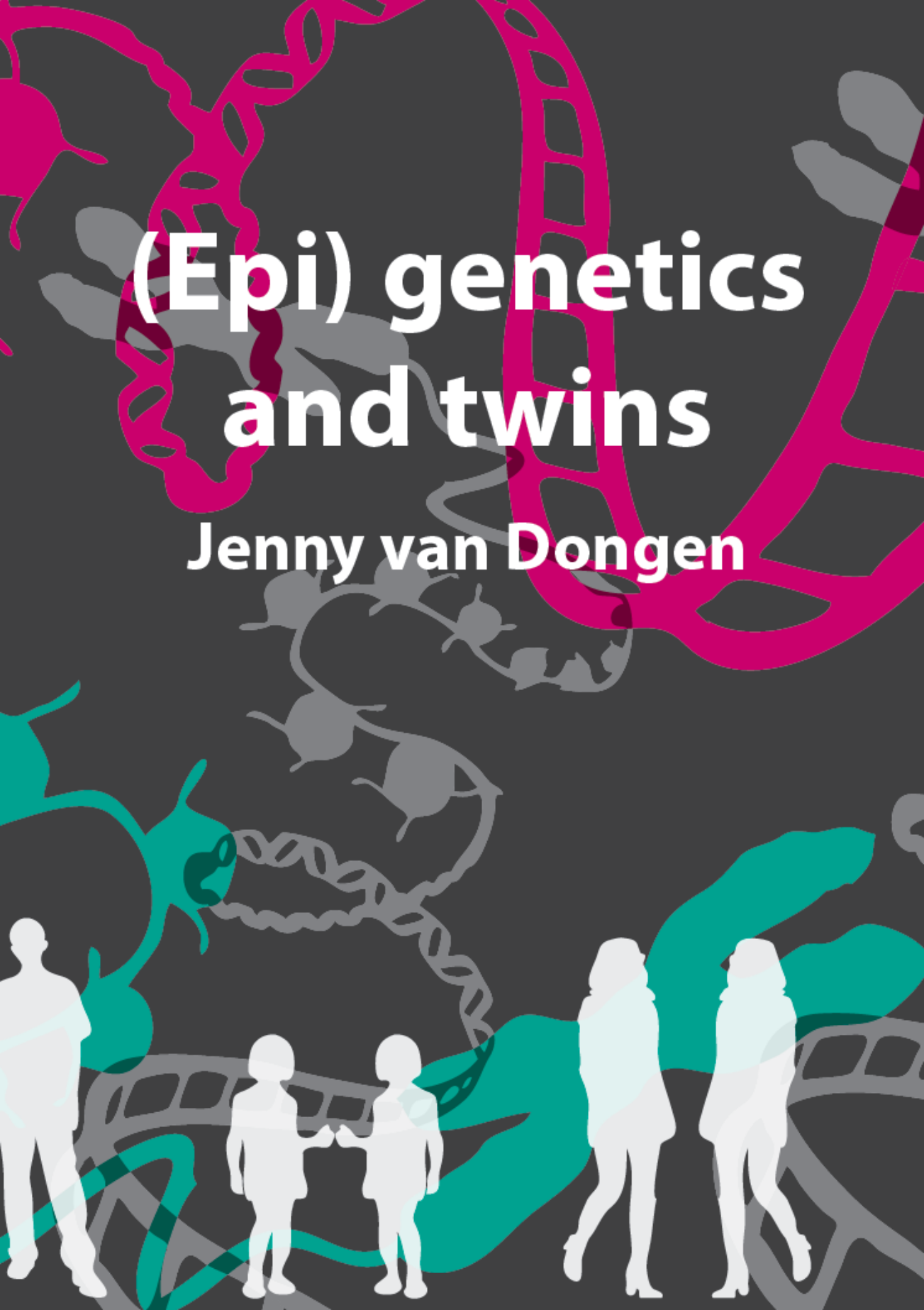
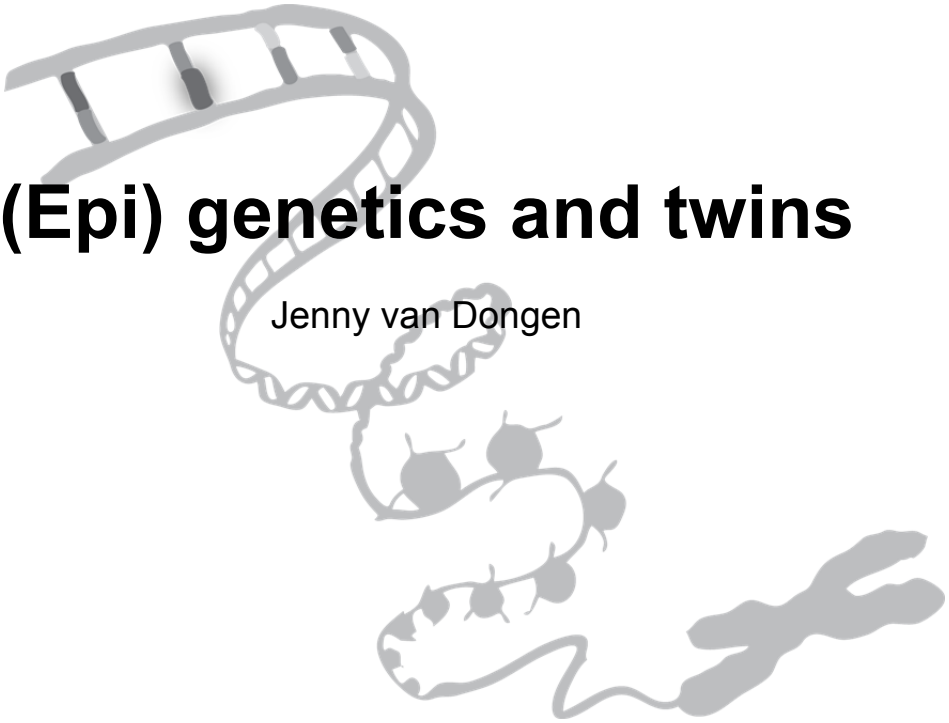
(Epi) genetics and twins

Jenny van Dongen





(Epi) genetics and twins

Jenny van Dongen

Reading comittee:

Dr. M. Beekman
Dr. E.A. Ehli
Dr. L. Franke
Prof.dr. J.C.N. de Geus
Prof.dr. J. Kaprio
Dr. J.B.J. van Meurs
Prof.dr C. Schuengel

Paranimfen:

Jorien Treur
Sue Mudde

Acknowledgements
This work was supported by the European Research Council [ERC-230374: Genetics of Mental Illness: A lifespan approach to the genetics of childhood and adult neuropsychiatric disorders and comorbid conditions. Data collection in the Netherlands Twin Register was supported by multiple grants from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development Grant (ZonMW). We acknowledge also support from the Biobanking and Biomolecular Resources Research Infrastructure (BBRMI-NL), the National Institute for Mental Health (NIMH), EMGO+ Institute for Health and Care Research and the Neuroscience Campus Amsterdam.

ISBN: 978-94-6259-608-5
Printed by: Ipskamp Drukkers, Amsterdam
Cover design: Rosie Paulissen

This thesis can be accessed online with all figures in full color at dare.ubvu.vu.nl

VRIJE UNIVERSITEIT

(Epi) genetics and twins

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan

de Vrije Universiteit Amsterdam,

op gezag van de rector magnificus

prof.dr. F.A. van der Duyn Schouten,

in het openbaar te verdedigen

ten overstaan van de promotiecommissie

van de Faculteit der Psychologie en Pedagogiek

op donderdag 16 april 2015 om 9.45 uur

in de aula van de universiteit,

De Boelelaan 1105

door

Jenny van Dongen

geboren te Haarlem

promotoren:	prof.dr. D.I. Boomsma
	prof.dr. P.E. Slagboom
copromotoren:	dr. B.T. Heijmans
	dr. A.H.M. Willemsen

Table of Contents

Introduction

To date, the leading causes of disease burden in moderate and high income countries (based on the disability-adjusted-life-year metric [1]) are complex diseases, with cardiovascular disease being the number one contributor followed by mental disorders, in particular depression. Complex diseases often cluster in families yet they show no clear pattern of inheritance and are thought to result from the interplay of environmental influences and genetic susceptibility acting on multiple biological pathways. The underlying disease mechanisms of such traits may be best understood if they are approached from a variety of different angles. In this thesis, multiple approaches are applied to study the etiology of human complex traits, with an emphasis on metabolic syndrome traits and inflammation.

The metabolic syndrome is a condition that involves various metabolic disturbances that are risk factors for cardiovascular disease (CVD) and type 2 diabetes (T2D), including central obesity, insulin resistance, dyslipidemia, and hypertension [2-4]. It is well established that the metabolic syndrome is associated with a state of chronic, low-grade inflammation that is thought to contribute to the pathogenesis of the disease and to the risk for CVD and T2D, although inflammation itself is generally not included in diagnostic guidelines [5, 6]. The most prevalent common mental disorder is major depressive disorder. Cardiovascular disease and depressive disorders at least to some extent touch upon common biological pathways as evidenced by the fact that biological risk factors for cardiovascular disease are also connected to depression, including BMI (either high or low), abdominal obesity, and dysregulated levels of lipids and inflammation biomarkers in blood (including pro-inflammatory cytokines, cytokine receptors and acute phase reactive proteins) [7-9]. Therefore, a better understanding of the biological mechanisms that contribute to individual differences in the level of metabolic and inflammation biomarkers in blood[i], and in particular the role of genomic variation, could ultimately be of benefit to both physical and mental health.

Known risk factors for the metabolic syndrome (e.g. physical inactivity, smoking and obesity) clearly emphasize the role of lifestyle factors in the etiology [10-13]. In fact, it is thought that interventions targeted at these risk factors are largely responsible for a decline in the prevalence of cardiovascular events that has been observed in several countries in the past decades, although

[i]In this thesis, I will use the term metabolic biomarkers to refer to the concentration of lipids, glucose and insulin in blood. The term inflammation biomarkers is used to refer to inflammation components measured in blood (e.g. blood level of cytokines and cytokine receptors).

improved treatment may also have contributed [14-16]. The high prevalence of major depressive disorder in wealthy societies is less well understood, and depression has been estimated to become the world leading cause of disease burden in 2030[17].

The research described in this thesis is characterized by two major themes: the first relates to the question to which extent differences between individuals in inflammation biomarkers and metabolic syndrome traits are caused by the impacts of genetic and environmental differences between people. The second important focus of this thesis is on examining the genetic and non-genetic sources of variation underlying individual differences in DNA methylation; an epigenetic mechanism that receives increasing attention as it may provide novel insights into human disease. These research themes are the subject of a series of chapters in which I analyze phenotypic, genetic and epigenetic information that was collected in twin families who participate in longitudinal research of the Netherlands Twin Register. The etiology of individual differences is examined with approaches from genetic epidemiology, including the classical twin design and the extended twin-family design, and by approaches that make use of measured genetic variants. Chapter 3 of this thesis illustrates how these different approaches can be combined to unravel the contribution of known genetic (or candidate) variants versus unidentified genetic variants to a complex trait (soluble interleukin-6 receptor levels in blood). In this introduction, I present an outline of the topics covered in this thesis. I shortly introduce the current knowledge regarding metabolic syndrome traits and inflammation biomarkers and especially focus on epigenetics. It is relatively recent that large scale epigenetic studies (genome-wide) have become feasible and have been applied to twins. In this chapter I therefore pay detailed attention to epigenetic mechanisms and their potential involvement in human complex disease.

The metabolic syndrome

A diagnosis of the metabolic syndrome is often given to a person if at least three of the following traits exceed a certain clinical threshold: waist circumference, body mass index, waist-to-hip-ratio, fasting plasma glucose, fasting insulin level, systolic blood pressure (SBP), diastolic blood pressure (DBP), HDL cholesterol, and/or triglycerides. There are multiple diagnostic guidelines for the metabolic syndrome that include slightly different combinations of traits. These guidelines were established with the primary aim to identify people at high risk of developing CVD and T2D. The metabolic syndrome is associated with a doubled risk of developing CVD and a more than 5-fold increased risk of T2D, but also with increased risk for many other diseases, including nonalcoholic fatty liver disease, reproductive disorders, depression and sleeping disorders[18, 19]. The pathophysiological mechanisms that characterize the metabolic syndrome and are thought to contribute to the comorbid conditions include excess adipose tissue mass, ectopic fat

deposition, excessive flux of fatty acids, and inflammation [3, 18, 20]. Although there is ongoing debate about the constituents and primary causal mechanism underlying the metabolic syndrome, it is generally acknowledged that obesity (particularly in the abdominal area) and insulin resistance are the major underlying risk factors for the metabolic syndrome. Insulin resistance is often secondary to obesity but may also present in individuals with a normal weight [21].

Overall, as much as 14% of individuals who live in the Netherlands [22] and 37% of individuals in the United States [23] are affected by the metabolic syndrome, but the prevalence varies greatly with sex and age and depends on the guideline used to characterize the metabolic syndrome [3]. In the Netherlands, the metabolic syndrome was reported to be more prevalent in males according to most but not all guidelines [24], and to be more prevalent in older people [22, 25]. Of Dutch individuals of 65 years or older 37% of individuals meet the metabolic syndrome criteria [25]. Although traditionally viewed as an age-related disease, the metabolic syndrome is becoming increasingly prevalent in all age groups and can even be present in childhood [26-28]. The rising prevalence of the metabolic syndrome is thought to be the consequence of the rising prevalence of obesity[29].

Global obesity trends and changes in lifestyle

In the period between 1980 and 2008, the worldwide prevalence of obesity almost doubled; from 6.4% to 12% [30], although there is considerable variation between countries[31]. In the Netherlands, the prevalence of overweight and obesity was 47.8% and 16.2% in adults aged 20+, respectively, in 2008[31]. There is no doubt that the outbreak of overweight- and obesity-related disease is related to modern day lifestyle. Worldwide, the average caloric intake per person increased with 450 kcal per day between 1960 and the late 1990s [32], and 31.1% of adult individuals (age 15+) worldwide were “physically inactive” according to a publication from 2012 [33]. It has been suggested that obesity-related disease and a lot of other modern day complex diseases as well are relate to a mismatch between our current lifestyle in comparison with our evolutionary history, a topic I will we discuss in more detail in chapter 8. Yet, while we are all surrounded by weight-promoting influences in wealthy countries nowadays (e.g. environments and jobs that limit the need to be physically active and high calorie food available to everyone), not all individuals develop obesity and not every obese individual develops obesity-related diseases. It has been reported that 65% of obese males and 56% of obese females meet the criteria for the metabolic syndrome [34], while the rest has been characterized as displaying “metabolically healthy obesity” [35]. By contrast, “metabolically unhealthy non-obese” individuals also exist [21, 36]. In the United States 23.% of individuals (age 20+) with a normal BMI show two or more metabolic abnormalities (e.g. insulin resistance, dyslipidemia, hypertension) [37].

Thus, individuals differ in their vulnerability to develop overweight, obesity and metabolic disease.

Relevant metabolic and inflammation biomarkers in the population

Of note, the metabolic syndrome characterizes individuals at the extreme end of the distribution of variation in metabolic traits. In the following chapters of my thesis, I study variation in individual metabolic syndrome traits in a population-based sample of individuals, who were unselected with respect to disease status or health. The aim was to characterize the causes of individual differences underlying population variation in metabolic syndrome traits. Of note, several other tools exist to assess the risk of developing coronary heart disease or cardiovascular disease within the 'normal population'. The most widely applied Framingham score for coronary heart disease is based on traits that are commonly referred to as "traditional risk factors": sex, age, total cholesterol or LDL cholesterol, systolic blood pressure, smoking and self-reported diabetes [38, 39]. In addition to the 'classical lipids' (i.e. total cholesterol, HDL, LDL, triglycerides), many other types of circulating lipid particles as well as other molecules (e.g. amino acids) may be informative to metabolic health: such markers may be studied through metabolomics technology[40].

Although metabolic syndrome criteria and other risk scores do not commonly include inflammation biomarkers, growing evidence suggests that a key mechanism behind the pathogenesis of the consequences of obesity and associated conditions (e.g. type II diabetes and cardiovascular disease) involves chronic over-activation of cellular stress signalling and inflammatory pathways in metabolic cells in response to excessive energy intake[41, 42]. In obese individuals, adipose tissue secretes a range of a pro-inflammatory cytokines such as TNF-alpha and IL-6 [43, 44]. Pro-inflammatory cytokines stimulate the release of so-called acute phase reactants (for example fibrinogen and CRP) [45]. Other molecules in the inflammation cascade that are elevated in obese individuals include so-called sensors of the innate immune system including the inflammosome and Toll-like receptors [46, 47]. It is now well-known that in addition to adipose tissue, the liver, pancreas and brain also respond to obesity or excess energy intake by increased levels of inflammation, and this inflammation is highly important in the development of obesity-related disease[48-50]. Systemic low-grade inflammation is also known as the pro-inflammatory state.

In the Netherlands Twin Register, the levels of inflammation biomarkers in blood and traits included in metabolic syndrome criteria have been assessed in a representative population-based sample. This study sample allowed us to study the entire distribution of these traits to examine the causes of individual differences in these traits at the population level. In this thesis, I study the following inflammation biomarkers: tumor necrosis factor-alpha (TNF-α), C-reactive protein (CRP), fibrinogen, interleukin-6 (IL-6), and the soluble IL-6 receptor (sIL-6R). Metabolic syndrome traits included systolic and diastolic

blood pressure, blood levels of glucose, insulin and lipids, and the anthropometric traits weight, BMI, waist circumference and waist-to-hip ratio (WHR).

Genetics of BMI and biomarkers

Feeding experiments conducted on initially lean male prisoners in the 1960s illustrated the existence of individual differences in the effects of over-eating (~10,000 kcal per day)[51]. While most of the 20 men who managed to complete this feeding regime for a period of 200 days quickly lost the weight they gained as soon as the feeding regime turned back to normal, two men had great difficulties losing the weight [52]. These men were initially the fastest to gain weight and had a family history of obesity, although they had not previously been overweight themselves.

In societies in which the majority of the population has access to sufficient calories, individual differences in BMI are to a large extent explained by genetic factors. Adoption studies show that the BMIs of adopted individuals at middle age are more similar to their biological parents' BMI than to their adoptive parents' [53]. Monozygotic (MZ) twins usually have very similar BMIs, irrespective of whether the twins grow up together or are adopted by different families [54]. A number of twin and family studies have estimated the proportion of variation in BMI between individuals that can be attributed to genetic effects (the heritability); these estimates vary between approximately 24% and 90% [55]. Similar results hold for inflammation biomarkers and for metabolic biomarkers including blood lipid levels, fasting insulin and glucose and blood pressure. A large body of research shows that variation in these phenotypes is also largely influenced by genes [56]. In the past decade or so, a number of variants in the DNA sequence (risk alleles) have been identified that contribute to individual differences in these traits in humans [57], many of which are located in regulatory regions of the genome [58]. For example, the Genetic Investigation of ANthropometric Traits (GIANT) consortium has reported 32 independent genetic variants for BMI [59] and the Gobal Lipids Genetics Consortium has published 157 loci for blood lipid levels (including LDL cholesterol, HDL cholesterol, triglycerides and total cholesterol)[60]. Sixteen loci that are involved in fasting glucose homeostasis have been identified by the meta-analyses conducted by the meta-analyses of Glucose and Insulin-related traits Consortium (MAGIC)[61].

Beyond investigations into the DNA sequence itself, large cohort studies are now also starting to address at a genome-wide resolution other types of molecular variation by which variation in human complex traits is created. Thus, recent studies have reported associations of BMI [62], T2D [63, 64] and vascular disease [65] with epigenetic variation, that is, variation in molecular mechanisms that regulate to which extent genes need to be expressed depending on the internal (such as the degree of adiposity) and external environment of the person.

Epigenetic mechanisms

Nearly every human cell contains the same DNA sequence (genome) inherited from the parents, although there is increasing recognition that mosaicism may occur (meaning that an individual has cell populations with distinct genotypes, due to a postzygotic mutation that arose in one cell lineage)[66]. Yet many different cell types and organs are formed with the same sequence information, requiring different genes to be activated or inactivated (switched on and off) in each cell at the right time. This regulation of gene activity is coordinated in each cell by numerous molecular mechanisms that collectively control chromatin structure; including chemical tags attached to the DNA molecule itself and to the histone proteins it is wrapped around (i.e. DNA methylation and histone modifications), and molecules interacting with the DNA or RNA transcripts (e.g. non-coding RNAs, transcription factors, and methyl-CpG-binding proteins) [67]. While the field of genetics is traditionally concerned with the study of the DNA sequence, epigenetics refers to the study of those molecular mechanisms that influence gene expression without changing the DNA sequence and that are transmitted from one cell generation to the next through cell division (mitosis and possibly meiosis) [68]. In practice, the term epigenetic regulation is often applied more broadly, to refer to DNA methylation plus all histone modifications, although it is expected that histone modifications are not fully transmitted during cell division [69].

Whereas a person's DNA sequence remains the same during the entire lifetime (except for occasionally occurring *de novo* mutations), DNA methylation and other epigenetic marks are dynamic; they may change during the lifetime as part of developmentally regulated process (related to e.g. tissue differentiation) [70] and aging [71], and may change in response to specific environmental exposures [72]. For example, multiple studies have reported altered methylation levels of the *AHRR* gene and several other genes in blood cells of smokers (see for example [73]), and in babies of mothers who smoked during pregnancy [74]. A study of middle-aged individuals who were in the womb when their mother experienced severe famine during the Dutch Hunger Winter demonstrated that this exposure had left persistent changes in the methylation patterns at a diversity of genes, illustrating that environmental exposures can have long-term effects on DNA methylation [75]. Because DNA methylation may change over the life time and can respond to environmental exposures, studying this molecular layer of information may shed light on disease mechanisms that would remain hidden when focusing on the DNA sequence only. This mechanism also may shed light on the frequent discordance of monozygotic twin pairs for many complex traits and disorders.

Chromatin

If the DNA molecule would not be condensed it would not fit into the cell nucleus. Chromatin refers to the complex of the chromosomal DNA molecule and all attached histone proteins that facilitate to package the entire human

genome (approximately 2 meters of DNA) into the cell nucleus (which has an average diameter of 6 micrometers [76]). When viewed under an electron microscope, the 3D structure of chromatin in regions where DNA transcription occurs resembles "a string of beads": a DNA string with so-called nucleosomes ('beads'), which are units of about 146 base pairs of DNA that are wrapped 1.65 times around a complex of histone proteins [77-79]. Between each nucleosome are little pieces of DNA (stretching on average 20 base pairs) called linker DNA. Because the histone proteins are positively charged while the DNA molecule is negatively charged, histones help to fold the DNA into a smaller volume through electrostatic interactions. At genomic regions that are not transcribed, chromatin is further compacted into a highly condensed conformation through higher order structuring of the nucleosomes. The exact structure of chromatin, also called 'chromatin state' is fundamental to the expression potential of a region and is intimately linked to covalent post-translational modifications of the histone proteins.

Histone modifications

The histone protein complex within nucleosomes, also called "histone octamer", consists of eight histone proteins: two copies of each of the histone types H2A, H2B, H3 and H4. A fifth type of histone (H1) can bind to the linker DNA between nucleosomes. Each histone protein is subject to numerous modifications at specific amino acids, including methylation, acetylation, phosphorylation and ubiquitination and many more. Most of these histone modifications occur at the "tails" (amino termini) of histones that stick out from the nucleosome complex. In total, there are more than 100 different histone modifications [80], most not well understood, which together control chromatin structure. For example, acetylation of the amino acid lysine is universally seen in transcriptionally active, accessible chromatin. Other modifications are associated with specific DNA elements. For example, histone H3 lysine 4 trimethylation (H3K4me3) is associated with promoters of actively transcribed genes [81]. It has been hypothesized that the exact chromatin state of a region lies written in the specific combination of histone modifications within that region ("the histone code")[82] .

Histone modifications may exert their influence on chromatin structure by changing the electric charge of the histone (thereby modifying the strength of the interaction between the histone protein and the DNA molecule), and by recruiting other molecules to the DNA. Histone modifications and chromatin structure are mediated by a number of proteins that fall in one or more of the following categories: writers (responsible for depositing histone modifications, for example histone acetyl transferases; HATs [83]), erasers (which can remove modifications, for example histone deacetylases; HDACs), and readers (proteins that 'interpret' the histone code by recruiting additional molecules that can stabilize or remodel specific chromatin states, upon recognition of specific histone modifications).

Chromatin states: Insight into the genome-wide expression landscape
Transcription of DNA requires that transcription factors can bind to specific target sequences in the DNA (i.e. promoters). Whether a transcription factor can bind to its target depends on whether interaction partners such as co-activators and repressors are present, and on whether the local chromatin structure permits such interactions to take place, in other words, whether the DNA molecule is accessible [84]. In each cell, the structure of chromatin varies along the genome. In genomic regions with 'inactive chromatin' or 'heterochromatin', the chromatin is highly condensed (nucleosomes are tightly packaged) and transcription is silenced due to DNA and histone methylation, while other regions are characterized by an open chromatin state ('active chromatin' or 'euchromatin'), where the distance between nucleosomes is larger, thereby permitting transcription to occur [80]. It is thought that, in addition to active and inactive chromatin, chromatin comes in a number of intermediate states. For example, in permissive or repressed states, DNA is accessible to become de-repressed or activated through interaction with for example transcription factors. A recent study described fifteen distinct chromatin states that were observed in nine human cell types [85].

In addition to the study of histone modifications, important insight into gene regulation has come from mapping DNAse hypersensitive sites (DHSs) in the genome; DNA that is sensitive for being cut by DNAse enzymes. DNAse enzymes can only cut DNA in accessible chromatin, i.e. DNA that is situated between nucleosomes ("linker DNA"). DNAse experiments have revealed that this nucleosome-free DNA contains regulatory elements including promoters, enhancers, silencer, insulators and locus control regions. Importantly, only ~5% of accessible chromatin occurs within 2.5kb of transcriptional start sites, whereas ~95% of accessible chromatin is located in distal intronic and intergenic regions in human cells[86].

DNA methylation
DNA methylation, the covalent attachment of a methyl-group (CH_3) to the DNA molecule, is one of the best studied epigenetic mechanisms in humans, and is currently the only epigenetic mark that is suited for measuring in large-scale human epidemiological studies. In vertebrates, DNA methylation occurs mostly at cytosines located next to guanines (CpG sites). Non-CpG methylation is quite prevalent in human embryonic stem cells, but is very rare or absent in differentiated somatic cells [87-89]. DNA methylation is established and maintained by enzymes from the methyltransferase family, including DNMT1, DNMT3a and DNMT3a [90]. Some studies have suggested that *de novo* DNA methylation represents a passive process that targets regulatory sequences in the DNA that are not occupied by transcription factors [86, 91]. In total, an estimated 70-80% of CpG sites in the genome are methylated in mammalian cells [92]. Many CpGs occur in clusters called CpG islands (CGIs). The

promoters of approximately 70% of human genes overlap with a CpG island. The classic view is that methylation of CpGs in the promoter area of genes is associated with repression of gene expression. It has now become clear that the effect of methylation on expression may vary depending on the exact location that is methylated [93] and there are indications that methylation at enhancers is more strongly related to the expression level of genes than methylation at promoters [94].

It has been postulated that methylation may impact on regulation of transcription through two main mechanisms. Firstly, methylation of specific sequences may prevent the binding of regulator proteins to the DNA (e.g. transcription factors [95] and insulators [96]). Secondly, methylated CpGs may attract methyl-CpG-binding domain proteins, which are regulatory proteins that recruit chromatin regulators such as histone deacetylases and chromatin remodeling complexes to the site [97, 98]. Thus, rather than acting as independent mechanisms, it is thought that the different layers of epigenetic information (e.g. DNA methylation, histone modifications) generally work together to regulate transcription. Although the presence of DNA methylation correlates with the presence of repressive histone marks [99], there are exceptions. For example, most CpGs within promoter CGIs are unmethylated, but genes can be repressed through repressive histone marks even if their CGIs are unmethylated. Methylation of promoter CGIs is thought to contribute to long-term repression, for example at inactive X-chromosome genes in females [100] and imprinted genes [101]. In comparison with other epigenetic marks, it has been suggested that DNA methylation may be best described as a “memory signal for the long-term maintenance of gene silencing”[102]. Thus, it has been shown in colon cancer cells that while drugs that target histone modifications(histone deacetylase inhibitors (HDACi)) can lead to transient re-activation of loci silenced by DNA methylation, permanent re-activation can only be induced by DNA-demethylating drugs [102].

Although multiple techniques exist to measure DNA methylation, the Infinium humanMethylation 450 array (Illumina 450k) has become a popular platform in recent years for assessing DNA methylation at a genome-wide scale in human cohorts [103, 104]. This array assesses methylation level at ~485,000 CpG sites across a variety of regions in the human genome, including regulatory with genes and in intergenic regions.

Epigenetic gene regulation: link between environmental exposures and disease?

The question is to what extent epigenetic variation, by itself or in interaction with genetic variation, influences the metabolic and inflammatory biomarkers mentioned above and thereby the risk of metabolic diseases. Convincing evidence that epigenetic dysregulation can cause obesity in humans comes from rare neurodevelopmental disorders that are associated with obesity and result from imprinting defects, including Prader Willi syndrome [105] and

Beckwith-Wiedemann Syndrome [106]. Importantly, epigenetic changes that arise as a result of obesity may be involved in the development of obesity-related disease. A recent study examined the relationship between BMI and a group of CpGs where DNA methylation is strongly related to age [71] (also known as "the epigenetic clock") in multiple tissues, and found that people with a higher BMI show accelerated 'epigenetic aging' of liver tissue and 'epigenetically older livers' showed differential expression of a number of genes [107]. The findings suggest that epigenetic dysregulation in the liver may connect obesity to the development of age- and liver-related disease such as insulin resistance, although this hypothesis remains to be examined. Many studies have indicated the importance of epigenetic regulation within relevant disease tissues for metabolic syndrome traits and inflammation [63, 64, 108]. For example, cholesterol homeostasis was shown to be epigenetically regulated by a microRNA (miR-33) that controls the expression of genes involved in cholesterol transport including *ABCA1* [109]. The expression of the TNF-alpha locus is regulated through development and in response to e.g. lipopolysaccharide stimulation (LPS) by various epigenetic marks including DNA methylation and histone modifications [110]. To identify novel loci where epigenetic variation is related to disease risk, a lot of researchers at the moment are working on Epigenome-Wide Association Studies (EWAS) that test a large number of epigenetic marks along the genome for association with a (disease-) phenotype.

Epigenetic regulation of gene expression thus is as an important candidate mechanism that may mediate the effects of external influences (e.g. environmental exposures) on disease development. However, the other way other around, DNA methylation may also be influenced by the DNA sequence itself [111], and this heritable epigenetic variation between individuals may represent another pathway contributing to human disease susceptibility. In addition to EWAS, important understanding of how the epigenome may mediate variation in disease susceptibility may come from studies that are able to delineate the extent to which epigenetic variation between individuals in accessible tissues such as blood and buccal cells can be attributed to heritable mechanisms versus environmental exposures and stochastic effects.

Outline of this thesis

In the first part of this thesis, I study the genetics of inflammation biomarkers circulating in blood. Chapter 2 describes a study of the heritability of various components of the pro-inflammatory state based on an extended twin-family design. Chapter 3 focuses on variation in soluble IL-6 receptor levels. This chapter investigates the total heritability, the heritability explained by measured DNA sequence variants (SNPs), and characterizes the relationship between such SNPs and gene expression as a possible mode of action of the genetic variants.

In the second part of this thesis, the importance of genetic and environmental influences on BMI and other metabolic syndrome traits is

examined. In chapter 4, the extended twin-family design is applied to estimate the heritability of metabolic syndrome traits. In chapter 5, I examine the prevalence and development of BMI discordance over time in MZ twins, and examine whether discordance for body composition (BMI) in genetically identical subjects is associated with differences in metabolic and inflammation biomarkers and gene expression.

The third part of this thesis is devoted to studies that look beyond the information that lies within the primary DNA sequence, by exploring variation in DNA methylation. In chapter 6, I examine the etiology of individual differences in DNA methylation in peripheral blood, based on data from MZ and dizygotic (DZ) twins and their parents. In chapter 7, I examine variation in DNA methylation in buccal cells from MZ twins.

The fourth part of this thesis consists of two reviews covering theoretical background and future directions for studies on human complex traits. Chapter 8 introduces a topic that is crucial for understanding the origin of common human disease; the topic of evolution. In this chapter, I review evolutionary perspectives of schizophrenia, a common mental disorder. This chapter explains the links between the evolutionary history of traits and their genetic architecture, and the implications thereof for gene finding studies. Chapter 9 explains the value of twins for studying the sources of variation in human traits and discusses the role that twin studies may play in future . Two designs based on data from twins that are used in this thesis; the estimation of heritability and the discordant MZ twin design, and their application in the context of 'omic' data, including studies of biomarkers, gene expression and epigenetics are reviewed in chapter 9.

In chapter 10, I summarize the most important results from all chapters and discusses these findings in the broader context of the current state and future directions of research on complex trait genetics.

Reference List

1. Mathers,C., Fat,D.M., & Boerma,J.T. *The global burden of disease: 2004 update*(World Health Organization,2008).
2. Alberti,K.G., Zimmet,P., & Shaw,J. The metabolic syndrome--a new worldwide definition. *Lancet* **366**, 1059-1062 (2005).
3. Eckel,R.H., Grundy,S.M., & Zimmet,P.Z. The metabolic syndrome. *The Lancet* **365**, 1415-1428 (2005).
4. Eckel,R.H., Alberti,K.G., Grundy,S.M., & Zimmet,P.Z. The metabolic syndrome. *Lancet* **375**, 181-183 (2010).
5. Haffner,S.M. The metabolic syndrome: inflammation, diabetes mellitus, and cardiovascular disease. *Am. J. Cardiol.* **97**, 3A-11A (2006).
6. Kassi,E., Pervanidou,P., Kaltsas,G., & Chrousos,G. Metabolic syndrome: definitions and controversies. *BMC. Med.* **9**, 48 (2011).
7. Lasserre,A.M. *et al.* Depression with atypical features and increase in obesity, body mass index, waist circumference, and fat mass: a prospective, population-based study. *JAMA Psychiatry* **71**, 880-888 (2014).
8. Penninx,B.W., Milaneschi,Y., Lamers,F., & Vogelzangs,N. Understanding the somatic consequences of depression: biological mechanisms and the role of depression symptom profile. *BMC. Med.* **11**, 129 (2013).
9. Vogelzangs,N., Comijs,H.C., Oude Voshaar,R.C., Stek,M.L., & Penninx,B.W. Late-life depression symptom profiles are differentially associated with immunometabolic functioning. *Brain Behav. Immun.* **41**, 109-115 (2014).
10. Dhingra,R. *et al.* Soft drink consumption and risk of developing cardiometabolic risk factors and the metabolic syndrome in middle-aged adults in the community. *Circulation* **116**, 480-488 (2007).
11. Lakka,T.A. *et al.* Sedentary lifestyle, poor cardiorespiratory fitness, and the metabolic syndrome. *Med. Sci. Sports Exerc.* **35**, 1279-1286 (2003).
12. McKeown,N.M. *et al.* Carbohydrate nutrition, insulin resistance, and the prevalence of the metabolic syndrome in the Framingham Offspring Cohort. *Diabetes Care* **27**, 538-546 (2004).
13. Park,Y.W. *et al.* The metabolic syndrome: prevalence and associated risk factor findings in the US population from the Third National Health and Nutrition Examination Survey, 1988-1994. *Arch. Intern. Med.* **163**, 427-436 (2003).
14. Berry,J.D. *et al.* Lifetime risks of cardiovascular disease. *N. Engl. J. Med.* **366**, 321-329 (2012).
15. Ford,E.S. *et al.* Explaining the decrease in U.S. deaths from coronary disease, 1980-2000. *N. Engl. J. Med.* **356**, 2388-2398 (2007).
16. Wijeysundera,H.C. *et al.* Association of temporal trends in risk factors and treatment uptake with coronary heart disease mortality, 1994-2005. *JAMA* **303**, 1841-1847 (2010).
17. Mathers,C.D. & Loncar,D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS. Med.* **3**, e442 (2006).
18. Cornier,M.A. *et al.* The metabolic syndrome. *Endocr. Rev.* **29**, 777-822 (2008).
19. Vogelzangs,N. *et al.* Metabolic depression: a chronic depressive subtype? Findings from the InCHIANTI study of older persons. *J Clin. Psychiatry* **72**, 598-604 (2011).
20. Grundy,S.M. The Metabolic Syndrome in *Atlas of Atherosclerosis and Metabolic Syndrome* (ed. Grundy,S.M.) 1-26 (Springer, New York, 2011).

21. Ruderman,N., Chisholm,D., Pi-Sunyer,X., & Schneider,S. The metabolically obese, normal-weight individual revisited. *Diabetes* **47**, 699-713 (1998).
22. Bos,M.B. *et al.* The prevalence of the metabolic syndrome in the Netherlands: increased risk of cardiovascular diseases and diabetes mellitus type 2 in one quarter of persons under 60. *Nederlands tijdschrift voor geneeskunde* **151**, 2382-2388 (2007).
23. Cameron,A.J., Shaw,J.E., & Zimmet,P.Z. The metabolic syndrome: prevalence in worldwide populations. *Endocrinology and metabolism clinics of North America* **33**, 351-375 (2004).
24. Dekker,J.M. *et al.* Metabolic syndrome and 10-year cardiovascular disease risk in the Hoorn Study. *Circulation* **112**, 666-673 (2005).
25. Oosterwerff,M.M., van Schoor,N.M., Lips,P., & Eekhoff,E.M. Osteocalcin as a predictor of the metabolic syndrome in older persons: a population-based study. *Clin. Endocrinol. (Oxf)* **78**, 242-247 (2013).
26. Weiss,R. *et al.* Obesity and the metabolic syndrome in children and adolescents. *N. Engl. J. Med.* **350**, 2362-2374 (2004).
27. Friend,A., Craig,L., & Turner,S. The prevalence of metabolic syndrome in children: a systematic review of the literature. *Metab Syndr. Relat Disord.* **11**, 71-80 (2013).
28. Fazeli,F.S., van der Aa,M.P., van der Vorst,M.M., Knibbe,C.A., & de,B.A. Global trends in the incidence and prevalence of type 2 diabetes in children and adolescents: a systematic review and evaluation of methodological approaches. *Diabetologia* **56**, 1471-1488 (2013).
29. National Institutes of Health. http://www.nhlbi.nih.gov/health/health-topics/topics/ms/, Accessed on 03/11/. 2014.
30. Stevens,G.A. *et al.* National, regional, and global trends in adult overweight and obesity prevalences. *Popul. Health Metr.* **10**, 22 (2012).
31. World Health Organization. http://www.who.int/gho/ncd/risk_factors/overweight/en/, Accessed on 12/06/. 2014.
32. Food and Agriculture Organization of the United Nations. World agriculture: towards 2015/2030 (Summary report). 2012. Rome.
33. Hallal,P.C. *et al.* Global physical activity levels: surveillance progress, pitfalls, and prospects. *Lancet* **380**, 247-257 (2012).
34. Ervin,R.B. Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006. *Natl. Health Stat. Report.*1-7 (2009).
35. Hamer,M. & Stamatakis,E. Metabolically healthy obesity and risk of all-cause and cardiovascular disease mortality. *J. Clin. Endocrinol. Metab* **97**, 2482-2488 (2012).
36. Conus,F., Rabasa-Lhoret,R., & Peronnet,F. Characteristics of metabolically obese normal-weight (MONW) subjects. *Appl. Physiol Nutr. Metab* **32**, 4-12 (2007).
37. Wildman,R.P. *et al.* The obese without cardiometabolic risk factor clustering and the normal weight with cardiometabolic risk factor clustering: prevalence and correlates of 2 phenotypes among the US population (NHANES 1999-2004). *Arch. Intern. Med.* **168**, 1617-1624 (2008).
38. Anderson,K.M., Wilson,P.W., Odell,P.M., & Kannel,W.B. An updated coronary risk profile. A statement for health professionals. *Circulation* **83**, 356-362 (1991).
39. Wilson,P.W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-1847 (1998).

40. Vaarhorst,A.A. *et al.* A metabolomic profile is associated with the risk of incident coronary heart disease. *Am. Heart J.* **168**, 45-52 (2014).
41. Ozcan,U. *et al.* Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* **306**, 457-461 (2004).
42. Stienstra,R., Tack,C.J., Kanneganti,T.D., Joosten,L.A., & Netea,M.G. The inflammasome puts obesity in the danger zone. *Cell Metab* **15**, 10-18 (2012).
43. Berg,A.H. & Scherer,P.E. Adipose tissue, inflammation, and cardiovascular disease. *Circ. Res.* **96**, 939-949 (2005).
44. Hotamisligil,G.S., Arner,P., Caro,J.F., Atkinson,R.L., & Spiegelman,B.M. Increased adipose tissue expression of tumor necrosis factor-alpha in human obesity and insulin resistance. *J. Clin. Invest* **95**, 2409-2415 (1995).
45. Shoelson,S.E., Lee,J., & Goldfine,A.B. Inflammation and insulin resistance. *J. Clin. Invest* **116**, 1793-1801 (2006).
46. Schroder,K., Zhou,R., & Tschopp,J. The NLRP3 inflammasome: a sensor for metabolic danger? *Science* **327**, 296-300 (2010).
47. Shi,H. *et al.* TLR4 links innate immunity and fatty acid-induced insulin resistance. *J. Clin. Invest* **116**, 3015-3025 (2006).
48. Cai,D. *et al.* Local and systemic insulin resistance resulting from hepatic activation of IKK-beta and NF-kappaB. *Nat. Med.* **11**, 183-190 (2005).
49. De Souza,C.T. *et al.* Consumption of a fat-rich diet activates a proinflammatory response and induces insulin resistance in the hypothalamus. *Endocrinology* **146**, 4192-4199 (2005).
50. Ehses,J.A. *et al.* Increased number of islet-associated macrophages in type 2 diabetes. *Diabetes* **56**, 2356-2370 (2007).
51. Sims,E.A. *et al.* Experimental obesity in man. *Trans. Assoc. Am. Physicians* **81**, 153-170 (1968).
52. Shell E.R. Hunger in *The hungry gene: the inside story of the obesity industry.* (Grove Press, New York, 2003).
53. Stunkard,A.J. *et al.* An adoption study of human obesity. *N. Engl. J. Med.* **314**, 193-198 (1986).
54. Stunkard,A.J., Harris,J.R., Pedersen,N.L., & McClearn,G.E. The body-mass index of twins who have been reared apart. *N. Engl. J. Med.* **322**, 1483-1487 (1990).
55. Elks,C.E. *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol. (Lausanne)* **3**, (2012).
56. Pilia,G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
57. Visscher,P.M., Brown,M.A., McCarthy,M.I., & Yang,J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7-24 (2012).
58. Hindorff,L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A* **106**, 9362-9367 (2009).
59. Speliotes,E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937-948 (2010).
60. Global Lipids Genetics Consortium Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274-1283 (2013).
61. Dupuis,J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105-116 (2010).

62. Dick,K.J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990-1998 (2014).
63. Dayeh,T. *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. Genet.* **10**, e1004160 (2014).
64. Nilsson,E. *et al.* Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes* **63**, 2962-2976 (2014).
65. Connelly,J.J. *et al.* Epigenetic regulation of COL15A1 in smooth muscle cell replicative aging and atherosclerosis. *Hum. Mol. Genet.* **22**, 5107-5120 (2013).
66. Biesecker,L.G. & Spinner,N.B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307-320 (2013).
67. Goldberg,A.D., Allis,C.D., & Bernstein,E. Epigenetics: a landscape takes shape. *Cell* **128**, 635-638 (2007).
68. Bird,A. Perceptions of epigenetics. *Nature* **447**, 396-398 (2007).
69. Bernstein,B.E., Meissner,A., & Lander,E.S. The mammalian epigenome. *Cell* **128**, 669-681 (2007).
70. Reik,W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425-432 (2007).
71. Horvath,S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
72. Feil,R. & Fraga,M.F. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* **13**, 97-109 (2011).
73. Philibert,R.A., Beach,S.R., Lei,M.K., & Brody,G.H. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clin. Epigenetics.* **5**, 19 (2013).
74. Joubert,B.R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* **120**, 1425-1431 (2012).
75. Heijmans,B.T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. U. S. A* **105**, 17046-17049 (2008).
76. Alberts,B. *et al. Molecular Biology of the Cell 4th Edition: International Student Edition*(Routledge,2002).
77. Luger,K., Mader,A.W., Richmond,R.K., Sargent,D.F., & Richmond,T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260 (1997).
78. Olins,A.L. & Olins,D.E. Spheroid chromatin units (v bodies). *Science* **183**, 330-332 (1974).
79. Olins,D.E. & Olins,A.L. Chromatin history: our view from the bridge. *Nat. Rev. Mol. Cell Biol.* **4**, 809-814 (2003).
80. Kouzarides,T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).
81. Santos-Rosa,H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407-411 (2002).
82. Jenuwein,T. & Allis,C.D. Translating the histone code. *Science* **293**, 1074-1080 (2001).
83. Brown,C.E., Lechner,T., Howe,L., & Workman,J.L. The many HATs of transcription coactivators. *Trends Biochem. Sci.* **25**, 15-19 (2000).

84. Li,B., Carey,M., & Workman,J.L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).
85. Ernst,J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49 (2011).
86. Thurman,R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
87. Laurent,L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res.* **20**, 320-331 (2010).
88. Lister,R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).
89. Ramsahoye,B.H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A* **97**, 5237-5242 (2000).
90. Goll,M.G. & Bestor,T.H. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* **74**, 481-514 (2005).
91. Stadler,M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490-495 (2011).
92. Bird,A., Taggart,M., Frommer,M., Miller,O.J., & Macleod,D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91-99 (1985).
93. Jones,P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484-492 (2012).
94. Aran,D., Sabato,S., & Hellman,A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14**, R21 (2013).
95. Watt,F. & Molloy,P.L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev.* **2**, 1136-1143 (1988).
96. Hark,A.T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486-489 (2000).
97. Nan,X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386-389 (1998).
98. Wade,P.A. *et al.* Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat. Genet.* **23**, 62-66 (1999).
99. Fuks,F. DNA methylation and histone modifications: teaming up to silence genes. *Curr. Opin. Genet. Dev.* **15**, 490-495 (2005).
100. Yasukochi,Y. *et al.* X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils. *Proc. Natl. Acad. Sci. U. S. A* **107**, 3704-3709 (2010).
101. Choufani,S. *et al.* A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. *Genome Res.* **21**, 465-476 (2011).
102. Raynal,N.J. *et al.* DNA methylation does not stably lock gene expression but instead serves as a molecular mark for gene silencing memory. *Cancer Res.* **72**, 1170-1181 (2012).
103. Bibikova,M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295 (2011).
104. Sandoval,J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics.* **6**, 692-702 (2011).

105. Ohta,T. *et al.* Imprinting-mutation mechanisms in Prader-Willi syndrome. *Am. J. Hum. Genet.* **64**, 397-413 (1999).
106. Weksberg,R. *et al.* Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Hum. Mol. Genet.* **11**, 1317-1325 (2002).
107. Horvath,S. *et al.* Obesity accelerates epigenetic aging of human liver. *Proc. Natl. Acad. Sci. U. S. A*(2014).
108. Sinclair,K.D. *et al.* DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status. *Proc. Natl. Acad. Sci. U. S. A* **104**, 19351-19356 (2007).
109. Rayner,K.J. *et al.* MiR-33 contributes to the regulation of cholesterol homeostasis. *Science* **328**, 1570-1573 (2010).
110. Sullivan,K.E. *et al.* Epigenetic regulation of tumor necrosis factor alpha. *Mol. Cell Biol.* **27**, 5147-5160 (2007).
111. Bell,J.T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).

Part 1 Characterizing the genetic architecture of inflammation biomarkers

Genetic architecture of the pro-inflammatory state in an extended twin-family design.

Abstract

In this study we examined the genetic architecture of variation in the pro-inflammatory state, using an extended twin-family design. Within the Netherlands Twin Register (NTR) Biobank, fasting Tumor Necrosis Factor-α (TNF-α), Interleukin-6 (IL-6), C-Reactive Protein (CRP) and fibrinogen levels were available for 3,534 twins, 1,568 of their non-twin siblings and 2,227 parents from 3,095 families. Heritability analyses took into account the effects of current and recent illness, anti-inflammatory medication, female sex hormone status, age, sex, BMI, smoking status, month of data collection, and batch processing. Moderate broad-sense heritability was found for all inflammatory parameters (39%, 21%, 45% and 46% for TNF-α, IL-6, CRP and fibrinogen, respectively). For all parameters, the remaining variance was explained by unique environmental influences and not by environment shared by family members. There was no resemblance between spouses for any of inflammatory parameters, except for fibrinogen. Also, there was no evidence for twin-specific effects. A considerable part of the genetic variation was explained by non-additive genetic effects for TNF-α, CRP and fibrinogen. For IL-6, all genetic variance was additive. This study may have implications for future genome-wide association (GWA) studies by setting a clear numerical target for genome-wide screens that aim to find the genetic variants regulating the levels of these pro-inflammatory markers.

Based on: 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; 1-10

* These authors contributed equally to the work

Introduction

Chronic low-grade inflammation plays an important role in numerous diseases including major depression and heart disease, and it has been implicated as one of the major causes for the comorbidity of these diseases [1-3]. The inflammatory response is activated by pro-inflammatory cytokines, of which TNF-α and IL-1 are the first to appear [4]. The inflammatory cascade is further promoted by the production of IL-6 that in turn stimulates the acute-phase response which is reflected in the synthesis of fibrinogen and CRP [5-7]. Elevations in TNF-α, IL-6, CRP and fibrinogen have been associated with an increased risk for both cardiac disease [7-15] as well as major depression [10, 11].

In spite of the obvious importance of these pro-inflammatory markers in depression and cardiovascular disease, which are both in the top 4 of burden of disease prediction for 2020 [16], very little is known about the etiology of the individual differences in TNF-α, IL-6, CRP and fibrinogen levels. A first important question is to what extent the variance in these biological parameters is innate, caused by environmental factors that are shared by family members, or caused by environmental factors unique to each individual member of a family. This question can be addressed by the classical twin design comparing the resemblance between monozygotic (MZ) and dizygotic (DZ) twins [17, 18]. A few twin studies in healthy samples have estimated the heritability of cytokines and acute phase reactants with estimates varying between 21% and 60% for fibrinogen [19-24], between 20% and 76% for CRP [21, 23-31], between 17% and 26% for TNF-α [21, 32], and between 15% and 61% for IL-6 [21, 23, 26, 28, 31-33]. With a few exceptions heritability estimates of the aforementioned studies have been based on relatively small twin samples. Such studies are fairly accurate in estimating broad-sense heritability but they lack precision and power to estimate the contribution of non-additive genetic effects or shared family environment like the dietary habits or neighborhood factors shared by parents and offspring. As the average sample size of previous studies was around 400 individuals, these studies were not sufficiently powered to detect an effect of shared environmental factors explaining less than 40% of the variance or to discriminate between additive and non-additive genetic factors [34]. Also, the relatively small sample sizes may explain the large range of heritability estimates based on previous studies.

Here we extend the classical twin design, including only MZ and DZ twin pairs, by including non-twin siblings, and their parents in the largest set of twin- and family data on TNF-α, IL-6, CRP, and fibrinogen described to date. Inclusion of non-twin siblings increases statistical power and offers the possibility to assess twin-specific effects. The inclusion of parents allows to take into account assortative (non-random) mating effects, which can influence heritability estimates. Data from parents also allow for the examination of shared household effects in spouses who share a household, but are not biologically related (e.g.[35, 36]). The availability of a large sample size allowed for the exclusion of subjects with current and recent illness and for the examination

of a number of health-related variables and methodological factors that could affect the reliability of the assessment of plasma levels of inflammatory variables, while retaining adequate power to detect shared environmental factors and to discriminate between additive and non-additive genetic factors.

Methods

Subjects

The data were obtained from the NTR Biobank study that was conducted among twins and their family members registered with the Netherlands Twin Register in the period of 2004-2008 [37]. Subjects were visited between 7 a.m. and 10 a.m. at home or, when preferred, at work, to collect blood and urine samples. Subjects were instructed to fast from the evening before, to abstain from physical exertion and, if possible, not to take medication at the day of the home visit, and to refrain from smoking one hour before the home visit. Fertile women were visited on the 2^{nd}-4^{th} day of their menstrual cycle or, if they took oral contraceptives, in their pill-free week. During the visit, a brief interview was conducted on health status, including an inventory of medication use, illness (last time occurrence, duration and type of illness), and adherence to the protocol.

The study consisted of 9,405 subjects with data on at least one of the four pro-inflammatory parameters of interest. Values exceeding 15 pg/ml for IL-6 and TNF-α, 15 mg/L for CRP and/or 6 g/L for fibrinogen were set to missing, leading to the exclusion of 11 subjects. Subjects who were on anti-inflammatory medication, medication impacting on the Hypothalamic Pituitary Adrenal (HPA)-axis, or both, were excluded from further analyses (N=408). We also excluded subjects suffering from a cold, the flu, inflammation, or allergy at the time of blood sampling (N=1,013). The remaining subjects (N=7,973) served as the reference group to quantify the effects of the various covariates and to compute residual scores for every immune parameter.

For the twin-family analyses we additionally excluded non-biological parents and siblings (N=35), spouses of twins (N=409), subjects under 18 years of age (N=87), the third member of triplets, and additional twins from families with more than one twin pair (N=4). When zygosity was missing for a twin pair and both twins participated in the study, we randomly selected one of the two to be excluded (N=10). To simplify the genetic model fitting procedure, we included a maximum of two singleton brothers and two singleton sisters per family and randomly selected two siblings from families with more than two same-sex siblings (N=99 excluded). The final sample was comprised of 3,095 families with 7,329 family members of which 3,534 subjects were twins, more specifically 590 MZ male (MZM), 320 DZ male (DZM), 1,281 MZ female (MZF), 624 DZ female (DZF) and 719 dizygotic opposite-sex (DOS) twins. The following number of complete twin pairs were included: 201 MZM, 96 DZM, 466 MZF, 211 DZF and 217 DOS. Furthermore, 464 non-twin male siblings, 1,104 non-twin female siblings, 1,003 fathers of twins and 1,224 mothers of twins

were included. Zygosity of twins was determined by DNA typing for 85.1% of the same-sex twin pairs. For the other same-sex pairs, zygosity was based on survey questions on physical similarity and the frequency of confusion of the twins by parents, other family members, and strangers. Agreement between zygosity based on these items and zygosity based on DNA was 96.1% [38].

Assessment of TNF-α, IL-6, CRP and fibrinogen

During the home visit, eight blood tubes were collected in the following order; 2 × 9 ml EDTA, 2 × 9 ml heparin, 1 × 4.5 ml CTAD, 1 × 2 ml EDTA, 1 x 4.5 ml serum. To prevent clotting, all tubes were inverted gently 8–10 times immediately after collection (for detail, see [37]).

Tumor Necrosis Factor-α (TNF-α) and Interleukin-6 (IL-6) were measured in EDTA plasma, obtained from one of the 9 ml tubes. During transport this tube was stored in melting ice and upon arrival at the laboratory, it was centrifuged for 20 minutes at 2000x *g* at 4°C. EDTA plasma, buffy coat, and red blood cells were harvested and aliquoted (0.5 ml), snap-frozen in dry ice, and stored at –30°C. Plasma levels of TNF-α and IL-6 were determined using an UltraSensitive ELISA (R&D systems, Minneapolis, USA, Quantikine HS HSTA00C). The inter-assay coefficient of variation (CV) for TNF-α was < 12.8%, for IL-6 the inter-assay CV was < 11.6%.

C-reactive protein (CRP) was obtained from one of the 9 ml heparin tubes. The tube was stored in melting ice during transport. At the laboratory the tube was centrifugated for 15 minutes at 1000x *g* at 4°C, after which heparin plasma was obtained and divided into 8 subsamples of 0.5 ml, snap-frozen and stored at –30°C. The processing took place in a sterile flow cabinet. CRP level in heparin plasma was determined using the Immulite 1000 CRP assay (Diagnostic Product Corporation, USA). The inter-assay CV was < 5.1%.

Fibrinogen. Fibrinogen level was obtained from the 4.5 ml CTAD tube, which was stored in melting ice during transport. Upon arrival at the laboratory, it was centrifuged for 20 minutes at 2000x *g* at 4°C, after which citrated plasma was harvested from the buffy coat and red blood cells, aliquoted (0.5 ml), snap-frozen in dry ice, and stored at –30° C. Fibrinogen levels in CTAD plasma were determined on a STA Compact Analyzer (Diagnostica Stago, France), using STA Fibrinogen (Diagnostica Stago, France). The inter-assay CV was < 6.1%. Fibrinogen values were normally distributed whereas data on the other variables were skewed. Therefore, we took the natural logarithm of these values.

Assessment of covariates

For the heritability analyses, we took into account the effects of age, sex, health-related covariates known to be associated with inflammatory parameters (body mass index (BMI), smoking status), and several methodological covariates that could lead to inflation of family correlations (month of blood sampling and batch effect). During the home visit, height and weight were

assessed and BMI was calculated. Subjects were also asked about their past and current smoking behavior and were categorized into one of five groups (never smoked, ex-occasional smoker, ex-regular smoker, current occasional smoker, current regular smoker). The month of blood sampling was used to correct for the effects of time of year on the four pro-inflammatory markers. For the cytokines, we also took into account differences in values due to the plate on which the samples were processed, by using the plate mean value for the cytokines as a covariate. The levels of the acute phase reactants were determined on a per sample basis, so plate effects for these variables are not applicable. Previous research suggested that when using the ELISA assay of R&D systems, individuals with blood group O may show higher TNF-α and IL-6 levels than other ABO blood groups, which may in part be due to assay-specific cross-reactivity with ABO antigens [39, 40]. To investigate this potential confounding effect we used a SNP (rs644234) that showed the strongest association with TNF-α and IL-6 in our data, of all SNPs in the ABO gene region plus/minus a 10 Kb border. The rs644234 SNP explained 7% and 4 % of TNF-α and IL-6 values, respectively. Data on this SNP were available for 5,950 healthy subjects with TNF-α data and for 5,947 subjects with IL-6 data. Because the twin-family models yielded similar results with and without taking the effect of the ABO SNP into account, we only report the analyses on the full sample.

Statistical analyses

Data preparation, sample selection and tests for the effects of covariates were conducted using IBM Statistical Package of Social Sciences 20.0. The covariates were included in a multiple regression analysis (forced entry) and the residual scores were saved for the heritability analyses. As there was a significant age-by-sex interaction for CRP, fibrinogen and IL-6, regression coefficients for age were estimated separately for men and women for these variables. Genetic models were fitted to the data using structural equation modeling (SEM) in the software package Mx [41]. First, a fully parameterized, or saturated, model was fitted and a goodness-of-fit statistic based on minus twice the logarithm of the likelihood (-2LL) was calculated. Next, the fully saturated model was simplified to a more restricted model to test whether constraints were allowed to be put on the data. The comparison of fit of a restricted model to the full model is performed by means of likelihood-ratio (χ^2) tests in which the difference in -2LL between the two models is calculated. When the likelihood-ratio test is significant ($p < .01$), the nested model is considered to fit significantly worse to the data than the fuller model it is tested against.
First, we tested if constraints on the means and variances for men and women were allowed and if different types of family correlations were equal. In the full model, 19 parameters were estimated: separate means, standard deviations, and regression coefficients for age on the phenotype for men and women, and 13 family correlations (for MZM, MZF, DZM, DZF, DOS twin pairs, male siblings,

female siblings, opposite-sex siblings and for mother-daughter, mother-son, father-daughter, father-son and one spouse correlation). Quantitative sex differences, indicating that the heritability of a trait is different in men and women, were assessed by testing whether correlations in male-male and female-female pairs of first-degree relatives (DZ twins and non-twin siblings) were equal. Next, we tested if the same genes regulate cytokine and acute phase reactant levels in men and women [42]. When correlations for a trait are the same in same-sex and opposite-sex pairs of family members, there is no evidence for qualitative sex differences in the genetic architecture. When the correlations in DZ twin pairs are of similar magnitude as the correlations in sib-sib pairs, there is no evidence for twin-specific resemblance. Generation effects were tested by equating parent-offspring correlations to the correlations between all other first-degree relatives (DZ twins and non-twin siblings). If this constraint is allowed, there is no evidence that gene expression changes with age. Spousal resemblance was assessed by testing if the correlation between the parents of the twins was significantly different from zero. The most parsimonious model with the maximal number of allowable restrictions was carried forward to the genetic structural equation analyses. In these analyses, the family covariance structure is used to estimate the relative contribution of latent additive (A) and non-additive or dominant (D) genetic factors and common (C) and unique (E) environmental factors to the phenotypic variance. Based on the variance estimates from the full genetic model, sequentially constrained submodels were compared to the fit of the full model to arrive at the most parsimonious genetic model describing the total phenotypic variance best (see figure 1 for a schematic representation of the extended twin-family model).

Results

Descriptive statistics for the four immune parameters of interest in the twins, siblings and parents are given in table 1. Table 2 presents the amount of variance explained by the various technical and biological covariates that were taken into account.

The effect of age on all parameters was significant and positive (p's < .01). Sex differences were present only for CRP and fibrinogen with women having significantly larger mean values than men (p's < .01). For TNF-α and CRP values, standard deviations were significantly larger in females than in males. Table 3 shows the family correlations for each of the immune parameters estimated for the values only adjusted for age and sex and for the values additionally adjusted for the other covariates.

Figure 1. Path diagram of an extended twin family showing four subjects belonging to one family: father (Pfa), mother (Pmo), twin 1 (Pt1) and twin 2 (Pt2), where P is phenotype. In the genetic model (without constraints) a maximum of 10 parameters is estimated. Path loadings a, d, c and e represent the influence of additive and non-additive genetic effects and shared and unique environmental factors on the phenotype. Spousal resemblance is modeled as a function of phenotypic assortment, denoted by the delta path (Δ). When the delta path coefficient is larger than zero, this indicates non-random mating. The residual variance of additive genetic effects in parents (Ra) will then become larger than 1 to account for increased transmission of additive genetic effects from parents to children under conditions of non-random mating. Finally, two means and two age regression coefficients (for males and females separately) are estimated. Fixed parameters include the .5 genetic relatedness of the parents and their children (path running from the A of both parents to the A of their children with a path loading of .5). The residual variance of A in the offspring is constrained at .5. Finally, MZ twins correlate 1 regarding A and D factor scores (because they are assumed to share all of their genetic material), whereas DZ or sibling pairs correlate .5 and .25, respectively.

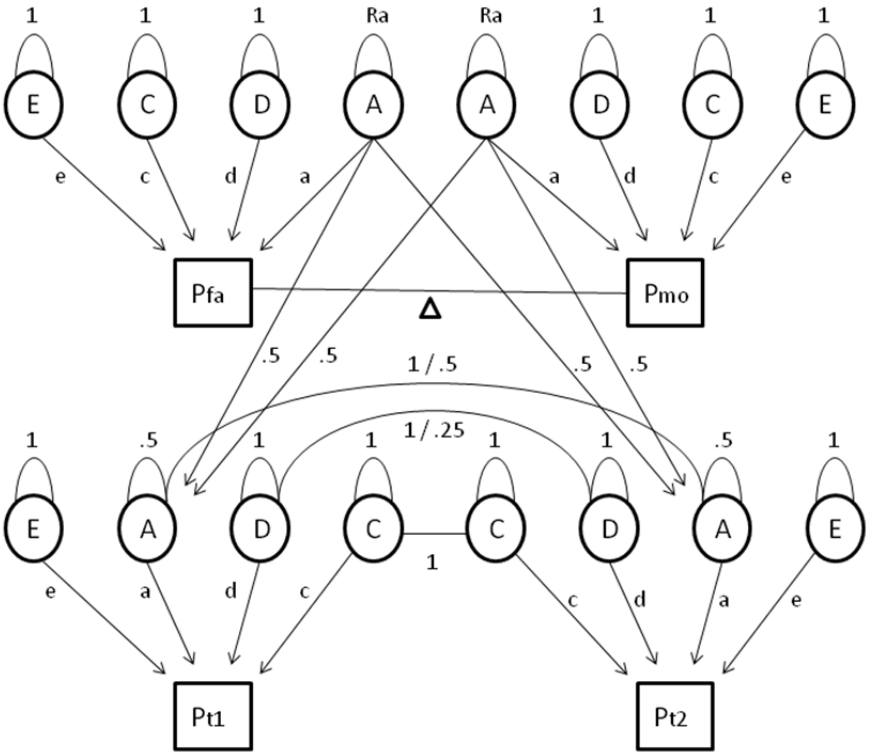


Table 1. Mean values (and standard deviations) and mean age (range) for Tumor Necrosis Factor-α (TNF-α), Interleukin-6 (IL-6), C-Reactive Protein (CRP) and fibrinogen.

Marker	N total	Mean (sd)	Mean age (range)
TNF-α (pg/ml)			
Fathers	987	1.21 (1.25)	61 (33-89)
Mothers	1,215	1.20 (1.10)	60 (26-89)
Male twins/siblings	1,594	1.02 (0.85)	37 (18-82)
Female twins/siblings	3 218	1.07 (1.14)	38 (18-90)
Total	7,014	1.10 (1.10)	45 (18-90)
IL-6 (pg/ml)			
Fathers	984	2.10 (1.77)	61 (33-88)
Mothers	1,213	1.91 (1.50)	60 (26-89)
Male twins/siblings	1,590	1.37 (1.38)	37 (18-82)
Female twins/siblings	3,220	1.41 (1.27)	38 (18-90)
Total	7,007	1.59 (1.44)	45 (18-90)
CRP (mg/L)			
Fathers	975	2.47 (2.58)	61 (33-89)
Mothers	1,171	2.73 (2.75)	60 (26-89)
Male twins/siblings	1,672	1.77 (2.23)	35 (18-82)
Female twins/siblings	3,244	2.59 (2.91)	38 (18-90)
Total	7,062	2.40 (2.71)	44 (18-90)
Fibrinogen (g/L)			
Fathers	983	2.94 (0.70)	61 (33-89)
Mothers	1,188	3.02 (0.68)	60 (26-89)
Male twins/siblings	1,550	2.51 (0.59)	37 (18-82)
Female twins/siblings	3,136	2.69 (0.65)	38 (18-90)
Total	6,857	2.74 (0.68)	45 (18-90)

For all parameters male and female MZ correlations did not differ significantly and same-sex and opposite- sex DZ twin and non-twin sibling correlations were also similar in all cases (p's > .01), so no quantitative and qualitative sex differences were present, nor did we find evidence for twin-specific environmental effects. Parent-offspring correlations were not significantly different from DZ twin and non-twin sibling correlations ($p > .01$), except for the fibrinogen values adjusted for age and sex only ($p = .003$). This effect was not present in the fully adjusted fibrinogen values. These results suggest that genetic regulation of cytokine and acute phase reactant levels does not change significantly across age. Overall, adjustment for BMI, smoking, month and plate effects in addition to age and sex tended to reduce all familial correlations, including the spouse correlations. Only the spouse correlation for the fully adjusted fibrinogen values was significantly different from zero ($r = .16$, $p < .01$),

which indicates that the effects of assortative mating or sharing a household without being biologically related are negligible, except for the small resemblance found for fibrinogen.

For the heritability analyses on the fully adjusted values, contributions of A, D, C and E factors to the total phenotypic variance were constrained over sex while taking into account sex differences in phenotypic variance in TNF-α, CRP and IL-6. Assortative mating was only modeled for fibrinogen. Table 4 shows the genetic models that were fitted to the data, supplemented with the proportions of the phenotypic variance that can be explained by the different genetic and environmental factors for both the full ADCE model and the model that provided the most parsimonious fit.

The broad-sense heritability was 39%, 21%, 45% and 46% for TNF-α, Il-6, CRP and fibrinogen, respectively. The models that provided the best fit to the data on TNF-α, CRP and fibrinogen included additive and non-additive genetic factors, and unique environmental factors. Non-additive genetic effects explained 22% of the variance of TNF-α, 18% of the variance of CRP, and 16% of the variance of fibrinogen. For CRP and fibrinogen, a small amount of variation was attributed to sibling-shared environmental factors in the full ADCE model, but an ADE model without shared environmental factors did not fit significantly worse. For IL-6, a model with additive genetic factors and unique environmental factors explained the data best, with no role for non-additive genetic factors, nor for shared environmental factors.

Table 2. Proportion of variance that is explained by the covariate with the number of subjects within brackets .

Covariates	TNF-α	IL-6	CRP	Fibrinogen
Age	.021** (7,566)	.118** (7,559)	.024** (7,684)	.097** (7,397)
Sex	.000 (7,566)	.000 (7,559)	.014** (7,684)	.008** (7,397)
BMI	.013** (7,521)	.098** (7,515)	.139** (7,645)	.100** (7,357)
Smoking	.001* (7,553)	.030 (7,546)	.007** (7,678)	.009** (7,392)
Plate effect	.106** (7,558)	.079** (7,551)	-	-
Month of blood sampling	.006** (7,566)	.007** (7,559)	.002** (7,684)	.009** (7,397)

* = p < .05, ** = p < .01

Table 3. Family correlations (and 95% confidence intervals) as estimated in the saturated model for the cytokines and the acute pha reactants, with the levels of the pro-inflammatory markers only adjusted for age and sex, and adjusted for all covariates.

	TNF-α[1]	TNF-α[2]	IL-6[1]	IL-6[2]	CRP[1]	CRP[2]	Fibr[1]	Fibr[2]
MZ twins								
MZM	.44 (.29-.55)	.44 (.29-.56)	.23 (.09-.35)	.19 (.04-.32)	.42 (.31-.51)	.39 (.27-.49)	.50 (.37-.60)	.46 (.31-.57)
MZF	.38 (.31-.45)	.38 (.31-.45)	.39 (.31-.46)	.35 (.26-.42)	.50 (.43-.56)	.46 (.39-.52)	.48 (.41-.55)	.45 (.37-.52)
DZ/siblings male								
DZM	-.02 (-.19-.15)	-.04 (-.22-.14)	.13 (-.06-.31)	.20 (.01-.36)	.23 (.06-.39)	.15 (-.03-.32)	.39 (.14-.57)	.31 (.00-.52)
MZM/DZM/brother-brother	.14 (.01-.26)	.11 (-.02-.23)	.05 (-.05-.16)	.00 (-.10-.11)	.27 (.16-.38)	.22 (.10-.33)	.41 (.27-.53)	.34 (.18-.47)
DZ/siblings female								
DZF	.17 (.05-.28)	.14 (.02-.25)	.19 (.06-.30)	.13 (.00-.26)	.34 (.22-.44)	.32 (.19-.43)	.30 (.15-.43)	.26 (.09-.40)
MZF/DZF/sister-sister	.19 (.11-.26)	.20 (.12-.27)	.14 (.07-.21)	.10 (.03-.17)	.23 (.16-.30)	.20 (.12-.27)	.25 (.17-.32)	.22 (.14-.29)
Opposite sex siblings								
DOS	.10 (-.03-.24)	.12 (-.01-.25)	.12 (-.01-.25)	.09 (-.04-.22)	.22 (.10-.33)	.23 (.11-.33)	.15 (.01-.28)	.12 (-.04-.27)
DOS/brother-sister	.12 (.04-.20)	.10 (.02-.18)	.09 (.01-.17)	.07 (-.01-.15)	.21 (.14-.28)	.18 (.10-.25)	.28 (.20-.36)	.24 (.15-.32)
Parent-offspring								
Mother-daughter	.14 (.07-.20)	.12 (.05-.18)	.14 (.08-.20)	.06 (.00-.12)	.16 (.09-.22)	.10 (.03-.16)	.22 (.16-.28)	.18 (.12-.24)
Mother-son	.14 (.05-.22)	.09 (.00-.18)	.20 (.13-.27)	.17 (.10-.25)	.14 (.06-.22)	.09 (.00-.17)	.17 (.08-.25)	.15 (.06-.23)
Father-son	.10 (.01-.18)	.13 (.04-.22)	.16 (.07-.24)	.18 (.09-.26)	.22 (.13-.30)	.22 (.13-.30)	.26 (.17-.34)	.26 (.17-.34)
Father-daughter	.08 (.02-.14)	.06 (.00-.12)	.11 (.05-.18)	.03 (-.03-.10)	.20 (.14-.27)	.15 (.08-.21)	.14 (.08-.20)	.14 (.08-.20)
Parents								
Mother-father	.13 (.06-.20)	.09 (.02-.16)	.13 (.05-.20)	.05 (-.03-.13)	.12 (.03-.20)	.05 (-.04-.14)	.18 (.12-.24)	.16 (.10-.22)

[1] after adjustment for the effects of age and sex.
[2] after adjustment for the effects of all covariates (age, sex, month, BMI, smoking and for TNF-α and IL-6 plate as well).

Table 4. Genetic model fit statistics of the pro-inflammatory markers after adjustment for sex, age, BMI, smoking, plate and month of sampling.

Marker	Model	df	**Model**	**-2LL**	**vs**	**$\Delta\chi^2$**	**Δ df**	**p**	**a^2**	**d^2**	**c^2**	**e^2**
TNF-α	1	6951	ADCE	19404.352					.17 (.10-.24)	.22 (.12-.31)	.00 (.00-.05)	.61 (.55-.67)
	2	6953	AE	19425.261	1	20.909	2	.000				
	3	**6952**	**ADE**	**19404.352**	**1**	**0.000**	**1**	**1**	**.17 (.10-.24)**	**.22 (.13-.31)**	**-**	**.61 (.55-.67)**
IL-6	1	6944	ADCE	19352.234					.20 (.09-.25)	.00 (.00-.00)	.00 (.00-.06)	.79 (.70-.84)
	2	**6946**	**AE**	**19352.248**	**1**	**0.014**	**2**	**.993**	**.21 (.16-.25)**	**-**	**-**	**.79 (.75-.84)**
CRP	1	7016	ADCE	19296.178					.25 (.18-.33)	.13 (.03-.24)	.05 (.00-.11)	.56 (.51-.62)
	2	7018	AE	19314.325	1	18.148	2	.000				
	3	7017	ACE	19302.277	1	6.100	1	.014				
	4	**7017**	**ADE**	**19298.509**	**1**	**2.331**	**1**	**.127**	**.27 (.20-.34)**	**.18 (.09-.26)**	**-**	**.55 (.50-.61)**
Fibrinogen	1	6807	ADCE	18524.411					.30 (.23-.36)	.11 (.00-.23)	.05 (.00-.12)	.55 (.49-.62)
	2	6809	AE	18538.390	1	13.979	2	.001				
	3	6808	ACE	18527.761	1	3.349	1	.067				
	4	**6808**	**ADE**	**18525.974**	**1**	**1.563**	**1**	**.211**	**.30 (.24-.37)**	**.16 (.09-.24)**	**-**	**.54 (.48-.60)**

Abbreviations: df = degrees of freedom, Model = specification of the model that is tested, -2LL = minus twice the logarithm of the likelihood, vs = the model against which this submodel is tested, $\Delta\chi^2$ = model fit statistic: difference in -2LL of two nested models' Δdf = the difference in the number of parameters between the models, p = p-value (was regarded significant when < .01), a^2, d^2, c^2, e^2 = proportions of variance explained by additive, non-additive, shared and unique environmental effects.
The 95% confidence intervals are depicted within brackets. The most parsimonious model is printed in bold.

Discussion

This is the most comprehensive twin-family study of the genetic architecture of the pro-inflammatory state that has been performed thus far. Results replicate the importance of genetic factors in pro-inflammation observed before [19-33] and extend the findings of previous studies by showing that genetic non-additivity is an important factor in explaining individual differences in TNF-α, CRP and fibrinogen levels and by ruling out a large role for environmental factors shared by family members.

There have only been three previous heritability studies employing a sample size of over 1,000 twins for CRP [25, 30], IL-6 and TNF-α [32]. For fibrinogen, the study with the largest sample size included 962 subjects [20]. None of these studies systematically corrected for recent illness, medication use, menstrual cycle, oral contraceptives use, batch effects, month of sampling, BMI, and smoking status as done in the present study. In spite of the more strict correction for confounders, our heritability estimate for CRP was of comparable magnitude to these previous studies. For CRP, we confirmed the importance of non-additive genetic factors that was found in the largest of the two previous twin studies [30], whereas the smaller of the two [25] only detected additive effects, likely reflecting insufficient power. For fibrinogen, our broad-sense heritability estimate was comparable to the estimate reported by de Lange and colleagues [20], but our study additionally indicated that a significant proportion of the heritability was due to non-additive genetic effects.

For TNF-α and IL-6 our results do not completely support the results of the only large (N > 1,000) previous twin study [32]. For both cytokines, Sas and colleagues [32] found substantial family resemblance, but they could not discriminate between genetics and shared environment as the source of that resemblance. The twin correlations reported in a smaller study on IL-6 [28] were suggestive of genetic factors and a potential role of shared environment with the MZ correlation being less than twice as high as the DZ correlation. Our study clearly shows that shared genetic make-up rather than shared family environment is the major source of familial resemblance in these parameters. Furthermore, we show a significant effect of non-additive factors on TNF-α. Four smaller twin studies of TNF-α and IL-6 values in healthy unchallenged subjects are consistent with our findings, as the MZ correlations found in those studies were about twice as high as the DZ correlations, in elderly subjects [21], young adults [33], and middle-aged twins [23, 26]. In their sample of young adult subjects, Grunnet and colleagues [33] even found the MZ correlations for TNF-α to be more than twice as high than the DZ correlations.

Taken together, our results and those from previous studies suggest that about a third of the variation in these core pro-inflammatory cytokines and acute phase reactants in healthy subjects with values in the non-extreme range is explained by genetic variation. This means that some individuals are more susceptible than others to have higher levels of pro-inflammatory markers and this increased susceptibility is, at least partly, due to genetic differences

between individuals. Large scale collaborative attempts to find the actual genes that underlie this genetic variation are under way. In 2011, a meta-analysis of GWA studies of CRP in over 80,000 subjects identified several genes implicated in immune system functioning and inflammation (*CRP, IL6R, NLRP3, IL1F10, IRF1, PPP1R3B, SALL1, PABPC4, ASCL1, RORA* and *BCL7B*) and the metabolic syndrome (*APOC1, HNF1A, LEPR, GCKR, HNF4A* and *PTPN2*) to be associated with circulating CRP levels [43]. In a meta-analysis of six GWA studies on fibrinogen in over 22,000 subjects, significant genome-wide hits were found in the *FGB, IRF1, PCCB* and *NLRP3* genes [44]. For IL-6 and TNF-α no meta-analysis has been published to our knowledge. A single large GWA study on IL-6 (N=6,145) found significant hits in the IL6R and ABO genes [40]. With our study we have accomplished a clear numerical target for these ongoing genome-wide screens that aim to find the actual genetic variants regulating the levels of these pro-inflammatory markers. Heritability studies conducted in large representative samples continue to be valuable, because the heritability of traits can vary between populations and can change across generations. We should also keep in mind that 54 to 79% of the variation found was due to unique environmental factors that are not shared between family members. This estimate may derive from unique environmental factors or measurement error, but it may also result from gene-by-environment interactions that may inflate estimates of unique environmental effects. Unravelling the genetics of these pro-inflammatory parameters may greatly contribute to our understanding of the aetiology of cardiac disease and major depression since chronic low-grade inflammation has repeatedly been shown to be associated with both [7-15].

Because of the large sample size and the extended twin-family design this study had sufficient power to decompose the variance in the levels of an important set of pro-inflammatory markers into additive and non-additive genetic factors, and shared and unique environmental factors. The inclusion of parents and siblings allowed us to detect and correct for assortative mating, quantitative and qualitative sex differences, and effects of age that could potentially affect the heritability estimates. The sample size also allowed us to employ strict exclusion criteria concerning the recent health status and medication use of the subjects without losing power so that we were ensured analyses were run on healthy individuals only. Furthermore, our study design controlled for female sex hormone status.

This study also had limitations. First, we selected only a subset of the many immune parameters that co-determine the pro-inflammatory state, including IL-1 and interferon-γ, and we did not take into account the action of the soluble receptors for the cytokines, levels of which may be substantial heritable. Secondly, we used the ELISA assay by R&D systems that may yield higher TNF-α and IL-6 values in individuals of blood group O levels than other ABO blood groups which may in part by due to assay specific cross-reactivity with ABO antigens [39, 40]. We indeed found a significant association of SNPs in the

ABO region to TNF-α and IL-6. Although it explained only a small amount of variance in TNF-α (7%) and IL-6 (4%) compared to the larger effects of plate and BMI, they may cause overestimation of non-additivity or underestimation of shared environment since the shared blood group O will make MZ twins appear more alike than DZ twins or non-twin siblings. Thirdly, we tested whether different genes are expressed at different ages by testing whether parent-offspring correlations and correlations in first-degree relatives (DZ twins and non-twin siblings) were of comparable magnitude. Because there was a partial overlap in age between the parent and the offspring generation, we cannot completely rule out the possibility that the expression of pro-inflammatory genes changes across age.

In conclusion, the familial resemblance in these core pro-inflammatory cytokines and acute phase reactants is explained by genetic variation and not by the shared family environment. For three out of four markers, both additive and non-additive genetic factors contribute to the heritability.

Reference List

1. Capuron,L. *et al.* Depressive Symptoms and Metabolic Syndrome: Is Inflammation the Underlying Link? *Biol. Psychiatry*(2008).
2. Vaccarino,V. *et al.* Association of Major Depressive Disorder with Serum Myeloperoxidase and Other Markers of Inflammation: A Twin Study. *Biol. Psychiatry*(2008).
3. Vaccarino,V. *et al.* Depression, inflammation, and incident cardiovascular disease in women with suspected coronary ischemia: the National Heart, Lung, and Blood Institute-sponsored WISE study. *J. Am. Coll. Cardiol.* **50**, 2044-2050 (2007).
4. Tracey,K.J. The inflammatory reflex. *Nature* **420**, 853-859 (2002).
5. Gabay,C. Interleukin-6 and chronic inflammation. *Arthritis Res. Ther.* **8 Suppl 2**, S3 (2006).
6. Gabay,C. & Kushner,I. Acute-phase proteins and other systemic responses to inflammation. *N. Engl. J. Med.* **340**, 448-454 (1999).
7. Packard,R.R. & Libby,P. Inflammation in atherosclerosis: from vascular biology to biomarker discovery and risk prediction. *Clin. Chem.* **54**, 24-38 (2008).
8. Danesh,J. *et al.* Long-term interleukin-6 levels and subsequent risk of coronary heart disease: two new prospective studies and a systematic review. *PLoS. Med.* **5**, e78 (2008).
9. Danesh,J. *et al.* C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *N. Engl. J. Med.* **350**, 1387-1397 (2004).
10. O'Brien,S.M., Scott,L.V., & Dinan,T.G. Cytokines: abnormalities in major depression and implications for pharmacological treatment. *Hum. Psychopharmacol.* **19**, 397-403 (2004).
11. Penninx,B.W. *et al.* Inflammatory markers and depressed mood in older persons: results from the Health, Aging and Body Composition study. *Biol. Psychiatry* **54**, 566-572 (2003).
12. Libby,P. & Theroux,P. Pathophysiology of coronary artery disease. *Circulation* **111**, 3481-3488 (2005).
13. Cesari,M. *et al.* Inflammatory markers and cardiovascular disease (The Health, Aging and Body Composition [Health ABC] Study). *Am. J. Cardiol.* **92**, 522-528 (2003).

