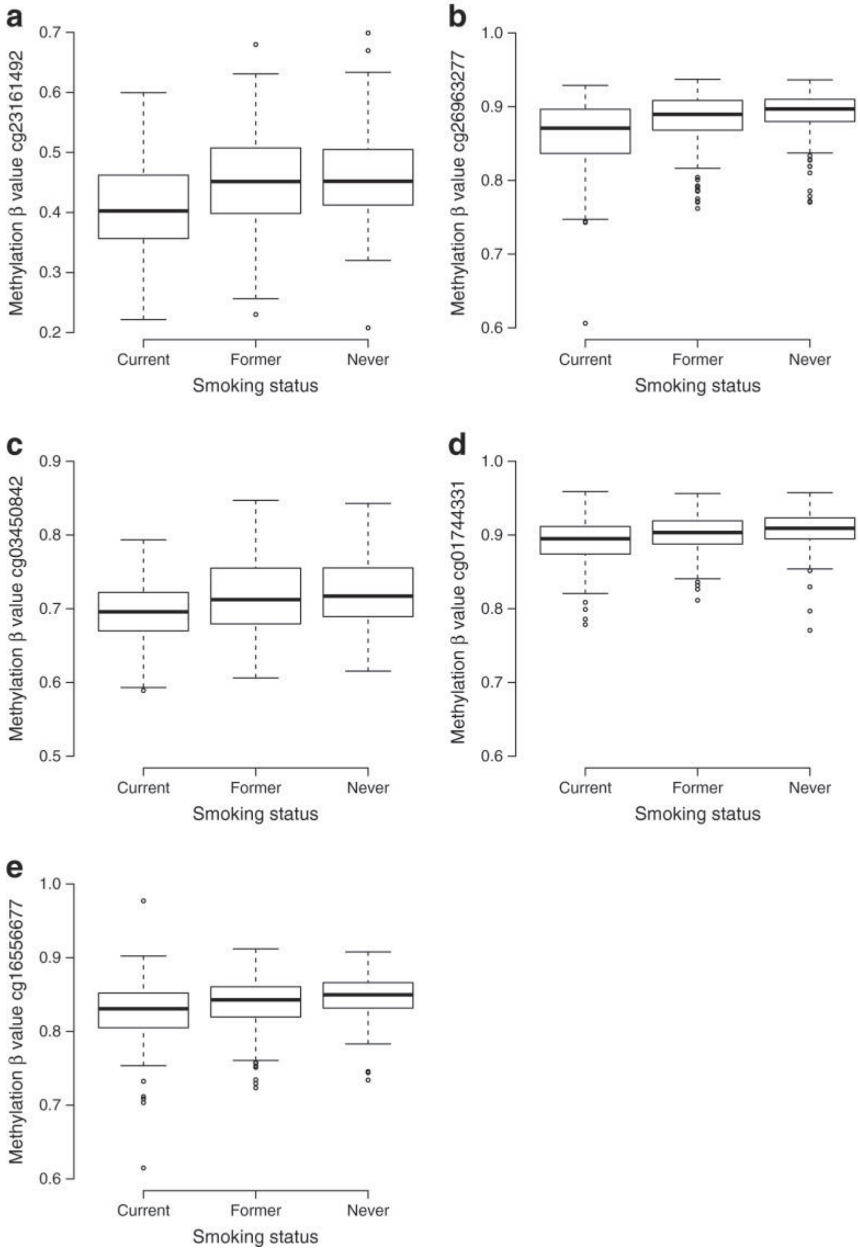
Discussion

Our findings suggest that tobacco smoking is associated with differential methylation of CpG sites within the type 2 diabetes risk genes *ANPEP, KCNQ1* and *ZMIZ1*. The associations were robust to adjustment for potential confounders and the effect of tobacco smoking appeared to be reversible after smoking cessation. In addition, methylation within *ANPEP* was significantly associated with gene expression levels of *ANPEP*. Furthermore, methylation at *KCNQ1* was associated with fasting insulin levels, and genetic data supported the role of methylation at *KCNQ1* in the development of diabetes. This study provides further insight into potential biological mechanisms underlying the association between tobacco smoking and an excess risk of type 2 diabetes.

In contrast to the findings comparing current to never smokers, we found similar DNA methylation levels at the replicated CpG sites comparing former to never smokers. Furthermore, four significant CpG sites were associated with time since smoking cessation, suggesting a return after smoking cessation to DNA methylation levels similar to never smokers. This is in agreement with previous studies investigating the role of smoking cessation on DNA methylation[7, 27, 28]. DNA methylation may return to levels similar to never smokers at some sites, whilst other sites stay differentially methylated. Our results are in agreement with a potential beneficial effect of smoking cessation on DNA methylation at risk loci for diabetes. Furthermore, we observed for four CpG sites a dose-dependent effect of smoking underscoring the importance of cumulative tobacco exposure over time.

We identified three CpG sites within intron 11 of *KCNQ1* (potassium channel, voltage gated KQT-like subfamily Q, member 1) to be differentially methylated in smokers compared to never smokers. Previous studies have reported differential DNA methylation at the *KCNQ1* locus in pancreatic islets and adipose tissue of diabetes cases and non-diabetes controls[29, 30]. Adjustment analyses suggested that cg26963277 is the driving CpG site associated with current smoking at this locus. Furthermore, we found the metQTL (rs231356) for

Figure 1. Boxplots depicting the methylation values in the replicated CpG sites in current, former, and never smokers.



cg26963277 to be associated with the risk of diabetes. More specifically, the T-allele of rs231356 is associated with lower methylation of cg26963277 and an increased odds of type 2 diabetes. In agreement with this observation, tobacco smoking lowers methylation at cg26963277 and is associated with an increased risk of diabetes. Additionally, our data suggest an association between cg26963277 and fasting insulin levels: increasing methylation was putatively associated with increasing fasting insulin levels. Although we did not observe an association between DNA methylation at cg26963277 and expression of *KCNQ1*, our results provide evidence that smoking may increase the risk of diabetes through decreased methylation at *KCNQ1* and a subsequent decrease in fasting insulin levels. Further, current tobacco smoking was associated with a 4.4% decrease in methylation at cg23161492 located near the 5' UTR of the gene *ANPEP* and this decreased methylation was correlated with increased gene expression levels of *ANPEP.* The *ANPEP* gene encodes the protein alanine aminopeptidase, a widely expressed enzyme with various cellular processes including cell proliferation, differentiation and apoptosis[31]. The observation that current smoking, which increases the risk of type 2 diabetes, may lead to higher gene expression levels of *ANPEP* is in line with the observation of Locke and colleagues[32]. The risk allele of the SNP rs2007084 identified in the DIAGRAM consortium is also associated with increased gene expression of *ANPEP* in islet cells[32]. This suggests that increased expression of *ANPEP* leads to an increased risk of type 2 diabetes. The observation that DNA sequence variation and DNA methylation at this locus is associated with increased expression levels of *ANPEP* suggests a role for *ANPEP* in the pathogenesis of type 2 diabetes, rather than the gene *AP3S2* proposed by prior GWAS[3].

We further identified the CpG cg03450842 near the 5' UTR of *ZMIZ1,* to be differentially methylated among smokers compared to never smokers. The CpG cg03450842 has been identified previously to be associated with smoking[11].Unfortunately, we had no expression data available in our samples for this gene and could therefore not study the effect of methylation at cg03450842 on gene expression of *ZMIZ1*.

The strength of the current study is the large sample size with available data on DNA methylation, gene expression and genetic variants which allowed in detail investigation of the interrelationship between tobacco smoking, DNA methylation and gene expression. A limitation of the current work is the use of whole blood samples for the quantification of DNA methylation and gene expression. As both methylation and expression may be tissue specific, we might have overlooked potential associations between tobacco smoking and differential methylation of diabetes related genes in other tissues, for instance liver, fat, pancreas or muscle tissue. Furthermore, observed associations may not be generalizable to other tissues. Another limitation is the challenge of gene annotation in GWAS. GWAS locate DNA sequence variants for phenotypes, but the underlying causal gene might be difficult to designate. To minimize this problem we limited our analysis to genes annotated to in-gene variants and known cis-eQTL effects. Therefore the diabetes risk genes selected in our study are more plausible to be the causal gene for diabetes.

In summary, our study suggests an effect of tobacco smoking on DNA methylation of the diabetes-related genes *ANPEP*, *KCNQ1* and *ZMIZ1.* Our study provides further insight into potential mechanisms linking tobacco smoking to an excess risk of type 2 diabetes.

References

1. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics* 2008; 40(5): 638-45.

2. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics* 2010; 42(7): 579-89.

3. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* 2012; 44(9): 981.

4. Marullo L, Moustafa JSE-S, Prokopenko I. Insights into the genetic susceptibility to type 2 diabetes from genome-wide association studies of glycaemic traits. *Current diabetes reports* 2014; 14(11): 1-17.

5. Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *Jama* 2007; 298(22): 2654-64.

6. Xie X-t, Liu Q, Wu J, Wakui M. Impact of cigarette smoking in type 2 diabetes development. *Acta Pharmacologica Sinica* 2009; 30(6): 784-7.

7. Zeilinger S, Kühnel B, Klopp N, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PloS one* 2013; 8(5): e63812.

8. Shenker NS, Polidoro S, van Veldhoven K, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Human molecular genetics* 2012: dds488.

9. Steenaard RV, Ligthart S, Stolk L, et al. Tobacco smoking is associated with methylation of genes related to coronary artery disease. *Clinical Epigenetics* 2015; 7(1): 54.

10. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* 2012; 13(7): 484-92.

11. Besingi W, Johansson Å. Smoke related DNA methylation changes in the etiology of human disease. *Human molecular genetics* 2013: ddt621.

12. Hofman A, Brusselle GGO, Murad SD, et al. The Rotterdam Study: 2016 objectives and design update. *European journal of epidemiology* 2015; 30(8): 661-708.

13. Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011; 6(6): 692-702.

14. van Iterson M, Tobi EW, Slieker RC, et al. MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics* 2014; 30(23): 3435-7.

15. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics* 2013; 14(1): 293.

16. Westra H-J, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics* 2013; 45(10): 1238-43.

17. Schurmann C, Heim K, Schillert A, et al. Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PloS one* 2012; 7(12): e50938.

18. Zhi D, Aslibekyan S, Irvin MR, et al. SNPs located at CpG sites modulate genome-epigenome interaction. *Epigenetics* 2013; 8(8): 802-6.

19. Chen Y-a, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013; 8(2): 203-9.

20. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0; 2012.

21. Koestler DC, Christensen BC, Karagas MR, et al. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* 2013; 8(8): 816-26.

22. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. *The Lancet* 2014; 383(9933): 1990-8.

23. Florath I, Butterbach K, Müller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Human molecular genetics* 2014; 23(5): 1186-201.

24. Zhang FF, Cardarelli R, Carroll J, et al. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics* 2011; 6(5): 623-9.

25. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* 2012; 13(1): 86.

26. Grundberg E, Meduri E, Sandling JK, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *The American Journal of Human Genetics* 2013; 93(5): 876-90.

27. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *The American Journal of Human Genetics* 2011; 88(4): 450-7.

28. Tsaprouni LG, Yang T-P, Bell J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* 2014; 9(10): 1382-96.

29. Nilsson E, Jansson PA, Perfilyev A, et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes* 2014; 63(9): 2962-76.

30. Dayeh T, Volkov P, Salö S, et al. Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS genetics* 2014; 10(3): e1004160.

31. Mina-Osorio P. The moonlighting enzyme CD13: old and new functions to target. *Trends in molecular medicine* 2008; 14(8): 361-71.

32. Locke JM, Hysenaj G, Wood AR, Weedon MN, Harries LW. Targeted allelic expression profiling in human islets identifies cis-regulatory effects for multiple variants identified by type 2 diabetes genome-wide association studies. *Diabetes* 2014: DB_140957.

Supplementary material

Table S1. Clinical characteristics of the replication population.

		Smoking category			
	Total	**Current**	**Former**	**Never**	**P-value[a]**
N	674	68	368	238	
Age, years	67.4 (6.0)	66.3 (6.2)	67.7 (5.7)	67.2 (6.2)	0.27
Sex, male (%)	277 (41%)	24 (35%)	176 (48%)	77 (32%)	0.76
Body mass index, kg/m^2	27.5 (4.0)	26.0 (3.5)	28.0 (4.1)	27.1 (3.9)	0.03
Fasting glucose, mmol/l	5.42 (0.57)	5.32 (0.52)	5.49 (0.57)	5.35 (0.56)	0.74
Systolic blood pressure, mmHg	144.7 (22.2)	139.9 (19.4)	145.9 (23.0)	144.3 (21.4)	0.11
Diastolic blood pressure, mmHg	84.5 (11.7)	81.3 (10.0)	85.6 (12.3)	83.9 (11.0)	0.07
Total cholesterol, mmol/l	5.60 (0.99)	5.62 (1.05)	5.59 (1.00)	5.59 (0.97)	0.30
HDL-cholesterol, mmol/l	1.55 (0.44)	1.60 (0.56)	1.52 (0.43)	1.58 (0.42)	0.77
Triglycerides, mmol/l	1.42 (0.79)	1.44 (0.75)	1.50 (0.88)	1.28 (0.60)	0.11
Fasting[b], yes (%)	673 (99.9%)	68 (100%)	367 (99.7%)	238 (100%)	NA

Data are mean (SD) or n (%).

[a] Current versus never smokers.

[b] The subjects who provided blood after an overnight fast.

Table S2. Significant associations between tobacco smoking and methylation of diabetes genes, adjusted for potential confounding factors.

CpG site	**Beta**	**SE**	**P-value**	**Gene**
cg23161492	-0.040	0.006	2.6×10^{-10}	*ANPEP*
cg26963277	-0.025	0.004	2.5×10^{-8}	*KCNQ1*
cg03450842	-0.012	0.003	5.8×10^{-4}	*ZMIZ1*
cg01744331	-0.013	0.003	1.4×10^{-5}	*KCNQ1*
cg16556677	-0.015	0.004	3.8×10^{-5}	*KCNQ1*

Adjusted for age, sex, body mass index, houseman estimated white blood cell proportions, batch effects, systolic blood pressure, total cholesterol, HDL-cholesterol, triglycerides (natural logarithm), alcohol consumption and C-reactive protein (natural logarithm).

Table S3. Association between identified CpG sites and expression of nearby genes.

CpG site	Gene	ILMNID	Beta	SE	P-value
cg01744331	*CDKN1C*	ILMN_1718565	4.824	2.044	0.02
cg01744331	*NAP1L4*	ILMN_1804327	0.988	1.096	0.37
cg01744331	*SLC22A18*	ILMN_2382505	-0.005	1.704	1.00
cg16556677	*NAP1L4*	ILMN_1804327	0.775	0.867	0.37
cg16556677	*SLC22A18*	ILMN_2382505	-0.295	1.347	0.83
cg16556677	*CDKN1C*	ILMN_1718565	-0.321	1.623	0.84
cg23161492	*ANPEP*	ILMN_1763837	-4.352	0.973	8.9×10^{-06}
cg23161492	*AP3S2*	ILMN_1731596	0.109	0.35	0.76
cg23161492	*C15ORF38*	ILMN_2189406	-0.111	0.387	0.77
cg26963277	*CDKN1C*	ILMN_1718565	2.998	1.359	0.03
cg26963277	*SLC22A18*	ILMN_2382505	2.007	1.13	0.08
cg26963277	*NAP1L4*	ILMN_1804327	-0.631	0.728	0.39

Residual expression after adjustment for age, sex, batch effects, houseman estimated white blood cell proportions, erythrocytes and platelet cell counts, fasting state and RNA quality score associated with residual methylation after adjustment for age, sex, houseman estimated white blood cell proportions and batch effects. Estimates are changes in residual expression per percentage residual methylation increase. Bonferroni corrected threshold of significance: 0.05/12=4.2x10^{-3}.

Table S4. Results for the associations between the replicated CpG sites and fasting serum glucose and insulin levels.

CpG site	Effect	P-value
cg23161492	0.0015	0.58
cg26963277	-0.0012	0.47
cg03450842	-0.0010	0.47
cg01744331	-0.0003	0.79
cg16556677	-0.0006	0.69
cg23161492	0.0051	0.10
cg26963277	0.0039	0.04
cg03450842	-0.0002	0.93
cg01744331	0.011	0.45
cg16556677	0.0031	0.09

Effect represents the effect in methylation beta-value per 1-unit increase in fasting glucose (mmol/l) or natural logarithm of insulin (pmol/l). *P* represents the unadjusted p-value.