14. Woods,A., Brull,D.J., Humphries,S.E., & Montgomery,H.E. Genetics of inflammation and risk of coronary artery disease: the central role of interleukin-6. *Eur. Heart J.* **21**, 1574-1583 (2000).
15. Humphries,S.E., Cooper,J.A., Talmud,P.J., & Miller,G.J. Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men. *Clin. Chem.* **53**, 8-16 (2007).
16. Mathers,C.D. & Loncar,D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS. Med.* **3**, e442 (2006).
17. Boomsma,D., Busjahn,A., & Peltonen,L. Classical twin studies and beyond. *Nat Rev Genet* **3**, 872-882 (2002).
18. van,D.J., Slagboom,P.E., Draisma,H.H., Martin,N.G., & Boomsma,D.I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **13**, 640-653 (2012).
19. Reed,T., Tracy,R.P., & Fabsitz,R.R. Minimal genetic influences on plasma fibrinogen level in adult males in the NHLBI twin study. *Clin. Genet.* **45**, 71-77 (1994).
20. de,L.M., Snieder,H., Ariens,R.A., Spector,T.D., & Grant,P.J. The genetics of haemostasis: a twin study. *Lancet* **357**, 101-105 (2001).
21. de Maat,M.P. *et al.* Genetic influence on inflammation variables in the elderly. *Arterioscler. Thromb. Vasc. Biol.* **24**, 2168-2173 (2004).
22. de,L.M. *et al.* Genetic influences on fibrinogen, tissue plasminogen activator-antigen and von Willebrand factor in males and females. *Thromb. Haemost.* **95**, 414-419 (2006).
23. Su,S. *et al.* Genetic and environmental influences on systemic markers of inflammation in middle-aged male twins. *Atherosclerosis* **200**, 213-220 (2008).
24. Jermendy,G. *et al.* Effect of genetic and environmental influences on cardiometabolic risk factors: a twin study. *Cardiovasc. Diabetol.* **10**, 96 (2011).
25. Wang,G. *et al.* C-reactive protein in adolescent twins: patterns and relationship to adiposity. *J. Clin. Endocrinol. Metab* **96**, 3226-3233 (2011).
26. Su,S. *et al.* Common genetic contributions to depressive symptoms and inflammatory markers in middle-aged men: the Twins Heart Study. *Psychosom. Med.* **71**, 152-158 (2009).
27. MacGregor,A.J., Gallimore,J.R., Spector,T.D., & Pepys,M.B. Genetic effects on baseline values of C-reactive protein and serum amyloid a protein: a comparison of monozygotic and dizygotic twins. *Clin. Chem.* **50**, 130-134 (2004).
28. Worns,M.A., Victor,A., Galle,P.R., & Hohler,T. Genetic and environmental contributions to plasma C-reactive protein and interleukin-6 levels--a study in twins. *Genes Immun.* **7**, 600-605 (2006).
29. Wessel,J. *et al.* C-reactive protein, an 'intermediate phenotype' for inflammation: human twin studies reveal heritability, association with blood pressure and the metabolic syndrome, and the influence of common polymorphism at catecholaminergic/beta-adrenergic pathway loci. *J. Hypertens.* **25**, 329-343 (2007).
30. Rahman,I. *et al.* Genetic dominance influences blood biomarker levels in a sample of 12,000 Swedish elderly twins. *Twin. Res. Hum. Genet.* **12**, 286-294 (2009).
31. Su,S. *et al.* Pleiotropy of C-reactive protein gene polymorphisms with C-reactive protein levels and heart rate variability in healthy male twins. *Am. J. Cardiol.* **104**, 1748-1754 (2009).
32. Sas,A.A. *et al.* The age-dependency of genetic and environmental influences on serum cytokine levels: a twin study. *Cytokine* **60**, 108-113 (2012).

33. Grunnet,L., Poulsen,P., Klarlund,P.B., Mandrup-Poulsen,T., & Vaag,A. Plasma cytokine levels in young and elderly twins: genes versus environment and relation to in vivo insulin action. *Diabetologia* **49**, 343-350 (2006).
34. Posthuma,D. & Boomsma,D.I. A note on the statistical power in extended twin designs. *Behav. Genet.* **30**, 147-158 (2000).
35. Distel,M.A. *et al.* Familial resemblance for loneliness. *Behav. Genet.* **40**, 480-494 (2010).
36. Rebollo,I. & Boomsma,D.I. Genetic and environmental influences on type A behavior pattern: evidence from twins and their parents in the Netherlands Twin Register. *Psychosom. Med.* **68**, 437-442 (2006).
37. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin. Res. Hum. Genet.* **13**, 231-245 (2010).
38. Willemsen,G. *et al.* The adult Netherlands twin register: twenty-five years of survey and biological data collection. *Twin. Res. Hum. Genet.* **16**, 271-281 (2013).
39. Melzer,D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS. Genet.* **4**, e1000072 (2008).
40. Naitza,S. *et al.* A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS. Genet.* **8**, e1002480 (2012).
41. Neale,M.C., Boker,S.M., Xie,G., & Maes,H.H. *Mx: Statistical Modeling*(VCU, Richmond,2006).
42. Vink,J.M. *et al.* Sex differences in genetic architecture of complex phenotypes? *PLoS. One.* **7**, e47371 (2012).
43. Dehghan,A. *et al.* Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* **123**, 731-738 (2011).
44. Dehghan,A. *et al.* Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. *Circ. Cardiovasc. Genet.* **2**, 125-133 (2009).

The contribution of the functional *IL6R* polymorphism rs2228145, eQTLs and other genome-wide SNPs to the heritability of plasma sIL-6R levels

Abstract

The non-synonymous SNP rs2228145 in the *IL6R* gene on chromosome 1q21.3 is associated with a wide range of common diseases, including asthma, rheumatoid arthritis, type 1 diabetes and coronary heart disease. We examined the contribution of this functional *IL6R* gene polymorphism rs2228145 versus other genome-wide SNPs to the variance of sIL-6R levels in blood plasma in a large population-based sample (N~5000), and conducted an expression QTL (eQTL) analysis to identify SNPs associated with *IL6R* gene expression. Based on data from 2360 twin families, the broad heritability of sIL-6R was estimated at 72%, and 51% of the total variance was explained by the functional SNP rs2228145. Converging findings from GWAS, linkage, and GCTA analyses indicate that additional variance of sIL-6R levels can be explained by other variants in the *IL6R* region, including variants at the 3'end of *IL6R* tagged by rs60760897 that are associated with *IL6R* RNA expression.

Based on: van Dongen J, Jansen R, Smit D, Hottenga JJ, Mbarek H, Willemsen G, Kluft C, AAGC collaborators[ii], Penninx BW, Ferreira MA, Boomsma DI, de Geus EJ. The Contribution of the Functional IL6R Polymorphism rs2228145, eQTLs and Other Genome-Wide SNPs to the Heritability of Plasma sIL-6R Levels. Behav.Genet. 2014; 44:368-382

[ii] Australian Asthma Genetics Consortium Collaborators

Introduction

The interleukin-6 receptor (IL-6R) forms part of the ligand-receptor complex that mediates the activities of interleukin-6 (IL-6). So-called classical IL-6 signaling occurs in hepatocytes and some leukocyte subtypes, which express a trans-membrane form of IL-6R on their surface (membrane-bound IL-6R, or mIL-6R)[1]. A second type of IL-6 signaling called IL-6 trans-signaling is capable of stimulating a variety of different cell types as it is mediated by a soluble form of the IL-6R (soluble IL-6R, or sIL-6R)[2, 3]. IL-6 trans-signaling allows cells lacking mIL-6R to respond to IL-6, as long as they express the trans-membrane signal transducer protein gp130 on their surface (which is thought to be expressed ubiquitously)[4]. On target cells, the complex of IL-6 and IL-6R binds to two molecules of gp130, thereby activating several signal transduction pathways [5], leading to cellular responses such as proliferation, differentiation and inflammatory processes. IL-6 trans-signaling plays a key role in several autoimmune diseases and inflammatory conditions, including asthma [6], rheumatoid arthritis [7], chronic inflammatory bowel disease [8], some types of cancer (e.g. multiple myeloma)[9, 10] and peritonitis[11].

Two isoforms of sIL-6R have been identified in blood plasma of healthy individuals that are generated through different mechanisms [12, 13]. The majority of sIL-6R is thought to be produced by a process called shedding, referring to the proteolytic cleavage of mIL-6R and subsequent release of the ligand-binding ectodomain into the extracellular space [14]. A second isoform is produced by translation of an alternatively spliced mRNA lacking a 94-bp sequence coding for part of the transmembrane domain [15, 16] that anchors the receptor to the cell membrane. The process of shedding is affected by a non-synonymous SNP (Asp358Ala or rs2228145 (A > C), previously also known as rs8192284) that occurs within the region encoding the proteolytic cleavage site, in exon 9 of the interleukin-6 receptor gene, *IL6R*, on chromosome 1q21.3 [17]. The SNP causes a striking difference in IL-6R concentrations between carriers of different alleles, with reduced concentrations of mIL-6R and increased concentrations of sIL-6R in carriers of the minor allele (C) [18-22]. Although some previous studies found no association of rs2228145 with overall *IL6R* RNA expression [23] or expression of the RNA transcript encoding mIL-6R [18], a positive association has been reported for the rs2228145 C allele and expression level of the alternatively spliced mRNA [10, 18], and we and others found a negative association between the rs2228145 C allele and overall expression level [24].

Variants in *IL6R* are associated with the risk of a wide spectrum of common diseases, with the rs2228145 C allele increasing susceptibility to asthma [25], and decreasing susceptibility to other diseases including rheumatoid arthritis [26], coronary heart disease [23], and type 1 diabetes [18]. Two consortia reported a protective effect of the C allele on the risk of coronary heart disease and emphasized the potential of tocilizumab, a monoclonal antibody against IL-6R used for treatment of chronic inflammatory disease, as a novel therapeutic

strategy to prevent cardiovascular disease [23, 27]. Growing interest in the role of *IL6R* in complex disease and in therapeutic strategies targeting the IL-6R pathway highlight the value of novel insights into genetic determinants of IL-6R level. The associations with disease reported for multiple variants in the *IL6R* gene are generally ascribed to LD with rs2228145. Two recent studies showed that some additional variance in sIL-6R level is explained by other SNPs within *IL6R* [18, 24], but it is unknown how much of the variance of sIL-6R levels in total is due to variants other than rs2228145 that remain to be identified.

We describe a series of analyses conducted in a population-based sample of ~5000 Dutch individuals (Supplementary Table S1) aimed at evaluating the contribution of the known functional polymorphism rs2228145 to the variation in sIL-6R levels, and the contribution of other genetic variants within the *IL6R* region and the rest of the genome. We analyzed plasma sIL-6R levels in a large sample of twins and their family members to estimate the total heritability of sIL-6R, and conducted GWA, linkage, and eQTL analyses to identify additional genetic variants influencing sIL-6R level. Using the data from twin families and the classic biometrical model, we estimate the broad heritability of sIL-6R level to be 72 % and show that rs2228145 accounts for 51% of the total variance of sIL-6R level. Results from linkage analysis corroborate this and indicate that genetic variation within the *IL6R* region on chromosome 1 explains 69% of the variation in sIL-6R, of which 19% is due to genetic variation other than rs2228145. Results from eQTL analysis point towards a role of genetic variants at the 3'end of *IL6R* contributing to the levels of sIL-6R and *IL6R* RNA in blood. In passing, we provide empirical evidence that different methods based on the same corpus of genetic theory, including the classic biometrical model, the twin-family model, linkage analysis at the *IL6R* locus, and genome-wide SNP sharing in unrelated individuals (GCTA) all converge to the same conclusion.

Materials and Methods

Subjects

Plasma sIL-6R level data were available for 8929 participants from the Netherlands Twin Register (NTR)[28], of which 5945 individuals also had genome-wide SNP data. Data from an additional 1966 participants from the Netherlands Study of Depression and Anxiety (NESDA[29]) were included in the eQTL analysis. Individuals using anti-inflammatory medication or medication influencing the HPA (Hypothalamic-Pituitary-Adrenal)-axis (NTR: N = 426/4.8%, NESDA: N=538/26%) at the time of blood sampling, or with a sIL-6R level > 100.000 pg/mL (N=6/0.07% of NTR subjects) were excluded from all analyses. For a detailed description of the characteristics of the subjects included in each analysis, see Supplementary Methods. NTR and NESDA studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the US Office of Human Research Protections (IRB

number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180, NESDA 03-183). All subjects provided written informed consent.

Blood sampling
Blood sampling procedures have been described previously [30, 31]. In short; venous blood samples were drawn in the morning after an overnight fast. For RNA measurement, heparinized whole blood from NESDA en NTR participants was transferred within 20 minutes of sampling into PAXgene Blood RNA tubes (Qiagen) and stored at -20°C. For sIL-6R measurement, EDTA plasma tubes were collected from NTR participants, and were centrifuged for 20 minutes at 2000x *g* at 4°C. EDTA plasma, buffy coat and red blood cells were harvested and aliquoted (0.5 ml), snap-frozen in dry ice, and stored at -30°C.

sIL-6R level
sIL-6R level was measured in EDTA plasma samples (diluted 1:100) using the Quantikine Elisa Human IL-6 sR assay of R&D systems. The inter-assay and intra-assay coefficient of variation were < 15 %. In all analyses, sIL-6R level was adjusted for sex and age by inclusion of covariates or by analysis of residualized levels. sIL-6R level was on average higher in males (4.25×10^{-8} g/mL) compared to females (4.14×10^{-8} g/mL) and increased with age (1 SD of age (14 years) was associated with an increase in sIL-6R level of 0.17×10^{-8} g/mL).

Genotype data
DNA extraction has been described before [32]. Genotyping was done on multiple chip platforms, for several partly overlapping subsets of participants. The following platforms were used: Affymetrix Perlegen 5.0, Illumina 370, Illumina 660, Illumina Omni Express 1M and Affymetrix 6.0. After array specific data analysis, genotype calls were made with the platform specific software (Genotyper, Beadstudio). In total, genotype data were available for 12,133 subjects from NTR and NESDA. The extensive genotyping quality control steps and imputation procedures are described in the Supplementary Methods. All analyses were performed on 1000G-imputed data (phase I Interim release All panel (sequence data freeze 23/11/2010), release June 2011, https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_interim.html.

IL6R expression
RNA extraction [30, 31] and expression QC procedures have been described in detail previously [33]. PAXgene tubes were shipped to the Rutgers University Cell and DNA Repository (RUCDR), USA. RNA was extracted using Qiagen Universal liquid handling system (PAXgene extraction kits following the manufacturer's protocol). Total RNA was measured by spectroscopy (Trinean DropSense) to determine purity and concentration while RNA fidelity was

measured by the Agilent Bioanalyzer analysis. RNA samples were hybridized to Affymetrix U219 array plates (GeneTitan), which contain 530,467 probes, each 25 bases in length. Array hybridization, washing, staining, and scanning were carried out in an Affymetrix GeneTitan System following the manufacturer's protocol. Non-uniquely mapping probes (hg19) and probes containing a polymorphic SNP based on snp137 (UCSC) were removed. Expression values were obtained using RMA normalization implemented in Affymetrix Power Tools (APT, v 1.12.0). Probes targeting *IL6R* (Supplementary Table S4) were selected for analysis.

Statistical analysis

The heritability of sIL-6R level in extended twin families

To estimate the broad- and narrow-sense heritability of sIL-6R and to examine the contribution of rs2228145, genetic structural equation models were fitted in Mx[34] to sIL-6R data from mono- and dizygotic twins, siblings and parents. The model included additive genetic influences (A), non-additive genetic influences (D), sibling-shared environmental influences (C) and unique environment (E).These methods are outlined in full in the Supplementary Methods.

Variance explained by chromosome-wide SNPs and SNPs in the IL6R region using GCTA

The variance in sIL-6R level explained by all SNPs was estimated in GCTA (Genome-wide Complex Trait Analysis [35, 36]) separately for each chromosome. For a full description of the methods, see Supplementary Methods.

GWA analysis

GWA analysis was performed in PLINK [37] on dosage data. PLINK accounts for familial relations by performing a stratified association analysis with clusters based on family id using the --family option. This option implements generalized estimating equations with an independence model [38] and robust standard errors obtained using the sandwich correction for the family clustering [39]. The analyses included one randomly selected twin of each MZ pair.

Biometrical model

The variance due to additive effects (V_A) and dominance effects (V_D) of rs2228145 were estimated by applying the biometrical model [40] to the allele frequency estimates and the mean sIL-6R level corresponding to each genotype group. According to biometrical model, V_A and V_D are calculated as follows:

$$V_A = 2pq\,[a + d\,(q - p)]^2$$

$$V_D = (2pqd)^2$$

Where p=frequency of allele 1, q=frequency of allele 2, a= genotypic value; half the distance between the mean phenotype level of the two homozygotes, and

d=dominance deviation; deviation of the mean phenotype level of heterozygotes from the midpoint of the two homozygotes.
Mean sIL-6R level per rs2228145 genotype, corrected for age and sex, were obtained in SPSS version 19. Allele frequency estimates were obtained with Sib-pair ("http://genepi.qimr.edu.au/staff/davidD/#sib-pair") using the best linear unbiased estimator (BLUE) option, which accounts for familial relatedness [41].

Heritability explained by rs2228145

To assess the contribution of rs2228145 to the heritability of sIL-6R level, models were fitted to the sIL-6R data from twin families with and without adjustment of sIL-6R levels for rs2228145 genotype. In the model with adjustment for rs2228145, sIL-6R level was modeled as follows:

$$\text{sIL-6R level} = \alpha + \beta_{age} * \text{age} + \beta_{sex} * \text{sex} + \beta_{rs2228145} * \text{rs2228145 genotype} + \varepsilon$$

Where α=intercept, age= age at blood sampling (z-score), sex= coded as 0 for males and 1 for females, $\beta_{rs2228145}$ = additive effect of rs2228145, rs2228145 genotype = observed genotype at rs2228145 (coded as 0, 1, 2 – corresponding to the number of minor alleles) and ε=residual.

When rs2228145 genotype is not accounted for, the effect of this SNP on the resemblance of sIL-6R level among family members is included in the total genetic influences (A and D, see Supplementary Methods). When the effect of rs2228145 is incorporated (by estimating $\beta_{rs2228145}$), sIL-6R levels are adjusted for rs2228145 genotype and the variance of residual levels (ε) is partitioned into unmeasured genetic and environmental factors (A, C, D and E). Thus, the total broad heritability (H^2) of sIL-6R can be written as:

$$H^2(\text{sIL-6R}) = a^2 + d^2 = a^2_{rs2228145} + a^2_{residual} + d^2_{rs2228145} + d^2_{residual}$$

Where a^2 and d^2 are the proportions of the variation in sIL-6R level due to total additive genetic and non-additive genetic effects, as estimated from the twin-family data, $a^2_{residual}$ and $d^2_{residual}$ represent all remaining unmeasured additive and non-additive genetic effects that are not captured by rs2228145 (expressed as a proportion of the total variation in sIL-6R level): these components were estimated in a model that included adjustment of sIL-6R levels for SNP effects. $a^2_{rs2228145}$ and $d^2_{rs2228145}$ are the proportion of the total variation in sIL-6R that can be explained by additive and non-additive genetic effects of rs2228145, which can be inferred from the difference between the variance estimates from the total heritability model (without correction for rs2228145) and the variance estimates from the model with adjustment for rs2228145.

Combined linkage and association analysis

Analysis of linkage while simultaneously modeling association, as suggested by Fulker *et al* [42] was conducted in QTDT [43]. The data came from nuclear families of which both parents and offspring had data on genome-wide SNPs and sIL-6R level, where the offspring consisted of DZ twins or non-twin sib-pairs, or a single MZ twin + sibling(s). The analysis was performed for all imputed SNPs with a MAF > 0.2 in the *IL6R* gene +/- 10 MB (*IL6R* gene location build 37/hg19, chr1: 154377669 - 154441926), leading to a selection of 13751 SNPs (chr1: 144377669-164441926). IBD probabilities were estimated in Merlin [44] using multipoint estimation, which takes into account the correlated structure of markers and is therefore suited for dense SNP data [45]. To create a centiMorgan (cM) map of the region, cM distances between SNPs were inferred from base-pair distances following the assumption that a distance of 1 million basepairs between SNPs corresponds to a distance of 1cM. The analysis was performed on residual sIL-6R levels after taking out the effects of all covariates (Supplementary Methods). First, the evidence for linkage was evaluated and next, these results were compared to the results obtained when modeling linkage and association simultaneously (using the –at option to specify the total association model, which is not a TDT test). The test for linkage while simultaneously modeling association involves comparing

H0: sIL-6R = μ + βSNP and Variance (sIL-6R) = $V_E + V_G$

against H1: sIL-6R = μ + βSNP and Variance (sIL-6R) = $V_E + V_G + V_A$;

Where μ=intercept (corrected for covariates), V_E=Non-shared environmental variance, V_G=Additive polygenic variance, which is estimated from the phenotypic covariance of relatives following the assumption that on average 50% of V_G is shared among first degree relatives. V_A=Additive major gene effect, which represents the additive effect of linkage to a major gene and is based on the pi-hat measure derived from the IBD matrix of relatives. H0=null-hypothesis. H1=Alternative hypothesis.

eQTL analysis

Inverse quantile normal transformation was applied to the individual probe data to obtain normal distributions. For each SNP-probe combination, a linear mixed model was fitted with expression level as dependent variable, and with fixed effects: genotype, sex, age, body mass index, smoking status, technical covariates (covering e.g. plate and well differences [33]), three principle components (PCs) from the genotype data (Supplementary Methods) and five PCs from the transformed expression data. Random effects included family ID and zygosity to account for family and twin relations [46]. Cis-eQTLs are expression-associated SNPs with a distance < 1Mb to the gene, and trans-eQTLs are the complementary set of SNPs. In an initial analysis of genome-wide SNPs, no trans-eQTLs were observed. The cis-eQTL analysis yielded 36 (N probes)*2731(N SNPs) = 98316 tests. To correct for multiple testing, a conservative *P* value threshold of 0.01/98316 ~ 1×10^{-7} was applied. The

conditional eQTL analysis was performed using the same model and *P* value threshold on expression data that had been residualized for the effect of rs7512646 in advance. To examine the relationship between sIL-6R level and *IL6R* expression level, the Pearson correlation between sIL-6R level and expression level was computed for all probes.

Table 1: Familial correlations of sIL-6R level

	Complete pairs	Correlations, no correction for SNP effects[d]		Correlations, correction for rs2228145[e]	
	N	r	95% CI	r	95% CI
Monozygotic twins					
MZ male twins	208	0.805	(0.7607 - 0.8388)	0.560	(0.4530 - 0.6416)
MZ female twins	520	0.684	(0.6421 - 0.7207)	0.358	(0.2852 - 0.4248)
Male siblings					
DZ male twins	93	0.407	(0.2083 - 0.5566)	0.249	(0.0258 - 0.4324)
Male sibs[a]	126	0.243	(0.0882 - 0.3859)	0.271	(0.1033 - 0.4129)
Female siblings					
DZ female twins	216	0.373	(0.2627 - 0.4688)	0.258	(0.1445 - 0.3612)
Female sibs[b]	313	0.298	(0.1784 - 0.4018)	0.121	(0.0116 - 0.2250)
Opposite-sex siblings					
DZ opposite-sex	224	0.317	(0.1884 - 0.4282)	0.153	(0.0191 - 0.2782)
Opposite-sex sibs[c]	337	0.428	(0.3276 - 0.5121)	0.208	(0.0748 - 0.3250)
Parent-offspring					
Mother-daughter	342	0.335	(0.2543 - 0.4076)	0.180	(0.0975 - 0.2581)
Mother-son	204	0.438	(0.3434 - 0.5174)	0.292	(0.1700 - 0.3969)
Father-son	171	0.360	(0.2626 - 0.4447)	0.202	(0.0708 - 0.3190)
Father-daughter	314	0.316	(0.2311 - 0.3915)	0.188	(0.0928 - 0.2748)
Spouses					
Mother-Father	374	0.118	(0.0269 - 0.2075)	0.247	(0.1563 - 0.3317)

[a]Non-twin brother-brother pairs and pairs of brother-male twin, [b]Sister-sister and Sister-female twin, [c] Sister-brother, brother-female twin and sister-male twin. [d]Correlation from a saturated model with covariates age and sex. [e]Correlation from a saturated model with covariates age, sex and rs2228145 genotype. CI =Confidence interval. Note that the correlations of sIL-6R levels adjusted for the effect of rs2228145 provide information about the proportions of residual variance due to genetic and environmental effects. The proportion of the total phenotypic variance that is due to genetic effects beyond rs2228145 is reflected in the difference between the correlations with and without adjustment for rs2228145 genotype. MZ= monozygotic, DZ= dizygotic.

Results

The heritability of sIL-6R level in extended twin families

To examine the overall contribution of genetic and environmental influences to the variation in sIL-6R level, genetic structural equation modeling was performed on sIL-6R level data of 4980 subjects from 2360 twin families. This approach allowed us to estimate the variance due to total heritable genetic effects (broad-sense heritability) and additive genetic effects (narrow-sense heritability). Based on the pattern of phenotypic correlations for sIL-6R level among monozygotic (MZ) twins, dizygotic (DZ) twins, siblings, and parent-offspring pairs (Table 1), the broad-sense heritability of sIL-6R level was estimated at 72% ($H^2=(V_A+V_D)/V_{total}=(0.89+0.08)/1.35$) and the narrow-sense heritability was estimated at 66% ($h^2= V_A/V_{total}=0.89/1.35$). The remaining variance (28%) was ascribed to environmental factors not shared among family members (unique environment: $e^2=V_E/V_{total}=0.38/1.35$).

Table 2: Variance of sIL-6R level explained by chromosome-wide SNPs estimated using GCTA.

		Unrelated subjects (N=2875)		Unrelated and Related subjects (N=4846)	
Chromosome	**N SNPs**	**V_G/V_P**	**SE**	**V_G/V_P**	**SE**
1 *IL6R* region	42268	0.547	0.025	0.533	0.019
1 Rest	584291	0.003	0.025	0.000	0.015
2	692964	0.056	0.028	0.028	0.016
3	590258	0.000	0.024	0.009	0.014
4	605730	0.000	0.023	0.002	0.013
5	541738	0.000	0.023	0.005	0.013
6	530265	0.017	0.021	0.030	0.014
7	483988	0.000	0.021	0.009	0.013
8	447709	0.001	0.021	0.010	0.013
9	350386	0.000	0.021	0.000	0.012
10	418731	0.010	0.021	0.017	0.013
11	410817	0.000	0.014	0.005	0.009
12	393609	0.001	0.019	0.010	0.011
13	307137	0.003	0.017	0.007	0.011
14	265369	0.000	0.017	0.000	0.010
15	227463	0.013	0.017	0.000	0.009
16	242058	0.006	0.018	0.002	0.010
17	202299	0.014	0.017	0.000	0.010
18	234349	0.002	0.018	0.000	0.010
19	158942	0.003	0.014	0.001	0.008
20	180019	0.016	0.016	0.007	0.010
21	109887	0.000	0.013	0.000	0.007
22	105190	0.003	0.012	0.010	0.008

SE=standard error

Variance explained by chromosome-wide SNPs and SNPs in the IL6R region using GCTA

We next used GCTA on 1000Genomes-imputed SNP data to examine how much of the variance in sIL-6R level can be explained by all SNPs in the *IL6R* region (*IL6R* +/- 10MB), how much by other SNPs on chromosome 1 and how much by each of the other chromosomes (Table 2, Supplementary Figure S1) based on the similarity across SNPs among 2875 unrelated subjects. This method gives insight into the total contribution of additive genetic effects tagged by all genotyped and imputed SNPs together, providing an estimate of the total variance that could be identified by GWAS on this set of SNPs, given sufficient power to detect individual SNP effects. SNPs in the *IL6R* region with a MAF > 0.001 (*N* SNPs = 42268) together explained 54.7 % (SE= 2.5%) of the variance of sIL-6R levels, while the rest of chromosome 1 did not contribute to the variance. Some additional variance was captured by SNPs on chromosome 2 (5.6%, SE=2.8). When the analysis was repeated with inclusion of related subjects (including siblings, DZ twins and parent-offspring pairs, total *N* subjects=4846) the estimate for chromosome 2 was somewhat lower (2.8%, SE= 1.6), whereas otherwise highly similar results were obtained (Table 2).

GWA analysis

GWA analysis of sIL-6R level was conducted on imputed SNP data from 4846 subjects. The genomic inflation factor indicated no effect of population stratification (λ= 1.015). 680 Genome-wide significant hits were found ($P < 5x10^{-8}$), which were all located on chromosome 1q21.3 (significant hits; chr 1: 153389207-154697624), except for four SNPs (MAF 0.01 – 0.09) on chromosomes 5, 8 and 20 (Supplementary Table S2). The top SNPs ($P < 1 \times 10^{-176}$) were located within *IL6R* and were all in LD with rs2228145.

Biometrical model

The minor (C) allele of rs2228145 occurred at a frequency of 0.39. To estimate how much of the variance in sIL-6R level is explained by rs2221845, we applied the classic biometrical model to the data. The average sIL-6R level was 5.698 (10^{-8} g/mL) in individuals homozygous for the minor allele (CC), 4.418 in heterozygotes (AC), and 3.238 in individuals with the AA genotype, giving an overall mean level of 4.17. The observed variance of sIL-6R levels in the total sample was 1.35. Applying the biometrical model to these data, it follows that:

$a = 0.5*(sIL\text{-}6R_{CC} - sIL\text{-}6R_{AA}) = 0.5*(5.698-3.238) = 1.23$

$d = sIL\text{-}6R_{AC} - (sIL\text{-}6R_{AA} + a) = 4.418 - (3.238 + 1.23) = -0.05$

$V_A = 2pq\,[a + d\,(q-p)]^2 = 2*0.39*0.61*[1.23-0.05*(0.61-0.39)]^2 = 0.71$

$V_D = (2pqd)^2 = (2*0.39*0.61*-0.05)^2 = 5.66 \times 10\text{-}4$

Thus, the allelic effect was almost completely additive, which illustrates that the average sIL-6R level of heterozygous individuals lies perfectly in the middle of the levels of the two homozygous groups (Figure 1A).

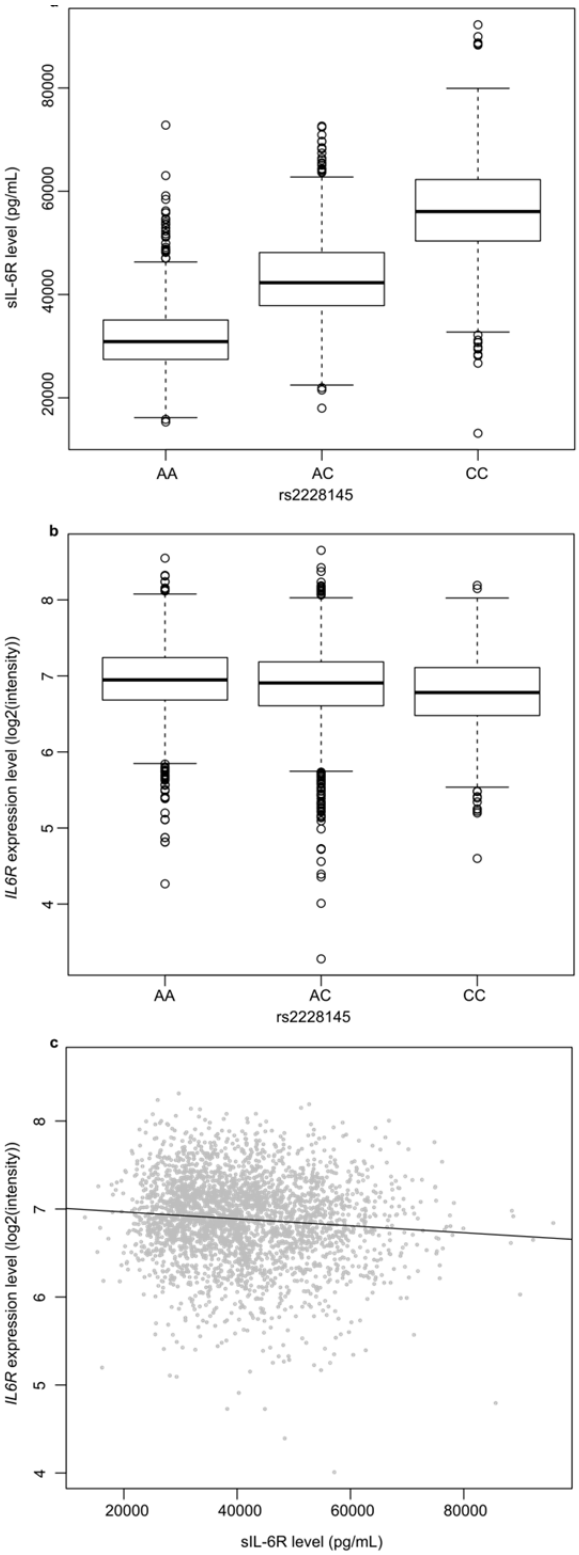


Figure 1. Associations between sIL-6R level, *IL6R* expression and rs2228145 genotype. A: Boxplots of plasma sIL-6R level (pg/mL) for each rs2228145 genotype. B: Boxplots of *IL6R* expression level in blood as measured by expression probe 582_132 for each rs2228145 genotype (P = 3.14 x 10^{-19}). C: Scatterplot of sIL-6R level (pg/mL) versus *IL6R* expression level for probe 582_132 (r = -0.092, P = 1.42 x 10^{-6}).

Heritability explained by rs2228145

The data from twin families allowed us to examine how much of the heritability is explained by rs2228145, and how much genetic variance is left after adjusting sIL-6R levels for rs2228145. The contribution of rs2228145 to the heritability of sIL-6R level is illustrated by the drop of the correlations of sIL-6R level among family members after correcting sIL-6R level for rs2228145 genotype (Table 1). When the allelic effect of rs2228145 on sIL-6R level was

taken into account in the twin family model (Table 3), the total variance of sIL-6R level dropped from 1.35 to 0.66 (difference= 0.69), illustrating that this SNP accounted for 51% of the total variance in sIL-6R level (H^2_{SNP}= 0.69/1.35=0.51). Residual genetic effects not tagged by rs2228145 ($V_{G\text{-}residual}$=0.26) accounted for 19% of the total variance of sIL-6R ($H^2_{residual}$= 0.26/1.35=0.19).

The contribution of other additive SNP effects in the IL6R region
When we repeated the analysis of unrelated subjects in GCTA while correcting for the effect of rs2228145, SNPs in the *IL6R* region on chromosome 1 still explained 5.6% of the total variance of sIL-6R level (SE=1.8%, $P = 6.0 \times 10^{-5}$). This suggests that part of the residual heritability that was estimated based on twin family analysis can be explained by other variation in the *IL6R* region that is tagged by genotyped and imputed SNPs.

Combined linkage and association analysis
In the GWA analysis of sIL-6R levels, highly significant hits were found across the *IL6R* region, which may include SNPs that merely capture the effect of rs2228145 (through LD) and SNPs tagging other causal variants. To search for genetic variants that explain additional variation in sIL-6R levels beyond the effect of rs2228145, we simultaneously modeled linkage and association using data from 355 nuclear families (total *N* subjects = 1254) for SNPs in the *IL6R* region (*IL6R* +/- 10MB, Figure 2a). This region covers all genome-wide significant GWA hits on chromosome 1. If the linkage test is not significant while simultaneously modeling association for a SNP (while the linkage test is significant when association is not modeled simultaneously), this indicates that the respective SNP is in high LD with a causal variant. On the other hand, the linkage signal will not be fully impaired when association is modeled for SNPs that are in lower LD with a causal variant, or when multiple causal variants in partial LD contribute to variation in sIL-6R levels.

In a model that included linkage (V_A, additive variance attributable to the locus estimated using IBD), additive polygenic variance (V_G, estimated from the familial resemblance in sIL-6R following the assumption that on average 50% of V_G is shared among first degree relatives) and unique environment (V_E), but not incorporating association, the linkage signal around rs2228145 explained 69% (V_A/V_{total}) of the total variation in sIL-6R levels (X^2=34.12, df=1, $P = 5 \times 10^{-9}$) and V_G was estimated at 0 (Table 3). Comparison of the estimate of variance explained by linkage at *IL6R* (69%) to the broad-sense heritability (72%) and narrow-sense heritability (66%) estimated by the twin family model suggests that the entire narrow-sense heritability of sIL-6R level and nearly the entire broad-sense heritability of sIL6-R level can be captured by modeling the covariance of sIL6-R level among relatives as a function of IBD-sharing at the *IL6R* locus.

Table 3: Variance of sIL-6R level due to rs2228145, total heritability of sIL-6R level, and estimates from linkage analysis

Observed variance due to rs2228145 genotype

Geno-type	Frequen-cy[a]	sIL-6R level (10^{-8} g/mL)[b]	Geno-typic Value	Frequency* sIL-6R	(deviation from μ)^2	Frequency* squared deviation
CC	p^2= 0.15	5.698	+ a	0.15*5.698 = 0.87	(5.698- μ)^2 = 2.32	0.15*2.32 = 0.35
AC	2pq=0.48	4.418	d	0.48*4.418 = 2.10	(4.418- μ)^2 = 0.06	0.48*0.06 = 0.03
AA	q^2= 0.37	3.238	- a	0.37*3.238 = 1.20	(3.238- μ)^2 = 0.88	0.37*0.88 = 0.33
Total			μ=0.87+2.10+1.20 = 4.17		V_{SNP}=0.35+0.03+0.33 =0.71	
						(95% CI: 0.68-0.74)

Extended twin family model estimates

	V_{sIL-6R}	V_A	V_D	V_C	V_E	V/V_{total}	h^2	d^2	c^2	e^2
ACDE model[c]	V_{total} = 1.35	0.89	0.08	0.00	0.38	1.00	0.66	0.06	0.00	0.28
ACDE + rs2228145[d]	$V_{residual}$ = 0.66	0.24	0.02	0.01	0.39	$V_{residual}/V_{total}$=0.49	0.18	0.01	0.01	0.29
Effect of rs2228145[e]	V_{SNP} = 0.69	0.65	0.06	-0.01	-0.01	V_{SNP}/V_{total} = 0.51	0.48	0.05	0.01	0.01

Linkage analysis estimates

	V_{sIL-6R}	$V_{A-linkage}$	$V_{A-polygenic}$	V_D	V_C	V_E	V/V_{total}	h^2	d^2	c^2	e^2
Linkage modeled at rs2228145	V_{total} =1.39	0.96	0	-	-	0.43	1.00	0.69	-	-	0.31
Linkage+ association modeled at rs2228145	$V_{residual}$ =0.67	0.26	0	-	-	0.41	$V_{residual}/V_{total}$ =0.48	0.19	-	-	0.30

[a] p=minor allele frequency=0.39 and q=major allele frequency=0.61

[b] Mean sIL-6R level, corrected for age and sex, for each genotype group.
μ= Mean sIL-6R level in the population, estimated from the genotype frequencies and corresponding sIL-6R levels for each genotype group.

[c] ACDE model without SNP effects, in which $\mathbf{V_{total}}$ is decomposed into $\mathbf{V_A}$, $\mathbf{V_D}$, $\mathbf{V_C}$, and $\mathbf{V_E}$.

[d] Model in which the additive effect of rs2228145 on sIL-6R level is modeled, and the residual variance of sIL-6R ($\mathbf{V_{residual} = V_{total} - V_{SNP}}$) is decomposed into $\mathbf{V_A}$, $\mathbf{V_C}$, $\mathbf{V_D}$ and $\mathbf{V_E}$.

[e] The effect of rs2228145 was inferred from the difference between model [c] and model [d]

$\mathbf{V_{total}}$= Total (phenotypic) variance of age- and sex-adjusted sIL-6R levels. $\mathbf{V_{SNP}}$ = Variance of sIL-6R level attributable to rs2228145, $\mathbf{V_{SNP}/V_{total}}$ = Proportion of total (phenotypic) variance sIL-6R level explained by rs2228145, $\mathbf{V_{residual}/V_{total}}$= Proportion of total (phenotypic) variance sIL-6R level not explained by rs2228145, $\mathbf{V_A}$=Additive genetic variance, $\mathbf{V_D}$=Non-additive genetic variance, $\mathbf{V_C}$=Sibling-shared environmental variance, $\mathbf{V_E}$=Unique environmental variance, $\mathbf{h^2}$= V_A/ V_{total} , $\mathbf{d^2}$= V_D/ V_{total}, $\mathbf{c^2}$= V_C/ V_{total}, $\mathbf{e^2}$= V_E/ V_{total}.

$\mathbf{V_{A-linkage}}$ =Variance due to linkage, which is based on the pi-hat measure at rs2228145, derived from the IBD matrix of relatives

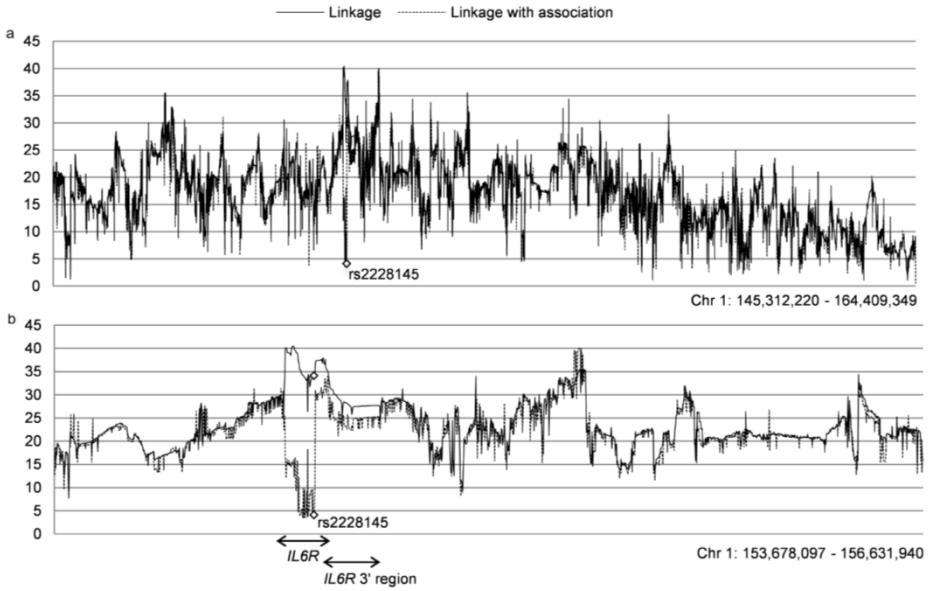
$\mathbf{V_{A-polygenic}}$ = Variance due to all additive genetic effects that are not captured within the linkage component (i.e. additive genetic variance that is not linked to the *IL6R* locus), which was estimated from the phenotypic covariance of relatives following the assumption that on average 50% of $V_{A-polygenic}$ is shared among first degree relatives.

When the allelic effects of SNPs on the means were incorporated in the model (i.e., association was added), 19% of the total variance in sIL-6R level (V_A/V_{total}) was still explained by the linkage component when modeling association at rs2228145 (similar to the estimate of residual heritability from the twin family model), although this signal was borderline significant only (X^2=4.13, df=1, P = 0.042). Significant linkage was initially observed across the entire region, but when association was modeled, linkage only dropped to the level of borderline significance for SNPs in high LD with rs2228145 (r^2 > 0.93 among 21 SNPs with P > 0.01) indicating that rs2228145 explains most, though not the entire linkage signal. The linkage signal was also attenuated when modeling association for SNPs within an LD block at the 3' end of *IL6R* containing the *IL6R* 3'UTR and the adjacent genes *SHE* and *TDRD10* (location indicated in Figure 2b). This region contains both SNPs in low LD (r^2 ~0.1) with rs2228145 and SNPs in high LD with rs2228145. When we conducted a conditional GWAS to test for association between chromosome 1 SNPs and sIL-6R levels after taking out the effect of rs2228145 (by including this SNP as a covariate in the model), significant associations were still observed for SNPs within *IL6R* including SNPs at the 3'end (Supplementary Table S3). These results support the presence of additional causal variants influencing sIL-6R level in the *IL6R* gene and 3'end.

eQTL analysis

To identify genetic variants related to *IL6R* RNA expression levels, eQTL analysis was performed on data from 4467 subjects for 36 probes (Supplementary Table S4) measuring *IL6R* RNA. A total of 341 significant associations (Supplementary Table S5) were found for 5 of these probes (4 probes targeting exon 7 and 1 probe targeting the 3'end). These associations involved 179 SNPs that were all in *cis* (SNP positions; chr1q21.3: 154395125-154521584). The most significant association was between rs7512646 and expression probe 582_132 targeting exon 7 (P = 2.84 x 10^{-22}). This SNP (intronic) is in high LD with rs2228145 (r^2 =0.94) and rs2228145 itself showed significant associations with all four probes targeting exon 7. Probe 308_15 targeting the 3' end was the only probe outside exon 7 with significant eQTLs (total N associated SNPs= 92, top hit= rs4072391 located in the *IL6R* 3'UTR, P = 1.32 x 10^{-11}, r^2 with rs2228145=0.142). Of the 680 significant hits identified in the GWAS of sIL-6R, 157 SNPs were significantly associated with *IL6R* expression level. For all of these SNPs, the allele associated with higher sIL-6R levels was associated with lower *IL6R* expression level.

Figure 2. Results from combined linkage and association analysis of the *IL6R* region on chromosome 1. The solid black line connects the X^2 values from the linkage test at each SNP without modeling SNP association and the dotted black line connects the X^2 values from the linkage test at each SNP when association is simultaneously modeled. The analysis was performed for all imputed SNPs with a MAF > 0.2 in *IL6R* +/-10MB, chr1 (build 37): 145312220-164409349 (x-axis, Figure A). Figure B zooms in to the region from basepair position 153678097 to 156631940. The arrows mark the location of the *IL6R* gene, and the location of an LD block stretching across the 3'UTR of *IL6R* and adjacent genes SHE and TDRD10, where significant hits were found in the eQTL analysis for SNPs in partial LD with rs2228145 (r^2 ~0.1).



Relationship between rs2228145, IL6R RNA and sIL-6R levels

As a measure of the relationship between plasma sIL-6R levels and the abundance of *IL6R* RNA in whole blood, the correlation between sIL-6R level and expression level was computed for all probes targeting *IL6R* transcripts (Supplementary Table S6). All five expression probes with significant eQTL hits showed small but statistically significant negative correlations with sIL-6R level (e.g. probe 582_132: r = -0.092, $P = 1.42 \times 10^{-6}$, Table 4). The negative correlations indicate that across individuals, higher sIL-6R level was associated with lower overall *IL6R* expression (Figure 1c). Correlations between *IL6R* RNA and sIL-6R levels became non-significant when expression levels were corrected for the effect of rs2228145 (e.g., probe 582_132: r = 0.019, P = 0.314, Table 4), which indicates that at the population level, the relationship between *IL6R* expression and sIL-6R level can be explained by the effect of

rs2228145. Indeed, the minor C allele of rs2228145 was associated with higher sIL-6R level and with lower *IL6R* expression level (Figure 1). When computed separately within rs2228145 genotype classes, correlations between IL6R expression and sIL-6R levels were not significant (Table 4).

Taken our eQTL design, we could not distinguish between alternative *IL6R* transcripts but observed a significant negative association between the rs2228145 C allele and total levels of *IL6R* RNA. In another eQTL analysis based on expression in peripheral blood (N=1469)[47], one of the probesets targeted *IL6R* exon 9, which is spliced out from the alternative RNA transcript that is presumed to directly code for sIL-6R. For this probeset, a negative relation between the minor allele of rs4845623 (r^2 with rs2228145=0.93) and expression level was observed ($P < 1 \times 10^{-7}$). We also downloaded the exon-specific expression data from HapMap lymphoblastoid cell lines (GEO accession nr = GSE9703, *N*=162) and corresponding genotype data[48]. In this dataset, no significant association between rs2228145 and exon 9 expression level was found (possibly due to the small sample size). When we computed the ratio of exon 9 expression level over mean *IL6R* transcript expression, we found that the proportion of *IL6R* transcripts containing exon 9 decreased with each copy of the rs2228145 C allele ($P < 0.01$, Supplementary Figure S2), suggesting that rs2228145 is associated with the ratio of normal and alternatively spliced transcripts. Note that if the rs2228145 C allele has differential effects on alternative RNA transcripts, a negative association with total RNA levels (as assessed in our own study) will emerge if the increasing effect on levels of the alternative splice variant [10, 18] is smaller compared to the decreasing effect on other transcripts.

Conditional eQTL analysis

When the eQTL analysis was conducted after adjusting expression levels for the effect of the most significant eQTL SNP from the initial analysis (rs7512646), significant associations (Supplementary Table S7) were observed between probe 308_15 (targeting the 3' end of *IL6R* transcripts) and 80 SNPs located in an LD block covering the 3'UTR of *IL6R* and the adjacent 3'region. This region was also highlighted by the analysis of linkage and association of sIL-6R levels (location indicated in Figure 2b). The most significantly associated conditional eQTL SNPs with a *P* value of 1.89×10^{-9} were rs60760897, rs60255122, rs61698846 and rs61275241 (Located in *TDRD10*, r^2 with rs2228145=0.133, MAF=0.21), followed by rs4072391 ($P = 2.10 \times 10^{-9}$, located in the *IL6R* 3'UTR, r^2 with rs2228145=0.142; these SNPs are in high LD with each other and for each of the SNPs, the minor allele was associated with higher expression level and with lower sIL-6R level (*P* value (GWAS) for rs60760897= 1.49 x 10-^{121}, and *P* value (conditional GWAS) for rs60760897= 8.04×10^{-4}). All the initially observed associations for exon 7 expression probes were no longer significant after correcting for the most significant eQTL, suggesting that genetic variation tagged by rs2228145 and variation tagged by

rs60760897 (*IL6R* 3'region) are both independently associated with *IL6R* expression level.

Table 4: Correlation between sIL-6R level and *IL6R* expression level for all probes with significant eQTL hits.

IL6R probe	Probe location		Overall correlation (N=2727)	Correction for rs2228145 (N=2727)	Genotype AA (N=991)	Genotype AC (N=1278)	Genotype CC (N=458)
323_134	Exon 7	r	-0.086	0.001	0.032	0.015	-0.023
		P	7.48×10^{-6}	0.955	0.312	0.601	0.618
201_497	Exon 7	r	-.054	0.011	0.047	0.009	0.002
		P	0.005	0.577	0.136	0.755	0.974
582_132	Exon 7	r	-0.092	0.019	0.024	0.005	-0.063
		P	1.42×10^{-6}	0.314	0.449	0.857	0.178
202_497	Exon 7	r	-0.071	0.005	0.072	-0.016	-0.017
		P	1.94×10^{-4}	0.791	0.023	0.569	0.724
308_15	3'-end	r	-0.055	-0.003	0.031	-0.043	-0.088
		P	0.004	0.894	0.335	0.124	0.059

r= Pearson correlation between sIL-6R level (corrected for age and sex) and expression level (corrected for technical covariates, age and sex). *P*= P-value. Colum 5 only is based on expression levels that were additionally corrected for the additive allelic effect of rs2228145. Columns 6 to 8 show the correlations between sIL-6R level and expression level computed separately within rs2228145 genotype classes. Correlations for all *IL6R* expression probes are given in Supplementary Table S6.

Discussion

Based on the analysis of three different types of familial relations in extended twin families (monozygotic twins, sibling pairs/dizygotic twins, and parent-offspring pairs), the broad heritability of sIL-6R level was estimated at 72% and the narrow-sense heritability was estimated at 66%. Linkage analysis closely recaptured this, with 69% of the variance in sIL-6R levels explained by all variation in the *IL6R* region that is captured by IBD. Both estimates are similar to the heritability of sIL-6R levels reported by a previous study conducted in middle-aged male twins (h^2=0.68) [49]. The estimate of variance explained by additive SNP effects in the *IL6R* region in unrelated subjects (GCTA) was slightly lower at 54.7%, which suggests that the heritability of sIL-6R cannot be fully ascribed to additive effects of currently genotyped and imputed SNPs. Whereas the variance explained by all SNPs in unrelated subjects (in which LD extends over relatively short distances) specifically provide information about

effects tagged by a given set of SNPs under an additive allelic model, linkage and twin analyses, which are based on IBD-sharing among close relatives, capture more variation contributing to the similarity of family members. Estimates from linkage may include variation in a region that is not (fully) captured by additive SNP effects, such as rare variants segregating in families, structural variation and effects of interacting loci (epistasis). Overall, the three different methods show the expected convergence.

After establishing the heritability, we examined the contribution of the functional *IL6R* polymorphism rs2228145, a known candidate SNP for sIL-6R [19-22, 50], to the population variance in plasma sIL-6R levels. Using the classic biometrical model and twin family analysis, we showed that rs2228145 explained 51% of the total variance of sIL-6R level, corresponding to 71% of the broad heritability (51/72=0.71). The estimate of variance explained by this SNP in our sample is comparable to an estimate previously reported for subjects of European descent (49% explained by rs2228145 [50]), and is larger compared to several other previous reports (33% explained by rs2228145 in African Americans [50], 29% explained by rs2228145 [18], 30% explained by rs4129267; a SNP in LD [24], and 20% explained by rs4537545; another SNP in LD [22]). The variation between studies might be related to differences in the lab protocol (e.g. differences related to sIL-6R detection assay and dilution of samples) and to differences between study populations. Whereas our study and the studies by Reich et al[50] and Rafiq et al[22] were conducted on population-based cohorts, the studies by Ferreira et al [18] and Revez et al[24] included patient populations. In contrast to all other studies, we excluded individuals using anti-inflammatory medication, which could have led to a slightly healthier population compared to previous studies, and it could be hypothesized that the variance explained by rs2228145 may vary with health status, as the levels of sIL-6R may rise 2-3 fold within a person during inflammation [51]. Although the allele frequencies of rs2228145 differ between individuals of European versus African descent, variation related to ancestry only seems explain the different estimates observed in the study by Reich et al [50], as all other studies focused on individuals of European descent and reported similar allele frequencies for rs2228145.

Functional studies have indicated that amino-acid mutations at the position encoded by rs2228145 can influence the production of sIL-6R through shedding of membrane-bound receptors [17]. The effects of rs2228145 and other SNPs on *IL6R* expression are less well characterized. To gain insight into the regulatory impact of genetic variation on *IL6R* expression, we studied the relation between *IL6R* expression level and genome-wide SNPs. We found two clusters of SNPs within the *IL6R* region that were associated with expression level, one of which included rs2228145, suggesting that this SNP also influences *IL6R* expression, though the signal may also come from another causal variant in LD. A previous analysis of gene expression in multiple samples and tissues revealed no significant association between rs2228145

genotype and overall expression of *IL6R* RNA transcripts (possibly due to limited sample size)[23]. Two other studies found that the rs2228145 C allele was associated with higher levels of an alternative *IL6R* mRNA splice variant [10, 18], which lacks a 94-bp sequence encoding part of the trans-membrane domain that anchors the membrane-bound IL-6R to the cell membrane and is presumed to directly code for the sIL-6R. In our study and in the study by Revez et al [24], the rs2228145 C allele was associated with lower overall *IL6R* RNA level and our analysis of the HAPMAP expression data suggested that rs2228145 (or a variant in LD) has an effect on the relative abundance of alternative *IL6R* RNA transcripts.

Though the opposite effects of rs2228145 on sIL-6R levels and overall *IL6R* expression level may appear contradictory, this finding is not unexpected given that rs2228145 has a strong effect on alternative splicing. As the C allele is associated with increased splicing of exon 9 [10], this allele is likely to decrease the level of the full-length RNA transcript. It can also be hypothesized that rs2228145 may have secondary effects on *IL6R* expression by impacting on feedback systems that control *IL6R* expression. In hepatocytes, a positive feedback circuit has been identified through which increased membrane-bound receptor-mediated classic IL-6 signaling triggers microRNA (miRNA)-mediated regulatory pathways that stimulate increased *IL6R* expression[52]. The rs2228145 C allele has been shown to impair classical IL-6 signaling, as indicated by reduced levels of mIL-6R and decreased IL-6 responsiveness of CD4+ T cells and monocytes from C allele carriers [18]. As the rs2228145 C allele weakens classical IL-6 signaling, positive feedback on *IL6R* expression may be reduced in individuals with the C allele, leading to lower RNA expression levels compared to individuals with the A allele. The associations between SNPs and variation in RNA and soluble IL-6 receptor abundance observed in our study provide guidance for functional studies into the mechanistic relationships among polymorphisms in *IL6R*, *IL6R* RNA expression and IL-6 receptor levels.

While the variance in sIL-6R levels that can be explained by rs2228145 is very large compared to single SNP effect sizes generally observed for quantitative traits, our study showed that other genetic effects also make an important contribution to the heritability of sIL-6R levels. Linkage analysis and analysis of the variance explained by chromosome wide SNP effects in unrelated subjects (GCTA) indicated that the remaining heritability of sIL-6R level appears to be primarily accounted for by other variation in the *IL6R* region on chromosome 1. Linkage analysis showed that genetic effects in this region not tagged by rs2228145 account for 19% of the total variance in sIL-6R levels. We therefore tried to identify other SNPs in the *IL6R* region that explain additional variance in this clinically important soluble cytokine receptor.

Evidence for association with sIL-6R level and *IL6R* expression was seen for SNPs located in the *IL6R* 3' region including the *IL6R* 3'UTR, which is an important region of regulatory control. Genetic variation in the 3'UTRs of

genes can affect transcript levels in several ways [53], including effects on mRNA stability, translation efficiency, and by affecting regulatory control by miRNAs. Previous studies have shown that the *IL6R* 3'UTR contains binding sites for several miRNAs and that *IL6R* transcript levels are (down)regulated by binding of these miRNAs to their 3'UTR target in *IL6R* mRNA [52, 54-56]. Alterations in components of this regulatory pathway have been reported in cancer tissues [52, 54-56] and in synovial fibroblasts from patients with rheumatoid arthritis [57]. Our novel finding that SNPs in the *IL6R* 3'region are associated with *IL6R* expression level generates novel hypotheses about the role of genetic variation in regulatory pathways of *IL6R* expression, and into the contribution of such pathways to complex disease susceptibility.

SNP rs4129267 (r^2 with rs2228145=0.97) has been identified as a risk variant for asthma; the minor allele that increases sIL-6R level is associated with increased asthma risk (OR 1.09)[25]. We looked at the effect of the top 3' eQTL hits for *IL6R* (rs60760897 and SNPs in LD (r^2 >= 0.3)) on asthma in data from the Australian Asthma Genetics Consortium [25] (AAGC, *N*=2110 cases, *N*=3857 controls) and the GABRIEL consortium [58] (*N*=10,365 cases, *N*=16,110 controls). rs60760897 was not associated with asthma risk (Tables S9 and S10) but a suggestive association signal was found for SNPs in modest LD (r^2=0.46) with rs60760897 (rs4478801; GABRIEL: OR=0.95, *P* =0.014 and AAGC: OR=0.94, *P*=0.045). The minor G allele of rs448801 that showed a trend of lower asthma risk in the AAGC and GABRIEL study was associated with lower sIL-6R level in our study (GWAS $P = 1.27 \times 10^{-145}$) and with higher *IL6R* expression (eQTL $P = 4.87 \times 10^{-8}$).

Compared to previous reports that common SNPs together generally explain less than 50% of the total variation of complex traits [36], it seems remarkable that a single common variant in the *IL6R* gene alone accounts for such a large part of the variation in sIL-6R levels. What makes the level of circulating sIL-6R different from other quantitative traits such as height and BMI? An important factor is that sIL-6R can be produced through two mechanisms (receptor cleavage/shedding and translation of an alternatively spliced mRNA), and that the *IL6R* SNP rs2228145 has a major impact on both mechanisms. In contrast, it seems likely, although this is actually unknown, that many complex traits result from the integration of numerous different processes that each make a small contribution to the endpoint, in which case there could be many genetic variants in different pathways that each contribute to a small portion of the total variation in the trait.

In conclusion, we have shown that sIL-6R levels are highly heritable and that variants in *IL6R* other than the well-known functional SNP make an important contribution to the heritability. Our findings shed novel light on the effects of rs2228145 and SNPs at the *IL6R* 3'end on *IL6R* expression. At the same time, we demonstrated that results from different methods based on the classic biometrical model are in agreement and converge to the same conclusion.

Reference List

1. Rose-John,S., Scheller,J., Elson,G., & Jones,S.A. Interleukin-6 biology is coordinated by membrane-bound and soluble receptors: role in inflammation and cancer. *J. Leukoc. Biol.* **80**, 227-236 (2006).
2. Mackiewicz,A., Schooltink,H., Heinrich,P.C., & Rose-John,S. Complex of soluble human IL-6-receptor/IL-6 up-regulates expression of acute-phase proteins. *J. Immunol.* **149**, 2021-2027 (1992).
3. Taga,T. *et al.* Interleukin-6 triggers the association of its receptor with a possible signal transducer, gp130. *Cell* **58**, 573-581 (1989).
4. Saito,M., Yoshida,K., Hibi,M., Taga,T., & Kishimoto,T. Molecular cloning of a murine IL-6 receptor-associated signal transducer, gp130, and its regulated expression in vivo. *J. Immunol.* **148**, 4066-4071 (1992).
5. Taga,T. & Kishimoto,T. Gp130 and the interleukin-6 family of cytokines. *Annu. Rev. Immunol.* **15**, 797-819 (1997).
6. Doganci,A. *et al.* The IL-6R alpha chain controls lung CD4+CD25+ Treg development and function during allergic airway inflammation in vivo. *J. Clin. Invest* **115**, 313-325 (2005).
7. Kotake,S. *et al.* Interleukin-6 and soluble interleukin-6 receptors in the synovial fluids from rheumatoid arthritis patients are responsible for osteoclast-like cell formation. *J. Bone Miner. Res.* **11**, 88-95 (1996).
8. Atreya,R. *et al.* Blockade of interleukin 6 trans signaling suppresses T-cell resistance against apoptosis in chronic intestinal inflammation: evidence in crohn disease and experimental colitis in vivo. *Nat. Med.* **6**, 583-588 (2000).
9. Becker,C. *et al.* TGF-beta suppresses tumor progression in colon cancer by inhibition of IL-6 trans-signaling. *Immunity.* **21**, 491-501 (2004).
10. Stephens,O.W. *et al.* An intermediate-risk multiple myeloma subgroup is defined by sIL-6r: levels synergistically increase with incidence of SNP rs2228145 and 1q21 amplification. *Blood* **119**, 503-512 (2012).
11. Hurst,S.M. *et al.* Il-6 and its soluble receptor orchestrate a temporal switch in the pattern of leukocyte recruitment seen during acute inflammation. *Immunity.* **14**, 705-714 (2001).
12. Jones,S.A., Horiuchi,S., Topley,N., Yamamoto,N., & Fuller,G.M. The soluble interleukin 6 receptor: mechanisms of production and implications in disease. *FASEB J.* **15**, 43-58 (2001).
13. Muller-Newen,G. *et al.* Purification and characterization of the soluble interleukin-6 receptor from human plasma and identification of an isoform generated through alternative splicing. *Eur. J. Biochem.* **236**, 837-842 (1996).
14. Müllberg,J. *et al.* The soluble interleukin-6 receptor is generated by shedding. *Eur. J. Immunol.* **23**, 473-480 (1993).
15. Horiuchi,S. *et al.* Soluble interleukin-6 receptors released from T cell or granulocyte/macrophage cell lines and human peripheral blood mononuclear cells are generated through an alternative splicing mechanism. *Eur. J. Immunol.* **24**, 1945-1948 (1994).
16. Lust,J.A. *et al.* Isolation of an mRNA encoding a soluble form of the human interleukin-6 receptor. *Cytokine* **4**, 96-100 (1992).
17. Müllberg,J. *et al.* The soluble human IL-6 receptor. Mutational characterization of the proteolytic cleavage site. *J. Immunol.* **152**, 4958-4968 (1994).

18. Ferreira,R.C. *et al.* Functional IL6R 358Ala Allele Impairs Classical IL-6 Receptor Signaling and Influences Risk of Diverse Inflammatory Diseases. *PLoS. Genet.* **9**, e1003444 (2013).
19. Galicia,J.C. *et al.* Polymorphisms in the IL-6 receptor (IL-6R) gene: strong evidence that serum levels of soluble IL-6R are genetically influenced. *Genes Immun.* **5**, 513-516 (2004).
20. Lourdusamy,A. *et al.* Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum. Mol. Genet.* **21**, 3719-3726 (2012).
21. Melzer,D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* **4**, e1000072 (2008).
22. Rafiq,S. *et al.* A common variant of the interleukin 6 receptor (IL-6r) gene increases IL-6r and IL-6 levels, without other inflammatory effects. *Genes Immun.* **8**, 552-559 (2007).
23. IL6R Genetics Consortium and Emerging Risk Factors Collaboration. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *Lancet* **379**, 1205-1213 (2012).
24. Revez,J.A. *et al.* A new regulatory variant in the interleukin-6 receptor gene associates with asthma risk. *Genes Immun.* **14**, 441-446 (2013).
25. Ferreira,M.A. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006-1014 (2011).
26. Eyre,S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336-1340 (2012).
27. The Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* **379**, 1214-1224 (2012).
28. Boomsma,D.I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res. Hum. Genet.* **9**, 849-857 (2006).
29. Penninx,B.W. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17**, 121-140 (2008).
30. Spijker,S., van de Leemput,J.C., Hoekstra,C., Boomsma,D.I., & Smit,A.B. Profiling gene expression in whole blood samples following an in-vitro challenge. *Twin Res.* **7**, 564-570 (2004).
31. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res. Hum. Genet.* **13**, 231-245 (2010).
32. Boomsma,D.I. *et al.* Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur. J. Hum. Genet.* **16**, 335-342 (2008).
33. Jansen,R. *et al.* Sex differences in the human peripheral blood transcriptome. *BMC. Genomics* **15**, 33 (2014).
34. Neale,M.C., Boker,S.M., Xie,G., & Maes,H.H. *Mx: Statistical Modeling.*(Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, 2006).
35. Yang,J., Lee,S.H., Goddard,M.E., & Visscher,P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).
36. Yang,J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519-525 (2011).
37. Purcell,S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

38. Dobson,A. *An introduction to generalized linear models*(Chapman & Hall/CRC, London, 2002).
39. Williams,R.L. A note on robust variance estimation for cluster-correlated data. *Biometrics* **56**, 645-646 (2000).
40. Falconer,D.S. *Introduction to quantitative genetics*(Ronald Press Co, New York, 1960).
41. McPeek,M.S., Wu,X., & Ober,C. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* **60**, 359-367 (2004).
42. Fulker,D.W., Cherny,S.S., Sham,P.C., & Hewitt,J.K. Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259-267 (1999).
43. Abecasis,G.R., Cardon,L.R., & Cookson,W.O. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279-292 (2000).
44. Abecasis,G.R., Cherny,S.S., Cookson,W.O., & Cardon,L.R. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97-101 (2002).
45. Abecasis,G.R. & Wigginton,J.E. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet.* **77**, 754-767 (2005).
46. Visscher,P.M., Benyamin,B., & White,I. The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin. Res.* **7**, 670-674 (2004).
47. Fehrmann,R.S. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS. Genet.* **7**, e1002197 (2011).
48. Zhang,W. *et al.* Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum. Genet.* **125**, 81-93 (2009).
49. Raggi,P. *et al.* Heritability of renal function and inflammatory markers in adult male twins. *Am. J. Nephrol.* **32**, 317-323 (2010).
50. Reich,D. *et al.* Admixture mapping of an allele affecting interleukin 6 soluble receptor and interleukin 6 levels. *Am. J. Hum. Genet.* **80**, 716-726 (2007).
51. Scheller,J., Garbers,C., & Rose-John,S. Interleukin-6: From basic biology to selective blockade of pro-inflammatory activities. *Semin. Immunol.*(2013).
52. Hatziapostolou,M. *et al.* An HNF4α-miRNA inflammatory feedback circuit regulates hepatocellular oncogenesis. *Cell* **147**, 1233-1247 (2011).
53. Kwan,T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40**, 225-231 (2008).
54. Gong,J. *et al.* MicroRNA-125b promotes apoptosis by regulating the expression of Mcl-1, Bcl-w and IL-6R. *Oncogene* doi: 10.1038/onc.2012.318 (2012).
55. Jia,H.Y. *et al.* MicroRNA-125b Functions as a Tumor Suppressor in Hepatocellular Carcinoma Cells. *Int. J. Mol. Sci.* **13**, 8762-8774 (2012).
56. Zhu,L.H. *et al.* MicroRNA-23a promotes the growth of gastric adenocarcinoma cell line MGC803 and downregulates interleukin-6 receptor. *FEBS J.* **277**, 3726-3734 (2010).
57. de la Rica,L. *et al.* Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. *J. Autoimmun.* **41**, 6-16 (2013).
58. Moffatt,M.F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211-1221 (2010).

Part 2 Genetic and environmental influences on BMI and other metabolic syndrome traits

The heritability of metabolic syndrome traits in a large population-based sample

Abstract

Heritability estimates of metabolic syndrome traits vary widely across studies. Some studies have suggested that the contribution of genes may vary with age or sex. We estimated the heritability of 11 metabolic syndrome-related traits and height as a function of age and sex in a large population-based sample of twin families (N=2792– 27021, for different traits). A moderate to high heritability was found for all traits (from H^2=0.47 (insulin) to H^2=0.78 (BMI)). The broad-sense heritability (H^2) showed little variation between age groups in women, and differed somewhat more in men (e.g. for glucose, H^2=0.61 in young females, H^2=0.56 in older females, H^2=0.64 in young males, and H^2= 0.27 in older males). While non-additive genetic effects explained little variation in the younger subjects, non-additive genetic effects became more important at a higher age. Our findings show that in an unselected sample (age range: ~18-98 years), the genetic contribution to individual differences in metabolic syndrome traits is moderate to large in both sexes and across age. While the prevalence of the metabolic syndrome has greatly increased in the past decades due to lifestyle changes, our study indicates that most variation in metabolic syndrome traits between individuals is due to genetic differences.

Based on: van Dongen J., Willemsen G, Chen WM, de Geus EJ, Boomsma DI. Heritability of metabolic syndrome traits in a large population-based sample. J.Lipid Res. 2013; 54:2914-2923

Introduction

The metabolic syndrome refers to a combination of traits, including central obesity, insulin resistance, dyslipidemia, and hypertension[1], associated with an increased risk of cardiovascular disease (CVD) and type 2 diabetes (T2D). The underlying pathophysiological mechanisms are thought to include excess adipose tissue mass, ectopic fat deposition, excessive flux of fatty acids, and inflammation [2-4]. Several clinical guidelines have been proposed for the diagnosis of the metabolic syndrome (i.e. ATPIII[5], WHO 1999[6], EGIR 1999[7], IDF 2006[1]), which most commonly include criteria for waist circumference, fasting plasma glucose, systolic blood pressure (SBP), diastolic blood pressure (DBP), HDL cholesterol and triglycerides. Other traits that have been included in individual guidelines are body mass index (BMI, WHO 1999)[6], waist-to-hip-ratio (WHR, WHO 1999)[6] and fasting insulin level (EGIR 1999)[7]. In all guidelines, metabolic syndrome is defined by the levels of multiple traits exceeding a certain threshold.

Although it has been questioned whether the combination of metabolic characteristics referred to as metabolic syndrome represents a biologically meaningful entity in itself [8-10], the definition is widely applied as a tool for risk prediction. The metabolic syndrome is associated with a doubled risk of developing CVD and a more than fivefold increased risk of T2D, with risk estimates varying somewhat depending on the classification criteria used to define the metabolic syndrome [2]. The metabolic syndrome is also associated with a higher risk of non-alcoholic fatty liver disease, reproductive disorders, depression, sleeping disorders and other conditions [2, 11]. The prevalence of the metabolic syndrome is globally increasing [12], presenting a major health problem particularly in Western countries (e.g. the prevalence has been estimated at 14% in the Netherlands [13] and at 37% in the US [14]). The rise is generally attributed to changes in lifestyle, while at the same time individual differences in metabolic traits have been shown to be to an important extent heritable. These observations may be explained by the fact that the expression of risk genotypes depends on the environment or that the exposure to lifestyle factors is genetically influenced. Indeed, several important risk factors for the metabolic syndrome that are considered environmental such as exercise behavior [15] and dietary patterns [16] are moderately heritable.

The importance of genetic influences ("the heritability") to the variation in susceptibility to the metabolic syndrome and associated traits has been estimated based on the similarity of family members for these traits, using twin and family data (reviewed by: [17-20]). If closer relatives resemble each other more for metabolic syndrome traits than more distant relatives do, this indicates that familial factors, including genetic factors or family-shared environmental influences, are important for these traits. The heritability estimates observed in these studies vary widely, with heritability ranging from 24-90% for BMI [18, 21-23], 10-75% for fasting glucose [17, 20, 22-25], 20-55% for fasting insulin [17, 20, 22, 23, 26], 0.03-72% for triglycerides [17, 19, 20, 22-25, 27], 25-98% for LDL-cholesterol [17, 19, 20, 22,

[23, 27], 30-80% for HDL cholesterol [17, 19, 20, 22, 24, 25, 27], 30-74% for total cholesterol [17, 19, 20, 22-24, 27], 20-71% for SBP [17, 20, 22-24] and 10-50% for DBP [17, 20, 22-24]. Such differences may be associated with study designs or represent meaningful variation because heritability can depend on context, sex [28] or age [29].

Quantitative differences refer to differences in the overall impact of genetic or environmental influences on the variance of a trait, such as differences in heritability between males and females or between different age groups. A study based on twin pairs from eight European countries found consistent sex differences in the heritability of BMI, but the direction of effect was not the same across countries and age strata[30]. In a similar comparison, no sex differences were observed for blood pressure[31]. For height, the heritability was slightly lower in females compared to males in several countries [32]. In a family study of body composition measures, a higher heritability was found for waist circumference and WHR in women, while no significant difference was observed for BMI [33]. A study conducted in Sardinian pedigrees found sex differences in the phenotypic variance for many metabolic traits, but a difference in heritability was only evident for weight and triglycerides (for both, the heritability was higher in females)[22]. A review of sex differences in the etiology of metabolic syndrome traits concluded that sex differences in heritability are most often reported for body composition and measures related to glucose homeostasis, while most studies of blood pressure and lipid levels found no sex differences in the heritability [28]. With regard to age effects, a large meta-analysis reported a decrease of the heritability of BMI with age[21], and age differences in heritability were also reported for several metabolic syndrome traits in the analysis of Sardinian pedigrees[22]. In the latter study, the heritability of BMI, lipids, glucose and insulin was found to be lower in older individuals (age > 42), while the heritability of blood pressure was found to be higher in older people.

Qualitative differences arise if a trait is influenced by different genes or aspects of the environment in different groups; for example, different genes may be responsible for the variation of a phenotype in males versus females or at different ages, while the overall impact of genes (the heritability) is the same. Mixed findings have been reported regarding qualitative sex and age effects on metabolic syndrome traits across studies. The results from two large-scale studies that examined a wide range of phenotypes suggested that qualitative genetic differences between the sexes or across age are not common. No evidence was found for qualitative age or sex effects for any of the traits in the study of Sardinian pedigrees [22]. Likewise, in a study of a wide range of phenotypes in metabolic, cardiovascular, physiological and psychiatric domains conducted in a large sample of Dutch twin pairs, no qualitative sex differences were observed for the large majority of traits[34].

If genetic influences depend on age, heritability estimates based on the comparison of twins (who always have the same age and therefore share age-specific genetic influences) may differ from estimates obtained using non-twin

relatives (who may share varying degrees of age-specific genetic influences). The mode of action of the underlying genes (additive versus non-additive genetic effects) may also play a role. The term non-additive genetic effect is generally used to refer to the effects of interacting alleles at a locus (dominance) or at different loci (epistasis). Related to this distinction, the term broad-sense heritability (H^2) refers to the variation of a trait due to total heritable genetic effects, while narrow-sense heritability (h^2) refers to the proportion of variation due to additive genetic effects. Though some studies have indicated that non-additive genetic influences contribute to metabolic syndrome traits [18, 22, 27], most heritability studies have not separated additive and non-additive genetic influences. The consequence of not taking genetic dominance into account when estimating the heritability of a trait while dominance effects are present depends on the study design, since dominance effects are shared to some extent by twins and siblings but not between parents and offspring or between more distant relatives. Family studies have reported lower heritability estimates for metabolic syndrome traits compared to twin studies [21], which could be related to age trends in heritability or to differences in the coverage of non-additive genetic effects between twin and family studies.

Twin and family studies examine the importance of genetic and environmental influences by comparing the resemblance of individuals who share different degrees of genetic or environmental influences. The classical twin study compares the resemblance of monozygotic (MZ) twins to the resemblance of dizygotic (DZ) twins, while family studies typically include parent-offspring pairs or sibling pairs. Since one chromosome from each chromosome pair of a parent is transmitted to a child, a parent always shares one copy of each gene (allele) with his or her child and thus in total shares 50% of additive genetic effects and no genetic dominance effects. For each chromosome pair, DZ twins and siblings can share 2, 1 or 0 chromosomes with each other that was inherited from the same parent, thus sharing on average 50% of additive genetic effects and 25% of dominance genetic effects. MZ twins share all their genetic material because they are derived from one zygote. Importantly, family members may share both genetic and environmental influences. The classical twin design allows distinguishing between heritable genetic influences and influences of the shared family environment ("common environment"): A larger phenotypic correlation in MZ twins than in DZ twins indicates that a phenotype is influenced by genetic factors, because influences of the family environment are shared equally in both types of twins. Most twin studies of metabolic syndrome traits have found no significant effect of the common environment in adults [18, 19, 27, 30, 35, 36]. It should be noted that in classical twin studies the assessment of common environment, by definition, tends to be limited to the effects of early family environment, and does not necessarily capture all environmental influences common to people who share a household, because adult twins often live separately. Insights into the role of

shared household effects in adult subjects may be obtained by studying, for example, the similarity of spouse pairs.

To summarize, variation in heritability estimates of metabolic syndrome traits across previous studies may be related to differences in study population and design. To obtain a representative estimate of heritability in the population and to examine interactions with age and sex, studies should cover a broad age range and include multiple types of familial relations. Therefore, we examined the etiology of metabolic syndrome traits using an extended twin-family design [37-39], combining data from a large population-based sample of MZ twins, DZ twins, non-twin siblings and parents registered with the Netherlands Twin Register (NTR), with a broad age range (18-98 years). Measurements included BMI, waist circumference, WHR, LDL, HDL, total cholesterol, triglycerides, fasting glucose, fasting insulin, SBP, DBP and height. This large dataset with a variety of familial relations allowed us to assess the contribution of additive and non-additive genetic influences (i.e. estimating narrow-sense and broad-sense heritability) and environmental influences (including shared household effects) to the variation in metabolic syndrome traits and height, and to examine variation in the heritability and expression of genes (qualitative effects) across age groups and sex. This study comprises one of the most extensive family-based datasets on metabolic syndrome traits described thus far and represents an elaborate assessment of quantitative and qualitative variation in genetic and environmental effects on metabolic syndrome traits across age and sex.

Materials and Methods

Subjects

The subjects in this study are Dutch twin families registered with the NTR [40]. Most twins were recruited through City Councils between 1990 and 1993 when the twins were adolescents or young adults. Since 1993, adult twins are recruited through a variety of other approaches as well. Every two to five years since 1991, twins and their families are invited to complete a survey (i.e. in 1991, 1993, 1995, 1997, 2000, 2002, 2004 and 2009[40, 41]. In each survey, participants were asked to report their height and current weight. Twin families are also regularly invited to participate in projects in which biological samples, anthropometric traits, and cardiovascular measures are collected. The current analyses are based on data from adult participants (age >= 18), including twins (max. one pair per family), brothers (max. two per family), sisters (max. two per family) and parents. For a detailed description of the characteristics of subjects, see Supplemental Table 1. Informed consent was obtained from participants, and study protocols were approved by the Medical Ethics Committee of the VU University Medical Center. Zygosity determination was based on DNA markers in 84.2 % - 88.7% of twins (range is for different phenotypes), except for BMI (47% of twins) and height (45%), for which a larger proportion of the data came from subjects who had only participated in survey studies. If DNA was not

available, zygosity determination was based on validated questionnaire items. Only subjects with complete information on sex and age (or birth year, for the analysis of height) were included in the analyses.

Procedure biobank project

Metabolic biomarkers and anthropometric measures were collected in a large-scale biobank project in which 9530 individuals participated, including 4259 twins, 2704 biological parents and 2052 biological siblings. Data from non-biological parents and siblings, spouses of twins and siblings, children of twins and siblings and second-degree relatives (e.g. grandparents, uncles and aunts) were not included in the current analyses (N=515). Participants were visited in the morning, usually at their home. At the visit, weight, waist circumference and hip circumference were measured, information about health, medication use, fasting status, and height was collected, and fasting blood and morning urine samples were collected, from which cell lines, biomarkers, DNA and RNA were obtained. For a detailed description of the study procedure, see Willemsen *et al*[42].

Lipid profiles and glucose metabolism

Total cholesterol, high-density lipoprotein cholesterol (HDL) and triglyceride levels were measured in heparin plasma using the Vitros 250 total cholesterol assay, the Vitros 250 direct HDL cholesterol assay and the Vitros 250 Triglycerides assay (Johnson & Johnson, Rochester, USA). Low-density lipoprotein cholesterol (LDL) was calculated using the Friedewald Equation [43]. Glucose and insulin were measured in blood plasma using the Vitros 250 Glucose assay (Johnson & Johnson, Rochester, USA) and the Immulite 1000 Insulin Method (Diagnostic Product Corporation, Los Angeles, USA). The following numbers of subjects had missing data due to sampling issues or technical reasons: glucose, N=253; insulin, N=315; total cholesterol, N=160; LDL, N=180; HDL, N=161; triglycerides, N=159.

Because the distributions of insulin and triglycerides were skewed, an LN-transformation was applied. For the analyses of all lipids, glucose and insulin, individuals who had not fasted from 12 PM the evening before blood collection were excluded (N=706, 7.4%). For the analyses of lipids, individuals using lipid-lowering medication were excluded (N=642, 6.7 %). For the analyses of glucose and insulin, individuals were excluded if they used diabetes medication (N=249, 2.6%) or if they had a fasting glucose level higher than 7 (N=400, 4.2 %). For HDL, one individual with an extreme value was excluded (HDL= 6.56 mmol/L, i.e. >13 SD above the mean). Based on the above inclusion criteria, the following sample sizes were obtained: triglycerides N=7469 (3105 families), total cholesterol: N=7468 subjects (3105 families), HDL: N=7466 subjects (3104 families), LDL: N=7453 subjects (3102 families), glucose: N=7563 (3102 families) and insulin: N= 7510 subjects (3088 families).

Anthropometric Traits

BMI was calculated from height and weight obtained in laboratory-based NTR projects or, if no lab-based data were available, from data obtained in surveys: BMI=weight (kg)/ height^2 (m). For subjects who completed multiple surveys at age 18+, height data were checked for consistency over time. If the difference in height reported by an individual across time did not exceed 2 cm, reported values were averaged to obtain one measure of adult body height. If different self-reports differed by 3 cm or more, the most deviating report was removed and the remaining values were averaged if the difference between remaining values was smaller than 3 cm. If the difference in height reported by an individual at different surveys was 3 or 4 cm after removal of two outlier values, remaining values were averaged to obtain one measure of adult height and if the difference still exceeded 4 cm after removal of two outlier values, height data for that subject were considered unreliable and were excluded. Waist and hip circumference were measured in various projects. For weight, waist circumference and hip circumference, the most recent measure was selected for subjects with multiple data points. Data from twins were selected from the same project or survey where possible. WHR was calculated as: waist (cm)/hip (cm). For WHR, 3 individuals with a WHR > 1.5 (i.e. > 8 SD above the mean) were excluded. Data from the following number of subjects were analyzed: height; N= 24904 (9513 families), BMI; N= 27021 (9793 families), waist circumference; N=8965 (3834 families), WHR; N=8962 (3834 families).

Blood pressure

Blood pressure was measured in a subset of NTR participants as part of several projects that used similar methodology (e.g. Hottenga *et al* [44]). Here, we analyze SBP and DBP measured at rest. For subjects who participated in multiple projects, the first measure was selected. SBP and DBP were corrected for antihypertensive medication use by adding drug-class specific average treatment effects to the measured values [45-48]. In total, data from 2792 subjects (1334 families) were analyzed.

Statistical analysis

The variance of a trait (Phenotypic variance, or V_P) can be divided into genetic variance (V_G), due to genetic differences between individuals, common environmental variance (V_C), due to environmental factors that are shared within families, and unique environmental variance (V_E), caused by environmental factors that are not shared within families [49]. V_E also includes measurement error. Genetic variance can be subdivided into variance due to additive effects of alleles (additive genetic variance, V_A) and variance due to non-additive effects of alleles, which includes interactions among alleles at a single locus (dominance variance, V_D) or at different loci (epistasis). Thus, the variance of a trait may be represented as: $V_P = V_G + V_C + V_E$, where $V_G = V_A + V_D$. The proportion of the phenotypic variance that is due to additive genetic

effects is called narrow-sense heritability ($h^2=V_A / V_P$) and the proportion of variance due to all genetic effects is called broad-sense heritability ($H^2= (V_A + V_D) / V_P$). Using model-fitting approaches, V_A, V_D, V_C and V_E can be estimated from the covariance or correlation of a trait between different types of relatives who differ in genetic relatedness (see Supplemental Methods).

In total, 12 traits were studied; 11 metabolic syndrome traits and height. For metabolic syndrome traits, we examined the heritability and variation with age and sex and for height, we examined the heritability and variation with birth year and sex. Phenotypic correlations among family members were estimated from the observed data after taking sex and age (for metabolic syndrome traits) or birth year (for height) effects into account in Mx[50]. Mx was also used to test for sex differences in means and variances, to test for age (or birth year) effects on means and to test whether the effect of age (or birth year) differed between males and females. To get a first impression of differences in heritability across age and sex (for metabolic syndrome traits) or across birth year and sex (for height), we used SPSS version 17.0 to obtain sex- and age -stratified phenotypic correlations among twins, using the median age or birth year as a cut-off (birth year < 1973 or >= 1973 for height, and age <32 years and >=32 years for all other traits). Next, three types of analyses were performed in POLY ("http://www.sph.umich.edu/csg/chen/public/software/poly")[22, 51]. In a first set of analyses, V_A, V_D and V_E were estimated for each trait in the entire cohort to obtain an overall estimate of heritability. In the second set of analyses, age differences in heritability were examined. V_A, V_D and V_E were estimated and V_A and V_E were allowed to differ between age groups, and the correlation between additive genetic effects among family members belonging to different age groups was estimated to test if different genes are expressed at different ages. Age groups were defined according to the median age of the subjects (age < median and >=median age) for each trait. For height, the heritability was examined as a function of birth year instead of age (birth year < median and >= birth year). In the third series of analyses, we assessed age-specific sex effects; here data from two age groups were analyzed separately. Between age groups, V_A, V_D and V_E could differ and within age groups, V_A and V_E could differ between males and females. Qualitative sex effects were assessed by estimating the genetic correlation among family members of different sex. This third set of analyses provided four estimates of heritability (for younger and older males and females), and an assessment of qualitative sex differences, separately for younger and older subjects. Height was analyzed as a function of sex and birth year instead of age. In all analyses, observed trait values were adjusted for sex and age (using standardized age scores (z-scores)) or birth year (height) by linear regression.

Statistical significance of effects was assessed by comparison of the log likelihood of sub models (e.g. to assess qualitative effects, the log likelihood of a model in which the correlation between additive genetic effects was

estimated was compared to the log likelihood of a model in which it was constrained at its theoretical value when there are no qualitative differences). An alpha level of 0.01 was applied to assess the statistical significance of correlations and to assess the significance of sex differences in means, phenotypic variances and age trends in trait values.

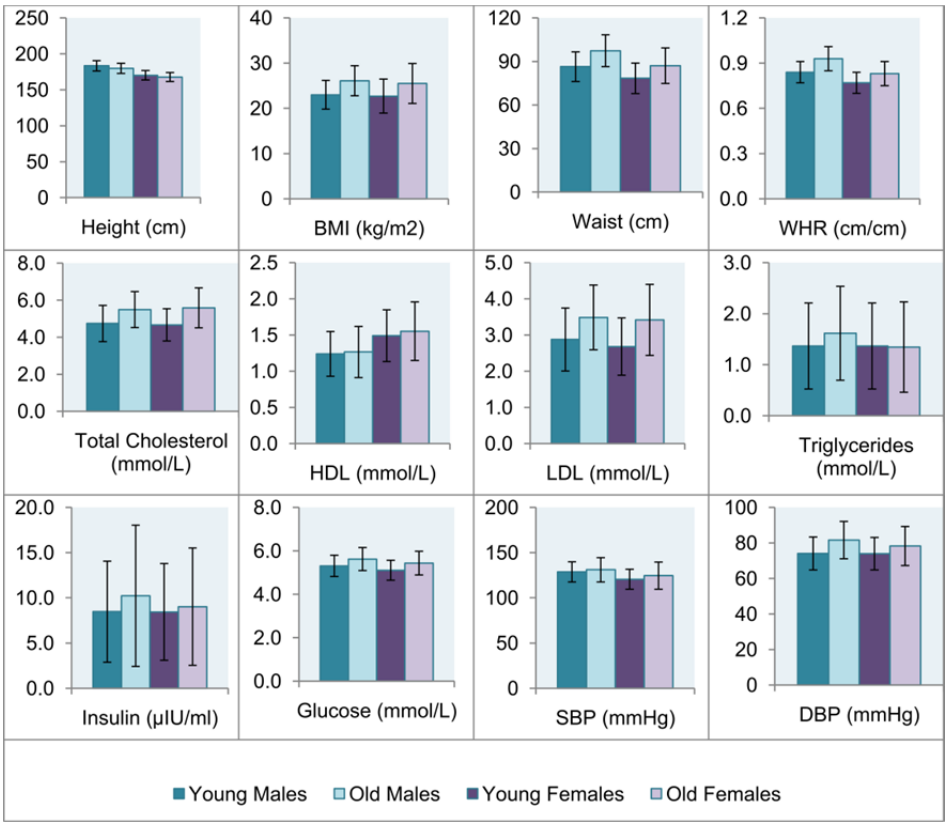
Finally, the data from spouses were used to examine the importance of shared household effects (environmental effects that contribute to the similarity of people who share a household). For each metabolic syndrome trait, the correlation between the mean age of spouses (as a measure of the duration of their relationship) and the absolute trait difference between spouses (as a measure of their similarity) was computed. A negative correlation suggests that spouses become more similar (the difference between spouses becomes smaller) with increasing duration of their relationship, which is suggestive of shared household effects.

Results

Variation of metabolic syndrome traits with age and sex

The mean age of subjects varied slightly for different traits; for parents the mean age ranged from 45.28 (SD=5.70) to 61.88 (SD=7.05) and the age of the offspring (twins and siblings) ranged from 28.79 (SD=11.93) to 42.12 (SD=10.58) for different traits. For a detailed description of the age, sex and numbers of family members for which data on different traits were available, see Supplemental Table 1. All metabolic syndrome traits showed a significant increase with age ($p < 0.001$, Table 1). Significant sex differences in means ($p < 0.01$, Table 1) were observed for all traits except for insulin ($p = 0.315$). On average, males had an unfavorable metabolic profile compared to females, with a higher mean BMI, waist circumference, WHR, SBP and DBP, and higher levels of LDL, triglycerides and glucose. Females had higher levels of HDL and total cholesterol. Many traits showed a significant sex difference in the effect of age on trait level ($p < 0.01$, Table 1), with males having on average a steeper increase in WHR and DBP with age compared to females, and females showing a steeper increase in the levels of total cholesterol, LDL, triglycerides, and glucose with age. Sex differences in the total variance ($p<0.01$, Table 1) were evident for BMI, waist, and HDL (larger variance in females) and for triglycerides (larger variance in males). No significant sex difference in variance was observed for WHR, total cholesterol, LDL, glucose, insulin, SBP and DBP. Height showed a significant increase with birth year, and the mean and variance of height and the effect of birth year on height were larger in males compared to females ($p < 0.01$, Table 1). Figure 1 shows the means of each trait in males and females, separately for the two age groups in which heritability analyses were conducted.

Figure 1: Mean trait values stratified by age and sex. Age categories were defined based on the median age of subjects (young=subjects with an age below the median and older=subjects with an age above the median). The median age was 35 years for SBP and DBP, 39 years for BMI, 40 years for HDL, LDL, total cholesterol and triglycerides, and 41 for insulin, glucose, waist and WHR. For height only, birth year (median=1966) instead of age was used to define categories. Error bars represent standard deviations.



Familial resemblance for metabolic syndrome traits

The familial correlations (Table 2) indicated a substantial role of genetic influences on metabolic syndrome traits, with higher correlations in MZ twins than in DZ twins, non-twin siblings and parent-offspring pairs. Scatterplots illustrating the similarity of sib-pairs are shown in Supplemental figure 1. Looking at the correlations of MZ twin pairs, male MZ twins were slightly more similar compared to female MZ twins for WHR, total cholesterol, HDL, LDL and triglycerides, and insulin, while female MZ twins showed slightly larger similarity for SBP and glucose.

Table 1: Sex differences in mean values, standard deviation and the effect of age on metabolic syndrome traits and height.

		N	Mean[a]	SD[a]	β age[b]
Height (cm)	F	15095	168.85*	6.54*	1.40*[c]
	M	9809	181.31	7.37	2.12[c]
BMI (kg/m^2)	F	16312	24.290*	4.389*	1.636
	M	10709	24.772	3.606	1.583
Waist (cm)	F	5730	83.400*	12.350*	0.532
	M	3235	92.170	11.940	0.556
WHR (cm/cm)	F	5728	0.802*	0.081	0.039*
	M	3234	0.891	0.084	0.044
Total Cholesterol (mmol/L)	F	4840	5.185*	1.088	0.542*
	M	2628	5.142	1.045	0.376
HDL (mmol/L)	F	4839	1.521*	0.385*	0.025
	M	2627	1.255	0.335	0.012
LDL (mmol/L)	F	4833	3.097*	0.972	0.445*
	M	2620	3.207	0.936	0.310
Triglycerides (mmol/L, LN)	F	4841	0.094*	0.450*	0.124*
	M	2628	0.269	0.514	0.088
Glucose (mmol/L)	F	4866	5.293*	0.536	0.197*
	M	2697	5.483	0.533	0.145
Insulin (µIU/ml, LN)	F	4835	1.975	0.703	0.056
	M	2675	2.003	0.738	0.093
SBP (mmHg)[d]	F	1665	122.689*	13.301	2.764
	M	1127	129.844	12.347	1.496
DBP(mmHg)[d]	F	1665	76.097*	10.263	2.495*
	M	1127	77.986	10.511	4.044

F= Females, M= males

[a] Obtained from a model in Mx without correction for age .

[b] Beta for the regression of metabolic syndrome traits on age. The effect of age was significant for all traits ($P < 0.001$, for dropping βage in both sexes).

[c] For height only, the β in this column represents the effect of birth year rather than age.

[d] Corrected for medication use.

*Significant difference in estimate between males and females ($P < 0.01$).

LN: Values in table refer to LN-transformed values.

Table 2: Familial correlations and spousal resemblance for metabolic syndrome traits and height.

	Height	BMI	Waist	WHR	TC	HDL	LDL	Trig	Insulin	Glu	SBP	DBP
MZ twins												
MZ males	0.90	0.79	0.76	0.60	0.74	0.75	0.72	0.67	0.53	0.49	0.58	0.60
MZ females	0.90	0.78	0.76	0.49	0.65	0.65	0.67	0.54	0.46	0.52	0.61	0.60
Male twins/sibs												
DZ males	0.50	0.46	0.45	0.48	0.30	0.57	0.27	0.38	0.39	0.34	0.25	0.21
Brother-brother[a]	0.50	0.28	0.33	0.27	0.30	0.34	0.30	0.27	0.33	0.33	0.28	0.41
Female twins/sibs												
DZ females	0.51	0.40	0.37	0.25	0.43	0.36	0.39	0.23	0.20	0.34	0.51	0.46
Sister-sister[a]	0.47	0.33	0.34	0.09	0.31	0.29	0.37	0.19	0.22	0.33	0.24	0.42
Opposite sex twins/sibs												
DZ opposite-sex	0.46	0.30	0.26	0.25	0.25	0.24	0.26	0.18	0.30	0.32	0.32	0.32
Sister-brother[a]	0.46	0.23	0.20	0.10	0.20	0.25	0.23	0.19	0.12	0.24	0.33	0.32
Parent-offspring												
Mother-daughter	0.50	0.27	0.22	0.21	0.29	0.16	0.30	0.17	0.19	0.21	-0.08	0.01
Mother-son	0.48	0.20	0.23	0.16	0.23	0.16	0.20	0.20	0.23	0.14	0.04	0.11
Father-son	0.49	0.21	0.20	0.17	0.28	0.26	0.25	0.18	0.22	0.22	0.26	0.19
Father-daughter	0.46	0.24	0.24	0.14	0.32	0.29	0.31	0.19	0.16	0.27	0.04	0.27
Spouses												
Mother-Father	0.26	0.22	0.22	0.18	0.13	0.10	0.13	0.09	0.31	0.24	0.01	-0.07
Age-Abs. difference[b]	n.a.	-0.04	-0.05	0.01	0.01	0.01	-0.03	-0.05	0.08	0.02	-0.08	-0.03

[a]Correlations among non-twin sibling pairs and between twins and their non-twin sibling.
[b]Correlations between the mean age of spouses and the absolute difference in their trait values.
TC=Total cholesterol; Trig= Triglycerides, Glu=Glucose

Correlations between parents and offspring were generally a bit smaller compared to correlations between DZ twins and siblings for all metabolic traits, particularly for SBP and DBP. This pattern is suggestive age-specific expression of effects or the presence of non-additive genetic effects. Age-stratified twin correlations (Supplemental Table 2) overall did not suggest large differences in heritability in the younger and older age group for most variables, though some variation between age groups was present. For HDL, total cholesterol and LDL, dizygotic opposite sex (DOS) correlations were consistently smaller in the younger group of twins, which is suggestive of

qualitative sex effects (e.g. differences in the set of genes influencing a trait in males and females).

Spousal resemblance for metabolic syndrome traits

Significant spouse correlations were observed for most traits. Spouses were most similar for anthropometric traits, glucose and insulin (correlations ranging from r=0.18 for WHR to r=0.31 for insulin, $P < 0.01$). For lipids, spouse correlations ranged from 0.10 to 0.13. Spouse correlations were not statistically significant for triglycerides (r=0.09, P =0.026), SBP (r=0.01, P =0.911), and DBP (r=-0.07, P =0.408). Correlations between the mean age of spouses and the difference in their trait values (Table 2) were not significant for any trait, which suggests that spousal resemblance for metabolic syndrome traits does not increase with increasing duration of the relationship.

Heritability of metabolic syndrome traits

Table 3 shows the heritability estimates based on all data (sexes and age groups combined). Moderate to high heritabilities were evident for all traits, with the highest estimates for height (H^2=0.90), BMI (H^2=0.78), and waist circumference (H^2=0.76). The heritability of WHR was lower compared to other body composition measures (H^2=0.49). For lipids, broad heritability estimates ranged from 0.59 (triglycerides) to 0.67 (total cholesterol). Heritabilities were estimated at 0.47 and 0.53 for insulin and glucose and at 0.61 and 0.60 for DBP and SBP. While the narrow-sense heritability (h^2=0.81) and broad-sense heritability (H^2=0.90) of height indicated that additive genetic effects explain most of the total heritability of height, metabolic syndrome traits generally showed a larger discrepancy between broad- and narrow-sense heritability. For BMI and waist circumference, almost half of the broad-sense heritability was ascribed to non-additive genetic effects.

For all traits except for insulin, the phenotypic variance was larger in the older subjects (see methods for definition) compared to younger subjects (Supplemental Table 3). The median age was 35 years for SBP and DBP, 39 years for BMI, 40 years for HDL, LDL, total cholesterol and triglycerides, and 41 for insulin, glucose, waist and WHR. Comparing the heritability estimates in the younger versus older group, no large differences were evident, although broad-sense heritability estimates were a bit lower in the older group for BMI, waist circumference, total cholesterol, LDL, and glucose, while the heritabilities of WHR, HDL, and triglycerides were estimated a bit higher in the older age group. No significant qualitative genetic differences between age groups were found.

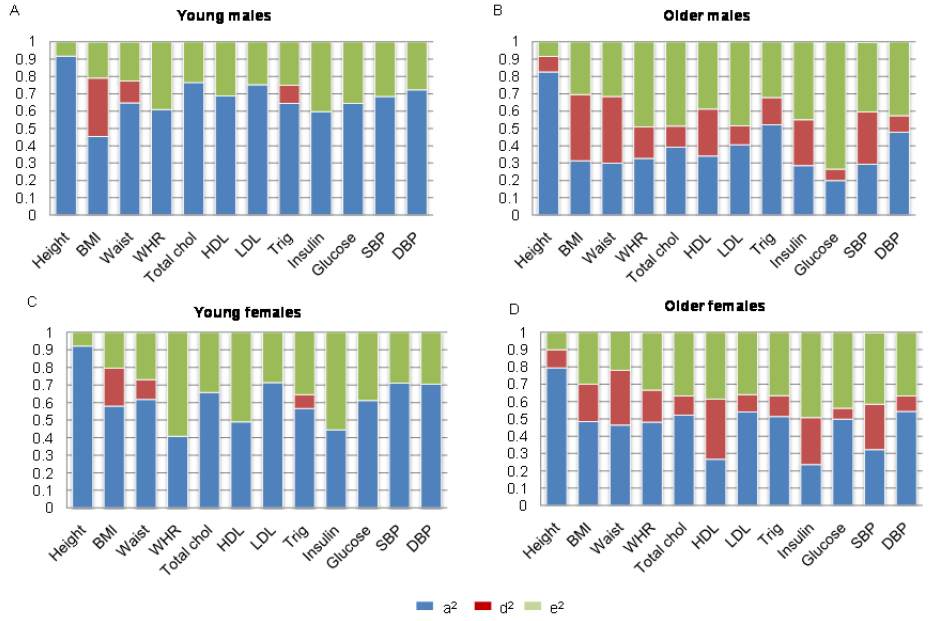
Table 3: Heritability estimates of metabolic syndrome traits and height.

	Height	BMI	Waist	WHR	Total chol	HDL	LDL	Trig	Insulin	Glu-cose	SBP	DBP
N Ss	24904	27021	8965	8962	7468	7466	7453	7469	7510	7563	2792	2792
V_P	42.89	14.14	119.58	0.005	0.93	0.13	0.77	21.25	0.50	0.25	161.21	97.48
h^2	0.81	0.41	0.39	0.31	0.51	0.40	0.51	0.33	0.31	0.38	0.37	0.53
d^2	0.09	0.37	0.37	0.18	0.16	0.27	0.18	0.25	0.16	0.15	0.24	0.08
H^2	0.90	0.78	0.76	0.49	0.67	0.67	0.69	0.59	0.47	0.53	0.60	0.61
SE	0.01	0.02	0.03	0.06	0.03	0.04	0.03	0.03	0.04	0.03	0.08	0.09

All trait values were corrected for age and sex by linear regression, and overall heritability estimates were obtained based on the entire sample, assuming no differences in heritability between sexes or age groups. N Ss= Total number of subjects included in the analysis. TC= Total cholesterol, Trig=Triglycerides. V_p = Total phenotypic variance, after taking out age and sex effects. h^2= narrow-sense heritability (V_A/V_P), d^2= proportion of variation due to non-additive genetic effects (V_D/V_P), H^2=broad-sense heritability ($(V_A+V_D)/V_P$). SE=Robust standard error of the broad-sense heritability estimate from poly.

When the heritability was analyzed separately by age and sex (allowing for differences in A,D and E between age groups and differences in A and E between males and females), age-related differences were more pronounced in males compared to females (Figure 2). This pattern was most evident for LDL (Young males H^2 =0.75, Older males H^2 = 0.52), total cholesterol (Young males H^2 =0.77, Older males H^2 = 0.51) and glucose (Young males H^2 =0.64, Older males H^2 = 0.27) and was primarily due to larger unique environmental variance in the older group. Consequently, these traits showed slightly lower heritability in males compared to females in the older age group, while this pattern was not observed in the younger age group (e.g. for glucose: young females H^2=0.61, older females H^2=0.56). The results suggested no important role of qualitative sex differences; significant qualitative sex differences were only observed in the younger age group for BMI, total cholesterol and HDL and in the older age group only for insulin. The most consistent finding across traits that emerged when analyzing the two age groups separately was a divergence of broad and narrow-sense heritability estimates (Figure 2). Whereas most of the heritable variation across traits in the younger age group was explained by additive genetic effects, non-additive genetic effects accounted for a considerable portion of the heritability of all traits in the older age group.

Figure 2: Heritability estimates stratified by age and sex.
a^2=narrow-sense heritability= V_A/V_P, d^2= V_D/V_P, e^2=V_E/V_P. Age categories were defined based on the median age of subjects (young=subjects with an age below the median and older=subjects with an age above the median). The median age was 35 years for SBP and DBP, 39 years for BMI, 40 years for HDL, LDL, total cholesterol and triglycerides, and 41 for insulin, glucose, waist and WHR. For height only, birth year (median=1966) instead of age was used to define categories. a: Young males. b: older males. c: young females. d: older females.



Discussion

We examined the contribution of genetic and environmental influences to the variation in metabolic syndrome traits in a large population-based sample of twin families, representing one of the most extensive family-based datasets on metabolic syndrome traits described thus far. Our twin-family design allowed for representative estimation of narrow-sense and broad-sense heritability in the general population and to examine whether genetic influences on metabolic syndrome traits interact with age and sex. In summary, moderate to high broad-sense heritability estimates were evident for all traits, ranging from H^2= 0.47 for insulin to H^2= 0.78 for BMI. Averaging the estimates of broad-sense heritability over all metabolic syndrome traits; genetic variation accounted for 62 % of the phenotypic variation of metabolic syndrome traits on average. Although these heritability estimates showed some variation across age groups and sex, it can be concluded that overall, heritability estimates were consistently high. These findings emphasize the importance of heritable influences on individual differences in susceptibility to the metabolic syndrome.

The rising prevalence of the metabolic syndrome is most likely related to changes in lifestyle, including an increase in the consumption of high energy food and a decrease in physical activity. Our findings indicate that even though lifestyle changes are driving an increase in the prevalence of metabolic syndrome, the heritability of the underlying traits continues to be high. This finding may reflect that within wealthy countries including the Netherlands, environmental conditions contributing to these traits are homogenously distributed, with for example high caloric meals being available to every adult individual. Even in this uniform 'obesogenic' environment, not every individual develops an unhealthy metabolic profile and our study indicates that these individual differences in metabolic syndrome traits are largely explained by genetic differences between individuals.

Analysis of qualitative genetic differences across age and sex suggested that the same set of genes account for the variation in metabolic syndrome traits in subjects with an age above the median age in our sample (around 40 years) and subjects with an age below the median age. We also found that the same genes generally accounted for metabolic syndrome traits in men and women, with a few exceptions: For BMI, HDL, and total cholesterol, the data from opposite-sex relatives suggested that the set of genes influencing variation in these traits was not entirely the same for men and women in the younger age group, and for insulin, different genes contributed to the variation in men and women in the older age group. One possible explanation for our finding that the genes influencing BMI in the younger age group differs between males and females could be that this finding is related to pregnancy-related changes in BMI in women, as most women in the reproductive age range were categorized within the younger age group.

While thus far only a few studies have reported that non-additive genetic effects contribute to the variation in metabolic syndrome traits [18, 22, 27], an important part of the heritability of all metabolic syndrome traits was ascribed to non-additive genetic effects in our study, particularly in the older age group. To our knowledge, our study is the first that specifically looked at the variation of non-additive genetic effects across age groups. The overall estimates of narrow-sense and broad-sense heritability from our study are largely similar to the estimates reported previously based on an analysis of extended Sardinian pedigrees that included siblings, parent-offspring pairs, grandparent-grandchild pairs, avuncular relationships and more distant relatives (e.g. for BMI; H^2=0.78 and h^2=0.41 in our study, versus H^2=0.78 and h^2=0.36 in the study by Pilia *et al*[22]). In our study, the information about non-additive genetic effects comes from the difference between the phenotypic correlation in MZ twins compared to DZ twins and siblings and from the difference between the phenotypic correlation among parents and offspring compared to DZ twins and siblings. The variation ascribed to non-additive genetic effects may include genetic dominance effects and epistasis, as well as other genetic effects not acting in an additive manner (causing e.g. lower

parent-offspring similarity compared to sib-pair similarity), such as interactions between genetic effects and age or generation. We found no evidence for a difference in the set of genes influencing metabolic syndrome traits in subjects of different age, suggesting that interactions between genetic effects and age or generation are not a likely explanation for the non-additive effects observed in this study. Our finding that part of the broad-sense heritability of metabolic syndrome traits is ascribed to non-additive genetic effects may explain some of the variation between heritability estimates reported by previous studies.

When looking at the data from spouse pairs, we noticed significant spouse correlations for all traits except for triglycerides and hypothesized that the similarity between spouses may reflect shared household effects (e.g. similarity in diet, leisure time activities etc. in subjects who live together). To explore this hypothesis, we tested whether the similarity of spouses increases over time, but found no evidence for an increase in the similarity of spouse pairs with increasing duration of their relationship. This observation does not rule out, however, that shared household effects do account for the similarity of metabolic syndrome traits in spouse pairs, without inducing an increase in the resemblance of spouses over time, which could be interesting to assess in future studies. In addition to shared household effects, spousal similarity for anthropometric traits may, at least to some extent, be related to assortative mating. Assortative mating refers to the phenomenon that individuals tend to choose a partner whose phenotype is similar to their own and has been demonstrated to contribute to spousal similarity in height and BMI [52].

To summarize, our findings indicate that in a representative population-based sample including multiple types of family relationships, individual metabolic syndrome traits are moderately to highly heritable. Representative heritability estimates are informative to obtain an estimate of the total genetic variation of traits that can be explained by currently identified loci based for example on genome-wide association studies (GWAS). Significant GWAS hits identified to date together explain between 1 and 2 % of the variation in BMI, 10% of the variation in height and 10 % of the variation in HDL cholesterol [53]. Accordingly, a large part of the heritability of these traits still remains to be identified.

Reference List

1. Alberti,K.G., Zimmet,P., & Shaw,J. Metabolic syndrome--a new world-wide definition. A Consensus Statement from the International Diabetes Federation. *Diabet. Med* **23**, 469-480 (2006).
2. Cornier,M.A. *et al.* The metabolic syndrome. *Endocr. Rev.* **29**, 777-822 (2008).
3. Eckel,R.H., Grundy,S.M., & Zimmet,P.Z. The metabolic syndrome. *The Lancet* **365**, 1415-1428 (2005).
4. Grundy,S.M. The Metabolic Syndrome in *Atlas of Atherosclerosis and Metabolic Syndrome* (ed. Grundy,S.M.) 1-26 (Springer, New York, 2011).
5. National Heart Lung and Blood Institute Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**, 3143-3421 (2002).
6. Alberti,K.G.M.M. & Zimmet,P. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. *Diabetic medicine* **15**, 539-553 (1998).
7. Balkau,B. & Charles,M.A. Comment on the provisional report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR). *Diabet. Med.* **16**, 442-443 (1999).
8. Després,J.P. & Lemieux,I. Abdominal obesity and metabolic syndrome. *Nature* **444**, 881-887 (2006).
9. Ding,E.L., Smit,L.A., & Hu,F.B. The metabolic syndrome as a cluster of risk factors: is the whole greater than the sum of its parts?: comment on "The metabolic syndrome, its component risk factors, and progression of coronary atherosclerosis". *Arch. Intern. Med* **170**, 484-485 (2010).
10. Kahn,R., Buse,J., Ferrannini,E., & Stern,M. The metabolic syndrome: time for a critical appraisal. *Diabetes care* **28**, 2289-2304 (2005).
11. VOGELZANGS,N. *et al.* Metabolic depression: a chronic depressive subtype? Findings from the InCHIANTI study of older persons. *J Clin. Psychiatry* **72**, 598-604 (2011).
12. Cameron,A.J., Shaw,J.E., & Zimmet,P.Z. The metabolic syndrome: prevalence in worldwide populations. *Endocrinology and metabolism clinics of North America* **33**, 351-375 (2004).
13. Bos,M.B. *et al.* The prevalence of the metabolic syndrome in the Netherlands: increased risk of cardiovascular diseases and diabetes mellitus type 2 in one quarter of persons under 60. *Nederlands tijdschrift voor geneeskunde* **151**, 2382-2388 (2007).
14. Ford,E.S. Prevalence of the metabolic syndrome defined by the International Diabetes Federation among adults in the U.S. *Diabetes care* **28**, 2745-2749 (2005).
15. Stubbe,J.H. *et al.* Genetic influences on exercise participation in 37.051 twin pairs from seven countries. *PLoS One* **1**, e22 (2006).
16. Teucher,B. *et al.* Dietary patterns and heritability of food choice in a UK female twin cohort. *Twin Res. Hum. Genet.* **10**, 734-748 (2007).
17. Argyropoulos,G., Smith,S., & Bouchard,C. Genetics of the Metabolic Syndrome in *Insulin Resistance : Insulin Action and Its Disturbances in Disease* (eds. Kumar,S. & O'Rahilly,S.) 401-450 (John Wiley & Sons, Ltd, Chichester,UK, 2005).
18. Maes,H.H.M., Neale,M.C., & Eaves,L.J. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet* **27**, 325-351 (1997).

19. Snieder,H., van Doornen,L.J.P., & Boomsma,D.I. Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies. *Arterioscler. Thromb* **19**, 2826-2834 (1999).
20. Teran-Garcia,M. & Bouchard,C. Genetics of the metabolic syndrome. *Appl. Physiol Nutr. Metab* **32**, 89-114 (2007).
21. Elks,C.E. *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol. (Lausanne)* **3**, (2012).
22. Pilia,G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
23. Elder,S.J. *et al.* Genetic and environmental influences on factors associated with cardiovascular disease and the metabolic syndrome. *J. Lipid Res.* **50**, 1917-1926 (2009).
24. Li,S. *et al.* Heritability of eleven metabolic phenotypes in Danish and Chinese twins: A cross-population comparison. *Obesity*(2013).
25. Zarkesh,M. *et al.* Heritability of the metabolic syndrome and its components in the Tehran Lipid and Glucose Study (TLGS). *Genet. Res. (Camb.)* **94**, 331-337 (2012).
26. Poulsen,P., Kyvik,K.O., Vaag,A., & Beck-Nielsen,H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia* **42**, 139-145 (1999).
27. Rahman,I. *et al.* Genetic dominance influences blood biomarker levels in a sample of 12,000 Swedish elderly twins. *Twin Res Hum Genet* **12**, 286-294 (2009).
28. McCarthy,J.J. Gene by sex interaction in the etiology of coronary heart disease and the preceding metabolic syndrome. *Nutr. Metab Cardiovasc. Dis.* **17**, 153-161 (2007).
29. Snieder,H., van Doornen,L.J., & Boomsma,D.I. The age dependency of gene expression for plasma lipids, lipoproteins, and apolipoproteins. *Am. J. Hum. Genet.* **60**, 638-650 (1997).
30. Schousboe,K. *et al.* Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. *Twin Research* **6**, 409-421 (2003).
31. Evans,A. *et al.* The genetics of coronary heart disease: the contribution of twin studies. *Twin Res.* **6**, 432-441 (2003).
32. Silventoinen,K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin. Res.* **6**, 399-408 (2003).
33. Zillikens,M.C. *et al.* Sex-specific genetic effects influence variation in body composition. *Diabetologia* **51**, 2233-2241 (2008).
34. Vink,J.M. *et al.* Sex differences in genetic architecture of complex phenotypes? *PLoS One* **7**, e47371 (2012).
35. Evans,A. *et al.* The genetics of coronary heart disease: the contribution of twin studies. *Twin Res.* **6**, 432-441 (2003).
36. Snieder,H. Familial aggregation of blood pressure in *Clinical hypertension and vascular disease: Pediatric hypertension* (eds. Portman,R.J., Sorof,J.M., Ingelfinger & J.R.) 265-277 (Humana Press, Totowa, NJ, 2004).
37. Eaves,L. *et al.* Comparing the biological and cultural inheritance of personality and social attitudes in the Virginia 30,000 study of twins and their relatives. *Twin Res* **2**, 62-80 (1999).
38. Maes,H.H. *et al.* Flexible Mx specification of various extended twin kinship designs. *Twin Res Hum Genet* **12**, 26-34 (2009).

39. Neale,M.C., Walters,E.E., Eaves,L.J., Maes,H.H., & Kendler,K.S. Multivariate genetic analysis of twin-family data on fears: Mx models. *Behav. Genet* **24**, 119-139 (1994).
40. Boomsma,D.I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* **9**, 849-857 (2006).
41. Willemsen,G. *et al.* The adult Netherlands twin register: twenty-five years of survey and biological data collection. *Twin. Res. Hum. Genet.* **16**, 271-281 (2013).
42. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* **13**, 231-245 (2010).
43. Friedewald,W.T., Levy,R.I., & Fredrickson,D.S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18**, 499-502 (1972).
44. Hottenga,J.J. *et al.* Heritability and stability of resting blood pressure. *Twin Res Hum Genet* **8**, 499-508 (2005).
45. de Geus,E.J.C., Kupper,N., Boomsma,D.I., & Snieder,H. Bivariate genetic modeling of cardiovascular stress reactivity: Does stress uncover genetic variance? *Psychosomatic Medicine* **69**, 356-364 (2007).
46. Kupper,N. *et al.* Heritability of daytime ambulatory blood pressure in an extended twin design. *Hypertension* **45**, 80-85 (2005).
47. Mancia,G. & Parati,G. Office compared with ambulatory blood pressure in assessing response to antihypertensive treatment: a meta-analysis. *Journal of hypertension* **22**, 435 (2004).
48. Wright,J.M., Lee,C.H., & Chambers,G.K. Systematic review of antihypertensive therapies: Does the evidence assist in choosing a first-line drug? *Canadian Medical Association Journal* **161**, 25-32 (1999).
49. Falconer,D.S. *Introduction to quantitative genetics*(Ronald Press Co., New York, 1960).
50. Neale,M.C., Boker,S.M., Xie,G., & Maes,H.H. *Mx: Statistical Modeling.*(Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, 2006).
51. Chen,W.M. & Abecasis,G.R. Estimating the power of variance component linkage analysis in large pedigrees. *Genet. Epidemiol.* **30**, 471-484 (2006).
52. Silventoinen,K., Kaprio,J., Lahelma,E., Viken,R.J., & Rose,R.J. Assortative mating by body height and BMI: Finnish twins and their spouses. *Am. J. Hum. Biol.* **15**, 620-627 (2003).
53. Visscher,P.M., Brown,M.A., McCarthy,M.I., & Yang,J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7-24 (2012).

Chapter 5

Longitudinal weight differences, gene expression, and blood biomarkers in BMI discordant identical twins

Abstract

Background: BMI discordant monozygotic (MZ) twins allow an examination of the causes and consequences of adiposity in a genetically controlled design. Few studies have examined longitudinal BMI discordance in MZ pairs.

Objectives: To study the development over time of BMI discordance in adolescent and adult MZ twin pairs, and to examine lifestyle, metabolic, inflammatory, and gene expression differences associated with concurrent and long-term BMI discordance in MZ pairs.

Subjects/Methods: BMI data from 2775 MZ twin pairs, collected in eight longitudinal surveys and a biobank project between 1991 and 2011, were analyzed to characterize longitudinal discordance. Lifestyle characteristics were compared within discordant pairs (ΔBMI $\geq$ 3 kg/m^2) and biomarkers (lipids, glucose, insulin, CRP, fibrinogen, IL-6, TNF-α and sIL-6R and liver enzymes AST, ALT and GGT) and gene expression were compared in peripheral blood from discordant pairs who participated in the NTR biobank project.

Results: The prevalence of discordance ranged from 3.2% in 1991 (mean age=17, SD=2.4) to 17.4% (N=202 pairs) in 2009 (mean age=35, SD=15), and was 16.5% (N=174) among pairs participating in the biobank project (mean age=35, SD=12). Of 699 MZ with BMI data from 3-5 time points, 17 pairs (2.4%) were long-term discordant (at all available time points; mean follow-up range=6.4 years). Concurrently discordant pairs showed significant differences in self-ratings of which twin eats most ($p=2.3x10^{-13}$), but not in leisure time exercise activity ($p=0.28$) and smoking ($p>0.05$). Ten out of 14 biomarkers showed significantly more unfavorable levels in the heavier of twin of the discordant pairs (p-values < 0.001); most of these biomarker differences were largest in longitudinally discordant pairs. No significant gene expression differences were identified, although high ranking genes were enriched for Gene Ontology (GO) terms highlighting metabolic gene regulation and inflammation pathways.

Conclusions: BMI discordance is uncommon in adolescent identical pairs but increases with higher pair-mean of BMI at older ages, although long-term BMI discordance is rare. In discordant pairs, the heavier twin had a more unfavorable blood biomarker profile than the genetically matched leaner twin, in support of causal effects of obesity.

Based on: van Dongen J, Willemsen G., Heijmans B.T., Neuteboom J., Kluft C, Jansen R, Penninx BWJ, Slagboom PE, Geus EJ, Boomsma DI. Longitudinal weight differences, gene expression, and blood biomarkers in BMI discordant identical twins *(accepted for publication by the International Journal of Obesity)*.

Introduction

Even for highly heritable traits, there can be substantial discordance in monozygotic (MZ) twin pairs [1, 2]. The causes for discordance may include unequal environmental exposures [3, 4], post-twinning DNA mutations [5], stochastic factors [6], and epigenetic differences between twins [7]. For body-mass index (BMI), heritability estimates tend to be high [8]. During foetal life, the heritability of body size increases between the second and third trimester [9]. After birth, the heritability of BMI continues to increase with age during childhood, but decreases with age in adulthood [10]. Similar variation has been demonstrated for the effects of genetic variants on BMI. For example, the association between FTO and BMI strengthens with age during childhood and adolescence to a maximum effect at age 20, but declines after this age [11]. These observations suggest that genetic influences are not deterministic and that the impact of heritable factors at least partly depends on non-genetic factors.

Monozygotic (MZ) twins are genetically (nearly) identical [12] and therefore give insight into the potential range of variation in body weight at a given genetic background. Previous studies, however, indicated that MZ twins with large BMI discordance are rare [13]. To date, at least three studies of BMI-discordant MZ twins have been described, including a well-characterized group of obesity-discordant MZ twins from Finland [13-25], a group of overweight-discordant MZ twins from the United States [26], and a BMI-discordant group from Belgium [27]. In studies that reported the height of BMI-discordant twins, no significant difference in height was evident in these pairs [24, 27]. In the longitudinal cohort of 658 Finnish MZ twin pairs, 14 obesity-discordant MZ twin pairs with an intra-pair difference > 4 kg/m^2 were identified at age 22-27 years [13, 24]. Retrospective data showed that the discordance had emerged around age 18 [24]. Importantly, many BMI discordant pairs did not continue to be discordant when followed over time [14].

Overweight and obesity are commonly regarded as an indicator of excessive energy intake and have been linked to adverse metabolic and cardiovascular changes and to conditions including the metabolic syndrome, type 2 diabetes, coronary artery disease and depression [28-34]. Growing evidence suggests that a key mechanism behind the pathogenesis of the consequences of obesity and associated conditions involves chronic over-activation of cellular stress signalling and inflammatory pathways in response to energy intake that strongly exceeds energy expenditure [35-38]. Associations of overweight and obesity with blood levels of metabolic and inflammatory biomarkers are well-established based on epidemiological studies (see for example [39], but a limitation in population-based studies is that results can be (partly) confounded by genetic factors, because weight, lipid levels and glucose metabolism can be influenced by common underlying genetic influences [40]. Several genes with pleiotropic effects on birth weight and type 2 diabetes have been identified [41, 42], where the allele associated with lower birth weight and

subsequent increased postnatal weight gain also increases the risk of Type 2 diabetes in adulthood.

Studying biomarker levels in BMI discordant MZ twins has the advantage that the relationship between differences in BMI and biomarker levels can be revealed under an identical genetic background and age. The current study had two main aims. Firstly, we aimed to examine the prevalence of BMI differences in MZ twin pairs and their development over time, analyzing BMI data from 2775 MZ twin pairs collected throughout adolescence and adulthood over a period of up to 20 years. Secondly, in subsets of these data, we examined whether differences in life style factors, metabolic and inflammatory biomarkers, and gene expression in peripheral blood are present in concurrently BMI discordant and long-term discordant MZ twin pairs.

Materials and methods

Subjects

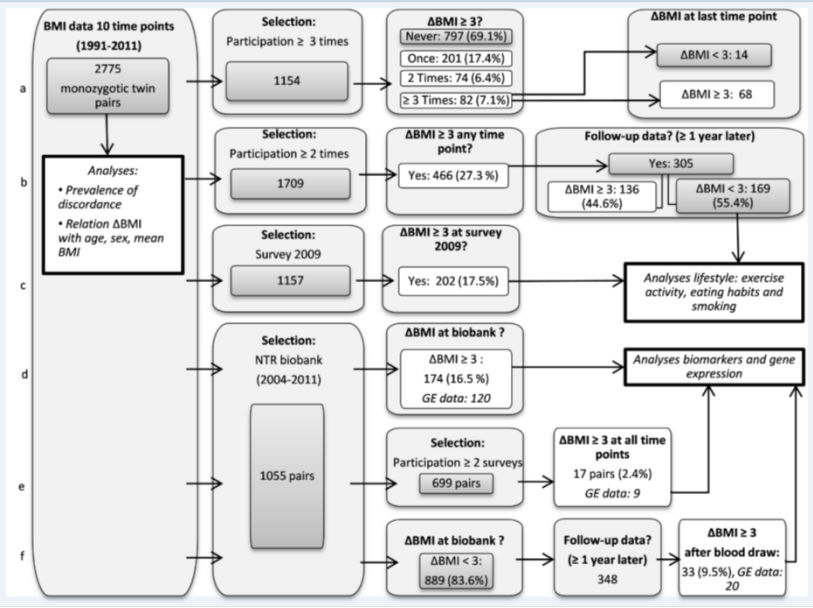
MZ twins from the Netherlands Twin Register (NTR)[43] took part in eight longitudinal survey studies between 1991-2009 and between 2004-2011 a subgroup also participated in the NTR biobank project [44, 45]. BMI data were available for 2775 MZ twin pairs (including 6 pairs who were part of triplets): 1709 pairs participated in the survey studies only and 1066 pairs participated also in NTR biobank, of which 1044 pairs participated once, and 22 pairs participated twice in biobanking (interval: 3 – 7 years, mean= 5 years). Of the pairs who participated in NTR biobank, eleven pairs (including 3 pairs who were discordant for BMI) were excluded because one twin was pregnant. After quality control, data on gene expression and cell counts were available for 634 pairs. Zygosity assessment is described in the Supplemental Methods. Informed consent was obtained from participants and study protocols were approved by the Medical Ethics Committee of the VU University Medical Centre.

Anthropometric, health and lifestyle measures

Data on height and weight were obtained in eight surveys (self-report) and were measured in the NTR biobank project by a calibrated balance and a stadiometer. BMI was calculated as: weight (kg)/ (height $(m)^2$). Self-reported height data were checked for consistency over time (Supplemental Methods). Surveys also contained questions regarding demographic and lifestyle characteristics, including cigarette smoking, eating habits, leisure time exercise activities and birth weight (Supplemental Methods) . Waist and hip circumference were assessed in the NTR biobank project with measurement tape. Additional measures collected at blood draw for the NTR biobank project included information regarding lipid-lowering and diabetes medication, menopause and pregnancy status. BMI difference (ΔBMI) was computed for each MZ pair as the difference between the heavier and the lighter twin, for all data points (for N pairs at each survey and the NTR biobank project, see table

1). BMI discordance was defined as ΔBMI ≥3 kg/m^2, in line with the threshold applied in previous studies of BMI discordant pairs [13, 26, 27].

Figure 1: Flowchart of the selection procedure of MZ twin pairs included in each analysis. All numbers in this figure represent numbers of MZ twin pairs. GE= Gene expression. Each row (a to f) illustrates the available data and selection criteria for MZ pairs included in a particular analysis. a) Frequency of BMI discordance at one, two or more longitudinal time points in MZ pairs with longitudinal BMI data. b) Number of MZ pairs who are discordant across all projects and the number of pairs who are still discordant at the first next available follow-up time point. c) Discordant pairs included in the analyses of life style data. d) MZ pairs who were discordant at blood draw and were included in the analyses of biomarkers and gene expression. e). MZ pairs who were discordant at all time points of participation and were included in the analyses of biomarkers and gene expression. f) MZ pairs who became discordant after blood draw and who were studied to examine biomarkers and gene expression difference before BMI discordance onset.



Blood biomarker profiles

Blood samples were collected as part of the NTR biobank project after overnight fasting [44] to assess total cholesterol, HDL cholesterol, triglycerides, LDL cholesterol, glucose, insulin, TNF-α, IL-6, sIL-6R, fibrinogen, CRP, aspartate aminotransferase (AST), alanine aminotransferase (ALT), and gamma glutamyl transferase (GGT), as described in detail previously [44]. Data on blood biomarkers were available for 878 (83%) – 966 (92%) complete MZ

pairs who participated in NTR biobank (range is for different biomarkers). For additional information, see Supplemental Methods.

Gene expression profiles

RNA extraction [44], expression profiling, and expression quality control procedures have been described in detail previously [46]. In short, gene expression in whole blood drawn for the NTR biobank project was measured with Affymetrix U219 arrays (GeneTitan), which contain 530,467 probes for 49,293 transcripts. For further information, see Supplemental Methods.

Statistical analyses BMI, lifestyle and biomarkers

The selection of twins for each analysis is illustrated in Figure 1. Associations of ΔBMI with age, sex and BMI level were tested by linear regression analysis in SPSS version 21 with ΔBMI as outcome, and sex, age, and mean BMI of co-twins (BMI twin 1 + BMI twin 2)/ 2) as predictors. Here, observations corresponded to twin pairs, and one measure of BMI was included for each pair (2775 pairs in total), which was selected from the most recent time point at which both twins participated, with a preference for biobank measures of BMI. To examine the progression of BMI discordance over time, we studied data from MZ pairs with BMI data available from ≥ 3 time points (N= 1154 pairs, figure 1a) and MZ pairs with BMI data available from ≥ 2 time points (N=1709, figure 1b). Lifestyle data (exercise activity, eating habits and smoking) collected in survey 8 (2009) were studied in all pairs who were discordant for BMI at that time point (figure 1c), and blood biomarkers, gene expression, and additional measures collected as part of the NTR biobank project were examined in all MZ pairs who were discordant at blood draw (figure 1d), and in a subset of longitudinally discordant pairs, who were selected out of 699 MZ pairs who participated in the NTR biobank plus in at least two surveys (Figure 1e). Finally, to verify whether differences within MZ pairs are present before BMI discordance, biomarkers and gene expression differences were tested in a separate group of MZ pairs. These pairs were not yet discordant when blood samples were collected or at prior surveys but they became discordant ≥ 1 year after blood draw. Thus for these pairs, data on gene expression and biomarkers were not available during discordance (Figure 1f). Differences between the heavier and leaner twin from discordant pairs were tested with Wilcoxon Signed Ranks tests (ordinal data), McNemar tests (dichotomous data) and paired t-tests (continuous data) in SPSS. In total, nine lifestyle variables were compared within discordant pairs who participated in surveys and 24 variables (excluding gene expression) were compared within discordant pairs who participated in the NTR biobank. To account for multiple testing, a p-value < 0.002 (=0.05/27) was considered significant in comparisons of biomarkers and lifestyle, where 27 represents the number of independent dimensions in the data, estimated with the online software program MatSpD (http://gump.qimr.edu.au/general/daleN/matSpD/ ; Supplemental Methods). To

rule out that small differences in age between co-twins (related to variation in the response time to questionnaires and because a subset of MZ co-twins who participated in the NTR biobank were not assessed on the same day) influenced the within-pair comparison of BMI, we tested for differences in age at assessment; there were no differences in age between discordant twins in any of the groups.

Statistical analyses gene expression

Gene expression levels corrected for a number of covariates (neutrophil , basophil, eosinophil, lymphocyte and monocyte cell counts, smoking status, age, sex, hemoglobin, hour of blood sampling, days between blood sampling and RNA extraction, plate and location on the plate, see ref [47]) were compared within discordant pairs by applying a paired t-test to all probe sets (44 241 probe sets after quality control) in R [48]. All probe sets were ranked by p-value to test for enrichment of Gene Ontology (GO) terms and for enrichment of a set of candidate genes for BMI, based on loci identified by a GWAS of BMI [49]. Enrichment analyses were conducted with the software packages *GOrilla* [50] and GSEA [51, 52], as described in the Supplemental Methods. All gene expression analyses accounted for multiple testing by controlling for the false discovery rate (FDR). An FDR q-value < 0.05 was considered significant. The FDR q-value for probe sets was computed with the R function qvalue() with default settings.

Results

Prevalence of BMI discordance and relation with age, sex, and mean BMI

The mean age of twins ranged from 17 years in 1991 (first survey) to 35 years in 2009 (last survey, from which lifestyle variables were analyzed). At all time points the majority of MZ twins had highly similar BMIs (Table 1), with 87.7-89.0% of pairs in surveys 1-3 showing a BMI difference < 2 kg/m^2. The percentage of discordant MZ pairs (ΔBMI ≥ 3) ranged from 3.2% in survey 1 to 17.4% in survey 8, when relatively more older pairs were included. To illustrate: if both twins have a height of 175 cm, a ΔBMI of 1 between co-twins corresponds to a weight difference of 3.1 kg, ΔBMI of 2 corresponds to a weight difference of 6.1 kg and ΔBMI of 3 to 9.2 kg.

Within-pair differences increased at each successive survey, together with the age and BMI of twins. In a linear regression analysis with ΔBMI as outcome and sex and age as predictors, ΔBMI was larger in female pairs and increased significantly with the age of twins (P_{sex}= 7.2 x 10^{-3}, P_{age}= 4.6 x 10^{-21}). However, when mean BMI of the twins was added as predictor, ΔBMI was only significantly associated with sex and mean BMI (P_{sex}= 4.9 x 10^{-5}, $P_{meanBMI}$=8.7 x 10^{-84}, P_{age}=0.22). On average, ΔBMI was 0.8 (SD=0.7) in pairs with a mean BMI in the underweight range (BMI<18), 1.3 (SD=1.3) in pairs with a 'normal' mean BMI (18-25), 2.5 (SD=2.4) in pairs with a mean BMI in the overweight range (25-30) and 3.9 (SD=3.9) in obese pairs (BMI >30), suggesting that

ΔBMI increases with the mean BMI of a pair. We next ranked the twin pairs on the basis of the BMI-class (underweight, normal weight, overweight, or obese) of the leaner twin. A BMI difference ≥ 3 between co-twins was evident in 10.1% of pairs where the leaner twin had a BMI in the underweight range, 12.3% of pairs where the leaner twin had a normal BMI, 21.7% of pairs where the leaner twin was overweight, and 45.5% of pairs where the leaner twin was obese.

Table 1. Average BMI difference between MZ twins and frequencies of various degrees of concordance and discordance at each survey and NTR biobank project.

					% of All MZ pairs					% male pairs	% female pairs
Survey no/ Project, year	N pairs[A]	% Female	Mean Age (SD)	Mean ΔBMI (SD)	Δ BMI 0-1	Δ BMI 1-2	Δ BMI 2-3	Δ BMI 3-4	Δ BMI ≥4	Δ BMI ≥3	Δ BMI ≥3
1, 1991	590	58.6	17.2 (2.4)	0.9 (0.9)	67.5	20.2	9.2	1.5	1.7	4.1	2.6
2, 1993	771	60.7	19.5 (8.4)	0.9 (1.4)	68.5	20.5	7.7	1.6	1.8	4.0	3.0
3, 1995	626	61.5	19.2 (3.2)	1.0 (1.0)	65.3	23.6	6.9	2.4	1.8	4.1	4.2
4, 1997	563	67.5	25.4 (10.3)	1.4 (1.4)	53.5	25.2	9.8	6.6	5.0	7.1	13.7
5, 2000	827	73.6	30.1 (12.0)	1.6 (1.6)	45.9	26.4	15.4	5.2	7.1	9.6	13.3
6, 2002	804	73.0	33.1 (12.1)	1.6 (1.7)	47.4	24.5	13.3	6.7	8.1	12.4	15.7
7, 2004	1155	75.5	36.2 (13.1)	1.7 (1.6)	42.9	26.8	13.9	7.2	9.4	15.9	16.7
8, 2009	1157	75.4	34.6 (15.0)	1.7 (1.7)	45.4	25.0	12.2	8.2	9.2	14.7	18.3
BB,2004-2011	1055	69.1	34.9 (12.4)	1.8 (1.9)	42.1	26.3	15.2	7.1	9.4	16.6	16.5

Δ BMI = Difference between the BMIs of co-twins.

[A] Number of complete MZ pairs.

Percentages in the table represent the percentage of MZ twin pairs with a certain BMI difference, relative to the total number of MZ twin pairs participating at each individual time point. BB=Biobank.

Progression of BMI discordance over time

To examine the progression of BMI discordance over time, we studied data from pairs who participated in at least three NTR projects (Supplemental Table 1), and found that 30.9% was discordant at least once, but only 7.1 % of all pairs had a Δ BMI ≥3 at three or more time points; still, some of these pairs did not cross the threshold for discordance at the most recent project in which they participated (N=14, Figure 1a). These findings suggest that it is not uncommon for MZ twins to show episodes of discordance and converge later, while long-term BMI discordance is rare. Across all time points, we identified 305 BMI-discordant MZ twin pairs (Δ BMI ≥3, at any time point) with follow-up data after

being identified as discordant (on average 3 years later, Figure 1b). At follow-up, the average BMI difference between co-twins had decreased, and 169 pairs (55.4%) were no longer discordant ("converging pairs"), due to weight gain of the leaner twin (mean 5.2 kg, SD=6.1) and weight loss of the heavier twin (-2.9 kg, SD=6.8; Supplemental Table 2). The following combinations were observed among converging pairs: leaner twin gained weight and heavier twin lost weight (45.6%), both twins gained weight (27.2%), both twins lost weight (11.8%), the leaner twins' weight was stable while the heavier twin lost weight (7.7%), or the heavier twin's weight was stable while the leaner twin gained weight (7.7%). Overall, 80.4% of initially leaner twins from converging pairs gained weight, and 65.1% of initially heavier twins lost weight. For 98 converging pairs, we also had BMI data before they became discordant (on average 3 years earlier). These data showed that 79.6% of the heavier twins and 76.5% of the leaner twins had ended up heavier after discordance in comparison to their BMI before discordance (average weight change in all converging pairs over the entire period of on average 6.5 years: heavier twin; mean=+4.8kg, SD=6.5, leaner twin; mean=+4.9kg, SD=6.9), and suggest that discordance mainly reflects one twin starting out earlier on a trajectory of weight gain.

Lifestyle

To assess whether BMI discordance in MZ twins is related to lifestyle differences, survey data collected in 2009 from 202 discordant pairs were studied (Figure 1c; Table 2). Heavier and leaner twins equally often reported to participate in leisure time exercise regularly (62.9% of leaner twins and 57.1% of heavier twins participated in exercise on a regular basis, P=0.28). Discordant twins also did not differ in the number of reported hours of exercise per week (P=0.58) but a difference was noticed in response to a question that asked twins about their relative food intake (P=2.3 x 10^{-13}). To the question "Who of you eats most?" 50.3% of the heavier twins responded with "I eat most" versus 6.3% of the leaner twins. 43.2% of the leaner twins reported that their co-twin eats most, while 3.1% of the heavier twins reported that their co-twin eats most. Heavier twins also reported to go on a diet more often compared to their leaner co-twins (P=2.1 x 10^{-5}). Discordant twins did not differ significantly in smoking status ($P_{\text{current smoking}}$=0.054, $P_{\text{ever smoked}}$=0.50).

We hypothesized that changes in smoking status may potentially contribute to shifts in BMI discordance and therefore compared smoking status in pairs who were initially discordant but concordant after a period of on average 3 years ("converging pairs" described in the previous section). We observed a larger percentage of individuals who quit smoking (14.1%) among the initially leaner twins compared to the initially heavier twins (4.3%). Of the initially leaner twins who quit smoking, all except for one twin gained weight (mean change=+9.8 kg, range=-2 kg – +27 kg). Of the (initially) heavier twins, 14.5% had started smoking (of which 70 % lost weight: mean change=-2.56 kg, range= -10 kg - +8kg) *versus* 2.8% of the initially leaner twins. Although this

Table 2. Leisure time exercise activity and eating habits of BMI discordant pairs (ΔBMI ≥ 3) who participated in NTR survey 8 (2009).

	Heavier twins	Leaner twins	P-value[A]
N pairs	202	202	
N male/ female pairs	42/160	42/160	
Age (years)	40.2(16.0)		
BMI (kg/m^2)	28.0 (4.2)	23.4 (3.6)	1.6×10^{-89}
Do you participate in leisure time exercise?			0.28
Yes	57.1%	62.9%	
No	42.9%	37.1%	
Frequency leisure time exercise activity			0.58
Almost never	30.7%	31.6%	
1-5 hours per week	56.1%	51.8%	
5-10 hours per week	11.10%	13%	
> 10 hours per week	2.10%	3.6%	
Ever been on a diet			2.1×10^{-5}
Never	40.1%	54.8%	
A few times	29.2%	25.6%	
Multiple times	16.8%	11.6%	
Often	10.9%	5.5%	
Always on a diet	3.0%	2.5%	
Fear to gain weight			0.01
Not scared	29.2%	37.2%	
A little bit scared	42.6%	38.2%	
Quite scared	17.3%	18.6%	
Very scared	8.4%	6.0%	
Extremely scared	2.5%	0.0%	
How fast do you normally eat?			0.64
Very slow	0.5%	1.0%	
Slowly	6.0%	8.0%	
Medium	60.0%	56.2%	
Fast	30.5%	33.8%	
Very fast	3.0%	1.0%	
Do you normally eat until you feel full?			0.07
I stop with eating before I feel full	34.5%	40.1%	
I stop with eating when I feel full	57.5%	55.4%	
I continue eating, even when I feel full	8.0%	4.5%	
Who of you eats most?			2.3×10^{-13}
I do	50.3%	6.5%	
We eat just as much	29.2%	27.6%	
My co-twin	3.1%	43.2%	
I do not know	17.4%	22.6%	
Smoking			
Current smoker	27.4%	34.8%	0.05
Ever smoked	69.9%	72.3%	0.50

[A]P-value from a paired t-test (BMI), McNemar test (Regular sports and smoking), or Wilcoxon signed rank test (all others). Mean (SD) or percentages are displayed.

pattern is in line with changes in BMI discordance being related to changes in smoking status of twins, the difference in smoking status over time in converging pairs was nominally significant only (p=0.012), and a similar trend of quitting smoking was noticed among the leaner twins from pairs of who were still BMI discordant at follow-up (Supplemental Table 2).

Blood biomarkers

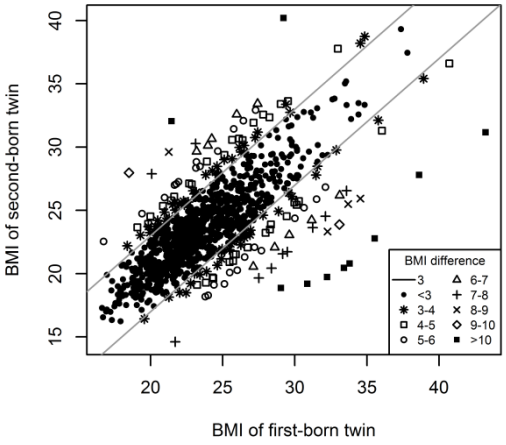
There were 174 MZ twin pairs with a BMI difference ≥ 3 BMI kg/m^2 at blood draw (16.5%). Their average age was 38.5 years and 69 % were female. The BMI of the heavier twins was on average 5.1 kg/m^2 (22%) larger compared to the leaner co-twins (range= 3-13, Figure 2a), and the weight, waist circumference and hip circumference of heavier twins were on average 14.7 kg, 11.0 cm and 8.4 cm larger respectively, compared to their leaner co-twins (Table 3, first four columns). Heavier twins had significantly ($p < 0.002$) higher levels of glucose, insulin, total cholesterol, LDL, triglycerides, CRP, IL-6, sIL-6R, and GGT, a lower level of HDL cholesterol (p-values: 6.0×10^{-13}-0.001), and a nominally significant trend of higher fibrinogen ($p=2.2 \times 10^{-3}$), compared to their leaner co-twins. Discordant twins did not differ significantly ($p \geq 0.002$) in height, birth weight, plasma levels of TNF-α, AST and ALT, menopause status, and use of lipid-lowering medication or diabetes medication (p-values: 0.04-0.88). These findings illustrate that the heavier twins from genetically identical BMI-discordant pairs show less favourable biomarker profiles, a pattern that is in line with reports from population-based studies on the relationships between BMI and biomarkers.

Gene expression

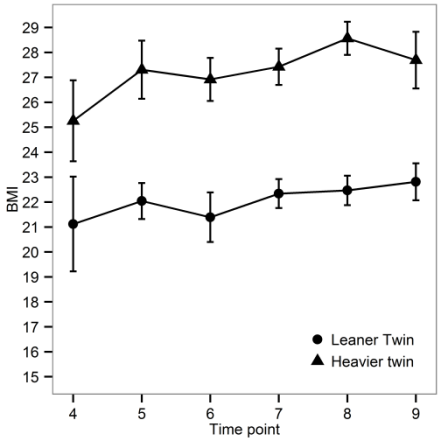
Of the 174 pairs who were discordant at blood draw, whole blood gene expression data were available for 120 pairs. None of the probe sets identified a difference in expression between discordant twins reaching genome-wide significance (FDR q-values > 0.05, for the top 100 probes see Supplemental Table 3), and discordant twins also showed no difference in the expression of BMI candidate genes from GWAS (Supplemental Table 3). GO enrichment analysis based on p-value rank from the gene expression comparison of discordant twins highlighted significant enrichment of a number of GO terms (FDR q-value < 0.05) related to broad metabolic categories (e.g. regulation of metabolic process, cellular macromolecule metabolic process), suggesting that BMI discordance is associated with small but widespread differences in the expression in blood of genes related to metabolism (Supplemental Table 3). Other GO processes significantly enriched among high ranking genes included hepatocyte differentiation and negative regulation of type I interferon production, and enriched GO components included Golgi apparatus, NLRP3 inflammasome complex, and mitochondrial outer membrane, amongst others.

Figure 2: BMI of MZ twin pairs who participated in the NTR biobank project (N=1055 pairs) and longitudinal BMI of 17 longitudinally discordant MZ pairs. a) Concordant pairs (Δ BMI < 3) are denoted by filled circles (N=881) and discordant pairs (Δ BMI ≥ 3 kg/m^2) are indicated by other symbols (N=174). The grey lines indicate the threshold for discordance. B) Mean BMI of longitudinally discordant twins across NTR projects, with data from the heavier twins denoted by triangles and data from the leaner twins denoted by circles. c) Mean age of longitudinally discordant twins across NTR projects. Error bars represent standard errors. Time points 4-7 and 9 are surveys and time point 8 represents the time point of blood draw (NTR biobank). Note: None of these pairs participated in surveys 1-3. The following number of twins participated at each time point (N=leaner twin/heavier twin): 4: N=3/3, 5: N=10/10, 6: N=7/10, 7: N=15/15, 8: N=17/17, 9: N=14/11.

a



b



c

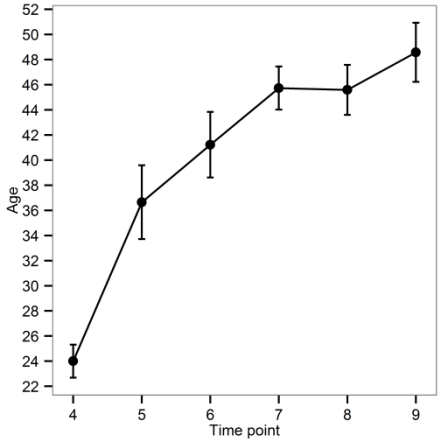


Table 3. Characteristics at the moment of blood draw of all BMI discordant MZ twin pairs (Δ BMI ≥ 3 kg/m^2) who participated in the NTR biobank project, and for the longitudinally discordant subset.

	All discordant pairs				Longitudinally discordant subset			
	Heavier Twin	Leaner Twin	Mean diff*	P-value	Heavier Twin	Leaner Twin	Mean diff*	P-value
N	174	174			17	17		
N male/female pairs	54/120	54/120			1/16	1/16		
Age (years)	38.5 (12.6)				45.6 (11.6)			
Birth weight (g)	2394 (600)	2378 (614)	16	0.16	2366 (742)	2538 (752)	-172	0.27
BMI (kg/m^2)	28.4 (4.2)	23.3 (3.7)	5.1	$8.7x10^{-71}$	28.6 (2.7)	22.5 (2.4)	6.1	$5.1x10^{-7}$
Weight (kg)	84.3 (14.8)	69.6 (13.9)	14.7	$3.7x10^{-68}$	80.1 (9.6)	63.1 (9.2)	17.0	$1.9x10^{-7}$
Height (cm)	172.1 (8.2)	172.4 (8.4)	0.3	0.32	167.4(5.6)	167.3 (5.3)	0.1	0.84
Waist (cm)	91.2 (12.0)	80.1 (11.3)	11.1	$1.5x10^{-47}$	92.2 (10.3)	76.6 (8.5)	15.6	$3.9x10^{-8}$
Hip (cm)	109. 2 (8.3)	100.8 (8.0)	8.4	$7.2x10^{-47}$	109.7 (5.8)	98.6 (6.9)	11.1	$8.1x10^{-10}$
WHR (cm/cm)	0.84 (0.08)	0.79 (0.08)	0.05	$1.1x10^{-15}$	0.84 (0.08)	0.78 (0.07)	0.06	$6.6x10^{-5}$
Glucose (mmol/L)	5.4 (0.8)	5.1 (0.6)	0.3	$3.4x10^{-8}$	5.4 (0.6)	5.1 (0.4)	0.3	0.06
Insulin (μIU/ml)	12.3 (12.4)	7.4 (4.5)	4.9	$5.6x10^{-13}$	10.1 (6.0)	5.3 (2.6)	4.8	0.03
Total Chol (mmol/L)	5.2 (1.2)	4.9 (1.1)	0.3	$2.0x10^{-4}$	5.9 (1.5)	5.2 (1.4)	0.7	0.03
LDL (mmol/L)	3.2 (1.1)	2.9 1.0)	0.3	$5.4x10^{-4}$	3.6 (1.4)	3.2 (1.5)	0.4	0.15
HDL (mmol/L)	1.3 (0.3)	1.5 (0.3)	-0.2	$2.2x10^{-7}$	1.5 (0.3)	1.6 (0.3)	-0.1	0.20
Triglycerides (mmol/L)	1.5 (0.9)	1.2 (0.5)	0.3	$8.1x10^{-11}$	1.5(0.9)	0.9 (0.4)	0.6	$9.2x10^{-4}$
CRP (mg/L)	4.5 (5.6)	2.7 (4.2)	1.8	$4. 6x10^{-11}$	4.4(4.6)	1.2 (1.4)	3.2	$4.9x10^{-4}$
TNFα (pg/ml)	1.1 (1.3)	1.1(1.4)	0.0	0.41	1.95 (3.7)	1.05 (1.3)	0.9	0.86
IL-6 (pg/ml)	1.7 (1.3)	1.5 (1.4)	0.2	$1.1x10^{-3}$	2.3 (1.9)	1.4 (0.8)	0.9	0.12
sIL-6R (pg/mL)	42506.5 (11168.4)	40093.9 (11051.6)	2412.6	$4.6x10^{-4}$	42645 (13375)	38672 (13340)	3973	0.14
Fibrinogen (g/L)	3.0 (0.7)	2.8 (0.8)	0.2	$2.2x10^{-3}$	3.3 (0.7)	2.6 (0.6)	0.7	$1.3x10^{-5}$
AST (U/L)	22.4 (6.3)	21.5 (6.5)	0.9	0.04	22.9 (6.9)	22.4 (10.0)	0.5	0.32
ALT (U/L)	12.3 (7.7)	11.4 (5.2)	0.9	0.68	12.6 (6.6)	12.0 (5.4)	0.6	0.90
GGT (U/L)	32.7 (23.4)	27.6 (19.6)	5.1	$5.4x10^{-4}$	29.4 (18.0)	20.6 (10.6)	8.8	$1.9x10^{-3}$
N (%) using lipid-lowering med	5 (2.9%)	7 (4%)	-2	0.73	1 (5.9%)	1 (5.9%)	0	0.99
N (%) using diabetes med	1 (0.6%)	1 (0.6%)	0	0.99	0(0%)	0 (0%)	0	0.99
N (% of female twins) menopause	20 (16.7%)	16 (13.3%)	4	0.22	6(37.5)	6 (37.5%)	0	0.99

*heavier -leaner twin. Numbers in the table represent Mean (SD) or N (%).Waist= Waist circumference, Hip= Hip circumference, WHR=Waist-to-hi ratio, Total Chol=Total cholesterol, AST= aspartate aminotransferase, ALT= alanine aminotransferase, GGT= gamma glutamyl transferase. Med=medication

Table 4: Characteristics of MZ twin pairs who participated in the NTR biobank project and who became discordant (Δ BMI ≥3 kg/m^2) after blood draw.

	MZ pairs who became discordant after blood draw			
	Heavier Twin	Leaner Twin	Mean	P-value
N	33	33		
N male/female pairs	7/26	7/26		
Age post-blood draw (years)[A]	34.7 (11.0)			
BMI post-blood draw(kg/m^2) [A]	27.1 (3.2)	23.0 (3.1)	4.1	2.2×10^{-25}
Weight post-blood draw(kg)[A]	80.9 (11.2)	69.0 (9.3)	11.9	2.9×10^{-18}
Age at blood draw (years)	31.5 (11.2)			
BMI at blood draw (kg/m^2)	24.9 (3.9)	23.8 (3.8)	1.1	1.5×10^{-6}
Weight at blood draw (kg)	74.5 (13.7)	71.3 (12.4)	3.2	4.2×10^{-5}
Height at blood draw (cm)	172.8 (8.0)	173.2 (8.0)	-0.4	0.37
Waist at blood draw (cm)	84.0 (11.2)	80.8 (10.6)	3.2	7.1×10^{-4}
Hip at blood draw (cm)	104.2 (7.1)	101.6 (7.7)	2.6	1.4×10^{-3}
WHR at blood draw (cm/cm)	0.8 (0.1)	0.8 (0.1)	0	0.26
Birth weight (g)	2638.6 (643.4)	2516.7 (707.7)	121.9	0.72
Glucose (mmol/L)	5.3 (0.5)	5.0 (0.6)	0.3	0.11
Insulin (μIU/ml)	9.1 (5.5)	8.4 (5.4)	0.7	0.16
Total Chol (mmol/L)	4.9 (1.0)	4.8 (0.9)	0.1	0.55
LDL (mmol/L)	2.9 (0.9)	2.9 (0.8)	0.0	0.78
HDL (mmol/L)	1.4 (0.4)	1.4 (0.3)	0.0	0.37
Triglycerides (mmol/L)	1.1 (0.5)	1.2 (0.6)	-0.1	0.71
CRP (mg/L)	5.7 (17.7)	2.8 (3.4)	2.9	0.72
TNFα (pg/ml)	0.9 (0.3)	1.1(1.0)	-0.2	0.32
IL-6 (pg/ml)	2.0 (3.1)	1.4 (0.9)	0.6	0.55
sIL-6R (pg/mL)	40664.0 (12019.4)	42012.5 (13394.6)	-1348.5	0.34
Fibrinogen (g/L)	2.6 (0.8)	2.5 (0.7)	0.1	0.28
AST (U/L)	20.0 (6.7)	19.1 (4.8)	0.9	0.66
ALT (U/L)	11.0 (6.7)	9.8 (3.8)	1.2	0.51
GGT (U/L)	24.2 (10.3)	22.8 (8.1)	1.4	0.39
N (%)lipid-lowering medication	1 (3%)	0 (0%)	1	0.99
N (%)diabetes medication	0(0%)	0 (0%)	0	n.a.
N (% of female twins) menopause	2 (7.7%)	1 (3.8%)	1	0.99

Characteristics of MZ twin pairs who were concordant for BMI (Δ BMI < 3) at blood draw and at all time points of participation prior to blood draw, but who became discordant (Δ BMI ≥ 3) ≥ 1 year after blood draw. The selection procedure of these twins is illustrated in figure 1f. All characteristics in this table refer to data collected at the moment of blood draw (when the twins were concordant for BMI), except for birth weight (based on data from multiple longitudinal surveys), the classification of "heavier" and "leaner" twins (which is based on the moment when twins were discordant), and variables marked with [A].

[A] Characteristics based on information collected ≥ 1year after blood draw (range 1-6, mean=3.1 years); this is the time point at which the BMI difference of these twins first passed the threshold of discordance (Δ BMI ≥ 3).

*Heavier-leaner twin

Waist= Waist circumference, Hip= Hip circumference, WHR=Waist-to-hip ratio, Total Chol=Total cholesterol, AST= aspartate aminotransferase, ALT= alanine aminotransferase, GGT= gamma glutamyl transferase. Numbers in the table represent Mean (SD) or N (%).

Biomarkers and gene expression related to prolonged BMI discordance
To examine biomarker and gene expression differences related to long-term BMI discordance, we studied a sub-group of 17 pairs who had a BMI difference ≥ 3 at all NTR projects in which they participated (3-5 time points, stretching on average 6.4 years, range 3-12 years; Figure 2b and c). These pairs showed similar differences in blood biomarkers (Table 3, last three columns), although not all effects were statistically significant in this smaller sample. For ten of the fourteen biomarkers, the effect size was larger in longitudinally discordant pairs than in the total group of pairs who were BMI discordant at blood draw (Supplemental Figure 1). This pattern was strongest for fibrinogen (Mean Difference in all discordant pairs=0.2 g/L, P=2.2 x10^{-3}; Mean Difference in longitudinally discordant pairs = 0.7 g/L, P=1.3 x 10^{-5}). Comparison of gene expression profiles (Supplemental Table 4), which were available for a subset of 9 longitudinally discordant pairs, revealed no genome-wide significant differences for individual probe sets, and no significant GO terms were found. Longitudinally discordant pairs also showed no difference in the expression of candidate genes for BMI. Effect sizes for genome-wide probe sets (mean difference in expression of the heavier twin – leaner twin) were moderately correlated with the effect sizes observed in all discordant pairs (r=0.31, p< 0.001, Supplemental Figure 2), suggesting partial correspondence of gene expression effects in all discordant pairs versus longitudinally discordant pairs.

Blood biomarkers and gene expression before onset of BMI discordance
Finally, we examined whether differences in molecular profiles precede BMI discordance, by studying 33 MZ pairs who were not yet discordant at blood draw, but who became discordant afterwards (mean=3.1 years after blood draw, range 1-6, figure1f). When first identified as discordant, the heavier and leaner twins had mean BMIs of 27.1 and 23.0, respectively, while their BMIs were on average 24.9 and 23.8, respectively, when blood samples were collected. Prior to BMI discordance, none of the blood biomarkers differed significantly (Table 4). A comparison of the effect sizes in the three groups illustrates that within-pair differences were largest for most biomarkers in the longitudinally discordant pairs while they were smallest in MZ pairs before BMI discordance (Supplemental Figure 1), suggesting that the adverse blood profile observed in heavier twins from BMI discordant pairs (Table 3) represents a consequence of the higher BMI. No significant differences in gene expression (data available for 20 pairs) were evident in MZ pairs before BMI discordance (Supplemental Table 5). No significant differences in gene expression (data available for 20 pairs) were evident in MZ pairs before BMI discordance (Supplemental Table 5). Even though not statistically significant, we explored whether genome-wide expression differences between twins before discordance (n twins pairs=20) were of comparable magnitude as the expression differences observed between twins during discordance (n twin pairs =120). We computed the correlation between effect sizes observed

before and during discordance (where effect sizes refer to the mean expression differences at 44 241 probe sets between the heavier and leaner twin). This correlation (r=-0.04, p< 0.001, Supplemental Figure 3) indicated that MZ twins do not exhibit comparable differences in gene expression prior to BMI discordance as observed during BMI discordance.

Discussion

We described longitudinal BMI data collected between 1991 and 2011 in adolescent and adult MZ twin pairs to examine the prevalence of BMI discordance (defined as ΔBMI ≥3 kg/m^2) in genetically identical individuals and its development over time, and examined possible associations of BMI discordance with discordance in lifestyle factors, biomarkers, and gene expression profiles. The majority of MZ twin pairs was highly concordant for BMI, but temporary differences in BMI were not uncommon, particularly when mean BMI increased at higher ages. However, when followed over time, only a minority of MZ twins stayed discordant for a prolonged time period. Of the pairs who were identified as discordant at any NTR project, 55.4% were no longer discordant after a period of on average 3 years, and in a group of 699 MZ twin pairs who participated ≥ 3 times in longitudinal NTR projects (including the NTR biobank), only 17 pairs (2.4%) were discordant at all time points (over a period of on average 6.5 years). These findings illustrate the difficulty to find long-term BMI discordant MZ twins and suggest that BMI discordance is generally not a stable characteristic of MZ twins. This observation carries an important message regarding the etiology of BMI: the fact that large differences in BMI in most MZ twin pairs do not last long emphasizes the important impact of genetic influences on weight regulation. Based on our findings and previous reports of convergence among initially BMI-discordant pairs [14] we conclude that genetically identical individuals who exhibit stable lifetime BMI discordance will be very rare.

The fact that (temporary) BMI discordance in MZ twins nonetheless occurs highlights the role of non-genetic influences on BMI. We found that food intake, assessed by asking each twin which of them eats most, showed the strongest difference between BMI-discordant twins of the lifestyle variables, and found no difference in self-reported frequency of leisure time exercise. This finding suggests that large BMI differences between genetically identical subjects are more strongly related to food intake than to current voluntary exercise participation. A limitation of our study is that the assessment of relative food intake, as well as other lifestyle measures and a subset of our BMI data, were based on self-report. The comparison of relative food intake among co-twins in particular might be biased by the twin's perception of their weight difference. Nonetheless, it has also been reported that self-report of twins about their relative food intake may provide more reliable information with regard to which twin eats most in comparison to self-report of absolute food intake. Thus, in a previous study of obesity-discordant twins that assessed

three measures; self-reported absolute food intake, self-reported relative food intake of twins, and measured energy turn-over (using double-labeled water technology), it was found that discordant twins showed no difference in self-reported absolute food intake (because obese twins tended to under-report their own energy intake, as suggested by data on measured energy intake). By contrast, the data on self-reported relative intake for several types of food (which twin eats most) corresponded well with predicted relative food intake of twins based on their difference in measured energy intake [53].

We found that convergence of the BMIs of initially discordant twins was related to both weight gain of the leaner twins and weight loss of the heavier twins, but an interesting question that remains to be examined by future studies is to which extent becoming BMI-discordant and converging after discordance in MZ twins is due to intentional efforts to lose or gain weight in one of the twins. Previous studies have shown that after weight loss following caloric restriction or an increase in physical exercise, most individuals tend to regain the lost weight [54, 55]. In addition to the difficulty to lose weight, "overeating" experiments have shown that most lean subjects eventually return to their original weight following diet-induced weight gain [56]. This tendency of individuals to return to a certain "set-point" of body weight has been attributed to homeostatic regulation of body fat mass, which triggers for example compensatory responses (e.g. increased appetite and energy efficiency) when the brain senses a reduction of energy stores through changes in circulating levels of adipocyte-secreted signals such as the hormone leptin [57]. Genetic regulation of this homeostatic system may explain our observation that most MZ twin pairs eventually converge to a similar BMI, after a period of discordance.

We observed that BMI discordance occurred more frequently among heavier twins. A significant relationship between the mean trait value of MZ co-twins and the difference of that trait between co-twins may reflect genotype-by-environment interaction [58]. When the impact of environmental influences on BMI depends on the genetic vulnerability of the twins, MZ twins who are highly vulnerable to the impact of environmental influences that promote weight gain (as reflected by a high mean BMI of co-twins) are expected to show the largest divergence in response to unequal environmental exposures. A second possible explanation is that the BMI of heavier persons may fluctuate more, because people often respond to weight gain by efforts to lose weight (e.g. by going on a diet). Unless such fluctuations in BMI occur at exactly the same time in MZ co-twins, they will lead to the observation of a greater percentage of discordant pairs in the higher BMI range. Finally, the larger percentage of BMI discordance among heavier twins might be related to the pathogenesis of obesity. Obesity is associated with a deterioration of homeostatic weight regulation [59]. It could thus be hypothesized that increasing variation of BMI between MZ co-twins in the higher BMI range is related to the decreasing capacity of the twins' bodies to regulate body weight after repeated weight gain.

It is well-established that obesity is associated with adverse changes in blood levels of biomarkers that are reflective of an increased risk of developing cardiovascular disease and type 2 diabetes. These markers include dysregulated blood lipid levels and glucose homeostasis, and a pro-inflammatory state. Studying variation in these markers in BMI-discordant twins has the advantage that genetic pleiotropy is ruled out as a potential explanation for the association by design. If the association between a higher BMI and adverse blood biomarkers in the population would solely exist because genetic variants that predispose to a high BMI also cause the adverse changes in biomarkers, MZ twins who are discordant for BMI should show equal levels of these biomarkers because MZ twins have the same genetic vulnerability. However, we found that MZ twin pairs who were discordant at the moment of blood draw exhibited significant differences in all metabolic biomarkers (with the heavier twin having an unfavorable metabolic profile) and heavier twins had significantly higher blood levels of IL-6, sIL-6R, CRP and GGT. Effect sizes ranged from a 0.14 standard deviation increase in IL-6 levels in the heavier twins to a 1.1 SD increase in insulin. Differences in biomarkers were more pronounced in twins with longitudinal evidence for discordance with effect sizes ranging up to a 2.3 SD increase in CRP in heavier twins of longitudinally discordant pairs. To illustrate the meaning of these differences, we assessed whether individuals had elevated fasting levels of triglycerides or glucose, or reduced levels of HDL cholesterol – according to the revised criteria from the National Cholesterol Education Program's Adult Treatment Panel III (NCEP ATP III) [60]. Of all discordant pairs, 35.7% of heavier twins versus 22.4% of leaner twins had elevated fasting glucose levels (or used glucose-lowering medication), 31.4% of heavier twins versus 16.0% of leaner twins had high triglyceride levels (or used lipid-lowering medication) and 40.8% of heavier twins versus 30.8% of leaner twins had low HDL levels (or used lipid-lowering medication). Of the longitudinally discordant pairs, the percentages were as follows: elevated glucose; 41.2% of heavier twins and 12.5% of leaner twins, elevated triglycerides; 37.5% of heavier twins and 6.7% of leaner twins, low HDL levels; 43.8% of heavier twins and 26.7% of leaner twins. Importantly, MZ pairs who became discordant after blood draw showed no significant differences in biomarker levels prior to BMI discordance. This pattern suggests that adverse changes in these biomarkers are caused by a change in weight and worsen over time as a consequence of larger adiposity in the heavier twins from discordant MZ pairs. The fact that no differences in biological markers were present in genetically identical subjects before their weight difference emerged highlights that these biomarkers are not predictive of BMI but that a high BMI is driving unfavorable changes in these biomarkers.

While MZ twins share the same genome, it is possible that, in addition to differences in lifestyle, differences in the regulation of genes between twins may contribute to differences in their BMIs. Although we noticed significant enrichment of various Gene Ontology terms among genes with a stronger

expression difference in 120 discordant pairs, highlighting metabolic regulation and processes that have been previously linked to the pathogenesis of obesity (e.g. type I interferon signaling [61] and NLRP3 inflammasome complex [62]), we found no statistically significant associations for the expression level of individual probe sets with BMI discordance. We next zoomed in to a set of candidate genes for BMI that were identified through genome-wide association analysis of BMI, which implies that genetic variants (single nucleotide polymorphisms) in or nearby these genes are associated with variation in BMI in the population. Since genetic variation at these loci is shared within MZ pairs, differences in expression between BMI-discordant twins would indicate a role of environmental influences or epigenetic mechanisms in the regulation of these genes. Yet, we did not find differences in the expression of these well-established BMI loci, thus finding no evidence for BMI discordance being the result of differential gene regulation, however, we only studied gene expression in peripheral blood, and it is possible that causal regulatory pathways underlying BMI discordance are confined to other tissues. Of note, previous studies have reported differences in the expression of candidate genes related to e.g. lipid metabolism in peripheral blood between overweight versus normal weight children [63] and between at-risk obese versus metabolically healthy obese adult individuals [64].

In conclusion, the prevalence of BMI discordance in MZ pairs is low and ranged from 3.2% in 1991 (mean age=17, SD=2.4) to 17.4% in 2009 (mean age=35, SD=15). Only 2.4% of MZ pairs showed a stable long-term difference in BMI (3-5 time points, average range 6.5 years). Comparing the heavier and leaner twins, we found significant differences in self-reported food intake relative to the co-twin, and found clinically meaningful differences in metabolic and inflammatory biomarkers that arise as a consequence of the difference in adiposity and that are more pronounced in long-term discordant pairs. Other lifestyle variables such as smoking, voluntary exercise and gene expression did not differ between the heavier and leaner twins.

Reference List

1. van Dongen,J., Slagboom,P.E., Draisma,H.H., Martin,N.G., & Boomsma,D.I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **13**, 640-653 (2012).
2. Zwijnenburg,P.J.G., Meijers Heijboer,H., & Boomsma,D.I. Identical but not the same: The value of discordant monozygotic twins in genetic research. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153**, 1134-1149 (2010).
3. Halfvarson,J. *et al.* Environmental factors in inflammatory bowel disease: a co-twin control study of a Swedish-Danish twin population. *Inflamm. Bowel. Dis.* **12**, 925-933 (2006).
4. Lehn,H. *et al.* Attention problems and attention-deficit/hyperactivity disorder in discordant and concordant monozygotic twins: evidence of environmental mediators. *J. Am. Acad. Child Adolesc. Psychiatry* **46**, 83-91 (2007).
5. Vadlamudi,L. *et al.* Timing of de novo mutagenesis--a twin study of sodium-channel mutations. *N. Engl. J. Med.* **363**, 1335-1340 (2010).
6. Wong,C.C. *et al.* A longitudinal twin study of skewed X chromosome-inactivation. *PLoS One.* **6**, e17873 (2011).
7. Weksberg,R. *et al.* Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Hum. Mol. Genet.* **11**, 1317-1325 (2002).
8. Schousboe,K. *et al.* Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. *Twin Research* **6**, 409-421 (2003).
9. Mook-Kanamori,D.O. *et al.* Heritability estimates of body size in fetal life and early childhood. *PLoS One.* **7**, e39901 (2012).
10. Elks,C.E. *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol. (Lausanne)* **3**, 29 (2012).
11. Hardy,R. *et al.* Life course variations in the associations between FTO and MC4R gene variants and body size. *Hum. Mol. Genet.* **19**, 545-552 (2010).
12. Ye,K. *et al.* Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin. Res. Hum. Genet.* **16**, 1026-1032 (2013).
13. Naukkarinen,J., Rissanen,A., Kaprio,J., & Pietilainen,K.H. Causes and consequences of obesity: the contribution of recent twin studies. *Int. J. Obes. (Lond)* **36**, 1017-1024 (2012).
14. Graner,M. *et al.* Epicardial fat, cardiac dimensions, and low-grade inflammation in young adult monozygotic twins discordant for obesity. *Am. J. Cardiol.* **109**, 1295-1302 (2012).
15. Hakala,P., Rissanen,A., Koskenvuo,M., Kaprio,J., & Ronnemaa,T. Environmental factors in the development of obesity in identical twins. *Int. J. Obes. Relat Metab Disord.* **23**, 746-753 (1999).
16. Kannisto,K. *et al.* Overexpression of 11beta-hydroxysteroid dehydrogenase-1 in adipose tissue is associated with acquired obesity and features of insulin resistance: studies in young adult monozygotic twins. *J. Clin. Endocrinol. Metab* **89**, 4414-4421 (2004).
17. Kaye,S.M. *et al.* Obesity-related derangements of coagulation and fibrinolysis: a study of obesity-discordant monozygotic twin pairs. *Obesity (Silver. Spring)* **20**, 88-94 (2012).