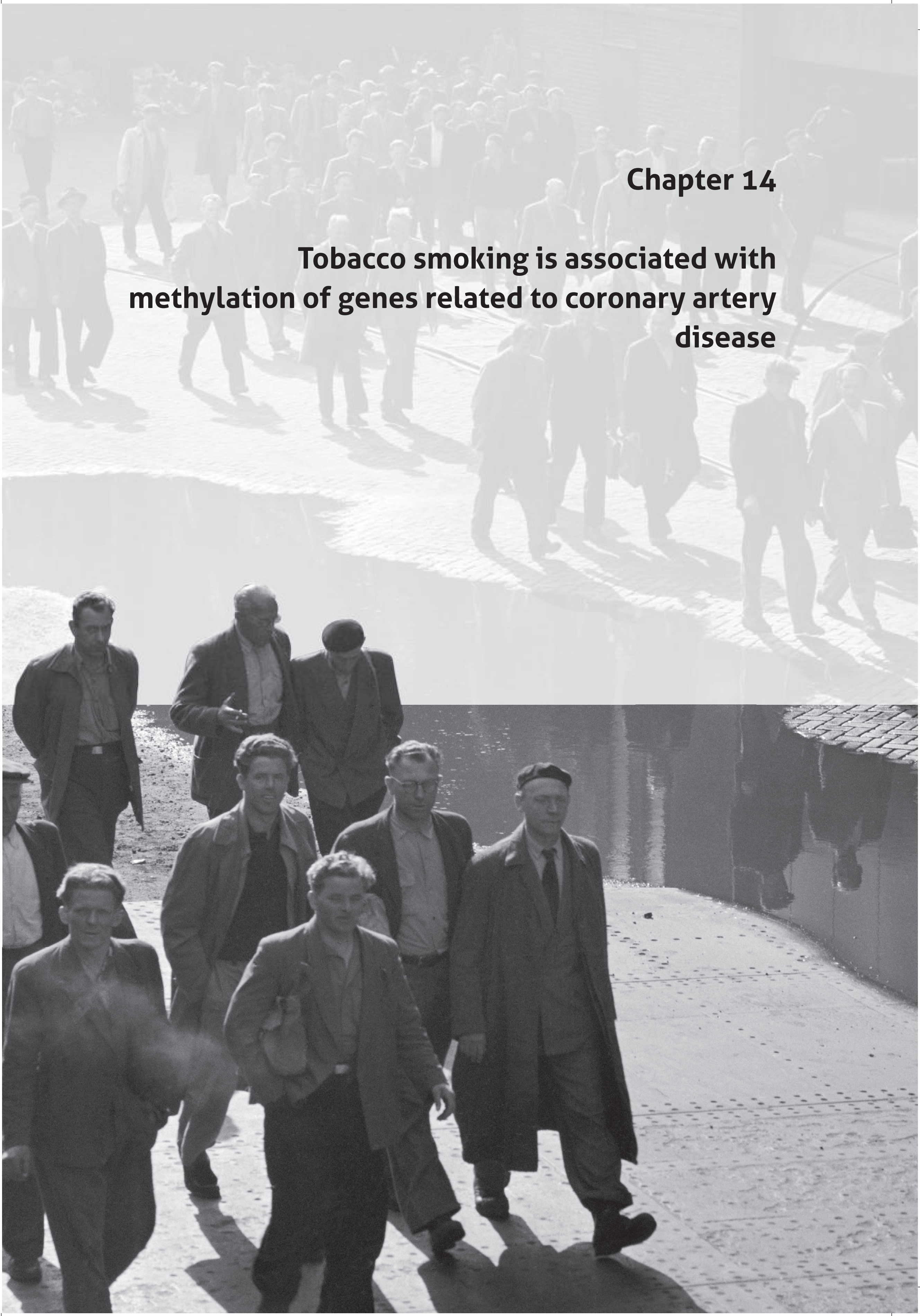
Table S5. Replicated CpG sites, met-QTLs and association with type 2 diabetes.

CpG site	SNP	Distance[a]	Effect Allele	met-QTL beta	met-QTL P-value	T2D OR	T2D P-value
cg23161492	rs11073891	6,792	C	0.023	1.5×10^{-14}	1.00	0.88
cg26963277	rs231356	17,065	T	-0.010	8.9×10^{-6}	1.06	1.3×10^{-5}
cg01744331	rs231356	17,015	T	-0.011	2.5×10^{-6}	1.06	1.3×10^{-5}
cg16556677	rs2283194	265	G	-0.011	3.2×10^{-5}	1.00	0.95

[a]Distance between CpG sites and met-QTL SNP.

Chapter 14

Tobacco smoking is associated with methylation of genes related to coronary artery disease



Background: Tobacco smoking, a risk factor for coronary artery disease (CAD), is known to modify DNA methylation. We hypothesized that tobacco smoking modifies methylation of the genes so far identified for CAD by genome-wide association study (GWAS).

Methods: We selected genomic regions based on 150 single-nucleotide polymorphisms (SNPs) identified in the largest GWAS on CAD. We investigated the association between current smoking and the CpG sites within and near these CAD related genes. Methylation was measured with the Illumina Human Methylation 450K array in whole blood of 724 Caucasian subjects from the Rotterdam Study, a Dutch population based cohort study. Significant CpG sites were then checked for association with mRNA expression of nearby CAD genes.

Results: A total of 3669 CpG sites within 169 CAD related genes were studied for association with current compared to never smoking. Fifteen CpG sites were significantly associated after correction for multiple testing (Bonferroni corrected P-value<1.4×10^{-5}). These sites were located in the genes *TERT*, *SARS, GNGT2, SMG6*, *SKI*, *TOM1L2*, *SIPA1*, *MRAS*, *CDKN1A, LRRC2, FES* and *RPH3A* . In twelve sites, current smoking was associated with a decreased methylation compared to never smoking and in three sites it was associated with increased methylation. The effect estimates decreased in nine of the CpG sites when comparing current to former smoking. One CpG site, cg05603984 (*SKI*) was found to be associated with expression of nearby CAD-related gene *PRKCZ*.

Conclusions: Our study provides examples of CAD related genes of which differential methylation is associated with tobacco smoking.

Introduction

Coronary artery disease (CAD) is a worldwide health problem with a high mortality rate and disease burden[1]. In recent years, large genome-wide association studies (GWAS) have been conducted to identify genetic risk factors for a vast amount of diseases including CAD. These GWAS have successfully identified tens of single-nucleotide polymorphisms (SNPs) located in genes and their vicinity that might play a role in the pathophysiology of CAD. The CARDIoGRAMplusC4D consortium is the largest CAD GWAS consortium comprising 63,746 CAD cases and 130,681 controls[2]. This consortium has found 46 susceptibility loci significantly associated with the risk of CAD and 104 loci suggestive for CAD.

One of the major risk factors for CAD is tobacco smoking which accounts for 10-15% of the risk[3]. Recent studies have shown that smoking can interact with genetic variation to increase the risk of CAD[4,5]. One of the potential mechanisms for this interaction is DNA methylation. DNA methylation is the attachment of a methyl-group to a nucleotide which occurs most often at the cytosine nucleotide of CpG dinucleotides. Methylation has varying functions at different locations in the human genome including influence on gene expression[6].

Since studies have established an important role for smoking in DNA methylation[7,8,9], we hypothesized that tobacco smoking changes DNA methylation of genes near genetic loci identified for CAD, which in turn could alter gene expression of these genes. We therefore investigated the association between methylation of genes near CAD-GWAS loci in whole blood and tobacco smoking in the Rotterdam Study. Furthermore, we investigated the correlation between methylation and expression of CAD related genes nearby the identified differentially methylation sites.

Methods

Study population

The study was conducted using data from the third cohort of the Rotterdam Study. The design of the Rotterdam Study has previously been described elsewhere[10]. Briefly, all inhabitants from the neighborhood Ommoord in Rotterdam aged 45 years and over were invited to participate. During the center visit, 3934 participants were examined between February 2006 and December 2008. We performed the analyses on a random subset of 747 Caucasian subjects from the center visit. The study was approved by the medical ethics committee at Erasmus University Rotterdam, Rotterdam, the Netherlands, and all examined participants gave written informed consent.

Data collection

Data on tobacco smoking was collected during home interviews. Participants were asked about past and present cigarette, cigar and pipe smoking behavior and where then categorized into current, former and never tobacco smokers. Seven participants had missing smoking status and were therefore excluded from any analysis. During the center visit, weight and height were measured in standing position wearing normal cloths. All participants had blood samples taken during the visit to quantify DNA methylation, messenger RNA (mRNA) expression levels and other blood measurements.

DNA methylation data

DNA was extracted from whole peripheral blood (stored in EDTA tubes) by standardized salting out methods. Genome-wide DNA methylation levels were measured using Illumina Human Methylation 450K array[11]. In short, samples (500ng of DNA per sample) were first bisulfite treated using the Zymo EZ-96 DNA-methylation kit (Zymo Research, Irvine, CA, USA). Next, they were hybridized to the arrays according to the manufacturers protocol. The methylation percentage of a CpG site was reported as a beta-value ranging between 0 (no methylation) and 1 (full methylation).

Quality control of the samples was done with Genome Studio. A total number of 16 samples were removed: 7 had a sample call rate below 99%; 5 had incomplete bisulfite conversion and 4 had gender swaps. Quality control of the probes was done based on the detection p-value calculated with Genome Studio. Probes with a detection p-value of more than 0.01 in more than 1% of the samples were excluded. This resulted in a total set of 474,528 probes which were normalized using the Dasen option of the WateRmelon R-package[12].

mRNA expression data

Whole-blood was collected (PAXGene Tubes – Becton Dickinson) and total RNA was isolated (PAXGene Blood RNA kits - Qiagen). To ensure a constant high quality of the RNA preparations, all RNA samples were analysed using the Labchip GX (Calliper) according to the manufacturer's instructions. Samples with an RNA Quality Score > 7 were amplified and labelled (Ambion TotalPrep RNA), and hybridized to the Illumina HumanHT12v4 Expression Beadchips as described by the manufacturer's protocol. Processing of the Rotterdam Study RNA samples was performed at the Genetic Laboratory of Internal Medicine, Erasmus University Medical Center Rotterdam. The RS-III expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE 33828: 881 samples are available for analysis.

Illumina gene expression data was quantile-normalized to the median distribution and subsequently log2-transformed. The probe and sample means were centered to zero. Genes were declared significantly expressed when the detection p-values calculated by

GenomeStudio were <0.05 in more than 10% of all discovery samples. A total number of 21,238 probes with a detection p-value of less than 0.05 in more than 10% of the probes were included. Quality control was done using the eQTL mapping pipeline.

Statistical analysis

The characteristics of the study population were compared between current and never smokers using IBM SPSS Statistics version 21.0.0.1 (IBM Corp.). The p-values were calculated using independent sample T-test for continuous variables and Chi-square test for dichotomous variables.

Of the 150 SNPs discovered by CARDIoGRAMplusC4D, 96 were located within a gene[2]. In addition, 58 SNPs had known effect on expression of a gene within 1Mb as found in a large publically available blood cis-expression-quantitative trait loci (eQTL) database (FDR<0.05)[13]. We annotated these SNPs to 85 genes with an in-gene variant and 84 cis eQTL genes (Table S1). The methylation probes within and near these CAD related genes as provided by Illumina were included in the analysis. We excluded probes from the Infinium HD methylation SNP list with a minor allele frequency above 1% as provided by Illumina, since variations in these SNPs can cause bias in the methylation measurement[14]. We further excluded known cross-reactive probes, since they can introduce bias in the results[15]. The remaining 3,669 methylation probes were checked for association with tobacco smoking using a linear mixed model with the LME4 package in R version 3.1.0 with Dasen normalized beta-values of the CpG sites as outcome measure[16]. We first compared current to never smokers and then performed a sensitivity analysis on the identified CpG sites comparing current to former smokers. Covariates were selected based on known association with DNA methylation and different distributions between current and never smokers in our samples. The selected covariates with fixed effects were age, sex and BMI[17,18,19,20]. Houseman estimated white blood cell proportions were used as fixed effects to correct for cell mixture distribution[21]. Array number and position on array were added in the model as covariates with random effects to correct for batch effects. We corrected for multiple testing using a robust Bonferroni corrected p-value of 1.4×10^{-5} as the threshold for significance (0.05/3,669 probes).

To decease the possibility of confounding in our association, we further adjusted model in a second analysis for other possible confounders and mediators. This analysis included total cholesterol, triglyceride levels, systolic blood pressure, alcohol intake and type 2 diabetes mellitus.

Functional analysis

Since DNA methylation may have an effect on gene expression, we tested the association between DNA methylation and mRNA expression of nearby CAD related genes. In the 724 individuals under study, 16 genes out of 26 candidates were found to be expressed in blood. First, we regressed out the Houseman estimated white blood cell proportions, the erythrocytes and platelets cell counts, fasting state, RNA quality score, plate number, age and sex on the mRNA expression levels using a linear mixed model. We then regressed out the Houseman estimated white blood cell proportions, age, sex, array number and position on array on the beta-values of the CpG sites using a linear mixed model. The residuals of the mRNA expression levels and the residuals of the beta-values of the CpG sites were checked for association using a linear regression model. The p-value threshold for association was 1.4×10^{-3} (0.05/35 tests). Significant associations were verified with a mediation analysis with current versus never smoking as exposure, beta-values of the CpG site a mediator and mRNA expression levels as outcome using the mediation package in R[22].

Results

Characteristics of the participants under study are summarized in Table 1. Of the 724 subjects in the study, 195 were current smokers and 201 were never smokers. The mean age was 59.9 years. Among the current smokers 50% was male, among the never smoker 37% (p-value=0.008).

The 150 SNPs identified in the CAD GWAS were annotated to 85 genes with an in-gene variant and 84 cis eQTL genes within 1Mb of the identified SNPs (Table S1). These genes had 3669 methylation sites measured on the array within and near the gene as provided by Illumina. After correction for multiple testing, 15 CpG sites were significantly associated with current smoking (p-value<$1.4 \text{x} 10^{-5}$) (Table 2). Current tobacco smoking was associated with a 1.2 to 2.4 percent lower DNA methylation compared to never smoking in 12 of the CpG sites. In three CpG sites, current tobacco smoking was associated with a 1.2 to 1.8 percent higher DNA methylation. The effect estimates of the associations did not change when we further adjusted for total cholesterol, triglyceride levels, systolic blood pressure, alcohol intake and type 2 diabetes mellitus as potential confounders or mediators. In a sensitivity analysis in current smokers, two sites were significantly associated with cumulative exposure to tobacco smoking (Table 3).

When comparing current to former smokers, the effect estimates were lower and the differences were no longer significant in 10 of the 15 CpG sites (Table 2). This was confirmed in a sensitivity analysis in former smokers, which showed that cessation time was associated with differences in methylation level in three of the identified CpG sites (Table 3). The two top CpG sites, cg24908166 and cg12324353, were annotated to *TERT* (Table 3). The two CpG sites were positively correlated with each other (r=0.27, p-value<0.001) and were located within 1kb from each other in an intron of *TERT*. Two other CpG sites cg05603985 and cg09469355 were located within 1kb from each other in the first exon and intron of *SKI*.

Table 1. Characteristics of study population.

	Total	Current smokers	Former smokers	Never smokers	P-value*
N	724	195	319	210	
Age (years)	59.9 (8.2)	58.1 (6.7)	61.5 (8.6)	59.2 (8.4)	0.15
Sex (male)	334 (45.8%)	98 (50%)	155 (49%)	78 (37%)	0.008
BMI (kg/m^2)	27.6 (4.6)	26.8 (4.3)	28.0 (4.5)	27.7 (4.9)	0.05
Proportion CD8+ T-cells[†]	0.09 (0.05)	0.10 (0.05)	0.09 (0.06)	0.10 (0.05)	0.81
Proportion CD4+ T-cells[†]	0.26 (0.08)	0.27 (0.07)	0.26 (0.08)	0.26 (0.08)	0.33
Proportion NK cells[†]	0.14 (0.06)	0.11 (0.06)	0.14 (0.06)	0.14 (0.06)	<0.001
Proportion B-cells[†]	0.10 (0.04)	0.10 (0.04)	0.10 (0.03)	0.10 (0.04)	0.59
Proportion monocytes[†]	0.08 (0.03)	0.08 (0.03)	0.08 (0.03)	0.08 (0.03)	0.55
Proportion granulocytes[†]	0.37 (0.11)	0.38 (0.11)	0.37 (0.12)	0.37 (0.11)	0.12
Nr. erytrocytes (10^9/L)	4.90 (0.37)	4.86 (0.36)	4.93 (0.37)	4.88 (0.38)	0.67
Nr. platelets (10^9/L)	283.8 (63.9)	293.2 (69.3)	280.2 (61.0)	280.4 (62.4)	0.05
Fasting (yes)	717 (99%)	192 (99%)	315 (99%)	210 (100%)	0.07
Alcohol (g/d)	18.2 (11.3)	19.3 (12.8)	18.9 (11.5)	16.0 (8.9)	0.002
RNA quality score	8.38 (0.51)	8.36 (0.50)	8.39 (0.49)	8.38 (0.54)	0.66
Type 2 diabetes mellitus	53 (7.3%)	7 (3.6%)	34 (10.6%)	12 (5.7%)	0.31
Total cholesterol	5.6 (1.3)	5.6 (1.1)	5.6 (1.5)	5.6 (1.1)	0.88
Triglycerides	1.49 (0.87)	1.68 (1.14)	1.43 (0.68)	1.39 (0.80)	0.005
Systolic blood pressure	135.7 (61.7)	132.3 (61.7)	137.5 (64.3)	136.3 (57.6)	0.50
Diastolic blood pressure	84.9 (63.3)	83.5 (61.7)	85.6 (66.7)	85.1 (59.7)	0.79

Data are mean (SD) or n (%).