18. Mustelin,L. *et al.* Acquired obesity and poor physical fitness impair expression of genes of mitochondrial oxidative phosphorylation in monozygotic twins discordant for obesity. *Am. J. Physiol Endocrinol. Metab* **295**, E148-E154 (2008).
19. Pietilainen,K.H. *et al.* Acquired obesity is associated with increased liver fat, intra-abdominal fat, and insulin resistance in young adult monozygotic twins. *Am. J. Physiol Endocrinol. Metab* **288**, E768-E774 (2005).
20. Pietilainen,K.H. *et al.* Acquired obesity increases CD68 and tumor necrosis factor-alpha and decreases adiponectin gene expression in adipose tissue: a study in monozygotic twins. *J. Clin. Endocrinol. Metab* **91**, 2776-2781 (2006).
21. Pietilainen,K.H. *et al.* Effects of acquired obesity on endothelial function in monozygotic twins. *Obesity (Silver. Spring)* **14**, 826-837 (2006).
22. Pietilainen,K.H. *et al.* Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects--a monozygotic twin study. *PLoS One.* **2**, e218 (2007).
23. Pietilainen,K.H. *et al.* Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity. *PLoS Med.* **5**, e51 (2008).
24. Pietiläinen,K.H. *et al.* Growth patterns in young adult monozygotic twin pairs discordant and concordant for obesity. *Twin Res.* **7**, 421-429 (2004).
25. Rissanen,A. *et al.* Acquired preference especially for dietary fat and obesity: a study of weight-discordant monozygotic twin pairs. *Int. J. Obes. Relat Metab Disord.* **26**, 973-977 (2002).
26. Mitchell,K.S. *et al.* Characteristics of monozygotic male and female twins discordant for overweight: a descriptive study. *Eat. Behav.* **9**, 366-369 (2008).
27. Souren,N.Y. *et al.* DNA methylation variability at growth-related imprints does not contribute to overweight in monozygotic twins discordant for BMI. *Obesity (Silver. Spring)* **19**, 1519-1522 (2011).
28. Dandona,P., Aljada,A., Chaudhuri,A., Mohanty,P., & Garg,R. Metabolic syndrome: a comprehensive perspective based on interactions between obesity, diabetes, and inflammation. *Circulation* **111**, 1448-1454 (2005).
29. Dantzer,R., O'Connor,J.C., Freund,G.G., Johnson,R.W., & Kelley,K.W. From inflammation to sickness and depression: when the immune system subjugates the brain. *Nat. Rev. Neurosci.* **9**, 46-56 (2008).
30. Festa,A., D'Agostino,R., Jr., Tracy,R.P., & Haffner,S.M. Elevated levels of acute-phase proteins and plasminogen activator inhibitor-1 predict the development of type 2 diabetes: the insulin resistance atherosclerosis study. *Diabetes* **51**, 1131-1137 (2002).
31. Hansson,G.K. Inflammation, atherosclerosis, and coronary artery disease. *N. Engl. J. Med.* **352**, 1685-1695 (2005).
32. Hotamisligil,G.S., Shargill,N.S., & Spiegelman,B.M. Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. *Science* **259**, 87-91 (1993).
33. Miller,A.H., Maletic,V., & Raison,C.L. Inflammation and its discontents: the role of cytokines in the pathophysiology of major depression. *Biol. Psychiatry* **65**, 732-741 (2009).
34. Nomiyama,T. *et al.* Osteopontin mediates obesity-induced adipose tissue macrophage infiltration and insulin resistance in mice. *J. Clin. Invest* **117**, 2877-2888 (2007).
35. Hotamisligil,G.S. Endoplasmic reticulum stress and the inflammatory basis of metabolic disease. *Cell* **140**, 900-917 (2010).

36. Ozcan,U. *et al.* Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* **306**, 457-461 (2004).
37. Stienstra,R., Tack,C.J., Kanneganti,T.D., Joosten,L.A., & Netea,M.G. The inflammasome puts obesity in the danger zone. *Cell Metab* **15**, 10-18 (2012).
38. Strowig,T., Henao-Mejia,J., Elinav,E., & Flavell,R. Inflammasomes in health and disease. *Nature* **481**, 278-286 (2012).
39. Srinivasan,S.R., Bao,W., Wattigney,W.A., & Berenson,G.S. Adolescent overweight is associated with adult overweight and related multiple cardiovascular risk factors: the Bogalusa Heart Study. *Metabolism* **45**, 235-240 (1996).
40. Elder,S.J. *et al.* Genetic and environmental influences on factors associated with cardiovascular disease and the metabolic syndrome. *J. Lipid Res.* **50**, 1917-1926 (2009).
41. Freathy,R.M. *et al.* Type 2 diabetes risk alleles are associated with reduced size at birth. *Diabetes* **58**, 1428-1433 (2009).
42. Horikoshi,M. *et al.* New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat. Genet.* **45**, 76-82 (2013).
43. Boomsma,D.I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* **9**, 849-857 (2006).
44. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* **13**, 231-245 (2010).
45. Willemsen,G. *et al.* The adult Netherlands twin register: twenty-five years of survey and biological data collection. *Twin. Res. Hum. Genet.* **16**, 271-281 (2013).
46. Jansen,R. *et al.* Sex differences in the human peripheral blood transcriptome. *BMC. Genomics* **15**, 33 (2014).
47. Wright,F.A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430-437 (2014).
48. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 2013.
49. Speliotes,E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937-948 (2010).
50. Eden,E., Navon,R., Steinfeld,I., Lipson,D., & Yakhini,Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC. Bioinformatics.* **10**, 48 (2009).
51. Mootha,V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267-273 (2003).
52. Subramanian,A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A* **102**, 15545-15550 (2005).
53. Pietilainen,K.H. *et al.* Inaccuracies in food and physical activity diaries of obese subjects: complementary evidence from doubly labeled water and co-twin assessments. *Int. J. Obes. (Lond)* **34**, 437-445 (2010).
54. Knowler,W.C. *et al.* 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* **374**, 1677-1686 (2009).
55. Safer,D.J. Diet, behavior modification, and exercise: a review of obesity treatments from a long-term perspective. *South. Med. J.* **84**, 1470-1474 (1991).

56. Sims,E.A. & Horton,E.S. Endocrine and metabolic adaptation to obesity and starvation. *Am. J. Clin. Nutr.* **21**, 1455-1470 (1968).
57. Rosen,E.D. & Spiegelman,B.M. Adipocytes as regulators of energy balance and glucose homeostasis. *Nature* **444**, 847-853 (2006).
58. Jinks,J.L. & Fulker,D.W. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of the human behavior. *Psychol. Bull.* **73**, 311-349 (1970).
59. Hussain,S.S. & Bloom,S.R. The regulation of food intake by the gut-brain axis: implications for obesity. *Int. J. Obes. (Lond)* **37**, 625-633 (2013).
60. Alberti,K.G. *et al.* Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation* **120**, 1640-1645 (2009).
61. Wang,X.A. *et al.* Interferon regulatory factor 7 deficiency prevents diet-induced obesity and insulin resistance. *Am. J. Physiol Endocrinol. Metab* **305**, E485-E495 (2013).
62. Vandanmagsar,B. *et al.* The NLRP3 inflammasome instigates obesity-induced inflammation and insulin resistance. *Nat. Med.* **17**, 179-188 (2011).
63. Sanchez,J. *et al.* Blood cells as a source of transcriptional biomarkers of childhood obesity and its related metabolic alterations: results of the IDEFICS study. *J. Clin. Endocrinol. Metab* **97**, E648-E652 (2012).
64. Telle-Hansen,V.H. *et al.* Altered expression of genes involved in lipid metabolism in obese subjects with unfavourable phenotype. *Genes Nutr.* **8**, 425-434 (2013).

Part 3 Beyond DNA sequence: Epigenetic variation

Chapter 6

The heritability of DNA methylation in peripheral blood: influences of common SNPs and variability of genetic and environmental variance with age and sex

Abstract

Heritable and environmentally mediated epigenetic variation between individuals may be an important contributor to individual differences in disease susceptibility. We estimated the genetic and environmental variance of DNA methylation level at genome-wide sites in whole blood taking advantage of the resemblance of monozygotic (MZ) twin pairs (N=769) and dizygotic (DZ) twin pairs (N=424) with the classical twin method, and by using estimates of identity-by-descent (IBD) among closely related and distantly related individuals (N=2603, mean age=37.2 years (SD=13.3), 62% females) derived from genome-wide SNPs. The average heritability across genome-wide methylation sites was 19%, of which on average 37% was explained by genome-wide SNPs (MAF > 0.01). Polygenic genotype X sex and genotype X age interaction analysis identified 2654 methylation sites with evidence for sex-specific heritability and 39194 sites with age-specific heritability, and revealed that the between-individual variance due to environmental or stochastic influences increased with age at a large proportion of these sites. The question remains whether these sites affect age and sex-dependent disease susceptibility.

This chapter is based on: Jenny van Dongen [*], Bastiaan T. Heijmans[*], Michel G. Nivard[*], Gonneke Willemsen, Jouke-Jan Hottenga, Quinta Helmer, Conor V. Dolan, Erik A. Ehli, Gareth Davies, BIOS Consortium[iii], H. Eka Suchiman, Rick Jansen, Joyce B. van Meurs, P. Eline Slagboom, Dorret I. Boomsma. The heritability of DNA methylation in peripheral blood: influences of common SNPs and variability of genetic and environmental variance with age and sex. (*manuscript in preparation*)

* These authors contributed equally to the work.

[iii] The Biobank-based Integrative Omics Study (BIOS) Consortium

Introduction

Epigenetic variation between individuals may represent an important contributor to individual differences in disease susceptibility[1]. Of the many epigenetic marks and mechanisms that exist[2], DNA methylation is thought to contribute to stable long-term gene expression regulation and tissue-differentiation[3, 4], and is currently the only that can be assessed at a genome-wide scale in large human epidemiological studies. Epigenome-Wide Association Studies (EWAS) in which a trait, disease or exposure is tested for association with DNA methylation show promising results[5-10]. DNA methylation variation between individuals may result from environmental and stochastic variation or from genetic influences (due to variation in the DNA sequence). Increasing evidence suggests that DNA-sequence-mediated epigenetic variation between individuals contributes to human disease susceptibility[8, 11-13]. Notably, methylation differences have been observed between the sexes[14-16] and across age[17-20], suggesting that epigenetic regulation may also be involved in the widely observed age and sex differences in the aetiology of complex diseases[21, 22]. Studies of prenatal dietary exposures illustrate that the impact of environmental influences on DNA methylation may depend on the sex of the individual[23, 24], while sex-specific SNP effects on DNA methylation have also been described[25].

It is well-known that genetically identical model organisms such as cloned animals[26], isogenic plants[27] and inbred mice[28] exhibit epigenetic and phenotypic differences. Human identical twins provide insight into the extent to which epigenetic regulation in humans may vary due to environmental and stochastic influences. The overall contribution of genetic and environmental differences to variation in DNA methylation between individuals in a population can be estimated by contrasting the correlation between DNA methylation levels of monozygotic (MZ) and dizygotic (DZ) twins, who share 100% and 50% of segregating genetic variants that contribute to methylation differences, respectively (the classical twin design). The average twin-based heritability of DNA methylation across genome-wide CpGs varies between 5% and 19% for different tissues[17, 29-31]. MZ twins already show differences in DNA methylation at birth[31, 32] and some studies have indicated that certain epigenetic marks including DNA methylation at specific loci may diverge in twin pairs with ageing[33, 34], although evidence for such effects is not always seen[35]. Changes in the heritability of DNA methylation with age or between the sexes can be assessed by polygenic genotype X age (or sex) interaction analysis[36].

Here, we report on the largest twin study to date of genome-wide DNA methylation (Illumina 450k array) in whole blood. The study design allows the estimation of the heritability of DNA methylation based on the classical twin method and based on measured genetic relationships between individuals (identity-by-descent, IBD), and to estimate the variance of DNA methylation explained by common SNPs. To obtain insight into potential differences in epigenetic regulation between the sexes and across age, we investigated

polygenic genotype X sex and genotype X age interaction effects on DNA methylation at genome-wide sites. To examine the stability of DNA methylation in blood and the correlation of DNA methylation across two accessible tissues that are suited for human epidemiological studies (blood and buccal), longitudinal and cross-tissue analyses were performed using data from a small subset of subjects.

Materials and Methods

Subjects and samples

The subjects in this study participated in the Netherlands Twin Register (NTR)[37, 38] . In the biobank project, venous blood samples were drawn in the morning after an overnight fast, and separate EDTA tubes were collected for isolation of DNA and assessment of haematological profiles. Blood sampling and buccal sample collection procedures were described in detail previously[39].

Most subjects were twins, but the sample also included parents of twins, siblings of twins or spouses of twins, as described in detail in Appendix 4. In total, 3264 blood samples from 3221 NTR participants were assessed for genome-wide methylation, of which 3089 samples from 3057 subjects passed quality control. Only samples with good quality DNA methylation data and for which data on white blood cell counts were available were kept in the analyses, leaving 3006 samples from 2975 subjects. This dataset included 769 MZ and 424 DZ pairs, and for 31 subjects longitudinal methylation data were available (two time points). All analyses that included genome-wide SNP data were performed on data from a subset of subjects who were genotyped and who were of Dutch origin (N=2603).

For a small subset of 11 MZ pairs (male pairs=3, female pairs=8, age: 18 years), genome-wide methylation data were available for two types of samples: blood (as described above) and buccal. The buccal samples from 10 twins were assessed in 2013, as described by van Dongen *et al*[40]. The 12 additional buccal samples were assessed using the same protocol in 2014. Buccal and blood samples were collected shortly after each other.

All subjects provided written informed consent and study protocols were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180).

Cell counts

The following subtypes of white blood cells were counted in blood samples: neutrophils, lymphocytes, monocytes, eosinophils, and basophils (see Willemsen et al[39]). Lymphocyte and neutrophil percentages were strongly negatively correlated (r=-0.93). Of these two white blood cell subtypes, the percentage of neutrophils showed the strongest correlation with DNA

methylation levels (as evidenced by the correlation with Principal Components (PCs) from the raw genome-wide methylation data). Basophil percentage showed little variation between subjects, with a large number of subjects having 0% of basophils. Therefore, the percentages of neutrophils, monocytes and eosinophils were used to adjust DNA methylation data for inter-individual variation in white blood cell proportions.

Genome-wide SNP data
Genome-wide SNPs were used to construct a Genetic Relationship Matrix (GRM), which summarizes overall genetic relatedness between all subjects (N=2603), based on all genotyped autosomal SNPs (MAF > 0.01) with the software program Genome-wide Complex Trait Analysis (GCTA)[41]. Details on genome-wide SNP data are provided in the Supplementary Methods.

Infinium HumanMethylation450 BeadChip data
DNA methylation was assessed with the Infinium HumanMethylation450 BeadChip Kit (Illumina, Inc.)[42]. 500ng of genomic DNA from whole blood was bisulfite treated using the Zymo Research 96-well plate using the standard protocol for Illumina 450K micro-arrays, by the department of Molecular Epidemiology from the Leiden University Medical Center (LUMC), The Netherlands. Subsequent steps (i.e sample hybridization, staining, scanning) were performed by the Erasmus Medical Center micro-array facility, Rotterdam, The Netherlands. Quality control and processing of the blood methylation dataset is described in detail in the supplementary methods. In short, a number of sample-level and probe level quality checks were performed. Probes were set to missing in a sample if they had an intensity value of exactly zero, detection *P*-value > 0.01, or bead count < 3. Probes were excluded from all samples if they mapped to multiple locations in the genome[43], had a SNP within the CpG site (at the C or G position) irrespective of minor allele frequency in the Dutch population[44], or if they had a success rate < 0.95 across samples. Only autosomal sites were kept in the analyses. Blood methylation data were normalized with Functional Normalization[45], and normalized intensity values were converted into beta-values (β) and M-values[46]; β -values were used for descriptive purposes only because of their biological interpretability, while M-values were used as input for all analyses. The DNA methylation protocol and data processing of the buccal samples have been described previously [40].

Heritable and environmental influences on DNA methylation levels in blood
To facilitate computations, missing methylation data (0.04%-2.14% of genome-wide probes per individual, mean=0.1%) were imputed with the R package impute. Prior to analysis, the normalized methylation M-values were corrected for sex, age, array row, 96-wells plate (dummy coded), white blood cell percentages (neutrophils, monocytes and eosinophils; assessed at sample collection), and the first ten PCs derived from the genotype data, with the lm

function in R. All analyses that included genome-wide SNP data were performed on the residuals derived after correcting for these covariates. All other analyses (i.e. twin correlations, longitudinal analyses and blood-buccal comparison) were performed on the residuals derived after correcting for the before mentioned covariates minus the genotype PCs.

The proportion of variance in DNA methylation attributable to total additive genetic effects (h^2), the proportion attributable to the additive effects of all measured SNPs (h^2_{SNPs}), and interactions of total additive genetic effects and environmental effects with age and sex were assessed by the classical twin model, and by modeling the DNA methylation data as a function of measured genetic relatedness between subjects. At each CpG site ($CpGi$), the classical twin heritability was computed as:

$h^2(CpGi)$= 2 * (*r*MZ - *r*DZ),

Where *r*MZ and *r*DZ are the correlations of DNA methylation level at one CpG site between the MZ, and between the DZ twins, respectively.

In all other analyses, h^2 was estimated by fitting a linear mixed model in which the covariance of DNA methylation between individuals was modeled as a function of measured genetic relationships based on SNP data. These methods are described in detail by Nivard *et al* (*Nivard M.G., Middeldorp C.M., Lubke G., Hottenga J.J., Abdellaoui A., Boomsma D.I., Dolan C.V. Continuous gene – environment interaction in attention problems, anxious depression, body mass index and height leveraging twin, pedigree and genotype data. manuscript in preparation*). In short, the approach outlined by Zaitlen *et al*[47] was applied, which allows for (simultaneous) estimation of h^2 and h^2_{SNPs} in study samples that include both closely and distantly related individuals. The method makes use of two GRMs: a GRM describing the relationships between all individuals ($\mathrm{GRM}^{\mathrm{IBS}}_{\mathrm{n*n}}$) and a second GRM in which all genetic relationships < 0.05 IBS (distant genetic relationships) are set to zero ($GRM^{IBS>0.05}_{n*n}$), making the estimates of genetic relatedness equivalent to the proportion in the genome shared identity-by-descent (IBD), as explained by Zaitlen *et al*[47]. For each CpG, we jointly estimated the total additive genetic variance (σ^2_{IBD}) and the variance explained by genome-wide SNPs (σ^2_{SNPs}) as follows:

$$var(CpGi) = GRM^{IBS}_{n*n} \otimes \sigma^2_{SNPs} + GRM^{IBS>0.05}_{n*n} \otimes (\sigma^2_{IBD} - \sigma^2_{SNPs}) + I_{n*n} \otimes \sigma^2_e,$$

Where $\mathrm{var(CpGi)}$ is the variance of DNA methylation at CpG i, adjusted for covariates, σ^2_{SNPs} is the variance explained by all SNPs, the term ($\sigma^2_{IBD} - \sigma^2_{SNPs}$) denotes the difference between the total genetic variance and the variance explained by SNPs, and σ^2_e reflects the variance attributable to residual effects (“unique environment”, which may include environmental

influences unique to each individual, stochastic influences and measurement error). The total heritability (h^2_{IBD}) was calculated as: $h^2_{IBD}=\sigma^2_{IBD}/(\sigma^2_{IBD}+\sigma^2_e)$. The proportion of variance explained by genome-wide SNPs was calculated as: $h^2_{SNPs}=\sigma^2_{SNPs}/(\sigma^2_{IBD}+\sigma^2_e)$ and the proportion of the heritability explained by SNPs was calculated as: $h^2_{SNPs}/ h^2_{IBD.}$

Polygenic genotype X sex interaction effects on DNA methylation were investigated with the following model:

$$var(CpGi) = GRM^{IBS>0.05}_{n*n} \otimes \left(\sigma_{IBD} + \beta_{IBD-sex} * Sex\right)^2 + I_{n*n} \otimes \left(\sigma_e + \beta_{e-sex} * Sex\right)^2,$$

Where $\beta_{IBD-sex}$= regression coefficient for the interaction of genetic variance with sex (coded as 0/1), and β_{e-sex}= regression coefficient for the interaction of residual variance with sex (coded as 0/1). This parameterization of the interaction effect is equivalent to how polygenic genotype-by-environment interaction is commonly assessed within the classical twin model, as proposed by Purcell[36].

Polygenic genotype X age interaction effects on DNA methylation were investigated with the following model:

$$\mathrm{var(CpGi)} = \mathrm{GRM}^{\mathrm{IBS>0.05}}_{\mathrm{n*n}} \otimes \left(\sigma_{\mathrm{IBD}} + \beta_{\mathrm{IBD\text{-}Age}} * \mathrm{Age}\right)^2 + \mathrm{I}_{\mathrm{n*n}} \otimes \left(\sigma_{\mathrm{e}} + \beta_{\mathrm{e\text{-}Age}} * \mathrm{Age}\right)^2,$$

Where $\beta_{IBD-Age}$= regression coefficient for the interaction of genetic variance with age (z-transformed), and β_{e-Age}= regression coefficient for the interaction of residual variance with age (z-transformed).

Prior to the analyses based on genome-wide SNP data, methylation data were standardized (z-transformation) to facilitate computations. A small proportion of CpGs for which a model did not run successfully were discarded (see results). The p-values of each of the four interaction effects (genetic and environmental variance by age and sex) were derived with a chi-square test (1 degrees of freedom), where X^2= (beta/se)^2. Statistical significance of interaction p-values was assessed after Bonferroni correction for the number of CpGs for which estimates were successfully obtained. The correspondence between twin-based heritability and heritability estimated on the basis of actual IBD was evaluated by computing the correlation between the value of h^2 for all CpGs based on the classical twin approach and the corresponding value of h^2_{IBD}.

Longitudinal correlation and correlation between blood and buccal methylation

Data from individuals for whom two longitudinal blood samples were collected were used to calculate the correlation between DNA methylation level at time point 1 and DNA methylation level at time point 2 for each CpG site. After

obtaining an estimate of heritability and a longitudinal correlation for each CpG, the correlation between genome-wide estimates of (twin-based) heritability and genome-wide estimates of the longitudinal correlation was estimated to examine the relationship between longitudinal stability and the heritability of DNA methylation level. Data from individuals with 450k methylation data from blood samples and buccal samples were used to calculate the correlation between DNA methylation level in blood and buccal for each CpG. Prior to this analysis, the buccal methylation data (M-values) were corrected for sex, array row and assessment batch (2 levels). Blood-buccal correlations for all CpGs were correlated with the twin-based estimate of h^2 to examine the relationship between the heritability in blood and the extent to which between-individual variation in DNA methylation level is shared across tissues.

Table 1: Characteristics of the subjects.

Sub-group/Analysis	N[A]	Mean age (SD), range	Sex	Interval (years)[C]
MZ twin pairs	769	36.1 (12.4), 18-78	F: 541, M:228	
DZ twin pairs	424	33.9 (10.5), 17-79	FF: 180, MM: 93, FM:151	
Subjects with genome-wide SNP data	2603	37.2 (13.3),17-79	F: 1613, M:889	
Subjects with longitudinal methylation data	31	34.4 (6.1),26-50[B]	F: 24, M:7	5.2 (1.1), 2-7
Subjects with blood and buccal methylation data	22	18[D]	F:16, M:6	

F= Female, M=Male. [A]Number of subjects or twin pairs. [B]Age at first blood sample collection. [C]Time between first and second blood sample. [D]All subjects were 18 years when blood and buccal samples were collected.

Results

Characteristics of the study sample

Genome-wide methylation data were available for 769 MZ twin pairs and 424 DZ twin pairs. The age and sex of subjects is described in Table 1. The analyses with genome-wide SNPs were performed on data from 2603 individuals, including twins (N=2373), parents of twins (N=212), siblings of twins (N=16) and spouses of twins (N=3). Longitudinal data on DNA methylation in peripheral blood samples collected with an interval of on average 5 years were available for 31 individuals and data from blood and buccal samples were available for 22 individuals. Methylation levels at 411169 autosomal CpGs were analyzed. Prior to analyses, DNA methylation levels

were adjusted for age at sample collection, sex, white blood cell percentages, principal components (PCs) from the genotype data, and technical aspects as described in the methods. Together, this set of predictors explained on average 17% of the variance in DNA methylation for genome-wide sites (Supplementary Figure 1).

Table 2 Twin correlations, classical twin heritability, heritability based on IBD, and variance explained by SNPs for DNA methylation level at genome-wide CpGs.

Classical twin approach					Heritability based on GRM				
Parameter	Min	Median	Mean	Max	Parameter	Min	Median	Mean	Max
*r*MZ	-0.14	0.12	0.20	0.99	h_{total}^2	0.00	0.12	0.19	0.99
*r*DZ	-0.25	0.06	0.09	0.89	h_{SNPs}^2	0.00	0.01	0.07	0.98
h^2	-1.56	0.16	0.22	1.65	h_{SNPs}^2 / h_{total}^2	0.00	0.22	0.37	1.00

*r*MZ= Correlation between DNA methylation levels of monozygotic (MZ) twins. *r*DZ= Correlation between DNA methylation levels of dizygotic (DZ) twins. h^2= Heritability of DNA methylation level based on the classical twin method. h_{total}^2= Total IBD-based heritability of DNA methylation level. h_{SNPs}^2= Proportion of variance of DNA methylation explained by genome-wide SNPs (MAF > 0.01). GRM= Genetic Relationship Matrix.

Heritable and environmental influences on DNA methylation

On average across genome-wide CpGs, the pattern of twin correlations (Table 2) suggests that additive genetic influences account for the resemblance of twins for DNA methylation level; the average correlation in MZ twins (*r*=0.20) was approximately twice as large as the average correlation in DZ twins (*r*=0.09). We estimated h^2, the proportion of variance in DNA methylation that can be explained by total additive genetic effects, at each CpG as twice the difference between the MZ and DZ twin correlation. The average heritability estimated by this approach was 0.22. These results are similar to the twin correlations and h^2 based on 450k methylation from peripheral blood published previously[17, 29]. Most CpGs showed little inter-individual variation in DNA methylation (Supplementary Figure 2). The twin correlations and h^2 were on average larger at CpGs with a larger variance (Supplementary Table 1). For example, at the most variable CpGs (defined here as CpGs with a SD of the beta-value >= 0.05, N=28078 CpGs), the average twin-based h^2 was 0.64.

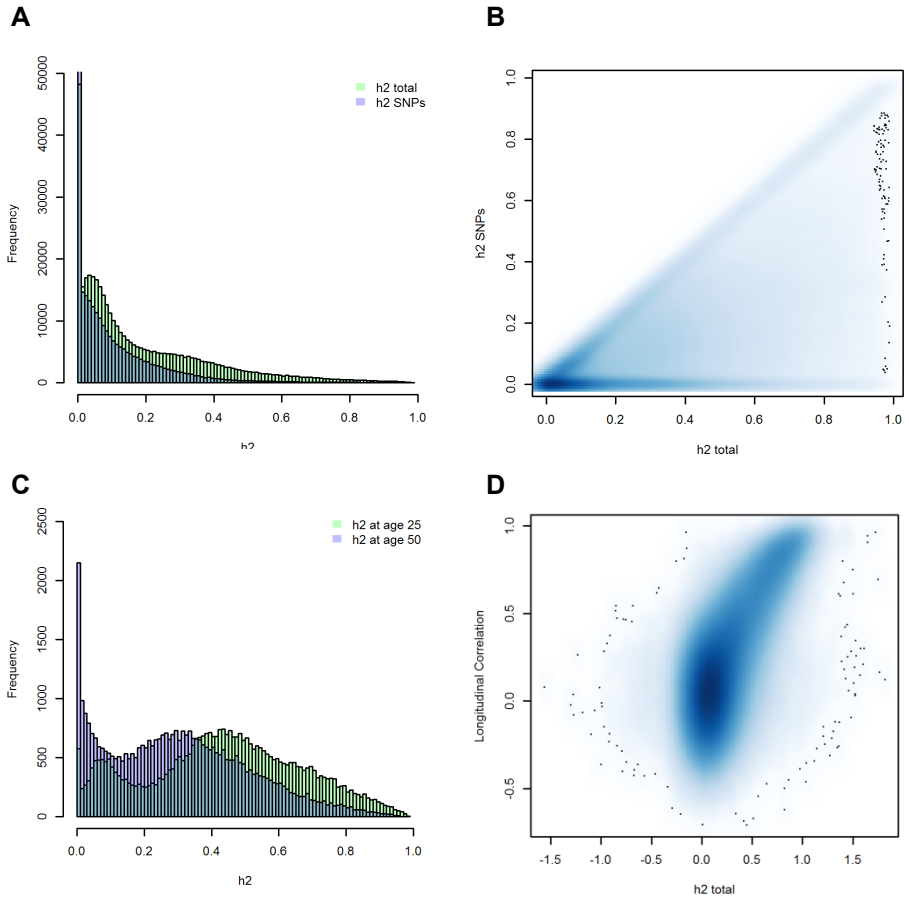
Another approach to estimate the heritability of a trait is to model the trait covariance between subjects as a function of the actual proportion of the genome that they share identity-by-descent (IBD)[48], which can be estimated for example based on genome-wide SNPs. Genome-wide SNP data also allow for estimating the proportion of variation of a trait that can be explained by all

genotyped common SNPS (h^2_{SNPs})[49]. We estimated h^2 and h^2_{SNPs} simultaneously by fitting linear mixed models with two Genetic Relationship matrices (GRMs) to the methylation data of all subjects; one GRM describing genome-wide IBD relationships between closely related (including MZ and DZ twins and their family members) and distantly related individuals (among which IBD is 0), and one GRM describing the genetic relationships between all individuals, respectively[47]. Estimates were successfully obtained for 405010 CpGsites (98.5%). The genome-wide average IBD-based heritability of DNA methylation (h^2=0.19) was similar to the classical twin estimate of h^2, and the estimates from the two methods were strongly correlated (*r*=0.83).

Common SNPs explain a large portion of the heritability of DNA methylation
The average h^2_{SNPs} across all CpGs was 0.07 (Figure 1A), and the proportion of total heritability explained by SNPs (h^2_{SNPs}/ h^2) was on average 0.37 (Table 2). At a relatively large proportion of CpGs, the proportion of total heritability that was explained by SNPs was (almost) 0% or (almost) 100% (CpGs with h^2_{SNPs}/ h^2 < 0.01: 158367 (39%), CpGs with h^2_{SNPs}/ h^2> 0.99=73820 (18%), Figure 1B). These findings indicate that for many CpGs, a relatively large proportion of total estimated heritable variation in DNA methylation can be explained by common SNP effects but there is also a substantial number of CpG sites in the genome where DNA methylation is heritable but cannot be explained by common SNPs. In total, there were 199340 CpG sites (49%) where common SNPs explained at least 1% of the variance of DNA methylation level, and 356741 CpGs (88%) with a total (IBD-based) heritability of at least 1%.

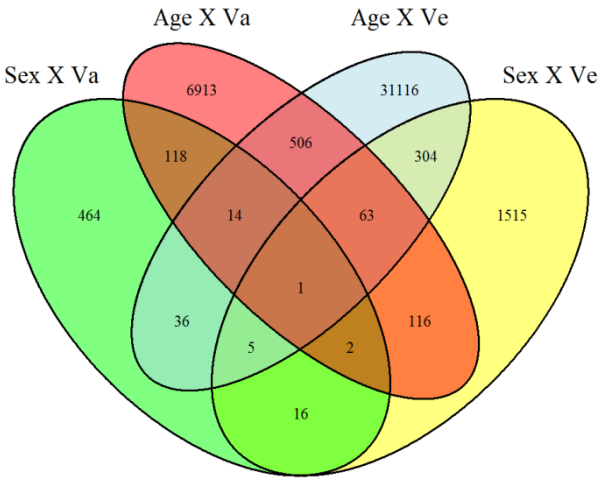
Variation of genetic and environmental influences on DNA methylation by gender and with age
To assess sex differences in the total heritability of DNA methylation, the IBD-based model was extended to allow for interaction between the total genetic variance and sex and interaction between unique environmental variance and sex. Sex interaction models ran successfully for 388950 CpGs (95%). At the genome-wide level, the average heritability was nearly identical in males (mean h^2=0.199, median=0.13) and females (mean h^2=0.198, median=0.13). Significant interaction between sex and the genetic or environmental variance was evident at 2654 CpGs. Of CpGs with significant sex interaction effects, 1562 had a lower heritability in females (59%, Supplemental Figure 3). Compared to all genome-wide sites, CpGs with significant sex interaction had a higher overall heritability on average (mean=0.37, median=0.29) and a larger variance explained by SNPs (mean=0.15, median=0.01), and were overall more variable between subjects (mean SD=0.04 for sites with significant sex interaction, versus mean SD=0.025 for all CpGs). In a similar way, we ran models that included age as a continuous interaction term and found significant interaction of the genetic or environmental variance of DNA methylation with

Figure 1: Heritability of DNA methylation, variance explained by SNPs and longitudinal stability. A): Histogram of total (IBD-based) heritability (h^2total) of DNA methylation (green) and variance explained by genome-wide SNPs (h^2 SNPs, purple) for genome-wide CpGs. The y-axis is truncated. B) Smooth scatterplot of the total (IBD-based) heritability (h^2 total, x-axis) of DNA methylation versus the variance explained by genome-wide SNPs (h^2 SNPs) for genome-wide CpGs. The density of data points is denoted by the color intensity (darker blue=higher density of data points). C) Histogram of IBD-based heritability at age 25 (green) and at age 50 (purple) for 39194 CpGs with significant interaction between age and genetic variance or between age and unique environmental variance. Dark blue denotes the overlap of green and purple bars. D) Smooth scatterplot of (twin-based) total heritability (h^2 total, x-axis) versus longitudinal correlation (y-axis) for genome-wide CpGs.



age at 39194 CpGs. Age interaction models ran successfully for 375783 CpGs (91%). CpGs with significant age interaction were more variable on average between subjects (mean SD age moderated CpGs= 0.036), had a higher total heritability on average (mean=0.37, median=0.36, based on IBD; mean=0.47, median=0.51, based on twins), and a larger proportion of variance explained by SNPs (mean=0.11, median=0.05). To illustrate the direction of effect, heritability of DNA methylation at age 25 and 50 is plotted for sites with significant age interaction effects in Figure 1C. For 90% of sites with significant age interaction, the heritability was lower at age 50 than at age 25. There were 2147 CpGs where the heritability was < 0.01 at age 50, suggesting that genetic variation contributes very little to the variation in methylation at these sites at a higher age, while the heritability of methylation at these sites at age 25 was on average 0.21. For most sites, the change in heritability was modest (Supplementary Figure 4), but large differences also occurred. For example, there were 103 CpGs where the change in heritability was larger than 0.5 between age 25 and age 50. For 32045 sites (82 %) with significant age moderation and 2022 sites (76%) with significant sex moderation, the difference in heritability (which is expressed as the ratio of additive genetic effects over the sum of additive genetic effects plus unique environmental effects) was related to a change in the unique environmental component. The overlap of age and sex moderation effects for individual CpGs is illustrated in Figure 2.

Figure 2: Venn diagram of the number of CpGs with significant interaction between sex and genetic variance (Sex X Va), between sex and unique environmental variance (Sex X Ve), between age and genetic variance (Age X Va) and between age and unique environmental variance (Age X Ve).



Longitudinal correlation of DNA methylation in blood

The longitudinal correlation was on average 0.21 (median=0.16), and the longitudinal correlation was strongly correlated with the h^2 of methylation at the same CpG (*r*=0.70), which suggests that sites with a larger heritability tend to be more stable across the time interval studied, as previously observed[50]. Highly heritable methylation sites ($h^2 > 0.8$, N= 17871) were highly stable over time (mean *r*=0.79, median *r*=0.84). Of note, sites with a high longitudinal correlation and a low h^2 also exist (Figure 1D): at these sites, longitudinal stability may reflect stability of environmentally driven variation or stability of variation that initially arose stochastically.

DNA methylation correlation between blood and buccal cells

Of 405487 CpGs common to the blood and buccal dataset, methylation level was positively correlated between blood and buccal at 244703 CpGs (mean *r*=0.22, median *r*=0.19) and negatively correlated at 160784 CpGs (mean *r*=-0.16, median *r*=0.14) between blood and buccal samples. Across all CpGs, the heritability in blood was weakly correlated with the size of the cross-tissue correlation (*r*=0.32). Focusing only on the most variable methylation sites in blood (SD of the β-value >= 0.05), heritability in blood was moderately correlated (*r*=0.54) with the blood-buccal correlation. Of note, CpGs where methylation level was strongly correlated between blood and buccal ($r > 0.8$, N=1015 CpGs, or $r < -0.8$, N=2 CpGs) were highly heritable and highly stable over time in blood (mean h^2=0.91, median=0.93; mean longitudinal *r*=0.83, median *r* =0.90). These findings suggest that when inter-individual variation in methylation level is strongly correlated between blood and buccal cells, genetic influences are main the cause of this correlation.

Discussion

We assessed genome-wide DNA methylation in whole blood in a large population-based twin cohort (including a small group of family members of twins) and estimated the heritability of DNA methylation level based on the classical twin method and based on measured genetic relationships between individuals (IBD). The two methods provided similar estimates: the genome-wide average heritability was 0.22 based on the classical twin approach and 0.19 based on IBD. Across genome-wide CpGs, the average proportion of variance of DNA methylation level explained by common SNPs was 0.07, and SNPs explained on average 37 % of the total heritability of methylation level. Our results indicate that for 49% of CpGs targeted by the 450k array at least 1% of the variation of DNA methylation in whole blood between individuals can be ascribed to the effects of common SNPs. These findings provide guidance for methylation QTL analyses by giving an indication of the number of CpGs that may be identified through methylation QTL analyses on whole blood samples (given sufficient power). Interestingly, at 18% of 450k targeted CpGs, > 99% of the heritability was explained by common SNPs. Yet, at the genome-

wide level there was a discrepancy between total heritability and variance explained by common SNPs. The proportion of DNA methylation heritability that cannot be explained by common SNPs may derive from genetic variants that are not or incompletely tagged by common SNPs, including rare variants and structural variation. CpGs with a higher heritability of DNA methylation tended to show larger stability across a time interval of on average 5 years. Also, CpGs where between-individual variation in methylation level was strongly correlated between blood and buccal cells were characterized by a high heritability in blood, suggesting that genetic influences underlie strong cross-tissue correlations.

We identified 2654 sites with evidence for sex-specific heritability and 39194 methylation sites with age-specific heritability. In support of previous indications that certain epigenetic marks between monozygotic twins may diverge as the twins age[33, 34] (a phenomenon referred to as epigenetic drift), our study indicated that the majority of methylation sites with significant age interaction showed a decreasing heritability with age due to increasing importance of environmental or stochastic influences on DNA methylation. The question remains whether these sites affect age and sex-dependent disease susceptibility.

Reference List

1. Mill,J. & Heijmans,B.T. From promises to practical strategies in epigenetic epidemiology. *Nat. Rev. Genet.* **14**, 585-594 (2013).
2. Bird,A. Perceptions of epigenetics. *Nature* **447**, 396-398 (2007).
3. Cedar,H. & Bergman,Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10**, 295-304 (2009).
4. Meissner,A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).
5. Dick,K.J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990-1998 (2014).
6. Hidalgo,B. *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* **63**, 801-807 (2014).
7. Irvin,M.R. *et al.* Epigenome-wide association study of fasting blood lipids in the Genetics of Lipid-lowering Drugs and Diet Network study. *Circulation* **130**, 565-572 (2014).
8. Liu,Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31**, 142-147 (2013).
9. Suter,M. *et al.* Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics.* **6**, 1284-1294 (2011).
10. Zeilinger,S. *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS. One.* **8**, e63812 (2013).
11. Gamazon,E.R. *et al.* Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol. Psychiatry* **18**, 340-346 (2013).

12. Zhang,X. *et al.* Linking the genetic architecture of cytosine modifications with human complex traits. *Hum. Mol. Genet.* **23**, 5893-5905 (2014).
13. Shi,J. *et al.* Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat. Commun.* **5**, 3365 (2014).
14. Boks,M.P. *et al.* The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS. One.* **4**, e6767 (2009).
15. Tapp,H.S. *et al.* Nutritional factors and gender influence age-related DNA methylation in the human rectal mucosa. *Aging Cell* **12**, 148-155 (2013).
16. Liu,J., Morgan,M., Hutchison,K., & Calhoun,V.D. A study of the influence of sex on genome wide methylation. *PLoS. One.* **5**, e10028 (2010).
17. Bell,J.T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS. Genet.* **8**, e1002629 (2012).
18. Horvath,S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13**, R97 (2012).
19. Horvath,S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
20. Teschendorff,A.E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440-446 (2010).
21. Kaminsky,Z., Wang,S.C., & Petronis,A. Complex disease, gender and epigenetics. *Ann. Med.* **38**, 530-544 (2006).
22. Bjornsson,H.T., Fallin,M.D., & Feinberg,A.P. An integrated epigenetic and genetic approach to common human disease. *Trends Genet.* **20**, 350-358 (2004).
23. Lillycrop,K.A., Phillips,E.S., Jackson,A.A., Hanson,M.A., & Burdge,G.C. Dietary protein restriction of pregnant rats induces and folic acid supplementation prevents epigenetic modification of hepatic gene expression in the offspring. *J. Nutr.* **135**, 1382-1386 (2005).
24. Tobi,E.W. *et al.* DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum. Mol. Genet.* **18**, 4046-4053 (2009).
25. Flores,K.G. *et al.* Sex-specific association of sequence variants in CBS and MTRR with risk for promoter hypermethylation in the lung epithelium of smokers. *Carcinogenesis* **33**, 1542-1547 (2012).
26. Rideout,W.M., III, Eggan,K., & Jaenisch,R. Nuclear cloning and epigenetic reprogramming of the genome. *Science* **293**, 1093-1098 (2001).
27. Cortijo,S. *et al.* Mapping the epigenetic basis of complex traits. *Science* **343**, 1145-1148 (2014).
28. Morgan,H.D., Sutherland,H.G., Martin,D.I., & Whitelaw,E. Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.* **23**, 314-318 (1999).
29. McRae,A.F. *et al.* Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol.* **15**, R73 (2014).
30. Grundberg,E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876-890 (2013).
31. Gordon,L. *et al.* Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res.* **22**, 1395-1406 (2012).

32. Martino,D. *et al.* Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biol.* **14**, R42 (2013).
33. Fraga,M.F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *PNAS* **102**, 10604-10609 (2005).
34. Talens,R.P. *et al.* Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell* **11**, 694-703 (2012).
35. Bocklandt,S. *et al.* Epigenetic predictor of age. *PLoS. One.* **6**, e14821 (2011).
36. Purcell,S. Variance components models for gene-environment interaction in twin analysis. *Twin. Res.* **5**, 554-571 (2002).
37. Boomsma,D.I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* **9**, 849-857 (2006).
38. Willemsen,G. *et al.* The adult Netherlands twin register: twenty-five years of survey and biological data collection. *Twin. Res. Hum. Genet.* **16**, 271-281 (2013).
39. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* **13**, 231-245 (2010).
40. van Dongen,J. *et al.* Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. *Genes (Basel)* **5**, 347-365 (2014).
41. Yang,J., Lee,S.H., Goddard,M.E., & Visscher,P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).
42. Bibikova,M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295 (2011).
43. Chen,Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* **8**, 203-209 (2013).
44. Genome of the Netherlands Consortium Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**, 818-825 (2014).
45. Fortin,J.P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *bioRxiv*(2014).
46. Du,P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC. Bioinformatics.* **11**, 587 (2010).
47. Zaitlen,N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS. Genet.* **9**, e1003520 (2013).
48. Visscher,P.M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS. Genet.* **2**, e41 (2006).
49. Yang,J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519-525 (2011).
50. Shah,S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* **24**, 1725-1733 (2014).

Epigenetic variation in monozygotic twins: A genome-wide analysis of DNA methylation in buccal cells

Abstract

DNA methylation is one of the most extensively studied epigenetic marks in humans. Yet, it is largely unknown what causes variation in DNA methylation between individuals. The comparison of DNA methylation profiles of monozygotic (MZ) twins offers a unique experimental design to examine the extent to which such variation is related to individual-specific environmental influences and stochastic events or to familial factors (DNA sequence and shared environment). We measured genome-wide DNA methylation in buccal samples from ten MZ pairs (age 8-19) using the Illumina 450k array, and examined twin correlations for methylation level at 420,921 CpGs after QC. After selecting CpGs showing most variation in methylation level between subjects, the mean genome-wide correlation (rho) was 0.54. The correlation was higher, on average, for CpGs within CpG islands (CGIs), compared to CGI shores, shelves, and non-CGI regions, particularly at hypomethylated CpGs. This finding suggests that individual-specific environmental and stochastic influences account for more variation in DNA methylation in CpG-poor regions. Our findings also indicate that it is worthwhile to examine heritable and shared environmental influences on buccal DNA methylation in larger studies that also include dizygotic twins.

Based on: van Dongen J., Ehli EA, Slieker RC, Bartels M, Weber ZM, Davies GE, Slagboom PE, Heijmans BT, Boomsma DI. Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. Genes (Basel) 2014; 5:347-365

Introduction

To date, hundreds of genetic risk variants for complex traits and diseases have been identified, although for most of these variants the biological mechanisms remain to be elucidated [1]. Interestingly, the majority of disease-associated genetic variation is located in regulatory regions of the genome [2], including transcription-factor-occupied regions and DNase I hypersensitive sites (which correspond to open chromatin) [3]. This suggests that mechanisms that control the activity of genes including epigenetic mechanisms may represent an important link between DNA sequence variation and common disease susceptibility [4]. Trying to unravel the molecular biology underlying complex traits and disease, much attention has been drawn recently to these epigenetic mechanisms; non-DNA sequence-based regulation of gene expression by DNA methylation, histone modification, microRNAs etc [5]. DNA methylation is one of the most extensively studied epigenetic mechanisms in human populations and tissues, and is the focus of this paper.

In humans, DNA methylation occurs almost exclusively at cytosines that are part of CpG dinucleotides. The relationship between DNA methylation and expression varies depending on the genomic context: CpG methylation at promoter regions is generally thought to repress gene expression, while gene body methylation is generally associated with active gene expression and has been suggested to regulate splicing [6-8]. In most cell types, the majority of CpGs in the genome (on average 70-80%) is typically methylated [9]. Of the unmethylated CpG sites in the genome, most occur in areas of clustered CpGs called CpG islands, which are often present in promoter regions. Yet, DNA methylation patterns may vary, and differential methylation has been demonstrated to occur across age [10], cell types, tissues [8, 11], and disease states [12, 13], and it has become clear that widespread variation in methylation patterns exist between individuals [14]. Accumulating evidence suggests that DNA methylation patterns can be affected by genetic variants (mQTLs) [15], environmental exposures [16], and stochastic factors [17, 18], but it is largely unknown how much each of these factors account for overall variation between individuals in DNA methylation across the genome. Twin studies provide insight into the proportion of inter-individual variation in DNA methylation that is due to genetic variation, environmental effects and stochastic variation [19].

Because MZ twins derive from a single zygote and therefore have (nearly) identical DNA sequences (see for example Ye *et al* 2013 [20]), the comparison of DNA methylation patterns of MZ twins allows to examine the extent to which differences in methylation between human individuals are related to environmental and stochastic events. Previous studies have highlighted that various tissues of MZ twins already show differences in DNA methylation at birth [21, 22], and that differences between twins for average genome-wide DNA methylation, total histone acetylation levels and methylation at certain loci increase with age (referred to as “epigenetic drift”) [23]. Although a cross-sectional study of DNA methylation discordance in saliva from 34 MZ pairs (age

range 21-55 years) found no evidence for larger differences in DNA methylation in older MZ pairs [24], results from a cross-sectional analysis based on 230 MZ pairs (age range 18-89 years) suggested a gradual increase of DNA methylation discordance in MZ twins from early adulthood to advanced age at various candidate loci, which was supported by longitudinal data from 19 elderly MZ pairs [25].

In the past few years, various studies have examined DNA methylation at a set of candidate genes or particular genomic regions in MZ and dizygotic (DZ) twins [26-31], usually reporting greater similarity of MZ twins compared to DZ twins, suggesting that heritable influences contribute to DNA methylation variation at specific regions. While CpG sites at some imprinted loci showed evidence for moderate to high heritability in blood samples from adolescent and middle-aged twins [29], other genomic regions including the major histocompatibility complex (MHC) region showed little evidence for genetic influences on DNA methylation variation [28]. Twin studies also highlighted variation between tissues in the importance of genetic influences on methylation of candidate loci at birth [30]. A longitudinal classical twin study of three candidate genes (*DRD4, SLC6A4/SERT* and *MAOA)* based on buccal cells indicated that changes in methylation of these genes within individuals between age 5 and 10 are mostly attributable to non-shared environmental influences and stochastic variation [31]. Clearly, twin studies of candidate regions suggest that there is broad variation in the importance of heritable influences, and environmental or stochastic variation to DNA methylation at different regions.

To date, only a few genome-scale analyses of DNA methylation have been performed using the classical twin design, including a study of ~12,000 CpG sites within islands [32], two studies that used a promoter-specific array targeting ~27,000 CpG sites (Illumina 27K) [21, 33], and two studies that used the Infinium HumanMethylation 450 array (Illumina 450k) [22, 34], which assesses ~485,000 CpG sites across a variety of regions in the genome, including gene bodies and intergenic regions [35]. The studies that assessed heritability consistently reported that the average heritability of methylation level at CpGs across the genome is low to moderate when all sites are considered, although the heritability of individual CpGs ranges between 0% and 100%. The following estimates of average heritability across genome-wide CpGs have been reported to date (based on all analysed CpGs): 18 % in blood from 32 to 80 year old twins (21 MZ pairs and 31 DZ pairs) [33], 5% in placenta, 7% in human umbilical vascular endothelial cells (HUVEC) and 12% in cord-blood mononuclear cells (CBMC) from neonatal twins (22 MZ and 12 DZ pair [21]), and 19% in adipose tissue from adult female twins (97 MZ pairs and 162 DZ pairs [34]). In two studies of neonatal twin tissues, methylation discordance in MZ and DZ twins increased with increasing distance from CpG islands (CGIs) for certain probes (type I), i.e., differences were larger in the shores and shelves that flank CGIs [21, 22]. In the study of adipose tissue, it was noted that the

average genome-wide heritability of DNA methylation was higher when restricting to the most variable CpG sites (for the top 10% CpGs of which methylation level varied most between subjects, the average heritability was 37%) [34]. It was also found that gene body and intergenic regions showed higher average methylation levels, more variation between subjects, and higher heritability compared to promoter regions in adipose tissue [34].

To summarize, there is great interest in unraveling the factors that contribute to variation in DNA methylation between persons, but most previous twin studies of DNA methylation have been limited to candidate genes or a subset of regulatory regions in the genome (mostly promoter regions and CGIs). Two earlier studies used the Illumina 450k to collect genome-wide data in MZ and DZ twins; one in adipose tissue in adults [34] and one in DNA isolated from buccal cells in infants (10 MZ pairs and 5 DZ pairs, longitudinal design) [22]. In line with earlier findings suggesting divergence of DNA methylation profiles with age in MZ twins (mostly based on data from adult twins, cross-sectional comparisons, and limited genomic coverage), Martino *et al* [22] showed that widespread DNA methylation changes occur across the genome in buccal cells between birth and 18 months, and that some MZ and DZ pairs already show divergence of DNA methylation profiles, whereas other pairs show stable difference levels or became more similar within the first 18 months after birth. In this paper, we analyzed genome-wide DNA methylation profiles (Illumina 450k) from buccal epithelium. We focused on 10 young and adolescent MZ twin pairs (age 8-19). The aim of our study was to examine how similar the DNA methylation profiles of buccal cells from genetically identical subjects are in childhood and adolescence, and whether MZ twin similarity varies between different genomic regions.

Previous studies have highlighted differences in mean methylation level, differences in the effect of methylation level on gene expression, and differences in the effect size and direction of effect on methylation for disease associations across different regions in the genome [6]. These findings indicate that the establishment and maintenance of DNA methylation is differentially regulated in different regions, and that a given change in methylation in different areas may have different downstream effects, suggesting that DNA methylation in some regions may be more tightly controlled than in others. We questioned whether these regional differences are also accompanied by differences in the importance of environmental and stochastic influences versus familial factors (genetic variation and shared environment) to inter-individual variation in methylation levels. Therefore, we describe the MZ twin correlations of individual CpGs as a function of various genomic classifications, including position relative to CGIs (CGI regions, shores, shelves, and non-CGI regions), genes (distal to promoter, proximal to promoter, gene body and intergenic), and ENCODE regulatory regions (DNaseI hypersensitive sites (DHS) and transcription factor binding sites (TFBS)). Hereby, our study gives valuable insight into the factors influencing inter-individual genome-wide DNA

methylation variation in buccal cells in childhood and adolescence, and into the degree to which these influences vary across functional regions in the genome.

Experimental Section

Subjects

Ten monozygotic twin pairs who take part in longitudinal studies of the Netherlands Twin Register (NTR) were selected for the current study. There were five young twin pairs [36] whose buccal samples were collected when the twins were between ages 8 and 10 years and five adolescent pairs [37] who were aged 18-19 years at the time of sample collection. In the young group, there were three male pairs and two female pairs, and in the adolescent group there were two male pairs and three female pairs. The twins were unselected with respect to phenotypic characteristics. Informed consent was obtained from the parents (children) or from the twins themselves (adolescents). The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180). Participants could indicate if they wished to be informed of the results of zygosity testing. Zygosity testing, based on a set of SNPs and VNTRs as described in Van Beijsterveldt *et al*. 2013 [36], confirmed that all pairs were MZ. In addition to the twin samples, a single sample was used as a genomic DNA control. This DNA sample (CEPH) was derived from a stable cell line (female) from the HapMap project and was run in four replicates on the methylation beadchip arrays.

Buccal DNA collection

The procedures of buccal swab collection [38] and genomic DNA extraction [39] have been described previously. In short, 16 cotton mouth swabs were individually rubbed against the inside of the cheek by the participants and placed in four separate 15 mL conical tubes (four swabs in each tube) containing 0.5 mL STE buffer (100 mM Sodium Chloride, 10 mM Tris Hydrochloride [pH 8.0], and 10 mM Ethylenediaminetetraacetic acid) with proteinase K (0.1 mg/mL) and Sodium Dodecyl Sulfate (SDS) (0.5%) per swab. Individuals were asked to refrain from eating or drinking 1 hour prior to sampling. High molecular weight genomic DNA was extracted from the swabs using a high salt (KAc) precipitation followed by a standard chloroform/isoamyl alcohol (24:1) extraction. The DNA samples were quantified using absorbance at 260 nm with a Nanodrop ND-1000 (Nanodrop Technologies, Wilmington, DE, USA).

Infinium HumanMethylation450 BeadChip Data Generation

The epigenome-wide methylation data was generated using the Infinium HumanMethylation450 BeadChip Kit (Illumina, Inc.). The Infinium

HumanMethylation450 BeadChip is able to interrogate over 450,000 methylation sites across the entire genome including 99% of RefSeq genes. Content was selected to include gene regulatory regions such as the promoter, 5' UTR, first exon, gene body, and the 3' UTR. Additionally, bead probes were also designed to cover regions adjacent to the CpG islands such as the shores and shelves [35].

The Infinium DNA methylation assay was performed at the Avera Institute for Human Genetics. The assay was completed exactly as denoted in the manufacturer's protocol. The concentration of genomic DNA used in the Infinium DNA methylation assay was determined by comparing the binding of PicoGreen to known standards (λ DNA) and to the sample DNA. Briefly, 500 ng of genomic DNA was used for bisulfite conversion using the Zymo EZ DNA methylation kit (Zymo Research). 5 µl of bisulfite-converted DNA was whole genome amplified which was followed by enzymatic end-point fragmentation. The resulting fragments were purified using an isopropanol precipitation and the resuspended genomic DNA was denatured and hybridized to the beadchip arrays for 18 hours. Extension, staining, and washing were completed manually in flow cells followed by imaging using the iScan system (Illumina, Inc.). The raw data were extracted as *idat* files and were used in the downstream analysis.

Quality Control, Normalization and data processing

The raw intensity files (*idat*) were imported into the R environment [40], where further processing, quality control and normalization took place. The performance of bisulfite control probes confirmed successful bisulfite conversion for all samples. For each sample, we compared the overall (median) methylated signal intensity to the overall unmethylated signal intensity across all probes, and compared the overall signal intensity from all CpG probes to the overall background signal ('noise') as assessed using negative control probes. The overall signal from CpG probes was good and well-separated from background signal for all samples. As a final quality check of the samples, cluster analysis was performed (cluster method=complete linkage) based on the Euclidean distance between samples, which was calculated from the pair-wise correlations between samples using the most variable probes (probes with a SD of the β-value across all 24 samples > 0.10, with probes on the X and Y chromosomes and probes containing SNPs, as described in the next paragraph excluded; Nprobes= 38,359). Results of the cluster analysis were visualized in a dendrogram (see results section), which showed no outlier samples and illustrated tight clustering of the four replicate measures of control DNA.

Several probe-level QC steps were performed to filter out probes with low performance. For all samples, ambiguously-mapped probes were excluded, based on the definition of an overlap of at least 47 bases per probe from Chen *et al*. [41], and all probes containing a SNP, identified in the Dutch population [42],

within the CpG site (at the C or G position) were excluded, irrespective of minor allele frequency. For each sample individually, probes with an intensity value of zero (not present on the array of a particular sample), probes with a detection *P* value > 0.01 (calculated using the function *detectionP* from the *minfi* package), and probes with a bead count < 3 were excluded. After these steps, probes with a success rate < 0.95 across samples were removed from all samples and the success rate across probes for each sample was computed (Range of per sample success rate: 0.9990-0.9998).

After QC, background and red/green color adjustment were applied to the raw probe intensity values using quantile normalization. Normalized intensity values were converted into beta-values (β). The β-value, which represents the methylation level at a CpG for an individual and ranges from 0 to 1, is calculated as:

$$\beta = \frac{M}{M+U+100}$$

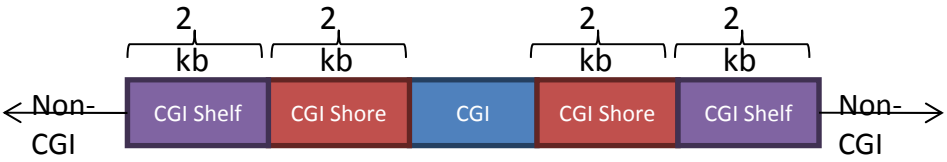
where M=Methylated signal, U=Unmethylated signal, and 100 represents a correction term to control the β –value of probes with very low overall signal intensity (i.e. probes for which M+U ~0 after background subtraction).

Finally, in anticipation of our categorization of CpGs based on the mean β-value across samples, β-values were adjusted to account for (intra-sample) differences in the distributions of methylation values derived from type I probes (two bead types per CpG site) versus type II probes (one bead type per CpG site) using the beta-mixture quantile normalization method (BMIQ) [43].

Genomic annotations

CpGs that passed QC criteria (N=420,921) were mapped to genomic features, DNase I hypersensitive sites (DHS), and transcription factor binding sites (TFBS) as described by Slieker *et al.* [8]. The genomic feature annotation is based on first assigning CpGs to one of five gene-centric regions: intergenic region (>10 kb from the nearest transcription start site (TSS)), distal promoter (-10 kb to -1.5 kb from the nearest TSS), proximal promoter (−1.5 kb to +500 bp from the nearest TSS), gene body (+500 bp to 3' end of the gene) and downstream region (3' end to +5 kb from 3' end). Next, CpGs were mapped to CGIs (CG content > 50%, length > 200bp and observed/expected ratio of CpGs > 0.6; locations were obtained from the UCSC genome browser [44]), CGI shore (2kb region flanking CGI), CGI shelf (2kb region flanking CGI shore), or non-CGI regions (Figure 1). According to the gene-annotations, 14.4% of all CpGs was located in intergenic regions, 4.7% mapped to distal promoter, 40.4% to proximal promoter, 38.6% to gene body, and 1.9% to downstream region. 33.0% of CpGs was located within CGIs, 23.8% in shores, 9.2% in shelves and 34.0% outside CGIs. The locations of DHS and TFBS, which were described by the ENCODE project [3], were downloaded from the UCSC genome browser. Finally, CpGs were mapped to imprinted genes that were described by Yuen *et al* [45].

Figure 1. Illustration of a CpG island (CGI) with surrounding CGI shores, CGI-shelves and non-CGI regions.



Statistical analysis of twin data

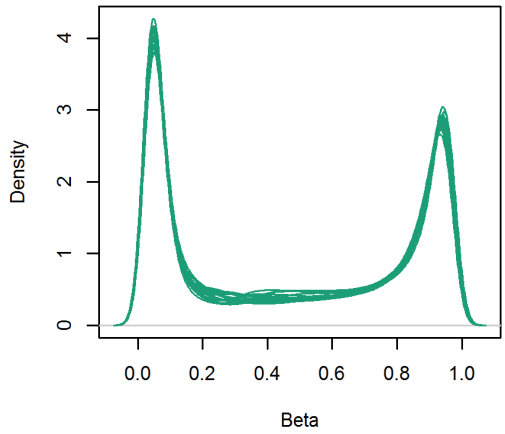
To examine the similarity of DNA methylation profiles of MZ twins, we computed correlations between the normalized β-values of MZ co-twins using the following two approaches: (1) For each MZ twin pair, the Spearman correlation (rho) was computed between the β-values of twin 1 and the β-values of twin 2 (across all CpGs, i.e. CpGs are cases), as a measure of the overall similarity of the methylation profiles of each twin pair. (2) For each CpG, the Spearman correlation (rho) was computed between the β-value of twin 1 and the β-value of twin 2 (across all 10 MZ twin pairs, i.e. MZ twin pairs are cases), as a measure of the similarity of the methylation level of a CpG in MZ twins. For scenario 2, we describe the range of correlations for the most variable CpGs. The most variable CpGs were additionally grouped by genomic annotations, and average methylation level. For each CpG, the average methylation level (β-value) and the standard deviation (SD) were computed across subjects (20 MZ twins). Based on the average β, CpGs were classified as hypomethylated (mean β < 0.3), intermediately methylated (mean β >=0.3-0.7), or hypermethylated (mean β >=0.7). Based on the SD, CpGs were classified as "most variable CpGs" if they had an SD >= 0.05.

Results and Discussion

DNA methylation level across the genome

After QC of the methylation data, 420,921 CpGs from 10 monozygotic twin pairs were analyzed. The methylation level across genome-wide CpGs showed the typical bimodal distribution for each subject (Figure 2). Based on our β-value cut-offs (see experimental section); 184,765 CpGs (43.9%) were classified as hypomethylated, 64,829 CpGs (15.4%) were intermediately methylated, and 171,327 CpGs (40.7%) were hypermethylated. CGIs were on average hypomethylated, with CGIs in proximal promoter regions showing a narrow range of average methylation levels across individual CpGs, and CGIs in gene bodies, downstream regions and intergenic regions showing a broader range of methylation levels across individual CpGs (see Figure 3). Compared to CGIs, the shores, shelves and non-CGI regions on average had a higher methylation level, except for proximal promoter shores. Shores generally showed the widest range of average methylation levels across individual CpGs, when compared to CGIs, shelves and non-CGI regions (Figure 3).

Figure 2. Density of β-values after normalization for all twin samples.



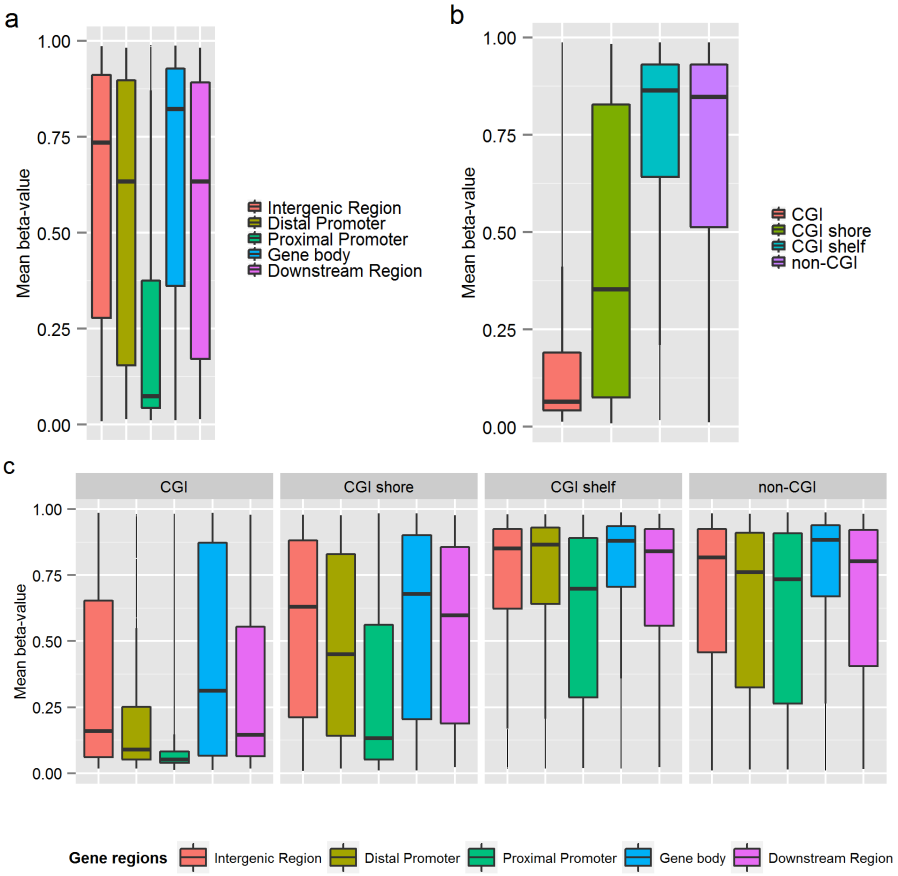
Similarity of genome-wide methylation profiles of MZ twins

A cluster analysis of the methylation data revealed that all but one MZ twin clustered closely together with their co-twin (Figure 4), which could be related to differences in the cellular composition of the samples of this twin pair. Buccal swab samples are mainly composed of buccal epithelial cells with a small proportion of leukocytes, but the exact proportions may vary between persons. Using information from a reference 450k methylation dataset [8], we examined potential variation between twin samples in the proportion of buccal versus blood cells, by clustering the twin data based on methylation values at CpGs that showed a large difference in methylation between blood and buccal samples in the reference dataset (See Supplemental Methods). Although some variation was indicated by this approach, exclusion of twin samples with putatively deviant cellular proportions yielded similar results for the correlation analyses (see Table S2 and Figure S2), and we therefore decided to keep all samples in the analyses reported in this paper.

Figure 5 shows a typical scatterplot of genome-wide CpG methylation levels in buccal cells from an MZ twin pair. It illustrates that overall, buccal DNA methylation profiles of MZ pairs are highly concordant when all CpGs are considered (rho= 0.981-0.994 for different MZ pairs, mean rho=0.991), however, these correlations are to a large extent driven by invariable CpGs that are hypomethylated or hypermethylated in both twins. For pairs of unrelated subjects, the mean correlation was 0.983 (range: 0.970-0.992). When comparing only the most variable CpGs (SD of $\beta \geq 0.05$), the correlations ranged from 0.869 to 0.989 (mean rho=0.966) in MZ twins (and mean rho=0.859, range: 0.608-0.963 for unrelated subjects). Thus, when looking only at CpGs that may vary between individuals, the overall pattern of methylation across CpGs is still highly similar within MZ pairs on average, but more variation between individual pairs becomes visible, as the methylation level at variable CpGs overall was more strongly correlated for some MZ pairs than for others. This finding is in line with the results from Martino *et al* based on buccal

cells from twins at birth and at the age of 18 months [22], which also indicated that some MZ pairs are more similar than other pairs with respect to their DNA methylation profiles.

Figure 3. Average methylation level of individual CpGs across gene regions (a), CpG islands (CGI) and non-CGI regions (b), and for each genomic feature separately (c).



Similarity of the methylation level at individual CpGs in MZ twins

Although all ten MZ twin pairs showed high overall similarity of methylation across genome-wide CpGs, some CpGs differed within MZ twin pairs (Figure 5), and we questioned how similar the methylation level at individual CpGs is when summarized across all MZ pairs. To this end, we computed for each CpG the correlation between methylation values of MZ twins. A high MZ twin correlation for a CpG suggests that MZ co-twins consistently show similar methylation levels at this CpG, indicating little stochastic and environmental variation (including measurement error) at this site, whereas a low MZ twin correlation for a CpG suggests dissimilar methylation levels in co-twins, which is indicative of a large degree of stochastic and environmental influences.

Figure 4. Cluster dendrogram of all twin and control samples. From left to right, the first two branches separate the control samples (HapMap cell line DNA) from the buccal samples from twins.

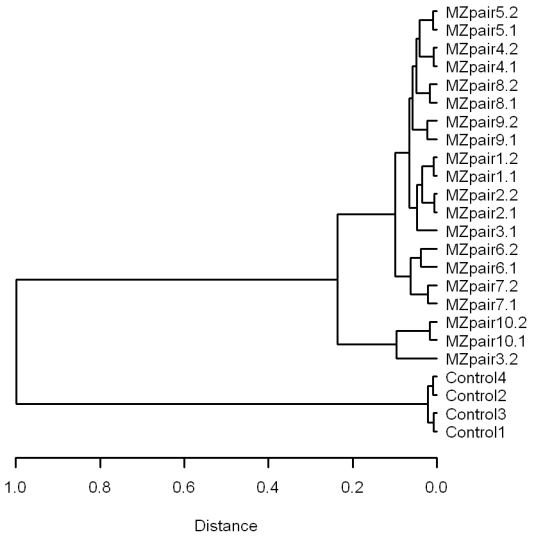
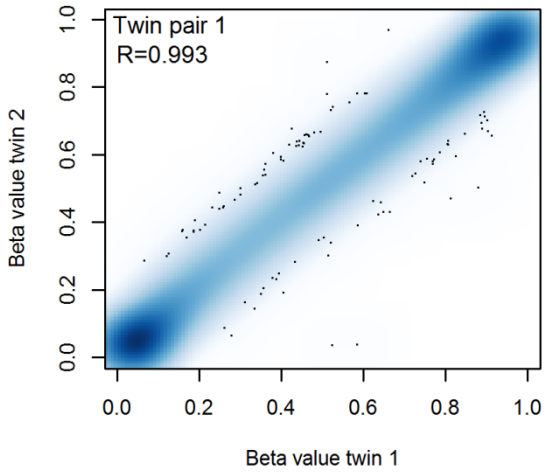


Figure 5. Smooth scatterplot of DNA methylation levels (β-values) at 420,921 CpGs in buccal cells from a monozygotic twin pair.



Summarizing the individual CpG correlations over all 420,921 CpGs, the average MZ twin correlation was 0.31(median=0.35, range: -0.963-1), which is in line with the low heritability across genome-wide CpGs reported by previous studies [21, 33, 34]. However, as the majority of CpGs showed very little variation in methylation level between subjects, all subsequent analyses were conducted using only the most variable sites (N=59,041), which showed an average genome-wide correlation of 0.54 (median=0.54, range: -0.661-1) in MZ twins.

These findings suggest that while the large majority of CpGs are either hyper- or hypomethylated and show little between-individual variation in DNA methylation in buccal samples, a small portion does vary markedly, and these CpGs are on average moderately to strongly correlated in MZ twins.

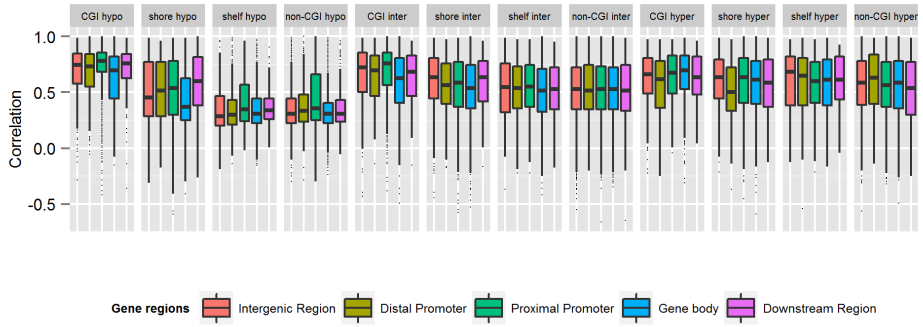
Table 1 describes the MZ twin correlations separately for various genomic regions, and separately for hypomethylated, intermediately methylated and hypermethylated CpGs. Comparing the different gene-centric classifications, the average MZ twin correlation was highest for CpGs in proximal promoter areas (mean rho=0.57) and lowest for gene body CpGs (mean rho=0.51). The MZ twin correlation of methylation values was also lower on average in CGI shores (mean rho=0.54), shelves (mean rho=0.50), and non-CGI regions (mean rho=0.49) compared to CGIs (mean rho=0.66). Looking at the MZ twin correlations across genome annotations separately for hypomethylated (29.8% of variable CpGs), intermediately methylated (50.0% of variable CpGs) and hypermethylated CpGs (20.2% of variable CpGs), the median MZ twin correlation was consistently lower in the shelves, shores and non-CGI regions compared to CGIs – for all genic and intergenic regions, and this difference was most pronounced for hypomethylated CpGs (Figure 6). This observation suggests that the relative influence of familial versus individual-specific influences differs between these regions, with regions of low CpG density showing more variation due to individual-specific environmental and stochastic factors compared to CpG dense regions. Larger methylation discordance of MZ twins in CGI shores and shelves was also previously indicated by studies of neonatal twins [21, 22]. Our results thus replicate previous findings and add to these findings that the pattern previously observed in MZ twins at birth is also visible in childhood and adolescence.

The most strongly correlated CpGs in MZ twins (mean rho=0.73) were hypomethylated CpGs located in proximal promoter CGIs (N=2,547 CpGs, constituting 4.3 % of the most variable CpGs, and 2.9% of all CpGs located in proximal promoter CGIs), while MZ twin correlations on average were lowest in hypomethylated non-CGI gene body CpGs (N=2,972 CpGs constituting 5.0% of the most variable CpGs, Mean rho=0.34). In combination with our observation that most proximal promoter CpGs are on average hypomethylated (Figure 3), these findings indicate that DNA methylation variation is generally depleted in proximal promoter CGIs. Yet, a small proportion of CpGs in proximal promoter CGIs does show marked variation in young and adolescent individuals, and the high average MZ twin correlations at these sites suggest that this variation may be to a large extent under genetic control.

Table 1: Spearman correlation between the methylation values of MZ twins for individual CpGs. Results are based on the most variable CpGs (N=59,041).

Category	N CpGs	Mean rho	Median rho	Min rho	Max rho
All CpGs	59041	0.54	0.54	-0.661	1
Gene-centric annotations	**N CpGs (%)**	**Mean rho**	**Median rho**	**Min rho**	**Max rho**
Intergenic (>10 kb from TSS)	11430 (19.4%)	0.52	0.53	-0.56	1
Distal Promoter (-10kb to -1.5 kb from TSS)	3193 (5.4%)	0.53	0.53	-0.54	1
Proximal Promoter (-1.5 kb to +500bp from TSS)	17880 (30.3%)	0.57	0.62	-0.66	1
Gene Body (+500 bp to 3'end)	25163 (42.6%)	0.51	0.50	-0.59	1
Downstream region (3'end to +5kb from 3'end)	1375 (2.3%)	0.55	0.55	-0.66	1
CGI annotations	**N CpGs (%)**	**Mean rho**	**Median rho**	**Min rho**	**Max rho**
CGI	10576 (17.9%)	0.66	0.73	-0.49	1
CGI shore	14803 (25.1%)	0.54	0.55	-0.59	1
CGI shelf	6001 (10.2%)	0.50	0.49	-0.54	1
Non-CGI	27661 (46.9%)	0.49	0.47	-0.66	1
Methylation level	**N CpGs (%)**	**Mean rho**	**Median rho**	**Min rho**	**Max rho**
Hypomethylated (average beta < 0.3)	17581 (29.8)	0.48	0.42	-0.59	1
Intermediately methylated (average beta >=0.3-0.7)	29519 (50.0)	0.55	0.56	-0.66	1
Hypermethylated (average beta > 0.7)	11941 (20.2)	0.58	0.61	-0.59	1

Figure 6. MZ twin correlations for individual CpGs grouped by genomic region and average methylation level. Hypo=Hypomethylated. Inter=intermediate methylation. Hyper=Hypermethylated. Results are



MZ twin resemblance at CpGs in ENCODE regulatory regions

To further examine DNA methylation at regulatory regions in the genome, we focused specifically on CpGs located within DNase I hypersensitive sites (DHS) and CpGs within transcription factor binding sites (TFBS) identified by the ENCODE project. It has previously been described that these regions are enriched among disease-associated genetic variants [3], but it has not yet been studied to which extent heritable versus other sources of variation account for variation in DNA methylation in these regions. We found that both DHS and TFBS were on average hypomethylated as expected for transcriptionally active DNA (DHS: Mean β=0.27, Median=0.09; TFBS: Mean β=0.24, Median=0.08). The most variable CpGs in these areas (representing 16.2% of all CpGs in DHS, and 13.7% of CpGs in TFBS) showed a mean correlation of 0.52 (DHS) and 0.53 (TFBS), respectively, in MZ twins. These results suggest that buccal cells overall show little variation in methylation level at the majority of CpGs within DHS and TFBS. A small proportion of CpGs in DHS and TFBS, however, does show variation between individuals, and these sites were moderately to strongly correlated in MZ twins, suggesting that these sites may be of particular interest to follow-up in future studies of heritability.

MZ twin resemblance at CpGs in imprinted genes

At imprinted genes, one of the alleles is typically methylated to repress expression, while the other allele is unmethylated – depending on the parent from whom the allele was inherited. This results in a methylation level of around 50% at imprinted CpGs when the two alleles are measured simultaneously. A previous twin study demonstrated moderate to high heritability at CpGs at two imprinted loci [29], suggesting that CpGs within imprinted genes may on average show more heritable variation compared to most other genome-wide CpGs. In our dataset, 346 CpGs were located in DMRs (Differentially Methylated Regions) of 59 imprinted genes, described by Yuen *et al* [45]. These genes were

identified as imprinted in human placental tissue, and although some of these genes showed similar methylation patterns in one or multiple fetal tissues, including muscle, brain and kidney, it is unknown whether these genes are also imprinted in buccal cells. From the Yuen *et al* set, 144 CpGs in 46 genes (see Supplemental Table 3) showed a methylation level indicative of imprinting in our data (intermediate methylation; mean β >=0.3-0.7). The average MZ twin correlation for this set of CpGs was 0.47 (median rho=0.50), suggesting that MZ twin correlations at imprinted gene CpGs on average are comparable to the MZ twin correlation at intermediately methylated CpGs in general.

Interpretation and future directions

The average twin correlation of methylation values for MZ twins at individual CpGs was low across all measured genome-wide CpGs, but it was moderate to large on average when focusing only on variably methylated CpGs. This is in line with results from a heritability analysis of DNA methylation in adipose tissue, which showed that the average heritability across all CpGs was higher for the top 10% of CpGs with the largest standard deviation of methylation level across subjects [34].

Importantly, as well as effects of environmental and stochastic influences, differences in DNA methylation within MZ twin pairs may result from variation in the cellular composition of samples and from technical variation (including measurement error). Buccal swab samples are mainly composed of buccal epithelial cells with a small proportion of leukocytes, but the exact proportions may vary between persons, which could lead to methylation variation within MZ twin pairs that mainly tag differences in cell type composisition. We examined the impact of variation in the proportion of buccal versus leukocytes on our data by studying methylation patterns of all twin samples at CpGs with a large methylation level difference between buccal and blood samples (see Supplementary Methods). Exclusion of four twin pairs, for which this approach indicated a more deviant cellular composition in one or both twins (lower proportion of buccal epithelial cells; see Supplemental Figure 1 and Supplemental Table 1), however, had very little impact on the average MZ twin correlations reported in this paper and led to the same conclusions (See Supplemental Table 2 and Supplemental Figure 2 for results based on exclusion of the putatively more heterogeneous samples).

With respect to technical variation, it is important to note that if the actual methylation status at a particular site is either completely unmethylated (0%) or completely methylated (100%) without true biological variation between subjects, some variability between the measured values of individuals is expected due to technical variation [46]. It is therefore likely that at sites that were on average hypomethylated or hypermethylated in our data, technical variation may account for a large part of the observed variation (although true biological variation may of course also account for part of the variation at these sites). An interesting question that largely remains to be examined is what types of

environmental influences can induce changes in DNA methylation, and thereby possibly impact on gene expression. Although our study design does not provide insight with regard to which of the observed differences between twins are the result of different environmental exposures and which differences have arisen due to stochastic variation in molecular processes, future studies of MZ twins who are discordant for environmental exposures should allow to examine the effects of such influences on DNA methylation. Our finding that many CpGs in the genome show dissimilar methylation levels in young and adolescent MZ twins indicates that it is of interest for further studies to specifically search for regions in the genome where differential methylation in MZ twin coincides with differential exposures. As we observed that DNA methylation in MZ twins is overall less similar at CpGs in non-CGI regions, CGI shores and shelves, these regions are of particular interest to studies examining environmental exposures, as these regions may show the strongest effects of environmental influences.

To check whether the lower average MZ twin correlation at hypomethylated sites is not merely related to the distribution of β-values being truncated at 0 (and 1) by definition, we also ran the analyses on M-values ($M= \log_2\left(\frac{\beta}{1-\beta}\right)$), which have better statistical properties but reduced biological interpretability compared to β-values [47]. The MZ twin correlations based on M-values were highly similar to those based on β-values and showed a similar genome-wide average (Table S4), and a similar pattern across regions and mean methylation categories (Figure S3). Irrespective of whether lower resemblance of MZ twins mainly reflects that these sites harbor more biological variation that is unique to MZ twins, or reflects that more variation at these sites is related to measurement error, our findings provide useful information for future heritability and mQTL studies. CpGs that are very weakly correlated between MZ twins are not likely to show high heritability or strong effects of DNA variants on methylation level.

A limitation of our study is the modest study size, which limited the scope of our analyses to the description of the major patterns (i.e. averages) of twin correlations across the genome. A second limitation is that we did not include DZ twins. The correlation between the phenotypes of MZ twins summarizes the contribution of heritable influences and shared environmental factors to phenotypic variation. It thus remains to be established whether CpGs that were strongly correlated in MZ twin pairs are strongly affected by heritable influences or whether shared environmental influences are also important at these sites. Of interest, a previous twin study of DNA methylation in adipose tissue identified a number of CpGs with evidence for shared environmental effects on DNA methylation [34]. Future studies that include data from both MZ and DZ twin pairs are needed to separate the effects of heritable effects and shared environment on genome-wide DNA methylation profiles in buccal cells. Our results indicate that such studies are worthwhile, as we have shown that

methylation at a number of CpGs is strongly correlated between MZ twins in buccal cells.

We studied DNA methylation extracted from buccal samples, which may be easier to collect than blood samples in e.g. young children, and are therefore well-suited for large-scale studies in humans. A relevant question is how representative DNA methylation extracted from these samples is for DNA methylation variation in other tissues, and whether methylation studies of buccal vs. blood-derived DNA would lead to similar insights. Although DNA methylation patterns are to a large extent tissue-specific [8], and epigenetic changes arising later in life in one tissue may not be detectable in others, epigenetic variation that is established early in development is more likely to be reflected in multiple tissues [4]. Yet, methylation patterns of buccal cells are likely to be more informative to the methylation state of other ectoderm-derived tissues, whereas methylation patterns in blood may be more comparable to other mesoderm-derived tissues. Finally, it may be regarded as an advantage that compared to blood, which consists of many different cell types, buccal samples represent a relatively homogenous sample type [48], in the sense that it consists of only two major cell types, which potentially makes correction for cell types more straightforward. On the other hand, an advantage of blood samples is that they may provide more insight into DNA methylation variation related to immune-system mediated processes in the body, which are important in many diseases. To conclude, blood and buccal samples are both valuable for gaining insight into the overall importance of heritable and environmental factors to DNA methylation variation in the genome, and our study showed that the average genome-wide MZ twin correlation for DNA methylation in buccal cells is similar to the average correlation previously reported for peripheral blood [33].

Conclusions

To summarize, we computed genome-wide MZ twin correlations for buccal DNA methylation level at individual CpGs. Methylation levels in MZ twins were moderately to strongly correlated at CpGs with the largest inter-individual variation, which constituted a relatively small proportion of the CpGs that were measured. The average MZ twin correlation across all CpGs was relatively low (mean rho=0.31), which is similar to findings from previous twin studies [21, 33]. Although most CpGs within CGIs were on average hypomethylated, some of them showed large variation in methylation levels. We observed that CpGs with variable methylation levels were more strongly correlated in MZ twins when located in CGIs compared to CpGs in shores and shelves. CpGs in DHS and TFBS were generally hypomethylated as expected for regulatory active DNA, but CpGs in these regions that were more variably methylated were moderately to strongly correlated in MZ twin pairs, in line with our findings for variably methylated CpGs in general. To conclude, we have shown that in buccal samples from young and adolescent MZ twins, most CpGs show an average methylation level close to 0 or 100% and little inter-individual variation, and a

subset of CpGs show larger variability with evidence for a familial component (DNA sequence variation or shared environment). These findings are relevant for future heritability studies of DNA methylation and for mQTL studies.

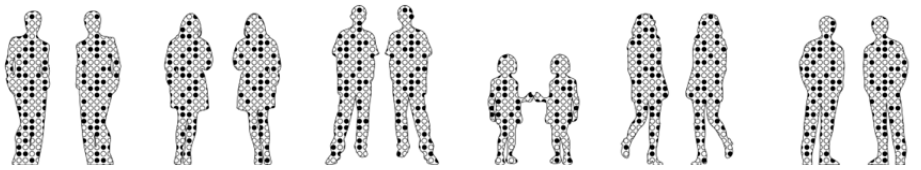
Reference List

1. Hindorff L.A *et al.* A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed January 27 2014.
2. Hindorff,L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* **106**, 9362-9367 (2009).
3. The ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
4. Mill,J. & Heijmans,B.T. From promises to practical strategies in epigenetic epidemiology. *Nat. Rev. Genet.* **14**, 585-594 (2013).
5. Goldberg,A.D., Allis,C.D., & Bernstein,E. Epigenetics: a landscape takes shape. *Cell* **128**, 635-638 (2007).
6. Jones,P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484-492 (2012).
7. Maunakea,A.K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253-257 (2010).
8. Slieker,R.C. *et al.* Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics. Chromatin.* **6**, 26 (2013).
9. Bird,A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6-21 (2002).
10. Horvath,S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13**, R97 (2012).
11. Ziller,M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-481 (2013).
12. Dempster,E.L. *et al.* Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Hum. Mol. Genet.* **20**, 4786-4796 (2011).
13. Kuehnen,P. *et al.* An Alu element-associated hypermethylation variant of the POMC gene is associated with childhood obesity. *PLoS. Genet.* **8**, e1002543 (2012).
14. Talens,R.P. *et al.* Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. *FASEB J.* **24**, 3135-3144 (2010).
15. Bell,J.T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).
16. Heijmans,B.T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *PNAS* **105**, 17046-17049 (2008).
17. Jeffries,A.R. *et al.* Stochastic choice of allelic expression in human neural stem cells. *Stem Cells* **30**, 1938-1947 (2012).
18. Waterland,R.A. *et al.* Maternal methyl supplements increase offspring DNA methylation at Axin Fused. *Genesis.* **44**, 401-406 (2006).

19. van Dongen J., Slagboom,P.E., Draisma,H.H., Martin,N.G., & Boomsma,D.I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **13**, 640-653 (2012).
20. Ye,K. *et al.* Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin. Res. Hum. Genet.* **16**, 1026-1032 (2013).
21. Gordon,L. *et al.* Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res.* **22**, 1395-1406 (2012).
22. Martino,D. *et al.* Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biol.* **14**, R42 (2013).
23. Fraga,M.F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *PNAS* **102**, 10604-10609 (2005).
24. Bocklandt,S. *et al.* Epigenetic predictor of age. *PLoS. One.* **6**, e14821 (2011).
25. Talens,R.P. *et al.* Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell* **11**, 694-703 (2012).
26. Boks,M.P. *et al.* The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS. One.* **4**, e6767 (2009).
27. Coolen,M.W. *et al.* Impact of the genome on the epigenome is manifested in DNA methylation patterns of imprinted regions in monozygotic and dizygotic twins. *PLoS. One.* **6**, e25590 (2011).
28. Gervin,K. *et al.* Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res.* **21**, 1813-1821 (2011).
29. Heijmans,B.T., Kremer,D., Tobi,E.W., Boomsma,D.I., & Slagboom,P.E. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Hum. Mol. Genet.* **16**, 547-554 (2007).
30. Ollikainen,M. *et al.* DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Hum. Mol. Genet.* **19**, 4176-4188 (2010).
31. Wong,C.C. *et al.* A longitudinal study of epigenetic variation in twins. *Epigenetics.* **5**, 516-526 (2010).
32. Kaminsky,Z.A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* **41**, 240-245 (2009).
33. Bell,J.T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS. Genet.* **8**, e1002629 (2012).
34. Grundberg,E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876-890 (2013).
35. Bibikova,M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295 (2011).

36. van Beijsterveldt,C.E. *et al.* The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin. Res. Hum. Genet.* **16**, 252-267 (2013).
37. Estourgie-van Burk,G.F. *et al.* A twin study of cognitive costs of low birth weight and catch-up growth. *J. Pediatr.* **154**, 29-32 (2009).
38. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* **13**, 231-245 (2010).
39. Meulenbelt,I., Droog,S., Trommelen,G.J., Boomsma,D.I., & Slagboom,P.E. High-yield noninvasive human genomic DNA isolation method for genetic studies in geographically dispersed families and populations. *Am. J. Hum. Genet.* **57**, 1252-1254 (1995).
40. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 2013.
41. Chen,Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* **8**, 203-209 (2013).
42. Boomsma,D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22**, 221-227 (2014).
43. Teschendorff,A.E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics.* **29**, 189-196 (2013).
44. Kent,W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).
45. Yuen,R.K., Jiang,R., Penaherrera,M.S., McFadden,D.E., & Robinson,W.P. Genome-wide mapping of imprinted differentially methylated regions by DNA methylation profiling of human placentas from triploidies. *Epigenetics. Chromatin.* **4**, 10 (2011).
46. Pan,H. *et al.* Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics.* **7**, 1173-1187 (2012).
47. Du,P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC. Bioinformatics.* **11**, 587 (2010).
48. Thiede,C., Prange-Krex,G., Freiberg-Richter,J., Bornhauser,M., & Ehninger,G. Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants. *Bone Marrow Transplant.* **25**, 575-577 (2000).

Part 4: Twin studies and complex traits: Future and further considerations



The evolutionary paradox and the missing heritability of schizophrenia

Abstract
Schizophrenia is one of the most detrimental common psychiatric disorders, occurring at a prevalence of approximately 1 %, and characterized by increased mortality and reduced reproduction, especially in men. The heritability has been estimated around 70% and the genome-wide association meta-analyses conducted by the Psychiatric Genomics Consortium have been successful at identifying an increasing number of risk loci. Various theories have been proposed to explain why genetic variants that predispose to schizophrenia persist in the population, despite the fitness reduction in affected individuals, a question known as the evolutionary paradox. In this review, we consider evolutionary perspectives of schizophrenia and of the empirical evidence that may support these perspectives. Proposed evolutionary explanations include balancing selection, fitness trade-offs, fluctuating environments, sexual selection, mutation-selection balance and genomic conflicts. We address the expectations about the genetic architecture of schizophrenia that are predicted by different evolutionary scenarios and discuss the implications for genetic studies. Several potential sources of 'missing' heritability, including gene-environment interactions, epigenetic variation, and rare genetic variation are examined from an evolutionary perspective. A better understanding of evolutionary history may provide valuable clues to the genetic architecture of schizophrenia and other psychiatric disorders, which is highly relevant to genetic studies that aim to detect genetic risk variants.

Based on: van Dongen J., Boomsma DI. The evolutionary paradox and the missing heritability of schizophrenia. Am.J.Med.Genet.B Neuropsychiatr.Genet. 2013; 162B:122-136

Introduction

The evolutionary paradox of common psychiatric disorders

Common psychiatric disorders can be highly detrimental and many are associated with a shorter lifespan [1-4]. Unlike common somatic disorders, common psychiatric disorders often emerge early in the reproductive age [3, 5-9], conferring a substantial reproductive disadvantage [10-14]. Twin and adoption studies have indicated that genetic differences between individuals explain an important part of the variation in risk for many psychiatric disorders [15-20]. Estimates from 'unrelated' subjects suggest that a significant part of the variation in risk can be explained by genome-wide SNPs, with 23% for schizophrenia [21], 38% for bipolar disorder [22], and 32% for major depressive disorder [23]. According to evolution theory the process of natural selection preserves genetic variants associated with survival and reproductive advantage (fitness), while genetic variants associated with low fitness are eliminated from the gene pool [24]. Given that genetic variants associated with reduced fitness are under negative selection pressure, why is it that natural selection has not eliminated genetic variants that predispose to psychiatric disorders? This question has been addressed by many and is known as the evolutionary paradox of psychiatric disorders.

The paradox is most evident for disorders that have a high heritability and are associated with a large fitness reduction. Why are harmful psychiatric disorders that are largely genetic in origin so common? Schizophrenia is among the most heritable psychiatric disorders (heritability~70% [20, 25]) and also among the most severe. It is characterized by positive symptoms (i.e. hallucinations, delusions and racing thoughts), negative symptoms (i.e. poor social functioning, apathy and lack of emotion), and cognitive symptoms (disorganized thoughts, concentration problems, memory problems and difficulty with completing tasks) [26]. The onset is typically in early adulthood, with an earlier onset in men than in women by on average three to four years [27], and is usually followed by a lifelong course of social and professional impairment. Individuals with schizophrenia show lower reproductive success (on average 30-80% relative to controls, with affected males showing a larger reduction in reproductive success than affected females [28]) and suffer from increased mortality due to natural and unnatural causes (suicide in particular) [29]. These disadvantages suggest that risk alleles for schizophrenia should be under negative selection. Yet, the disorder is surprisingly prevalent, affecting approximately 1% of individuals worldwide [3]. A variety of evolutionary hypotheses has been proposed to explain the high prevalence of schizophrenia, despite the associated fitness reduction and high heritability [30, 31].

The evolutionary paradox and the missing heritability: A common ground?

Although schizophrenia appears to be highly heritable, most of the genetic variants remain to be identified. Genetic linkage studies have pointed at various

loci, but these loci were rarely replicated across populations [32]. It was hypothesized that susceptibility to schizophrenia may be mediated by common genetic variants with small individual effects, a view known as the common disease-common variant (CDCV) hypothesis [33, 34]. Genome-wide association studies (GWAS) have identified an increasing number of common variants that appear to modify the risk of schizophrenia but these variants together explain only a small fraction of the total amount of genetic variation that is assumed to underlie the disorder [35-38]. Recent findings from sequencing studies suggest that at least part of the genetic risk for schizophrenia may be related to rare genetic variants that are difficult to detect in GWAS [39-41]. Evidence has emerged for a role of rare structural variants [37, 42] and *de novo* single nucleotide mutations [43].

The genetic properties of populations are the result of natural selection in the past, together with mutation and random drift [44]. A better understanding of the evolutionary history of diseases may provide valuable insights into their genetic architecture and into suitable strategies to identity risk alleles. The hypothesis that common genetic variants influence the risk for common psychiatric disorders relies upon important assumptions about the evolutionary history of psychiatric disorders. Common susceptibility alleles must be evolutionary ancient and cannot have been subject to continuous strong negative selection pressure, since such variants should have reached fixation and no longer contribute to heritable variation in traits [44]. The evolutionary paradox of common psychiatric disorders and the difficulty to identify susceptibility genes ('the missing heritability' [45]) may be closely linked. The difficulty in finding replicable genetic associations for psychiatric disorders that account for a substantial part of disease risk may be explained by a characteristic genetic architecture that has been shaped by evolutionary history [46].

In this review, we address the question how genetic risk for schizophrenia may persist in the population. We consider evolutionary perspectives on schizophrenia, evaluate the usefulness of these theories in terms of explaining the persistence of heritable variation, and discuss various aspects of the genetic architecture of schizophrenia that are predicted under different evolutionary scenarios. Although schizophrenia is the central theme of most evolutionary theories, many of these theories may apply to a broader concept of psychotic illness. Non-organic psychoses (i.e. psychoses in the absence of organic brain disorder) have traditionally been divided into two diagnostic categories; schizophrenia and bipolar disorder [26]. A combination of symptoms that is intermediate of these two categories may be classified as schizoaffective disorder, however, many clinical signs are shared across all psychotic disorders and it is unclear to which degree the different diagnostic categories are aetiologically distinct. There is increasing support that there is at least partial overlap between genetic risk variants for schizophrenia and bipolar

disorder [35, 36] and such overlap may even extend to other psychiatric disorders as well [37].

Schizophrenia as an evolutionary adaptation

It has been suggested that psychiatric disorders should not be regarded as merely harmful conditions. Rather, these conditions, associated traits, or underlying genes may provide certain advantages to affected individuals or their relatives, which may have been favoured throughout evolutionary history [47-50]. Most of such advantages were proposed for cognitive domains. For example, Kellet [48] suggested that personality traits associated with schizophrenia such as inventiveness and the ability to tolerate low levels of stimulation while remaining alert may offer good territorial instincts, which could have been advantageous to territorial animals. Although this theory might explain the persistence of such traits in ancestral times, modern allele frequencies depend mainly on fitness in recent times. Irrespective of whether schizophrenia or associated traits were indeed beneficial in the past, such theories do not clarify why risk alleles for schizophrenia exist today.

Various other advantageous correlated phenotypes have been proposed, which may also be of benefit in modern times, including social skills, creativity, musical skills, intelligence and exceptional abilities. Associations of creativity with psychosis and schizotypy are well-supported by empirical evidence; individuals with a high score on schizotypy or with a history of psychosis on average score higher on measures of creativity and vice versa [51, 51, 52], and individuals with schizophrenia and bipolar disorder are overrepresented in creative and artistic occupations [53].

Several authors have suggested a link between schizophrenia and advantageous somatic characteristics. Among one of the earliest, Huxley *et al* [54] proposed that the disadvantage of schizophrenia susceptibility may be outweighed by advantages such as higher resistance to infection, heat shock and allergies. Several studies have indeed demonstrated immune-related differences in schizophrenia [54-57], but it remains unclear to which degree this characteristic is advantageous. While several studies have indicated that schizophrenia is associated with a lower risk of rheumatoid arthritis [58-60], the incidence of various other autoimmune diseases, including thyrotoxicosis, celiac disease, acquired hemolytic anemia, interstitial cystitis, and Sjogren's syndrome, has been found to be elevated [61].

Another often cited potential benefit associated with schizophrenia is lower susceptibility to cancer. Besides a possible protective effect of antipsychotic drugs in cancer development [62], several characteristics of schizophrenia itself have been proposed to provide an inherent biological protection against cancer, including a protective effect of excess dopamine [63], increased apoptosis [64], and enhanced natural killer cell activity [65]. Some studies have found a lower incidence of cancer among schizophrenic patients compared to the general population [66-68], also after correcting for risk factors

such as age, race, gender, marital status, education and smoking [66]. Yet, other studies have reported a similar or even higher incidence of several types of cancer in schizophrenia [67, 69] and mixed findings for the association between schizophrenia and cancer may be related to numerous confounders that have not been accounted for in all studies [70, 71], of which smoking could be most important [72].

Although schizophrenia may be associated with some positive aspects, these aspects obviously do not outweigh the negative effects of this disorder to affected individuals. Neurocognitive studies have highlighted that schizophrenic patients generally show marked cognitive deficits [73]. Most importantly, data on lifespan and reproduction of patients with schizophrenia show that any benefit experienced by these patients is apparently not enough to prevent them from having a lower life expectancy and fertility compared to the general population. Therefore, possible cognitive or somatic benefits to affected individuals cannot explain the survival of genetic risk variants for schizophrenia.

Schizophrenia as a fitness trade-off at the extreme end of variation

Several authors have suggested that schizophrenia may have arisen as an unfavourable but inevitable (by-) product of human brain evolution. These theories have in common that schizophrenia is approached in the traditional perspective as a disorder, that is as a phenotype that is purely disadvantageous to the affected individual. According to this view, the high prevalence of the disorder is explained by positive selection for genetic variants that allowed for higher-order cognitive functions throughout evolutionary history, despite the cost of predisposing to schizophrenia.

Schizophrenia could represent the extreme end of normal variation in cognitive skills. Farley [74] proposed that schizophrenia may be regarded as an outlier on the normal continuum of social behaviour and as the toll that humans pay for the benefit of adaptive social skills genes. Crow referred to the link between schizophrenia, language dysfunction, and cerebral flexibility to hypothesize that schizophrenia reflects the extreme end of variation underlying language capacity [75-77]. According to Crow, positive selection for cerebral flexibility during human evolution allowed for the emergence of language; however, a by-product of cerebral flexibility was the associated variation in psychological functioning, resulting in personality disorders and schizophrenia at the extremes. Dodgson and Gordon [78] proposed that certain types of hallucinations may be regarded as the evolutionary by-products of a cognitive system designed to detect threat. From an evolutionary perspective, it might be better to mistakenly believe being threatened by an approaching predator than to fail to recognize it if one is really in danger.

Randall emphasized the role of neural connections in the evolution of human brain functions [79, 80]. According to Randal, the random establishment of novel neural pathways throughout development may produce advantageous supernormal connections or non-adaptive misconnections. This “biological trial

and error of connections" may give rise to a range of behavioural variants, including schizophrenia. Horrobin focused on the biochemistry underlying such neural connections, emphasizing the role of phospholipids biochemistry in the evolution of the human brain and in disorders such as schizophrenia [81-83]. According to this theory, a boost of neuronal membrane phospholipid metabolism resulting from the introduction of a larger amount of essential fatty acids in the early human's diet triggered the evolution of enhanced neuronal micro-connectivity. Though increased micro-connectivity may have allowed for the emergence of traits such as creative thinking, the authors propose that increased neuronal connectivity may also predispose to unwanted side-effects such as schizophrenia. Finally, schizophrenia has been proposed to result from delayed cerebral maturation, which may represent a disadvantageous phenotype within the boundaries of normal variation in cerebral maturation [84].

Although the disorder is approached from a different angle, theories that consider schizophrenia to be by-product of evolution actually rely on a similar principle as theories in which schizophrenia is considered as an evolutionary adaptation; both assume that the disorder is somehow linked to beneficial characteristics. If schizophrenia has arisen as a by-product of evolution at the extreme end of variation in 'normal' traits, the question that still remains is why this extreme and maladaptive phenotype persists, or from a genetic perspective; why the genetic variants responsible for this disorder are maintained in the population.

Balancing selection

Most evolutionary perspectives on psychiatric disorders rely on "balancing selection", which refers to a situation where multiple alleles may be maintained in the gene pool, if the genotypes are under different selection pressures, or if different selection pressures act upon an individual allele under different circumstances. One example of balancing selection is presented by antagonistic pleiotropy, where the effect of a genetic variant is associated with both advantageous and disadvantageous traits within the same individual, making it selectively neutral. An example is the P53 gene, which suppresses cancer, thereby increasing survival at younger ages, but also suppresses stem cell proliferation, thereby contributing to the process of aging [85]. The hypothesis that schizophrenia risk alleles are potentially protective to cancer is an example of antagonistic pleiotropy.

Another example of balancing selection is presented by balanced polymorphisms (heterozygote advantage). Thus, it has been proposed that genetic variants that predispose to fitness-reducing psychiatric disorders in homozygotes are maintained in the population because they are associated with a fitness-increasing trait in a large number of carriers (heterozygotes) [47, 54]. A classic example of a disease that is related to a balanced polymorphism is sickle-cell anemia, a severe disease that presents in individuals who are recessively homozygous for the β-hemoglobin gene [86]. The recessive allele is

maintained in the gene pool because it confers resistance to malaria in heterozygous carriers. Likewise, schizophrenia risk alleles could be maintained in the population because they provide beneficial cognitive or somatic traits in unaffected carriers of these alleles. Of note, a genome-wide scan to identify loci that have been subject to balancing selection indicated that balanced polymorphisms are probably rare [87].

Several authors have suggested that schizophrenia persists due to a benefit experienced by family members of affected individuals; e.g. schizophrenia may present in homozygous individuals, while their heterozygous relatives experience superior social skills [49], creativity [47, 50, 88] or academic success [89], thereby enjoying a selective advantage compared to the general population. Support for the link with creativity has been demonstrated for healthy siblings of schizophrenia patients [53]. Yet, similar cognitive deficits as seen in schizophrenia are often present, although to a milder degree, in relatives without a diagnosis of schizophrenia [90, 91]. Both affected individuals and their relatives perform worse than healthy controls on a range of cognitive tasks, with the most pronounced deficits observed in verbal memory, executive functioning and attention [90, 91]). With respect to cancer, conflicting findings have been found in unaffected relatives of schizophrenic patients, with some studies reporting a lower incidence of cancer in relatives [69, 92] and others reporting a higher incidence [93]. Some studies have reported increased fertility in relatives of schizophrenic patients [94, 95], however, most studies have concluded that the fertility of relatives is not sufficient to outweigh the reproductive cost of schizophrenia [11, 13, 96, 97]. So far, there is thus no evidence for the hypothesis that the fitness cost of schizophrenia is outweighed by an advantage experienced by close relatives, although one might argue that a true instance of "heterozygote advantage" is difficult to detect when it is unknown which relatives are heterozygous for the responsible genetic variant.

Cliff-edged fitness refers to the increase in fitness associated with increased expression of a trait up to a certain threshold, above which increased expression of the trait is associated with a sharp drop in fitness [98]. A classic example of the cliff-edged fitness model is provided by the tendency of some birds to lay fewer eggs than they are capable of; birds that lay fewer eggs avoid the risk that all offspring die under conditions of nutritional scarcity [99]. One theory that relies on the cliff-edged fitness model to explain the persistence of schizophrenia addresses the link between schizophrenia and synaptic pruning. Pruning; the selective elimination of weak neuronal connections, is a normal developmental process that occurs predominantly throughout childhood and adolescence. The elimination of little-used synapses improves mental efficiency; however, excessive reduction of synaptic connectivity (over-pruning) may result in spontaneous and autonomous cerebral activity, causing hallucinations and other positive symptoms [100]. The optimum level of pruning might lie just below the threshold above which psychosis may be induced.

Therefore, evolutionary processes may select towards maximal neuritic pruning, despite the potential risk of over-pruning.

The cliff-edged fitness model has also been applied to explain the persistence of schizophrenia at the level of the underlying genes. Thus, a small number of susceptibility alleles may be beneficial to the individual, for example by providing good social skills and theory of mind capacity [98]. Too many susceptibility alleles, however, may be maladaptive and increase the risk of schizophrenia. The cliff-edged fitness model also offers a potential mechanism for the other theories stating that schizophrenia has arisen as a by-product of evolution at the extreme-end of variation in some trait. Yet, the cliff-edged fitness model does not actually solve the paradox, because it is not clear why natural selection would maintain a number of harmful alleles in the population that can lead to schizophrenia in a subset of individuals, and has not rather selected a set of alleles that is beneficial to all individuals.

Another type of balancing selection is frequency-dependent selection, where the fitness of a phenotype depends on its frequency relative to other phenotypes in the population. Positive frequency-dependent selection refers to the situation in which the fitness of a phenotype is increased as it becomes more common. For example, bright warning (aposematic) coloration in a poisonous species is associated with higher fitness when it is common, since predators are more likely to avoid brightly colored individuals if most individuals are brightly colored [101]. Negative frequency-dependent selection refers to the situation where the fitness of a phenotype increases as it becomes less common. For example, female fruit flies prefer males with a rare phenotype, which is called the “rare male advantage” [102].

Selfish gene theory and group advantages

In evolutionary biology, group selection theory refers to natural selection favoring a trait that confers an advantage to the species as a whole, regardless of the effect of the trait on the fitness of individuals within the group [103]. Similarly, the selfish gene theory emphasizes that the preservation of a gene in the gene pool is determined by its ability to proliferate in the population, even if it predisposes the individual who carries it to self-sacrificing behavior [104]. Group selection has for example been put forward as an explanation for the (apparently) altruistic behavior of “helper birds” observed in many bird species, which delay their own reproductive efforts to help raising the offspring of close relatives [105]. Although such behavior decreases the individual birds’ reproductive success and survival, it promotes the survival of young relatives, thereby stimulating the propagation of the family's genes.

The group selection approach has been adopted to explain the persistence of schizophrenia, by suggesting that the characteristics of some affected individuals may confer an advantage to the group. For example, Price and Stevens [106] proposed the group-splitting hypothesis of schizophrenia, which states that schizotypal traits may reflect an ancient form of behavioural specialization for hunting and gathering tribes. This hypothesis relies on the

assumption that in ancient times, proliferating tribal communities had to split from time to time to maintain optimum numbers. According to Price and Stevens, schizotypal traits in certain prominent individuals may have been advantageous to ensure survival of an offshoot group. To illustrate their hypothesis, they suggested that schizotypal traits are often found in charismatic leaders, who use "delusions, paranoia and religious themes to fraction disaffected groups and to seed new cultures". This type of leadership could be regarded as an altruistic behaviour that is maintained by group selection.

A second group selection theory of schizophrenia has highlighted the link between schizophrenia and shamanism [107]. Schizophrenia and associated traits may be advantageous for shamans to perform religious rituals. Since religious rituals and shamans are universally observed across all cultures, this activity may have a genetic basis, and may be relevant to the survival of humankind. This theory was proposed to be supported by the numerous reports of religious-based delusions in psychotic individuals [108].

A third group selection hypothesis of schizophrenia relies on the mechanism of frequency-dependent balancing selection [109]). This hypothesis states that individuals with some genetic susceptibility to schizophrenia may have a survival advantage by possessing a greater sense of individuality, and the ability to "resist shared biases and misconceptions of the group". The authors of this hypothesis propose that the integrity of a group can sustain some betrayal if there are some nonconformists; however, too many would hinder a harmonious society. It was also suggested that this theory may explain why schizophrenia is more prevalent in modern industrial societies [3]. Complex societies are more tolerant to individuals with a greater sense of individuality and may benefit from a modest number of individuals with such characteristics. Yet, as for the other group selection theories, it is difficult if not impossible to assess whether group selection mechanisms contribute to the persistence of risk alleles for schizophrenia.

Sexual selection and the evolution of fitness indicators

Traits with a high heritability that appear puzzling from an evolutionary perspective have been explained by the theory of sexual selection, which refers to the evolutionary selection of traits associated with a reproductive advantage, rather than survival advantage and includes selection due to differences in intra-sexual competitive abilities (intra-sexual competition), and selection due to mate preferences of the choosier sex (intersexual selection). Intersexual selection may stimulate the evolution of traits such as attractive bright plumage in male birds, despite the survival costs that are often associated with such traits[110]. For example, male peacocks have enormous tails that are energetically costly to grow, prevent the bird from flying, and make it an easy target for predators; however, the tails have been favored by sexual selection because they attract females[111].

Several theories have been proposed to explain why such traits attract the opposite sex. The good-genes theory states that individuals of the choosier sex prefer features of mates that advertise genetic fitness[112]. The fitness indicator theory is slightly broader, and states that mate preference traits reveal to potential mates an individual's underlying genetic quality (e.g. mutational load) and condition (e.g., nutritional status and parasite load)[113]. Selection pressures favor individuals who prefer mates with high-quality fitness indicators, since such mates are more likely to successfully produce offspring with high fitness. Fitness indicators may comprise behavioral features such as the courtship songs of birds[114]. Theoretically, the most informative fitness indicators are to a large degree influenced by genetic variation (to allow for advertising genetic fitness) and are at the same time highly sensitive to the environment (to allow for advertising overall fitness).

Visible human body traits may have evolved as fitness indicators, including male facial structure, muscularity and height[115], and female breasts[116]. Shaner *et al* [117] proposed that human mental and behavioral characteristics may also have evolved as fitness indicator. This fitness indicator may involve verbal courtship behaviors (e.g., "attracting mates by telling funny stories with creativity, social sensitivity, and emotional expressiveness"). In individuals with good genes and a favorable prenatal and postnatal environment, neurodevelopmental processes influencing these mental characteristics result in successful courtship behavior. A poor genetic background (due to harmful alleles) or environmental background, however, leads to unsuccessful courtship behavior that repels potential mates. According to Shaner *et al* schizophrenia represents the unattractive and dysfunctional extreme of a highly variable trait shaped by sexual selection. A computational model developed by Del Giudice demonstrates that the sexual selection model is compatible with reduced fertility in families of schizophrenic patients[28]. Yet, the existence of traits that advertise genetic fitness only makes sense as long as harmful genetic variants ('bad genes') are present in the population, and the question is why such variants (still) exist if they produce an 'unattractive' phenotype with lower fitness.

Recessive alleles and epistasis

One factor that is of major importance for the outcome of natural selection on a phenotype is the mode of action of the underlying genes, for example whether causal alleles act in an additive, dominant, or recessive manner [44]. If a maladaptive phenotype results from a dominant allele or additive gene effect at a locus, fitness will be decreased in all carriers of the risk allele and the risk allele will go extinct while the other allele will reach fixation. On the other hand, if a maladaptive phenotype results from the action of a recessive allele, selection will only act on individuals that are homozygous for this allele. In heterozygous individuals, the maladaptive phenotype will not come to expression and the recessive allele will be 'invisible' to natural selection. The

persistence of risk alleles for schizophrenia may therefore in part be explained by the action of recessive alleles that are difficult to eliminate by natural selection. Natural selection similarly acts more slowly on maladaptive traits that result from the interaction of multiple loci (epistasis), in which case the fitness of individuals is determined by the combination of alleles at each locus. For maladaptive traits that arise from epistatic interactions among multiple loci, a 'risk allele' is only associated with reduced fitness in individuals with the specific combination of alleles, while the same allele can be harmless in individuals with other combinations.

Negative selection pressures on recessive alleles can explain the well-described phenomenon of inbreeding depression, where a drop of fitness is observed in the offspring of related parents, because inbreeding increases the chance that offspring are homozygous for deleterious recessive alleles [118, 119]. Schizophrenia is more prevalent in populations with higher levels of inbreeding [120, 121], which supports the role of rare recessive variants, and suggests that these variants may have been subject to negative selection, although other factors, e.g. demographic, social, and economic ones may also influence such outcomes. Additional support for the role of recessive alleles in schizophrenia comes from a study of runs of homozygosity (ROH, long stretches of homozygous polymorphisms), which showed that ROH were more common in schizophrenia patients and found that several specific ROH were present in schizophrenia patients that were very rare in healthy subjects [122].

Fitness trade-offs

From an evolutionary perspective, all phenotypes can be regarded as compromises. Evolution does not strive for perfection. Rather, it drives traits towards an optimum level where fitness and trade-offs are balanced. For example, "our immune systems could be more aggressive, but only at the cost of damaging our own tissue" [123]. Perhaps schizophrenia could have been eliminated by natural selection, but at the expense of loosing valuable cognitive traits. The capacity of natural selection to optimize traits is bounded by some important constraints [123]. Firstly, natural selection represents a stochastic process; certain mutations that could be of benefit to a species may never occur, while harmful mutations can go to fixation by mere chance. Another important constraint is path dependence, which refers to humans being the result of evolutionary forces acting on a continuous lineage from one-celled organism with no fresh start. Therefore, most aspects of the human body depend on aspects that evolved earlier in a way that suboptimal characteristics may not be set straight.

An example of a suboptimal morphological characteristic that has been suggested to reflect path dependence in evolutionary history is the recurrent (inferior) laryngeal nerve, which is a branch of the vagus nerve (tenth cranial nerve) that supplies motor function and sensation to the larynx [124]. The nerve takes a remarkable detour to reach its target: it descends from the brain into

the thorax, loops around the aorta, and travels back to innervate the laryngeal muscles in the neck. This pathway does not seem to make sense but is thought to reflect a design that originates from an ancient ancestor in which major blood vessels were located much closer to the target of this nerve. Thus, the recurrent nerve is also present in fish, in which it is the fourth branch of the vagus nerve innervating one of the posterior gills. The example illustrates that once selection has shaped a trait into a certain direction over evolutionary time, evolution cannot go back in time to reverse it if the trait becomes suboptimal later in evolutionary time. The evolution of the human brain may also have been limited by such constraints.

The triune brain concept is a model in which the human brain contains the evolutionary remnants of three ancestral brains: the reptilian brain (upper brain stem), the paleomammalian brain (limbic area), and the neomammalian brain (cortical region) [125, 126]. According to this model, each successive brain area that was introduced incorporates and modifies previous functions. Millar proposed that the introduction of each successive brain feature may have come with difficulties connecting pre-existing and novel parts, and hypothesized that schizophrenia may reflect a failure of integration between different parts, in particular between the limbus and cortex, an error that may have resulted from a suboptimal brain design due to evolutionary constraints [127]. However, the benefits of having a more complex brain that allowed for novel functions such as language may have outweighed the disadvantage that the design is sensitive to errors. But how can path dependence in the history of brain evolution explain the survival of heritable risk factors for schizophrenia? Why do errors in brain development only lead to problems in some individuals, and if there is a genetic cause for this why is it not wiped out by selection?

If a developmental outcome (brain function) is determined by the interaction of multiple areas and the development of each area is guided by its own genetic information, this suggests that the effect of an allele on an individual's outcome may depend on the presence of (many) other alleles, which implies epistasis. Yet, if there is even a very small difference in the fitness between different combinations of alleles, natural selection generally favours the most fit set of alleles, thus to maintain genetic variation there must be additional factors that play a role. Keller and Miller [128] illustrated the biological network of mechanisms that ultimately produce behaviour by using a watershed analogy: a huge number of "upstream" biological processes (e.g. neuron proliferation, dendritic pruning etc.) eventually flows into all sorts of "downstream" processes (e.g. language, learning capacity etc). One mutation in an upstream process can affect many downstream processes, and one downstream process can be influenced by mutations in many different upstream processes. Keller and Miller [128] suggested that the watershed model predicts that downstream fitness-related traits such as psychiatric disorders have a high heritability because they result from the integration of so many processes and are therefore highly polygenic. Support for the watershed model

comes from the finding that most rare structural variants that contribute to therisk of schizophrenia also increase risk of autism, developmental delay, intellectual disability, epilepsy, somatic dysmorphism, and extremes of body mass and head size [37]. Of interest to the developmental perspective, it was found that a large proportion of *de novo* mutations in schizophrenic patients were present within genes with a higher expression in the early and mid-stage fetal period [43].

Mutation selection balance

Harmful mutations are removed from the gene pool at a rate proportional to their effect on fitness. Yet, novel mutations occur all the time. The polygenic mutation-selection balance hypothesis states that the persistence of schizophrenia and other heritable common mental disorders may be ascribed to the continuous occurrence of new mutations [128, 129]. These mutations are harmful and under negative selection pressure; however, the elimination of fitness reducing mutations may be balanced by the continuous arrival of new mutations. The rate of *de novo* mutations is low (around 1.2×10^{-8} per nucleotide per generation [130]), but mental health may be influenced by many mutations, since the brain depends on the functioning of a large number of genes and their regulatory sequences. It has been estimated that human individuals carry on average 500 mutations with fitness-reducing effects on brain function that have not yet been removed by selection [128].

Polygenic mutation selection balance appears to be the most likely evolutionary explanation for the maintenance of genetic variation for psychiatric disorders with a remarkable reproductive disadvantage, such as schizophrenia [32, 128, 131]. The hypothesis may be supported by the paternal age effect that has been observed for schizophrenia, i.e. the risk of schizophrenia in offspring increases with increasing paternal age [132], and sequencing studies have shown that the age of the father at conception is associated with the number of *de novo* mutations in offspring [43, 130]. The number of *de novo* mutations was shown to increase with two extra mutations per year under a linear model, or doubled every 16.5 years under an exponential model [130].

Developmental instability and phenotypic plasticity

Schizophrenia has been proposed to represent a failure to express precisely an 'intended' developmental design due to perturbations caused by deleterious environmental influences and mutations. The developmental instability model states that during development, environmental and genetic perturbations, including pathogens, toxins, and harmful mutations introduce random effects and imprecision in developmental pathways [133-135]. An example of a feature that is thought to reflect developmental instability is fluctuating asymmetry, which is indexed, for example, by differential ear length. Though the left and right ears are on average of equal size in the population, the ears of an individual may differ slightly, and this could reflect 'noise' in development. Fluctuating

asymmetry may also affect developmental processes in the brain; an example that is thought to illustrate this is hand preference. Schizophrenic patients show greater dermatoglyphic fluctuating asymmetry and more often show atypical (mixed) handedness [136, 137]. Natural selection should favor individuals that are capable of buffering perturbations of developmental pathways, but it has been suggested that an important part of the genetic variation in developmental instability may consist of genetic variation in the ability to resist pathogens [135]. Such variation can be maintained in populations by the process of host-parasite co-evolution [138]. Of interest to this theory, the strongest genetic association for schizophrenia that has thus far emerged from GWAS is in the major histocompatability complex (MHC) region [37].

Rather than being a pathological maladaptation to developmental insults, schizophrenia has also been suggested to represent an adaptively programmed phenotype that is induced by environmental adversity. Many organisms express strikingly variable morphologies in response to variable environmental conditions encountered during development, many of which are thought to represent alternative survival or reproductive strategies. The phenotypic plasticity hypothesis states that exposure to adverse environmental cues during early development may induce alterations in the expression of genes, resulting in a phenotype that is better suited for a stressful or deprived environment [139]. According to Reser [140], some of the core characteristics of schizophrenia that predict social and vocational disabilities in modern times, such as the inability to calm instinctual drives, ignore arousing stimuli, and inhibit transient desires may represent a “defensive, vigilance-based behavioral strategy that alerts the organism to salient, potentially informative stimuli and permits it to be more impulsive and vigilant”. Thus, schizophrenia may be related to physiological and behavioral characteristics that created a fitness advantage in the ancestral environment under conditions of nutritional scarcity and severe environmental stress. The link between schizophrenia and environmental adversity may be supported by several observations. Brain areas in the hippocampus and frontal lobes that become hypometabolic in schizophrenia [141-143] have also been demonstrated to become adaptively hypometabolic in response to starvation, stress and variations in ecological rigor in other mammals and birds [144, 145]. Furthermore, schizophrenia has been linked to exposure to stress during development. Thus, maternal malnutrition [146], maternal stress [147], multiparity [148], short birth interval [149] and stressful postnatal events [150] are all risk factors for schizophrenia, and certain neurophysiological characteristics of schizophrenia can be induced in animals through exposure to prenatal and postnatal stressors [151, 152].

The mismatch hypothesis

While schizophrenia may present a fitness cost in modern societies, this might not have been the case throughout the entire evolutionary history of humankind [153]. This mismatch hypothesis is supported by the fact that the prevalence of

schizophrenia seems to be quite variable across different locations [3], with the highest rates generally found in urban areas [154]. The mismatch hypothesis has been translated in various ways. Firstly, genetic variants that predispose to schizophrenia in modern times may have been adaptive in ancient environments (ancestral adaptation [155]. Secondly, schizophrenia may have been selectively neutral throughout most of human evolutionary history (ancestral neutrality hypothesis [156]. Thirdly, schizophrenia may persist due to variable selective pressures as a result of fluctuating environmental conditions [157].

Epigenetic variation as an evolutionary adaptation to fluctuating environments

How can the persistence of variation in genetic risk be explained if the outcome depends on environmental factors, as predicted by the phenotypic plasticity model? Feinberg and Irizarry [157] proposed a framework that describes how genotypes may influence fitness by regulating the variability of a trait in a population, rather than by influencing the average level of the trait. Thus, the authors suggested that certain loci regulate the plasticity of development of individuals by influencing stochastic variation in gene expression. The suggested loci are CpG islands, the density of which may vary between individuals due to genetic polymorphisms, presenting an inherited basis of variably methylated regions (VMRs). Epigenetic variation at VMRs influence variation in the expression of nearby genes [158], which can give rise to large stochastic variability in phenotypes under a given genetic background.

This type of genetically inherited stochastic variation may provide a powerful mechanism for evolutionary adaptation to variable environments. Using simulations, the authors demonstrate that under fluctuating environmental conditions, a genetically inherited propensity to phenotypic variability increases fitness of a population despite increasing disease susceptibility. In other words, the fitness of individuals may be determined in part by the ability to vary around a certain phenotypic level (or disease risk), rather than by the phenotypic level itself, as this may be the best strategy for the population when environmental factors are not constant. The model of genetically inherited stochastic epigenetic variation provides a molecular mechanism for the phenotypic plasticity paradigm, and may explain how the persistence of disadvantageous traits such as schizophrenia may be stimulated by the pressures of variable environments and at the same time have a heritable basis. Of interest, several studies have reported epigenetic alterations in schizophrenia [159-162]. The paternal age effect is also compatible with the implicated role of epigenetic mechanisms, as increasing paternal is associated with increased risk of epigenetic abnormalities [163].

Sexual and genomic conflicts at imprinted genes

Sexual conflict arises when the two sexes of a species have conflicting optimal reproductive strategies, leading to an evolutionary arms race between males and females [164]. In many species, reproduction is characterized by differential investment of the sexes in their offspring. In mammals, the mother is predominantly responsible for providing resources to offspring pre- and perinatally. As a result, the fitness of maternally derived alleles favors smaller demand on maternal resources, anticipating on the survival of future offspring, than paternally derived alleles, which are associated with high fitness if offspring exploit as much resources from the mother as possible. It is thought that the level of expression that maximizes the fitness of an allele depends on whether the allele was present in a male or a female in the previous generation.

At imprinted genes, the expression pattern of an allele depends on its parent of origin [165]. Typically, one allele is expressed, while the other is transcriptionally silent. The kinship theory of imprinting states that the evolution of imprinted gene expression originates from the conflict of interests between maternally and paternally derived alleles at a locus. Paternally derived alleles favor higher growth rates of offspring and greater demand on maternal resources than maternally derived alleles. Therefore, growth promoting loci are often maternally silenced through imprinting, whereas loci that suppress growth are often paternally silenced [166].

A well-studied example of an imprinted gene is the *IGF2* gene, which encodes a growth promoting factor that is only expressed from the paternal allele. In humans, imprinting defects that activate the silenced maternal allele result in Beckwith-Wiedemann syndrome, an over-growth syndrome characterized by a 50% increase in birth weight [167]. Conversely, imprinting defects that cause the silencing of both alleles give rise to an under-growth syndrome called Silver-Russell syndrome [168]. Imprinting is thought to be particularly important for genes expressed in the placenta, but is also frequently observed for genes with a role in brain development [169, 170]. Thus, the genetic conflict over maternal investment may also affect behavior, cognition and personality of offspring [171].

Badcock and Crespi suggest that the “genetic war” at imprinted genes for brain development may give rise to mental disorders if expression is pushed too far towards the benefit of one of the parental alleles. Paternally biased expression of genes involved in brain development may give rise to a self-oriented child that is highly demanding to its mother, extreme cases being recognized as autism. In line with this theory, Beckwith-Wiedemann patients have an increased risk of autism [172], and individuals with autism tend to show increased *IGF2* expression [173]. Badcock and Crespi hypothesized that small deviations in imprinted gene expression towards a maternal bias may lead to offspring that are energetically ‘cheaper’ and easier behaviorally to mothers, i.e. more placid, less demanding and better capable of interpreting and

understanding the mental states of others. Large maternally biased deviations may lead to psychosis. The authors suggest that several characteristics of autism and psychosis may be regarded as opposites in the context of parental demand, i.e., autistic spectrum conditions are characterized by deficits in theory-of-mind skills, or 'hypo-mentalism', whereas psychotic spectrum conditions involve the opposite: 'hyper-mentalism'. For example, people with autism are characterized by a defective detection of gaze and inability to appreciate what goes on in groups, while individuals with schizophrenia may experience paranoid delusions of conspiracies and being watched by others.

Some empirical support for the theory of Badcock and Crespi is provided by findings in a region that contains several imprinted genes on chromosome 15. Paternally biased expression of this region causes Angelman syndrome, a disorder that is highly comorbid with autism, while maternally biased expression of the same region causes Prader-Willi syndrome, a condition that is often accompanied by psychotic symptoms [174]. Several genes have been found to contribute to risk of autism, schizophrenia and bipolar disorder at the same time [175], but it remains to be established whether these genes are imprinted and whether the expression of the genes may differ across disorders. To conclude, fluctuations in imprinted gene expression that result from the ongoing conflict between reproductive strategies of males and females may contribute to the persistence of fitness decreasing conditions such as schizophrenia. Because epigenetic mechanisms that regulate imprinting can be influenced by genetic variation [157], this theory is compatible with the persistence of heritable variation.

Wilkins addressed the situation of imprinted genes with pleiotropic effects, and suggested that natural selection can systematically cause a loss of fitness and fixation of maladaptive phenotypes due to genomic conflicts [176]. At imprinted loci, selection is driven exclusively by the fitness of the active allele. When a phenotype is influenced by multiple oppositely imprinted loci, an interlocus conflict arises, because any given level of the phenotype will be associated with differential fitness effects for the underlying maternally vs paternally expressed loci. Using a mathematical model to describe a pair of antagonistic imprinted genes (one paternally expressed and one maternally expressed) with pleiotropic phenotypic effects (i.e., both genes influence multiple phenotypic aspects), it was demonstrated that the genomic conflicts that arise can cause natural selection to drive phenotypes away from their optimum values, resulting in a maladaptive, but selectively favored, evolutionary trajectory. According to this theory, mental disorders that occur at high frequencies despite reducing individual fitness, such as schizophrenia, may be related to pleiotropic effects of imprinted gene expression in the brain.

Conclusions

We have presented a variety of evolutionary perspectives of schizophrenia and addressed how they might explain the persistence of genetic risk variants for

schizophrenia in the population. The different evolutionary scenarios make different assumptions about the genetic architecture of schizophrenia, which is relevant to genetic studies that aim to identify genetic variants.

The polygenic mutation-selection balance model offers an explanation for how fitness-reducing genetic variation is maintained in the population; harmful mutations are under negative selection but variation persists, because the removal of alleles is balanced by the occurrence of new mutations in the population. Of interest, a study of the rate of *de novo* occurrence and overall frequency of ten large and rare recurrent DNA copy number variants (CNVs) that have been associated with schizophrenia and other neurodevelopmental disorders indicated that all of these variants are under strong negative selection [177]. The highest selection coefficients were observed for the rarest CNVs, and given the observed selection pressures, *de novo* CNVs at these loci appear to persist in the population for only a few generations. To date, various studies have identified rare SNPs and structural variants that are associated with the risk of schizophrenia [37, 42, 43], and the polygenic mutation-selection model predicts that many more rare genetic variants are likely to contribute to the risk of schizophrenia in the population.

Theories that propose that schizophrenia is in some way linked to adaptive traits, such as social skills, creativity or pathogen resistance suggest that genetic risk variants persist through balancing selection; alleles that confer risk to schizophrenia are maintained in the population because they are of benefit to unaffected individuals. These theories imply that genetic risk variants are common. Other mechanisms that may account for the maintenance of common genetic variants that contribute to disease risk include genomic and sexual conflicts, and the maintenance of genetic variation at CpG sites as an adaptation to fluctuating environmental pressures. It is often thought that environmental approaches cannot explain the paradox of psychiatric disorders, because environmental explanations do not seem to be compatible with the high heritability. However, we have discussed how environmental pressures may in fact contribute to the maintenance of heritable variation in areas that regulate gene expression.

An important difference exists between perspectives that assume that the fitness cost associated with schizophrenia is balanced by increased fitness in relatives, and those that see schizophrenia as a maladaptive by-product of evolution, or fitness trade-off that persists at the benefit of humankind. Distinct mechanisms have been proposed to account for these alternative scenarios, which make different predictions about the genetic architecture of schizophrenia. The “heterozygote advantage” model proposed to account for increased fitness in relatives may be the most convenient evolutionary scenario for genetic association studies (e.g. GWA studies), as cases and controls are expected to differ at common polymorphisms (with adaptive heterozygote genotypes being overrepresented among controls). Yet, current data on fitness of relatives does not appear to support this model, nor does the fact that

genetic variants with large effects on the risk of schizophrenia have not emerged from GWAS.

Perspectives in which schizophrenia is considered to represent a maladaptive by-product of genetic variants required for complex cognitive functions suggest that every individual carries some genetic susceptibility to schizophrenia, and whether individuals are affected may depend on the number of susceptibility alleles they carry (cliff-edged fitness), or on the combination of alleles across multiple loci (epistasis). The cliff-edged fitness paradigm predicts that inter-individual variation in cognitive characteristics is not so much determined by the particular genotype at each locus. Although susceptibility alleles generally give rise to favorable cognitive traits, too many alleles result in schizophrenia. Since many combinations of susceptibility alleles may predispose to schizophrenia, as long as the total number of alleles is large enough, this scenario is a difficult one for case-control association studies or linkage studies. The scenario would in fact be in line with the variable linkage results that have been reported, since affected individuals might be distinguished from non-affected relatives by the additional presence of any copy from the total pool of susceptibility alleles. Thus, although the genetic architecture of schizophrenia under the cliff-edged fitness paradigm could be in line with the CDCV hypothesis, it may explain the limited success of gene finding studies, since the success of detecting susceptibility alleles under this scenario critically depends on the study design. Epistasis likewise implies that single-SNP tests as usually conducted in GWAS are not the optimal strategy to identify common risk variants for schizophrenia, although main effects are expected to exist for individual alleles, which should be identified when sample sizes are large enough. Yet, as the overall effects of these alleles on fitness are expected to be very small from an evolutionary perspective (otherwise they would have been eliminated by natural selection), the individual effects of these alleles when compared between cases and controls are likewise expected to be very small.

The hypothesis that genetic variants that predispose to schizophrenia may have been favoured by natural selection is supported by some empirical evidence. Several genes that have been linked to schizophrenia appear to show signs of positive selection in the human lineage, including Disrupted in Schizophrenia 1 (*DISC1),* Dystrobrevin Binding Protein 1 (*DTNBP1)* and Neuregulin 1 (*NRG1)*, each of which is thought to play an important role in brain development [178]. Several genes related to energy metabolism that have been implicated in the pathophysiology of schizophrenia also appear to have undergone rapid changes in the human lineage [179]. It thus seems that at least some of the variants associated with schizophrenia may have been favoured by natural selection.

Perhaps the strongest evidence for the role of common genetic variants for schizophrenia comes from the estimate that 23% of the variation in disease risk can be explained by all genome-wide SNPs from SNP arrays

together [21]. Though some of this signal may come from rare genetic variants, strong evidence for the importance of common genetic variants is implicated. Common variants can only persist in the population if they are maintained by some sort of balancing mechanism (e.g. antagonistic pleiotropy, fluctuating environments or genomic conflicts), or if the individual effect of risk variants on fitness is so small that relatively high frequencies (e.g. higher than 5 %) can result from random drift. The latter scenario is not unlikely if the risk of schizophrenia in the population is determined by thousands of genetic variants.

Though some theories cannot on their own explain the maintenance of harmful genetic variation in the population, they do provide a framework that allows us to understand how evolution has shaped the brain, and that it is not strange from an evolutionary perspective that this design can be sensitive to errors (e.g. path dependence, fitness trade-offs and developmental instability). Several theories imply that environmental exposures are important, including the mismatch hypothesis, the phenotypic plasticity hypothesis, the fitness indicator theory and the theory of stochastic epigenetic variation. These perspectives are closely linked to each other, and they all predict that gene-environment interactions and epigenetic variation contribute to the etiology of schizophrenia. Genetic variants that predispose to schizophrenia may confer risk to the condition by increasing environmental sensitivity [180] and may therefore be associated with the amount of variation in the phenotype rather than with a specific mean level. Detection of such genetic variants will require novel methodologies and statistical approaches. This evolutionary scenario is also in line with variable linkage and association results across different populations, since different populations may show different levels of the relevant environmental exposures. Part of the heritability of schizophrenia may reflect genetic variation that contributes to the exposure to certain environments (gene-environment correlation), as several 'environmental risk factors' of schizophrenia, for example smoking [181] and cannabis use [182], are known to be to heritable to some extent.

An important point of critique that has been raised in response to evolutionary approaches of schizophrenia is that most take for granted that schizophrenia represents a trait that is 'visible' to natural selection. Thus, one of the core assumptions of evolutionary psychiatry and biomedical psychiatry in general is that schizophrenia and other mental disorders are natural kinds, i.e. bounded entities with discrete biological causes [183]. Given the phenotypic heterogeneity of schizophrenia and the assumed underlying genetic heterogeneity [184, 185], the 'construct' schizophrenia may not have a discrete biological cause, but may rather represent an umbrella concept that covers a heterogeneous group of disorders. The heterogeneity hypothesis predicts that genetic studies of schizophrenia may benefit from focusing on underlying mechanisms with a more homogeneous biological foundation, rather than disease status (affected versus unaffected) as determined by clinical guidelines. Several authors have proposed a unitary model of psychosis [186-188]

and this hypothesis should also be kept in mind when considering the evolutionary history of schizophrenia. In fact, both from evolutionary and genetic perspective, diagnostic categories of psychiatric disorders can be arbitrary, and it seems likely that many genetic variants may contribute to the risk of multiple disorders at the same time.

An important general issue in evolutionary biology is the debate over the level of selection, which refers to the question which level of the biological hierarchy is touched by natural selection. Does natural selection act on organisms, genes, groups, populations or species? Classical Darwinian theory states that it is the differential survival and reproduction of individual organisms that drives the evolutionary process [24]. However, natural selection can operate simultaneously at different levels of the biological hierarchy (multi-level selection theory [189]. In fact, the direction of selection may differ between different hierarchical levels. For example, a trait may be selectively disadvantageous to individuals, but selectively advantageous at the group level. This issue is also important for evolutionary psychiatry and the debate over the evolutionary paradox, i.e. if fitness is reduced in schizophrenia and the fitness of relatives is equal to that of the general population, should it be concluded that schizophrenia risk alleles are merely maladaptive? The pluralist view of natural selection states that the distinction between different levels is a conceptual mistake; different levels of selection represent a matter of perspective rather than empirical fact [190]. Psychiatric disorders are probably subject to a combination of selective pressures, and different evolutionary perspectives may shed light on different aspects and levels of selection.

To conclude, we have discussed a variety of theories that contribute to our understanding of how heritable risk factors for schizophrenia persist in the population, providing insight into the genetic architecture of the disorder and into useful strategies for gene finding. Many of the evolutionary perspectives of schizophrenia may to some extent also apply to other common psychiatric disorders.

Reference List

1. Hiroeh,U., Appleby,L., Mortensen,P.B., & Dunn,G. Death by homicide, suicide, and other unnatural causes in people with mental illness: a population-based study. *Lancet* **358**, 2110-2112 (2001).

2. Joukamaa,M. *et al.* Mental disorders and cause-specific mortality. *Br. J. Psychiatry* **179**, 498-502 (2001).

3. McGrath,J., Saha,S., Chant,D., & Welham,J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **30**, 67-76 (2008).

4. Mouridsen,S.E., Bronnum-Hansen,H., Rich,B., & Isager,T. Mortality and causes of death in autism spectrum disorders: an update. *Autism* **12**, 403-414 (2008).

5. Andrade,L. *et al.* The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *Int. J. Methods Psychiatr. Res.* **12**, 3-21 (2003).

6. Bebbington,P. & Ramana,R. The epidemiology of bipolar affective disorder. *Soc. Psychiatry Psychiatr. Epidemiol.* **30**, 279-292 (1995).
7. Hoek,H.W. Incidence, prevalence and mortality of anorexia nervosa and other eating disorders. *Curr. Opin. Psychiatry* **19**, 389-394 (2006).
8. Kessler,R.C. *et al.* Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62**, 593-602 (2005).
9. Rutter,M. Incidence of autism spectrum disorders: changes over time and their meaning. *Acta Paediatr.* **94**, 2-15 (2005).
10. Baron,M., Risch,N., & Mendlewicz,J. Differential fertility in bipolar affective illness. *J. Affect. Disord.* **4**, 103-112 (1982).
11. Haukka,J., Suvisaari,J., & Lonnqvist,J. Fertility of patients with schizophrenia, their siblings, and the general population: a cohort study from 1950 to 1959 in Finland. *Am. J. Psychiatry* **160**, 460-463 (2003).
12. King,R.B. Subfecundity and anxiety in a nationally representative sample. *Soc. Sci. Med.* **56**, 739-751 (2003).
13. Svensson,A.C., Lichtenstein,P., Sandin,S., & Hultman,C.M. Fertility of first-degree relatives of patients with schizophrenia: a three generation perspective. *Schizophr. Res.* **91**, 238-245 (2007).
14. Williams,K.E., Marsh,W.K., & Rasgon,N.L. Mood disorders and fertility in women: a critical review of the literature and implications for future research. *Hum. Reprod. Update.* **13**, 607-616 (2007).
15. Bulik,C.M. *et al.* Prevalence, heritability, and prospective risk factors for anorexia nervosa. *Arch. Gen. Psychiatry* **63**, 305-312 (2006).
16. Kan,K.J. *et al.* Genetic and Environmental Stability in Attention Problems Across the Lifespan: Evidence From the Netherlands Twin Register. *JAACAP*(2013).
17. Lundstrom,S. *et al.* Autism spectrum disorders and autistic like traits: similar etiology in the extreme end and the normal variation. *Arch. Gen. Psychiatry* **69**, 46-52 (2012).
18. McGuffin,P. *et al.* The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch. Gen. Psychiatry* **60**, 497-502 (2003).
19. Sullivan,P.F., Neale,M.C., & Kendler,K.S. Genetic epidemiology of major depression: review and meta-analysis. *Am. J. Psychiatry* **157**, 1552-1562 (2000).
20. Sullivan,P.F., Kendler,K.S., & Neale,M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187-1192 (2003).
21. Lee,S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet* **44**, 247-250 (2012).
22. Lee,S.H., Wray,N.R., Goddard,M.E., & Visscher,P.M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
23. Lubke,G.H. *et al.* Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biol. Psychiatry* **72**, 707-709 (2012).
24. Darwin,C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*(John Murray, London, 1859).
25. Lichtenstein,P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234-239 (2009).

26. Tamminga,C.A. & Holcomb,H.H. Phenotype of schizophrenia: a review and formulation. *Mol. Psychiatry* **10**, 27-39 (2005).
27. Hafner,H., Maurer,K., Loffler,W., & Riecher-Rossler,A. The influence of age and sex on the onset and early course of schizophrenia. *Br. J Psychiatry* **162**, 80-86 (1993).
28. Del Giudice,M. Reduced fertility in patients' families is consistent with the sexual selection model of schizophrenia and schizotypy. *PLoS One* **5**, e16040 (2010).
29. Brown,S. Excess mortality of schizophrenia. A meta-analysis. *Br. J Psychiatry* **171**, 502-508 (1997).
30. Brüne,M. Schizophrenia-an evolutionary enigma? *Neurosci. Biobehav. Rev.* **28**, 41-53 (2004).
31. Polimeni,J. & Reiss,J.P. Evolutionary perspectives on schizophrenia. *Can. J Psychiatry* **48**, 34-39 (2003).
32. Ng,M.Y. *et al.* Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol. Psychiatry* **14**, 774-785 (2009).
33. Lander,E.S. The new genomics: global views of biology. *Science* **274**, 536-539 (1996).
34. Risch,N. & Merikangas,K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517 (1996).
35. International Schizophrenia Consortium Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).
36. Schizophrenia Psychiatric Genome-Wide Association Study Consortium Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet* **43**, 969-976 (2011).
37. Sullivan,P.F., Daly,M.J., & O'Donovan,M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet* **13**, 537-551 (2012).
38. Wang,K.S., Liu,X.F., & Aragam,N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophr. Res.* **124**, 192-199 (2010).
39. McClellan,J. & King,M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210-217 (2010).
40. Tennessen,J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69 (2012).
41. Veltman,J.A. & Brunner,H.G. De novo mutations in human genetic disease. *Nat. Rev. Genet* **13**, 565-575 (2012).
42. International Schizophrenia Consortium Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237-241 (2008).
43. Xu,B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet* **44**, 1365-1369 (2012).
44. Fisher,R.A. *The genetical theory of natural selection*(Clarendon Press, Oxford, 1930).
45. Maher,B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).
46. Uher,R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. *Mol. Psychiatry* **14**, 1072-1082 (2009).
47. Karlsson,J.L. Genetic association of giftedness and creativity with schizophrenia. *Hereditas* **66**, 177-182 (1977).

48. Kellet,J.M. Evolutionary theory for the dichotomy of the functional psychoses. *The Lancet* **301**, 860-863 (1973).
49. Kuttner,R.E., Lorincz,A.B., & Swan,D.A. The schizophrenia gene and social evolution. *Psychol. Rep.* **20**, 407-412 (1967).
50. Waddell,C. Creativity and mental illness: is there a link? *Can. J. Psychiatry* **43**, 166-172 (1998).
51. Nettle,D. & Clegg,H. Schizotypy, creativity and mating success in humans. *Proc. Biol. Sci.* **273**, 611-615 (2006).
52. O'Reilly,T., Dunbar,R., & Bentall,R. Schizotypy and creativity: an evolutionary connection? *Personality and Individual Differences* **31**, 1067-1078 (2001).
53. Kyaga,S. *et al.* Creativity and mental disorder: family study of 300,000 people with severe mental disorder. *Br. J Psychiatry* **199**, 373-379 (2011).
54. Huxley,J., Mayr,E., Osmond,H., & HOFFER,A. Schizophrenia as a genetic morphism. *Nature* **204**, 220-221 (1964).
55. Muller,N., Riedel,M., Gruber,R., Ackenheil,M., & Schwarz,M.J. The immune system and schizophrenia. An integrative view. *Ann. N. Y. Acad. Sci.* **917**, 456-467 (2000).
56. Schwarz,M.J., Riedel,M., Gruber,R., Ackenheil,M., & Muller,N. Antibodies to heat shock proteins in schizophrenic patients: implications for the mechanism of the disease. *Am. J. Psychiatry* **156**, 1103-1104 (1999).
57. Strous,R.D. & Shoenfeld,Y. Schizophrenia, autoimmunity and immune system dysregulation: a comprehensive model updated and revisited. *J. Autoimmun.* **27**, 71-80 (2006).
58. Gorwood,P. *et al.* Rheumatoid arthritis and schizophrenia: a negative association at a dimensional level. *Schizophr. Res.* **66**, 21-29 (2004).
59. Oken,R.J. & Schulzer,M. At issue: schizophrenia and rheumatoid arthritis: the negative association revisited. *Schizophr. Bull.* **25**, 625-638 (1999).
60. Rubinstein,G. Schizophrenia, rheumatoid arthritis and natural resistance genes. *Schizophr. Res.* **25**, 177-181 (1997).
61. Eaton,W.W. *et al.* Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *Am. J. Psychiatry* **163**, 521-528 (2006).
62. Carrillo,J.A. & Benitez,J. Are antipsychotic drugs potentially chemopreventive agents for cancer? *Eur. J. Clin. Pharmacol.* **55**, 487-488 (1999).
63. Basu,S. & Dasgupta,P.S. Role of dopamine in malignant tumor growth. *Endocrine.* **12**, 237-241 (2000).
64. Catts,V.S. & Catts,S.V. Apoptosis and schizophrenia: is the tumour suppressor gene, p53, a candidate susceptibility gene? *Schizophr. Res.* **41**, 405-415 (2000).
65. Yovel,G. *et al.* Higher natural killer cell activity in schizophrenic patients: the impact of serum factors, medication, and smoking. *Brain Behav. Immun.* **14**, 153-169 (2000).
66. Cohen,M.E., Dembling,B., & Schorling,J.B. The association between schizophrenia and cancer: a population-based mortality study. *Schizophr. Res.* **57**, 139-146 (2002).
67. Goldacre,M.J., Kurina,L.M., Wotton,C.J., Yeates,D., & Seagroat,V. Schizophrenia and cancer: an epidemiological study. *Br. J. Psychiatry* **187**, 334-338 (2005).
68. Mortensen,P.B. The incidence of cancer in schizophrenic patients. *J. Epidemiol. Community Health* **43**, 43-47 (1989).

69. Catts,V.S., Catts,S.V., O'Toole,B.I., & Frost,A.D. Cancer incidence in patients with schizophrenia and their first-degree relatives - a meta-analysis. *Acta Psychiatr. Scand.* **117**, 323-336 (2008).
70. Hodgson,R., Wildgust,H.J., & Bushe,C.J. Cancer and schizophrenia: is there a paradox? *J Psychopharmacol.* **24**, 51-60 (2010).
71. Bushe,C.J. & Hodgson,R. Schizophrenia and cancer: in 2010 do we understand the connection? *Can. J Psychiatry* **55**, 761-767 (2010).
72. de Leon,J. & Diaz,F.J. A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. *Schizophr. Res.* **76**, 135-157 (2005).
73. Heinrichs,R.W. & Zakzanis,K.K. Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology.* **12**, 426-445 (1998).
74. Farley,J.D. Phylogenetic adaptations and the genetics of psychosis. *Acta Psychiatr. Scand.* **53**, 173-192 (1976).
75. Crow,T.J. A theory of the evolutionary origins of psychosis. *Eur. Neuropsychopharmacol.* **5 Suppl**, 59-63 (1995).
76. Crow,T.J. Is schizophrenia the price that Homo sapiens pays for language? *Schizophr. Res.* **28**, 127-141 (1997).
77. Crow,T.J. Schizophrenia as the price that homo sapiens pays for language: a resolution of the central paradox in the origin of the species. *Brain Res. Brain Res. Rev.* **31**, 118-129 (2000).
78. Dodgson,G. & Gordon,S. Avoiding false negatives: are some auditory hallucinations an evolved design flaw? *Behav. Cogn Psychother.* **37**, 325-334 (2009).
79. Randall,P.L. Schizophrenia, abnormal connection, and brain evolution. *Med. Hypotheses* **10**, 247-280 (1983).
80. Randall,P.L. Schizophrenia as a consequence of brain evolution. *Schizophr. Res.* **30**, 143-148 (1998).
81. Horrobin,D.F. Schizophrenia as a membrane lipid disorder which is expressed throughout the body. *Prostaglandins Leukot. Essent. Fatty Acids* **55**, 3-7 (1996).
82. Horrobin,D.F. Schizophrenia: the illness that made us human. *Med. Hypotheses* **50**, 269-288 (1998).
83. Horrobin,D.F. Lipid metabolism, human evolution and schizophrenia. *Prostaglandins Leukot. Essent. Fatty Acids* **60**, 431-437 (1999).
84. Saugstad,L.F. A lack of cerebral lateralization in schizophrenia is within the normal variation in brain maturation but indicates late, slow maturation. *Schizophr. Res.* **39**, 183-196 (1999).
85. Rodier,F., Campisi,J., & Bhaumik,D. Two faces of p53: aging and tumor suppression. *Nucleic Acids Res.* **35**, 7475-7484 (2007).
86. Ashley-Koch,A., Yang,Q., & Olney,R.S. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. *Am. J. Epidemiol.* **151**, 839-845 (2000).
87. Bubb,K.L. *et al.* Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics* **173**, 2165-2177 (2006).
88. Post,F. Creativity and psychopathology. A study of 291 world-famous men. *Br. J. Psychiatry* **165**, 22-34 (1994).
89. Jeste,D.V., Harless,K.A., & Palmer,B.W. Chronic late-onset schizophrenia-like psychosis that remitted: revisiting Newton's psychosis? *Am. J. Psychiatry* **157**, 444-449 (2000).

90. Sitskoorn,M.M., Aleman,A., Ebisch,S.J., Appels,M.C., & Kahn,R.S. Cognitive deficits in relatives of patients with schizophrenia: a meta-analysis. *Schizophr. Res* **71**, 285-295 (2004).
91. Snitz,B.E., Macdonald,A.W., III, & Carter,C.S. Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: a meta-analytic review of putative endophenotypes. *Schizophr. Bull.* **32**, 179-194 (2006).
92. Lichtermann,D., Ekelund,J., Pukkala,E., Tanskanen,A., & Lonnqvist,J. Incidence of cancer among persons with schizophrenia and their relatives. *Arch. Gen. Psychiatry* **58**, 573-578 (2001).
93. Dalton,S.O., Laursen,T.M., Mellemkjaer,L., Johansen,C., & Mortensen,P.B. Risk for cancer in parents of patients with schizophrenia. *Am. J. Psychiatry* **161**, 903-908 (2004).
94. Avila,M., Thaker,G., & Adami,H. Genetic epidemiology and schizophrenia: a study of reproductive fitness. *Schizophr. Res* **47**, 233-241 (2001).
95. Srinivasan,T.N. & Padmavati,R. Fertility and schizophrenia: evidence for increased fertility in the relatives of schizophrenic patients. *Acta Psychiatr. Scand.* **96**, 260-264 (1997).
96. Bassett,A.S., Bury,A., Hodgkinson,K.A., & Honer,W.G. Reproductive fitness in familial schizophrenia. *Schizophr. Res* **21**, 151-160 (1996).
97. MacCabe,J.H., Koupil,I., & Leon,D.A. Lifetime reproductive output over two generations in patients with psychosis and their unaffected siblings: the Uppsala 1915-1929 Birth Cohort Multigenerational Study. *Psychol Med* **39**, 1667-1676 (2009).
98. Nesse,R.M. Cliff-edged fitness functions and the persistence of schizophrenia. *Behav. Brain Sci.* **27**, 862-863 (2004).
99. Liou,L.W., Price,T., Boyce,M.S., & Perrins,C.M. Fluctuating environments and clutch size evolution in great tits. *Am. Nat.* **141**, 507-516 (1993).
100. Hoffman,R.E. & McGlashan,T.H. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated "voices" in schizophrenia. *Am. J. Psychiatry* **154**, 1683-1689 (1997).
101. Endler,J.A. Frequency-dependent predation, crypsis and aposematic coloration. *Philos. Trans. R. Soc. Lond B Biol. Sci.* **319**, 505-523 (1988).
102. Som,A. & Singh,B.N. Evidence for minority male mating success and minority female mating disadvantage in Drosophila ananassae. *Genet. Mol. Res.* **4**, 1-17 (2005).
103. Wilson,D.S. A theory of group selection. *Proc. Natl. Acad. Sci. U. S. A* **72**, 143-146 (1975).
104. Dawkins,R. *The Selfish gene*(Oxford University Press, New York, 2006).
105. Emlen,S.T. & Wrege,P.H. The role of kinship in helping decisions among white-fronted bee-eaters. *Behav. Ecol. Sociobiol.* **23**, 305-315 (1988).
106. Price,J.S. & Stevens,A. The human male socialization strategy set: cooperation, defection, individualism, and schizotypy. *Evol. Hum. Behav.* **19**, 57-70 (1998).
107. Polimeni,J. & Reiss,J.P. How shamanism and group selection may reveal the origins of schizophrenia. *Med. Hypotheses* **58**, 244-248 (2002).
108. Maslowski,J., Jansen van,R.D., & Mthoko,N. A polydiagnostic approach to the differences in the symptoms of schizophrenia in different cultural and ethnic populations. *Acta Psychiatr. Scand.* **98**, 41-46 (1998).
109. Allen,J.S. & Sarich,V.M. Schizophrenia in an evolutionary perspective. *Perspect. Biol. Med.* **32**, 132-153 (1988).

110. Emlen,S.T. & Oring,L.W. Ecology, sexual selection, and the evolution of mating systems. *Science* **197**, 215-223 (1977).
111. Petrie,M., Halliday,T., & Sanders,C. Peahens prefer peacocks with elaborate trains. *Anim. Behav.* **41**, 323-332 (1991).
112. Houle,D. & Kondrashov,A.S. Coevolution of costly mate choice and condition-dependent display of good genes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **269**, 97-104 (2002).
113. Kokko,H., Brooks,R., Jennions,M.D., & Morley,J. The evolution of mate choice and mating biases. *Proc. Biol. Sci.* **270**, 653-664 (2003).
114. Nowicki,S., Hasselquist,D., Bensch,S., & Peters,S. Nestling growth and song repertoire size in great reed warblers: evidence for song learning as an indicator mechanism in mate choice. *Proc. Biol. Sci.* **267**, 2419-2424 (2000).
115. Perrett,D.I., May,K.A., & Yoshikawa,S. Facial shape and judgements of female attractiveness. *Nature* **368**, 239-242 (1994).
116. Barber,N. The evolutionary psychology of physical attractiveness: Sexual selection and human morphology. *Ethol. Sociobiol.* **16**, 395-424 (1995).
117. Shaner,A., Miller,G., & Mintz,J. Schizophrenia as one extreme of a sexually selected fitness indicator. *Schizophr. Res.* **70**, 101-109 (2004).
118. Keller,M.C., Visscher,P.M., & Goddard,M.E. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* **189**, 237-249 (2011).
119. Wright,S. *Evolution and the Genetics of Populations, Vol. 3: Experimental Results and Evolutionary Deductions*(University of Chicago Press, Chicago, 1977).
120. Bittles,A.H. & Black,M.L. Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. *Proc. Natl. Acad. Sci. U. S. A* **107 Suppl 1**, 1779-1786 (2010).
121. Mansour,H. *et al.* Consanguinity and increased risk for schizophrenia in Egypt. *Schizophr. Res.* **120**, 108-112 (2010).
122. Lencz,T. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A* **104**, 19942-19947 (2007).
123. Nesse,R.M. Darwinian medicine and mental disorders. *International Congress Series* **1296**, 94 (2006).
124. Dawkins,R. *The greatest show on earth: the evidence for evolution.*(Free Press, New York, 2009).
125. MacLean,P.D. The triune brain in conflict. *Psychother. Psychosom.* **28**, 207-220 (1977).
126. MacLean,P.D. Evolutionary psychiatry and the triune brain. *Psychol. Med.* **15**, 219-221 (1985).
127. Millar,T.P. Schizophrenia: an etiological speculation. *Perspect. Biol. Med.* **30**, 597-607 (1987).
128. Keller,M.C. & Miller,G. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behav. Brain Sci.* **29**, 385-404 (2006).
129. Pritchard,J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124-137 (2001).
130. Kong,A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).

131. McClellan,J.M., Susser,E., & King,M.C. Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190**, 194-199 (2007).
132. Malaspina,D. *et al.* Paternal age and sporadic schizophrenia: evidence for de novo mutations. *Am. J. Med. Genet.* **114**, 299-303 (2002).
133. Markow,T.A. Genetics and developmental stability: an integrative conjecture on aetiology and neurobiology of schizophrenia. *Psychological medicine* **22**, 295-305 (1992).
134. Markow,T.A. Evolutionary ecology and developmental instability. *Annu. Rev. Entomol.* **40**, 105-120 (1995).
135. Yeo,R.A., Gangestad,S.W., Edgar,C., & Thoma,R. The evolutionary genetic underpinnings of schizophrenia: the developmental instability model. *Schizophr. Res.* **39**, 197-206 (1999).
136. Mellor,C.S. Dermatoglyphic evidence of fluctuating asymmetry in schizophrenia. *Br. J Psychiatry* **160**, 467-472 (1992).
137. Reilly,J.L. *et al.* Dermatoglyphic fluctuating asymmetry and atypical handedness in schizophrenia. *Schizophr. Res* **50**, 159-168 (2001).
138. Woolhouse,M.E., Webster,J.P., Domingo,E., Charlesworth,B., & Levin,B.R. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet* **32**, 569-577 (2002).
139. Feinberg,A.P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**, 433-440 (2007).
140. Reser,J.E. Schizophrenia and phenotypic plasticity: schizophrenia may represent a predictive, adaptive response to severe environmental adversity that allows both bioenergetic thrift and a defensive behavioral strategy. *Med. Hypotheses* **69**, 383-394 (2007).
141. Andreasen,N.C. *et al.* Thalamic abnormalities in schizophrenia visualized through magnetic resonance image averaging. *Science* **266**, 294-298 (1994).
142. Carter,C.S. *et al.* Functional hypofrontality and working memory dysfunction in schizophrenia. *Am. J. Psychiatry* **155**, 1285-1287 (1998).
143. Tamminga,C.A. *et al.* Limbic system abnormalities identified in schizophrenia using positron emission tomography with fluorodeoxyglucose and neocortical alterations with deficit syndrome. *Arch. Gen. Psychiatry* **49**, 522-530 (1992).
144. Jacobs,L.F. The economy of winter: phenotypic plasticity in behavior and brain structure. *Biol. Bull.* **191**, 92-100 (1996).
145. Planel,E., Yasutake,K., Fujita,S.C., & Ishiguro,K. Inhibition of protein phosphatase 2A overrides tau protein kinase I/glycogen synthase kinase 3 beta and cyclin-dependent kinase 5 inhibition and results in tau hyperphosphorylation in the hippocampus of starved mouse. *J. Biol. Chem.* **276**, 34298-34306 (2001).
146. Susser,E. *et al.* Schizophrenia after prenatal famine. Further evidence. *Arch. Gen. Psychiatry* **53**, 25-31 (1996).
147. Khashan,A.S. *et al.* Higher risk of offspring schizophrenia following antenatal maternal exposure to severe adverse life events. *Arch. Gen. Psychiatry* **65**, 146-152 (2008).
148. Hultman,C.M., Sparen,P., Takei,N., Murray,R.M., & Cnattingius,S. Prenatal and perinatal risk factors for schizophrenia, affective psychosis, and reactive psychosis of early onset: case-control study. *BMJ* **318**, 421-426 (1999).
149. Smits,L., Pedersen,C., Mortensen,P., & van Os,J. Association between short birth intervals and schizophrenia in the offspring. *Schizophr. Res.* **70**, 49-56 (2004).

150. Norman,R.M. & Malla,A.K. Stressful life events and schizophrenia. I: A review of the research. *Br. J. Psychiatry* **162**, 161-166 (1993).
151. McClure,W.O., Ishtoyan,A., & Lyon,M. Very mild stress of pregnant rats reduces volume and cell number in nucleus accumbens of adult offspring: some parallels to schizophrenia. *Brain Res. Dev. Brain Res.* **149**, 21-28 (2004).
152. Schneider,M.L., Roughton,E.C., Koehler,A.J., & Lubach,G.R. Growth and development following prenatal stress exposure in primates: an examination of ontogenetic vulnerability. *Child Dev.* **70**, 263-274 (1999).
153. Gluckman,P. & Hanson,M. *Mismatch: Why our bodies no longer fit our world*(Oxford University Press, Oxford, New York, 2006).
154. Kirkbride,J.B. *et al.* Heterogeneity in incidence rates of schizophrenia and other psychotic syndromes: findings from the 3-center AeSOP study. *Arch. Gen. Psychiatry* **63**, 250-258 (2006).
155. Di Rienzo,A. & Hudson,R.R. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet.* **21**, 596-601 (2005).
156. Tooby,J. & Cosmides,L. On the universality of human nature and the uniqueness of the individual: the role of genetics and adaptation. *J Pers.* **58**, 17-67 (1990).
157. Feinberg,A.P. & Irizarry,R.A. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. U. S. A* **107 Suppl 1**, 1757-1764 (2010).
158. Jones,P.A. & Takai,D. The role of DNA methylation in mammalian epigenetics. *Science* **293**, 1068-1070 (2001).
159. Huang,H.S. & Akbarian,S. GAD1 mRNA expression and DNA methylation in prefrontal cortex of subjects with schizophrenia. *PLoS. One.* **2**, e809 (2007).
160. Iwamoto,K. *et al.* DNA methylation status of SOX10 correlates with its downregulation and oligodendrocyte dysfunction in schizophrenia. *J. Neurosci.* **25**, 5376-5381 (2005).
161. Tochigi,M. *et al.* Methylation status of the reelin promoter region in the brain of schizophrenic patients. *Biol. Psychiatry* **63**, 530-533 (2008).
162. Veldic,M. *et al.* Epigenetic mechanisms expressed in basal ganglia GABAergic neurons differentiate schizophrenia from bipolar disorder. *Schizophr. Res.* **91**, 51-61 (2007).
163. Perrin,M.C., Brown,A.S., & Malaspina,D. Aberrant epigenetic regulation could explain the relationship of paternal age to schizophrenia. *Schizophr. Bull.* **33**, 1270-1273 (2007).
164. Chapman,T. Sexual conflict and sex allocation. *Biol. Lett.* **5**, 660-662 (2009).
165. Reik,W. & Walter,J. Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.* **2**, 21-32 (2001).
166. Haig,D. Genomic imprinting and kinship: how good is the evidence? *Annu. Rev. Genet.* **38**, 553-585 (2004).
167. Weksberg,R., Shen,D.R., Fei,Y.L., Song,Q.L., & Squire,J. Disruption of insulin-like growth factor 2 imprinting in Beckwith-Wiedemann syndrome. *Nat. Genet.* **5**, 143-150 (1993).
168. Gicquel,C. *et al.* Epimutation of the telomeric imprinting center region on chromosome 11p15 in Silver-Russell syndrome. *Nat. Genet.* **37**, 1003-1007 (2005).
169. Gregg,C., Zhang,J., Butler,J.E., Haig,D., & Dulac,C. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329**, 682-685 (2010).

170. Wilkinson,L.S., Davies,W., & Isles,A.R. Genomic imprinting effects on brain development and function. *Nat. Rev. Neurosci.* **8**, 832-843 (2007).
171. Badcock,C. & Crespi,B. Battle of the sexes may set the brain. *Nature* **454**, 1054-1055 (2008).
172. Kent,L., Bowdin,S., Kirby,G.A., Cooper,W.N., & Maher,E.R. Beckwith Weidemann syndrome: a behavioral phenotype-genotype study. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, 1295-1297 (2008).
173. Mills,J.L. *et al.* Elevated levels of growth-related hormones in autism and autism spectrum disorder. *Clin. Endocrinol. (Oxf)* **67**, 230-237 (2007).
174. Nicholls,R.D., Saitoh,S., & Horsthemke,B. Imprinting in Prader-Willi and Angelman syndromes. *Trends Genet.* **14**, 194-200 (1998).
175. Carroll,L.S. & Owen,M.J. Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome Med.* **1**, 102 (2009).
176. Wilkins,J.F. Antagonistic coevolution of two imprinted loci with pleiotropic effects. *Evolution* **64**, 142-151 (2010).
177. Rees,E., Moskvina,V., Owen,M.J., O'Donovan,M.C., & Kirov,G. De novo rates and selection of schizophrenia-associated copy number variants. *Biol. Psychiatry* **70**, 1109-1114 (2011).
178. Crespi,B., Summers,K., & Dorus,S. Adaptive evolution of genes underlying schizophrenia. *Proc. Biol. Sci.* **274**, 2801-2810 (2007).
179. Khaitovich,P. *et al.* Metabolic changes in schizophrenia and human brain evolution. *Genome Biol.* **9**, R124 (2008).
180. van Os,J., Kenis,G., & Rutten,B.P. The environment and schizophrenia. *Nature* **468**, 203-212 (2010).
181. Li,M.D., Cheng,R., Ma,J.Z., & Swan,G.E. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction* **98**, 23-31 (2003).
182. Agrawal,A. & Lynskey,M.T. The genetic epidemiology of cannabis use, abuse and dependence. *Addiction* **101**, 801-812 (2006).
183. Adriaens,P.R. Debunking evolutionary psychiatry's schizophrenia paradox. *Med. Hypotheses* **70**, 1215-1222 (2008).
184. Fanous,A.H. & Kendler,K.S. Genetic heterogeneity, modifier genes, and quantitative phenotypes in psychiatric illness: searching for a framework. *Mol. Psychiatry* **10**, 6-13 (2005).
185. Fanous,A.H. & Kendler,K.S. Genetics of clinical features and subtypes of schizophrenia: a review of the recent literature. *Curr. Psychiatry Rep.* **10**, 164-170 (2008).
186. Craddock,N. & Owen,M.J. Rethinking psychosis: the disadvantages of a dichotomous classification now outweigh the advantages. *World Psychiatry* **6**, 84-91 (2007).
187. Nuevo,R. *et al.* The continuum of psychotic symptoms in the general population: a cross-national study. *Schizophr. Bull.* **38**, 475-485 (2012).
188. van Os,J. Is there a continuum of psychotic experiences in the general population? *Epidemiol. Psichiatr. Soc* **12**, 242-252 (2003).
189. Damuth,J. & Heisler,I.L. Alternative formulations of multilevel selection. *Biology and Philosophy* **3**, 407-430 (1988).
190. Wilson,R.A. Pluralism, Entwinement and the Levels of Selection. *Philos. Sci.* **70**, 531-552 (2003).

The continuing value of twin studies in the omics era

Abstract

The classical twin study has been a powerful heuristic in biomedical, psychiatric and behavioral research for decades. Twin registries worldwide have collected biological material and longitudinal phenotypic data on tens of thousands of twins, providing a valuable resource to study complex phenotypes and their underlying biology. In this review, we consider the continuing value of twin studies in the current era of molecular genetic studies. We conclude that classical twin methods combined with novel technologies represent a powerful approach to identify and understand the molecular pathways underlying complex traits.