*Current versus never.

[†]Houseman estimated white blood cell proportions.

Table 2. Significant associations between tobacco smoking and methylation of CAD genes.

	Current-Never		*Adjusted model*		*Current-Former*	
CpG Site	**Estimate (se)**	**P-value**	**Estimate (se)**	**P-value**	**Estimate (se)**	**P-value**
cg24908166	-0.014 (0.003)	1.1×10^{-7}	-0.015 (0.003)	1.6×10^{-8}	-0.006 (0.002)	0.02
cg12324353	-0.012 (0.003)	1.3×10^{-7}	-0.012 (0.002)	3.5×10^{-7}	-0.007 (0.002)	2.8×10^{-3}
cg03725309	-0.024 (0.005)	2.0×10^{-7}	-0.022 (0.005)	3.8×10^{-6}	-0.017 (0.004)	4.2×10^{-5}
cg00980784	-0.015 (0.003)	2.5×10^{-7}	-0.014 (0.003)	3.5×10^{-6}	-0.015 (0.003)	2.4×10^{-8}
cg13916835	-0.021 (0.004)	3.0×10^{-7}	-0.021 (0.004)	1.3×10^{-6}	-0.013 (0.004)	1.5×10^{-3}
cg09469355	-0.017 (0.003)	7.7×10^{-7}	-0.017 (0.003)	8.5×10^{-7}	-0.015 (0.003)	3.8×10^{-6}
cg05603985	-0.014 (0.003)	8.4×10^{-7}	-0.013 (0.003)	3.6×10^{-5}	-0.014 (0.003)	1.3×10^{-7}
cg04324276	-0.015 (0.003)	3.2×10^{-6}	-0.015 (0.003)	1.2×10^{-5}	-0.010 (0.003)	2.4×10^{-4}
cg25468516	-0.015 (0.003)	3.6×10^{-6}	-0.014 (0.003)	3.0×10^{-5}	0.011 (0.003)	2.3×10^{-4}
cg22907952	-0.011 (0.002)	4.5×10^{-6}	-0.010 (0.003)	6.9×10^{-5}	-0.007 (0.002)	1.1×10^{-3}
cg15474579	-0.017 (0.004)	5.0×10^{-6}	-0.013 (0.004)	5.2×10^{-4}	-0.016 (0.003)	2.0×10^{-6}
cg20496896	0.012 (0.003)	6.7×10^{-6}	0.013 (0.003)	4.6×10^{-6}	0.005 (0.002)	0.06
cg09397246	0.018 (0.004)	1.0×10^{-5}	0.018 (0.004)	3.1×10^{-5}	0.017 (0.003)	5.6×10^{-7}
cg26405020	0.014 (0.003)	1.2×10^{-5}	0.014 (0.003)	1.4×10^{-5}	0.008 (0.003)	2.6×10^{-3}
cg18236066	-0.014 (0.003)	1.3×10^{-5}	-0.015 (0.003)	2.5×10^{-5}	-0.011 (0.003)	1.4×10^{-4}

Bonferroni corrected threshold 1.4×10^{-5}. Current-Never: adjusted for age, sex, BMI, Houseman estimates, batch effects. Adjusted model: adjusted for age, sex, BMI, Houseman estimates, batch effects, systolic blood pressure, total cholesterol, triglycerides, daily alcohol intake, type 2 diabetes mellitus. Current-Former: adjusted for age, sex, BMI, Houseman estimates, batch effects.

Table 3. Association between packyears and cessation time and methylation of significant CpG sites.

	Packyears[a]		*Cessation time*[b]	
CpG site	**Estimate (SE)**	**P-value**	**Estimate (SE)**	**P-value**
cg24908166	−0.001 (0.001)	0.24	−0.003 (0.001)	0.02
cg12324353	−0.003 (0.001)	2.3×10^{-3}	0.002 (0.001)	0.05
cg03725309	−0.002 (0.002)	0.15	0.004 (0.002)	0.02
cg00980784	−0.003 (0.001)	0.01	−0.001 (0.001)	0.69
cg13916835	−0.002 (0.002)	0.14	0.002 (0.002)	0.24
cg09469355	−0.003 (0.001)	0.04	−0.006 (0.001)	6.0×10^{-5}
cg05603985	−0.002 (0.001)	0.03	0.004 (0.001)	1.2×10^{-3}
cg04324276	−0.002 (0.001)	0.22	0.002 (0.001)	0.17
cg25468516	−0.001 (0.001)	0.20	0.003 (0.001)	0.06
cg22907952	−0.003 (0.001)	1.8×10^{-3}	0.002 (0.001)	0.12
cg15474579	−0.004 (0.001)	0.02	0.005 (0.001)	1.5×10^{-3}
cg20496896	0.002 (0.002)	0.12	−0.003 (0.001)	3.9×10^{-3}
cg09397246	0.001 (0.002)	0.35	−0.001 (0.002)	0.38
cg26405020	0.002 (0.001)	0.06	−0.002 (0.001)	0.05
cg18236066	−0.002 (0.001)	0.18	0.003 (0.001)	0.04

Bonferroni-corrected threshold 3.3×10^{-3}.

[a] In current smokers, per 10 packyears, adjusted for age, sex, BMI, Houseman estimates, batch effects.

[b] In former smokers, per 10 years of smoking cessation, adjusted for age, sex, BMI, Houseman estimates, batch effects.

Methylation of these two sites was positively correlated with each other (r=0.57, p-value<0.001).

CpG sites cg09397246 and cg26405020 were located within 1500 basepairs from the transcription start site of *FES.* These sites were within 2 basepairs from each other and had a positive correlation (r=0.88, p-value<0.001). The other significant hits were annotated to *SARS, GNGT2, SMG6*, *TOM1L2*, *SIPA1*, *MRAS*, *CDKN1A, LRRC2* and *RPH3A*. The beta-value distributions for all identified CpG sites stratified by the three smoking categories can be found in Figure S1.

The associations between the 15 CpG sites and mRNA expression of nearby CAD genes are shown in Table S2. Increased methylation of cg05603985 (*SKI*) was associated with increased expression of cis eQTL gene *PRKCZ* (estimate=0.035, p-value=1.4×10^{-4}). The mediation analysis however, was not significant (proportion mediated 0.24, p-value=0.79). The other CpG sites were not associated with gene expression.

Table 4. Characteristics of significant CpG sites.

					Average Beta Values (SD)		
CpG site	**Gene**	**Position***	**Placement**	**Island status**	**Current smokers**	**Former smokers**	**Never smokers**
cg24908166	*TERT*	5: 1268800	Body	N Shore	0.893 (0.031)	0.900 (0.029)	0.908 (0.024)
cg12324353	*TERT*	5: 1269197	Body	N Shore	0.846 (0.027)	0.851 (0.027)	0.855 (0.024)
cg03725309	*SARS*	1: 109757585	Body	S Shore	0.292 (0.057)	0.297 (0.060)	0.311 (0.063)
cg00980784	*GNGT2*	17: 47287577	TSS1500		0.331 (0.048)	0.340 (0.048)	0.341 (0.048)
cg13916835	*SMG6*	17: 2025181	Body		0.796 (0.049)	0.812 (0.047)	0.822 (0.039)
cg09469355	*SKI*	1: 2161886	Body	S Shore	0.500 (0.041)	0.516 (0.045)	0.520 (0.040)
cg05603985	*SKI*	1: 2161049	1st Exon	Island	0.347 (0.037)	0.362 (0.038)	0.364 (0.037)
cg04324276	*TOM1L2*	17: 17817462	Body		0.502 (0.047)	0.502 (0.046)	0.507 (0.047)
cg25468516	*SIPA1*	11: 65408028	5'UTR	N Shore	0.215 (0.048)	0.218 (0.051)	0.226 (0.054)
cg22907952	*MRAS*	3: 138121287	3'UTR		0.806 (0.033)	0.809 (0.034)	0.812 (0.030)
cg15474579	*CDKN1A*	6: 36753790	Body	N Shore	0.706 (0.058)	0.718 (0.052)	0.726 (0.047)
cg20496896	*LRRC2*	3: 46579532	Body		0.728 (0.056)	0.727 (0.055)	0.712 (0.056)
cg09397246	*FES*	15: 91427361	TSS1500†	N Shore	0.296 (0.066)	0.285 (0.065)	0.286 (0.062)
cg26405020	*FES*	15: 91427363	TSS1500†	N Shore	0.464 (0.057)	0.459 (0.059)	0.457 (0.055)
cg18236066	*RPH3A*	12: 113293823	Body		0.702 (0.040)	0.709 (0.038)	0.713 (0.041)

*Based on genome build 37, chromosome: position.

†1500bp from transcription start site.

Discussion

The results of the current study suggest an association between tobacco smoking and DNA methylation of 12 genes suggested to be associated to CAD via GWAS. One of these CpG sites was found to be associated with expression of nearby CAD-related gene *PRKCZ*.

We found that the effect estimates of tobacco smoking on DNA methylation decreased in 10 of the 15 CpG sites when comparing current to former smokers as compared to never smokers. This suggests that the effect of tobacco smoking on DNA methylation of these CpG sites is relatively sustained after smoking cessation.

The top two CpG sites, cg24908166 and cg12324353, are located within *TERT* (telomerase reverse transcriptase). High levels of *TERT* expression are found in macrophages of human atherosclerotic lesions[23].Two other CpG sites, cg09397246 and cg26405020, were located near the transcription start site of *FES* (FES proto-oncogene, tyrosine kinase), which has been identified by GWAS to be associated with blood pressure and hypertension[24]. Smoking was further associated with methylation of cg09469355 and cg05603985 within *SKI* (avian sarcoma viral oncogene homolog) which is a repressor of TGF-beta activity. Decreased TGF-beta activity is associated with atherosclerosis development and plaque instability[25,26]. This could be a plausible pathway through which smoking can increase the occurrence of CAD since smoking has already been associated with decreased plasma levels of TGF-beta and decreased expression of TGF-beta in bronchial cell lines[27,28].

Methylation of cg05603985, located in the first exon of *SKI*, was positively associated with the expression of the nearby CAD related gene *PRKCZ* even though this association did not survive the mediation analysis*. PRKCZ* is a known cis eQTL gene for the CAD SNP rs10797416 within *SKI* and is located approximately 100kb upstream of *SKI*. According to ENCODE (GSM788075, Farnham – USC, PBMC cells), cg05603985 (*SKI*) is located within a regulatory region which might suggest that the CpG site lies within an enhancer of *PRKCZ*[29]. *PRKCZ* (protein kinase C, zeta) is involved in proliferation, differentiation and secretion of almost all cell types including myocardial cells. Apart from human height and antipsychotic treatment response, it has not been related to any diseases in large GWAS studies[30,31].

None of the other CpG sites were associated with the expression of nearby CAD related genes. The lack of an association does not necessarily mean that methylation of these sites has no effect on expression but could result from an insufficient statistical power. This also applies for the non-significant mediation analysis. Furthermore, mRNA expression is tissue specific and an association can therefore not be found in whole blood. Finally, not all methylation sites in the human genome have an effect on mRNA expression. It could be that these methylation sites function through histone modification or DNA stability which could not be studied in the current work. Last, it could be that these sites are merely biomarkers of tobacco smoking[6].

The availability of DNA methylation and mRNA expression data from the same samples is a major strength of this study. Therefore, we were able to conduct an in depth exploration of the association between smoking, DNA methylation and mRNA expression of CAD related genes. Our study involved methylation and expression data from whole blood samples and not from vascular or lung tissue. This could be a limitation, since methylation and expression might be tissue specific. However, the relationship between smoking and DNA methylation has been confirmed in other tissues including lung tissue[32]. A second limitation is the challenge of gene annotation in GWAS. GWAS locate risk variants for the phenotype under study, but the underlying causal gene might be difficult to designate. To minimize this problem we limited our analysis to in-gene variants and variants with known cis eQTL effects. Therefore the CAD related genes in our study are more plausible to be actual causal variants for CAD, thus making the results more convincing.

Our study provides examples of CAD related genes of which differential methylation is associated with tobacco smoking. Whether or not these genes are in the causal pathway between smoking and coronary artery disease needs further elucidation as well as further efforts in large samples.

References

1. Alwan A. Global status report on noncommunicable diseases 2010: World Health Organization. 2011.

2. CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013; 45(1): 25-33.

3. Cheng S, Claggett B, Correia AW, et al. Temporal trends in the population attributable risk for cardiovascular disease: the atherosclerosis risk in communities study. *Circulation* 2014; 130(10): 820-8.

4. Vecoli C, Adlerstein D, Shehi E, et al. Genetic score based on high-risk genetic polymorphisms and early onset of ischemic heart disease in an Italian cohort of ischemic patients. *Thromb Res* 2014; 133(5): 804-10.

5. Niemiec P, Nowak T, Iwanicki T, et al. The -930A>G polymorphism of the CYBA gene is associated with premature coronary artery disease. A case-control study and gene-risk factors interactions. *Mol Biol Rep* 2014; 41(5): 3287-94.

6. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012; 13(7): 484-92.

7. Zeilinger S, Kuhnel B, Klopp N, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One* 2013; 8(5): e63812.

8. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet* 2011; 88(4): 450-7.

9. Joubert BR, Haberg SE, Nilsen RM, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect* 2012; 120(10): 1425-31.

10. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *European journal of epidemiology* 2013; 28(11): 889-926.

11. Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011; 6(6): 692-702.

12. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 2013; 14: 293.

13. Westra H-J, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. 2013; 45(10): 1238-43.

14. Zhi D, Aslibekyan S, Irvin MR, et al. SNPs located at CpG sites modulate genome-epigenome interaction. *Epigenetics* 2013; 8(8): 802-6.

15. Chen YA, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013; 8(2): 203-9.

16. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* 2014.

17. Koestler DC, Christensen B, Karagas MR, et al. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* 2013; 8(8): 816-26.

18. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* 2014.

19. Florath I, Butterbach K, Muller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum Mol Genet* 2014; 23(5): 1186-201.

20. Zhang FF, Cardarelli R, Carroll J, et al. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics* 2011; 6(5): 623-9.

21. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012; 13: 86.

22. Tingley D. YT, Keele L., Imai K. Mediation: R package for causal mediation analysis.

23. Gizard F, Heywood EB, Findeisen HM, et al. Telomerase activation in atherosclerosis and induction of telomerase reverse transcriptase expression by inflammatory stimuli in macrophages. *Arterioscler Thromb Vasc Biol* 2011; 31(2): 245-52.

24. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011; 478(7367): 103-9.

25. Lebastchi AH, Qin L, Khan SF, et al. Activation of human vascular cells decreases their expression of transforming growth factor-beta. *Atherosclerosis* 2011; 219(2): 417-24.

26. Mallat Z, Gojova A, Marchiol-Fournigault C, et al. Inhibition of transforming growth factor-beta signaling accelerates atherosclerosis and induces an unstable plaque phenotype in mice. *Circ Res* 2001; 89(10): 930-4.

27. Kamio K, Ishii T, Motegi T, et al. Decreased serum transforming growth factor-beta1 concentration with aging is associated with the severity of emphysema in chronic obstructive pulmonary disease. *Geriatr Gerontol Int* 2013; 13(4): 1069-75.

28. Samanta D, Gonzalez AL, Nagathihalli N, Ye F, Carbone DP, Datta PK. Smoking attenuates transforming growth factor-beta-mediated tumor suppression function through downregulation of Smad3 in lung cancer. *Cancer Prev Res (Phila)* 2012; 5(3): 453-63.

29. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489(7414): 57-74.

30. McClay JL, Adkins DE, Aberg K, et al. Genome-wide pharmacogenomic study of neurocognition as an indicator of antipsychotic treatment response in schizophrenia. *Neuropsychopharmacology* 2011; 36(3): 616-26.

31. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; 467(7317): 832-8.

32. Shenker NS, Polidoro S, van Veldhoven K, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet* 2013; 22(5): 843-51.

Supplementary material

Table S1. CARDIoGRAMplusC4D SNPs and annotated genes.

SNP	Chr	Position[a]	Within Gene	cis-eQTL gene
rs10797416	1	2172202	*SKI*	*C1orf86, PRKCZ*
rs11206510	1	55268627		
rs17114036	1	56735409	*PPAP2B*	
rs1490738	1	88910193		*GTF2B, PKN2*
rs4268379	1	109579760	*SARS*	
rs602633	1	109623034		
rs12127701	1	109639787	*MYBPHL*	
rs7515901	1	109641419	*MYBPHL*	
rs11806316	1	115555005		
rs11204666	1	148809752		*CTSS, CTSK*
rs4845625	1	152688691	*IL6R*	*IL6R, AQP10*
rs12125501	1	167540432	*NME7*	
rs6700559	1	198912696		*DDX59*
rs2292096	1	199093392	*CAMSAP1L1*	*DDX59*
rs2820315	1	200138887	*LMOD1*	*IPO9*
rs17465637	1	220890152	*MIA3*	
rs16986953	2	19805954		
rs515135	2	21139562		
rs7561273	2	24101018		
rs10495907	2	43852230		*DYNC2LI1*
rs6544713	2	43927385	*ABCG8*	
rs1561198	2	85663500		*VAMP8, USP39, VAMP5, GNLY*
rs2252641	2	145517931		
rs816889	2	151033541	*RND3*	
rs2351524	2	203589237	*NBEAL1*	
rs2571445	2	218391399	*TNS1*	*TNS1*
rs4566357	2	227630259	*COL4A4*	*COL4A4*
rs11718455	3	44031902		
rs11710224	3	46561282	*LRRC2*	
rs7642590	3	48074754	*MAP4*	*NME6*
rs11916151	3	88363366		
rs1393786	3	137336725	*PPP2R3A*	*PCCB*
rs2306374	3	139602642	*MRAS*	
rs4301033	3	151525308		*TSC22D2*
rs17655141	4	44814329		
rs17083481	4	54351705	*PDGFRA*	
rs17087335	4	57533340	*C4orf14*	
rs7356185	4	120386559	*USP53*	
rs1429141	4	148507517		

Table S1 (continued). CARDIoGRAMplusC4D SNPs and annotated genes.

SNP	Chr	Position^{a}	Within Gene	cis-eQTL gene
rs4469055	4	148605171		
rs6841581	4	148620640		
rs4690974	4	156613091		
rs7692387	4	156854759	*GUCY1A3*	
rs2736100	5	1339516	*TERT*	
rs10051876	5	87300520		
rs273909	5	131695252	*SLC22A4*	*SLC22A4, SLC22A5*
rs246600	5	142497090	*ARHGAP26*	
rs2294461	6	6559500	*LY86*	
rs9472428	6	12830159	*PHACTR1*	
rs883947	6	12895700	*PHACTR1*	
rs12526453	6	13035530	*PHACTR1*	
rs13211739	6	13070981	*PHACTR1*	
rs12205331	6	35006433	*ANKS1A*	
rs1321309	6	36746614		*CDKN1A*
rs3778448	6	39271179	*KCNK5*	*C6orf64*
rs10947789	6	39282900	*KCNK5*	
rs4613862	6	82668990		
rs17062853	6	134142738		
rs12190287	6	134256218	*TCF21*	
rs12663498	6	151045533	*PLEKHG1*	
rs2048327	6	160783522	*SLC22A3*	
rs6926458	6	160939856	*LPA*	
rs4252120	6	161063598	*PLG*	
rs1247351	6	161283909		
rs2023938	7	19003300	*HDAC9*	
rs972158	7	26301532	*SNX10*	
rs217	7	27917546	*JAZF1*	
rs1167800	7	75014132	*HIP1*	*HIP1*
rs2395858	7	106751669	*COG5*	*COG5, HBP1*
rs11556924	7	129450732	*ZC3HC1*	*KLHDC10*
rs4591971	7	130996735		
rs10237377	7	139403605	*PARP12*	*TBXAS1*
rs264	8	19857460	*LPL*	*LPL*
rs6984210	8	22089560	*BMP1*	
rs17485781	8	27943481	*C8orf80*	
rs2954029	8	126560154		
rs10962774	9	16958831		
rs3217992	9	21993223	*CDKN2B, CDKN2BAS1, MTAP*	

Table S1 (continued). CARDIoGRAMplusC4D SNPs and annotated genes.

SNP	Chr	Position[a]	Within Gene	cis-eQTL gene
rs16905599	9	22059144	*CDKN2BAS1*	
rs10965228	9	22072380	*CDKN2BAS1*	
rs1333049	9	22115503		
rs495828	9	135144688		*GBGT1, SURF6*
rs2505083	10	30375128	*KIAA1462*	
rs11238956	10	44069860		
rs17155842	10	44072639		
rs501120	10	44073873		
rs3748242	10	81904767		*ANXA11, SFTPD*
rs7074064	10	88673102	*BMPR1A*	
rs2246833	10	90995834	*LIPA*	*IFIT5, LIPA, IFIT1*
rs11191447	10	104642313	*C10orf32, AS3MT*	*NT5C2, C10orf32, ARL3*
rs12765878	10	105659612	*OBFC1*	
rs93139	11	9716184	*SWAP70*	
rs7116641	11	43653493		*HSD17B12*
rs12801636	11	65147893	*PCNXL3*	*SIPA1*
rs590121	11	74951798	*SERPINH1*	*GDPD5*
rs606452	11	74953826	*SERPINH1*	
rs974819	11	103165777		
rs9326246	11	116116943		
rs683800	11	125688966	*DCPS*	*FOXRED1, SRPR*
rs4762911	12	20052013		
rs4149033	12	21209077	*SLCO1B1*	
rs2681472	12	88533090	*ATP2B1*	*WDR51B*
rs6490029	12	110182840	*CUX2*	*SH2B3*
rs3184504	12	110368991	*SH2B3*	*SH2B3*
rs3809274	12	110528716		*SH2B3, ATXN2*
rs17630235	12	111076069		*TMEM116*
rs2891403	12	111621955	*RPH3A*	*OAS1*
rs2244608	12	119901371	*HNF1A*	*OASL, C12orf43*
rs11057841	12	123882696	*SCARB1*	
rs9319428	13	27871621	*FLT1*	
rs9316753	13	54365930		
rs10507753	13	68180277		
rs11617955	13	109616103	*COL4A1*	
rs7139492	13	109713796	*COL4A1*	
rs12873154	13	109718853	*COL4A1*	
rs4773144	13	109758713	*COL4A2*	
rs11619057	13	109806392	*COL4A2*	
rs9515201	13	109838799	*COL4A2*	

Table S1 (continued). CARDIoGRAMplusC4D SNPs and annotated genes.

SNP	Chr	Position[a]	Within Gene	cis-eQTL gene
rs9515203	13	109847624	*COL4A2*	
rs2895811	14	99203695	*HHIPL1*	
rs2146238	14	99242482	*CYP46A1*	
rs6494488	15	62811257		*RBPMS2, ANKDD1A*
rs11072794	15	76793637		*PSMA4*
rs7173743	15	76928839		*CTSH*
rs7181240	15	76932181		
rs2880765	15	83857466	*AKAP13*	*AKAP13*
rs17514846	15	89217554	*FURIN*	*FES, FURIN, MAN2A2*
rs2521501	15	89238392	*FES*	*FES, FURIN, UNC45A, MAN2A2, RCCD1*
rs7496815	15	89862501		
rs2281727	17	2064695	*SMG6*	*SRR, TSR1*
rs12936587	17	17484447		*SREBF1, PEMT, RASD1*
rs4299203	17	17818884	*LRRC48*	*SREBF1, C17orf39, DRG2, ATPAF2, TOM1L2*
rs2071167	17	39643045	*UBTF*	*ASB16, C17orf65, SLC4A1, SLC25A39, G6PC3, C17orf53, UBTF, RUNDC3A*
rs15563	17	44360192	*UBE2Z*	*ATP5G1, CALCOCO2, UBE2Z*
rs16948048	17	44795465		*GNGT2, PHOSPHO1*
rs4793721	17	47166312	*CA10*	
rs2070783	17	59760703	*PECAM1*	
rs4410190	18	18274198		
rs1122608	19	11024601	*SMARCA4*	*C19orf52, CARM1, SMARCA4*
rs892115	19	11124650	*SPC24*	*KANK2*
rs17318596	19	46628935		*BCKDHA, B3GNT8*
rs2075650	19	50087459	*TOMM40*	*PVRL2*
rs2288911	19	50141124	*APOC2, APOC4*	
rs8111989	19	50501048		*VASP, KLC3, CKM*
rs6088638	20	32934175	*ACSS2*	*GGT7, EDEM2*
rs867186	20	33228215	*PROCR, EDEM2*	*EIF6, ACSS2*
rs2832227	21	29454947	*C21orf7*	
rs9982601	21	34520998		*MRPS6*
rs1034565	22	18364211	*ARVCF*	
rs9608859	22	28997277		*SF3A1, MTFP1*

Chr denotes chromosome.

[a]Based on genome build 37.

Table S2. Association between methylation and expression of CAD and cis-eQTL genes.

CpG site	Gene	Probe	Estimate (se)	P-value
cg03725309	*SARS*	ILMN_1786972	0.003 (0.006)	0.57
cg00980784	*GNGT2*	ILMN_1671237	0.012 (0.008)	0.12
cg13916835	*SMG6*	ILMN_1695280	-0.009 (0.005)	0.11
	TSR1[a]	ILMN_1775761	-0.0009 (0.006)	0.88
	TSR1[a]	ILMN_2092232	-0.001 (0.006)	0.82
cg09469355	*SKI*	ILMN_1710598	-0.0004 (0.006)	0.95
	C1orf86[b]	ILMN_2097790	0.015 (0.009)	0.10
	PRKCZ[b]	ILMN_2253286	0.005 (0.006)	0.34
	PRKCZ[b]	ILMN_2386982	0.0009 (0.009)	0.92
	PRKCZ[b]	ILMN_1697267	0.012 (0.073)	0.09
cg05603985	*SKI*	ILMN_1710598	-0.0009 (0.007)	0.89
	C1orf86[b]	ILMN_2097790	0.013 (0.011)	0.24
	PRKCZ[b]	ILMN_2253286	0.006 (0.007)	0.41
	PRKCZ[b]	ILMN_2386982	0.009 (0.011)	0.42
	PRKCZ[b]	ILMN_1697267	0.035 (0.009)	1.4×10^{-4}
cg15474579	*CDKN1A*	ILMN_1784602	-0.006 (0.013)	0.65
cg04324276	*TOM1L2*	ILMN_1686261	0.005 (0.005)	0.36
	TOM1L2	ILMN_1711109	-0.005 (0.005)	0.41
cg25468516	*SIPA1*	ILMN_1682930	-0.010 (0.011)	0.35
	SIPA1	ILMN_2415536	0.044 (0.098)	0.65
cg09397246	*FES*	ILMN_1693650	-0.008 (0.009)	0.40
	FURIN[c]	ILMN_1790228	-0.001 (0.007)	0.16
	UNC45A[c]	ILMN_1726434	0.007 (0.006)	0.21
	UNC45A[c]	ILMN_2395932	-0.005 (0.007)	0.49
	MAN2A2[c]	ILMN_1815148	-0.010 (0.006)	0.12
cg26405020	*FES*	ILMN_1693650	-0.030 (0.012)	0.01
	FURIN[c]	ILMN_1790228	-0.018 (0.010)	0.09
	UNC45A[c]	ILMN_1726434	0.007 (0.008)	0.42
	UNC45A[c]	ILMN_2395932	-0.008 (0.009)	0.40
	MAN2A2[c]	ILMN_1815148	-0.010 (0.009)	0.24
cg18236066	*RPH3A*	ILMN_1693717	0.014 (0.007)	0.05
	RPH3A	ILMN_1663356	0.014 (0.007)	0.06
	OAS1	ILMN_1675640	-0.033 (0.025)	0.19
	OAS1	ILMN_2410826	-0.027 (0.025)	0.29
	OAS1	ILMN_1658247	0.010 (0.020)	0.62

Bonferroni p-value threshold 9.2×10^{-4}. Model: residual expression after adjustment for age, sex, batch effects, houseman estimated white blood cell proportions, erythrocytes and platelet cell counts, fasting state and RNA quality score associated with residual methylation after adjustment for age, sex, houseman estimated white blood cell proportions and batch effects. Estimates are changes in residual expression per percentage residual methylation increase.

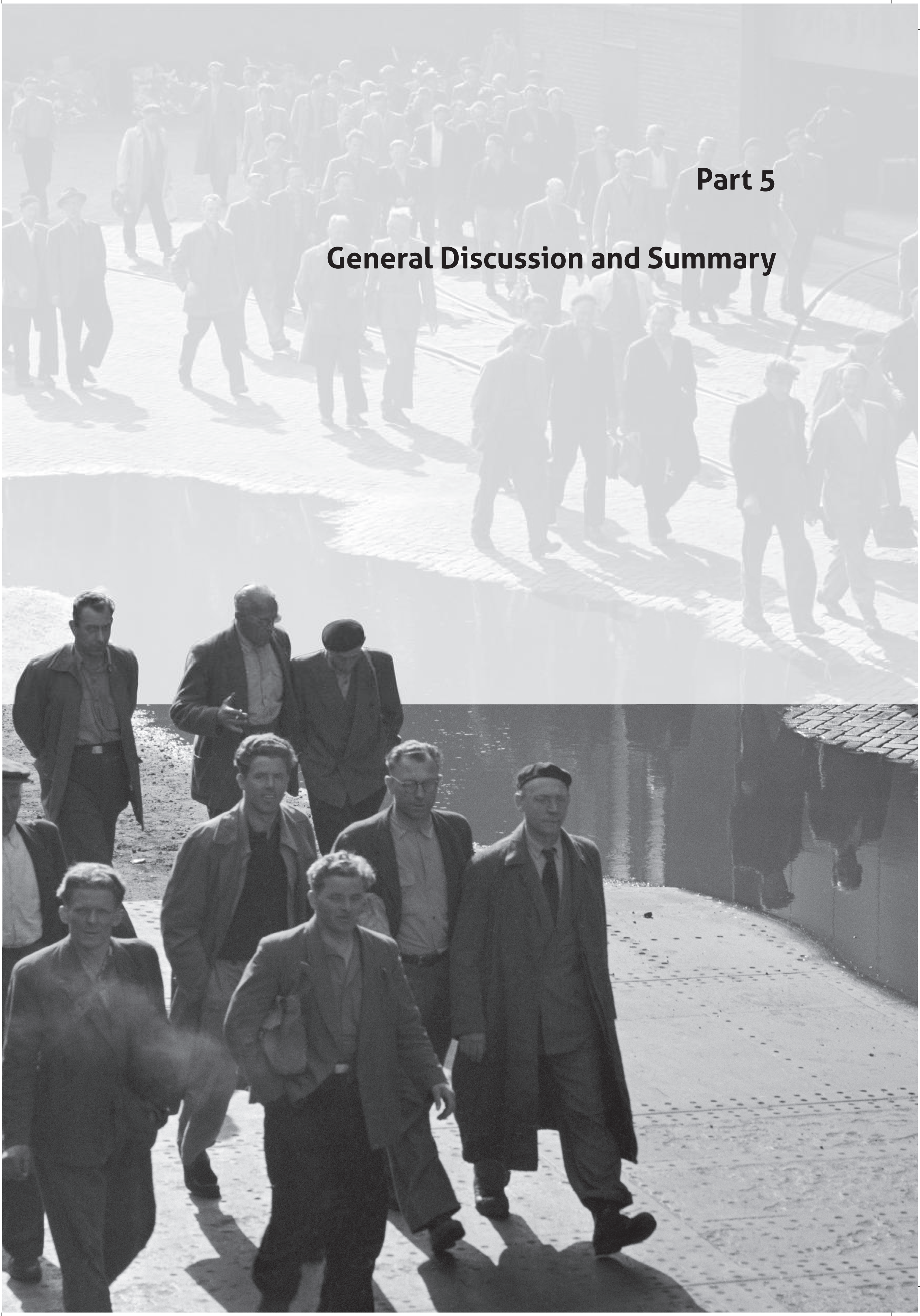
[a] Gene with cis-eQTL FDR<0.05 with rs2281727 (CARDIoGRAMplusC4D SMG6).

[b] Gene with cis-eQTL FDR<0.05 with rs10797416 (CARDIoGRAMplusC4D SKI).

[c] Gene with cis-eQTL FDR<0.05 with rs2521501/rs17514846 (CARDIoGRAMplusC4D FES/FURIN).

Part 5

General Discussion and Summary



Chapter 15

General discussion

The objective of this thesis was to provide further insights in the cause and consequence of inflammation in relation to diabetes and cardiovascular disease (CVD). I used advanced descriptive methods in epidemiology to study the population risk of type 2 diabetes and applied molecular epidemiology approaches including genomics, epi-genomics, and inflammation markers to study the link between chronic inflammation and diabetes and CVD. I performed a GWAS to identify genetic variants for circulating C-reactive protein (CRP) levels, and sought to unravel the causal role of chronic inflammation in cardiometabolic diseases. Next, I analysed DNA methylation data to identify methylation changes related to chronic inflammation, and studied smoking related methylation changes in genes identified for type 2 diabetes and CHD.

Here, I will discuss the main findings of this thesis and I will address important methodological issues that were encountered. Furthermore, future directions in the research of diabetes, CVD and molecular epidemiology are presented.

Main findings and interpretation

Lifetime risk of diabetes

In Chapter 2 of this thesis I estimated a lifetime risk of type 2 diabetes in the Netherlands comparable to estimations in the USA and Australia. I reported that 1 in 3 individuals are at risk of type 2 diabetes through their life. The numbers were even higher for prediabetes, showing that 1 in 2 will develop prediabetes at some point in their life, and 3 in 4 individuals with prediabetes will eventually progress to overt diabetes. These numbers illustrate the high lifetime risk of diabetes and indicate the importance of prevention early in life, even before prediabetes occurs. In agreement with data from the USA[1], I found that the lifetime risk of diabetes is more than 50% in severely obese individuals. With respect to treatment for diabetes, we estimated that eventually 1 in 10 individuals will require insulin treatment to successfully lower blood sugar levels. Altogether, these estimates underscore the public health problem of diabetes in Western society, and highlights the role of weight management in prevention of diabetes.

It has previously been shown that genetic information adds to the discriminative ability to identify individuals with high risk of diabetes[2,3,4]. In chapter 3, I could show that genetic information is useful in lifetime risk prediction of diabetes. Furthermore, individuals at high genetic risk that adhere to a normal weight have a substantial lower risk of diabetes compared to their obese counterparts (22% versus 58%). This observation again underscores the importance of weight management in prevention of diabetes, and suggest that adherence to a normal weight can offset high genetic risk of diabetes.

Inflammatory markers for diabetes and coronary heart disease

Inflammation is known to play an important role in the pathophysiology of diabetes and CHD[5,6], but CRP is unlikely to be causal[7,8]. It is thought that certain inflammatory processes contribute to cardiometabolic diseases and increase CRP levels. In part 2, I show the potential of exploring novel inflammation markers to provide a further understanding of the link between inflammation and disease. Using a panel of inflammation markers, I observed that higher EN-RAGE (Extracellular Newly identified Receptor for Advanced Glycation End-products binding protein) was associated with an increased risk of prediabetes, and higher IL-13 (interleukin-13) with an increased risk of diabetes. Higher IL-17 (interleukin-17) was associated with a lower risk of type 2 diabetes. The associations between EN-RAGE, IL-13, and IL-17 with prediabetes and diabetes were independent from the association between CRP and incidence of prediabetes and diabetes. In chapter 5, I found that individuals in the highest tertile of EN-RAGE had a 2.5-fold higher risk of CHD compared to the individuals in the lowest tertile. This association also appeared to be independent from CRP and other inflammatory markers.

Inflammation: genetics and beyond

In chapter 6, in a GWAS of CRP I confirmed 16 genetic variants previously identified for CRP, and found 42 novel distinct genetic variants. The identified genes were mainly annotated to liver metabolism and immune pathways, and mostly appeared to be independent from BMI. Although BMI increases CRP levels, this finding suggests that most of the CRP variation explained by genetics is independent from body fat mass. Interestingly, in the Mendelian randomization (MR) analysis I did observe a causal association between CRP and the risk of schizophrenia and bipolar disorder. Genetically higher CRP is associated with a decreased risk of schizophrenia and an increased risk of bipolar disorder. In chapter 8 and 9, I succeeded to identify shared genetic risk variants between CRP levels and cardiometabolic phenotypes, mainly related to liver metabolism (for example *HNF1A*, *HNF4A*, *APOC1*, and *GCKR)*. At several loci, I observed horizontal pleiotropy, i.e. the association with CRP appeared to be independent of the metabolic phenotype (*HNF1A* and *HNF4A*). Other loci were associated with CRP mediated through the metabolic phenotype (vertical pleiotropy), for instance the BMI-associated loci *FTO* and *TMEM18*. These pleiotropic findings provide further insights into the complex association between inflammation and metabolic phenotypes. Furthermore, in chapter 10 I carried out a bidirectional MR study on the association between vitamin D and CRP. I confirmed the association between vitamin D and CRP in observational data, but could not find any evidence for a causal association in neither direction. I may conclude that this association is likely to be confounded by shared risk factors.

Epigenetic landscape of inflammation

I conducted an EWAS on CRP levels in chapter 11 in order to describe the epigenetic landscape of chronic inflammation. I found robust evidence for associations between DNA methylation and circulating CRP levels at 45 genetic loci. DNA methylation mainly affect gene expression, and I could demonstrate associations between DNA methylation and nearby gene expression at 16% of the findings. Furthermore, several CpG sites were associated with risk of clinical diseases, including CHD. Also, in chapter 12 I performed an EWAS on the blood levels of the proinflammatory cytokine tumor necrosis factor α (TNFα). I demonstrated that DNA methylation at two genetic loci, namely *NLRC5* and *DTX3L-PARP9*, two immune response related genes, was associated with circulating TNFα levels. DNA methylation at *NLRC5* and *DTX3L-PARP9* was also associated with incident CHD. DNA methylation correlated with lower TNFα levels was associated with reduction of CHD risk. The CRP and TNFα EWAS show the potential of EWAS to identify epigenetic changes related to inflammation and open the way for further studies investigating the pathways that relate inflammation to changes in DNA methylation.

In addition, in chapter 13 and 14, I investigated the role of DNA methylation in the increased risk of type 2 diabetes and CHD in smokers. I observed that tobacco smoking is associated with differential methylation of *ANPEP*, *KCNQ1*, and *ZMIZ1* which are identified by GWAS for risk of type 2 diabetes. For CHD, I also observed associations between tobacco smoking and DNA methylation at CHD risk loci. Altogether, these data suggest that smoking contributes to an increased risk of disease through alterations in DNA methylation.

CRP and complex disease: cause, consequence, or epiphenomenon?

CRP and type 2 diabetes

Inflammation is known to play a key role in the pathogenesis of type 2 diabetes and inflammatory markers are shown to predict risk of type 2 diabetes[5]. First reported in 2001, Pradhan et al. investigated the association between serum CRP levels and risk of type 2 diabetes[9]. After adjustment for conventional diabetes risk factors, individuals in the highest quartile of CRP had a 4.2 times higher risk of diabetes compared to individuals in the lowest quartile. Numerous studies confirmed the association between serum CRP levels and risk of diabetes[10,11,12]. However, it has been debated whether the association between CRP and diabetes is causal or just represents an observation. One study has reported a positive association between a CRP haplotype and incidence of diabetes[13]. However, the finding was not replicated in subsequent studies[7,14], concluding that a causal role for CRP is unlikely. Furthermore, there is substantial evidence that inflammation is a consequence of obesity[15,16], which in turn is an important risk factor for diabetes[17]. In chapter 8 I focused on the shared genetics of inflammation and metabolic phenotypes, and observed genetic pleiotropy between CRP and glucose metabolism suggesting that shared genetics play a role in the association reported in observational data. In addition, the genetic pleiotropy between CRP and metabolic phenotypes appeared to be highly complex. First, several pleiotropic variants had a mediated effect on CRP levels, i.e. they have an effect on metabolic disturbances and subsequently affect CRP levels (vertical pleiotropy). Secondly, other pleiotropic variants are independently associated with CRP levels, thus not mediated through an intermediate phenotype (horizontal pleiotropy). Thirdly, I observed for some pleiotropic SNPs an opposite direction of effect between the effect allele and CRP from what is expected from observational data. For instance, although higher CRP is associated with an increased risk of diabetes, at the *APOC1* gene, the rs4420638-A allele is associated with higher CRP levels and lower type 2 diabetes risk. Altogether, from the findings of chapter 8 it is likely that CRP is both a consequence of metabolic disturbances and an epiphenomenon in the association with diabetes, rather than a cause. I may further conclude that revealing the shared genetics of associated phenotypes may provide understanding of the nature of the observed association.

CRP and coronary heart disease

There is ample evidence in observational studies that CRP is associated with risk of CHD[18]. However, several studies have rejected the hypothesis that CRP is causal to CHD[8,19]. In chapter 6, I performed MR analyses to study the causal role of CRP in the pathogenesis of CHD. There appeared to be an association between genetically elevated CRP and risk of CHD in the MR-Egger analysis. However, this might be due to the fact that the MR-Egger estimate relies on the InSIDE assumption. The InSIDE assumption denotes that the strength of the association between the genetic variants and CRP is independent from the strength of the direct pleiotropic effect of the genetic variants on CHD (outcome). The InSIDE assumption may be violated when the genetic variants are associated with a confounder of the CRP-CHD association, especially when the genetic variants are associated with a phenotype that is causally upstream of CRP. For the CRP-CHD association, this might be interleukin 6 or other phenotypes. In the weighted median (WM) and penalized weighted median (PWM) MR analyses, the InSIDE assumption is relaxed[20]. For CHD, the WM and PWM MR analyses showed evidence against a causal association between CRP and CHD. Finally, the single rs2794520 *CRP* variant was not associated with CHD. This association is not as powerful as others, however, is least likely to be affected by pleiotropy. Altogether, it is more likely that CRP is a consequence of other CHD risk factors, rather than a cause of CHD. In chapter 8, further evidence is provided for shared genetics between CRP and CHD, for instance at the *IL6* and *IL6R* loci. Interleukin 6 is the main determinant of CRP production in liver[21], and is identified in GWAS of CHD[22]. Altogether, these data support the hypothesis that CRP is not causal to CHD, but rather a consequence of CHD and CHD risk factors.

CRP and schizophrenia

Increased CRP levels are associated with increased risk of schizophrenia in observational studies[23]. Thus far, the general concept was that increased inflammation enhances the risk of schizophrenia[23]. However, in a recent MR effort using 18 genetic variants identified for CRP, the data suggested a causal role for CRP in the protection of schizophrenia[14]. In chapter 6, I extended the genetic score with further variants identified for CRP and observed a similar causal protective association. Also the single rs2794520 *CRP* variant showed a similar effect. Given that there is no genome-wide significant genetic risk variant for schizophrenia in linkage disequilibrium with the *CRP* locus, the observed protective effect seems to be genuine[24]. A hypothesis for this observation might be the immune response to infections early in life, i.e. a genetic profile that allows a stronger inflammatory response to stimuli might protect humans against infections that might induce schizophrenia later in life. A previous study has compared levels of acute-phase reactant proteins in dry blood spots collected at birth between patients with non-affective psychosis, which includes schizophrenia, and controls[25]. Cases of non-affective psychosis had lower levels of three acute-phase response proteins, namely serum amyloid P (SAP), tissue plasminogen activator (tPA), and procalcitonin, compared to controls, suggesting a less pronounced immune response at birth. In line with this observation, other studies have provided evidence that neonates with a history of severe infections have a higher risk of future schizophrenia[26,27]. Also, neonates that have been exposed to a maternal infection with cytomegalovirus or toxoplasma gondii, with low levels of acute-phase response proteins, have a higher risk of schizophrenia[28]. This evidence suggests that children with a less efficacious immune response may have chronic infection that over time contribute to the development of schizophrenia. Further research, possibly functional studies, are needed to elucidate the observed association between genetically elevated CRP levels and lower risk of schizophrenia.

Methodological considerations

Genetic studies

Since the first GWAS in 2005[29], GWAS have succeeded in discovering thousands of genetic loci associated with numerous phenotypes and clinical outcomes[30]. With the identification of novel genetic loci, further insight is provided into the biology underlying many diseases. By continuing the gene "hunting" GWAS, the number of associated loci is expected to continue increasing in the future by increasing the sample size and improving the imputation panels[31]. The first wave of GWAS have mainly identified common genetic variants with - relatively - larger effect sizes. Future GWAS will search for genetic loci that have smaller effect sizes or are less common. Some researchers have criticized the search for such variants as they consider them to be less important. It should be noted that such variants may play an equally important role in disease pathophysiology[32]. Novel variants with small effect sizes could provide insights into biological systems underlying phenotypic variance and may help in personalized risk prediction. As an example, the first GWAS in 1,087 participants on lipid levels did not identify *HMGCR*, a gene that is targeted by the most common lipid-lowering drug, i.e. statins[33]. Statins inhibit the 3-hydroxy-3-methylglutaryl-coenzyme A reductase, the protein product of *HMGCR*, and are highly effective in lowering cholesterol levels and the risk of cardiovascular disease[34]. It was not until 2010 when the *HMGCR* gene was identified in a GWAS meta-analysis including >100,000 individuals[35]. Individuals carrying the rs12916-C allele had on average 2.84 mg/dL higher total cholesterol, whereas other variants comprise >4mg/dL total cholesterol elevation. This example shows that novel loci with smaller effect sizes do not necessarily represent less important loci, but might even have the potential to significantly affect the disease pathogenesis.

Augmenting the sample size and denser mapping including low frequency variants may help to identify further genetic variants underlying complex phenotypes. In this thesis, I performed in chapter 6 two GWAS using different imputation panels (HapMap and 1000Genomes). The two CRP GWAS presented are not directly comparable, as the sample sizes were different between the two studies. In general, I identified more loci in the HapMap GWAS as the sample size was larger. Nevertheless, four loci were specific to the 1000Genomes GWAS, of which two were not available in the HapMap imputed data. Those two SNPs had low frequencies with minor allele frequency <0.05. The other two variants that were unique to the 1000Genomes GWAS were present in the HapMap GWAS, but showed less significant association. This may be due to the improved imputation of the variants in 1000Genomes compared to HapMap. I may conclude that the different imputation panels are complementary as they may identify specific loci, a finding which is in comparison to other studies[36,37]. Although I support the development of novel imputation panels such as the Haplotype Reference Consortium[38], the identification of novel loci is expected to be more noticeable when increasing sample sizes.

Most published GWAS to date are based on SNP arrays that tag common genetic variants across the genome. One of the major challenges in GWAS is the identification of the causal variant(s) at the associated genetic loci and the underlying mechanism by which the associated variants affect the phenotypic variance. Fine-mapping refers to the search for one or more causal variants at associated loci. With the increasing availability of whole-genome sequence data in large samples, imputation panels will be upgraded and the search for causal variants through fine-mapping will likely improve, but might be limited by available sample sizes. In line with the challenge of the identification of the causal variant is the follow up of GWAS findings in functional studies and translating them into targets for therapeutic interventions. This has been a major source of criticism on GWAS in last years.

However, with the advent of databases such as ENCODE[39] and the integration of gene expression[40], DNA methylation[41], histone modification amongst other, genes and causal pathways can be prioritized for follow-up studies such as wet lab experiments. It is probably a matter of time before novel treatments from GWAS findings are on the market to treat clinical diseases[42].

Epigenetic studies

Since the introduction of the Illumina® Beadchip DNA methylation assay technology for the quantification of DNA methylation at thousands of genetic loci across the human genome, the era of epigenome-wide association studies was started[43]. Illumina® first introduced the 27K array that assays DNA methylation at approximately 27,000 sites. A few years later Illumina® launched the Illumina® Infinium Human Methylation450K BeadChip. Now, already the third version of the Illumina® methylation chip, the MethylationEPIC BeadChip 850K, is available. As the human genome includes approximately 28 million CpG sites, the beadchips only cover a small part of the full epigenome, but include >99% of all genes. After the introduction of the arrays, many papers have focused on the pre-processing of the methylation array data[44,45,46]. Here, I will focus on the study design after the pre-processing, and discuss some methodological consideration of data analysis.

In this thesis, I analysed the methylation data using linear regression analysis. Traditionally, methylation has been analysed as the dependent variable. Depending on the research question of the researcher, one might be interested in the effect of DNA methylation on the phenotype or vice versa. For instance, when someone is interested in the influence of tobacco smoking on DNA methylation, smoking could be modelled as a determinant of DNA methylation. However, if someone is interested in the variance explained in serum CRP levels by a set of CpG sites, the DNA methylation should be introduced as the independent variable, and CRP as the dependent variable. Thus, depending on the research question, the DNA methylation variable is handled as an independent or dependent variable. I believe this is important in order to draw correct interpretation about the observed associations as effect estimates are different and not easy to convert. In chapter 11 I modelled DNA methylation as the dependent variable and CRP as an independent variable. However, for estimation of the variance explained in CRP by the methylation, I modelled CRP as the dependent variable. Note that power is not affected, thus the p-value should not be affected.