Based on: van Dongen J., Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nat.Rev.Genet. 2012; 13:640-653

Introduction

The **classical twin design** has been used for decades to estimate the importance of genetic and environmental influences on complex trait variation. Its results have contributed to the awareness that variation in almost every conceivable facet of the human condition is influenced by genetic variation (BOX 1). Traits include intrinsic physical, medical, and biochemical characteristics; life outcome variables such as income, divorce and mortality; and behavioral traits, including apparently trivial ones such as TV watching and internet use. In fact, for many human phenotypes, **heritability** estimates derived from twin studies initially encouraged the search for the responsible genetic variation. Through their collaboration in genome-wide association study (GWAS) consortia, large twin registries (TABLE 1; Supplementary information S1 (table)) are nowadays also making an important contribution to identifying the genetic variation underlying complex traits and disorders.

Twins offer unique opportunities to genetic research that extend beyond the analysis of phenotypic heritability (BOX 2). Twin designs can provide insight into the genetic etiology of disease development over time, and aid in the detection of biomarker profiles for medical conditions. For heritable traits, the comparison of **discordant monozygotic (MZ) twins** represents a powerful improvement over the traditional **case–control study** to search for disease-associated biological marks. The power of this design is illustrated in a recent study that compared the DNA methylation patterns of MZ twins discordant for systemic lupus erythematosus (SLE), which identified several genomic regions in which DNA methylation changes were associated with the disease[1]. Novel applications of the classical twin design can provide fundamental insights into the biological mechanisms underlying complex traits. For example, gene expression studies in monozygotic and dizygotic (DZ) twins have highlighted that variation in genome-wide expression between individuals is due to both genetic and environmental influences, and that the importance of these influences may vary across genes and tissues[2, 3].

This review addresses the continuing value of twin studies. We describe various twin study designs with examples of traditional applications, and we describe how twin approaches are now used for tracing disease-causing mutations and for studying a variety of other newly emerging phenotypes (*e.g.,* the **epigenome, transcriptome, metabolome, proteome** and **microbiome**). We address the use of discordant MZ twins to identify biological mechanisms associated with complex traits, for the inference of causality, and for the genome-wide analysis of genotype-by-environment (G×E) interaction at **variability gene**s. We also discuss various questions that can be addressed by contrasting data from MZ and DZ twins to establish the heritability of biological marks and to unravel the shared etiology of associated traits. A range of twin studies is presented, focusing on the initial level of the DNA sequence, down to its expression and intermediate phenotypes such as metabolites, and ultimately the clinical endpoints of interest**.**

Box 1: The history of the classical twin study

The scientific study of twins goes back to 1875, when Francis Galton published his seminal paper *The history of twins, as a criterion of the relative powers of nature and nurture*[104]. However, Galton was unaware of the distinction between monozygotic (MZ) and dizygotic (DZ) twins. The first papers to contrast the similarity of MZ and DZ twins were published by Poll (1914)[105] and Siemens (1924)[106], whose interest was pigmented nevi (common moles), a phenotype still being studied intensively today because of its importance as a risk factor for melanoma[107]. Not much later, the first twin registries were founded, and power calculations showing that very large sample sizes were needed to obtain reliable estimates of heritability stimulated the foundation of new large registries in the 1980s[108, 109]. Consolidation of these registries, new methods for **zygosity assessment**, and improved survey methods coincided with a growing awareness that genetic influences affected a wide range of traits of biomedical and social significance, and an increase in funding to mount large studies. Worldwide, many countries have now set up twin registries[110-112], which have established collections of longitudinal data in twins across age categories from birth[113] to death[34]. Within the last twenty years, very large twin studies have been carried out through mailed, telephone, and internet surveys. Methods linking twin registry data to national databases containing information on cancer and mortality[114], or outcomes of population screens[115] have provided population-based estimates of heritability on samples as large as 44,000 twin pairs.

The continuing importance of twin study designs

Quantitative analysis of genetic and environmental influences

The classical twin design has traditionally been used to study the heritability of disease-related phenotypes and clinical endpoints (TABLE 2). This design has also been widely applied to estimate the extent to which different traits are influenced by the same or different genetic and environmental factors[4].

Multivariate twin models of symptoms of anxiety and depression, for example, provided evidence that comorbidity of these disorders is due to genetic influences that affect the vulnerability to both disorders, but that different environments determine whether a vulnerable person develops major depression or generalized anxiety disorder[5, 6]. Longitudinal data can be analyzed in a similar way: genetic variation in IQ from age 1 to 16 is largely attributable to the same genetic influences[7], and the increase in heritability[8] is due to amplification of genetic effects with age. The classical twin design can be extended to model polygenic G×E interactions, by testing whether the heritability of a trait varies across different levels of environmental exposures[9]. The heritability of body mass index (BMI) is moderated by physical activity: the higher the level of physical activity, the smaller the genetic influence on BMI[10].

Box 2: The classical twin design

In the classical twin design, the extent to which phenotypic variation in a trait (V_P) is due to genetic (V_G) and environmental (V_E) influences is estimated: $V_P = V_G + V_E$. Genetic variance can be further decomposed into additive genetic variance (V_A) and variance due to non-additive genetic effects (dominance variance, V_D): $V_G=V_A+V_D$. Most twin studies, unless they are very large, consider the narrow-sense heritability (h^2), which refers to the proportion of variation that is due to additive genetic variance: $h^2=V_A/V_P$. Environmental influences (V_E) comprise those that are shared by family members ("the common environment", V_C) and influences that are unique to each individual ("the unique environment", V_U): $V_E=V_C+V_U$.

These unobserved variance components can be estimated from the observed resemblance (*i.e*. the phenotypic covariance) in MZ and DZ twin pairs. Monozygotic (MZ) twins are derived from a single fertilized egg cell and share (nearly) 100% of their segregating genes, while dizygotic (DZ) twins are derived from two distinct zygotes and share on average 50% of their segregating genes. Twins of both types share 100% of the common environment and 0% of the unique environment. Therefore, the phenotypic covariance of MZ twins is expected to equal $V_A + V_D + V_C$ and the phenotypic covariance of DZ twins is expected to equal $0.5V_A + 0.25V_D + V_C$. These expectations are the input (structural equations) for genetic structural equation modeling (GSEM), a technique by which **maximum likelihood** estimates of variance components are obtained from twin data. GSEM obtains the expected MZ and DZ covariances given the equations above, and compares the outcome to the covariances observed in the data. The maximum likelihood estimates of V_A, V_D, V_C, and V_E are those estimates that predict covariances that are most consistent with the observed data. With MZ and DZ data, V_C and V_D cannot be estimated simultaneously. V_D is estimated if there is stronger evidence for non-additive effects (if the MZ correlation is more than twice as large as the DZ correlation) and V_C is estimated if there is stronger evidence for common environmental effects (if the MZ correlation is less than twice as large as the DZ correlation). In extended-twin family designs, the information from additional types of family relations together with the information from twins allows for estimating V_A, V_D, V_C and V_E simultaneously.

In multivariate twin models, extending the set of equations for the expected covariances allows the modelling of the cross-twin–cross-trait covariance, i.e. the covariance of trait 1 in one twin with trait 2 in the co-twin. To estimate to which degree the clustering of different traits or comorbidity of disorders is explained by genetic and environmental influences, the same principles apply as for the expected covariances of twins (*e.g.*, MZ twins are expected to share 100% of genetic influences that overlap between traits while DZ twins are expected to share 50%, resulting in a larger cross-twin–cross-trait covariance for MZ twins if the association between traits has a genetic basis).

Extending twin models with data from other relatives (their parents, siblings, spouses or offspring) enhances statistical power[11] and allows for testing of a much wider range of hypotheses about the causes of human variation, including the role of cultural transmission, social interactions among relatives[12], **genetic non-additivity** and various mechanisms of **assortative mating**[13, 14]. The offspring-of-twins design is a powerful tool for studying intergenerational associations between environmental variables and outcomes in children[15]. Also, comparing the phenotypic similarity of children of female MZ twins (who are socially cousins but genetically half-siblings) to the similarity of children of male MZ twins gives insight into the differential importance of paternal and **maternal effects**; if paternal and maternal effects are equally important, children of male twins and female twins are expected to be equally similar. For birth weight, larger correlations have been observed in children of female twins compared to children of male twins, highlighting the importance of maternal effects for this trait[16].

Classical twin methods continue to be a valuable addition to genetic association studies, for example to establish the proportion of the heritability that can be explained by newly identified SNPs from GWAS[17]. The current discussion about "missing heritability" largely stems from the (often great) disparity between estimates of total heritability from twin studies and the proportion of variance accounted for by SNPs from GWAS[18-20], for which many explanations have been proposed[21] including implications that heritability estimates from twin studies may be too high. In our later section on testing classical assumptions, we discuss the relevance of recent molecular findings in twins in the light of the current discussion on "missing heritability".

The value of discordant twins

Data from MZ and DZ twins allow for the examination of causal relations in the comorbidity of traits. In this case, information from discordant twins is used in a design referred to as the **co-twin control method**. This method was first used to study the association between smoking and lung cancer[22], and has since been applied to investigate a wide variety of medical hypotheses, for example to provide evidence against the efficacy of vitamin C in preventing the common cold[23]. The value of the co-twin control design for distinguishing between associations that reflect causality and associations due to confounding effects of genes or environmental factors (i.e. if two traits are affected by the same genetic or environmental influences, rather than one trait causing the other) is further exemplified by several recent studies on complex traits, as described below.

Experimental studies in which depressive patients are exposed to various types of exercise regimes suggest that regular exercise causes a reduction in anxious and depressive symptoms. To examine whether this causal relationship is present at the population level, twins discordant for

exercise behavior were studied[24]. MZ twins who exercised more than their co-twin did not have fewer symptoms of anxiety and depression. The relationship between exercise behavior and depression was explained by shared genetic influences, rather than by a cause–effect relationship. In another twin study, a reciprocal causal relationship between depression and migraine was revealed[25]. In MZ pairs discordant for depression, only the depressed twin had an increased risk of migraine, and in MZ pairs discordant for migraine, only the twin with migraine had an increased risk of depression. Furthermore, a co-twin control study of anthropometric traits and cancer found a positive correlation between height and risk of breast and ovarian cancer and indicated correlations between BMI and several types of cancer in some population subgroups[26].

The comparison of discordant MZ twins offers an alternative to the traditional case–control study. Here, the primary interest is not the inference of causality, but to identify factors associated with a trait of interest that differ between cases and controls who are perfectly matched for age, sex, and genetic background, and partly matched for early environmental influences.

Molecular phenotypes and the causes of quantitative trait variation

Technological advances allow an assessment of the extent to which twins resemble each other at the level of molecular processes that contribute to their phenotypic similarity[27]. Thereby, the comparison of discordant MZ twins can lead us into novel pathways associated with disease. A unique advantage of the MZ twin design is the ability to study biological discordance against an equivalent genetic background. Divergence of epigenetic profiles in MZ twins depends on the locus and has been documented for both younger and older age groups[28-30]. In fact, differences in DNA methylation and gene expression are already evident in newborn MZ twins[31, 32]. Clearly, environmental and stochastic factors start *in utero* and operate throughout life.

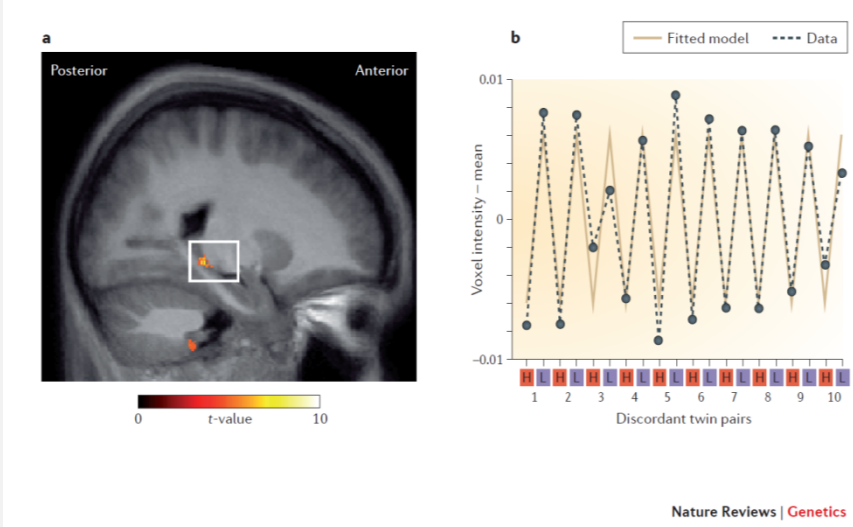
In addition to traditional organismal quantitative traits (such as height and BMI) molecular characteristics — such as gene expression levels, the methylation state of CpG sites in the DNA and the concentration of metabolites in blood and urine — may also be regarded as quantitative traits. Variation in molecular traits measured in groups of MZ and DZ twins can be analyzed using the classical twin method like any other phenotype. Multivariate twin analyses address questions that are not easily resolved in any other study design, such as: to what extent is the epigenetic regulation and expression of genes across genomic regions influenced by shared genetic factors and to what extent is each region influenced by unique factors? And: to what degree do common genetic and environmental mechanisms underlie biological variation across different cells and tissues[33]?

The availability of genome-wide DNA marker data allows for novel approaches to study G×E interactions in which MZ twins can play a vital role. By studying variation in a phenotypic trait of interest in MZ twins, it is possible to see not only whether some genotypes confer higher *levels* of risk for that

trait, but also whether some contribute to its *variability*; high variability in the expression of a trait from a common genetic background could explain phenotypic differences between MZ co-twins. Of interest, genetic and environmental factors may influence disease through different pathways (BOX 3). Twin studies can be used to identify aspects of disease that are most related to the underlying genetic liability of individuals, and thereby help to establish clinical criteria and phenotypic definitions that will facilitate the success of GWAS. Other approaches such as the offspring-of-twins design may provide insight into **trans-generational inheritance** of epigenetic regulation and the importance of maternal effects and **imprinting** on epigenetic marks, though such studies have not yet been published.

An important strength of twin registries lies in the extensive longitudinal collection of data on a variety of phenotypes. Twin studies have indicated that approximately 20-30% of the overall variation in adult lifespan is accounted for by genetic factors[34]. Longitudinal twin studies can be used to identify biomarkers associated with aging: a co-twin control analysis demonstrated that telomere length at advanced age is predictive of survival[35]. MZ twins with the shortest telomeres at baseline had a three-fold greater risk of death during a follow-up period of 7 years than their co-twins with the longest telomere measurements (relative risk (RR) = 2.8). The discordant MZ twin design and the classical twin design have received much interest in recent years for studying molecular biology. The following sections will provide an overview of findings from such studies.

Box 3: The value of twins in neuroimaging genetics



Imaging genetics is a form of association analysis in which the phenotype is a measure of brain structure or function (*e.g.*, physiological response of the brain during information processing)[116, 117]. Brain imaging studies in twins have

contributed substantially to the knowledge that individual differences in brain structure[118] and function[119] are highly heritable. A group of ten male MZ twin pairs and their non-twin brothers had their brains scanned in a functional magnetic resonance imaging (fMRI) study while they had to memorize a short span of digits (digit-memory task)[120]. Before they were asked to recall the digits they memorized, a distraction task was presented in which objects (*e.g.* fruit, vegetables and tools) had to be categorized. When they were distracted by the object categorization task, many men used brain areas associated with language for recalling the digits they had memorized. These men took longer to provide the answer than did those who resorted to a visual-spatial memory system to encode the numbers. MZ twins used the same strategy more often than their non-twin brothers, indicating that there are qualitative differences in how individuals think, and that these differences have a substantial genetic component.

Another design in imaging genetics compares disease-discordant and disease-concordant MZ twins to assess whether genetic and environmental risk factors for psychiatric disorders act on the same brain regions. Comparisons of discordant MZ twins can highlight brain regions that are susceptible to environmental risk factors. Contrasting MZ twins who both score high on the disease phenotype to those who both score low can be used to identify brain characteristics that are related to genetic risk for disease. An imaging study of bipolar disorder that made use of this design found that white matter pathology in the frontal lobe may be central to the genetic risk of developing bipolar disorder, whereas widespread grey matter abnormalities may be more related to environmental effects and reflect effects of the illness itself[121]. A study of MZ twins discordant or concordant for anxious depression found that environmental risk is highlighted in the left temporal lobe (see the figure)[122]. Most notable were the lower grey matter volumes in the left posterior hippocampus, which contains the main afferent and efferent connections of the hippocampus to the rest of the temporal lobe, in high-risk twins from discordant pairs. The Figure illustrates the striking differences in discordant MZ twins, both at the group and individual pair level. The boxed region in **panel A** shows the left parahippocampal area where a significant volume reduction was found in the high risk twin compared to the low risk co-twin from MZ twin pairs discordant for anxious depression. The reduction was not evident in MZ pairs concordant for high risk of depression, when compared to MZ twin pairs concordant for low risk of depression. The within-pair comparison of discordant MZ pairs most likely reveals differences related to environmental exposures, while the between-pair comparison of concordant-high and concordant-low pairs is more likely to reveal differences in genetic vulnerability. Therefore, changes in the left-parahippocampal area may be specific to an environmentally driven etiology of anxiety and depression. Colours represent the effect size (t-value from paired t-test) of the comparison of grey matter volume between discordant

twins. **Panel B** shows the relative responses (individual voxel intensity minus mean voxel intensity in all twins) of ten discordant twin pairs at the most significant voxel in the left parahippocampal area (H= twin with high risk of anxious depression, L=low risk co-twin). Although a significant overall volume reduction was found in the group of discordant pairs, this Figure illustrates that there is large variation in volume difference across individual discordant pairs. Figure is reproduced, with permission, from [122] © (2007) Elsevier.

Tracing the origin of new mutations

Identifying sequence differences between twins

Although MZ twins originate from one zygote, there is some evidence that their somatic cells are not always identical at the DNA sequence level[36]. A study of healthy MZ twins and singletons suggested that **copy number variations** (CNVs) may accumulate with aging in a dynamic fashion[37]. By comparing CNVs in longitudinally collected blood samples of MZ pairs, both increases and decreases in CNV content were found after ten years (between co-twins and within individual twins). This may reflect fluctuations in the proportions of peripheral blood cells carrying aberrant DNA. By comparing copy numbers in buccal cells of twins and their parents, Ehli *et al* found evidence for a pre-twinning *de novo* duplication in a healthy twin pair (present in both twins but not in their parents) and a post-twinning *de novo* deletion in one twin from a pair of twins concordant for attention problems[38]. A comparison of CNVs in the blood of MZ pairs discordant for **congenital diaphragmatic hernia** and **esophageal atresia** found no evidence for structural genomic differences between twins[39]. All these studies made use of microarrays, which cover a limited portion of the total content of structural variation in the genome[40]. The application of whole-genome sequencing techniques may unravel many more sequence differences between MZ twins, including single nucleotide substitutions.

In 2010, Baranzini *et al* published the first study that applied whole-genome-sequencing technology in discordant MZ twins [41]. The study entailed a combination of techniques — including whole-genome sequencing, RNA sequencing and genome-wide SNP microarrays — to measure multiple molecular marks in $CD4^+$ cells from female twins discordant for multiple sclerosis (MS). Only a small fraction of SNPs and structural variants differed within twin pairs, but no differences were replicated across methods. This study should be interpreted as exploratory, however, as only three discordant pairs were studied. Larger studies are needed to establish whether molecular differences may explain discordance for MS and other diseases in MZ twin pairs.

Table 1 | Twin registries worldwide (for a full list see Supplement)

Country	Twin Registry Name	Registry Characteristics	Age	Website	N twins/ subjects (approx.)[A]	N twins/subjects with DNA available (approx.)[A]	Biospecimens (available for at least subset of the sample)
Africa							
Guinea-Bissau	Bandim Health Project Twin Registry	Population-based with ongoing longitudinal data collection	0-30	www.bandim.org	2,500 (twins and singleton controls)	200 twin pairs	Whole blood, plasma
Asia and Australia							
Australia	Australian Twin Registry	Population-based with ongoing longitudinal data collection	0-90	www.twins.org.au	66,000	12,000 (twins and other family members)	Serum, plasma, buccal
China	Chinese National Twin Registry (CNTR)	Population-based with ongoing longitudinal data collection	All	cntr.bjmu.edu.cn	35,000 twin pairs	3,200	Serum, DNA
Korea	South Korean Twin Registry (SKTR)	Volunteer preschoolers, cohort of school children, volunteer young adults	0-30	www.ktrc.org	10,000 twin pairs	800 twin pairs	Hair, saliva
Japan	Keio Twin Registry	Adult and adolescent twins from the general population in the Tokyo area	14-30	totcop.keio.ac.jp; kts.keio.ac.jp; kotrec.keio.ac.jp	4,000 twin pairs (plus other family members)	600 twin pairs	Buccal, blood
Sri Lanka	Sri Lankan Twin Registry	Voluntary twin registry component and a population-based database with ongoing data collection	6-94	www.ird.lk/Twin%20Registry.php	35,000	Plans to collect DNA from 4,000	Buccal

Table 1 | continued							
Europe							
Denmark	The Danish Twin Registry (DTR)	Population-based with ongoing longitudinal data collection	0-107	www.sdu.dk/dtr	170,000	20,000	Serum, plasma, buffy coat, saliva, buccal, urine
Finland	Finnish Twin Cohort study	Population-based with ongoing longitudinal data collection	11-100+	www.twinstudy.helsinki.fi	45,000 (plus family members)	14,600 (twins and family members)	Whole blood, serum, plasma, saliva, urine, fat and muscle by biopsy
Netherlands	Netherlands Twin Register (NTR)	Population-based with ongoing longitudinal data collection	0-100	http://www.tweelingenregister.org/en/	87,500 (plus family members)	18,000	DNA, RNA, cell lines, serum, plasma, buccal, urine, stool
Norway	Norwegian Twin Registry (NTR)	Population-based with ongoing longitudinal data collection	18+	www.fhi.no/twins	40,000	4,800	Whole blood, buccal swabs, plasma
Sweden	Swedish Twin Register (STR)	Population-based with ongoing longitudinal data collection	5-100+	http://ki.se/ki/jsp/polopoly.jsp;jsessionid=acR0ziTHzWEcIO_cNC?l=en&d=9610	194,000	44,600	Whole blood, serum, saliva
UK	TwinsUK registry	Population-based with ongoing longitudinal data collection	18-90	www.twinsuk.ac.uk	12,000	7,000	Whole blood, serum, plasma, buffy coat, saliva, buccal, urine, skin, fat, muscle

Table 1 | continued							
North America							
USA	Mid-Atlantic Twin Registry (MATR)	Population-based, ascertained at birth	0-94	www.matr.vcu.edu	56	1,5	Whole blood, serum, plasma, buffy coat, saliva, buccal
USA	NAS-NRC	Male twins born between 1917-27, both of whom served in the military, mostly during World War II	85-95	iom.edu/Activities/Veterans/TwinsStudy.aspx	31,848	700+	Blood and other materials collected for various investigations
USA	Minnesota Twin Family Study	Ongoing population - based longitudinal study	11-47	mctfr.psych.umn.edu	5,000 (plus family)	10,000 (twins and family members)	Blood-derived or saliva-derived DNA
USA	Wisconsin Twin Panel (WTP)	Population based, longitudinal data, extensive phenotypic characterization, follow up of selected samples	0-23	www.waisman.wisc.edu/twinresearch	19,638 twins (plus parents and sibs)	3,489 (twins, parents, sibs)	Saliva, buccal
South America							
Cuba	Cuban Twin Registry	Population-based with ongoing longitudinal data collection	All	-	55.400 twin pairs	250 twin pairs	Blood-derived DNA

[A] Numbers refer to individual twins (rather than twin pairs) unless indicated otherwise. This table shows a selection of some of the large twin registries worldwide. For a more comprehensive table see Supplementary information S1 (table).

Table 2 | Heritability estimates from twin studies			
Trait	Heritability	Number of twin pairs (or study type for multiple data sets)*	Refs
Anthropometric			
Height	M: 0.87-0.93[A]; F: 0.68-0.90[A]	30111	126
Body Mass Index (BMI)	M: 0.65-0.84[A]; F: 0.64-0.79[A]	37000	127
Birth weight	0.42	2009[B]	128
Metabolic and cardiovascular			
Diabetes, Type 1	0.88	22650	129
Diabetes, Type 2	0.64	13888	130
Coronary heart disease	M: 0.57; F: 0.38	10483	131
Systolic blood pressure	0.42	1617[C]	132
Diastolic blood pressure	0.40	1617[C]	132
Markers for cardiovascular disease in blood		12000 twins	133
High density lipoprotein(HDL) level	0.66		
Low density lipoprotein (LDL) level	0.53		
Triglyceride level	0.54		
Glucose level	0.53		
C-reactive protein (CRP) level	0.43		
Diseases and characteristics of the brain and CNS, psychiatric disorders			
Alzheimer's Disease	0.48	662	134
Parkinson's Disease	0.34	46436 twins	135
Migraine	0.34-0.57[A]	29717	136
Multiple Sclerosis	0.25-0.76[A]	Review	137
Attention-Deficit Hyperactivity disorder	0.76	Review	138
Autism Spectrum Disorders	0.71	11535 twins	139
Schizophrenia	0.81	Meta-analysis	140
Major Depression	0.37	Meta-analysis	141
EEG measures of brain activity		Meta-analysis	119
Alpha power	0.79		
P300 amplitude	0.60		
MRI measures of brain structure		Review	118
Total Brain volume	0.66-0.97		

Table 2 | continued			
Frontal lobe volumes	0.90-0.95		
Hippocampal volumes	0.40-0.69		
Skeletal features and disorders			
Bone mineral density	0.60-0.80	Review	[142]
Osteoarthritis	0.40-0.70	Review	[143]
Rheumatoid arthritis	0.60	13502	[144]
Asthma and pulmonary function			
Asthma	0.60[D]	21135	[145]
Forced Expiratory Volume in one second (FEV(1))	0.61	4314 twins	[146]
Forced Vital Capacity (FVC)	0.55	4314 twins	[146]
Peak Expiratory Flow (PEF)	0.43	4314 twins	[146]
Cancer			
Prostate cancer	0.42	21000	[114]
Breast cancer (in females)	0.27	23788	[114]
Colorectal cancer	0.35	44788	[114]
Aging			
Mortality	0.25	Review	[34]
Telomere length	0.56	175	[35]
Lifestyle and life events			
Exercise participation	0.48-0.71[A]	37051	[89]
Dietary patterns	0.41-0.48	3262[C]	[90]
Smoking initiation	M: 0.37; F: 0.55	Meta-analysis	[147]
Smoking persistence	M: 0.59; F: 0.46	Meta-analysis	[147]
Alcohol abuse/dependence	0.50-0.70	Review	[148]
Stressful life events	0.28	Meta-analysis	[92]

Abbreviations: CNS, central nervous system; EEG, Electroencephalography; **F**,females; **M**,males;* Not that numbers refer to twin pairs unless stated otherwise and most heritability estimates refer to the narrow-sense heritability (h^2, Box 2)

[A] Range of heritabilities from different countries or study samples

[B] Female twin pairs with child (offspring-of-twin design)

[C] Only females

[D] The original paper reports estimates for various age categories from 3-71 years, separately for males and females.

Timing the occurrence of de novo mutations

A unique advantage of studying disease-causing mutations in MZ twins is that the developmental timing of *de novo* mutations[42] may be tracked if DNA from multiple cell lines is available for both twins. Vadlamudi *et al* were able to determine the timing of a mutation in the sodium channel α1 subunit gene (*SCN1A)* that causes **Dravet's syndrome**, by sequencing DNA from several embryonic tissue lineages from a pair of discordant MZ twins[43]. As the mutation was present in all analysed cell lines of the affected twin but not in those of the unaffected co-twin, it was concluded that the mutation had probably occurred at the two-cell stage in the pre-morula embryo. For any disease caused by *de novo* mutations, information about the timing of mutagenesis is of major importance for genetic counselling. Mutations that occurred in parental gametes are associated with a negligible risk of recurrence in additional offspring. By contrast, parental germ-line **mosaicism** for the mutation is associated with a high recurrence risk because many existing parental gametes will carry the mutation.

Phenotypic impact of epigenetic variation

DNA methylation and disease

Besides *de novo* mutations in the DNA, epigenetic variation may be another important source of phenotypic variation and discordance in MZ twins. The following example illustrates this point. In 1997, a pair of MZ girls was born; one of them was healthy but the other had a severe spinal malformation in which the spinal cord was duplicated. This defect resembled a condition in mice with a mutation in the *Axin* gene, but no mutation was found in this gene in the twins. Oates *et al,* however, found increased methylation of CpG sites at the *AXIN* promoter in the affected twin as compared to the unaffected, which may have suppressed gene expression and caused the malformation[44].

Although epigenetic variation has not yet been investigated in large twin studies, several small studies illustrate the promise of the discordant twin design for epigenetics, including studies of Alzheimer's Disease[45], autism[46], Bipolar Disorder[47, 48], birth weight [49], cancer[50], and systemic lupus erythematosus (SLE)[1]. In MZ twins discordant for autoimmune disorders (SLE, rheumatoid arthritis, and dermatomyositis), Javierre *et al* identified a global decrease in DNA methylation (hypomethylation) in SLE-affected twins, and regional DNA methylation changes at 49 genes that were enriched for immune function. Many of the genes that were hypomethylated in SLE-affected twins also showed increased expression compared to the healthy co-twin[1]. Integrated studies of DNA methylation and gene expression in discordant twins[51] are particularly valuable to identify loci at which epigenetic regulation may be associated with disease. Importantly, the dynamic nature of epigenetic variation makes results of epigenetic studies more difficult to interpret compared to genetic studies. Alternatively to being the cause of disease discordance, epigenetic differences may also reflect the effects of disease or the effect of an

event occurring in one twin that triggered both the disease and the epigenetic changes independently. Some twin registries have collected longitudinal biological samples, which allow for identifying epigenetic differences between twins that were already present prior to onset of discordance for some diseases. Functional studies will ultimately be required to verify the effect of epigenetic variation.

The classical twin design provides information about the importance of genetic influences on epigenetic variation: comparison of the level of DNA methylation at the imprinted *IGF2–H19* locus in MZ and DZ twins showed that variation in DNA methylation at this locus is mainly determined by heritable factors before middle age[52]. High heritability of epigenetic variation has also been observed for some other loci[53, 54], although the average heritability across all loci seems to be low[55].

Differential miRNA expression and disease.

The role of **non-coding RNA**s such as **microRNA** (miRNA)[56-58] is relatively unexplored. Sarachana *et al* measured miRNA expression in **lymphoblastoid cell lines** in a sample of MZ twins and sibling pairs discordant for autism and observed differential regulation of a number of miRNA transcripts[59]. For two differentially expressed brain-specific miRNAs, the putative target genes — *ID3* and *PLK2*, which have been implicated in circadian rhythm signaling and modulation of synapses — were validated by experiments involving knockdown or over-expression of these miRNAs. By combining miRNA data and mRNA expression data, dysregulation of miRNA expression was found to contribute to alterations in target gene expression, which in turn may contribute to disease pathology of autism. Te *et al* measured miRNA expression in MZ twins discordant for Lupus Nephritis and observed differential expression of several miRNAs[60]. Among the gene targets of the most important miRNAs were primarily genes with a role in **interferon (IFN) signaling**. Together, these studies indicate that the discordant MZ twin design will be a valuable approach to explore the role of miRNA expression in complex disease.

Gene expression variation: causes and disease links

There is wide variation in the heritability of transcript expression across the genome[2, 61]. To identify **expression quantitative trait loci** (eQTLs), variation in expression across tissues of healthy female twin pairs was investigated in a "matched co-twin analysis" [62]. In the initial stage, SNP associations were tested in one twin of each pair. Although this method of eQTL identification does not require twins the co-twins in this study served to replicate and validate the identified eQTLs, thus providing extra confidence in the findings.

A frequent use of twin studies is to identify gene expression alterations (on a shared genetic background) that are associated with various disease states; such genes may provide mechanistic insight into disease pathogenesis. A study of gene expression in subcutaneous fat of obesity-discordant MZ twins

detected differential expression of a range of genes[63]. Differentially expressed genes included those involved in inflammatory pathways (up-regulated in obese twins) and in mitochondrial branched-chain amino acid (BCAA) catabolism (down-regulated in obese twins). Interestingly, the largest increase in expression in obese twins was reported for the gene encoding the inflammatory cytokine osteopontin (*SPP1*), which has previously been associated with obesity and insulin resistance in mice. Other diseases for which gene expression changes have been identified in discordant MZ twins include rheumatoid arthritis[64], bipolar disorder[65], schizophrenia[66], and type 1 diabetes[67, 68]. A comparison of the skeletal muscle transcriptomes in MZ twins discordant for postmenopausal estrogen-based hormone replacement therapy (HRT) highlights the insights that may be obtained from MZ twins discordant for drug treatment, regarding the long-term effects of drug therapies[69]. Several pathways were differentially regulated in twins who received hormonal treatment, and expression differences correlated significantly with differences in muscle performance between the twins. Large twin studies estimating the heritability of expression of individual transcripts have not yet been published.

Metabolomics

Metabolites may serve as biomarkers of health and disease[70] and can be quantified in body fluids and tissue samples by approaches such as **mass spectrometry (MS)** and **^{1}H NMR spectroscopy.** Nicholson *et al* published the first metabolomics study based on ^{1}H NMR spectroscopic analysis of urine and blood plasma from MZ and DZ twin pairs[71], showing that familial factors (genetic influences and family environment) explain on average 42% of the variation in individual metabolite peak heights in plasma and 30% of the variation in urine. In two GWASs of metabolite profiles, data from twins allowed the proportion of variance in metabolite levels explained by significant SNPs to be compared with the proportion explained by the total genetic or familial variance[72, 73]. Heritability estimates of metabolic measures based on data from 221 MZ and 340 DZ twin pairs ranged between 23%–55% for amino acids and other small-molecule metabolites[72]. Estimates were higher for lipids (48%–62%) and lipoproteins (50%–76%). Although for most direct metabolite measures the total variance explained by significantly associated SNPs was 10% at most, higher estimates of explained variance were observed for certain metabolites ratios. The highest explained variance (25%) was observed for the ratio of linoleic acid to other polyunsaturated fatty acids (LA/PUFA). The twin based heritability for this ratio was 62%, implying that 40% of the total heritability can be ascribed to SNPs, which is high compared most other (clinical) phenotypes.

While traditional enzymatic methods usually provide composite measures of metabolites, ^{1}H NMR gives more detailed insight into the behavior of individual metabolites in pathways. In a direct comparison, similar estimates of heritability were found for most composite lipid measures based on either

enzymatic methods or ^{1}H NMR[72]. This supports the notion that high resolution metabolomics techniques are reliable.

Similarly to differentially expressed genes, differential levels of other molecules can be linked to disease pathogenesis. After detecting differences in serum and fat tissue lipid profiles in MZ twins discordant for obesity[74], Pietiläinen *et al* performed a simulation of **lipid bilayer dynamics** using **lipidomic** and gene expression data from the twins, providing novel functional insights into the biological pathways that underlie adipocyte expansion[75]. This study illustrates how findings from discordant twin studies may encourage and guide further functional or bioinformatic approaches to obtain in-depth mechanistic insights into the pathological mechanisms underlying complex traits and disease.

To date, there have been few proteomic studies in twins. A twin study of serum protein levels, as measured by antibody arrays, found that a relatively small proportion of the variation was attributable to familial factors; however, experimental variation in this study was relatively large[76].

Tissue-specificity of molecular variation

In concordance with the majority of molecular and genetic epidemiological studies, most twin studies have been based on peripheral blood. But how well does a molecular profile in blood cells reflect epigenetic and gene expression changes associated with different phenotypes and diseases in relevant tissues? Epigenetic changes arising at later stages of development and throughout life are more likely to be limited to specific tissues or even cells. DNA methylation profiles of MZ twins discordant for major psychosis suggest that epigenetic changes related to psychosis may be reflected in peripheral blood[77]. In this study, the most significant methylation change in psychosis-affected twins, *i.e.* hypomethylation at the promoter of the *ST6GALNAC1* gene, was also evident in postmortem brain tissues of some psychosis patients. However, large studies are warranted to establish how well molecular profiles in blood reflect those occurring in tissues relevant to disease, since molecular characteristics, particularly epigenetic and gene expression profiles are known to be largely tissue-specific. Although many of the relevant disease tissues are difficult if not impossible to obtain from large groups of living subjects, several twin registries are collecting biological samples from a variety of sources other than blood, including saliva, buccal cells, hair, skin, fat, muscle, urine and stool (Table 1; Supplementary information S1 (table)).

An issue of particular relevance to MZ twins, and possibly also DZ twins[78], is **chimerism**. Twins can exchange fetal blood through vascular connections between their circulatory systems. As a result, MZ twins can be hematopoietic chimaeras, and a variable fraction of cells derived from the hematopoietic stem cells (*e.g.* peripheral blood cells) in each twin may actually originate from the co-twin. This process can have implications for the detection of genetic or epigenetic events related to discordance originating *in utero*, since

some cells in unaffected twins may carry the genetic or epimutation of the co-twin. A study of twins discordant for transient neonatal diabetes mellitus type 1 (TNDM1) found that buccal cells only displayed hypomethylation of the *TNDM1* locus in the affected twins, while the same epigenetic change was evident in blood samples from both twins[79]. The issue may likewise influence the results of DNA sequence analysis of blood samples from MZ twins[80], although a study in healthy twins suggested that MZ twin concordance for SNPs and copy number in blood versus buccal cells is highly similar[81].

Host genetic influences on the microbiome

Studies of the human gut microbiome have revealed considerable variation in the composition of microbial communities between individuals. It remains to be established to what degree this variation is controlled by host genetics[82], but greater similarity has been observed in family members compared to unrelated individuals. A few studies have explored the role of host genetics by comparing various measures of the microbiome in small groups of MZ and DZ twins, but findings have so far been inconclusive, with some studies suggesting that the **microbiota** is slightly more similar in MZ twins compared to DZ twins[83, 84] and others observing comparable levels of similarity of the fecal microbiome of MZ and DZ twins[85]. An important factor in the comparison of similarity of individuals is the level that is compared: the overlap between relatives may be small at the organismal level, but might be larger at relevant functional levels (*e.g.* the degree to which microbial genes and metabolic pathways are shared).

A few studies in twins searched for microbial signatures associated with disease. A comparison of the fecal microbial communities in (concordant) obese and lean MZ twins showed that obesity is associated with various changes, including reduced bacterial diversity and differences in the representation of specific bacterial genes and metabolic pathways[85]. In MZ twins discordant for inflammatory bowel diseases, certain gastrointestinal bacterial populations differed in abundance among individuals with different clinical phenotypes of Crohn's disease, which is relevant to our understanding of the pathogenesis behind inflammatory bowel diseases[86]. MZ twins discordant for ulcerative colitis differed in the composition of the microbiota and in the expression of human RNA transcripts related to oxidative and immune responses in the mucosal epithelium[87]. In affected twins, fewer RNA transcripts correlated with bacterial genera than in unaffected twins, suggesting that ulcerative colitis may be associated with a loss of interaction between the mucosal transcriptional profile and the colonic microbiota.

The interplay of genes and environment

Genetic and environmental influences in many cases do not act independently. Gene–environment correlation (*r*GE) refers to the situation that exposure to certain environments is under genetic control[88]. For instance, twin but also adoption studies have found that lifestyle factors (*e.g.* exercise participation[89]

and diet[90]), life events (*e.g.*, divorce[91]) and life circumstances (*e.g.* family environment and social support[92]) are moderately heritable. Thus, influences that are usually considered as measures of 'environment' might often be better described as external factors that are partly under genetic control[93]. By contrast, G×E interaction refers to the scenario where different genotypes have different reactions to the same environmental exposure[94, 95]. By comparing differences in serum lipid levels in MZ twins across pairs with different genotypes, it was found that the Kidd (*JK*) blood group locus is associated with variability in total cholesterol level[96]. A similar approach was used to test whether interaction of the serotonin transporter gene (*SLC6A4*) length polymorphism with environmental stress is associated with MZ discordance for depression; no evidence was found for this hypothesis[97].

Testing classical assumptions

MZ twins share all their segregating genes while DZ twins share on average 50%

The assumptions of the classical twin method and the interpretation of results have always been a subject of debate (for a detailed discussion of the difficulties related to the concept of heritability, see[98]). A first assumption is that MZ twins are genetically identical, for which it has now been proven that there are exceptions to the rule. Still, the difficulty of various whole-genome sequencing efforts to find any replicable differences between MZ twins[39, 41] suggests that DNA sequence differences between MZ twins are not large, although an exact estimation of somatic sequence variation (given the nontrivial error rate in sequencing itself) has not been reported.

The availability of genome-wide marker data also allows us to address the assumption that DZ twins share on average 50% of their segregating genetic material, by estimating the true amount of genetic material that DZ twins inherited from the same parent (*i.e.* identity-by-descent (**IBD) sharing**). From genome-wide microsatellite marker data, Visscher *et al* demonstrated that the proportion of IBD sharing in most (95%) DZ twins and siblings lies within the range of 42-58%, with an average very close to 50%[99]. Using the empirical IBD measure instead of assumptions about genetic sharing, the heritability of height was estimated at 0.86, *i.e.* highly consistent with results from traditional twin studies, providing perhaps the most pertinent evidence to support the estimates of narrow-sense heritability from twin studies.

MZ twins share environmental influence to the same degree as DZ twins

Now that the classical twin design is being used to study epigenetic variation, it is becoming evident that novel attention has to be paid to the assumption that MZ twins share environmental influences to the same degree as DZ twins. Since MZ twins are derived from a single zygote, they may start out with more similar epigenomes than DZ twins, who originate from two zygotes with unique epigenetic profiles. DZ twins may thus start with more epigenetic differences

than MZ twins due to a cause that is not necessarily related to genetic differences. Although this hypothesis remains to be tested, an important observation in this light has been provided by a comparison of a small groups of MZ twins that were either monochorionicor dichorionic. The DNA methylation profiles of buccal epithelial cells were more similar in dichorionic MZ twins than in monochorionic MZ twins[55], and this may be related to the timing of splitting of the zygote. Thus, differences in epigenetic resemblance of monochorionic and dichorionic twins may be due to epigenetic divergence of embryonic cells that takes place after the blastomeric stage. Although this issue requires further study in larger samples, it illustrates that prenatal developmental processes related to twinning may influence the epigenetic resemblance of twins. Importantly, if MZ twins are epigenetically more similar than DZ twins due to non-genetic causes, the heritability of phenotypes that are epigenetically regulated may be overestimated.

Table 3 | **MZ and DZ twin concordance for complex disease**

	Probandwise concordance[A] (%)		
	MZ twins	DZ twins	refs
Type 1 diabetes	42.9	7.4	[129]
Type 2 diabetes	34	16	[130]
Multiple sclerosis	25.3	5.4	[149]
Crohn's disease	38	2	[150]
Ulcerative colitis	15	8	[150]
Alzheimer's disease	32.2	8.7	[134]
Parkinson's disease	15.5	11.1	[151]
Schizophrenia	40.8	5.3	[152]
Major depression	31.1[B]/47.6[C]	25.1[B]/42.6[C]	[153]
Attention-deficit hyperactivity disorder (ADHD)	82.4	37.9	[154]
Autism spectrum disorders	93.7	46.7	[155]
Colorectal cancer	11	5	[114]
Breast cancer	13[C]	9[C]	[114]
Prostate cancer	18	3	[114]

[A] Defined as 2C/(2C+D), where C is the number of concordant affected twin pairs and D is the number of discordant twin pairs. [B] Concordance in male twin pairs. [C] Concordance in female twin pairs.

Twin concordance and disease liability

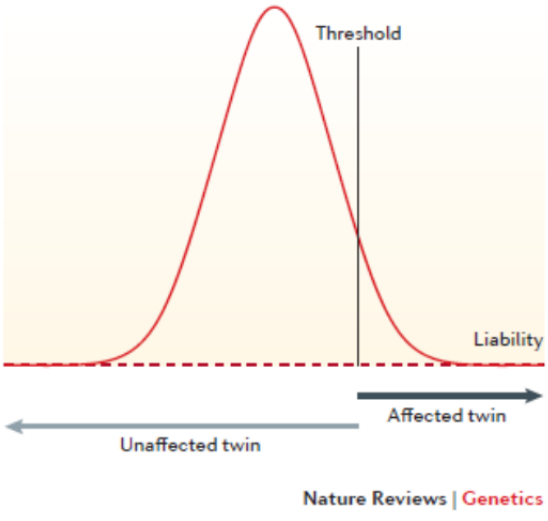
Relationship between heritability and discordance rates in MZ twins

A high concordance of MZ twins on its own does not imply a high heritability, as illustrated by concordance for measles. Before immunization was introduced,

concordance was close to 100% in both MZ and DZ twins[100]. This indicates that, despite the high concordance in MZ twins, genetic differences between individuals actually contribute little to differences in the vulnerability to this infectious disease. Likewise, a high rate of disease discordance in MZ twins does not rule out the importance of genetic influences. Although MZ twins are usually remarkably similar in appearance, MZ twins discordant for disease are often observed (TABLE 3). It is generally assumed that liability to disease is continuous, and disease becomes evident once a threshold is passed. The probability of observing discordant MZ twins thus depends on the heritability of the underlying liability and on the level of the threshold[101]. Especially for rare disorders (for which the threshold is high), many affected MZ twins are discordant even if the heritability is high (*e.g.*, schizophrenia, ADHD, autism, MS or type 1 diabetes,). From the dimensional view of disease liability it also follows that despite striking differences in clinical appearance, discordant MZ twins can be quite similar in terms of underlying disease liability (FIGURE 1).

Trait concordance in MZ twins, penetrance and disease-risk prediction.
The presence of disease-discordant twins indicates that genomes cannot completely predict the disease outcome of individuals, even if most variation in disease outcome *between* individuals is caused by genetic differences. For example, for schizophrenia, despite the high heritability of 80% the probandwise concordance between MZ co-twins is only 40-50%. The fact that MZ twin concordance for common disorders is not always high has important implications for genomic risk prediction and the ethical concerns that have been raised in this light. Even if we knew all the genetic variants that contribute to differences in disease risk between individuals, we would still not be able to predict with certainty the disease risk of all individuals based on their DNA sequence.

Figure 1| Liability threshold model and disease discordance in MZ twins. The liability threshold model assumes that multifactorial diseases result from an underlying continuous character (liability) that is normally distributed in the population[123]. If the combined effects of genetic and environmental influences push an individual's liability across a certain threshold level, the individual is affected. In the population, the proportion of individuals with a liability above the threshold is reflected in the disease prevalence. In discordant MZ twin pairs, only one twin has a liability above the threshold, although the liability of the unaffected twin may also be high. The red arrow displays the potential range of liabilities of affected twins from discordant MZ twin pairs, and the blue arrow displays the potential range of liabilities of unaffected twins. A comparison of MZ twins discordant for congenital diaphragmatic hernia and oesophageal atresia found no differences in genomic structural variation between co-twins[39]. However, structural events in relevant genomic regions that may have contributed to the genetic predisposition of both twins were detected in several pairs; these events were rarely observed in individuals from a healthy control population. A metabolomic study of MZ twins discordant for schizophrenia found that, relative to healthy individuals in concordant pairs, the unaffected twins from discordant female pairs showed similar (though smaller) metabolic changes than the affected co-twins[124]. These examples illustrate that the liability of unaffected twins from discordant pairs may also be elevated. However, this feature does not argue against the value of studying discordant MZ twin pairs to search for the molecular events that caused the affected twin to pass the threshold, or events that protected the unaffected twin. Of interest, a study of neurofibromatosis type 1 (NF1) in MZ twins with the same causal mutation in the *NF1* gene but highly variable disease phenotypes revealed considerable variation between twins in DNA methylation at the *NF1* gene[125].



Conclusions

Insights that can be obtained from twin studies extend far beyond the classical estimates of heritability. Traditional comparisons of the phenotypic resemblance of twins have been extended to studies of molecular variation across biological samples, providing functional insights into the underlying biology of heritable traits. The study of discordant MZ twins is a powerful method to identify DNA sequence variants, epigenetic variation, and metabolites associated with disease.

One might feel that there are few aspects of the human condition that have not been investigated in twins; however, new aspects emerge all the time. We have emphasized the value of twin studies to refine phenotypic and clinical definitions and to evaluate biomarkers for disease, but the use of twins can go even further. In recent years, political scientists, sociologists and even economists have become engaged in twin studies. A study of MZ twins who were infected with HIV through blood-transfusion at birth but who had strikingly different clinical outcomes used the identical genetic background of twins as a model to study the evolutionary processes and population dynamics that shape viral diversity[102].

In the coming years, longitudinal phenotypic information coupled with biological material collected by worldwide twin registries (TABLE1 ; Supplementary information S1 (table)) will be an important resource for large-scale molecular studies. To make optimal use of genetic data collected within twin registries, methods for family-based association analysis are being explored[103]. With the increasing interest in rare genetic variants, there may be renewed interest in linkage studies, in which DZ twins can play an important role. Linkage analysis in DZ twins, contrary to the analysis of non-twin siblings, is not affected by age differences within pairs and is less likely to suffer from non-paternity. Next-generation sequencing across multiple tissues and cell types will facilitate the detection of genome-wide SNPs, CNVs and epigenetic variation in discordant twins at an unprecedented scale, suggesting that twins will continue to provide valuable insights to human genetics.

Glossary

Classical twin design

The approach used to estimate the importance of genetic and environmental influences on complex trait variation. The estimate of heritability is based on a comparison of resemblance in monozygotic twins (who share nearly all of their genetic material) and dizygotic twins (who share, on average, half of their segregating genetic material).

Heritability

The proportion of variation in a trait that is due to heritable differences between individuals in a population, i.e. the proportion of variation due to additive genetic effects (narrow-sense heritability) or the proportion of variation due all genetic effects (broad-sense heritability).

Discordant monozygotic twins
(Discordant MZ twins). Twins who derive from a single fertilized egg cell but are dissimilar for a certain characteristic or disease. By contrast, concordant MZ twins are phenotypically similar.
Case–control study
The comparison of individuals with a trait or disease of interest (cases) to controls to identify genes or other aspects associated with the trait. Cases and controls can be unrelated or can be relatives (within-family case-control design).
Epigenome
The entire collection of epigenetic marks, including DNA methylation and histone-modifications, that regulate the expression of the genome. In contrast to the genome, the epigenome is specific to each cell.
Transcriptome
The total set of RNA transcripts that are produced in a cell or tissue by transcription of DNA.
Metabolome
The total set of small molecules (*e.g.*, lipids, amino acids, and sugars) that are the reactants, intermediate or end products of cellular metabolism and that are present in a cell, tissue, or complete organism.
Proteome
The entire complement of proteins that are present in a cell, tissue, or complete organism.
Microbiome
The entire set of genomes of micro-organisms (*e.g.* bacteria, fungi and viruses) that are present in a certain environment, for example in the human gut.
Multivariate twin models
Models used for the simultaneous analysis of multiple traits measured in MZ and DZ twins to estimate the importance of genetic and environmental influences shared ("overlapping") between traits in explaining their clustering, comorbidity or covariance.
Variability genes
A gene that contributes to the variation in a phenotype. The genotypes are associated with phenotypic variance rather than with the mean level or frequency of the trait.
Genetic non-additivity
Refers to genetic effects that contribute to the phenotypic variance in a non-additive manner. These include the effects of interacting alleles at a single locus (dominance) and interactions between different loci (epistasis).
Assortative mating
Refers to the situation whereby a trait is correlated in spouses because it influences partner choice (phenotypic assortment), or because it correlates with certain environments that influence partner choice (social homogamy). It is also called non-random mating.

Maternal effects
Effects that are transmitted from mother to offspring, including genetic effects. The phenotype in offspring can be influenced by: the maternal allele, mitochondrial inheritance, the effects of the prenatal environment (*e.g.* nutrient supply *in utero*), or the maternal supply of RNA or proteins to the egg cell.
Co-twin control method
Method to examine the associations between traits using discordant twins. If MZ twins discordant for trait 1 are also discordant for trait 2, the association between these traits is unlikely to be confounded by underlying shared genetic or early environmental influences.
Trans-generational inheritance
The transmission of a trait across generations (genetic or cultural inheritance). Epigenetic variation may also be transmitted across generations.
Imprinting
The mechanism that can occur at some loci to silence the expression of one of the two alleles, depending on the parent-of-origin of the allele.
Copy number variation
CNV. It refers to large DNA segments (over 1 kb) of which the number of copies is variable (*e.g.* between individuals or between cells within an individual), for example insertions, deletions and duplications.
Congenital diaphragmatic hernia
A birth defect that is characterized by malformation of the diaphragm, lung hypoplasia and pulmonary hypertension.
Oesophageal atresia
A congenital malformation of the esophagus in which the esophagus does not form an open passage to the stomach and may be connected to the trachea.
Dravet's syndrome
A childhood-onset epileptic encephalopathy, also called severe myoclonic epilepsy of infancy.
Mosaicism
The situation where the tissue of an individual consists of two or more genetically distinct cell lines due to somatic mutation, but originally derived from one (genetically homogeneous) zygote.
Non-coding RNAs
RNA transcripts that are not translated into protein but probably serve a regulatory function.
microRNA
A type of non-coding RNA with an average length of 22 nucleotides that has been suggested to play an important role in post-transcriptional gene regulation networks.
Lymphoblastoid cell lines
Cell lines derived from lymphoblasts, which are immortalized, cultured and stored to provide a renewable source of DNA and RNA.

Interferon (IFN) signaling
A signaling system for communication between cells that is involved in the immune response to pathogens and tumors
Expression quantitative trait loci
(eQTLs). Genomic regions that are associated with the level of expression of an RNA transcript. eQTLs can be tissue-specific.
Mass spectrometry
Technique to determine the mass-to-charge ratio of ions (charged particles) based on their separation in an electromagnetic field. The measured ratios and their relative intensities provide information about both identity and abundance of the molecules that gave rise to the ions.
^{1}H NMR spectroscopy
Metabolomics technique providing information about structure and quantity of hydrogen-containing molecules. It is based on the absorption and emittance of radiofrequent energy by hydrogen atoms when placed in a strong magnetic field, with wavelengths depending on the atoms' position in the molecule.
Lipid bilayer dynamics
The dynamic properties of lipid bilayer membranes such as thickness, fluidity, and permeability, that influence the physiological properties of a cell.
Lipidomics
The comprehensive study of the entire set of lipids in biological systems, such as cells, tissues and organs, using metabolomics techniques
Microbiota
The collection of all micro-organisms living in a certain environment, for example the human gut.
Identity-by-descent (IBD) sharing
(IBD sharing). Refers to the proportion of alleles in two individuals that are derived identical by descent from a common ancestor
Monochorionic twins
Twins that share the outer membrane (chorion) surrounding the embryos *in utero*. Monochorionic monozygotic twins result when the zygote splits after day 3 after fertilization.
Dichorionic twins
Twins that do not share the chorion surrounding the embryos *in utero*. Dizygotic twins are always dichorionic. Dichorionic monozygotic twins result when the zygote splits early after fertilization.
Chimerism
The situation where an individual carries some of the genetic material originating from another individual (*e.g.,* originating from the co-twin or originating from the mother).
Zygosity assessment
The assessment whether same-sex twins are monozygotic or dizygotic is often based on the comparison of DNA markers, or alternatively on standardized questionnaires.

Maximum likelihood

Maximum likelihood estimation obtains estimates of population parameters from a dataset by computing the probability of obtaining the observed data (likehood) for a range of different parameter values, and evaluating for which values the probability of observing the data is highest.

Reference List

1. Javierre,B.M. *et al.* Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.* **20**, 170-179 (2010).
2. McRae,A.F. *et al.* Replicated effects of sex and genotype on gene expression in human lymphoblastoid cell lines. *Hum. Mol. Genet.* **16**, 364-373 (2007).
3. York,T.P. *et al.* Epistatic and environmental control of genome-wide gene expression. *Twin Res. Hum. Genet.* **8**, 5-15 (2005).
4. Martin,N.G. & Eaves,L.J. The genetical analysis of covariance structure. *Heredity* **38**, 79-95 (1977).
5. Kendler,K.S., Heath,A.C., Martin,N.G., & Eaves,L.J. Symptoms of anxiety and symptoms of depression: same genes, different environments? *Arch. Gen. Psychiatry* **44**, 451-457 (1987).
6. Middeldorp,C.M., Cath,D.C., Van Dyck,R., & Boomsma,D.I. The co-morbidity of anxiety and depression in the perspective of genetic epidemiology. A review of twin and family studies. *Psychol. Med.* **35**, 611-624 (2005).
7. Brant,A.M. *et al.* The developmental etiology of high IQ. *Behav. Genet.* **39**, 393-405 (2009).
8. Haworth,C.M. *et al.* The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Mol. Psychiatry* **15**, 1112-1120 (2010).
9. Purcell,S. Variance components models for gene-environment interaction in twin analysis. *Twin Res.* **5**, 554-571 (2002).
10. Mustelin,L., Silventoinen,K., Pietiläinen,K., Rissanen,A., & Kaprio,J. Physical activity reduces the influence of genetic effects on BMI and waist circumference: a study in young adult twins. *Int. J. Obes.* **33**, 29-36 (2008).
11. Posthuma,D. & Boomsma,D.I. A note on the statistical power in extended twin designs. *Behav. Genet.* **30**, 147-158 (2000).
12. Eaves,L.J. Inferring the causes of human variation. *J. R. Stat. Soc. Ser. A* **140**, 324-355 (1977).
13. Reynolds,C.A., Baker,L.A., & Pedersen,N.L. Models of spouse similarity: applications to fluid ability measured in twins and their spouses. *Behav. Genet.* **26**, 73-88 (1996).
14. van Grootheest,D.S., van den Berg,S.M., Cath,D.C., Willemsen,G., & Boomsma,D.I. Marital resemblance for obsessive-compulsive, anxious and depressive symptoms in a population-based sample. *Psychol. Med.* **38**, 1731-1740 (2008).
15. Magnus,P., Berg,K., & Bjerkedal,T. No significant difference in birth weight for offspring of birth weight discordant monozygotic female twins. *Early Hum. Dev.* **12**, 55-59 (1985).
16. Nance,W.E., Kramer,A.A., Corey,L.A., Winter,P.M., & Eaves,L.J. A causal analysis of birth weight in the offspring of monozygotic twins. *Am. J. Hum. Genet.* **35**, 1211-1223 (1983).

17. Vrieze,S.I. *et al.* An Assessment of the Individual and Collective Effects of Variants on Height Using Twins and a Developmentally Informative Study Design. *PLoS Genet.* **7**, e1002413 (2011).
18. Maher,B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).
19. Visscher,P.M., Brown,M.A., McCarthy,M.I., & Yang,J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7-24 (2012).
20. Yang,J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565-569 (2010).
21. Zuk,O., Hechter,E., Sunyaev,S.R., & Lander,E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A* **109**, 1193-1198 (2012).
22. Friberg,L., Cederlof,R., Lundman,T., & Olsson,H. Mortality in smoking discordant monozygotic and dizygotic twins. A study on the Swedish Twin Registry. *Arch. Environ. Health* **21**, 508-513 (1970).
23. Martin,N.G., Carr,A.B., Oakeshott,J.G., & Clark,P. Co-twin control studies: vitamin C and the common cold. *Prog. Clin. Biol. Res.* **103 Pt A**, 365-373 (1982).
24. de Moor,M.H., Boomsma,D.I., Stubbe,J.H., Willemsen,G., & de Geus,E.J. Testing causality in the association between regular exercise and symptoms of anxiety and depression. *Arch. Gen. Psychiatry* **65**, 897-905 (2008).
25. Ligthart,L., Nyholt,D.R., Penninx,B.W., & Boomsma,D.I. The shared genetics of migraine and anxious depression. *Headache* **50**, 1549-1560 (2010).
26. Lundqvist,E. *et al.* Co-twin control and cohort analyses of body mass index and height in relation to breast, prostate, ovarian, corpus uteri, colon and rectal cancer among Swedish and Finnish twins. *Int. J. Cancer* **121**, 810-818 (2007).
27. Bell,J.T. & Spector,T.D. A twin approach to unraveling epigenetics. *Trends Genet.* **27**, 116-125 (2011).
28. Fraga,M.F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10604-10609 (2005).
29. Talens R.P. *et al.* Epigenetic variation during the adult lifespan: crosssectional and longitudinal data on monozygotic twin pairs. Aging cell . 2012.
30. Wong,C.C. *et al.* A longitudinal study of epigenetic variation in twins. *Epigenetics.* **5**, 516-526 (2010).
31. Gordon,L. *et al.* Expression discordance of monozygotic twins at birth: effect of intrauterine environment and a possible mechanism for fetal programming. *Epigenetics.* **6**, 579-592 (2011).
32. Ollikainen,M. *et al.* DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Hum. Mol. Genet.* **19**, 4176-4188 (2010).
33. Powell,J.E. *et al.* Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.* **22**, 456-466 (2012).
34. Hjelmborg,J.B. *et al.* Genetic influence on human lifespan and longevity. *Hum. Genet.* **119**, 312-321 (2006).
35. Bakaysa,S.L. *et al.* Telomere length predicts survival independent of genetic influences. *Aging cell* **6**, 769-774 (2007).
36. Zwijnenburg,P.J.G., Meijers Heijboer,H., & Boomsma,D.I. Identical but not the same: The value of discordant monozygotic twins in genetic research. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153**, 1134-1149 (2010).

37. Forsberg,L.A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J Hum. Genet.* **90**, 217-228 (2012).
38. Ehli,E.A. *et al.* De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems. *Eur. J. Hum. Genet.* doi: 10.1038/ejhg.2012.49 (2012).
39. Veenma,D. *et al.* Copy number detection in discordant monozygotic twins of Congenital Diaphragmatic Hernia (CDH) and Esophageal Atresia (EA) cohorts. *Eur. J. Hum. Genet.* **20**, 298-304 (2012).
40. Alkan,C., Coe,B.P., & Eichler,E.E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363-376 (2011).
41. Baranzini,S.E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351-1356 (2010).
42. Veltman J.A. & Brunner H.G. *De novo* mutations in human genetic disease. Nature Rev.Genet. (doi: 10.1038/nrg3241). 2012.
43. Vadlamudi,L. *et al.* Timing of De Novo Mutagenesis: A Twin Study of Sodium-Channel Mutations. *N. Engl. J. Med.* **363**, 1335-1340 (2010).
44. Oates,N.A. *et al.* Increased DNA methylation at the AXIN1 gene in a monozygotic twin from a pair discordant for a caudal duplication anomaly. *Am. J. Hum. Genet.* **79**, 155-162 (2006).
45. Mastroeni,D., McKee,A., Grover,A., Rogers,J., & Coleman,P.D. Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for Alzheimer's disease. *PLoS One* **4**, e6617 (2009).
46. Nguyen,A., Rauch,T.A., Pfeifer,G.P., & Hu,V.W. Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. *FASEB J.* **24**, 3036-3051 (2010).
47. Kuratomi,G. *et al.* Aberrant DNA methylation associated with bipolar disorder identified from discordant monozygotic twins. *Mol. Psychiatry* **13**, 429-441 (2008).
48. Rosa,A. *et al.* Differential methylation of the X-chromosome is a possible source of discordance for bipolar disorder female monozygotic twins. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, 459-462 (2008).
49. Gao,Y. *et al.* Increased Expression and Altered Methylation of HERVWE1 in the Human Placentas of Smaller Fetuses from Monozygotic, Dichorionic, Discordant Twins. *PLoS One* **7**, e33503 (2012).
50. Galetzka,D. *et al.* Monozygotic twins discordant for constitutive BRCA1 promoter methylation, childhood cancer and secondary cancer. *Epigenetics.* **7**, 47-54 (2012).
51. Gervin,K. *et al.* DNA Methylation and Gene Expression Changes in Monozygotic Twins Discordant for Psoriasis: Identification of Epigenetically Dysregulated Genes. *PLoS Genet.* **8**, e1002454 (2012).
52. Heijmans,B.T., Kremer,D., Tobi,E.W., Boomsma,D.I., & Slagboom,P.E. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Hum. Mol. Genet.* **16**, 547-554 (2007).
53. Coolen,M.W. *et al.* Impact of the Genome on the Epigenome Is Manifested in DNA Methylation Patterns of Imprinted Regions in Monozygotic and Dizygotic Twins. *PLoS One* **6**, e25590 (2011).
54. Gertz,J. *et al.* Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation. *PLoS Genet.* **7**, e1002228 (2011).

55. Kaminsky,Z.A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* **41**, 240-245 (2009).
56. Amaral,P.P., Dinger,M.E., Mercer,T.R., & Mattick,J.S. The eukaryotic genome as an RNA machine. *Science* **319**, 1787-1789 (2008).
57. Kim,V.N. MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.* **6**, 376-385 (2005).
58. Mattick,J.S. & Makunin,I.V. Non-coding RNA. *Hum. Mol. Genet.* **15 Spec No 1**, R17-R29 (2006).
59. Sarachana,T., Zhou,R., Chen,G., Manji,H.K., & Hu,V.W. Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines. *Genome Med.* **2**, 23 (2010).
60. Te,J.L. *et al.* Identification of unique microRNA signature associated with lupus nephritis. *PLoS One* **5**, e10344 (2010).
61. Tan,Q. *et al.* Genetic dissection of gene expression observed in whole blood samples of elderly Danish twins. *Hum. Genet.* **117**, 267-274 (2005).
62. Nica,A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).
63. Pietiläinen,K.H. *et al.* Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity. *PLoS Med.* **5**, e51 (2008).
64. Haas,C.S. *et al.* Identification of genes modulated in rheumatoid arthritis using complementary DNA microarray analysis of lymphoblastoid B cell lines from disease-discordant monozygotic twins. *Arthritis Rheum.* **54**, 2047-2060 (2006).
65. Matigian,N. *et al.* Expression profiling in monozygotic twins discordant for bipolar disorder reveals dysregulation of the WNT signalling pathway. *Mol. Psychiatry* **12**, 815-825 (2007).
66. Kakiuchi,C. *et al.* Upregulation of ADM and SEPX1 in the lymphoblastoid cells of patients in monozygotic twins discordant for schizophrenia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147**, 557-564 (2008).
67. Beyan,H. *et al.* Monocyte gene-expression profiles associated with childhood-onset type 1 diabetes and disease risk: a study of identical twins. *Diabetes* **59**, 1751-1755 (2010).
68. Caramori,M.L. *et al.* Gene expression differences in skin fibroblasts in identical twins discordant for type 1 diabetes. *Diabetes* **61**, 739-744 (2012).
69. Ronkainen,P.H. *et al.* Postmenopausal hormone replacement therapy modifies skeletal muscle composition and function: a study with monozygotic twin pairs. *J. Appl. Physiol.* **107**, 25-33 (2009).
70. Ellis,D.I., Dunn,W.B., Griffin,J.L., Allwood,J.W., & Goodacre,R. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics* **8**, 1243-1266 (2007).
71. Nicholson,G. *et al.* Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol. Syst. Biol.* **7**, 525 (2011).
72. Kettunen,J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269-276 (2012).
73. Nicholson,G. *et al.* A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet.* **7**, e1002270 (2011).
74. Pietiläinen,K.H. *et al.* Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects - a monozygotic twin study. *PLoS One* **2**, e218 (2007).

75. Pietiläinen,K.H. *et al.* Association of lipidome remodeling in the adipocyte membrane with acquired obesity in humans. *PLoS Biol.* **9**, e1000623 (2011).
76. Kato,B.S. *et al.* Variance decomposition of protein profiles from antibody arrays using a longitudinal twin model. *Proteome. Sci.* **9**, 73 (2011).
77. Dempster,E.L. *et al.* Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Hum. Mol. Genet.* **20**, 4786-4796 (2011).
78. van Dijk,B.A., Boomsma,D.I., & de Man,A.J. Blood group chimerism in human multiple births is not rare. *Am. J. Med. Genet.* **61**, 264-268 (1996).
79. Laborie,L.B. *et al.* DNA hypomethylation, transient neonatal diabetes, and prune belly sequence in one of two identical twins. *Eur. J. Pediatr.* **169**, 207-213 (2010).
80. Erlich,Y. Blood Ties: Chimerism Can Mask Twin Discordance in High-Throughput Sequencing. *Twin Res. Hum. Genet.* **14**, 137-143 (2011).
81. Scheet,P. *et al.* Twins, tissue, and time: an assessment of SNPs and CNVs. *Twin. Res. Hum. Genet.* **15**, 737-745 (2012).
82. Spor,A., Koren,O., & Ley,R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* **9**, 279-290 (2011).
83. Stewart,J.A., Chadwick,V.S., & Murray,A. Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J. Med. Microbiol.* **54**, 1239-1242 (2005).
84. Zoetendal,E.G., Akkermans,A.D.L., Akkermans-van Vliet,W.M., de Visser,J.A.G.M., & de Vos,W.M. The host genotype affects the bacterial community in the human gastronintestinal tract. *Microb. Ecol. Health Dis.* **13**, 129-134 (2001).
85. Turnbaugh,P.J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2008).
86. Willing,B.P. *et al.* A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* **139**, 1844-1854 (2010).
87. Lepage,P. *et al.* Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* **141**, 227-236 (2011).
88. Kendler,K.S. & Eaves,L.J. Models for the joint effect of genotype and environment on liability to psychiatric illness. *Am. J. Psychiatry* **143**, 279-289 (1986).
89. Stubbe,J.H. *et al.* Genetic influences on exercise participation in 37.051 twin pairs from seven countries. *PLoS One* **1**, e22 (2006).
90. Teucher,B. *et al.* Dietary patterns and heritability of food choice in a UK female twin cohort. *Twin Res. Hum. Genet.* **10**, 734-748 (2007).
91. Middeldorp,C.M., Cath,D.C., Vink,J.M., & Boomsma,D.I. Twin and genetic effects on life events. *Twin Res. Hum. Genet.* **8**, 224-231 (2005).
92. Kendler,K.S. & Baker,J.H. Genetic influences on measures of the environment: a systematic review. *Psychol. Med.* **37**, 615-626 (2007).
93. Vinkhuyzen,A.A.E., Van Der Sluis,S., De Geus,E.J.C., Boomsma,D.I., & Posthuma,D. Genetic influences on 'environmental' factors. *Genes Brain Behav.* **9**, 276-287 (2010).
94. Caspi,A. & Moffitt,T.E. Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nat. Rev. Neurosci.* **7**, 583-590 (2006).
95. Caspi,A. *et al.* Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18860-18865 (2007).

96. Berg,K. Variability gene effect on cholesterol at the Kidd blood group locus. *Clin. Genet.* **33**, 102-107 (1988).
97. Wray,N.R. *et al.* Use of monozygotic twins to investigate the relationship between 5HTTLPR genotype, depression and stressful life events: an application of Item Response Theory. *Novartis Found. Symp.* **293**, 48-59 (2008).
98. Visscher,P.M., Hill,W.G., & Wray,N.R. Heritability in the genomics era - concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255-266 (2008).
99. Visscher,P.M. *et al.* Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am. J. Med. Genet.* **81**, 1104-1110 (2007).
100. Jörgensen,G. Erbfaktoren bei häufigen Krankheiten. in *Erbgefüge* (ed. Vogel F) 581-665 (Springer, Berlin Heidelberg New York, 1974).
101. Smith,C. Heritability of liability and concordance in monozygous twins. *Ann. Hum. Genet.* **34**, 85-91 (1970).
102. Tazi,L. *et al.* HIV-1 infected monozygotic twins: a tale of two outcomes. *BMC Evol. Biol.* **11**, 62 (2011).
103. Ott,J., Kamatani,Y., & Lathrop,M. Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* **12**, 465-474 (2011).
104. Galton,F. The history of twins, as a criterion of the relative powers of nature and nurture. *The Journal of the anthropological Institute of Great Britain and Ireland* **5**, 391-406 (1876).
105. Mayo,O. Early research on human genetics using the twin method: who really invented the method? *Twin Res. Hum. Genet.* **12**, 237-245 (2009).
106. Siemens,H.W. Die zwillingspathologie. *Mol. Gen. Genet.* **35**, 311-312 (1924).
107. Zhu,G. *et al.* A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions. *Eur. J Hum. Genet.* **15**, 94-102 (2007).
108. Jinks,J.L. & Fulker,D.W. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of the human behavior. *Psychol. Bull.* **73**, 311-349 (1970).
109. Martin,N.G., Eaves,L.J., Kearsey,M.J., & Davies,P. The power of the classical twin study. *Heredity (Edinb)* **40**, 97-116 (1978).
110. Boomsma,D.I. Twin registers in Europe: An overview. *Twin Res.* **1**, 34-51 (1998).
111. Busjahn,A. & Hur,Y.M. Twin registries: an ongoing success story. *Twin Res. Hum. Genet.* **9**, 705 (2006).
112. Peltonen,L. GenomEUtwin: A strategy to identify genetic influences on health and disease. *Twin Res.* **6**, 354-360 (2003).
113. Llewellyn,C.H., van Jaarsveld,C.H., Johnson,L., Carnell,S., & Wardle,J. Nature and nurture in infant appetite: analysis of the Gemini twin birth cohort. *Am. J. Clin. Nutr.* **91**, 1172-1179 (2010).
114. Lichtenstein,P. *et al.* Environmental and heritable factors in the causation of cancer: analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78-85 (2000).
115. Vink,J.M. *et al.* Cervix smear abnormalities: linking pathology data in female twins, their mothers and sisters. *Eur. J Hum. Genet.* **19**, 108-111 (2011).
116. De Geus,E.J.C. Introducing genetic psychophysiology. *Biol. Psychol.* **61**, 1-10 (2002).
117. Mattay,V.S., Goldberg,T.E., Sambataro,F., & Weinberger,D.R. Neurobiology of cognitive aging: insights from imaging genetics. *Biol. Psychol.* **79**, 9-22 (2008).

118. Peper,J.S., Brouwer,R.M., Boomsma,D.I., Kahn,R.S., & Hulshoff Pol,H.E. Genetic influences on human brain structure: a review of brain imaging studies in twins. *Hum. Brain Mapp.* **28**, 464-473 (2007).
119. Van Beijsterveldt,C.E.M. & Van Baal,G.C.M. Twin and family studies of the human electroencephalogram: a review and a meta-analysis. *Biol. Psychol.* **61**, 111-138 (2002).
120. Koten,J.W. *et al.* Genetic contribution to variation in cognitive function: an FMRI study in twins. *Science* **323**, 1737-1740 (2009).
121. van der Schot,A.C. *et al.* Influence of genes and environment on brain volumes in twin pairs concordant and discordant for bipolar disorder. *Arch. Gen. Psychiatry* **66**, 142-151 (2009).
122. De Geus,E.J.C. *et al.* Intrapair differences in hippocampal volume in monozygotic twins discordant for the risk for anxiety and depression. *Biol. Psychiatry* **61**, 1062-1071 (2007).
123. Falconer,D.S. *Introduction to quantitative genetics*(Ronald Press Co., New York, 1960).
124. Tsang,T.M., Huang,J.T., Holmes,E., & Bahn,S. Metabolic profiling of plasma from discordant schizophrenia twins: correlation between lipid signals and global functioning in female schizophrenia patients. *J. Proteome Res.* **5**, 756-760 (2006).
125. Harder,A. *et al.* Monozygotic Twins With Neurofibromatosis Type 1 (NF1) Display Differences in Methylation of NF1 Gene Promoter Elements, 5'Untranslated region, Exon and Intron 1. *Twin Res. Hum. Genet.* **13**, 582-594 (2010).
126. Silventoinen,K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res.* **6**, 399-408 (2003).
127. Schousboe,K. *et al.* Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. *Twin Res.* **6**, 409-421 (2003).
128. Clausson,B., Lichtenstein,P., & Cnattingius,S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG* **107**, 375-381 (2000).
129. Hyttinen,V., Kaprio,J., Kinnunen,L., Koskenvuo,M., & Tuomilehto,J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs. *Diabetes* **52**, 1052-1055 (2003).
130. Kaprio,J. *et al.* Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* **35**, 1060-1067 (1992).
131. Zdravkovic,S. *et al.* Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J. Intern. Med.* **252**, 247-254 (2002).
132. Zhang,S. *et al.* Genetic and environmental contributions to phenotypic components of metabolic syndrome: a population-based twin study. *Obesity (Silver Spring)* **17**, 1581-1587 (2009).
133. Rahman,I. *et al.* Genetic dominance influences blood biomarker levels in a sample of 12,000 Swedish elderly twins. *Twin Res. Hum. Genet.* **12**, 286-294 (2009).
134. Pedersen,N.L., Gatz,M., Berg,S., & Johansson,B. How heritable is Alzheimer's disease late in life? Findings from Swedish twins. *Ann. Neurol.* **55**, 180-185 (2004).
135. Wirdefeldt,K., Gatz,M., Reynolds,C.A., Prescott,C.A., & Pedersen,N.L. Heritability of Parkinson disease in Swedish twins: a longitudinal study. *Neurobiol. Aging* **32**, 1923-1928 (2011).

136. Mulder,E.J. *et al.* Genetic and environmental influences on migraine: a twin study across six countries. *Twin Res.* **6**, 422-431 (2003).
137. Hawkes,C.H. & MacGregor,A.J. Twin studies and the heritability of MS: a conclusion. *Mult. Scler.* **15**, 661-667 (2009).
138. Faraone,S.V. *et al.* Molecular genetics of attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **57**, 1313-1323 (2005).
139. Lundstrom,S. *et al.* Autism spectrum disorders and autistic like traits: similar etiology in the extreme end and the normal variation. *Arch. Gen. Psychiatry* **69**, 46-52 (2012).
140. Sullivan,P.F., Kendler,K.S., & Neale,M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187-1192 (2003).
141. Sullivan,P.F., Neale,M.C., & Kendler,K.S. Genetic epidemiology of major depression: review and meta-analysis. *Am J. Psychiatry* **157**, 1552-1562 (2000).
142. Peacock,M., Turner,C.H., Econs,M.J., & Foroud,T. Genetics of osteoporosis. *Endocr. Rev.* **23**, 303-326 (2002).
143. Spector,T.D. & MacGregor,A.J. Risk factors for osteoarthritis: genetics. *Osteoarthr. Cartil.* **12, Supplement**, 39-44 (2004).
144. MacGregor,A.J. *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* **43**, 30-37 (2000).
145. Thomsen,S.F., Van Der Sluis,S., Kyvik,K.O., Skytthe,A., & Backer,V. Estimates of asthma heritability in a large twin sample. *Clin. Exp. Allergy* **40**, 1054-1061 (2010).
146. Ingebrigtsen,T.S. *et al.* Genetic influences on pulmonary function: a large sample twin study. *Lung* **189**, 323-330 (2011).
147. Li,M.D., Cheng,R., Ma,J.Z., & Swan,G.E. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction* **98**, 23-31 (2003).
148. Agrawal,A. & Lynskey,M.T. Are there genetic influences on addiction: evidence from family, adoption and twin studies. *Addiction* **103**, 1069-1081 (2008).
149. Willer,C.J., Dyment,D.A., Risch,N.J., Sadovnick,A.D., & Ebers,G.C. Twin concordance and sibling recurrence rates in multiple sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12877-12882 (2003).
150. Halfvarson,J. Genetics in twins with Crohn's disease: less pronounced than previously believed? *Inflamm. Bowel Dis.* **17**, 6-12 (2011).
151. Tanner,C.M. *et al.* Parkinson disease in twins. *JAMA* **281**, 341-346 (1999).
152. Cardno,A.G. *et al.* Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch. Gen. Psychiatry* **56**, 162-168 (1999).
153. Kendler,K.S. & Prescott,C.A. A population-based twin study of lifetime major depression in men and women. *Arch. Gen. Psychiatry* **56**, 39-44 (1999).
154. Levy,F., Hay,D.A., McStephen,M., Wood,C., & Waldman,I. Attention-deficit hyperactivity disorder: a category or a continuum? Genetic analysis of a large-scale twin study. *J. Am. Acad. Child Adolesc. Psychiatry* **36**, 737-744 (1997).
155. Rosenberg,R.E. *et al.* Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Arch. Pediatr. Adolesc. Med.* **163**, 907-914 (2009).

Summary and discussion

The research described in this thesis was characterized by two major themes: the first was related to the question to which extent variation between individuals in inflammation biomarkers and metabolic syndrome traits are caused by the impact of genetic and environmental differences between people. The second important focus of this thesis went beyond the influence of the DNA sequence and examined epigenetic variation. I addressed the question how important genetic and non-genetic sources of variation are for individual differences in DNA methylation. DNA methylation is an epigenetic mechanism that receives increasing attention as it may provide novel insights into human disease and represents an extra layer causing differences between people. In this chapter, I summarize the most important results and discuss these findings in the broader context of the current state and future directions of research on complex trait genetics.

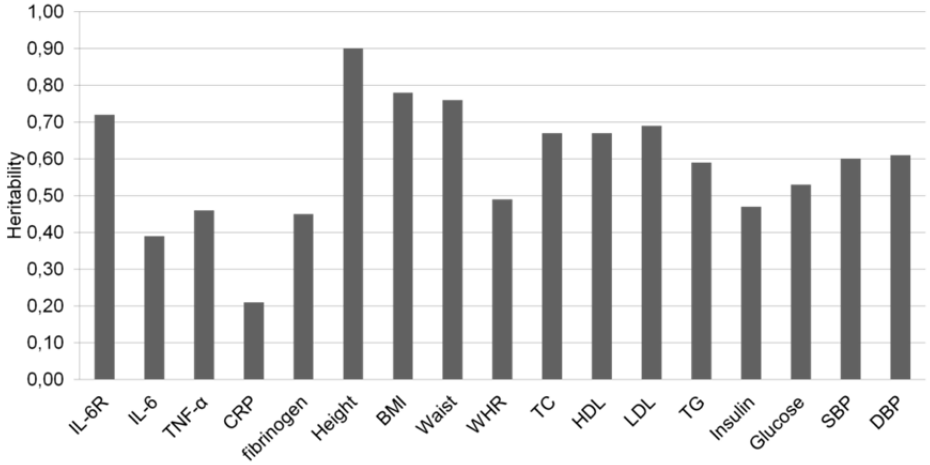
Part 1: Characterizing the genetic architecture of inflammation biomarkers

In the first part of this thesis (chapters 2 and 3), I examined the importance of genetic and environmental influences for individual differences in inflammation biomarkers, including pro-inflammatory cytokines (IL-6 and TNF-α), a soluble cytokine receptor (sIL-6R) and acute-phase proteins fibrinogen and CRP. Extended twin-family models showed that variation in all of these biomarkers of inflammation is to an important extent explained by genetic variation (figure 1). Moderate heritabilities were found for the concentrations of IL-6 (H^2=21%), TNF- α (H^2=39%), CRP (H^2=45%), and fibrinogen (H^2=46%). The levels of soluble IL-6 receptor levels in blood were highly heritable (H^2=72%). The heritability reflects the overall proportion of variation of a trait in the population that can be attributed to genetic variation, but the value of this statistic does not give insight in the number of genes involved, or in the molecular pathways that give rise to the inheritance of traits.

The importance of currently identified genetic variants

The contribution of particular genetic variants to the heritability of a trait can be examined if classical estimation of heritability is combined with the analysis of measured DNA-sequence variants. In chapter 3, I applied this combination of methods to the concentration of sIL-6R in blood. The variance of sIL-6R was largely explained by a single SNP in the IL6R gene that influences the production of sIL-6R (rs2228145, total variance explained=51%; 71% of the total heritability). Through linkage analysis, we found evidence that the remaining heritability is mostly attributable to other genetic variants within the

Figure 1: Heritability of inflammation biomarkers and metabolic syndrome traits.



Heritability=Broad-sense heritability, WHR=Waist-to-hip-ratio, TC=Total cholesterol, TG=Triglycerides, SBP=Systolic blood pressure, DBP=diastolic blood pressure.

IL6R gene region on chromosome 1 and detected novel SNPs at the 3'end that were associated with IL6R gene expression. Of the inflammation biomarkers examined in chapter 2, GWA studies have thus far been published for CRP, fibrinogen and IL-6. A genetic risk score based on 18 genome-wide significant SNPs for CRP explained 5% of the variance in CRP levels[1]. For fibrinogen, four genome-wide significant loci have been reported to date that together explain < 2% of the variance in fibrinogen levels [2]. For IL-6, two loci (*ABO* and *IL6R*) have thus far been reported. The top SNPs in these loci together explain 2.2% of the variance of IL-6 levels [3]. The findings illustrate the difference in the genetic architecture of sIL-6R (for which a single common SNP explains > 50% of the total variance) versus the other inflammation biomarkers. In fact, the variance of sIL-6R levels that can be explained by rs2228145 is very large compared to single SNP effect sizes observed for quantitative traits in general.