In population-based cohort studies, DNA methylation is mainly quantified in whole blood samples since blood samples can be collected in a non-invasive manner and can be easily obtained from the participants. DNA in whole blood samples is mostly composed of white blood cell DNA, since red blood cells and platelets do not contain DNA. However, there are numerous types of circulating white blood cells in whole blood (for instance monocytes, granulocytes, CD4+ cells, CD8+ cells), and DNA methylation patterns differ between the different types of white blood cells. Furthermore, white blood cell composition may be dependent on the outcome under investigation. Altogether, this may introduce confounding of the association between DNA methylation and the outcome under study ensuing false positive associations. To overcome the confounding by white cell composition, adjustment in the regression analysis for the different types of white blood cells is required. In some studies, different types of white blood cells are measured in the samples from which the DNA methylation is derived. These quantifications can be easily used for adjustment in the regression analysis. However, only a minority of the studies have measured different white blood cell types. To get an estimation of white blood cell composition, Houseman et al. introduced an imputation method to obtain white blood cell composition based on DNA methylation signatures[47]. Currently, the Houseman method is widely applied in EWAS to adjust for white blood cell composition. However, there are several limitations in the use of the white blood cell composition variables. First, Houseman et al. merely used five human adult samples to build the model for white blood cell composition estimation, and thus external validity may be limited. Second, the Houseman estimates have been used to impute cell counts in non-adult samples, for instance in the cord blood of neonates[48]. In this particular population, imputation performs much worse compared to older individuals[49]. Recently, a reference panel for cord blood cell proportions has been introduced[50], which shows much better performance compared to adult reference panel estimated cell compositions[51]. Altogether, inaccurate estimation of white blood cell composition may introduce serious confounding, especially when the outcome under study is related to white blood cell composition (e.g. immune-related phenotypes, auto-immune diseases). Further reference panels for specific populations with different characteristics are warranted.

In GWAS, cryptic relatedness should be accounted for in the association analysis to avoid false positive findings caused by confounding by population stratification. To detect population stratification in GWAS it is common to estimate the genomic control lambda (λ)[52]. The genomic control is based on the concept that only a few genetic variants are associated with the phenotype of interest, and the other variants follow the distribution under the null hypothesis of no association between the variant and the phenotype. After computing the λ, the association analyses are adjusted for the genomic control. Since most complex traits are highly polygenic, GWAS with large sample sizes show higher inflation that may present true associations. To distinguish polygenicity from confounding bias in GWAS, the relation between association results and linkage disequilibrium can be used to estimate the contribution of polygenicity and bias to the results using LD score regression[53]. When the first EWAS were published, the same methods were implemented to assess genomic inflation[54]. However, the interpretation of the inflation factor in EWAS is challenging since it is unknown how many CpG sites are expected to associate with the phenotype under study and there are no methods to estimate the contribution of bias to the inflation factor. In chapter 11, the QQ-plot showed inflation in the genome-wide DNA methylation in white blood cells with circulating CRP. As white blood cells are the main component of our immune system, and CRP is a sensitive marker of the overall immune response, I may expect many associations between DNA methylation sites and circulating CRP. Therefore, the assumptions underlying the genomic control hypothesis may not hold. I believe that future EWAS should provide a further understanding of the number of CpG sites that are expected to be associated with the phenotype under study in order to correctly interpret the inflation factor.

In the current thesis, I have performed both GWAS and EWAS of circulating CRP levels. GWAS and EWAS have certain similarities with respect to design, data processing, and analyses, however, there are definitely also differences with respect to design, analyses, and interpretation of the results. A major challenge in EWAS is the interpretation of the direction of the association results. In GWAS, as DNA sequence variants are inherited at random and do not alter during the life course (apart from somatic mutations), the causal inference is that genetic functional variants tagged by the associated genetic variants causally influence the phenotype. However, in EWAS this interpretation is more challenging since DNA methylation is affected by environmental factors and thus may change over time. To unravel the directionality of the association in EWAS, Mendelian Randomization (MR) methods can be used. Since I randomly inherit genes from our parents, genetic variant can be used to infer causality. As an example, in chapter 11 and 12 I studied the association between DNA methylation and circulating CRP and TNFα levels. The results could not be used to infer causality, i.e. it is not known whether cytokine levels are affecting DNA methylation, DNA methylation alters cytokine levels, or a common factor (such as obesity) is modifying both. I therefore applied MR analyses to infer causality. First I identified genetic variants associated with the DNA methylation sites (mQTLs), and subsequently associated the mQTLs with CRP and TNFα. I could not successfully assign a direction to the associations. This might be due to lack of power, since MR requires large samples to infer causality. Investigation of the association between baseline DNA methylation with prospective changes in the phenotype might provide further insights in the direction of the association[55], but is not conclusive. Future analytical techniques or wet lab experiments are necessary to infer correct causal inference of the associations observed in EWAS.

In genome- and epigenome-wide association studies researchers aim to find genes for the trait of interest through a hypothesis-free approach. In this thesis, one of the aims was to identify genetic loci for CRP levels. In both the EWAS and GWAS of CRP I identified genetic loci related to CRP levels. With respect to the findings in the EWAS and GWAS of CRP levels, I did not observe overlapping genetic loci. This is in line with findings from other studies on for instance BMI and glycemic phenotypes[55,56]. Also for TNFα, I found different genomic regions in the EWAS compared to the GWAS[57]. There are several possible explanation for the observation that GWAS and EWAS lack overlap of genetic loci. Methylation signals may truly target different genetic loci compared to GWAS and provide additional unique information. However, it could be that the findings from EWAS mostly represent DNA methylation changes attributable to the phenotypes, thus changes in DNA methylation due to inflammation, and rarely present causal associations as in GWAS in which the genetic variation alters inflammation. Furthermore, DNA sequence variation is present and similar in all tissues, whereas DNA methylation may differ between tissues and population-based DNA methylation studies merely study whole blood DNA since this is an easily accessible tissue. For instance for CRP levels, I learned from GWAS that genes active in liver tissue play an important role in CRP levels. Overlapping genetic loci in the liver may be missed by assessing DNA methylation in whole blood in relation to CRP.

Mendelian Randomization

Observational studies are limited by the fact that the observed associations cannot be automatically inferred to be causal[58]. Confounding, selection bias, or reverse causation may be a potential source of incorrect inference of associations in observational data[58]. One solution to overcome this shortcoming of observational studies are randomized controlled trials. In a randomized controlled trial individuals are randomly assigned to either receive the treatment under investigation or not. In this case, when all other variables are constant between the two groups, researchers may determine the effect of the treatment. However, commencing a randomized trial may be neither practical nor ethical to draw correct inference for associations observed in observational research. Therefore, alternative approaches to infer causality have been developed, such as MR[59]. According to the second law of Mendel, genes are randomly inherited from parents during meiosis. The basic principle of Mendelian randomization lays in the fact that, if the observed association is causal, genetic variants that affect the level of the exposure that itself is associated with the risk of the outcome, should be associated with the outcome. Owing to GWAS in large populations, over the last decade researchers have successfully identified many genetic factors that contribute to the risk of many diseases. These genetic risk variants may be used to construct powerful instrumental variables for MR analyses. As these data are now increasingly available to the scientific community, MR is now recognized as a valuable technique to draw inference of associations from observational studies.

In this thesis I applied MR analyses in several chapters. In chapter 6 I used different MR approaches to investigate the causal effect of CRP levels on multiple clinical outcomes. I observed a causal relation between CRP levels and schizophrenia, an association that I have extensively discussed above. Also in chapter 10, I used MR to infer causality regarding the association between vitamin D and CRP; no causal relationship could be established. Furthermore, in chapter 11 I applied MR to determine which CpG sites that were associated with CRP levels in a cross-sectional manner are a determinant of CRP levels. Unfortunately, I found no evidence for any causal association between the CpG sites as determinants for CRP levels.

Although MR is increasingly popular and has provided answers to several pending research questions, interpretation of MR studies should be done with caution. One of the major problems in MR is pleiotropic genetic effects, particularly when using multiple genetic variants. The use of multiple genetic variants as the instrumental variable has the advantage of increasing the variance explained of the exposure and thus increasing power[60]. However, this strategy increases the likelihood of pleiotropic effects. The variants used in the MR models may be associated with another intermediate phenotype, and thus the genetic variants serve as proxies for more than one intermediate factor. In this scenario of genetic pleiotropy, the MR assumption that the instrumental variable is independent from any confounding factor of the exposure-outcome association is violated. This problem also holds for CRP. As detailed in chapter 8, many genetic variants identified in the GWAS for CRP are associated with other metabolic phenotypes such as lipid levels, glycaemic phenotypes, and adiposity. Several statistical methods have been developed to overcome this issue, such as the MR-egger regression[61] and weighted median[20] approaches. However, although these techniques provide solutions to weaken the assumptions needed for a consistent estimation of the causal effect, they are not conclusive. The use of merely one genetic variant that directly influences the levels of the exposure, for instance the rs2794520 variant associated with CRP levels near to the *CRP* gene, will likely produce the most robust results. Furthermore, in MR studies individuals are randomized at conception. However, GWAS are commonly performed in large population-based or case-control studies that start at older ages, for instance 45 years of age for the Rotterdam Study. This may cause selection or survival bias. For instance, if individuals with higher CRP levels die prior to inclusion in the study and die prior to the development of chronic illnesses, than bias could occur because individuals with lower CRP levels live longer and may develop chronic illnesses such as CVD and schizophrenia.

Clinical implications

The first clinical implication of the findings from this thesis is the use of the lifetime risk estimates in diabetes risk communication to patients. Patients prefer absolute risks and long-term risks in the communication of disease risk. To get a sense of disease risk, absolute risks are easier to interpret compared to relative risk, and as a 40-year old is on average expected to live more than ten years, a long-term risk or lifetime risk provides a more comprehensive disease risk. The estimates provided in the first chapter are widely applicable in clinical settings. Furthermore, the lifetime risk estimates are highly useful information for public health services to estimate disease burdens in the general population. The addition of genetic information to disease prediction is already possible, but will likely further improve in the future with the discovery of novel and rare genetic risk variants increasing the discriminative ability.

Targeting inflammation with statins is successful in reducing CVD events[62]. Currently, promising studies are ongoing for inflammation lowering treatments in diabetes and CHD[63, 64]. In future, targeting specifically inflammation with immune-modulating agents may be an additional treatment to prevent cardiometabolic disease. Then, the findings from the CRP GWAS presented in this thesis could play a key role in the identification of genes and important causal pathways. Furthermore, the possible adverse consequences of immune treatment may be evaluated based on data I obtain from genetic pleiotropy studies. For instance, it has been suggested that long-term IL-1 inhibition for the treatment of rheumatic diseases may enhance the risk of future CVD[65]. Information about horizontal pleiotropic genetic effects may help to predict adverse effects of pharmacological treatments.

Genetic risk prediction is a promising strategy in the near future to identify individuals early in life that may benefit from an intervention or treatment lowering the risk of developing disease. In chapter 3 I showed that genetic data predicts the risk of developing diabetes. The identification of further variants associated with clinical disease may eventually improve risk stratification. It is waiting for studies that incorporate genetic data in risk prediction in clinical practice for common diseases.

Future directions

Today, GWAS have discovered hundreds to thousands of genetic variants associated with a diversion of phenotypes. Since the number of genetic variants in the genome is finite, it is likely that several causal variants overlap between distinct phenotypes. Further evidence for this hypothesis comes from the fact that causal mutations in Mendelian disorders are associated with different phenotypic features in the affected individual. Thus, it is likely that many genetic variants affect many phenotypes[32]. The increasing number of associated variants with phenotypes will extend the knowledge on genetic pleiotropy. I used published GWAS that probably reflect the "tip of the iceberg" with respect to associated genetic loci, and thus the extension of sample sizes in future GWAS will produce an incredible amount of information with respect to genetic pleiotropy. The identification of shared genes and pathways between associated phenotypes and diseases may alter classification and treatment of diseases. Resources such as the UK Biobank[66] and Million

Veteran Program[67] including hundreds of thousands of individuals will be of incredible value in the identification of further genetic variants for complex diseases and to improve the understanding of cellular networks[32].

Most genetic studies on inflammation are conducted in population-based cohort studies to study chronic inflammation in the general population. Not much is known about the genetic background of acute inflammation in the setting of an infectious illness such as sepsis that requires hospital admission. A previous study has failed to find robust evidence for genetic variants underlying sepsis mortality[68]. Well-designed genetic studies are warranted to find DNA sequence variants underlying acute inflammation.

Causal inference in EWAS remains challenging. In the majority of the published reports, the causal direction of the association between DNA methylation and the phenotype under study could not be established. A major limitation in unravelling the causality in DNA methylation studies is the lack of strong genetic instruments for MR analyses. Therefore, large studies are necessary to identify genetic instruments for DNA methylation and phenotypes to improve the power to successfully conduct robust MR analyses. Also, the findings in part 4 may be the start for wet lab experiments to unravel the role of the DNA methylation findings in inflammation and cardiometabolic disease. For instance the DNA methylation findings near the gene *AIM2* in the CRP epigenetic study, what is the role of *AIM2* in low-grade inflammation and may targeting *AIM2* help to lower inflammation?

Concluding remarks

In this thesis, I have found genetic and epigenetic markers for inflammation and sought to disentangle the complex interplay between inflammation and diabetes and CVD. In this chapter I gave an overview of the findings, discussed methodological issues, and provided my view on the future directions. Exciting times are coming as genetic and epigenetic findings are expected to being translated into clinical practice in coming years improving prediction, prevention, and treatment of disease.

References

1. Narayan KV, Boyle JP, Thompson TJ, Gregg EW, Williamson DF. Effect of BMI on lifetime risk for diabetes in the US. *Diabetes Care* 2007; 30(6): 1562-6.
2. Van Hoek M, Dehghan A, Witteman JC, et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies. *Diabetes* 2008; 57(11): 3122-8.
3. Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *The New England journal of medicine* 2008; 359(21): 2208-19.
4. Talmud PJ, Cooper JA, Morris RW, et al. Sixty-five common genetic variants and prediction of type 2 diabetes. *Diabetes* 2014: DB_141504.
5. Esser N, Legrand-Poels S, Piette J, Scheen AJ, Paquot N. Inflammation as a link between obesity, metabolic syndrome and type 2 diabetes. *Diabetes Res Clin Pract* 2014; 105(2): 141-50.
6. Libby P. Inflammation in atherosclerosis. *Arterioscler Thromb Vasc Biol* 2012; 32(9): 2045-51.
7. Brunner EJ, Kivimaki M, Witte DR, et al. Inflammation, insulin resistance, and diabetes—Mendelian randomization using CRP haplotypes points upstream. *PLoS Med* 2008; 5(8): e155.
8. C Reactive Protein Coronary Heart Disease Genetics Collaboration. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ* 2011; 342: d548.
9. Pradhan AD, Manson JE, Rifai N, Buring JE, Ridker PM. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA* 2001; 286(3): 327-34.
10. Barzilay JI, Abraham L, Heckbert SR, et al. The relation of markers of inflammation to the development of glucose disorders in the elderly. *Diabetes* 2001; 50(10): 2384-9.
11. Festa A, D'Agostino R, Tracy RP, Haffner SM. Elevated levels of acute-phase proteins and plasminogen activator inhibitor-1 predict the development of type 2 diabetes. *Diabetes* 2002; 51(4): 1131-7.
12. Hu FB, Meigs JB, Li TY, Rifai N, Manson JE. Inflammatory markers and risk of developing type 2 diabetes in women. *Diabetes* 2004; 53(3): 693-700.
13. Dehghan A, Kardys I, de Maat MP, et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. *Diabetes* 2007; 56(3): 872-8.
14. Prins BP, Abbasi A, Wong A, et al. Investigating the causal relationship of C-reactive protein with 32 complex somatic and psychiatric outcomes: a large-scale cross-consortium Mendelian randomization study. *PLoS Med* 2016; 13(6): e1001976.
15. Timpson NJ, Nordestgaard BG, Harbord RM, et al. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *International Journal of Obesity* 2011; 35(2): 300.
16. Holmes MV, Lange LA, Palmer T, et al. Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. *American journal of human genetics* 2014; 94(2): 198-208.

17. Prospective Studies Collaboration. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet (London, England)* 2009; 373(9669): 1083-96.

18. Emerging Risk Factors Collaboration. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet (London, England)* 2010; 375(9709): 132-40.

19. Elliott P, Chambers JC, Zhang W, et al. Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. *Jama* 2009; 302(1): 37-48.

20. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol* 2016; 40(4): 304-14.

21. Morrone G, Ciliberto G, Oliviero S, Arcone R, Dente L, Cortese R. Recombinant interleukin 6 regulates the transcriptional activation of a set of human acute phase genes. *J Biol Chem* 1988; 263(25): 12554-8.

22. CARDIoGRAMplusC4D Consortium. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013; 45(1): 25-33.

23. Singh B, Chaudhuri TK. Role of C-reactive protein in schizophrenia: an overview. *Psychiatry Res* 2014; 216(2): 277-85.

24. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; 511(7510): 421.

25. Gardner R, Dalman C, Wicks S, Lee B, Karlsson H. Neonatal levels of acute phase proteins and later risk of non-affective psychosis. *Transl Psychiatry* 2013; 3(2): e228.

26. Dalman C, Allebeck P, Gunnell D, et al. Infections in the CNS during childhood and the risk of subsequent psychotic illness: a cohort study of more than one million Swedish subjects. *Am J Psychiatry* 2008; 165(1): 59-65.

27. Koponen H, Rantakallio P, Veijola J, Jones P, Jokelainen J, Isohanni M. Childhood central nervous system infections and risk for schizophrenia. *Eur Arch Psychiatry Clin Neurosci* 2004; 254(1): 9-13.

28. Blomström Å, Gardner R, Dalman C, Yolken R, Karlsson H. Influence of maternal infections on neonatal acute phase proteins and their interaction in the development of non-affective psychosis. *Transl Psychiatry* 2015; 5(2): e502.

29. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; 308(5720): 385-9.

30. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *American journal of human genetics* 2017; 101(1): 5-22.

31. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* 2007; 17(10): 1520-8.

32. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 2017; 169(7): 1177-86.

33. Kathiresan S, Manning AK, Demissie S, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet* 2007; 8(1): S17.

34. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20 536 high-risk individuals: a randomised placebocontrolled trial. *Lancet (London, England)* 2002; 360(9326): 7-22.

35. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical, and population relevance of 95 loci for blood lipids. *Nature* 2010; 466(7307): 707.

36. de Vries PS, Sabater-Lleal M, Chasman DI, et al. Comparison of HapMap and 1000 genomes reference panels in a large-scale genome-wide association study. *PloS One* 2017; 12(1): e0167742.

37. Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet* 2012; 20(7): 801.

38. Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; 48(10): 1279-83.

39. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489(7414): 57.

40. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015; 47(9): 1091-8.

41. Kato N, Loh M, Takeuchi F, et al. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat Genet* 2015; 47(11): 1282.

42. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. *Nat Genet* 2015; 47(8): 856.

43. Bibikova M, Le J, Barnes B, et al. Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* 2009; 1(1): 177-200.

44. Pidsley R, Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 2013; 14(1): 293.

45. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 2012; 4(3): 325-41.

46. Liu J, Siegmund KD. An evaluation of processing methods for HumanMethylation450 BeadChip data. *BMC Genomics* 2016; 17(1): 469.

47. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012; 13(1): 86.

48. Joubert BR, Felix JF, Yousefi P, et al. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *American journal of human genetics* 2016; 98(4): 680-96.

49. Yousefi P, Huen K, Quach H, et al. Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. *Environ Mol Mutagen* 2015; 56(9): 751-8.

50. Bakulski KM, Feinberg JI, Andrews SV, et al. DNA methylation of cord blood cell types: applications for mixed cell birth studies. *Epigenetics* 2016; 11(5): 354-62.

51. Cardenas A, Allard C, Doyon M, et al. Validation of a DNA methylation reference panel for the estimation of nucleated cells types in cord blood. *Epigenetics* 2016; 11(11): 773-9.

52. Zheng G, Freidlin B, Gastwirth JL. Robust genomic control for association studies. *Am J Hum Genetic* 2006; 78(2): 350-6.

53. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. 2015; 47(3): 291-5.

54. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet (London, England)* 2014; 383(9933): 1990-8.

55. Demerath EW, Guan W, Grove ML, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet* 2015; 24(15): 4464-79.

56. Hidalgo B, Irvin MR, Sha J, et al. Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network study. *Diabetes* 2014; 63(2): 801-7.

57. Melzer D, Perry JR, Hernandez D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* 2008; 4(5): e1000072.

58. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004; 33(1): 30-42.

59. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003; 32(1): 1-22.

60. Pierce BL, Ahsan H, VanderWeele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol* 2010; 40(3): 740-52.

61. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015; 44(2): 512-25.

62. Ridker PM, Danielson E, Fonseca FA, et al. Reduction in C-reactive protein and LDL cholesterol and cardiovascular event rates after initiation of rosuvastatin: a prospective study of the JUPITER trial. *Lancet (London, England)* 2009; 373(9670): 1175-82.

63. Ridker PM, Thuren T, Zalewski A, Libby P. Interleukin-1β inhibition and the prevention of recurrent cardiovascular events: rationale and design of the Canakinumab Anti-inflammatory Thrombosis Outcomes Study (CANTOS). *Am Heart J* 2011; 162(4): 597-605.

64. Ridker PM. Testing the inflammatory hypothesis of atherothrombosis: scientific rationale for the cardiovascular inflammation reduction trial (CIRT). *J Thromb Haemost* 2009; 7(s1): 332-9.

65. Interleukin 1 Genetics Consortium. Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: a Mendelian randomisation analysis. *Lancet Diabetes Endocrinol* 2015; 3(4): 243-53.

66. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12(3): e1001779.

67. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016; 70: 214-23.

68. Rautanen A, Mills TC, Gordon AC, et al. Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study. *The Lancet Respiratory medicine* 2015; 3(1): 53-60.

Summary

Inflammation plays a pivotal role in the pathogenesis of type 2 diabetes and coronary heart disease (CHD). C-reactive protein (CRP) is a sensitive marker of inflammation and has been of major interest in epidemiological studies of diabetes and CHD. In this thesis, we aimed to study the link between inflammation and diabetes and CHD with the use of genetic and epigenetic information. For this, we used data from population based cohort studies and collaborated with numerous studies worldwide in large consortia.

The first part of this thesis is dedicated to the descriptive epidemiology of prediabetes and type 2 diabetes. In chapter two, we estimated that the lifetime risk of developing diabetes during the lifespan is approximately one in two for prediabetes, and one in three for over type 2 diabetes. The vast majority of individuals with prediabetes at the age of 45 will eventually develop overt diabetes. The lifetime risk is highly dependent on body mass index (BMI), and attenuates with advancing age. To investigate the effect of genetic information on the lifetime risk of diabetes, we constructed in chapter three a genetic score for diabetes based on genetic risk variants derived from large genome-wide association studies for diabetes. We found that genetic information adds to the lifetime risk prediction of diabetes, and the data showed that lean individuals at high genetic risk have a substantial lower risk of developing diabetes compared to their obese counterparts. This observation suggest that a healthy BMI may offset high genetic risks.

To explore other inflammatory markers with risk of diabetes, we performed in chapter four a biomarker study in which we tested several biomarkers with the risk of diabetes. We found that higher EN-RAGE and higher IL-13 were associated with a higher risk of diabetes. We also explored novel inflammatory markers for CHD, and observed that individuals with higher levels of EN-RAGE are at increased risk of CHD, mainly hard CHD. We suggest that EN-RAGE may have a role in the pathogenesis of diabetes and CHD, or may add to the risk prediction of these phenotypes.

Part three is devoted to the identification of genetic variants underlying circulating CRP, and overlapping genetic variants between CRP and cardiometabolic phenotypes. In Chapter six we performed a meta-analysis of genome-wide association studies of CRP levels. Both a HapMap imputed GWAS meta-analysis, and a 1000Genomes imputed meta-analysis were conducted. We confirmed 18 established genetic loci for CRP, and found 40 additional loci. Mendelian randomization analyses suggested a causal link between CRP and schizophrenia, whereas the association between CRP and CHD is likely to be confounded by pleiotropic genetic variants. The causal association between CRP and schizophrenia may be explained by the response of the immune system in early life as discussed in chapter seven. In chapter eight and nine, we studied the genetic overlap between CRP and cardiometabolic diseases, and observed numerous overlapping genetic variants. The pleiotropic genetic architecture underlying CRP and cardiometabolic phenotypes is highly complex, and further research will likely result in further insights in the shared pathways. In chapter ten, we sought to study the hypothesis that vitamin D is causally associated with levels of CRP. The data suggested no causal association, and therefore we may conclude that the association between vitamin D and CRP is likely to be confounded.

The role of DNA methylation in inflammation and cardiometabolic phenotypes is discussed in part four. In chapter 11, the association between epigenome-wide DNA methylation and circulating CRP levels was studied. We identified numerous CpG sites that were associated with CRP levels, and could identify associations between those CpG sites and nearby gene expression. The epigenome-wide association study of CRP provided further insights in the epigenetic landscape of general inflammation. Also, in chapter 12 we found CpG sites associated with circulating TNFα levels. The TNFα-associated methylation sites were associated with incident CHD. DNA methylation correlated with lower TNFα levels were associated with reduction of incident CHD, a finding that is in line with other observational studies. As smoking materially affects DNA methylation and increases the risk of diabetes and CHD, in chapter 13 and 14 we sought to identify DNA methylation changes attributable to smoking at genes known to be causally involved in the risk of diabetes and CHD. For both diabetes and CHD, we observed associations between smoking and DNA methylation nearby genes identified for diabetes and CHD, suggesting that smoking affects the risk of those disease through alterations in DNA methylation.

Samenvatting

Inflammatie speelt een cruciale rol in de pathogenese van type 2 diabetes en coronaire hartziekte. C-reactive protein (CRP) is een gevoelig meetinstrument voor inflammatie en in de afgelopen jaren is er veel epidemiologische onderzoek gedaan naar de relatie tussen CRP en diabetes en coronaire hartziekte. In deze thesis wilden we de relatie tussen inflammatie en diabetes en coronaire hartziekte bestuderen met gebruik van genetica en epigenetica. Om dit te bewerkstelligen hebben we data van bevolkingsonderzoeken over de hele wereld gebruikt en samengewerkt met andere universiteiten in grote internationale consortia.

Het eerste deel van deze thesis is gewijd aan de beschrijvende epidemiologie van prediabetes en type 2 diabetes. In hoofdstuk twee hebben we berekend dat het levenslang risico om prediabetes te ontwikkelen één op twee is, en dat het levenslang risico om type 2 diabetes te ontwikkelen één op drie is. Het overgrote merendeel van de mensen met prediabetes op de leeftijd van 45 jaar zal uiteindelijk type 2 diabetes ontwikkelen. Bovendien zagen we dat het levenslang risico op diabetes sterk afhankelijk is van de body mass index, en het levenslang risico daalt naarmate men ouder wordt. Om het effect van genetische informatie op het levenslang risico van diabetes te bestuderen hebben we in hoofdstuk drie een genetische score gemaakt gebaseerd op genetische risico varianten die in grootschalig genetisch onderzoek van diabetes zijn gevonden. We hebben aangetoond dat genetische informatie bijdraagt aan de voorspelling van het levenslange risico op diabetes, en de data liet zien dat magere mensen met een hoog genetisch risico een substantieel lager risico hebben op het ontwikkelen van diabetes in vergelijking met obese mensen met een hoog genetisch risico. Deze observatie suggereert dat een lage body mass index een hoog genetisch risico kan compenseren.

In de zoektocht naar potentiele nieuwe inflammatoire markers die geassocieerd zijn met het risico op diabetes hebben we in hoofdstuk vier een biomarker studie verricht waarin we verschillende biomarkers hebben onderzocht met het risico op diabetes. We vonden dat een hoger EN-RAGE en hoger IL-13 waren geassocieerd met een hoger risico op diabetes. We hebben ook nieuwe inflammatoire biomarkers onderzocht voor coronaire hartziekte. In deze studie vonden we dat hogere EN-RAGE waarden geassocieerd waren met een hoger risico op coronaire hartziekte. Deze resultaten wijzen op een mogelijke rol voor EN-RAGE in de pathogenese van diabetes en coronaire hartziekte en een potentiele meerwaarde in de risico voorspelling van deze beide ziekten.

Deel drie van deze thesis is gewijd aan het identificeren van genetische varianten voor CRP waarden in bloed, en het identificeren van overlappende genetische varianten tussen CRP en cardiometabole fenotypes. In hoofdstuk zes hebben we een meta-analyse van genoomwijde associatie studies naar CRP waarden uitgevoerd. We hebben zowel een HapMap geïmputeerde als een 1000Genomes geïmputeerde meta-analyse verricht. De studie bevestigde 18 eerder gevonden genetische locaties in het genoom die geassocieerd zijn met CRP, en resulteerde in 40 nieuwe associaties met CRP. Analyses waarin

Mendeliaans randomiseren werd toegepast suggereerde een causaal verband tussen CRP en schizofrenie, daar waar het causaal verband tussen CRP en coronaire hartziekte waarschijnlijk veroorzaakt wordt door pleiotrope genetische varianten. Het causale verband tussen CRP en schizofrenie wordt mogelijks verklaard door de reactie van het immuunsysteem vroeg in het leven, zoals beschreven in hoofdstuk zeven. In hoofdstuk acht en negen hebben we de genetische overlap tussen CRP en cardiometabole ziekten bestudeerd, waarbij we verscheidene overlappende genetische varianten hebben kunnen aanduiden. De pleiotrope genetische architectuur van CRP en cardiometabole fenotypes is zeer complex, en het is ter verwachten dat toekomstig onderzoek meer inzichten kan geven in de gedeelde biologie van CRP en cardiometabole fenotypes. In hoofdstuk tien hebben we de hypothese dat vitamine D een causaal verband heeft met CRP waarden bestudeerd. De data suggereerde geen causaal verband en daarom concluderen we dat het verband tussen vitamine D en CRP waarschijnlijk beïnvloed wordt door andere factoren.

De rol van DNA methylatie bij inflammatie en cardiometabole fenotypes hebben we in deel vier van deze thesis bekeken. In hoofdstuk 11 hebben we het verband tussen epigenoomwijde DNA methylatie en CRP waarden in bloed bestudeerd. We vonden verscheidene CpG-gebieden die geassocieerd waren met CRP, en we konden verbanden leggen tussen deze CpG-gebieden en genexpressie in de buurt van de CpG-gebieden. De epigenoomwijde associatie studie van CRP heeft nieuwe inzichten gegeven in het epigenetische landschap bij inflammatie. In hoofdstuk 12 hebben we ook CpG gebieden gevonden die geassocieerd waren met TNFα waarden in het bloed. De TNFα geassocieerde methylatie gebieden waren tevens geassocieerd met het vroegtijdig krijgen van coronaire hartziekte. DNA methylatie dat correleerde met lagere TNFα warden in het bloed was geassocieerd met een legere kans op het krijgen van coronaire hartziekte. Deze bevinding komt overeen met andere observationele studies. Gezien roken een sterk effect heeft op DNA methylatie en het risico op diabetes en coronaire hartziekte verhoogd, hebben we in hoofdstuk 13 en 14 het verband tussen roken en DNA methylatie veranderingen in genen die gekend zijn voor diabetes en coronaire hartziekte onderzocht. Voor zowel diabetes als coronaire hartziekte hebben verbanden gevonden tussen roken en de DNA methylatie van genen die gekend zijn voor diabetes en coronaire hartziekte. Dit suggereert dat roken mogelijk een effect op deze ziekten heeft via veranderingen in DNA methylatie.

Acknowledgements

Dit proefschrift zou niet tot stand gekomen zijn zonder de hulp van velen. Daarom wil ik deze hier van de gelegenheid gebruik maken om iedereen te bedanken voor hun tijd en bijdrage aan mijn proefschrift.

In alle artikelen zijn gegevens gebruikt die verzameld zijn binnen het kader van de Rotterdam studie. Daarom wil ik alle deelnemers van de Rotterdam studie bedanken voor hun deelname; zonder jullie is er geen onderzoek mogelijk. Het is prachtig om te zien hoeveel mensen gepassioneerd deelnemen aan dit grootschalig gezondheidsonderzoek.

Geachte prof. dr. M.A. Ikram, beste Arfan, bedankt dat je mijn promotor wilde zijn en bedankt voor jouw inzet tijdens het voltooien van mijn thesis. Ik heb goede herinneringen aan de lange gesprekken en discussies tijdens de CHARGE meetings en de heen- en terugvlucht achter in het vliegtuig. Je bent een briljant wetenschapper en ik ben er van overtuigd dat je op de juiste plek bent om je waardevolle ideeën uit te werken.