The difference between the population explained variance by common SNPs for sIL-6R and IL-6 could be related to the biological consequences of a certain degree of change in concentration of these molecules. Importantly, it is thought that the level of IL-6, and not sIL-6R determines the degree of IL-6 signalling under 'normal conditions'. There is typically a buffer of sIL-6R and soluble gp130 in blood that block the activity of IL-6: the concentrations of sIL-6R and sgp130 (which acts as an antagonist of the active membrane-bound transducer protein gp130) are generally 1000 times higher than the concentration of IL-6 [4]. Importantly, the dynamic range of sIL-6R concentration is more restricted compared to IL-6: during inflammation, sIL-6R levels may rise in the range of 2-3 fold[5]. By contrast, the levels of IL-6 can rise as much as 1 million-fold (under severe conditions)[6]. It may be hypothesized that a genetic

variant with an effect so large that it would explain 50% of the variance of IL-6 levels would have much more serious health consequences compared to the effect of SNP rs2228145, because it would cause serious chronic inflammation in part of the population. Of note, the top SNP for IL-6 may be just as important for disease outcomes as the top SNP of sIL-6R. These inflammation biomarkers exemplify the differences in genetic architecture that may exist between different complex traits (or between different biological pathways that together lead to a disease) and how these differences affect the effort to identify the underlying genes. Even though SNPs identified through GWAS of complex diseases typically explain only a small proportion of the variance, they may point at biological pathways that provide novel insight into disease mechanisms. For example, based on the finding that the *IL6R* SNP is associated with coronary heart disease, it was suggested that tocilizumab, a monoclonal antibody against IL-6R that was already used for treatment of rheumatoid arthritis, should be tested in randomised trials for prevention of coronary heart disease [7]. SNPs in *IL6R* are also associated with asthma risk, and trials investigating tocilizumab for treatment of asthma are currently underway in Australia[8, 9]. It is expected that GWAS findings will lead to novel drug targets for other complex diseases as well.

Part 2 Genetic and environmental influences on BMI and other metabolic syndrome traits

Heritability

In the second part of this thesis, the extended twin-family design was applied to examine the heritability of height and individual metabolic syndrome traits including BMI, waist circumference, waist-to-hip-ratio, metabolic biomarkers and blood pressure (chapter 4, figure 1). This study showed that height, BMI and waist circumference were the most heritable traits (H^2=0.90, H^2=0.78, and H^2=0.76, respectively). Lipids and blood pressure were also highly heritable (H^2=0.59- H^2=0.69), and WHR, insulin and glucose levels were moderately heritable (H^2=0.49, H^2= 0.47, H^2= 0.53, respectively).

An important finding of chapter 4 was that all metabolic syndrome traits were influenced by non-additive genetic effects. For example, the total (broad) heritability of BMI was 78%, with 41% of the variance explained by additive genetic effects and 37% of the variance explained by non-additive genetic effects. Possible explanations for the observed non-additive genetic effects may include the effects of dominant alleles, epistasis, or any other type of genetic effect not acting in an additive manner, causing lower phenotypic similarity between parents and offspring compared to the similarity of DZ twins and siblings. We examined qualitative age differences as a possible alternative explanation for the observed non-additive genetic variance, however, we found no significant qualitative genetic differences between age groups (i.e. different genes contributing to a trait at different ages) for any trait. We did find that the total heritability was slightly lower in older subjects for several traits and this

was mainly related to an increase in unique environmental variance at higher ages. The non-additive genetic effects in our study could also potentially involve genetic effects that depend on environmental influences that are shared to a greater extent between siblings and twins than between children and their parents (gene X (sibling-shared) environment interaction). For example, lifestyle conditions have changed between the time when the parents in our study grew up and the time when their offspring grew up[10]. If the impact of genetic influences on metabolic syndrome traits depends on environmental exposures (e.g. caloric intake) that have changed between generations, such effects may be to a greater extent shared between twins and sibs than between parents and offspring. A final possible explanation is that parents and offspring have been equally exposed to relevant environmental influences during their lives, but that the impact of these exposures depends on genotype plus developmental stage, age, or length of the exposure (the latter three are similar among siblings but differ between parents and children).

In chapter 5, the influence of genes and environment on BMI was studied from a different angle by examining longitudinal BMI discordance in MZ pairs who participated in NTR studies between 1991 and 2011. The most important finding was that large BMI discordance (within-pair BMI difference > 3 kg/m^2), especially long-term discordance, in MZ twin pairs is rare, although discordance became more frequent at later NTR surveys - as the mean BMI and age of twin pairs increased (mean age at survey 1=17.2, SD=2.4, mean age at survey 8=34.6, SD=15.0). We observed that of the MZ twins who became discordant at some point in their lives, most converged to the same weight quickly. These findings illustrated that the heritability of BMI, at least in part, reflects that people tend to have a certain set-point of BMI that is strongly genetically determined, and that this set-point may diffuse after substantial weight gain. Another important finding of this study was that BMI discordant twins show clinically relevant differences in metabolic and inflammation biomarkers. These differences were not present in MZ twins who developed BMI discordance later in their life, illustrating that the level of these biomarkers cannot predict whether a person will develop a high BMI but rather the levels of these biomarkers are changed in response to changes in BMI. The discordant MZ twin design will continue to provide insight into the causes of obesity-associated disease in the future. The EUroDiscoTwin consortium plans to characterize obesity- concordant twins with differences in metabolic biomarkers, including Dutch twins from the NTR, to gain insight into the causes of 'metabolically healthy' versus 'unhealthy obesity'.

Current state and future directions of genetic research on complex (metabolic) traits

Large genome-wide association meta-analyses for the traits studied in chapters 6 and 7 have identified a large number of loci to date [11-14]. For example, 32 loci have been published for BMI to date, which explain 1.45% of the variance of

BMI [13], and 164 loci have been identified in the largest to date meta-analysis of BMI based on 320,485 individuals (manuscript submitted[15]). Importantly, as I explained earlier in this chapter, a small proportion of variance explained by a variant in the population certainly does not mean that the pathway affected by this variant is not that important for the phenotype or disease, although extremely large samples may be required to identify all common genetic variation through GWAS [13]. Of note, for adult height, as much as 697 common variants in 423 loci have been identified with genome-wide significance to date[16], explaining one-fifth of the heritability of height (16% of the total variance). The height paper illustrates that larger samples continue to provide novel biological insights for the trait by pointing at novel relevant genes, and that phenotypic variation is related to multiple variants in the same gene or pathway, each with small individual effects[16].

While genome-wide significant SNPs from GWAS point at locations in the genome where something is happening, the molecular action of these SNPs is often not understood. The majority of currently identified SNPS associated with complex diseases are located outside gene-coding regions[17], and the pathways through which the risk allele influences disease risk is often unknown. This issue is exemplified by the *FTO* locus, which harbours the strongest genetic associations with BMI and obesity [13, 18]. The SNPs are located in intron 1 and 2 of the *FTO* gene. The biological relationship between this SNP and BMI was thought for years to be related to the expression of *FTO*, but a recent study demonstrated that the obesity-associated SNPs are associated with the expression of the homeobox gene *IRX3*, not *FTO* in human brain tissue[19]. Thus, the obesity-associated region within FTO contains enhancers that form long-range interactions with the promoter of *IRX3*. It has become clear that studies that provide insight into how DNA sequence and gene regulation are connected may be crucial to our understanding of complex diseases. The ENCODE project has greatly enhanced our knowledge of the regulatory regions encoded in the human genome by large-scale mapping of functional elements such as promoters, enhancers and open chromatin regions[20]. Further insight may come from studies that map genetic variants that impact on epigenetic regulation. For example, methylation QTL analysis aims to identify DNA sequence variants associated with variation in DNA methylation, and is expected to provide insight into the mechanisms affected by complex disease-associated SNPS that are identified through GWAS [21, 22]. The results of our study from chapter 6 in part 3 of my thesis give insight in the extent to which DNA methylation targeted by the now widely used Illumina 450k array shows variation due to common genotyped SNPs.

Part 3 Beyond DNA sequence: Epigenetic variation

A growing body of evidence highlights the importance of epigenetic regulation for the phenotypes examined in part 1 and 2 of my thesis (see for example [23-29]). DNA methylation is one of such epigenetic mechanisms, which may

mediate DNA sequence effects on complex traits as well as the effects of environmental exposures. In part 3 of this thesis, I examined the causes of variation between people in DNA methylation across the genome. In chapter 6, I describe that methylation at many CpG sites in blood is heritable and sites where a large proportion of variance is due to non-genetic influences (including environment, stochastic variation and measurement error) are also abundant. Of note, the average genome-wide variance of DNA methylation across individuals is low, because at a large number of CpGs, most individuals have very similar methylation levels. I also extended current knowledge on the heritability of DNA methylation with estimates of the variance explained by common genome-wide SNPs. These estimates of 'SNP heritability' suggested that DNA methylation level at a large number of CpGs measured by the 450k array displays genetic variation that is tagged by common genotyped SNPs, which provides guidance for mQTL studies and suggests that the 450k array will provide insight into disease-associated SNPs that act through methylation. In addition, I identified a large number of CpGs where the heritability of DNA methylation decreased with age, a small subset where the heritability increased with age, and a number of CpGs where the heritability of DNA methylation differed between males and females. The overall (twin-based) heritability and 'SNP' heritability of DNA methylation were on average 22 % and 7%, respectively, across genome-wide CpGs.

In chapter 7, I examined MZ twin correlations for DNA methylation measured using the 450k array in buccal samples. In many studies, for example those that focus on development and childhood traits and disorders, buccal is an attractive biological sample to study. I found that the MZ twin correlation is high at a number of CpGs sites in the genome (average correlation=0.31 across all CpGs, and average correlation=0.54 at sites where methylation showed the largest variance across subjects). In comparison, the average MZ twin correlation for DNA methylation level in blood was 0.20 across all genome-wide sites, and 0.57 for sites with the largest between-individual variance in blood. These findings illustrate that DNA methylation variation in buccal cells, similar to our finding in blood, is likely to be influenced by genetic variation. Of note, in chapter 7 I also found that the MZ twin correlation for DNA methylation level shows variation between genomic regions: Regions with low CpG density showed lower MZ twin correlations, suggesting that the variation in DNA methylation in these regions to a larger extent reflects non-genetic sources including environmental and stochastic influences. Although most studies of DNA methylation are currently performed using blood samples, our results from chapter 7 suggest that buccal samples are also suitable for obtaining genome-wide DNA methylation data.

Epigenetic studies and tissue

Large-scale epigenetic studies in living humans are only possible if DNA is obtained from easily accessible tissues. In chapter 6, I examined DNA

methylation in blood samples and in chapter 7 I studied DNA methylation levels based on DNA extracted from buccal swabs. Buccal samples are easier to collect than blood samples, particularly in (young) children. It has also been suggested that buccal-derived DNA may be better suited compared to blood-derived DNA for studying epigenetics in connection to behavioural/psychiatric traits [30]. A recent study that compared methylation measured with the 450k array based on DNA from human post-mortem brain tissues from four different regions, blood and saliva samples found that saliva samples, in particular those containing the largest proportion of buccal epithelial cells showed an overall genome-wide methylation pattern that was closest to the brain tissues [31]. This greater similarity may relate to the fact that buccal cells and brain cells are both derived from the ectodermal layer, whereas blood cells are derived from the mesodermal layer. Nevertheless, there are also indications that DNA methylation levels at relevant candidate genes in blood may respond in a similar way to relevant exposures compared to methylation in the relevant disease tissue: Allele-specific methylation of the *FKBP5* gene (which has been implicated as a mediator of the effects of childhood traumatic life events on stress-related psychiatric disorders in adulthood) responds similarly in blood and brain cells to stimulation by glucocorticoids [32]. Although well-suited for DNA methylation studies, a potential drawback of buccal samples is that, because they contain a relatively small number of cells and thus a small amount of proteins, they are currently not suited for studies of other epigenetic marks such as histone modifications. In conclusion, it is expected that both blood and buccal will contribute to insights in the role of DNA methylation in human complex traits. In the nearby future, DNA methylation will be measured in buccal samples from twins who participate in the NTR as part of an FP7 EU-funded ACTION project on aggression in children.

Part 4: Twin studies and complex traits: Future and further considerations

While GWASs for complex diseases and traits such as type II diabetes, insulin response and triglyceride levels were an instant success (see for example [33-35]), early GWASs for common mental disorders were initially less successful, however, these studies were conducted in much smaller cohorts. For example, early GWASs of schizophrenia did not find genome-wide significant SNPs [36-38]. The finding that SNPs identified through GWAS explained only a small proportion of the total variance fuelled the discussion about the “missing heritability”[39], which was at the time an important inspiration for chapter 8, in which I reviewed evolutionary perspectives on schizophrenia. It had been postulated that the difficulty to find genetic variants that explain a substantial part of the risk for psychiatric disorders may be related to the evolutionary history of these disorders [40]. In chapter 8, I reviewed evolutionary perspectives that have been proposed to explain why schizophrenia, one of the most detrimental common mental disorders, persists in the population despite

conferring a significant reproductive disadvantage to patients. Proposed evolutionary explanations include balancing selection, fitness trade-offs, fluctuating environments, sexual selection, mutation-selection balance and genomic conflicts. The evolutionary perspectives of schizophrenia may to some extent also apply to other common (psychiatric) disorders. In July 2014, a paper came out that reported 108 loci for schizophrenia, which were identified based on 36,989 cases and 113,075 controls and more loci will undoubtedly be found in even larger samples [41]. Yet, the principles discussed in chapter 8 are still relevant as the 108 identified loci explain just 3.4% of the variation of disease liability, and all common SNPs together have been estimated to account for 23% of the variation in liability to schizophrenia, while the total heritability of this disease is ~70%.

For decades, the classical twin study has provided insight into the importance of genes and environment to individual differences in human complex traits. In chapter 9, I described the types of questions that can be examined with twin studies and review studies that applied either the classical twin study or the discordant MZ twin design to complex molecular traits such as metabolomics data, gene expression, the microbiome and epigenetics. The chapter illustrates that twin studies will continue to be highly valuable in the future. In chapters 6 and 7, I studied the importance of genes and environment to the variance of methylation at individual CpG sites in the genome. In chapter 9, I mentioned other interesting questions that may be examined in future studies with multivariate twin designs, for example: to what extent is epigenetic regulation and gene expression across genomic regions influenced by shared genetic or environmental factors? And to what degree do common genetic and environmental mechanisms underlie biological variation across different cells and tissues? MZ twins, who are often discordant for complex diseases, are expected to provide novel insights into epigenetic mechanisms involved in complex disease in the future.

On the interpretation of the heritability of DNA methylation

In chapter 9, I highlighted several assumptions of the twin design that are important to (re) evaluate if this design is applied to molecular data, including the assumption that MZ twins share 100% of their genetic material and the assumption that MZ twins share environmental influences to the same extent as DZ twins. Two other important issues for the interpretation of heritability in the light of epigenetic studies are parental effects and the potential existence of trans-generational epigenetic inheritance. The question whether trans-generational epigenetic inheritance exists in humans is perhaps one of the hottest topics these days in human epigenetic research. Although the answer to this question is currently unknown, a non-trivial question is how parental and trans-generational effects would influence the interpretation of the research described in this thesis, including the heritability of DNA methylation (part 3) and the heritability of phenotypes such as BMI (part 1 and 2).

Parental effects on epigenetic regulation imply that an exposure of the parent causes an epigenetic change in the germ line that is transmitted to the offspring or that maternal exposure during pregnancy causes an epigenetic change in the embryo. These epigenetic changes in turn may cause phenotypic changes in the offspring. There is increasing evidence for the existence of such effects in mammals. For example, several studies have demonstrated that in-utero exposure of rodents to either maternal low protein diet or maternal high-fat diet causes an increased risk of obesity and metabolic syndrome in the offspring through changes in epigenetic regulation [42-46], and effects of maternal under-nutrition on epigenetic regulation in offspring have also been demonstrated in humans [47]. In rodents, similar effects have been demonstrated for paternal exposures [48, 49]. Although there is no doubt that these effects exist, an important question that remains to be unravelled is how much they contribute to variation in human phenotypes at the population level. Trans-generational epigenetic inheritance implies that epigenetic marks are transmitted through the germ line from parents to children, from these children to the grand children etc. (also in the absence of in-utero or germ line exposures in each generation, and independently of the DNA sequence). Rodent studies have shown that the phenotypic effects of in-utero exposure to maternal diet may be transmitted at least up to the third generation, depending on the sex of the offspring and the parent-of-origin [50]. Although the mechanisms behind the transmission of these effects are not well understood, a recent study of paternal prediabetes in rats that was transmitted to offspring revealed alterations in the fathers sperm methylome [51] and transmission of traits to offspring after paternal exposure have also been linked to microRNAs in sperm [52]. Parental and grandparental effects have also been observed in human epidemiological data. For example, effects of exposure to famine of paternal grandfathers is still evident in the grand-children, who have an increased risk of diabetes, cardiovascular disease and mortality [53, 54]. Yet, whether trans-generational inheritance of acquired epigenetic marks occurs in humans remains currently unknown.

If the classical twin model is applied to estimate the sources of variation underlying DNA methylation level, the effects of parental exposures on DNA methylation level that act independently of genotype are expected to end up in the component referred to as “the common environment”, because such effects are equally shared between members of both types of twins. However, if the effect of parental exposures (or other shared environmental influences) on DNA methylation depends on genotype of the offspring, then these effects will manifest as part of the genetic variance. Also, if the if the parental exposure affects DNA methylation in an allele-specific manner and this allele-specific methylation pattern is maintained in the offspring, methylation variation will be statistically correlated with genetic variation (even though genetic variation in this case is not the primary cause of the variation in methylation); such effects will be included in the heritability of DNA methylation. Examples of allele-

specific exposure effects, where a certain exposure alters only the methylation level of a particular genetic allele, have been published (see for example [32]), but it is unknown how often these effects occur. Allele-specific methylation implies that individuals have intermediate methylation levels and the fact that not that many CpGs show intermediate methylation levels suggests that allele-specific exposure effects do not have a great impact on DNA methylation at the genome-wide level, although it may affect a proportion of CpGs in the genome. A special but abundant type of allele-specific methylation occurs when the CpG site itself is a genetic variant (e.g. the C or G may 'contain' a SNP; only the "CG allele" has the potential to become methylated) [55]. These CpG-SNPs also have implications for GWA studies of complex traits: if disease risk depends on the methylation status of the "CG" allele, associations of such SNPs with complex phenotypes are statistically weak if methylation status is not taken into account. Such mechanisms may be identified by integrated analysis of DNA sequence variants and epigenetic variation [56, 57].

If DNA methylation is indeed influenced by trans-generational inheritance of epigenetic mechanisms, variation in DNA methylation attributable to this mechanism of inheritance is only included in the heritability of DNA methylation if transmitted in an allele-specific fashion. The existence of allele-specific trans-generational inheritance of DNA methylation is well-established in plants. For example, fruit-ripening in tomato plants is affected by an "epi-allele" involving methylation of a locus that shows a Mendelian inheritance pattern despite not being related to a nucleotide sequence difference [58]. Similar to human MZ twins, isogenic *Arabidopsis* lines can show substantial variation in DNA methylation [59]. Some differentially methylated regions that reside in isogenic lines (called epigenetic quantitative trait loci) account for substantial parts (60% to 90%) of the heritability of the complex traits flowering time and primary root length [59]. Although these findings by no means imply that similar mechanisms occur in humans (because in humans, DNA methylation is largely erased in the germ line and in the early embryo, whereas DNA methylation is not erased between generations in plants), the findings from plants illustrate the implications for the heritability of complex traits if trans-generational epigenetic inheritance (independent of the DNA sequence) would exist in humans.

Partitioning complex trait variation into genetic and epigenetic effects

It is clear that individual differences in complex traits are related to genetic variation and epigenetic variation, and that these components are connected to each other in complex ways. While it was already possible to estimate the proportion of variation in complex traits due to measured genetic variation (as illustrated in chapter 3), it is currently unknown how much of the variation in complex traits is related to epigenetic variation. To explore this question, my colleague Michel Nivard and I applied a novel method based on genome-wide SNP data and genome-wide blood methylation data, to simultaneously estimate the proportion of variation in BMI explained by SNPs and the proportion of

variation explained by DNA methylation in blood, while accounting for the correlation between the two. Preliminary results of this analysis indicated that 42% of the variation in BMI is explained by common genotyped SNPs, 39% is explained by DNA methylation variation in blood and the correlation between the effect of genome-wide SNPs on BMI and methylation variation associated with BMI was 0.36. Importantly, while the association between SNPs and the trait is inherently causal (the level of a trait cannot affect the genotype), the association between DNA methylation and the trait is more complex: this variance component may include effects of DNA methylation on the trait, effects of the trait on DNA methylation, and effects of a common underlying mechanisms on the trait and DNA methylation (with the exception of shared effects of common genotyped SNPs as these are included in the correlation of 0.36). Longitudinal phenotypic and epigenetic data will be highly valuable to distinguish between epigenetic variation that is induced by the phenotype/disease state and epigenetic variation that is causally related to the trait. In the future, the novel method may be applied to other complex traits to gain insight into the degree to which variation is related to DNA methylation variation in a target tissue, and into the extent to which genetic and epigenetic effects related to the trait of interest are correlated.

Conclusions

In conclusion, I found that inflammation biomarkers and metabolic syndrome traits are characterized by significant heritability. These findings illustrate that individual differences in the vulnerability to develop metabolic disease are to a large extent explained by differences in genetic susceptibility. Many loci have been identified where genetic or epigenetic variation is connected to these traits. I found that DNA methylation is itself to an important extent heritable, but environmental influences also account for a large part of the variation. Interestingly, I found that at a substantial number of CpGs in the genome the variation in DNA methylation due to environmental or stochastic influences increases with age. I also found that a substantial part of the heritability of DNA methylation can be explained by common SNPs, and that that blood and buccal samples are both informative to inter-individual differences in methylation due to genetic and environmental influences, although they may not hold the same information with respect to epigenetic variation involved in complex traits. The extent to which the phenotypes examined in this thesis (i.e. inflammation biomarkers and metabolic syndrome traits) are influenced by epigenetic variation, possibly in interaction with the DNA sequence and environment, awaits further examination. Future studies including Epigenome-Wide Association Studies (EWAS), and studies that integrate epigenetic and genetic information will provide further insight into the epigenetic mechanisms involved in complex traits and into the interplay of genetic and epigenetic variation within and across generations.

Reference List

1. Dehghan,A. *et al.* Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* **123**, 731-738 (2011).
2. Dehghan,A. *et al.* Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. *Circ. Cardiovasc. Genet.* **2**, 125-133 (2009).
3. Naitza,S. *et al.* A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS. Genet.* **8**, e1002480 (2012).
4. Rose-John,S. IL-6 trans-signaling via the soluble IL-6 receptor: importance for the pro-inflammatory activities of IL-6. *Int. J. Biol. Sci.* **8**, 1237-1247 (2012).
5. Scheller,J., Garbers,C., & Rose-John,S. Interleukin-6: From basic biology to selective blockade of pro-inflammatory activities. *Semin. Immunol.*(2013).
6. Waage,A., Brandtzaeg,P., Halstensen,A., Kierulf,P., & Espevik,T. The complex pattern of cytokines in serum from patients with meningococcal septic shock. Association between interleukin 6, interleukin 1, and fatal outcome. *J. Exp. Med.* **169**, 333-338 (1989).
7. IL6R Genetics Consortium and Emerging Risk Factors Collaboration. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *Lancet* **379**, 1205-1213 (2012).
8. Ferreira,M.A. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006-1014 (2011).
9. Revez,J.A. *et al.* A new regulatory variant in the interleukin-6 receptor gene associates with asthma risk. *Genes Immun.* **14**, 441-446 (2013).
10. Food and Agriculture Organization of the United Nations. World agriculture: towards 2015/2030 (Summary report). 2012. Rome.
11. Dupuis,J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105-116 (2010).
12. Global Lipids Genetics Consortium Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274-1283 (2013).
13. Speliotes,E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937-948 (2010).
14. Levy,D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* **41**, 677-687 (2009).
15. GIANT consortium. The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study (manuscript submitted). 2014.
16. Wood,A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173-1186 (2014).
17. Hindorff,L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A* **106**, 9362-9367 (2009).
18. Frayling,T.M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-894 (2007).
19. Smemo,S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-375 (2014).

20. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
21. Bell,J.T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).
22. Gibbs,J.R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS. Genet.* **6**, e1000952 (2010).
23. Dayeh,T. *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. Genet.* **10**, e1004160 (2014).
24. Dick,K.J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990-1998 (2014).
25. Nilsson,E. *et al.* Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes* **63**, 2962-2976 (2014).
26. Sinclair,K.D. *et al.* DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status. *Proc. Natl. Acad. Sci. U. S. A* **104**, 19351-19356 (2007).
27. Stenvinkel,P. *et al.* Impact of inflammation on epigenetic DNA methylation - a novel risk factor for cardiovascular disease? *J. Intern. Med.* **261**, 488-499 (2007).
28. Agarwal,S. & Rao,A. Modulation of chromatin structure regulates cytokine gene expression during T cell differentiation. *Immunity.* **9**, 765-775 (1998).
29. Plagemann,A. *et al.* Hypothalamic proopiomelanocortin promoter methylation becomes altered by early overfeeding: an epigenetic model of obesity and the metabolic syndrome. *J. Physiol* **587**, 4963-4976 (2009).
30. Lowe,R. *et al.* Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics.* **8**, 445-454 (2013).
31. Smith,A.K. *et al.* DNA extracted from saliva for methylation studies of psychiatric traits: Evidence tissue specificity and relatedness to brain. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*(2014).
32. Klengel,T. *et al.* Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. *Nat. Neurosci.* **16**, 33-41 (2013).
33. Saxena,R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-1336 (2007).
34. Scott,L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341-1345 (2007).
35. Steinthorsdottir,V. *et al.* A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39**, 770-775 (2007).
36. Lencz,T. *et al.* Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Mol. Psychiatry* **12**, 572-580 (2007).
37. O'Donovan,M.C. *et al.* Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat. Genet.* **40**, 1053-1055 (2008).
38. Sullivan,P.F. *et al.* Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Mol. Psychiatry* **13**, 570-584 (2008).
39. Maher,B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).
40. Uher,R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. *Mol. Psychiatry* **14**, 1072-1082 (2009).

41. Schizophrenia Working Group of the Psychiatric Genomics Consortium Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
42. Guo,F. & Jen,K.L. High-fat feeding during pregnancy and lactation affects offspring metabolism in rats. *Physiol Behav.* **57**, 681-686 (1995).
43. Khan,I.Y. *et al.* A high-fat diet during rat pregnancy or suckling induces cardiovascular dysfunction in adult offspring. *Am. J. Physiol Regul. Integr. Comp Physiol* **288**, R127-R133 (2005).
44. Lillycrop,K.A., Phillips,E.S., Jackson,A.A., Hanson,M.A., & Burdge,G.C. Dietary protein restriction of pregnant rats induces and folic acid supplementation prevents epigenetic modification of hepatic gene expression in the offspring. *J. Nutr.* **135**, 1382-1386 (2005).
45. Vucetic,Z., Kimmel,J., Totoki,K., Hollenbeck,E., & Reyes,T.M. Maternal high-fat diet alters methylation and gene expression of dopamine and opioid-related genes. *Endocrinology* **151**, 4756-4764 (2010).
46. Zambrano,E. *et al.* Sex differences in transgenerational alterations of growth and metabolism in progeny (F2) of female offspring (F1) of rats fed a low protein diet during pregnancy and lactation. *J. Physiol* **566**, 225-236 (2005).
47. Heijmans,B.T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. U. S. A* **105**, 17046-17049 (2008).
48. Carone,B.R. *et al.* Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell* **143**, 1084-1096 (2010).
49. Ng,S.F. *et al.* Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring. *Nature* **467**, 963-966 (2010).
50. Dunn,G.A. & Bale,T.L. Maternal high-fat diet effects on third-generation female body size via the paternal lineage. *Endocrinology* **152**, 2228-2236 (2011).
51. Wei,Y. *et al.* Paternally induced transgenerational inheritance of susceptibility to diabetes in mammals. *Proc. Natl. Acad. Sci. U. S. A* **111**, 1873-1878 (2014).
52. Gapp,K. *et al.* Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nat. Neurosci.* **17**, 667-669 (2014).
53. Kaati,G., Bygren,L.O., & Edvinsson,S. Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. *Eur. J. Hum. Genet.* **10**, 682-688 (2002).
54. Pembrey,M.E. *et al.* Sex-specific, male-line transgenerational responses in humans. *Eur. J. Hum. Genet.* **14**, 159-166 (2006).
55. Shoemaker,R., Deng,J., Wang,W., & Zhang,K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20**, 883-889 (2010).
56. Guintivano,J. *et al.* Identification and Replication of a Combined Epigenetic and Genetic Biomarker Predicting Suicide and Suicidal Behaviors. *Am. J. Psychiatry*(2014).
57. Reynard,L.N., Bui,C., Syddall,C.M., & Loughlin,J. CpG methylation regulates allelic expression of GDF5 by modulating binding of SP1 and SP3 repressor proteins to the osteoarthritis susceptibility SNP rs143383. *Hum. Genet.* **133**, 1059-1073 (2014).
58. Manning,K. *et al.* A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* **38**, 948-952 (2006).
59. Cortijo,S. *et al.* Mapping the epigenetic basis of complex traits. *Science* **343**, 1145-1148 (2014).

Nederlandse samenvatting

Nederlandse samenvatting

In landen met een middelmatig tot hoog inkomen vormen complexe ziekten, in het bijzonder hart- en vaatziekten en psychische stoornissen, de belangrijkste oorzaak van ziektelast[1]. Complexe ziekten vertonen geen duidelijk patroon van overerving en er wordt aangenomen dat ze het gevolg zijn van een samenspel tussen genetische factoren en omgevingsfactoren. De onderliggende ziektemechanismen kunnen waarschijnlijk het beste worden begrepen wanneer ze worden bestudeerd vanuit verschillende invalshoeken. In dit proefschrift worden verschillende methoden toegepast om complexe eigenschappen die aan ziekten gerelateerd zijn, met de nadruk op het metabool syndroom en inflammatoire markers, te bestuderen.

Het metabool syndroom

Het metabool syndroom is de benaming voor een combinatie van gezondheidsklachten die nauw samenhangen met overgewicht[2]. Personen met metabool syndroom hebben een verdubbeld risico op hart- en vaatziekten en een meer dan vijfvoudig verhoogd risico op type 2 diabetes[2]. Ook hebben zij een verhoogd risico op vele andere ziekten, waaronder depressie[3]. Een persoon krijgt de diagnose metabool syndroom als ten minste drie van de volgende kenmerken boven een bepaalde klinische drempel liggen: buikomvang, body-mass-index (BMI), taille-heup-ratio, glucoseniveau in bloed, insulineniveau in bloed, systolische bloeddruk, diastolische bloeddruk, HDL-cholesterol en/of triglyceriden. Het wordt algemeen verondersteld dat obesitas (in het bijzonder vet in de buikstreek) en insulineresistentie de belangrijkste onderliggende factoren zijn die het ziektebeeld van metabool syndroom veroorzaken. Meerdere studies suggereren dat de ziektes die verband houden met overgewicht (bijvoorbeeld type II diabetes en hart- en vaatziekten) het gevolg zijn van chronische over-activatie van cellulaire stress signaleringsmechanismen en ontstekingsreacties als gevolg van overmatige energie inname[4, 5]. Bij personen met obesitas scheidt het vetweefsel zogenaamde pro-inflammatoire cytokines zoals TNF-alfa en IL-6 af[6, 7]. Deze moleculen stimuleren vervolgens de afgifte van zogenaamde acutefase-eiwitten (bijvoorbeeld fibrinogeen en CRP)[8]. Als gevolg van obesitas en/of overmatige energieopname vinden ook ontstekingsreacties plaats in de lever, pancreas en hersenen die bijdragen aan obesitas gerelateerde aandoeningen [9-11]. In Nederland heeft 14% van de bevolking metabool syndroom[12] en in de Verenigde Staten zelfs 37%[13]. Het metabool syndroom komt vaker voor onder mannen[14] en onder oudere personen[12, 15]. De toenemende prevalentie van het metabool syndroom is waarschijnlijk het gevolg van de stijgende prevalentie van obesitas[16].

Obesitas en veranderingen in levensstijl

In de periode tussen 1980 en 2008 is de wereldwijde prevalentie van obesitas bijna verdubbeld; van 6,4% naar 12%[17]. In Nederland was de prevalentie van overgewicht en obesitas in 2008 respectievelijk 47,8% en 16,2% bij volwassenen (leeftijd 20+)[18]. Er is geen twijfel dat de uitbraak van overgewichtgerelateerde ziekten te maken heeft met de hedendaagse levensstijl. Wereldwijd is de gemiddelde calorie-inname per persoon toegenomen met 450 kcal per dag tussen de 1960 en eind jaren '90[19]. Volgens een publicatie uit 2012 is 31,1% van alle volwassen individuen wereldwijd (leeftijd 15+) "fysiek inactief"[20]. Men denkt dat obesitas en vele andere hedendaagse complexe ziekten het gevolg zijn van een 'mis-match' tussen onze huidige leefstijl en onze evolutionaire geschiedenis, een onderwerp dat ik toelicht in **hoofdstuk 8** van mijn proefschrift.

Tegenwoordig leven we in welvarende landen allemaal in een omgeving die gewichtstoename bevordert. Omgevingen en banen beperken bijvoorbeeld de noodzaak om fysiek actief te zijn en voedsel met veel calorieën is beschikbaar voor iedereen. Toch ontwikkelt niet iedere personen obesitas en heeft niet iedere zwaarlijvige individu metabool syndroom. Dit suggereert dat individuen verschillen in hun gevoeligheid voor het ontwikkelen van overgewicht, obesitas en metabole ziekten.

Erfelijkheid

In samenlevingen waarin de meerderheid van de bevolking voldoende voedsel ter beschikking heeft worden verschillen in BMI tussen personen grotendeels verklaard door genetische factoren[21]. Hetzelfde geldt voor inflammatoire markers (zoals TNF-alfa en IL-6), cholesterolniveau, bloedglucose en bloeddruk[22]. In de laatste jaren zijn verschillende varianten in de DNA-code (risico allelen) ontdekt die individuele verschillen in deze eigenschappen veroorzaken[23-25].

Naast het onderzoek naar de DNA-code zijn grote studies van menselijke populaties nu ook begonnen met het onderzoeken van andere vormen van moleculaire variatie die menselijke complexe eigenschappen beïnvloeden. Zo hebben recente studies gevonden dat BMI[26] , type 2 diabetes[27] en vaatziekten[28] samenhangen met zogenaamde epigenetische variatie.

Epigenetica en DNA methylatie

Bijna iedere cel in het menselijk lichaam bevat dezelfde DNA-code (het genoom) die een persoon erft van zijn/haar ouders. Op basis van dit ene genoom worden vele verschillende celtypes (bijvoorbeeld huidcellen en hersencellen) en organen gevormd doordat verschillende genen op het juiste moment worden aan- en uitgeschakeld in elke cel. Deze regulatie van genactiviteit ("genexpressie") wordt gecoördineerd in elke cel door talrijke mechanismen die gezamenlijk beïnvloeden hoe strak het DNA-molecuul

opgerold ligt[29]. DNA methylatie is één van deze mechanismen, die ik bestudeer in deel 3 van mijn proefschrift. DNA methylatie houdt in dat een methyl-groep (CH_3) op bepaalde plekken aan het DNA molecuul wordt bevestigd. De mate van methylatie van een gen kan bepalen hoe actief dit gen is. Terwijl de genetica traditioneel de DNA-code bestudeert, heeft de epigenetica betrekking op de studie van de moleculaire mechanismen die genexpressie beïnvloeden[30].

De DNA-code verandert nauwelijks gedurende het leven. DNA-methylatie en andere epigenetische mechanismen kunnen wel veranderen gedurende de ontwikkeling[31], als onderdeel van het ouder worden[32], en als gevolg van specifieke omgevingsfactoren, zoals roken[33]. Daarom kan het bestuderen van deze epigenetische mechanismen mogelijk licht werpen op ziektemechanismen die verborgen zouden blijven als we ons alleen focussen op de DNA-code. Epigenetische mechanismen kunnen mogelijk ook verklaren waarom eeneiige tweelingparen vaak “discordant” zijn voor complexe ziekten (wat betekent dat één van de twee tweelingen de ziekte heeft en de ander niet). Hoewel eeneiige tweelingen dezelfde DNA-code hebben, kunnen zij wel verschillen in de methylatie van hun DNA. Dit heb ik onderzocht in **hoofdstuk 7**.

Onderzoeksvragen

Het in dit proefschrift beschreven onderzoek wordt gekenmerkt door twee belangrijke thema's: het eerste heeft betrekking op de vraag in welke mate verschillen tussen personen in inflammatoire markers en kenmerken van het metabool syndroom worden veroorzaakt door genetische factoren en omgevingsfactoren. De tweede belangrijke focus van dit proefschrift ligt op het onderzoeken van de rol van genen en omgeving in variatie tussen personen in DNA-methylatie. Deze onderzoeksthema's zijn het onderwerp van een reeks hoofdstukken waarin ik gegevens analyseer die werden verzameld in tweelinggezinnen die deelnemen aan onderzoek van het Nederlands Tweelingen Register. De rol van genen en omgeving wordt onderzocht met behulp van statische analyses waarin de gelijkenis van eeneiige en twee-eiige tweelingen en hun familieleden wordt vergeleken, en met behulp van analyses die gebruik maken van gemeten genetische varianten in de DNA-code. **Hoofdstuk 3** van dit proefschrift (over interleukine-6 receptorconcentratie in bloed) laat zien hoe deze verschillende benaderingen kunnen worden gecombineerd om de bijdrage van reeds geïdentificeerde genetische varianten aan een complexe eigenschap te schatten versus genetische varianten die nog ontdekt moeten worden.

Deel 1: Erfelijkheid van inflammatoire markers

In het eerste deel van dit proefschrift (**hoofdstuk 2 en 3**) heb ik de invloed van genetische factoren en omgevingsfactoren op de concentratie van inflammatoire markers in bloed onderzocht, waaronder pro-inflammatoire cytokines (IL-6 en TNF-α), een cytokine receptor (sIL-6R) en de acutefase-

eiwitten fibrinogeen en CRP. Door concentraties van deze moleculen te vergelijken tussen tweelingen, en hun broers, zussen en ouders, heb ik de erfelijkheid van deze eigenschappen geschat. De erfelijkheid geeft weer hoeveel procent van de variatie in concentraties tussen personen wordt verklaard door genetische verschillen. Ik vond dat de variatie in al deze inflammatoire markers in belangrijke mate verklaard wordt door genetische variatie. De niveaus van sIL-6R in bloed bijvoorbeeld waren zeer erfelijk (72% van de verschillen in concentraties tussen personen kwam door genetische verschillen). De bijdrage van specifieke varianten in de DNA-code aan de erfelijkheid van een eigenschap kan worden onderzocht door de klassieke schatting van erfelijkheid te combineren met de analyse van gemeten varianten in de DNA-code. In **hoofdstuk 3** heb ik deze combinatie van methoden toegepast op de concentratie van sIL-6R in bloed. De variatie in sIL-6R concentratie tussen personen wordt grotendeels verklaard door één enkele variant (rs2228145) in het *IL6R* gen die de productie van sIL-6R beïnvloedt. Voor de meeste complexe eigenschappen is het juist zo dat de variatie tussen personen door heel veel DNA-varianten wordt beïnvloed wordt en dat ieder van deze varianten afzonderlijk maar een heel klein effect heeft.

Deel 2: Invloed van genetische factoren en omgevingsfactoren op BMI en andere kenmerken van het metabool syndroom

In het tweede deel van mijn proefschrift heb ik de erfelijkheid van lengte en kenmerken van het metabool syndroom onderzocht, waaronder BMI, buikomvang, taille-heup-ratio, lipiden, bloed glucose en insuline en bloeddruk (**hoofdstuk 4**). Lengte, BMI en middelomtrek waren de meest erfelijke eigenschappen (erfelijkheid = 90%, 78%, en 76%). Lipiden en bloeddruk waren ook zeer erfelijk (erfelijkheid = 59% en 69%) en taille-heup-ratio, insuline niveau en glucose niveau waren matig erfelijk (erfelijkheid = 49%, 47%, en 53%).

In **hoofdstuk 5** heb ik de invloed van genen en omgeving op BMI op een andere manier bestudeerd, namelijk door het BMI van eeneiige tweelingen die tussen 1991 en 2011 aan NTR onderzoek hebben meegedaan over de tijd heen te vergelijken. De belangrijkste bevinding was dat grote verschillen in BMI tussen eeneiige tweelingen ("BMI discordantie", gedefinieerd als een verschil van tenminste 3 kg / m^2), zeer zeldzaam zijn, zeker op de lange termijn. BMI discordantie kwam vaker voor op de latere tijdspunten van het onderzoek. Gedurende de periode van het onderzoek steeg ook het gemiddelde BMI en de leeftijd van tweelingparen (gemiddelde leeftijd bij vragenlijst 1 = 17 jaar, gemiddelde leeftijd bij vragenlijst 8 = 34 jaar). Ik vond dat van alle eeneiige tweelingen die op een bepaald punt in hun leven discordant werden, de meesten later al snel weer een vergelijkbaar BMI hadden. Deze bevindingen suggereren dat mensen een bepaald "set-point" van BMI hebben welke sterk genetisch bepaald is en dat dit set-point kan vervagen na grote gewichtstoename, waardoor eeneiige tweelingen naarmate zij zwaarder

worden ook meer van elkaar gaan verschillen. Een andere belangrijke bevinding van dit onderzoek was dat tweelingen die discordant zijn voor BMI ook verschillen vertoonden in de concentraties van metabole biomarkers (bijvoorbeeld bloedglucosewaarden) en inflammatoire markers in bloed. Deze verschillen waren niet aanwezig in eeneiige tweelingparen die pas later in hun leven discordant werden voor BMI. Deze bevindingen illustreren dat de concentraties van deze biomarkers niet voorspellen of een persoon een hoog BMI zal ontwikkelen, maar dat ongezonde niveaus van deze biomarkers het gevolg zijn van gewichtstoename.

Deel 3: Bovenop de DNA-code: Epigenetica

Een toenemend aantal studies wijst op het belang van epigenetische mechanismen, zoals DNA methylatie, voor de in deel 1 en 2 van mijn proefschrift onderzochte eigenschappen. In deel 3 van dit proefschrift heb ik de oorzaken van variatie tussen mensen in de mate van methylatie van hun DNA onderzocht. In **hoofdstuk 6** beschrijf ik dat het methylatie-niveau op vele plekken van het DNA in bloed erfelijk is en dat er ook veel locaties zijn waar een groot deel van de variatie is toe te schrijven aan niet-genetische invloeden, waaronder de omgeving van een persoon en stochastische processen. In dit hoofdstuk onderzocht ik niet alleen de 'totale erfelijkheid' van DNA-methylatie niveaus maar onderzocht ik ook hoeveel van de verschillen in DNA methylatie tussen personen verklaard kunnen worden op basis van alle varianten in de DNA-code die wij bij personen hebben gemeten. De totale erfelijkheid van DNA-methylatie niveau was gemiddeld 22%. Gemiddeld 7% van de variatie in DNA methylatie werd verklaard door alle gemeten varianten in de DNA-code. In hoofdstuk 6 liet ik ook zien dat er een groot aantal plekken in het genoom is waar de erfelijkheid van DNA methylatie afneemt met leeftijd, een kleiner aantal plekken waar de erfelijkheid toeneemt met leeftijd, en ik vond plekken waar de erfelijkheid van DNA methylatie verschilt tussen mannen en vrouwen. Deze bevindingen kunnen mogelijk inzicht geven in ziektes die pas op latere leeftijd ontstaan en ziektes die vaker in mannen of vrouwen voorkomen.

Grootschalige epigenetische studies in levende mensen zijn alleen mogelijk als DNA wordt verkregen uit gemakkelijk bereikbaar weefsels. In **hoofdstuk 6** onderzocht ik DNA-methylatie in bloedmonsters en in **hoofdstuk 7** bestudeerde ik methylatie van DNA verkregen uit wangslijmvlies van eeneiige tweelingparen. Wanguitstrijkjes zijn makkelijker te verzamelen dan bloedmonsters, zeker bij (jonge) kinderen. De resultaten uit **hoofdstuk 7** lieten zien dat DNA methylatie in wangslijmvlies, vergelijkbaar met mijn bevindingen in bloed, beïnvloed wordt door zowel genetische factoren als omgevingsfactoren. Ook vond ik in **hoofdstuk 7** dat methylatie van sommige locaties in het DNA in eeneiige tweelingen meer verschilt dan op andere locaties. Mogelijk zijn locaties die meer verschillen in eeneiige tweelingen gevoeliger voor veranderingen door de omgeving of voor stochastische invloeden.

Deel 4: Tweelingstudies en complexe eigenschappen: Toekomst en verdere overwegingen

Voor complexe ziekten en eigenschappen zoals type II diabetes, insuline niveau en cholesterol zijn al vele genen geïdentificeerd. Studies naar psychische stoornissen waren aanvankelijk minder succesvol in het identificeren van genen. Voor schizofrenie bijvoorbeeld, verklaren de tot nu toe ontdekte genetische varianten slechts een heel klein deel van de totale variatie. Mogelijk is het zo moeilijk om genetische varianten te vinden die een groot deel van het risico op psychiatrische stoornissen verklaren door de evolutionaire geschiedenis van deze aandoeningen. Hierover gaat **hoofdstuk 8**, waarin ik een samenvatting heb gemaakt van evolutionaire theorieën die zijn voorgesteld om te verklaren waarom schizofrenie, een van de meest ernstige en meest erfelijke psychische stoornissen, blijft voortbestaan in de populatie ondanks dat patiënten vaak geen (of minder) kinderen krijgen. Voorgestelde evolutionaire verklaringen zijn 'balancing seletion', 'fitness trade-offs', wisselende omgevingen, seksuele selectie, mutatie-selectie balans en genomische conflicten.

Al decennia lang geeft het klassieke tweelingmodel inzicht in de rol van genen en omgeving in verschillen in menselijke complexe eigenschappen, zoals ziekte en gedrag. In **hoofdstuk 9** beschrijf ik onderzoeksvragen die kunnen worden onderzocht door het klassieke tweelingmodel of het discordante tweelingmodel toe te passen op moleculaire karakteristieken, zoals "metabolomics", genexpressie, het microbioom en epigenetische mechanismen. Discordante eeneiige tweelingparen kunnen in de toekomst bijdragen aan nieuwe inzichten in de rol van epigenetische mechanismen in complexe ziekten.

Referentielijst

1. Mathers,C., Fat,D.M., & Boerma,J.T. *The global burden of disease: 2004 update*(World Health Organization,2008).
2. Cornier,M.A. *et al.* The metabolic syndrome. *Endocr. Rev.* **29**, 777-822 (2008).
3. Pan,A. *et al.* Bidirectional association between depression and metabolic syndrome: a systematic review and meta-analysis of epidemiological studies. *Diabetes Care* **35**, 1171-1180 (2012).
4. Ozcan,U. *et al.* Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* **306**, 457-461 (2004).
5. Stienstra,R., Tack,C.J., Kanneganti,T.D., Joosten,L.A., & Netea,M.G. The inflammasome puts obesity in the danger zone. *Cell Metab* **15**, 10-18 (2012).
6. Berg,A.H. & Scherer,P.E. Adipose tissue, inflammation, and cardiovascular disease. *Circ. Res.* **96**, 939-949 (2005).
7. Hotamisligil,G.S., Arner,P., Caro,J.F., Atkinson,R.L., & Spiegelman,B.M. Increased adipose tissue expression of tumor necrosis factor-alpha in human obesity and insulin resistance. *J. Clin. Invest* **95**, 2409-2415 (1995).
8. Shoelson,S.E., Lee,J., & Goldfine,A.B. Inflammation and insulin resistance. *J. Clin. Invest* **116**, 1793-1801 (2006).
9. Cai,D. *et al.* Local and systemic insulin resistance resulting from hepatic activation of IKK-beta and NF-kappaB. *Nat. Med.* **11**, 183-190 (2005).
10. De Souza,C.T. *et al.* Consumption of a fat-rich diet activates a proinflammatory response and induces insulin resistance in the hypothalamus. *Endocrinology* **146**, 4192-4199 (2005).
11. Ehses,J.A. *et al.* Increased number of islet-associated macrophages in type 2 diabetes. *Diabetes* **56**, 2356-2370 (2007).
12. Bos,M.B. *et al.* The prevalence of the metabolic syndrome in the Netherlands: increased risk of cardiovascular diseases and diabetes mellitus type 2 in one quarter of persons under 60. *Nederlands tijdschrift voor geneeskunde* **151**, 2382-2388 (2007).
13. Cameron,A.J., Shaw,J.E., & Zimmet,P.Z. The metabolic syndrome: prevalence in worldwide populations. *Endocrinology and metabolism clinics of North America* **33**, 351-375 (2004).
14. Dekker,J.M. *et al.* Metabolic syndrome and 10-year cardiovascular disease risk in the Hoorn Study. *Circulation* **112**, 666-673 (2005).
15. Oosterwerff,M.M., van Schoor,N.M., Lips,P., & Eekhoff,E.M. Osteocalcin as a predictor of the metabolic syndrome in older persons: a population-based study. *Clin. Endocrinol. (Oxf)* **78**, 242-247 (2013).
16. National Institutes of Health. http://www.nhlbi.nih.gov/health/health-topics/topics/ms/, Accessed on 03/11/. 2014.
17. Stevens,G.A. *et al.* National, regional, and global trends in adult overweight and obesity prevalences. *Popul. Health Metr.* **10**, 22 (2012).
18. World Health Organization. http://www.who.int/gho/ncd/risk_factors/overweight/en/. Accessed on 12/06/. 2014.
19. Food and Agriculture Organization of the United Nations. World agriculture: towards 2015/2030 (Summary report). 2012. Rome.
20. Hallal,P.C. *et al.* Global physical activity levels: surveillance progress, pitfalls, and prospects. *Lancet* **380**, 247-257 (2012).
21. Elks,C.E. *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol. (Lausanne)* **3**, (2012).

22. Pilia,G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
23. Dupuis,J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105-116 (2010).
24. Global Lipids Genetics Consortium Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274-1283 (2013).
25. Speliotes,E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937-948 (2010).
26. Dick,K.J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383**, 1990-1998 (2014).
27. Dayeh,T. *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. Genet.* **10**, e1004160 (2014).
28. Connelly,J.J. *et al.* Epigenetic regulation of COL15A1 in smooth muscle cell replicative aging and atherosclerosis. *Hum. Mol. Genet.* **22**, 5107-5120 (2013).
29. Goldberg,A.D., Allis,C.D., & Bernstein,E. Epigenetics: a landscape takes shape. *Cell* **128**, 635-638 (2007).
30. Bird,A. Perceptions of epigenetics. *Nature* **447**, 396-398 (2007).
31. Reik,W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425-432 (2007).
32. Horvath,S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
33. Philibert,R.A., Beach,S.R., Lei,M.K., & Brody,G.H. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clin. Epigenetics.* **5**, 19 (2013).

Appendices

Appendix 1

Supplement to chapter 3

Full author list

Jenny van Dongen,[1,2,*] Rick Jansen,[3,4] Dirk Smit,[1,4] Jouke-Jan Hottenga,[1,2] Hamdi Mbarek,[1,2] Gonneke Willemsen,[1,2] Cornelis Kluft,[5] David L. Duffy,[6] Melanie C. Matheson,[7] Guy B. Marks,[8] Jennie Hui,[9,10,11,12] Peter Le Souëf,[13] Patrick Danoy,[14] Svetlana Baltic,[15] Dale R. Nyholt,[6] Mark Jenkins,[7] Catherine Hayden,[13] John Beilby,[9,11,12] Faang Cheah,[15] Pamela A. Madden,[16] Andrew C. Heath,[16] John L. Hopper,[7] Bill Musk,[10,12,17,18] Stephen R. Leeder,[19] Eugene H Walters,[20] Nicholas G. Martin,[6] Alan James,[12,18,21] Graham Jones,[22] Michael J. Abramson,[23] Colin F. Robertson,[24] Shyamali C. Dharmage,[7] Matthew A. Brown,[14] Grant W. Montgomery,[6] Philip J. Thompson,[15] Brenda W.J. Penninx,[2,3,4] Manuel A. Ferreira,[6] Dorret I. Boomsma,[1,2,4,25] Eco J. C. de Geus,[1,2,25]

[1] Department of Biological Psychology, Van der Boechorststraat 1, 1081 BT, VU University Amsterdam, Amsterdam, The Netherlands
[2]EMGO institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081BT, Amsterdam, The Netherlands
[3]Department of Psychiatry, VU University Medical Center, A.J. Ernststraat 1187, 1081HL, Amsterdam, The Netherlands
[4]Neuroscience Campus Amsterdam, De Boelelaan 1085, 1081HV, Amsterdam, The Netherlands
[5]Good Biomarker Sciences, Zernikedreef 8, 2333CL, Leiden, The Netherlands
[6]Queensland Institute of Medical Research, Royal Brisbane Hospital, Locked Bag 2000, Herston QLD 4029, Australia
[7] Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Australia.
[8] Woolcock Institute of Medical Research, University of Sydney, Sydney, Australia.
[9] PathWest Laboratory Medicine of Western Australia (WA), Nedlands, Australia.
[10] School of Population Health, The University of WA, Nedlands, Australia.
[11] School of Pathology and Laboratory Medicine, The University of WA, Nedlands, Australia.
[12] Busselton Population Medical Research Foundation, Sir Charles Gairdner Hospital, Perth, Australia.
[13] School of Paediatrics and Child Health, Princess Margaret Hospital for Children, Perth, Australia.
[14] University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, Australia.
[15] Lung Institute of WA and Centre for Asthma, Allergy and Respiratory Research, University of WA, Perth, Australia.
[16] Washington University School of Medicine, St Louis, United States
[17] Department of Respiratory Medicine, Sir Charles Gairdner Hospital, Perth, Australia
[18] School of Medicine and Pharmacology, University of WA, Nedlands, Australia
[19] Australian Health Policy Institute, University of Sydney, Sydney, Australia
[20] School of Medicine, University of Tasmania
[21] Department of Pulmonary Physiology, West Australian Sleep Disorders Research Institute, Nedlands, Australia
[22] School of Science and Health, University of Western Sydney, Penrith, Australia
[23] Department of Epidemiology& Preventive Medicine, Monash University, Melbourne, Australia.
[24] Respiratory Medicine, Murdoch Children's Research Institute, Melbourne, Australia
[25] These authors contributed equally to the manuscript

Figure S1: Variance of sIL-6R level explained by genome-wide SNPs in unrelated individuals.
Estimates are expressed as a proportion of the total variance of sIL-6R level (V_A/V_P). Error bars represent standard errors. *IL6R*= All SNPs in the *IL6R* gene +/- 10 MB on chromosome 1 (1q21.3). 1 rest= All other SNPs on chromosome 1.

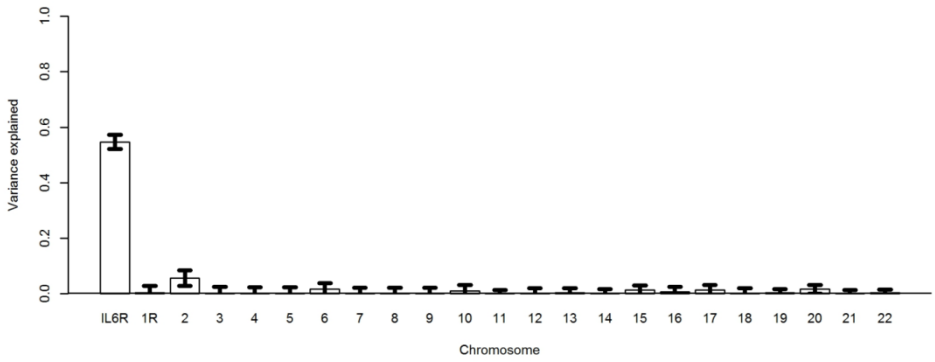


Figure S2: Association between rs2228145 genotype and the ratio of *IL6R* exon 9 expression over mean *IL6R* expression. Boxplots of the relative transcript abundance (transcripts containing exon 9/ mean expression of all *IL6R* transcripts) in lymphoblastoid cell lines from HAPMAP (GEO accession nr = GSE9703, N=162) are displayed for each rs2228145 genotype (P < 0.01).

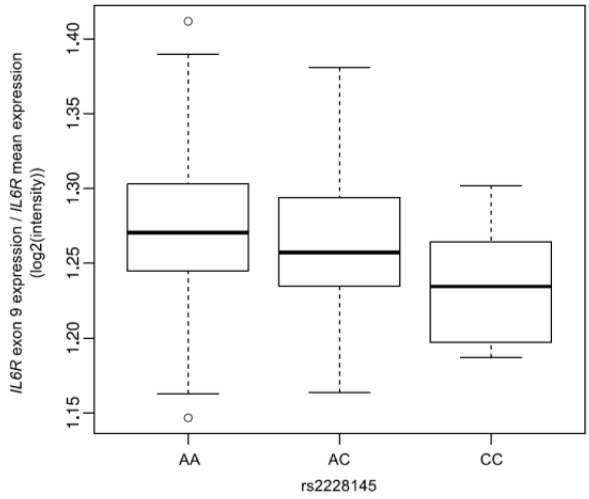


Table S1: Characteristics of subjects included in the analyses.

Analysis	N subjects	Mean age (SD), min-max	% Male	Cohort
Heritability analysis and biometrical model (MZ and DZ twins, siblings, and parents)	4980	42.7 (14.3), 18-89	36.2	NTR
GWA and GCTA (unrelated + related Ss)	4846	44.2 (14.4), 18-90	38.7	NTR
GCTA (unrelated Ss)	2875	46.5 (14.4), 18-89	38.8	NTR
Combined linkage and association analysis (Nuclear families)	1254	48.3 (15.7), 18-89	44.4	NTR
eQTL analysis (unrelated + related Ss)	4467	38.4 (13.0), 25-51	34.4	NTR + NESDA
Correlation between sIL-6R level and *IL6R* expression (unrelated + related Ss)	2727	37.5 (12.0), 18-79	34.5	NTR

Table S2: GWA results and information for all 680 SNPs reaching genome-wide significance in the GWAS of sIL-6R level.

SNP[A]	Chr	Position[B]	A1[C]	A2[D]	Frequency A1[E]	Info[F]	Beta[G]	SE[H]	*P*-value
rs11265611	1	154395125	G	A	0.42	0.894	-9748.6	236.8	< 1 E-176
rs4133213	1	154395212	C	A	0.58	0.893	-11519.4	219.2	< 1 E-176
rs6689306	1	154395946	A	G	0.45	0.937	-9896.3	226.5	< 1 E-176
rs12118721	1	154397416	T	C	0.45	0.961	-9916.3	222.9	< 1 E-176
rs12118770	1	154397589	T	C	0.44	0.959	-9827.9	223.7	< 1 E-176
rs12117832	1	154397610	A	G	0.45	0.959	-9890.3	222.6	< 1 E-176
rs10908836	1	154397932	C	T	0.46	0.913	-9625.9	230.7	< 1 E-176
rs35109459	1	154397933	G	A	0.47	0.917	-9778.4	228.7	< 1 E-176
rs10908838	1	154397984	T	G	0.44	0.963	-9822.9	223.4	< 1 E-176
rs6687726	1	154400320	A	G	0.45	0.969	-9890.7	222.1	< 1 E-176
rs6427658	1	154400799	T	C	0.45	0.965	-9877.7	221.5	< 1 E-176
rs6694817	1	154401972	T	C	0.44	0.965	-9923.9	221.6	< 1 E-176
rs7549250	1	154404336	C	T	0.45	0.970	-10068.0	219.3	< 1 E-176
rs7549338	1	154404380	C	G	0.44	0.970	-9995.9	219.8	< 1 E-176
rs56383622	1	154405024	A	G	0.60	0.969	-12091.3	204.4	< 1 E-176
rs4845619	1	154405058	T	G	0.45	0.970	-10088.9	219.2	< 1 E-176
rs59632925	1	154406540	T	G	0.44	0.970	-10014.3	219.7	< 1 E-176
rs4845620	1	154406656	A	G	0.61	0.969	-12111.3	203.9	< 1 E-176
rs7521458	1	154407713	T	C	0.61	0.969	-12123.6	203.6	< 1 E-176
rs4845371	1	154408340	T	C	0.44	0.969	-10034.3	219.6	< 1 E-176
rs6667434	1	154409100	A	G	0.44	0.969	-10040.3	219.6	< 1 E-176
rs4845621	1	154409730	G	A	0.61	0.970	-12152.8	203.3	< 1 E-176
rs4845622	1	154411419	A	C	0.61	0.970	-12171.5	203.1	< 1 E-176
rs4393147	1	154414037	C	T	0.60	0.973	-12194.9	202.5	< 1 E-176
rs4453032	1	154414086	A	G	0.60	0.973	-12195.2	202.5	< 1 E-176
rs6664201	1	154414296	C	T	0.60	0.973	-12196.2	202.5	< 1 E-176
rs4845372	1	154415396	C	A	0.60	0.974	-12019.9	204.0	< 1 E-176
rs12753254	1	154416935	G	A	0.61	0.977	-12294.5	199.9	< 1 E-176
rs12730036	1	154416969	C	T	0.61	0.977	-12296.6	199.8	< 1 E-176
rs11265612	1	154417044	A	G	0.44	0.975	-10170.5	217.6	< 1 E-176
rs4845373	1	154417829	C	T	0.61	0.979	-12291.5	200.7	< 1 E-176
rs6683206	1	154418088	T	C	0.44	0.969	-10228.0	218.1	< 1 E-176
rs11265613	1	154418415	T	C	0.61	0.983	-12322.0	198.5	< 1 E-176
rs4576655	1	154418749	C	T	0.60	0.989	-12163.1	199.6	< 1 E-176
rs6686750	1	154419843	A	G	0.45	0.990	-10227.6	215.8	< 1 E-176
rs12730935	1	154419892	G	A	0.61	0.985	-12393.1	198.6	< 1 E-176
rs61812598	1	154420087	G	A	0.61	0.993	-12357.8	197.2	< 1 E-176
rs7512646	1	154420402	G	C	0.60	0.991	-12186.8	198.9	< 1 E-176
rs7529229	1	154420778	T	C	0.60	0.991	-12189.7	198.8	< 1 E-176
rs12404927	1	154421438	T	C	0.48	0.931	-10032.6	229.5	< 1 E-176
rs12129500	1	154423764	T	C	0.45	0.992	-10227.7	215.7	< 1 E-176
rs7536152	1	154423909	A	G	0.45	0.992	-10228.0	215.6	< 1 E-176
rs7526131	1	154425135	G	A	0.45	0.993	-10228.4	215.6	< 1 E-176
rs12126142	1	154425456	G	A	0.61	0.997	-12361.6	196.6	< 1 E-176
rs6689393	1	154426097	A	G	0.45	0.993	-10229.8	215.7	< 1 E-176
rs12133641	1	154428283	A	G	0.61	0.993	-12381.4	196.2	< 1 E-176
rs6694258	1	154428505	C	A	0.45	0.991	-10267.7	214.6	< 1 E-176
rs6690230	1	154432877	C	G	0.40	0.990	-6873.5	232.7	2.58E-176
rs6695045	1	154432957	A	G	0.40	0.990	-6872.5	232.7	2.96E-176
rs12044132	1	154462360	C	T	0.85	0.853	-10008.7	341.2	3.70E-174
rs55676222	1	154435289	A	T	0.89	0.637	-13456.9	485.4	5.24E-157
rs10752641	1	154432042	G	C	0.24	0.973	-6740.6	248.2	3.75E-151
rs6698040	1	154432948	T	C	0.24	0.990	-6658.3	247.1	5.34E-149
rs7537291	1	154433407	G	A	0.24	0.990	-6650.5	247.1	1.11E-148
rs7546068	1	154433415	C	T	0.24	0.990	-6650.4	247.1	1.13E-148

rs7537316	1	154433466	G	A	0.24	0.990	-6649.5	247.1	1.23E-148
rs12023772	1	154483868	G	A	0.85	0.920	-8936.2	332.4	2.12E-148
rs72698169	1	154486799	A	C	0.85	0.920	-8930.3	332.3	2.66E-148
rs7546552	1	154433905	C	A	0.24	0.992	-6635.3	247.0	3.76E-148
rs7546555	1	154433911	C	T	0.24	0.992	-6635.2	247.1	3.83E-148
rs6427672	1	154435346	C	T	0.24	0.992	-6629.8	247.3	1.14E-147
rs6687597	1	154434936	G	A	0.24	0.992	-6618.1	247.5	5.38E-147
rs11265621	1	154442960	G	A	0.37	0.991	-6427.1	240.4	6.34E-147
rs10908839	1	154430798	C	G	0.24	0.986	-6641.6	248.8	1.73E-146
rs4509570	1	154436384	G	C	0.24	0.994	-6585.7	247.4	8.69E-146
rs4341355	1	154436404	C	G	0.24	0.994	-6585.8	247.4	9.25E-146
rs4638123	1	154505704	C	T	0.37	0.997	-6418.6	241.2	1.20E-145
rs12407048	1	154505270	C	T	0.37	0.997	-6418.6	241.2	1.20E-145
rs12037271	1	154507888	G	A	0.37	0.996	-6418.8	241.2	1.22E-145
rs4478801	1	154464572	G	A	0.37	0.994	-6413.3	241.0	1.27E-145
rs6698971	1	154475331	C	A	0.37	0.995	-6424.3	241.5	1.55E-145
rs11265632	1	154514067	G	A	0.37	0.995	-6413.2	241.1	1.74E-145
rs11265628	1	154496489	C	T	0.37	0.997	-6412.6	241.2	2.07E-145
rs10908843	1	154494480	C	G	0.37	0.997	-6412.5	241.2	2.09E-145
rs12136771	1	154511502	T	C	0.37	0.994	-6419.6	241.5	2.11E-145
rs10908845	1	154500898	C	T	0.37	0.996	-6412.9	241.4	3.43E-145
rs10908847	1	154504954	G	A	0.37	0.995	-6418.7	241.7	4.36E-145
rs4845647	1	154514331	C	A	0.37	0.995	-6407.6	241.3	4.90E-145
rs4845642	1	154498028	G	A	0.37	0.996	-6410.8	241.5	5.48E-145
rs4345797	1	154495675	C	T	0.37	0.997	-6406.9	241.4	5.76E-145
rs10908842	1	154492702	G	A	0.37	0.998	-6390.6	241.2	1.82E-144
rs6700296	1	154473660	T	C	0.37	0.998	-6388.9	241.2	1.85E-144
rs4845638	1	154490269	T	A	0.37	0.998	-6389.9	241.2	1.95E-144
rs12128408	1	154488533	A	G	0.37	0.998	-6389.8	241.2	1.97E-144
rs12740969	1	154487060	T	G	0.37	0.998	-6389.7	241.2	1.99E-144
rs12118018	1	154477440	G	A	0.37	0.999	-6388.7	241.2	2.00E-144
rs4845637	1	154490178	A	G	0.37	0.998	-6390.8	241.3	2.00E-144
rs10908841	1	154487763	C	T	0.37	0.998	-6389.5	241.2	2.10E-144
rs6664608	1	154479670	C	T	0.37	0.999	-6388.3	241.2	2.23E-144
rs4382717	1	154518610	C	T	0.37	0.989	-6410.5	242.1	2.40E-144
rs6689965	1	154470606	T	A	0.37	0.999	-6386.6	241.2	2.43E-144
rs6686276	1	154466188	C	A	0.37	0.998	-6385.8	241.2	2.66E-144
rs6684921	1	154464945	A	C	0.37	0.997	-6388.3	241.4	3.51E-144
rs12753680	1	154474900	G	A	0.37	0.998	-6385.9	241.3	3.98E-144
rs4845639	1	154490352	C	T	0.37	0.998	-6383.3	241.4	5.47E-144
rs7519499	1	154487926	G	A	0.37	0.998	-6383.1	241.4	5.56E-144
rs7518694	1	154484788	C	G	0.37	0.998	-6382.9	241.4	5.69E-144
rs9660786	1	154484017	A	T	0.37	0.998	-6382.8	241.4	5.72E-144
rs12119111	1	154478600	G	A	0.37	0.998	-6382.0	241.4	6.35E-144
rs6658175	1	154475330	T	A	0.37	0.997	-6377.4	241.6	1.65E-143
rs12753666	1	154474875	G	A	0.38	0.986	-6404.8	243.1	5.49E-143
rs9616	1	154555733	A	T	0.71	0.978	-7349.6	282.2	6.01E-140
rs11265616	1	154419980	T	C	0.60	0.702	-7757.1	301.5	7.72E-137
rs55668699	1	154435293	T	A	0.76	0.955	6435.5	253.5	1.80E-133
rs78664422	1	154499246	G	A	0.84	0.944	6973.1	286.3	1.13E-123
rs4072391	1	154438880	T	C	0.20	0.983	-6376.1	262.8	9.85E-123
rs112231452	1	154504900	G	A	0.79	0.988	6356.5	262.1	1.25E-122
rs61275241	1	154504887	T	G	0.79	0.988	6332.1	262.2	1.16E-121
rs60255122	1	154504702	A	G	0.79	0.988	6332.1	262.2	1.16E-121
rs60760897	1	154504849	C	T	0.79	0.988	6330.6	262.3	1.49E-121
rs61698846	1	154504854	C	T	0.79	0.988	6330.6	262.3	1.49E-121
rs61403567	1	154504927	T	C	0.79	0.988	6339.2	263.1	3.75E-121
rs77994623	1	154505106	C	T	0.84	0.933	6922.3	287.4	4.36E-121

rs60498985	1	154504172	C	T	0.79	0.995	6296.4	261.4	4.53E-121
rs7526293	1	154444209	T	C	0.21	0.990	-6278.9	261.2	1.25E-120
rs6669229	1	154444591	A	G	0.21	0.990	-6277.4	261.3	1.46E-120
rs58348886	1	154446198	T	C	0.79	0.990	6280.2	261.6	2.25E-120
rs59741504	1	154504601	T	C	0.79	0.991	6311.6	263.0	2.70E-120
rs12405637	1	154501285	C	T	0.79	0.995	6284.0	262.4	7.39E-120
rs57783436	1	154511709	A	G	0.79	0.995	6283.0	262.5	9.98E-120
rs73023346	1	154512077	C	T	0.79	0.995	6283.1	262.5	1.00E-119
rs12408461	1	154512710	T	C	0.79	0.994	6283.2	262.5	1.01E-119
rs73023348	1	154513418	G	A	0.79	0.994	6283.5	262.5	1.02E-119
rs73023349	1	154513440	T	C	0.79	0.994	6283.5	262.5	1.02E-119
rs72999415	1	154513712	C	T	0.79	0.994	6283.5	262.5	1.03E-119
rs72999419	1	154514092	T	C	0.79	0.994	6283.7	262.5	1.04E-119
rs41313910	1	154514203	C	G	0.79	0.994	6283.7	262.5	1.04E-119
rs72999422	1	154514942	G	A	0.79	0.994	6283.9	262.5	1.05E-119
rs58881140	1	154515193	C	T	0.79	0.994	6283.9	262.5	1.07E-119
rs6672087	1	154516122	A	C	0.79	0.994	6284.2	262.6	1.09E-119
rs6681207	1	154517217	T	C	0.79	0.993	6284.5	262.6	1.20E-119
rs112394421	1	154512008	T	C	0.79	0.995	6281.7	262.5	1.22E-119
rs3811449	1	154517504	G	A	0.79	0.993	6284.6	262.6	1.25E-119
rs41308419	1	154507197	T	C	0.79	0.996	6276.6	262.3	1.28E-119
rs60368585	1	154509340	T	C	0.79	0.995	6277.3	262.4	1.29E-119
rs111742980	1	154508764	T	A	0.79	0.995	6277.1	262.3	1.29E-119
rs73023339	1	154510852	C	T	0.79	0.995	6277.6	262.4	1.31E-119
rs112236096	1	154510790	C	T	0.79	0.995	6277.6	262.4	1.31E-119
rs59084843	1	154518286	A	G	0.79	0.993	6284.8	262.7	1.41E-119
rs61559765	1	154518349	T	C	0.79	0.993	6284.8	262.7	1.42E-119
rs60092776	1	154518496	A	G	0.79	0.992	6284.8	262.7	1.46E-119
rs60931472	1	154518642	T	C	0.79	0.992	6284.9	262.8	1.50E-119
rs112585956	1	154507767	T	C	0.79	0.996	6275.3	262.4	1.63E-119
rs4633282	1	154507899	T	C	0.79	0.996	6275.3	262.4	1.63E-119
rs73023331	1	154508210	G	A	0.79	0.996	6275.4	262.4	1.63E-119
rs4639752	1	154508974	T	A	0.79	0.996	6275.7	262.4	1.63E-119
rs112674677	1	154508792	C	T	0.79	0.996	6275.6	262.4	1.63E-119
rs113915076	1	154508812	G	A	0.79	0.996	6275.6	262.4	1.63E-119
rs7551873	1	154505857	C	T	0.79	0.995	6274.9	262.4	1.63E-119
rs58710275	1	154505010	C	T	0.79	0.996	6274.6	262.4	1.63E-119
rs111600849	1	154487489	G	A	0.79	0.997	6240.5	262.0	1.27E-118
rs11490956	1	154487258	G	T	0.79	0.997	6239.2	261.9	1.30E-118
rs41269915	1	154521584	T	A	0.79	0.986	6269.8	263.3	1.35E-118
rs77184252	1	154492533	A	G	0.79	0.997	6234.2	261.9	1.90E-118
rs4390168	1	154509096	A	G	0.79	0.992	6255.6	262.8	1.94E-118
rs12403537	1	154487726	G	A	0.79	0.997	6233.2	261.9	2.05E-118
rs73020246	1	154485640	A	G	0.79	0.997	6233.0	261.9	2.07E-118
rs73020234	1	154482767	C	A	0.79	0.997	6232.7	261.9	2.11E-118
rs73020232	1	154482669	C	T	0.79	0.997	6232.7	261.9	2.12E-118
rs59838898	1	154466301	C	T	0.79	0.997	6222.7	261.9	4.25E-118
rs73018293	1	154465577	C	T	0.79	0.997	6222.7	261.9	4.26E-118
rs12406822	1	154499484	C	T	0.79	0.983	6286.5	265.5	2.40E-117
rs59239860	1	154492107	G	T	0.80	0.994	6233.4	265.1	8.67E-116
rs6675472	1	154445503	T	C	0.22	0.972	-5986.3	264.6	8.64E-108
rs4291493	1	154436920	T	C	0.27	0.919	-5772.3	259.5	2.27E-104
rs35717427	1	154391882	G	A	0.86	0.685	-8971.6	429.1	5.53E-93
rs11579998	1	154580834	C	T	0.93	0.650	-11125.0	616.9	2.18E-70
rs11264224	1	154568086	A	C	0.80	0.802	-6930.6	384.4	2.25E-70
rs57569414	1	154380419	C	A	0.86	0.786	-7433.4	420.5	8.69E-68
rs3738030	1	154575315	A	C	0.87	0.965	5418.8	313.7	6.35E-65
rs7525477	1	154394297	G	A	0.53	0.708	5242.3	309.8	1.99E-62

rs10908421	1	154635885	G	T	0.71	0.912	-4865.2	297.1	1.05E-58
rs4601580	1	154394417	T	A	0.48	0.736	-5037.2	307.9	1.45E-58
rs11264237	1	154636276	G	A	0.71	0.955	-4654.5	293.6	3.40E-55
rs55873271	1	154610952	C	T	0.79	0.971	4327.6	273.6	5.49E-55
rs12132326	1	154376896	G	A	0.96	0.656	-11470.7	739.5	5.52E-53
rs59618219	1	154589931	C	T	0.79	0.995	4220.8	272.9	1.05E-52
rs41269913	1	154461480	C	T	0.97	0.612	-13738.6	888.3	1.07E-52
rs56020456	1	154591882	G	T	0.79	0.994	4216.1	272.9	1.38E-52
rs58655370	1	154580458	C	A	0.79	0.990	4207.2	272.6	1.76E-52
rs11264227	1	154596551	C	T	0.79	0.990	4209.9	273.2	2.44E-52
rs61811397	1	154606846	G	A	0.79	0.984	4217.1	273.9	3.01E-52
rs12739228	1	154426190	G	A	0.95	0.894	7822.9	509.1	4.68E-52
rs56019122	1	154596051	C	T	0.79	0.992	4198.0	273.3	5.28E-52
rs61811390	1	154592832	A	C	0.79	0.998	4187.5	273.2	8.44E-52
rs56064112	1	154602271	A	G	0.79	0.991	4197.7	274.0	9.06E-52
rs2335230	1	154589628	A	C	0.79	0.979	4207.0	275.2	1.53E-51
rs79438587	1	154342517	C	T	0.84	0.685	-6238.8	408.4	1.79E-51
rs12059682	1	154579585	T	C	0.79	0.995	4153.0	272.3	2.58E-51
rs59711031	1	154599500	G	C	0.79	0.993	4175.2	274.0	3.06E-51
rs3766926	1	154564417	T	C	0.79	0.994	4124.5	272.0	9.46E-51
rs17655362	1	154565914	G	A	0.79	0.995	4118.1	272.0	1.26E-50
rs9427106	1	154603322	T	C	0.54	0.973	-3974.1	265.4	1.39E-49
rs79778789	1	154409434	A	G	0.96	0.506	-14297.8	968.2	2.71E-48
rs9426831	1	154609657	A	T	0.54	0.983	-3891.2	263.8	3.30E-48
rs12403159	1	154421521	G	C	0.96	0.378	-19001.0	1289.3	4.15E-48
rs11264230	1	154608418	G	T	0.54	0.983	-3885.4	263.8	4.52E-48
rs66980031	1	154481153	C	G	0.96	0.868	7709.4	527.4	2.20E-47
rs113580743	1	154420333	G	A	0.95	0.640	8600.5	596.4	3.54E-46
rs9427109	1	154609009	A	G	0.56	0.932	-3904.4	273.2	2.14E-45
rs9427110	1	154616804	G	A	0.55	0.964	-3833.4	268.3	2.24E-45
rs9427092	1	154553722	T	C	0.80	0.937	4113.8	294.1	1.31E-43
rs9426829	1	154592201	T	C	0.51	0.988	-3630.9	259.9	1.68E-43
rs35504625	1	154599778	C	G	0.51	0.989	-3574.1	260.0	3.43E-42
rs7531982	1	154596122	A	T	0.51	0.992	-3568.6	259.7	3.46E-42
rs9426828	1	154591489	G	A	0.51	0.995	-3560.3	259.4	4.30E-42
rs9426827	1	154589965	T	C	0.51	0.995	-3560.0	259.3	4.34E-42
rs12125166	1	154582129	C	T	0.51	0.992	-3562.3	259.6	4.66E-42
rs9427102	1	154586957	A	G	0.51	0.994	-3556.4	259.3	5.11E-42
rs9427104	1	154589232	C	T	0.51	0.995	-3556.8	259.4	5.18E-42
rs9426826	1	154585480	C	G	0.51	0.995	-3555.7	259.3	5.34E-42
rs9427100	1	154583762	G	A	0.51	0.995	-3555.0	259.3	5.56E-42
rs61811388	1	154581231	G	C	0.51	0.994	-3553.2	259.4	6.41E-42
rs6702549	1	154583102	T	C	0.51	0.995	-3550.2	259.3	7.13E-42
rs6702449	1	154583210	A	C	0.51	0.995	-3550.2	259.4	7.18E-42
rs6699729	1	154582896	A	T	0.51	0.995	-3550.1	259.4	7.23E-42
rs11802588	1	154580895	C	G	0.51	0.993	-3550.1	259.5	8.18E-42
rs3766922	1	154577055	T	G	0.51	0.989	-3551.1	259.9	1.01E-41
rs6703672	1	154474557	T	C	0.20	0.902	-3891.2	287.0	4.08E-41
rs11580535	1	154364317	G	T	0.86	0.968	-5340.8	397.0	1.57E-40
rs72633650	1	154360838	T	C	0.86	0.969	-5339.5	396.9	1.59E-40
rs11264239	1	154653052	G	A	0.94	0.850	-8296.0	621.0	5.37E-40
	1	154353503	T	C	0.92	0.723	-7451.9	558.2	6.09E-40
	1	154430647	C	T	0.97	0.680	10346.0	782.2	3.00E-39
rs9427094	1	154557685	C	T	0.50	0.990	-3428.7	260.9	8.60E-39
rs10908419	1	154567699	G	A	0.50	0.995	-3394.3	260.6	3.88E-38
rs116710125	1	154502312	C	T	0.95	0.810	6463.0	497.2	5.35E-38
rs11264223	1	154565519	G	A	0.50	0.995	-3378.5	260.4	7.49E-38
rs77993403	1	154316996	G	A	0.96	0.697	-9378.5	729.3	3.12E-37

rs66654715	1	154376820	G	C	0.96	0.776	7326.8	571.7	5.31E-37
rs12122449	1	154283908	C	T	0.86	0.789	-5446.7	432.0	6.94E-36
rs9427103	1	154587817	A	G	0.80	0.980	3755.1	298.6	1.07E-35
rs11580217	1	154353499	T	C	0.91	0.750	-6712.8	538.3	3.76E-35
rs75456865	1	154478114	A	T	0.96	0.838	6376.5	511.8	4.28E-35
rs60517797	1	154642224	T	C	0.90	0.862	5033.9	409.8	3.55E-34
rs116059394	1	154483655	A	G	0.95	0.702	6876.6	568.4	3.31E-33
rs78042851	1	154627572	A	C	0.97	0.490	-12080.3	1011.0	1.90E-32
rs76289529	1	154516404	C	T	0.98	0.638	-13735.3	1153.8	3.17E-32
rs116306348	1	154653601	A	T	0.96	0.883	5926.5	498.1	3.38E-32
rs11265623	1	154454674	C	G	0.03	0.331	16001.3	1347.4	4.44E-32
rs77801962	1	154566932	T	C	0.97	0.489	-12434.9	1064.4	4.10E-31
rs4845626	1	154423485	G	T	0.84	0.985	3427.2	295.1	8.89E-31
rs114660934	1	154413904	G	A	0.96	0.774	6651.6	573.0	9.60E-31
	1	154509061	G	A	0.97	0.649	8678.9	750.3	1.52E-30
rs116141616	1	154416069	G	A	0.98	0.641	11083.8	958.5	1.60E-30
rs114879247	1	154434836	T	C	0.96	0.807	6463.7	560.6	2.34E-30
rs34450945	1	154289348	C	T	0.96	0.768	7042.2	611.4	2.66E-30
rs116323753	1	154586202	G	C	0.97	0.493	-12144.5	1056.2	3.34E-30
	1	154392851	G	A	0.98	0.605	-13591.7	1185.7	4.96E-30
rs112559935	1	154336175	G	T	0.96	0.599	8056.8	705.8	8.54E-30
rs6427720	1	154501756	G	A	0.17	0.980	-3346.6	293.4	9.57E-30
rs9427116	1	154631123	T	C	0.48	0.974	-2987.3	263.2	1.74E-29
rs10796931	1	154652119	G	A	0.48	0.944	-2994.4	264.1	2.00E-29
rs7554871	1	154650598	G	A	0.53	0.943	2991.1	264.3	2.55E-29
	1	154353507	T	C	0.95	0.625	-8057.0	712.9	2.98E-29
rs10908423	1	154644546	G	A	0.48	0.968	-2945.1	262.0	5.79E-29
rs9427117	1	154632231	C	G	0.48	0.983	-2935.8	261.6	7.30E-29
rs6426877	1	154651044	C	G	0.53	0.939	2972.0	264.9	7.64E-29
	1	154628059	G	A	0.97	0.492	-12081.3	1081.0	1.20E-28
rs9330261	1	154634998	T	A	0.48	0.986	-2911.4	260.7	1.33E-28
rs79925547	1	154443014	C	T	0.99	0.722	-11248.9	1009.8	1.79E-28
	1	154415848	C	G	0.98	0.753	8072.8	725.3	1.98E-28
rs10159236	1	154431405	C	A	0.84	0.983	3296.8	296.4	2.14E-28
rs12748485	1	154560471	C	T	0.96	0.748	7362.1	662.2	2.28E-28
rs6688149	1	154637946	C	G	0.47	0.987	-2898.1	260.9	2.51E-28
rs79480105	1	154445104	G	A	0.96	0.790	6367.1	573.4	2.66E-28
rs6703211	1	154637950	T	C	0.47	0.987	-2896.8	260.9	2.67E-28
rs61812599	1	154421554	G	A	0.84	0.988	3302.2	298.3	3.83E-28
rs111301013	1	154419584	C	A	0.84	0.988	3301.0	298.5	4.28E-28
rs7523890	1	154622234	T	C	0.84	0.969	3440.0	311.4	4.91E-28
	1	154353511	T	C	0.95	0.594	-8400.1	760.7	5.17E-28
rs55800510	1	154417187	C	T	0.84	0.976	3293.9	298.4	5.36E-28
rs4308966	1	154432622	T	C	0.84	0.994	3262.9	295.6	5.38E-28
rs4845374	1	154426947	T	A	0.84	0.992	3281.1	297.3	5.60E-28
rs11804305	1	154424497	C	T	0.84	0.992	3279.4	297.8	7.18E-28
rs11265634	1	154522248	A	G	0.17	0.984	-3229.3	293.3	7.40E-28
rs80033405	1	154416572	A	C	0.99	0.701	-11265.3	1024.0	8.06E-28
rs60682501	1	154351304	A	G	0.86	0.972	-4378.7	398.7	9.76E-28
rs116088025	1	154354350	C	T	0.86	0.972	-4381.2	399.1	1.04E-27
rs11579718	1	154353358	A	G	0.86	0.972	-4371.2	398.9	1.29E-27
rs7547072	1	154639255	C	T	0.48	0.950	-2922.1	266.7	1.30E-27
rs61812596	1	154414691	C	T	0.84	0.976	3265.0	298.2	1.43E-27
rs1889313	1	154351717	C	A	0.86	0.972	-4363.5	398.8	1.54E-27
rs60412881	1	154351323	T	C	0.86	0.972	-4362.3	398.8	1.58E-27
rs60767732	1	154351277	G	T	0.86	0.972	-4362.2	398.8	1.58E-27
rs57100877	1	154351247	C	G	0.86	0.972	-4362.0	398.8	1.59E-27
rs9803896	1	154347450	G	A	0.86	0.971	-4331.1	396.3	1.78E-27

rs11582433	1	154349605	C	T	0.86	0.972	-4356.2	398.7	1.82E-27
rs61812626	1	154436436	G	A	0.84	0.984	3245.8	297.1	1.86E-27
rs11264233	1	154625632	G	A	0.49	0.963	-2934.9	268.7	1.89E-27
rs59356432	1	154349159	A	G	0.86	0.972	-4354.6	398.7	1.89E-27
rs56264602	1	154349121	G	A	0.86	0.972	-4354.5	398.7	1.90E-27
rs9803950	1	154348513	G	C	0.86	0.972	-4347.0	398.1	1.94E-27
rs4345796	1	154436750	C	T	0.84	0.985	3227.5	296.5	2.82E-27
rs79753070	1	154652279	G	A	0.98	0.600	-12335.1	1133.9	3.02E-27
rs9427113	1	154626477	A	G	0.49	0.980	-2898.8	266.7	3.30E-27
rs41310893	1	154514267	C	T	0.96	0.781	6286.9	579.5	4.20E-27
rs61812594	1	154411628	A	G	0.84	0.974	3228.9	298.8	6.63E-27
rs61812593	1	154410490	T	C	0.84	0.973	3227.5	299.0	7.49E-27
rs61812592	1	154410482	G	A	0.84	0.973	3227.4	299.0	7.52E-27
rs61811421	1	154653951	C	T	0.78	0.866	3255.4	302.4	1.03E-26
rs12024175	1	154643443	G	A	0.48	0.958	-2847.7	264.6	1.03E-26
rs113325045	1	154413242	G	A	0.97	0.662	7827.7	727.8	1.12E-26
rs9427115	1	154627852	G	A	0.49	0.964	-2879.3	268.1	1.35E-26
rs79219014	1	154415675	G	T	0.97	0.566	-9849.7	917.5	1.39E-26
rs57502626	1	154408916	T	A	0.84	0.973	3214.6	299.6	1.49E-26
rs55826755	1	154409520	C	G	0.84	0.973	3211.7	299.5	1.55E-26
	1	154345962	C	A	0.86	0.973	-4259.7	399.3	2.79E-26
rs56047170	1	154528053	G	A	0.84	0.972	3162.4	298.2	5.40E-26
rs6689710	1	154470356	T	C	0.16	0.998	-3132.4	296.0	6.82E-26
rs56807273	1	154340230	C	T	0.86	0.974	-4222.4	399.2	7.23E-26
rs6662503	1	154477069	T	A	0.16	0.999	-3126.6	296.2	9.01E-26
rs6696177	1	154477263	G	A	0.16	0.999	-3126.9	296.2	9.06E-26
rs12124333	1	154478292	T	A	0.16	0.999	-3126.3	296.2	9.34E-26
rs7526247	1	154485039	T	C	0.16	0.999	-3128.0	296.4	9.42E-26
rs4288587	1	154492357	A	T	0.16	1.000	-3125.7	296.3	9.85E-26
rs4518898	1	154492432	T	C	0.16	1.000	-3125.7	296.3	9.86E-26
rs6687112	1	154497146	C	T	0.16	0.998	-3129.1	296.7	1.00E-25
rs6427716	1	154497329	C	A	0.16	0.998	-3125.1	296.6	1.10E-25
rs4845640	1	154497802	C	T	0.16	0.998	-3124.6	296.6	1.15E-25
rs4845641	1	154497930	C	G	0.16	0.998	-3124.6	296.6	1.15E-25
rs4845643	1	154498239	T	C	0.16	0.998	-3124.6	296.6	1.16E-25
rs6697103	1	154499926	C	T	0.16	0.998	-3124.6	296.6	1.16E-25
rs6697115	1	154499947	C	T	0.16	0.998	-3124.6	296.6	1.16E-25
rs6656395	1	154500016	T	C	0.16	0.998	-3124.6	296.6	1.16E-25
rs4131514	1	154500089	A	G	0.16	0.998	-3124.6	296.6	1.16E-25
rs4585969	1	154498957	C	T	0.16	0.998	-3124.5	296.6	1.16E-25
rs10908844	1	154500857	G	A	0.16	0.998	-3124.5	296.6	1.16E-25
rs4307560	1	154499345	G	A	0.16	0.998	-3124.5	296.6	1.16E-25
rs6701586	1	154499611	A	T	0.16	0.998	-3124.5	296.6	1.16E-25
rs12129484	1	154501204	C	A	0.16	0.998	-3124.5	296.6	1.16E-25
rs6427722	1	154502123	G	C	0.16	0.998	-3124.5	296.6	1.16E-25
rs6427724	1	154502499	T	A	0.16	0.998	-3124.5	296.6	1.16E-25
rs10752643	1	154503601	G	A	0.16	0.998	-3124.4	296.6	1.16E-25
rs4269769	1	154494082	G	C	0.16	0.999	-3124.1	296.6	1.17E-25
rs6670375	1	154494947	G	A	0.16	0.999	-3124.1	296.6	1.17E-25
rs4575077	1	154495011	C	A	0.16	0.999	-3124.1	296.6	1.17E-25
rs11265629	1	154498574	G	A	0.16	0.998	-3123.4	296.6	1.20E-25
rs10908846	1	154501364	C	A	0.16	0.998	-3123.4	296.6	1.20E-25
rs6691727	1	154497005	A	G	0.16	0.999	-3123.2	296.7	1.23E-25
rs11265627	1	154496286	C	A	0.16	0.998	-3122.2	296.6	1.26E-25
rs111885536	1	154461260	G	A	0.84	0.994	3126.2	297.1	1.28E-25
rs113624284	1	154339299	A	G	0.86	0.975	-4196.2	398.8	1.29E-25
rs14021	1	154452204	T	C	0.84	0.998	3116.5	296.2	1.30E-25
rs7550664	1	154528709	A	C	0.84	0.973	3139.7	298.5	1.36E-25

rs16836054	1	154462195	G	A	0.84	0.999	3113.4	296.1	1.38E-25
rs76518735	1	154454309	A	C	0.98	0.670	8211.9	781.0	1.40E-25
	1	154455249	T	C	0.98	0.670	8211.8	781.0	1.41E-25
rs12118074	1	154477585	G	A	0.16	0.999	-3113.9	296.2	1.42E-25
rs4845634	1	154464160	G	A	0.84	0.999	3113.0	296.1	1.43E-25
rs61812631	1	154467816	A	T	0.84	0.999	3112.8	296.1	1.44E-25
rs56233546	1	154471371	A	G	0.84	1.000	3112.7	296.2	1.46E-25
rs7513603	1	154481158	T	C	0.16	0.999	-3111.0	296.2	1.58E-25
rs4292956	1	154548946	C	T	0.93	0.945	4751.9	453.5	2.06E-25
rs72698167	1	154480157	G	A	0.84	1.000	3104.2	296.3	2.11E-25
rs45490696	1	154548992	C	T	0.93	0.940	4782.8	456.6	2.14E-25
rs6427729	1	154518874	G	A	0.16	0.995	-3109.1	296.9	2.21E-25
rs9427415	1	154508437	A	T	0.84	0.995	3107.0	297.0	2.42E-25
rs4845644	1	154501576	A	G	0.84	0.998	3103.5	296.7	2.46E-25
rs61812654	1	154502517	G	A	0.84	0.998	3103.5	296.7	2.46E-25
rs9427076	1	154503022	C	T	0.84	0.998	3103.0	296.7	2.50E-25
rs12135008	1	154384293	G	A	0.96	0.612	-8308.8	794.5	2.52E-25
rs4845648	1	154514338	A	T	0.84	0.997	3097.3	296.9	3.24E-25
rs61811362	1	154512495	G	A	0.84	0.997	3097.0	296.8	3.26E-25
rs12048091	1	154395077	A	G	0.84	0.910	3243.4	312.1	4.88E-25
rs67860750	1	154538917	G	C	0.85	0.712	-4805.5	463.1	5.78E-25
rs7553602	1	154529421	A	T	0.84	0.982	3080.5	297.0	6.10E-25
rs6680410	1	154550724	C	T	0.93	0.952	4679.1	451.6	6.88E-25
rs112095222	1	154627851	C	T	0.89	0.921	3981.7	384.6	7.41E-25
	1	154450510	A	C	0.98	0.669	8119.7	784.4	7.53E-25
rs10752605	1	154627213	A	G	0.11	0.926	-3953.9	382.3	8.20E-25
rs9427082	1	154523562	G	T	0.84	0.990	3068.3	297.2	1.01E-24
rs7539745	1	154526414	C	G	0.16	0.988	-3063.6	297.2	1.16E-24
rs41302545	1	154525999	C	T	0.84	0.989	3063.4	297.3	1.23E-24
	1	154450365	T	A	0.99	0.686	9532.3	927.5	1.59E-24
rs6664039	1	154337238	A	G	0.86	0.985	-4075.6	397.0	1.78E-24
rs72633647	1	154335609	T	A	0.86	0.988	-4050.4	396.5	2.96E-24
rs76093405	1	154411768	G	A	0.98	0.745	8061.6	790.8	3.72E-24
rs72633646	1	154334683	G	A	0.86	0.985	-4020.4	396.2	5.93E-24
rs77710113	1	154618623	T	C	0.94	0.859	5241.5	516.6	6.02E-24
	1	154412524	G	A	0.98	0.743	8065.8	795.9	6.74E-24
rs12023358	1	154399649	C	T	0.84	0.966	3098.0	305.9	7.34E-24
	1	154394484	G	A	0.98	0.607	-10548.2	1048.0	1.36E-23
rs11590203	1	154333569	G	T	0.86	0.981	-3989.9	396.9	1.53E-23
rs12048950	1	154394966	T	C	0.84	0.914	3138.0	314.1	2.84E-23
	1	154419517	T	A	0.98	0.575	-9871.0	989.2	3.19E-23
rs73001418	1	154585594	T	A	0.93	0.961	4507.2	452.2	3.60E-23
rs73001420	1	154586759	C	G	0.93	0.961	4506.6	452.3	3.67E-23
rs6686467	1	154329095	G	A	0.86	0.985	-3930.0	394.7	3.95E-23
rs72633645	1	154328434	C	T	0.86	0.985	-3931.5	394.9	3.97E-23
rs6698881	1	154332672	T	A	0.86	0.989	-3910.6	392.8	4.01E-23
rs61300392	1	154575685	A	T	0.93	0.973	4492.9	452.2	4.86E-23
rs12404936	1	154421530	T	C	0.98	0.337	-23260.1	2342.0	5.04E-23
rs72999485	1	154559360	G	C	0.93	0.972	4479.0	451.9	6.15E-23
rs78062588	1	154566225	T	C	0.94	0.864	5136.9	518.7	6.63E-23
rs115800464	1	154569895	G	C	0.94	0.864	5136.8	518.7	6.63E-23
rs12047973	1	154394766	A	G	0.84	0.912	3108.1	315.0	9.67E-23
rs113936582	1	154327737	C	G	0.86	0.973	-3925.4	400.2	1.67E-22
rs116805289	1	154510155	A	C	0.98	0.755	7740.3	790.7	2.03E-22
rs45478197	1	154422733	C	T	0.91	0.928	3634.5	371.4	2.07E-22
rs115617158	1	154405495	G	A	0.98	0.744	8056.2	827.2	3.30E-22
rs11576151	1	154327420	C	T	0.86	0.980	-3875.1	400.2	5.70E-22
rs116247632	1	154476522	C	T	0.98	0.612	-8942.5	926.9	7.93E-22