Dr. A. Dehghan, dear Abbas, first I would like to thank you for providing the means to start a PhD trajectory in the field of molecular epidemiology. You are a wonderful person with an admirable patience. Thank you for your perseverance when I didn´t see the light at the end of the tunnel. I am convinced you will have a great future in London together with Raha and Nick. You have been a great co-promotor and mentor.

Prof. dr. O.H. Franco, dear Oscar, thank you for all your support during my PhD trajectory. I highly appreciate your positive attitude and thank you for all the wise words such as “always say yes” and “there is nothing that can´t be done”.

Verder wil ik mijn kamergenoten bedanken met wie ik vele uren heb doorgebracht in Na-2901 van het Erasmus MC. Paul de Vries, je bent een geweldige collega geweest en uitstekend wetenschapper. Bedankt voor al je hulp met data analyses met het programma R. Ik heb altijd gezegd dat je nooit meer terugkomt uit de VS: mind my words! Sanaz Sedaghat, I am grateful for your help and assistance especially in the beginning of my thesis. You have dramatically changed the way I look at Iran, and I hope to visit Iran in the near future. Layal Chaker, ik heb genoten van de eindeloze gesprekken en discussies over het Midden-Oosten en meer specifiek Syrië. Ik heb van niemand zoveel geleerd over dit thema als van jou. Veel succes met je verdere carrière zowel in de kliniek als het onderzoek! Mohsen Ghanbari, your passion for micro-RNA is fantastic! I always enjoyed our conversations and I am sure you will have a fruitful career in science. Dear Jana Nano, you are a wonderful person, I wish you all the best with your further career fighting the diabetes epidemic.

Maarten Leening, bedankt voor alle wijze gesprekken bij de koffieautomaat. Als wij het voor het zeggen hadden in de wereld, zou de wereld er een stuk beter uitzien! Thijs van Herpt, ik heb genoten van de avondjes diabetes data coderen en de supervisie momenten in de kliniek. Je bent een goede clinicus en je enthousiasme en humor zijn goud waard! Ik hoop dat we in de toekomst nog veel mogen samenwerken. Janine Felix, bedankt voor het feit dat ik altijd even kon aankloppen voor een kopje koffie of een inhoudelijke vraag over het onderzoek. Ik bewonder je empathie en enthousiasme. Tim Korevaar, de vele kopjes koffie waren elke keer weer verhelderende momenten. Je bent een uitstekend wetenschapper, een voorbeeld voor velen! Verder wil ik alle collega´s van de cardiovasculaire epidemiologie groep en de Erasmus Age groep bedanken voor hun kritische mening alle jaren.

Verder wil ik een aantal mensen bedanken die achter de schermen de Rotterdam studie draaiende houden. Beste Jolande Verkroost, bedankt voor het beantwoorden van vele vragen over het verzamelen van de diabetes data. Mede dankzij jouw hulp zijn hoofdstuk 2 en 3 van dit proefschrift tot stand gekomen. Frank van Rooij, bedankt dat je altijd weer met veel geduld mijn vragen wilde beantwoorden met betrekking tot de Rotterdam studie data. Nana Suwarno, ik heb genoten van je lach, en bedankt voor alle computer-gerelateerde oplossingen. Jan Heeringa, jouw nuchtere visie is leerzaam geweest en ook jou wil ik bedanken voor de gesprekken bij de kopjes koffie. Verder wil ik Prof. dr. A. Hofman bedanken, uw jarenlange inzet heeft van de Rotterdam studie een toonaangevend populatie cohort gemaakt.

De genetische en epigenetische data in de Rotterdam studie zijn vooral tot stand gekomen dankzij de afdeling Interne Geneeskunde van prof. dr. André Uitterlinden. Beste André, bedankt voor de mogelijkheid om de waardevolle genetische, epigenetische en genexpressie data te gebruiken. Verder wil ik Joyce van Meurs, Marjolein Peters, en Lisette Stolk bedanken voor het genereren van de DNA methylatie en gen expressie data, en Carolina Medina-Gomez, Fernando Rivadineira en Jeroen van Rooij voor hun adviezen bij het gebruiken en toepassen van de GWAS data.

Veel artikelen in dit proefschrift zijn tot stand gekomen uit samenwerkingen met andere onderzoekers en universiteiten, in het bijzonder het CHARGE consortium. Ik ben van mening dat samenwerkingen in het wetenschappelijk onderzoek van onschatbare waarde zijn: het geheel is meer dan de som der delen. Het CHARGE consortium is een fantastisch platform voor genetisch onderzoek en een schoolvoorbeeld voor andere onderzoeksgebieden. Ik wil alle deelnemers van de CHARGE inflammatie werkgroep en de epigenetica werkgroep bedanken voor hun inzet en ideeën.

Sjoerd Duim, bedankt dat je mijn paranimf wil zijn en bedankt voor alle geweldige momenten die wij samen beleefd hebben van huttentochten in de Alpen tot late uren in de kroeg in Leuven. Ik kijk uit naar het volgende avontuur samen, en ik hoop dat dat niet lang op zich laat wachten.

Na mijn tijd op de epidemiologie afdeling heb ik met veel plezier gewerkt als arts-assistent in het Amphia ziekenhuis in Breda. Daarom wil ik mijn opleider Joost van Esser bedanken voor de mogelijkheid om mijn thesis tijdens mijn opleiding interne geneeskunde te finaliseren. Beste Joost, je bent een inspirerende opleider en ik ben zeer verheugd met het voorrecht dat ik onder jouw hoede de eerste jaren van mijn opleiding interne geneeskunde heb genoten.

Verder wil ik mijn ouders bedanken voor hun onvoorwaardelijke steun en wijze raad. Jullie nemen een bijzondere plek in in mijn leven en zonder jullie zou ik niet staan waar ik nu sta. Pap, ik kan genieten van de lange gesprekken aan de keukentafel over belangrijke levenskwesties en ziekenhuis gerelateerde onderwerpen. Mam, je bent er altijd voor me. Je liefde voor ons en Amélie is groot, ik geniet er elke dag weer van.

Lieve Leen, ik houd zielsveel van je. Ik bewonder je onvoorwaardelijke steun en de manier waarop je altijd de positieve kant van alles belicht. We hebben al veel mooie momenten samen en met Amélie beleefd en ik hoop dat er nog veel momenten zullen volgen.

PhD portfolio

PhD Portfolio

Name	Symen Ligthart
Project title	Molecular Epidemiology of Inflammation – Link with Type 2 Diabetes and Coronary Heart Disease
Department	Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands
Research School	Netherlands Institute for Health Sciences (NIHES)
PhD period	2012-2020
Supervisors	Prof. dr. M.A. Ikram Dr. Abbas Dehghan

Courses

2012-2014	MSc in Clinical Epidemiology, NIHES, Erasmus MC, Rotterdam, the Netherlands
2012	SNPs and Human Diseases, MOLMED, Erasmus MC, Rotterdam, the Netherlands

Scientific meetings

2016	American Diabetes Association meeting, New Orleans, LA, USA (poster presentation)
2015	CHARGE investigator meeting, Jackson, MS, USA (oral presentation)
2015	American Diabetes Association meeting, Boston, MA, USA (poster presentation)
2015	Netherlands Epidemiology Society Congress (WEON), Leiden, the Netherlands (oral presentation)
2014	CHARGE investigator meeting, Washington, DC, USA
2014	Netherlands Epidemiology Society Congress (WEON), Leiden, the Netherlands (oral presentation)
2014	European Society of Cardiology Congress, Barcelona, Spain (oral presentation)

2014	CHARGE investigator meeting, Los Angeles, CA, USA (poster presentation)
2013	European Society of Cardiology Congress, Amsterdam, the Netherlands
2013	CHARGE investigator meeting, Rotterdam, the Netherlands
2012	CHARGE investigator meeting, Houston, TX, USA

Teaching

Assistance

2013	Principals of Research in Medicine and Epidemiology (ESP01), Erasmus Summer Program. NIHES, Erasmus MC, Rotterdam, the Netherlands
2013-2014	Methodological Topics of Study Design, NIHES, Erasmus MC, Rotterdam, the Netherlands

Supervising MSc students

2013-2014	Marte C. Liefaard
2014-2015	Rebecca V. Steenaard

Miscellaneous

2014-2015	PhD representative for NIHES in Erasmus MC PhD Committee
2015	PhD representative Rotterdam Study Management Team
2015-2016	Organizer 2020 meeting, Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands
2014-2017	Peer review of articles in scientific journals (*JAMA, Lancet Diabetes and Endocrinology, Epigenetics, International Journal of Epidemiology, European Journal of Epidemiology, Epidemiology, Cardiovascular Diabetology*)

List of publications

List of publications in chronological order

1. Ligthart S. Commentary: CRP and schizophrenia: cause, consequence or confounding? *International journal of epidemiology* 2019; 48(5): 1514-5.

2. Agha G, Mendelson MM, Ward-Caviness CK, et al. Blood Leukocyte DNA Methylation Predicts Risk of Future Myocardial Infarction and Coronary Heart Disease. *Circulation* 2019; 140(8): 645-57.

3. Liu J, Carnero-Montoro E, van Dongen J, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. 2019; 10(1): 2581.

4. Kraja AT, Liu C, Fetterman JL, et al. Associations of Mitochondrial and Nuclear Mitochondrial Variants and Genes with Seven Metabolic Traits. *American journal of human genetics* 2019; 104(1): 112-38.

5. Ligthart S, Vaez A, Vosa U, et al. Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. *American journal of human genetics* 2018; 103(5): 691-706.

6. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature genetics* 2018; 50(11): 1505-13.

7. Mahajan A, Wessel J, Willems SM, et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature genetics* 2018; 50(4): 559-71.

8. Aslibekyan S, Agha G, Colicino E, et al. Association of Methylation Signals With Incident Coronary Heart Disease in an Epigenome-Wide Assessment of Circulating Tumor Necrosis Factor alpha. *JAMA cardiology* 2018.

9. Richard MA, Huan T, Ligthart S, et al. DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. *American journal of human genetics* 2017; 101(6): 888-902.

10. Muka T, Asllanaj E, Avazverdi N, et al. Age at natural menopause and risk of type 2 diabetes: a prospective cohort study. *Diabetologia* 2017.

11. Herder C, de Las Heras Gala T, Carstensen-Kirberg M, et al. Circulating Levels of Interleukin 1-Receptor Antagonist and Risk of Cardiovascular Disease: Meta-Analysis of Six Population-Based Cohorts. *Arteriosclerosis, thrombosis, and vascular biology* 2017; 37(6): 1222-7.

12. Nano J, Muka T, Ligthart S, et al. Gamma-glutamyltransferase levels, prediabetes and type 2 diabetes: a Mendelian Randomization study. *International journal of epidemiology* 2017.

13. Brahimaj A, Ligthart S, Ghanbari M, et al. Novel inflammatory markers for incident pre-diabetes and type 2 diabetes: the Rotterdam Study. *European journal of epidemiology* 2017; 32(3): 217-26.

14. Kieboom BC, Ligthart S, Dehghan A, et al. Serum magnesium and the risk of prediabetes: a population-based cohort study. *European journal of epidemiology* 2017; 60(5): 843-53.

15. de Vries PS, Sabater-Lleal M, Chasman DI, et al. Comparison of HapMap and 1000 Genomes Reference Panels in a Large-Scale Genome-Wide Association Study. *PloS one* 2017; 12(1): e0167742.

16. Brahimaj A, Ligthart S, Ikram MA, et al. Serum Levels of Apolipoproteins and Incident Type 2 Diabetes: A Prospective Cohort Study. *PloS one* 2017; 40(3): 346-51.

17. Ligthart S, Marzi C, Aslibekyan S, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome biology* 2016; 17(1): 255.

18. de Vries PS, van Herpt TT, Ligthart S, et al. ADAMTS13 activity as a novel risk factor for incident type 2 diabetes mellitus: a population-based cohort study. *Diabetologia* 2017; 60(2): 280-6.

19. Chaker L, Ligthart S, Korevaar TI, et al. Thyroid function and risk of type 2 diabetes: a population-based prospective cohort study. *BMC medicine* 2016; 14(1): 150.

20. Hanewinckel R, Drenthen J, Ligthart S, et al. Metabolic syndrome is related to polyneuropathy and impaired peripheral nerve function: a prospective population-based cohort study. *Journal of neurology, neurosurgery, and psychiatry* 2016; 87(12): 1336-42.

21. Dhana K, Nano J, Ligthart S, et al. Obesity and Life Expectancy with and without Diabetes in Adults Aged 55 Years and Older in the Netherlands: A Prospective Cohort Study. *PLoS medicine* 2016; 13(7): e1002086.

22. Ligthart S, Vaez A, Hsu YH, et al. Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. *BMC genomics* 2016; 17: 443.

23. Smith JG, Felix JF, Morrison AC, et al. Discovery of Genetic Variation on Chromosome 5q22 Associated with Mortality in Heart Failure. *PLoS genetics* 2016; 12(5): e1006034.

24. Muka T, Nano J, Voortman T, et al. The role of global and regional DNA methylation and histone modifications in glycemic traits and type 2 diabetes: A systematic review. *Nutrition, metabolism, and cardiovascular diseases : NMCD* 2016; 26(7): 553-66.

25. Joubert BR, Felix JF, Yousefi P, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *American journal of human genetics* 2016; 98(4): 680-96.

26. Joubert BR, den Dekker HT, Felix JF, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nature communications* 2016; 7: 10577.

27. Ligthart S, Steenaard RV, Peters MJ, et al. Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. *Diabetologia* 2016; 59(5): 998-1006.

28. Maksimovic A, Hanewinckel R, Verlinden VJ, et al. Gait characteristics in older adults with diabetes and impaired fasting glucose: The Rotterdam Study. *Journal of diabetes and its complications* 2016; 30(1): 61-6.

29. Ligthart S, van Herpt TT, Leening MJ, et al. Lifetime risk of developing impaired glucose metabolism and eventual progression from prediabetes to type 2 diabetes: a prospective cohort study. *The lancet Diabetes & endocrinology* 2016; 4(1): 44-51.

30. LeBlanc M, Zuber V, Andreassen BK, et al. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes With Several Cardiovascular Risk Factors. *Circulation research* 2016; 118(1): 83-94.

31. Liefaard MC, Ligthart S, Vitezova A, et al. Vitamin D and C-Reactive Protein: A Mendelian Randomization Study. *PloS one* 2015; 10(7): e0131740.

32. Freitag DF, Butterworth AS, Willeit P, et al. Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: a Mendelian randomisation analysis. *The lancet Diabetes & endocrinology* 2015; 3(4): 243-53.

33. Steenaard RV, Ligthart S, Stolk L, et al. Tobacco smoking is associated with methylation of genes related to coronary artery disease. *Clinical epigenetics* 2015; 7: 54.

34. de Vries PS, Kavousi M, Ligthart S, et al. Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: the Rotterdam Study. *International journal of epidemiology* 2015; 44(2): 682-8.

35. Ligthart S, de Vries PS, Uitterlinden AG, et al. Pleiotropy among common genetic loci identified for cardiometabolic disorders and C-reactive protein. *PloS one* 2015; 10(3): e0118859.

36. Ligthart S, Sedaghat S, Ikram MA, Hofman A, Franco OH, Dehghan A. EN-RAGE: a novel inflammatory marker for incident coronary heart disease. *Arteriosclerosis, thrombosis, and vascular biology* 2014; 34(12): 2695-9.

37. Kraja AT, Chasman DI, North KE, et al. Pleiotropic genes for metabolic syndrome and inflammation. *Molecular genetics and metabolism* 2014; 112(4): 317-38.

38. Korevaar TI, Steegers EA, Schalekamp-Timmermans S, et al. Soluble Flt1 and placental growth factor are novel determinants of newborn thyroid (dys)function: the generation R study. *The Journal of clinical endocrinology and metabolism* 2014; 99(9): E1627-34.

About the Author

Symen Ligthart was born on September 30th 1987 in the town of Dordrecht in the Netherlands. Most part of his youth he grew up in the small village Numansdorp which is located south of Rotterdam. He finished his Gymnasium at the Erasmiaans Gymnasium in Rotterdam in 2005. After graduating from the Erasmiaans Gymnasium, he moved south to study medicine at the Catholic University of Leuven in Belgium. He successfully finished his medical school magna cum laude in 2012, after which he commenced his PhD in epidemiology at the department of Epidemiology, Erasmus MC, Rotterdam under the supervision of Abbas Deghan and prof. dr. M. Arfan Ikram. In January 2016 he finished his work at the Epidemiology department of the Erasmus MC and pursued his medical studies as a resident in internal medicine at the Amphia Hospital in Breda and the Erasmus MC in Rotterdam. After four years of medical residency in Breda, Symen currently works as a resident at the intensive care department at the Erasmus MC to complete his training in intensive care medicine. Symen lives in Antwerp in Belgium with his beloved wife Leen and beautiful daughter Amélie.