	1	154519249	G	A	0.98	0.612	-8934.8	927.2	8.85E-22
rs79794939	1	154390932	C	T	0.93	0.745	4987.8	524.2	2.79E-21
	1	154301261	C	A	0.86	0.903	-3945.9	415.6	3.42E-21
rs34693607	1	154661369	C	G	0.77	0.760	-3261.0	343.6	3.51E-21
rs78739139	1	154675520	G	A	0.95	0.519	-7919.4	835.7	4.03E-21
rs11264247	1	154685607	G	A	0.95	0.909	-5966.8	631.7	5.40E-21
rs4845645	1	154502909	T	A	0.18	0.922	-2828.3	301.2	8.99E-21
rs35221765	1	154345344	G	A	0.74	0.977	2713.8	289.8	1.14E-20
rs34055426	1	154345519	G	A	0.74	0.976	2714.0	290.2	1.27E-20
rs4521987	1	154388668	T	C	0.14	0.834	-3542.7	379.0	1.34E-20
rs56100876	1	154496473	G	A	0.98	0.520	-11811.6	1264.0	1.37E-20
rs4845653	1	154551368	G	A	0.85	0.887	3153.8	337.6	1.41E-20
rs75631461	1	154498153	T	C	0.98	0.520	-11798.9	1263.8	1.49E-20
	1	154608755	C	T	0.97	0.740	6976.4	747.7	1.58E-20
rs4307559	1	154499123	G	C	0.17	0.953	-2865.9	307.6	1.79E-20
rs115880387	1	154585644	A	G	0.99	0.549	-14698.5	1579.7	2.00E-20
rs114292408	1	153641058	T	C	0.99	0.607	11437.5	1230.1	2.12E-20
rs113721028	1	154293096	G	A	0.95	0.634	5890.0	634.0	2.27E-20
rs111923597	1	154293099	G	C	0.95	0.634	5889.9	634.0	2.28E-20
rs4556348	1	154394296	C	T	0.87	0.761	3464.5	376.3	4.92E-20
	1	154534480	C	T	0.98	0.710	7860.8	853.9	4.96E-20
rs11264245	1	154682776	C	T	0.95	0.956	-5776.0	627.9	5.25E-20
rs56048555	1	154630668	C	T	0.89	0.921	3608.6	393.0	6.15E-20
rs17699328	1	154322473	A	G	0.88	0.992	-3883.9	423.3	6.61E-20
rs12134719	1	154687229	A	C	0.95	0.951	-5769.8	630.1	7.76E-20
rs35013837	1	154480123	C	T	0.85	0.761	-4117.4	450.2	8.61E-20
rs2010828	1	154314225	C	T	0.88	0.997	-3869.9	424.0	1.01E-19
rs56258967	1	154334023	C	T	0.99	0.434	-15877.4	1739.8	1.03E-19
rs6427575	1	154305560	C	T	0.88	0.995	-3867.3	424.5	1.18E-19
rs6691345	1	154302826	C	T	0.88	0.992	-3859.1	424.5	1.41E-19
rs115513320	1	154668908	G	T	0.99	0.700	-9373.6	1032.6	1.59E-19
rs2297607	1	154320942	A	G	0.75	0.952	2644.1	291.5	1.67E-19
rs58892549	1	154688488	T	C	0.95	0.929	-5715.9	630.3	1.73E-19
rs11265539	1	154290852	C	G	0.88	0.970	-3864.9	428.3	2.57E-19
rs6427568	1	154290535	G	A	0.88	0.968	-3864.3	429.3	3.17E-19
rs41310887	1	154321623	G	A	0.95	0.591	6276.5	697.9	3.38E-19
rs1194583	1	154320196	T	C	0.12	0.965	3794.1	422.7	4.00E-19
rs6673495	1	154289966	G	A	0.88	0.963	-3860.9	431.1	4.69E-19
rs6687971	1	154334253	G	C	0.74	0.971	2603.7	291.9	6.44E-19
rs6687939	1	154334176	G	A	0.74	0.971	2602.3	291.9	6.79E-19
rs3103309	1	154270877	C	T	0.65	0.777	-2870.8	324.4	1.22E-18
rs4845657	1	154630335	T	C	0.81	0.967	2650.5	304.3	4.08E-18
	1	154206013	C	A	0.97	0.632	-8870.2	1028.9	8.91E-18
rs72696248	1	154212541	G	A	0.97	0.632	-8869.7	1029.0	8.99E-18
rs6426875	1	154650870	T	C	0.03	0.767	-6098.6	709.1	1.06E-17
rs55675282	1	154167180	A	C	0.97	0.631	-8796.2	1025.8	1.32E-17
rs1931299	1	154283831	G	T	0.74	0.935	2525.6	295.2	1.55E-17
rs1931298	1	154283896	C	T	0.88	0.933	-3731.5	440.9	3.39E-17
rs733228	1	154639618	T	C	0.81	0.969	2590.7	306.2	3.52E-17
rs2335251	1	154634568	T	C	0.04	0.329	10199.1	1207.3	3.88E-17
rs11578646	1	154282917	C	A	0.88	0.930	-3726.1	441.1	3.90E-17
rs115306743	1	154463784	A	C	0.98	0.721	7285.6	862.7	3.98E-17
	1	154431123	T	A	0.93	0.824	3598.6	426.8	4.46E-17
rs11588248	1	154390575	A	G	0.83	0.782	3155.6	375.2	5.29E-17
	1	154454553	A	G	0.98	0.369	-17337.6	2065.3	6.08E-17
rs116037345	1	154404454	C	T	0.97	0.890	4985.5	597.5	9.29E-17
rs56120520	1	154650381	C	T	0.81	0.952	2586.0	310.3	1.02E-16
rs61811419	1	154643702	A	G	0.81	0.972	2556.6	306.8	1.02E-16

rs11800830	1	154274638	A	G	0.88	0.922	-3693.2	443.6	1.08E-16
rs56128567	1	154274672	C	T	0.88	0.919	-3687.3	443.9	1.27E-16
	1	154428309	G	T	0.99	0.710	7981.3	964.2	1.61E-16
rs111829145	1	154545240	C	T	0.99	0.682	7791.9	941.4	1.63E-16
rs72698115	1	154379369	A	C	0.89	0.935	3301.1	400.1	2.02E-16
rs35387092	1	154288680	T	C	0.75	0.956	2433.7	295.4	2.24E-16
rs56293184	1	154288351	C	T	0.75	0.955	2431.6	295.6	2.45E-16
rs6427561	1	154287733	T	A	0.75	0.953	2426.0	295.8	3.02E-16
rs4845658	1	154643382	G	A	0.82	0.944	2590.1	315.8	3.03E-16
rs6427560	1	154287687	A	G	0.75	0.953	2425.7	295.8	3.05E-16
rs34569656	1	154670136	A	G	0.94	0.787	4011.2	489.4	3.14E-16
rs35800693	1	154284894	A	G	0.75	0.951	2418.3	295.6	3.56E-16
rs3916566	1	154282146	C	T	0.75	0.943	2430.4	297.3	3.74E-16
rs56367957	1	154283217	A	G	0.75	0.950	2415.9	295.6	3.83E-16
rs7547298	1	154281680	G	A	0.75	0.950	2414.4	295.7	4.11E-16
rs115870735	1	154472809	T	C	0.99	0.711	8094.4	992.5	4.39E-16
rs35362987	1	154279756	G	A	0.75	0.948	2411.1	296.2	5.03E-16
rs61578738	1	154278943	T	C	0.75	0.947	2411.3	296.4	5.21E-16
rs12753821	1	154628799	C	T	0.94	0.845	3850.7	473.4	5.23E-16
rs34640592	1	154276010	G	A	0.75	0.944	2416.2	297.1	5.29E-16
rs2297606	1	154321084	G	T	0.72	0.927	2356.5	290.3	5.98E-16
rs73026617	1	154369981	C	T	0.89	0.991	3126.4	387.2	8.45E-16
rs113057497	1	154367754	C	T	0.89	0.993	3119.7	386.5	8.68E-16
rs111810442	1	154369683	C	G	0.89	0.991	3124.2	387.1	8.69E-16
	1	154416805	C	T	0.99	0.533	9284.7	1152.9	1.01E-15
rs12033701	1	154365886	C	T	0.89	0.994	3113.9	386.9	1.04E-15
rs7547455	1	154371753	T	C	0.73	0.929	2362.1	293.7	1.10E-15
rs9651053	1	154359411	G	A	0.89	0.996	3108.9	386.6	1.11E-15
rs72696301	1	154355501	G	C	0.89	0.994	3105.0	386.2	1.12E-15
rs9651036	1	154360492	C	A	0.89	0.995	3108.9	386.7	1.12E-15
rs9651055	1	154360684	A	G	0.89	0.995	3108.9	386.7	1.12E-15
rs111920902	1	154361226	C	T	0.89	0.995	3108.8	386.7	1.12E-15
rs12026876	1	154361788	G	A	0.89	0.995	3108.8	386.7	1.13E-15
	1	154363137	A	T	0.89	0.995	3108.0	386.7	1.15E-15
	1	154650925	C	T	0.82	0.926	2558.0	318.6	1.23E-15
	1	154431128	G	A	0.93	0.827	3565.4	445.0	1.41E-15
rs2135694	1	154651725	C	T	0.82	0.924	2553.3	318.8	1.45E-15
rs35735180	1	154325411	G	T	0.82	0.967	2605.9	325.6	1.51E-15
rs2335254	1	154652093	G	T	0.82	0.923	2554.1	319.2	1.52E-15
	1	154367898	C	T	0.89	0.980	3124.9	391.5	1.80E-15
rs77302258	1	154396203	A	G	0.91	0.898	3273.7	411.0	2.03E-15
rs112203594	1	154553430	C	A	0.98	0.643	7310.4	918.1	2.08E-15
	1	154335324	C	T	0.98	0.705	-9698.4	1227.3	3.37E-15
rs12075836	1	154371487	C	T	0.86	0.927	2881.1	365.8	4.15E-15
	1	154331320	C	T	0.98	0.700	-9663.9	1228.2	4.40E-15
rs111908494	1	154368095	A	G	0.90	0.963	3167.3	402.8	4.62E-15
rs61386354	1	154685551	G	A	0.97	0.458	8214.5	1045.6	4.84E-15
	1	154328261	C	T	0.98	0.699	-9658.2	1230.2	5.06E-15
rs113639721	1	154368098	T	C	0.89	0.978	3106.3	396.9	6.09E-15
	1	154532599	A	T	0.89	0.765	3126.8	401.1	7.85E-15
rs6427721	1	154501770	A	G	0.18	0.907	-2394.2	307.7	8.77E-15
	1	154536197	A	G	0.98	0.642	7580.3	977.5	1.07E-14
rs115697580	1	154517689	G	A	0.98	0.780	5697.6	736.3	1.22E-14
	1	154368224	C	T	0.89	0.971	3067.3	396.8	1.30E-14
	1	154368223	G	A	0.97	0.667	-6262.1	810.3	1.32E-14
rs78038982	1	154539156	G	T	0.98	0.642	7538.7	976.1	1.37E-14
	1	154479486	G	A	0.98	0.644	7513.1	973.2	1.41E-14
	1	154499328	G	A	0.98	0.657	-9007.7	1171.6	1.80E-14

	1	154469825	T	C	0.98	0.657	-9009.7	1172.3	1.83E-14
rs6688376	1	154326366	C	T	0.70	0.959	2211.3	287.9	1.91E-14
	1	154645937	T	C	0.98	0.702	7658.0	999.3	2.18E-14
rs41304075	1	154507190	T	C	0.98	0.666	-8979.1	1175.9	2.69E-14
	1	154418200	C	T	0.99	0.459	11666.9	1532.0	3.14E-14
	1	154333027	G	T	0.98	0.877	6145.5	812.7	4.74E-14
rs1626035	1	154311144	C	T	0.27	0.999	-2121.3	280.7	4.89E-14
rs1352333	1	154310677	C	A	0.27	0.999	-2121.0	280.7	4.94E-14
rs11809740	1	154319689	G	A	0.73	0.998	2123.0	281.0	5.03E-14
rs2481064	1	154309474	G	A	0.27	0.999	-2119.4	280.7	5.18E-14
rs2483711	1	154309444	T	A	0.27	0.999	-2119.3	280.7	5.19E-14
rs111237020	1	153842032	C	T	0.96	0.553	-6446.3	854.6	5.45E-14
rs1760796	1	154307664	A	G	0.27	0.999	-2117.6	280.7	5.45E-14
rs2274988	1	154316434	C	T	0.73	0.993	2121.8	281.4	5.54E-14
rs2988721	1	154308641	A	C	0.27	0.997	-2118.6	280.9	5.55E-14
rs2483710	1	154306180	G	A	0.27	0.997	-2118.0	280.9	5.58E-14
	1	154637113	G	A	0.98	0.660	-8326.1	1104.9	5.78E-14
rs115178836	1	154011624	C	A	0.99	0.538	8417.0	1117.4	5.90E-14
rs35463210	1	154313764	C	A	0.73	0.999	2116.0	281.0	6.00E-14
rs12131108	1	154657905	C	A	0.60	0.868	2058.6	274.3	7.28E-14
rs115200978	1	154374850	C	T	0.98	0.608	-10185.2	1360.2	8.29E-14
	1	154510957	T	C	0.98	0.365	-14964.3	2014.4	1.29E-13
rs115224285	1	154411473	C	T	0.98	0.522	9485.5	1278.2	1.37E-13
rs28654266	1	153942593	C	T	0.98	0.573	6918.5	938.0	1.92E-13
rs77490383	1	153899248	G	T	0.98	0.686	6495.3	884.8	2.48E-13
	1	154307776	C	T	0.98	0.880	6104.0	831.9	2.54E-13
rs11802757	1	154658849	G	C	0.60	0.871	1996.6	273.3	3.20E-13
	1	154433897	G	A	0.97	0.799	5351.8	733.3	3.40E-13
rs12090237	1	154389741	G	A	0.99	0.492	-10025.5	1374.2	3.46E-13
rs6426880	1	154655042	G	A	0.10	0.922	3381.4	464.2	3.76E-13
rs3887104	1	154376671	C	T	0.83	0.857	2629.8	361.2	3.87E-13
rs12138773	1	153843489	C	A	0.97	0.751	-6063.6	836.5	4.87E-13
rs72698126	1	154388686	A	G	0.99	0.495	-9938.7	1373.1	5.27E-13
rs1027809	1	154656371	C	G	0.10	0.915	3371.2	466.2	5.53E-13
	1	153870483	A	C	0.97	0.747	-6093.1	842.7	5.56E-13
rs72698128	1	154391819	G	A	0.99	0.510	-10025.0	1386.5	5.57E-13
rs114492020	1	154044615	A	G	0.98	0.676	5749.4	796.0	5.90E-13
rs6701860	1	154653632	A	G	0.10	0.924	3341.2	465.6	8.24E-13
rs72698123	1	154387025	T	C	0.99	0.498	-9793.7	1369.9	1.00E-12
rs11265607	1	154357678	A	G	0.71	0.997	-2094.9	293.2	1.04E-12
rs6427631	1	154370020	T	C	0.71	0.984	-2104.6	294.7	1.06E-12
rs6670683	1	154623950	A	G	0.04	0.309	8156.4	1147.9	1.38E-12
rs2481065	1	154311911	G	A	0.12	0.998	-2651.2	373.3	1.40E-12
rs1209264	1	154305010	A	G	0.12	0.997	-2649.6	373.3	1.45E-12
rs11265609	1	154364328	C	T	0.71	0.998	-2078.8	292.9	1.47E-12
rs9724691	1	154365235	T	A	0.71	0.998	-2075.2	292.9	1.61E-12
rs10908426	1	154657909	C	T	0.10	0.929	3290.6	466.3	1.95E-12
rs55886061	1	154383261	C	T	0.99	0.502	-9610.4	1364.6	2.16E-12
	1	154320191	C	T	0.88	0.997	2632.1	374.0	2.22E-12
rs1205591	1	154298374	T	G	0.12	0.992	-2631.9	374.0	2.24E-12
rs7411976	1	154382443	A	C	0.99	0.502	-9593.6	1364.1	2.31E-12
rs114821176	1	154004498	G	A	0.98	0.833	5901.4	839.8	2.41E-12
rs9662562	1	154364762	A	T	0.71	0.996	-2059.1	293.2	2.50E-12
rs61806853	1	154154587	T	C	0.95	0.786	-4601.0	655.6	2.56E-12
rs72698116	1	154380934	A	G	0.99	0.502	-9558.4	1362.7	2.64E-12
	1	154362445	C	T	0.72	0.982	-2079.3	296.8	2.77E-12
rs1395566	1	154658966	A	G	0.10	0.930	3259.9	465.9	2.96E-12
rs72698118	1	154381781	T	C	0.99	0.498	-9521.7	1360.8	2.98E-12

rs6672010	1	154292161	C	T	0.70	0.992	1989.4	285.1	3.40E-12
rs12727865	1	154303705	C	T	0.70	0.999	1984.4	284.8	3.64E-12
rs58548028	1	154347784	G	A	0.89	0.965	2751.1	394.9	3.69E-12
rs12077870	1	154305190	G	C	0.70	0.998	1982.2	284.8	3.86E-12
rs6702754	1	154303976	T	C	0.70	0.997	1981.0	284.8	4.01E-12
rs10737169	1	154653704	T	C	0.10	0.912	3286.7	472.6	4.02E-12
rs6685323	1	154295592	C	T	0.70	0.998	1975.1	284.3	4.21E-12
rs116828691	1	154080990	T	C	0.97	0.748	-6608.1	951.2	4.22E-12
rs3890153	1	154295089	C	G	0.70	0.999	1972.3	284.3	4.51E-12
rs12748146	1	154292982	T	A	0.70	0.998	1968.5	284.2	4.86E-12
rs6668968	1	154293675	G	A	0.70	0.998	1968.5	284.2	4.89E-12
	1	154075227	C	A	0.97	0.784	4924.3	711.0	4.92E-12
rs35950207	1	154291958	C	T	0.70	0.999	1965.7	283.9	4.94E-12
	1	154303783	G	A	0.98	0.846	5234.6	760.2	6.49E-12
	1	154445362	G	T	0.99	0.591	8179.2	1188.9	6.78E-12
	1	154475027	C	T	0.99	0.401	-10883.0	1595.5	1.02E-11
rs115188583	1	153849029	T	C	0.99	0.657	6958.0	1027.1	1.40E-11
rs11265606	1	154356459	C	T	0.71	0.985	-1986.1	293.5	1.48E-11
	1	154064138	G	A	0.89	0.857	2765.1	409.2	1.57E-11
rs2633438	1	154308508	G	A	0.28	0.982	-1890.1	280.2	1.70E-11
	1	154252188	A	G	0.99	0.732	-10114.6	1503.4	1.92E-11
	1	153389207	G	A	0.98	0.440	7330.4	1089.7	1.94E-11
rs114980362	1	154209281	C	T	0.98	0.864	5345.8	795.8	2.06E-11
rs114779784	1	154239549	G	C	0.98	0.695	5514.2	822.0	2.20E-11
rs6676680	1	154662203	G	A	0.60	0.828	1890.3	282.8	2.59E-11
rs36046176	1	154355875	A	G	0.92	0.867	3240.3	488.5	3.64E-11
	1	153962327	C	T	0.99	0.396	-13519.7	2048.6	4.57E-11
rs2230324	1	153954646	C	T	0.96	0.833	-4408.1	673.1	6.42E-11
rs41264642	1	153392663	G	A	0.98	0.440	7029.7	1075.9	7.08E-11
rs76401668	1	154610358	G	A	0.98	0.656	6181.9	949.6	8.26E-11
	1	153768598	G	A	0.98	0.869	5170.5	798.0	1.02E-10
rs12130660	1	154088258	T	C	0.95	0.812	-4282.6	662.3	1.11E-10
rs12135246	1	154618028	A	G	0.97	0.336	8098.6	1253.6	1.15E-10
rs115165870	1	154303228	G	A	0.97	0.854	4319.7	671.4	1.37E-10
rs10908840	1	154459477	C	T	0.52	0.953	-1763.2	274.1	1.39E-10
rs116039813	1	154185514	G	T	0.97	0.727	4905.2	764.0	1.49E-10
rs114639817	1	154186137	G	T	0.97	0.727	4904.8	764.0	1.50E-10
rs114772434	1	154182928	C	G	0.97	0.727	4905.2	764.1	1.50E-10
rs114003180	1	153820352	T	C	0.98	0.863	5221.9	814.3	1.57E-10
	1	154317077	G	A	0.99	0.649	6392.2	1003.5	2.07E-10
rs10797084	1	154264736	C	T	0.69	0.866	-1929.6	303.7	2.28E-10
	1	154011426	A	C	0.99	0.849	6189.6	975.2	2.40E-10
	1	154697624	C	G	0.97	0.495	5532.7	876.2	2.95E-10
	1	153985235	A	C	0.99	0.862	6119.5	971.6	3.28E-10
	1	153962470	A	G	0.99	0.865	6101.4	971.0	3.60E-10
rs114166921	1	154211108	G	A	0.99	0.720	-9853.8	1568.3	3.62E-10
	5	61040597	C	T	0.99	0.613	6938.6	1107.0	3.98E-10
rs12069266	1	154664175	G	T	0.60	0.858	1730.4	276.9	4.51E-10
rs75940909	1	153932075	G	A	0.99	0.875	6053.7	969.6	4.65E-10
rs6427641	1	154380486	G	A	0.45	0.850	1775.3	284.8	4.99E-10
	1	154072601	G	A	0.97	0.785	4279.2	686.7	5.01E-10
rs10908804	1	154262857	A	C	0.69	0.885	-1860.4	299.3	5.58E-10
rs111412860	1	154462988	T	A	0.96	0.331	-7693.3	1242.7	6.48E-10
	1	153885456	C	T	0.99	0.880	5970.4	967.0	7.20E-10
	1	153872026	G	A	0.99	0.869	6358.0	1035.4	8.89E-10
	1	154100908	C	T	0.99	0.770	6655.5	1089.2	1.07E-09
	1	154301378	T	C	0.88	0.490	3115.6	510.6	1.13E-09
rs78740769	1	153789150	C	A	0.96	0.539	-5519.3	908.2	1.32E-09

	1	153874665	G	A	0.99	0.868	6080.7	1002.8	1.43E-09
rs1979575	1	154318172	A	G	0.85	0.999	2160.6	357.9	1.68E-09
rs11580594	1	154291386	C	T	0.85	0.986	2167.2	359.4	1.76E-09
rs76789167	1	153805095	T	C	0.99	0.853	5890.5	977.7	1.82E-09
rs11580613	1	154291559	C	T	0.85	0.991	2154.2	359.3	2.17E-09
rs9660850	1	154317281	G	A	0.85	0.991	2141.2	358.4	2.48E-09
	1	153872788	A	C	0.99	0.741	-8821.5	1483.9	2.96E-09
rs116612897	1	154663267	G	A	0.98	0.683	5595.9	943.0	3.17E-09
	1	154006203	T	C	0.98	0.743	4767.5	808.0	3.87E-09
rs61812595	1	154411852	G	A	0.94	0.843	2886.9	492.5	4.88E-09
rs7547947	1	154651751	A	T	0.71	0.914	1804.5	308.0	4.97E-09
	1	154540865	G	T	0.98	0.543	4949.8	845.4	5.09E-09
	1	153850697	T	C	0.99	0.840	6613.1	1130.5	5.25E-09
	1	153828359	A	G	0.99	0.840	6612.2	1130.5	5.28E-09
rs1552483	1	154324898	G	A	0.15	0.980	-2078.2	356.4	5.88E-09
rs56211693	1	154355008	C	T	0.70	0.992	-1697.6	292.1	6.58E-09
rs41313936	1	153617749	C	T	0.98	0.558	5170.0	892.0	7.22E-09
	1	154351995	G	A	0.93	0.858	3034.8	524.7	7.74E-09
rs11580178	1	154353262	T	A	0.70	0.992	-1690.6	292.6	8.05E-09
rs76697909	1	153451003	C	T	0.99	0.450	7019.6	1215.4	8.16E-09
rs1889315	1	154351960	T	C	0.70	0.991	-1688.0	292.7	8.56E-09
rs34437053	1	154351199	G	A	0.70	0.991	-1686.1	292.7	8.96E-09
	1	153454043	G	A	0.99	0.449	7007.5	1217.2	9.10E-09
	1	153965150	C	G	0.99	0.710	6617.7	1149.6	9.13E-09
rs10908834	1	154349381	C	T	0.30	0.989	1678.3	292.6	1.03E-08
rs114603881	1	154293923	T	G	0.96	0.802	3522.3	615.2	1.10E-08
rs55874907	1	153903129	C	T	0.95	0.839	-3744.2	654.4	1.12E-08
rs1889314	1	154351880	G	A	0.30	0.990	1670.7	292.6	1.19E-08
rs112505856	1	154107960	C	T	0.96	0.495	-5734.3	1005.9	1.27E-08
rs116038461	1	153412147	G	A	0.98	0.455	6808.1	1195.8	1.32E-08
rs41308403	1	154294560	C	T	0.92	0.935	2612.1	459.0	1.34E-08
rs115891382	1	154277834	T	C	0.96	0.804	3513.0	617.4	1.35E-08
rs79438264	1	154611891	C	T	0.97	0.619	-5075.4	894.9	1.50E-08
rs4845387	1	154645691	G	A	0.30	0.932	-1727.5	305.0	1.56E-08
rs391831	8	90685350	G	A	0.08	0.574	-3089.2	547.2	1.75E-08
	1	154643536	C	T	0.29	0.946	-1712.1	303.6	1.80E-08
	1	153743937	T	C	0.98	0.885	4499.9	798.8	1.87E-08
	1	153737991	T	C	0.98	0.885	4488.6	796.8	1.87E-08
	1	153880288	T	C	0.99	0.725	-8730.0	1550.4	1.90E-08
rs116072155	1	153831877	G	A	0.99	0.725	-8730.2	1550.5	1.90E-08
rs2280781	1	154540468	C	T	0.92	0.929	2408.2	429.4	2.17E-08
	1	154257452	C	T	0.99	0.747	-8268.1	1494.8	3.35E-08
rs113191295	1	153912623	G	A	0.95	0.593	-4468.4	809.0	3.50E-08
rs71503474	8	90685351	A	C	0.09	0.562	-2976.2	540.6	3.87E-08
	20	429758	G	A	0.99	0.380	7743.4	1409.9	4.18E-08

[A]rsnumber is given where available. [B]Base pair position build 37/hg19. [C]Allele 1 (effect allele) [D]Allele 2. [E]Frequency of allele 1. [F]Info= Imputation quality metric from SNPtest. [G]Beta for the effect on sIL-6R levels associated with allele 1. [H]SE = Standard error of the beta estimate

Table S3: SNP information and Conditional GWAS results of all 293 SNPs reaching genome wide significance in the GWAS of sIL-6R level after correction for the effect of rs2228145.

SNP[A]	Chr	Position[B]	A1[C]	A2[D]	Frequency A1[E]	Info[F]	Beta[G]	SE[H]	*P* value
rs6694258	1	154428505	C	A	0.45	0.991	-3458.1	244.1	1.43E-44
rs6689393	1	154426097	A	G	0.45	0.993	-3420.2	244.7	1.85E-43
rs7526131	1	154425135	G	A	0.45	0.993	-3409.1	244.6	3.23E-43

rs7536152	1	154423909	A	G	0.45	0.992	-3404.6	244.6	4.15E-43
rs12129500	1	154423764	T	C	0.45	0.992	-3402.3	244.7	4.74E-43
rs6686750	1	154419843	A	G	0.45	0.990	-3387.8	245.0	1.35E-42
rs11265612	1	154417044	A	G	0.44	0.975	-3334.0	244.1	1.30E-41
rs12404927	1	154421438	T	C	0.48	0.931	-3176.9	232.9	1.61E-41
rs6683206	1	154418088	T	C	0.44	0.969	-3321.0	245.7	8.15E-41
rs7549250	1	154404336	C	T	0.45	0.970	-3270.3	243.9	3.45E-40
rs6667434	1	154409100	A	G	0.44	0.969	-3246.5	242.4	4.20E-40
rs4845371	1	154408340	T	C	0.44	0.969	-3241.7	242.3	4.74E-40
rs4845619	1	154405058	T	G	0.45	0.970	-3277.8	245.2	5.64E-40
rs59632925	1	154406540	T	G	0.44	0.970	-3222.6	241.8	9.19E-40
rs7549338	1	154404380	C	G	0.44	0.970	-3208.6	241.3	1.43E-39
rs6694817	1	154401972	T	C	0.45	0.965	-3155.8	238.8	4.12E-39
rs6427658	1	154400799	T	C	0.45	0.965	-3123.7	238.0	1.24E-38
rs6687726	1	154400320	A	G	0.45	0.969	-3147.8	242.1	5.87E-38
rs12117832	1	154397610	A	G	0.45	0.959	-3093.0	238.4	8.35E-38
rs12118721	1	154397416	T	C	0.45	0.961	-3150.0	242.8	8.54E-38
rs10908838	1	154397984	T	G	0.44	0.963	-3091.4	239.3	1.71E-37
rs12118770	1	154397589	T	C	0.44	0.959	-3071.4	238.8	3.52E-37
rs6689306	1	154395946	A	G	0.45	0.937	-3090.6	241.9	1.03E-36
rs10908836	1	154397932	C	T	0.46	0.913	-2956.2	231.5	1.09E-36
rs35109459	1	154397933	G	A	0.47	0.917	-3002.4	236.9	3.64E-36
rs4308966	1	154432622	T	C	0.84	0.994	-3047.7	242.6	1.41E-35
rs4345796	1	154436750	C	T	0.84	0.985	-3057.8	243.7	1.65E-35
rs61812626	1	154436436	G	A	0.84	0.984	-3061.8	244.1	1.81E-35
rs4845374	1	154426947	T	A	0.84	0.992	-3047.7	243.8	3.01E-35
rs10159236	1	154431405	C	A	0.84	0.983	-3037.9	243.5	4.00E-35
rs11804305	1	154424497	C	T	0.84	0.992	-3048.6	244.4	4.13E-35
rs61812599	1	154421554	G	A	0.84	0.988	-3032.3	246.3	2.93E-34
rs61811362	1	154512495	G	A	0.84	0.997	-2976.0	242.9	5.74E-34
rs4845648	1	154514338	A	T	0.84	0.997	-2974.6	242.9	6.18E-34
rs9427076	1	154503022	C	T	0.84	0.998	-2966.6	242.7	8.15E-34
rs4845644	1	154501576	A	G	0.84	0.998	-2965.7	242.7	8.45E-34
rs61812654	1	154502517	G	A	0.84	0.998	-2965.7	242.7	8.48E-34
rs14021	1	154452204	T	C	0.84	0.998	-2966.1	242.8	9.19E-34
rs6427729	1	154518874	G	A	0.16	0.995	2968.3	243.0	9.21E-34
rs9427415	1	154508437	A	T	0.84	0.995	-2967.4	243.1	9.80E-34
rs11265627	1	154496286	C	A	0.16	0.998	2961.6	242.6	1.01E-33
rs72698167	1	154480157	G	A	0.84	1.000	-2963.9	242.8	1.01E-33
rs6691727	1	154497005	A	G	0.16	0.999	2961.4	242.6	1.04E-33
rs6670375	1	154494947	G	A	0.16	0.999	2960.5	242.6	1.06E-33
rs4575077	1	154495011	C	A	0.16	0.999	2960.5	242.6	1.06E-33
rs4269769	1	154494082	G	C	0.16	0.999	2960.4	242.6	1.06E-33
rs56233546	1	154471371	A	G	0.84	1.000	-2960.5	242.7	1.13E-33
rs4845634	1	154464160	G	A	0.84	0.999	-2960.0	242.7	1.13E-33
rs16836054	1	154462195	G	A	0.84	0.999	-2959.4	242.6	1.13E-33
rs61812631	1	154467816	A	T	0.84	0.999	-2960.2	242.7	1.13E-33
rs7513603	1	154481158	T	C	0.16	0.999	2961.0	242.8	1.16E-33
rs111885536	1	154461260	G	A	0.84	0.994	-2969.8	243.5	1.18E-33
rs12118074	1	154477585	G	A	0.16	0.999	2960.9	242.8	1.18E-33
rs6687112	1	154497146	C	T	0.16	0.998	2957.3	242.7	1.34E-33
rs9427082	1	154523562	G	T	0.84	0.990	-2960.8	243.1	1.41E-33
rs4845640	1	154497802	C	T	0.16	0.998	2955.0	242.7	1.43E-33
rs4845641	1	154497930	C	G	0.16	0.998	2955.0	242.7	1.43E-33
rs4845643	1	154498239	T	C	0.16	0.998	2955.0	242.7	1.43E-33
rs4585969	1	154498957	C	T	0.16	0.998	2954.9	242.7	1.43E-33
rs4307560	1	154499345	G	A	0.16	0.998	2954.9	242.7	1.43E-33
rs11265629	1	154498574	G	A	0.16	0.998	2954.9	242.7	1.43E-33

rs6701586	1	154499611	A	T	0.16	0.998	2954.9	242.7	1.43E-33
rs6697103	1	154499926	C	T	0.16	0.998	2954.9	242.7	1.44E-33
rs6697115	1	154499947	C	T	0.16	0.998	2954.8	242.7	1.44E-33
rs6656395	1	154500016	T	C	0.16	0.998	2954.8	242.7	1.44E-33
rs4131514	1	154500089	A	G	0.16	0.998	2954.8	242.7	1.44E-33
rs10908844	1	154500857	G	A	0.16	0.998	2954.8	242.7	1.44E-33
rs12129484	1	154501204	C	A	0.16	0.998	2954.8	242.7	1.44E-33
rs10908846	1	154501364	C	A	0.16	0.998	2954.7	242.7	1.44E-33
rs6427722	1	154502123	G	C	0.16	0.998	2954.7	242.7	1.45E-33
rs6427724	1	154502499	T	A	0.16	0.998	2954.7	242.7	1.45E-33
rs10752643	1	154503601	G	A	0.16	0.998	2954.7	242.7	1.45E-33
rs4518898	1	154492432	T	C	0.16	1.000	2952.2	242.8	1.74E-33
rs4288587	1	154492357	A	T	0.16	1.000	2952.2	242.8	1.75E-33
rs6662503	1	154477069	T	A	0.16	0.999	2952.2	242.8	1.76E-33
rs6696177	1	154477263	G	A	0.16	0.999	2952.4	242.8	1.77E-33
rs6689710	1	154470356	T	C	0.16	0.998	2949.9	242.6	1.80E-33
rs12124333	1	154478292	T	A	0.16	0.999	2952.1	242.8	1.80E-33
rs7526247	1	154485039	T	C	0.16	0.999	2951.0	242.8	1.93E-33
rs41302545	1	154525999	C	T	0.84	0.989	-2955.6	243.3	2.04E-33
rs4845626	1	154423485	G	T	0.84	0.985	-2966.7	244.2	2.06E-33
rs6427716	1	154497329	C	A	0.16	0.998	2945.3	242.6	2.21E-33
rs7539745	1	154526414	C	G	0.16	0.988	2952.2	243.3	2.45E-33
rs111301013	1	154419584	C	A	0.84	0.988	-3013.3	248.7	3.03E-33
rs7553602	1	154529421	A	T	0.84	0.982	-2927.2	243.5	9.01E-33
rs7550664	1	154528709	A	C	0.84	0.973	-2937.6	245.0	1.28E-32
rs56047170	1	154528053	G	A	0.84	0.972	-2933.7	245.0	1.53E-32
rs4307559	1	154499123	G	C	0.17	0.953	3041.1	254.1	1.68E-32
rs12047973	1	154394766	A	G	0.84	0.912	-3133.2	262.8	2.85E-32
rs12048950	1	154394966	T	C	0.84	0.914	-3118.9	262.4	4.30E-32
rs55800510	1	154417187	C	T	0.84	0.976	-3005.1	253.2	5.31E-32
rs6427720	1	154501756	G	A	0.17	0.980	2861.2	241.7	7.39E-32
rs12048091	1	154395077	A	G	0.84	0.910	-3087.7	260.9	7.80E-32
rs61812592	1	154410482	G	A	0.84	0.973	-3001.8	254.1	1.03E-31
rs61812593	1	154410490	T	C	0.84	0.973	-3001.7	254.1	1.03E-31
rs12023358	1	154399649	C	T	0.84	0.966	-3036.1	257.3	1.16E-31
rs4845645	1	154502909	T	A	0.18	0.922	2847.6	241.3	1.19E-31
rs57502626	1	154408916	T	A	0.84	0.973	-3001.5	254.4	1.21E-31
rs61812594	1	154411628	A	G	0.84	0.974	-2998.0	254.2	1.28E-31
rs55826755	1	154409520	C	G	0.84	0.973	-2999.6	254.5	1.37E-31
rs61812596	1	154414691	C	T	0.84	0.976	-2990.4	253.9	1.48E-31
rs11265634	1	154522248	A	G	0.17	0.984	2835.5	240.9	1.62E-31
rs11265611	1	154395125	G	A	0.42	0.894	-2789.6	242.4	3.22E-30
rs6427721	1	154501770	A	G	0.18	0.907	2868.2	249.3	3.30E-30
rs6703672	1	154474557	T	C	0.20	0.902	2591.9	235.9	1.03E-27
rs4556348	1	154394296	C	T	0.87	0.761	-3240.3	316.1	2.22E-24
rs11265616	1	154419980	T	C	0.60	0.702	-2426.7	237.7	3.42E-24
rs12128984	1	154498406	G	A	0.23	0.772	2415.0	239.2	1.03E-23
rs10465961	1	154487354	T	A	0.26	0.729	2381.1	236.2	1.23E-23
rs6690230	1	154432877	C	G	0.40	0.990	1900.7	209.2	1.56E-19
rs6695045	1	154432957	A	G	0.40	0.990	1900.4	209.3	1.59E-19
	1	154431123	T	A	0.93	0.824	-3080.9	343.7	4.57E-19
rs45478197	1	154422733	C	T	0.91	0.928	-2594.7	306.1	3.12E-17
rs11265621	1	154442960	G	A	0.37	0.991	1777.8	210.2	3.68E-17
	1	154431128	G	A	0.93	0.827	-3023.8	357.5	3.70E-17
rs4478801	1	154464572	G	A	0.37	0.994	1770.2	210.5	5.46E-17
rs6658175	1	154475330	T	A	0.37	0.997	1757.8	210.7	9.62E-17
rs6684921	1	154464945	A	C	0.37	0.997	1758.6	211.0	1.05E-16
rs12119111	1	154478600	G	A	0.37	0.998	1753.8	210.6	1.09E-16

rs12133641	1	154428283	A	G	0.61	0.993	-13784.0	1656.1	1.13E-16
rs9660786	1	154484017	A	T	0.37	0.998	1751.3	210.6	1.19E-16
rs7518694	1	154484788	C	G	0.37	0.998	1751.2	210.6	1.19E-16
rs7519499	1	154487926	G	A	0.37	0.998	1751.1	210.6	1.20E-16
rs4845639	1	154490352	C	T	0.37	0.998	1751.0	210.6	1.20E-16
rs6686276	1	154466188	C	A	0.37	0.998	1752.8	210.9	1.24E-16
rs6700296	1	154473660	T	C	0.37	0.998	1752.7	210.9	1.25E-16
rs6689965	1	154470606	T	A	0.37	0.999	1752.3	210.9	1.26E-16
rs12753680	1	154474900	G	A	0.37	0.998	1751.5	210.9	1.31E-16
rs12118018	1	154477440	G	A	0.37	0.999	1750.3	210.8	1.33E-16
rs6664608	1	154479670	C	T	0.37	0.999	1749.1	210.7	1.35E-16
rs4845637	1	154490178	A	G	0.37	0.998	1748.9	210.8	1.40E-16
rs10908841	1	154487763	C	T	0.37	0.998	1748.4	210.7	1.40E-16
rs12740969	1	154487060	T	G	0.37	0.998	1748.0	210.7	1.42E-16
rs12128408	1	154488533	A	G	0.37	0.998	1747.9	210.7	1.42E-16
rs4845638	1	154490269	T	A	0.37	0.998	1747.9	210.7	1.43E-16
rs10908842	1	154492702	G	A	0.37	0.998	1747.6	210.7	1.45E-16
rs6698971	1	154475331	C	A	0.37	0.995	1754.1	211.6	1.48E-16
rs4345797	1	154495675	C	T	0.37	0.997	1730.8	211.5	3.54E-16
rs10908843	1	154494480	C	G	0.37	0.997	1728.2	211.6	4.06E-16
rs11265628	1	154496489	C	T	0.37	0.997	1728.0	211.6	4.10E-16
rs4845647	1	154514331	C	A	0.37	0.995	1719.8	212.3	7.00E-16
rs12136771	1	154511502	T	C	0.37	0.994	1720.9	212.5	7.27E-16
rs10908847	1	154504954	G	A	0.37	0.995	1723.3	213.0	7.57E-16
rs11265632	1	154514067	G	A	0.37	0.995	1717.0	212.4	8.05E-16
rs4845642	1	154498028	G	A	0.37	0.996	1715.0	212.1	8.06E-16
rs10908845	1	154500898	C	T	0.37	0.996	1714.4	212.2	8.38E-16
rs4382717	1	154518610	C	T	0.37	0.989	1712.5	212.4	9.50E-16
rs12407048	1	154505270	C	T	0.37	0.997	1710.7	212.4	1.01E-15
rs4638123	1	154505704	C	T	0.37	0.997	1710.6	212.4	1.02E-15
rs12037271	1	154507888	G	A	0.37	0.996	1710.6	212.4	1.02E-15
rs6688376	1	154326366	C	T	0.70	0.959	1516.3	189.5	1.55E-15
rs77302258	1	154396203	A	G	0.91	0.898	-2581.9	325.6	2.77E-15
rs55908418	1	154513387	A	C	0.94	0.926	-3002.6	378.8	2.82E-15
rs12753666	1	154474875	G	A	0.37	0.986	1666.1	210.9	3.54E-15
rs72698182	1	154533863	G	C	0.93	0.905	-2967.3	376.5	4.06E-15
rs12083537	1	154381103	A	G	0.79	0.850	1843.7	234.9	5.25E-15
rs1386821	1	154382049	T	G	0.80	0.804	1919.9	247.8	1.16E-14
rs61812598	1	154420087	G	A	0.61	0.993	-13155.7	1704.3	1.44E-14
rs1889312	1	154326473	G	A	0.57	0.978	1364.8	177.7	1.96E-14
rs61812595	1	154411852	G	A	0.94	0.843	-3156.7	422.4	9.38E-14
rs61811370	1	154537502	T	A	0.93	0.988	-2580.9	355.0	4.23E-13
rs61811371	1	154537505	T	C	0.93	0.988	-2580.7	355.0	4.25E-13
rs4845378	1	154544651	G	T	0.93	0.951	-2599.6	360.0	6.06E-13
rs77994623	1	154505106	C	T	0.84	0.933	1665.2	230.7	6.19E-13
rs78664422	1	154499246	G	A	0.84	0.944	1659.1	230.0	6.35E-13
rs6662003	1	154343588	C	T	0.86	0.921	1848.4	257.0	7.47E-13
rs12077265	1	154359915	T	G	0.84	0.955	1728.7	241.3	9.23E-13
rs61804489	1	154354643	C	T	0.85	0.979	1718.2	240.2	9.86E-13
rs61804488	1	154354370	A	G	0.85	0.979	1717.8	240.2	1.00E-12
rs61812565	1	154363305	C	T	0.85	0.972	1726.5	241.5	1.01E-12
rs68138282	1	154363441	C	T	0.85	0.972	1726.4	241.5	1.01E-12
rs112527081	1	154354257	G	A	0.85	0.979	1717.6	240.2	1.01E-12
rs36029470	1	154352194	T	C	0.84	0.974	1719.3	240.5	1.02E-12
rs34065906	1	154354129	G	A	0.85	0.979	1717.3	240.2	1.03E-12
rs34146123	1	154353846	C	A	0.85	0.979	1716.7	240.2	1.05E-12
rs35595831	1	154353743	C	T	0.85	0.979	1716.3	240.2	1.06E-12
rs35629421	1	154352252	C	A	0.85	0.980	1714.4	240.3	1.13E-12

rs35091500	1	154352310	G	A	0.85	0.980	1714.3	240.3	1.13E-12
rs35112510	1	154352165	T	C	0.85	0.980	1714.3	240.3	1.13E-12
rs4304577	1	154352133	T	C	0.85	0.980	1714.3	240.3	1.13E-12
rs1889316	1	154352017	C	T	0.85	0.980	1714.0	240.3	1.14E-12
rs16865579	1	154351073	T	C	0.85	0.980	1713.4	240.3	1.17E-12
rs3811451	1	154350589	T	C	0.85	0.980	1713.2	240.3	1.18E-12
rs61803414	1	154350238	C	G	0.85	0.980	1712.9	240.3	1.19E-12
rs6690468	1	154349639	T	A	0.85	0.980	1712.3	240.3	1.21E-12
rs34749203	1	154352641	C	T	0.85	0.977	1717.8	241.2	1.25E-12
rs6661998	1	154343576	C	T	0.88	0.848	2025.5	284.8	1.32E-12
rs2280781	1	154540468	C	T	0.92	0.929	-2504.5	352.5	1.41E-12
rs28410194	1	154348336	C	T	0.85	0.980	1701.6	240.2	1.61E-12
rs28530583	1	154348332	C	G	0.85	0.980	1701.5	240.2	1.61E-12
rs61811372	1	154543215	T	C	0.93	0.971	-2527.3	357.2	1.73E-12
rs61803412	1	154348199	A	G	0.85	0.980	1698.8	240.2	1.75E-12
rs71586008	1	154343577	A	G	0.87	0.855	1985.0	282.7	2.53E-12
rs4622059	1	154347662	C	T	0.85	0.981	1687.1	240.6	2.71E-12
rs4526605	1	154347467	C	T	0.84	0.982	1684.4	240.4	2.84E-12
rs9728024	1	154346161	C	T	0.85	0.966	1700.2	243.6	3.42E-12
rs9727988	1	154345857	C	T	0.84	0.942	1684.3	242.4	4.22E-12
rs4406622	1	154346416	A	G	0.84	0.982	1673.3	241.5	4.83E-12
rs6663105	1	154346950	G	C	0.84	0.982	1672.7	241.4	4.84E-12
rs12072348	1	154547292	A	C	0.91	0.855	-2340.7	337.8	4.86E-12
rs6662855	1	154346669	G	A	0.84	0.982	1672.3	241.4	4.92E-12
rs7530868	1	154345567	T	C	0.84	0.982	1671.7	241.4	5.02E-12
rs7530735	1	154345371	T	C	0.84	0.982	1671.5	241.4	5.04E-12
rs7523399	1	154345275	C	G	0.84	0.982	1671.4	241.4	5.05E-12
rs7511806	1	154344897	G	A	0.84	0.982	1671.2	241.4	5.09E-12
rs12066892	1	154344275	C	T	0.84	0.982	1670.6	241.4	5.17E-12
rs6701309	1	154343935	G	C	0.84	0.982	1670.5	241.4	5.19E-12
rs6667225	1	154343941	A	C	0.84	0.982	1670.5	241.4	5.19E-12
rs7518394	1	154342022	T	A	0.85	0.983	1669.6	241.4	5.32E-12
rs12058332	1	154341406	G	A	0.85	0.983	1669.1	241.4	5.40E-12
rs7552370	1	154340321	C	T	0.84	0.983	1668.4	241.4	5.50E-12
rs12063437	1	154340084	G	A	0.85	0.984	1667.0	241.4	5.72E-12
rs68031662	1	154339892	C	T	0.85	0.984	1665.6	241.4	5.92E-12
rs6662976	1	154338499	C	T	0.85	0.973	1665.5	242.7	7.70E-12
rs6672627	1	154346967	C	A	0.84	0.982	1658.3	241.6	7.72E-12
rs12068128	1	154338155	T	A	0.85	0.989	1651.5	241.1	8.39E-12
rs12749618	1	154337824	C	T	0.84	0.987	1649.4	241.1	8.87E-12
	1	154532599	A	T	0.89	0.765	-2175.8	318.0	8.91E-12
rs12739228	1	154426190	G	A	0.95	0.894	2705.9	396.3	9.84E-12
rs6427741	1	154535817	A	C	0.92	0.872	-2572.2	377.3	1.05E-11
rs12730935	1	154419892	G	A	0.61	0.985	-10941.5	1608.1	1.16E-11
rs1948143	1	154334264	G	C	0.84	0.988	1631.0	240.4	1.32E-11
rs10908840	1	154459477	C	T	0.52	0.953	1239.5	183.9	1.79E-11
rs11265594	1	154328508	G	A	0.85	0.997	1623.8	241.4	1.95E-11
rs4414034	1	154331186	A	G	0.84	0.977	1601.3	243.3	5.25E-11
rs3811452	1	154321802	C	T	0.85	0.996	1598.2	243.5	5.87E-11
rs34968830	1	154313689	G	A	0.85	0.999	1571.9	243.6	1.22E-10
rs66980031	1	154481153	C	G	0.95	0.868	2637.6	410.1	1.39E-10
rs35229198	1	154330289	C	T	0.85	0.978	1583.4	246.2	1.40E-10
rs6686621	1	154293925	T	A	0.85	0.992	1566.2	244.1	1.55E-10
rs6702754	1	154303976	T	C	0.70	0.997	1251.3	195.4	1.67E-10
rs4601580	1	154394417	T	A	0.48	0.736	-1511.5	236.2	1.72E-10
rs9787014	1	154308680	G	A	0.85	0.997	1558.2	243.5	1.73E-10
rs749966	1	154314408	C	T	0.85	0.997	1559.7	243.8	1.76E-10
rs6672010	1	154292161	C	T	0.70	0.992	1252.9	195.9	1.78E-10

rs6685323	1	154295592	C	T	0.70	0.998	1246.9	195.2	1.85E-10
rs12727865	1	154303705	C	T	0.70	0.999	1247.5	195.3	1.85E-10
rs35950207	1	154291958	C	T	0.70	0.999	1246.5	195.2	1.87E-10
rs3890153	1	154295089	C	G	0.70	0.999	1245.4	195.2	1.96E-10
rs12748146	1	154292982	T	A	0.70	0.998	1246.0	195.3	1.96E-10
rs12077870	1	154305190	G	C	0.70	0.998	1241.3	194.7	2.01E-10
rs6668968	1	154293675	G	A	0.70	0.998	1244.4	195.2	2.04E-10
rs34880256	1	154308089	C	T	0.85	0.998	1548.1	243.7	2.33E-10
rs12726220	1	154171550	A	G	0.96	0.813	2974.7	472.0	3.23E-10
	1	154540865	G	T	0.98	0.543	-4207.9	676.2	5.36E-10
	1	154430647	C	T	0.97	0.680	3732.3	607.2	8.64E-10
rs12730036	1	154416969	C	T	0.61	0.977	-6971.3	1134.6	8.76E-10
rs11265613	1	154418415	T	C	0.61	0.983	-8293.4	1352.3	9.41E-10
rs12753254	1	154416935	G	A	0.61	0.977	-6933.7	1130.6	9.41E-10
rs66654715	1	154376820	G	C	0.96	0.776	2774.4	454.2	1.09E-09
rs34450945	1	154289348	C	T	0.96	0.768	2759.7	469.8	4.56E-09
rs72700210	1	154665602	T	A	0.93	0.813	-2189.5	373.1	4.72E-09
rs72698179	1	154522394	C	A	0.98	0.663	-3800.1	652.8	6.27E-09
rs7529229	1	154420778	T	C	0.60	0.991	-5952.6	1032.7	8.76E-09
rs7512646	1	154420402	G	C	0.60	0.991	-5915.2	1028.5	9.46E-09
rs55997241	1	154671128	G	A	0.93	0.753	-2335.4	409.9	1.29E-08
rs4576655	1	154418749	C	T	0.60	0.989	-5555.9	977.3	1.39E-08
rs12729561	1	154321663	C	T	0.86	0.978	1482.1	260.7	1.39E-08
rs10218433	1	154330183	T	C	0.71	0.982	1174.8	207.7	1.64E-08
rs12143987	1	154330725	C	G	0.30	0.992	-1162.5	206.2	1.84E-08
rs67881926	1	153831229	C	G	0.94	0.885	2258.1	400.6	1.84E-08
rs1889314	1	154351880	G	A	0.30	0.990	-1175.7	208.7	1.89E-08
rs10908834	1	154349381	C	T	0.30	0.989	-1174.1	208.9	2.02E-08
rs12759389	1	153773326	G	C	0.94	0.887	2247.5	400.6	2.15E-08
rs1073907	1	154332520	C	T	0.70	0.995	1155.3	206.0	2.17E-08
rs11577266	1	154333216	A	G	0.70	0.994	1156.2	206.2	2.18E-08
rs12719998	1	154335417	A	G	0.70	0.993	1157.2	206.4	2.21E-08
rs34845300	1	153755335	C	A	0.94	0.852	2118.3	378.1	2.24E-08
rs34437053	1	154351199	G	A	0.70	0.991	1168.9	208.8	2.30E-08
rs11265593	1	154328464	T	C	0.70	0.991	1157.6	206.8	2.30E-08
rs7555891	1	154329595	A	G	0.70	0.991	1156.5	206.6	2.32E-08
rs1889315	1	154351960	T	C	0.70	0.991	1168.0	208.8	2.34E-08
rs11580178	1	154353262	T	A	0.70	0.992	1167.4	208.7	2.37E-08
rs11576193	1	154330749	T	G	0.70	0.993	1152.6	206.3	2.46E-08
rs6696089	1	154332870	A	G	0.70	0.990	1157.4	207.5	2.56E-08
rs2072661	1	154548880	G	A	0.76	0.883	1219.1	218.7	2.64E-08
rs12738097	1	154335378	C	T	0.71	0.995	1151.3	207.0	2.85E-08
rs6698143	1	154337209	G	T	0.71	0.993	1152.1	207.3	2.92E-08
rs56023096	1	154291416	C	T	0.86	0.973	1450.3	261.1	2.94E-08
rs6657938	1	154336126	A	G	0.71	0.995	1150.4	207.1	2.94E-08
rs56211693	1	154355008	C	T	0.70	0.992	1157.0	208.4	2.98E-08
rs6698130	1	154337177	G	A	0.71	0.993	1151.2	207.4	3.00E-08
rs12118634	1	154335146	C	T	0.30	0.972	-1139.4	205.4	3.08E-08
rs12061599	1	154344135	G	T	0.71	0.989	1152.9	208.3	3.32E-08
rs6661909	1	154343486	C	T	0.71	0.989	1152.4	208.3	3.33E-08
rs35982711	1	154339919	C	A	0.71	0.989	1152.6	208.3	3.34E-08
rs7525436	1	154345089	A	G	0.71	0.989	1152.2	208.3	3.35E-08
rs9728585	1	154346077	A	C	0.71	0.989	1152.0	208.3	3.35E-08
rs113445115	1	154122787	T	C	0.92	0.807	-2051.5	370.9	3.37E-08
rs6659828	1	154341479	A	G	0.71	0.989	1151.7	208.3	3.39E-08
rs2229857	1	154573967	T	C	0.29	0.977	-1109.9	200.9	3.48E-08
rs7544062	1	154554022	C	T	0.29	0.987	-1106.0	200.3	3.58E-08
rs11576181	1	154330659	T	G	0.71	0.984	1140.7	206.8	3.67E-08

[A]rsnumber is given where available

[B]Base pair position build 37/hg19. [C]Allele 1 (effect allele). [D]Allele 2. [E]Frequency of allele 1 [F]Info= Imputation quality metric from SNPtest. [G]Beta for the effect on sIL-6R levels associated with allele 1. [H]SE = Standard error of the beta estimate

Table S4: Location (Build 37/hg19) and sequence of expression probes targeting *IL6R* transcripts (Affymetrix U219 array).

Probe ID	Sequence	Chr	Start (basepair)	End (basepair)	Location (*IL6R*)
11736509_x_at;545;104	GGTTACCCAGTTAGCTCTCAAGTTA	1	154439782	154439807	3'-end
11736509_x_at;40;550	GATTTTAGACCCTATTGCTGCTTGA	1	154440035	154440060	3'-end
11736509_x_at;410;418	TCATTGCTGGGCTTGTCTGTAGATT	1	154440014	154440039	3'-end
11736509_x_at;308;15	AAATCAGCCGTGTAACCATGGACCC	1	154439837	154439862	3'-end
11736509_x_at;553;369	ACTTCAGCTGACTTTTCTGTCCGAG	1	154439733	154439758	3'-end
11736509_x_at;372;285	CCATGTTTTGTTTACGGTTTTCCAG	1	154439953	154439978	3'-end
11736509_x_at;495;603	GTTCCTTGAGTTGATTCAGCTCTGC	1	154439660	154439685	3'-end
11736509_x_at;61;280	ACTGTCTTCAGTAAGCCGTGATTTT	1	154439990	154440015	3'-end
11736509_x_at;551;225	GCTCTCAAGTTATCAGGGTATTCCA	1	154439795	154439820	3'-end
11736510_a_at;390;362	ACGGAGCCCTTATGACATCAGCAAT	1	154437807	154437832	3'-end
11736510_a_at;323;134	CAGGAGTCCTCCAGCTGAGAACGAG	1	154420604	154420629	Exon 7
11736510_a_at;609;718	AAGACAAGCATGCATCCGCCGTACT	1	154437649	154437674	3'-end
11736510_a_at;58;615	GGGTCTGACAATACCTCGAGCCACA	1	154437760	154437785	3'-end
11736510_a_at;46;386	AGACTACTTCTTCCCCAGATAGCTG	1	154437834	154437859	3'-end
11736510_a_at;484;717	GATTCTGCAAATGCGACAAGCCTCC	1	154422428	154422453	Exon 8
11736510_a_at;659;168	GCCACAACCGACCAGATGCCAGGGA	1	154437779	154437804	3'-end
11736510_a_at;201;497	TCCAGCTGAGAACGAGGTGTCCACC	1	154420613	154420638	Exon 7
11736510_a_at;435;552	TCAGCAATACAGACTACTTCTTCCC	1	154437824	154437849	3'-end
11741958_a_at;264;163	TACAGACTACGGTTTGAGCTCAGAT	1	154407569	154407594	Exon 5
11741958_a_at;582;132	CAGGAGTCCTCCAGCTGAGAACGAG	1	154420604	154420629	Exon 7
11741958_a_at;610;718	AAGACAAGCATGCATCCGCCGTACT	1	154437649	154437674	3'-end
11741958_a_at;483;717	GATTCTGCAAATGCGACAAGCCTCC	1	154422428	154422453	Exon 8
11741958_a_at;235;610	ACTGTGTCATCCACGACGCCTGGAG	1	154408463	154408488	Exon 6
11741958_a_at;202;497	TCCAGCTGAGAACGAGGTGTCCACC	1	154420613	154420638	Exon 7
11741958_a_at;705;156	CAGCTGGTCCCGGAGAGGCCTCGAC	1	154437682	154437707	3'-end
11741958_a_at;314;329	TACGGTTTGAGCTCAGATATCGGGC	1	154407576	154407601	Exon 5
11741959_x_at;505;200	CTGCTGACTGTTTTCTCTTGAGAGG	1	154438903	154438928	3'-end
11741959_x_at;361;694	AATATCCAATATTCGCTGTGTCAGC	1	154438932	154438957	3'-end
11741959_x_at;586;417	TTTACTGCAGCTTTGTTTGTTGTCA	1	154438780	154438805	3'-end
11741959_x_at;735;334	TGTGGTAGAGTGTGCCTGAAGTCCC	1	154439183	154439208	3'-end
11741959_x_at;109;540	TGTTGTCAGCTGAACCTGGGTAACT	1	154438797	154438822	3'-end
11741959_x_at;260;360	AGCACCATAACTTTGTTTAGCCCAA	1	154438986	154439011	3'-end
11741959_x_at;218;13	ATCAAAACGGTTTTACTGCAGCTTT	1	154438769	154438794	3'-end
11741959_x_at;267;582	CGCTGTGTCAGCATAGAAGTAACTT	1	154438945	154438970	3'-end
11741959_x_at;548;21	TCTCTTGAGAGGGTGGAATATCCAA	1	154438916	154438941	3'-end
11741959_x_at;7;323	GCTTTGTTTGTTGTCAGCTGAACCT	1	154438789	154438814	3'-end

Table S5: SNP information and eQTL *P*-values for all 341 significant eQTL associations.

Probe Name[A]	SNP	Chr	SNP position[B]	Alleles		MAF	Info[C]	eQTL *P*-value
582_132	rs7512646	1	154420402	G	C	0.40	0.991	2.84E-22
582_132	rs4845372	1	154415396	C	A	0.40	0.974	2.93E-22
582_132	rs7529229	1	154420778	T	C	0.40	0.991	3.41E-22
582_132	rs4576655	1	154418749	C	T	0.40	0.989	5.14E-22
323_134	rs4845372	1	154415396	C	A	0.40	0.974	6.44E-22
582_132	rs4845623	1	154415777	A	G	0.40	0.973	7.29E-22
582_132	rs4537545	1	154418879	C	T	0.39	0.976	1.30E-21
323_134	rs4845623	1	154415777	A	G	0.40	0.973	1.37E-21
323_134	rs7529229	1	154420778	T	C	0.40	0.991	2.62E-21
323_134	rs7512646	1	154420402	G	C	0.40	0.991	2.77E-21

323_134	rs4576655	1	154418749	C	T	0.40	0.989	4.29E-21
323_134	rs12753254	1	154416935	G	A	0.39	0.977	1.78E-20
323_134	rs12730036	1	154416969	C	T	0.39	0.977	1.78E-20
323_134	rs4537545	1	154418879	C	T	0.39	0.976	1.78E-20
323_134	rs6664201	1	154414296	C	T	0.39	0.973	1.90E-20
323_134	rs4845373	1	154417829	C	T	0.39	0.979	5.24E-20
323_134	rs4393147	1	154414037	C	T	0.39	0.973	5.81E-20
323_134	rs4453032	1	154414086	A	G	0.39	0.973	5.81E-20
323_134	rs11265613	1	154418415	T	C	0.39	0.983	6.31E-20
323_134	rs61812598	1	154420087	G	A	0.39	0.993	1.05E-19
582_132	rs12730935	1	154419892	G	A	0.38	0.985	1.10E-19
323_134	rs12133641	1	154428283	A	G	0.39	0.993	1.11E-19
323_134	rs12730935	1	154419892	G	A	0.38	0.985	1.14E-19
582_132	rs4393147	1	154414037	C	T	0.39	0.973	1.27E-19
582_132	rs4453032	1	154414086	A	G	0.39	0.973	1.27E-19
582_132	rs12753254	1	154416935	G	A	0.39	0.977	1.29E-19
582_132	rs12730036	1	154416969	C	T	0.39	0.977	1.29E-19
582_132	rs61812598	1	154420087	G	A	0.39	0.993	1.31E-19
582_132	rs6664201	1	154414296	C	T	0.39	0.973	1.40E-19
323_134	rs12126142	1	154425456	G	A	0.39	0.997	1.54E-19
323_134	rs4129267	1	154426264	C	T	0.38	0.995	1.59E-19
323_134	rs4845622	1	154411419	A	C	0.39	0.970	1.88E-19
582_132	rs11265613	1	154418415	T	C	0.39	0.983	1.89E-19
582_132	rs12133641	1	154428283	A	G	0.39	0.993	2.02E-19
323_134	rs2228145	1	154426970	A	C	0.39	0.967	2.25E-19
582_132	rs12126142	1	154425456	G	A	0.39	0.997	2.53E-19
582_132	rs4845373	1	154417829	C	T	0.39	0.979	2.77E-19
582_132	rs4845622	1	154411419	A	C	0.39	0.970	2.98E-19
582_132	rs2228145	1	154426970	A	C	0.39	0.967	3.14E-19
582_132	rs4129267	1	154426264	C	T	0.38	0.995	5.85E-19
323_134	rs4845621	1	154409730	G	A	0.39	0.970	1.21E-18
323_134	rs4845620	1	154406656	A	G	0.39	0.969	1.87E-18
323_134	rs56383622	1	154405024	A	G	0.39	0.969	2.27E-18
323_134	rs7521458	1	154407713	T	C	0.39	0.969	2.40E-18
582_132	rs4845621	1	154409730	G	A	0.39	0.970	2.48E-18
323_134	rs7518199	1	154407419	A	C	0.39	0.968	2.77E-18
582_132	rs7518199	1	154407419	A	C	0.39	0.968	5.36E-18
582_132	rs7521458	1	154407713	T	C	0.39	0.969	8.57E-18
202_497	rs4845372	1	154415396	C	A	0.40	0.974	1.19E-17
202_497	rs4845623	1	154415777	A	G	0.40	0.973	1.58E-17
582_132	rs4845620	1	154406656	A	G	0.39	0.969	1.77E-17
582_132	rs56383622	1	154405024	A	G	0.39	0.969	2.54E-17
323_134	rs6684439	1	154395839	C	T	0.39	0.926	2.79E-17
202_497	rs7529229	1	154420778	T	C	0.40	0.991	3.92E-17
202_497	rs7512646	1	154420402	G	C	0.40	0.991	4.06E-17
202_497	rs4576655	1	154418749	C	T	0.40	0.989	6.39E-17
202_497	rs6664201	1	154414296	C	T	0.39	0.973	1.40E-16
202_497	rs12753254	1	154416935	G	A	0.39	0.977	1.56E-16
202_497	rs12730036	1	154416969	C	T	0.39	0.977	1.56E-16
202_497	rs12730935	1	154419892	G	A	0.38	0.985	1.70E-16
202_497	rs11265613	1	154418415	T	C	0.39	0.983	1.89E-16
202_497	rs4393147	1	154414037	C	T	0.39	0.973	2.06E-16
202_497	rs4453032	1	154414086	A	G	0.39	0.973	2.06E-16
202_497	rs4845373	1	154417829	C	T	0.39	0.979	2.41E-16
202_497	rs4845620	1	154406656	A	G	0.39	0.969	2.42E-16
202_497	rs56383622	1	154405024	A	G	0.39	0.969	2.45E-16
202_497	rs7521458	1	154407713	T	C	0.39	0.969	3.29E-16
202_497	rs7518199	1	154407419	A	C	0.39	0.968	3.32E-16

202_497	rs61812598	1	154420087	G	A	0.39	0.993	4.08E-16
323_134	rs11265612	1	154417044	A	G	0.44	0.975	4.77E-16
202_497	rs4537545	1	154418879	C	T	0.39	0.976	4.90E-16
202_497	rs4845622	1	154411419	A	C	0.39	0.970	5.17E-16
202_497	rs4845621	1	154409730	G	A	0.39	0.970	5.82E-16
323_134	rs6683206	1	154418088	T	C	0.44	0.969	6.75E-16
323_134	rs4133213	1	154395212	C	A	0.42	0.893	1.33E-15
323_134	rs6686750	1	154419843	A	G	0.45	0.990	1.34E-15
202_497	rs4129267	1	154426264	C	T	0.38	0.995	1.39E-15
202_497	rs12133641	1	154428283	A	G	0.39	0.993	1.39E-15
323_134	rs12129500	1	154423764	T	C	0.45	0.992	1.64E-15
323_134	rs7536152	1	154423909	A	G	0.45	0.992	1.64E-15
202_497	rs12126142	1	154425456	G	A	0.39	0.997	1.64E-15
582_132	rs6684439	1	154395839	C	T	0.39	0.926	1.64E-15
323_134	rs7526131	1	154425135	G	A	0.45	0.993	1.67E-15
582_132	rs4133213	1	154395212	C	A	0.42	0.893	1.70E-15
323_134	rs6694258	1	154428505	C	A	0.45	0.991	1.70E-15
323_134	rs4845625	1	154422067	T	C	0.45	0.991	1.82E-15
323_134	rs6689393	1	154426097	A	G	0.45	0.993	2.61E-15
582_132	rs6686750	1	154419843	A	G	0.45	0.990	2.75E-15
582_132	rs4845625	1	154422067	T	C	0.45	0.991	2.80E-15
202_497	rs2228145	1	154426970	A	C	0.39	0.967	2.98E-15
582_132	rs11265612	1	154417044	A	G	0.44	0.975	4.12E-15
582_132	rs6694258	1	154428505	C	A	0.45	0.991	4.32E-15
582_132	rs6689393	1	154426097	A	G	0.45	0.993	4.34E-15
582_132	rs7526131	1	154425135	G	A	0.45	0.993	4.49E-15
582_132	rs12129500	1	154423764	T	C	0.45	0.992	4.89E-15
582_132	rs7536152	1	154423909	A	G	0.45	0.992	4.89E-15
582_132	rs6683206	1	154418088	T	C	0.44	0.969	5.91E-15
202_497	rs6684439	1	154395839	C	T	0.39	0.926	1.06E-14
202_497	rs4133213	1	154395212	C	A	0.42	0.893	2.55E-14
323_134	rs7549338	1	154404380	C	G	0.44	0.970	2.81E-14
323_134	rs59632925	1	154406540	T	G	0.44	0.970	3.15E-14
323_134	rs7553796	1	154404406	A	C	0.44	0.969	5.24E-14
323_134	rs4845371	1	154408340	T	C	0.44	0.969	6.95E-14
323_134	rs6667434	1	154409100	A	G	0.44	0.969	1.06E-13
582_132	rs7549338	1	154404380	C	G	0.44	0.970	1.26E-13
582_132	rs6667434	1	154409100	A	G	0.44	0.969	1.56E-13
323_134	rs10908838	1	154397984	T	G	0.44	0.963	1.57E-13
582_132	rs7553796	1	154404406	A	C	0.44	0.969	1.65E-13
323_134	rs6694817	1	154401972	T	C	0.44	0.965	1.71E-13
582_132	rs59632925	1	154406540	T	G	0.44	0.970	1.72E-13
582_132	rs4845371	1	154408340	T	C	0.44	0.969	1.83E-13
323_134	rs6689306	1	154395946	A	G	0.44	0.937	2.03E-13
323_134	rs6427658	1	154400799	T	C	0.45	0.965	2.11E-13
323_134	rs12118770	1	154397589	T	C	0.44	0.959	2.69E-13
201_497	rs4845623	1	154415777	A	G	0.40	0.973	3.35E-13
323_134	rs7549250	1	154404336	C	T	0.45	0.970	4.83E-13
323_134	rs4553185	1	154410955	C	T	0.45	0.966	5.71E-13
201_497	rs4845372	1	154415396	C	A	0.40	0.974	5.77E-13
323_134	rs4845619	1	154405058	T	G	0.45	0.970	6.12E-13
323_134	rs12404927	1	154421438	T	C	0.48	0.931	6.59E-13
323_134	rs10908836	1	154397932	C	T	0.46	0.913	7.31E-13
582_132	rs6694817	1	154401972	T	C	0.44	0.965	1.48E-12
201_497	rs7529229	1	154420778	T	C	0.40	0.991	1.55E-12
201_497	rs7512646	1	154420402	G	C	0.40	0.991	1.88E-12
323_134	rs12118721	1	154397416	T	C	0.45	0.961	2.79E-12
323_134	rs6687726	1	154400320	A	G	0.45	0.969	2.79E-12

582_132	rs12404927	1	154421438	T	C	0.48	0.931	3.04E-12
323_134	rs4845618	1	154400015	G	T	0.45	0.964	3.32E-12
323_134	rs12117832	1	154397610	A	G	0.45	0.959	3.51E-12
582_132	rs6427658	1	154400799	T	C	0.45	0.965	3.57E-12
201_497	rs4537545	1	154418879	C	T	0.39	0.976	3.84E-12
201_497	rs4576655	1	154418749	C	T	0.40	0.989	4.57E-12
582_132	rs10908838	1	154397984	T	G	0.44	0.963	5.16E-12
582_132	rs12118770	1	154397589	T	C	0.44	0.959	5.74E-12
582_132	rs6689306	1	154395946	A	G	0.44	0.937	6.49E-12
582_132	rs10908836	1	154397932	C	T	0.46	0.913	7.59E-12
582_132	rs4553185	1	154410955	C	T	0.45	0.966	9.04E-12
582_132	rs7549250	1	154404336	C	T	0.45	0.970	1.09E-11
323_134	rs35109459	1	154397933	G	A	0.47	0.917	1.11E-11
308_15	rs4072391	1	154438880	T	C	0.21	0.983	1.32E-11
201_497	rs6664201	1	154414296	C	T	0.39	0.973	1.48E-11
308_15	rs60255122	1	154504702	A	G	0.21	0.988	1.80E-11
308_15	rs60760897	1	154504849	C	T	0.21	0.988	1.80E-11
308_15	rs61698846	1	154504854	C	T	0.21	0.988	1.80E-11
308_15	rs61275241	1	154504887	T	G	0.21	0.988	1.80E-11
201_497	rs12753254	1	154416935	G	A	0.39	0.977	1.87E-11
201_497	rs12730036	1	154416969	C	T	0.39	0.977	1.87E-11
308_15	rs59239860	1	154492107	G	T	0.21	0.994	2.03E-11
308_15	rs112231452	1	154504900	G	A	0.21	0.988	2.03E-11
201_497	rs4393147	1	154414037	C	T	0.39	0.973	2.04E-11
201_497	rs4453032	1	154414086	A	G	0.39	0.973	2.04E-11
308_15	rs61403567	1	154504927	T	C	0.21	0.988	2.11E-11
582_132	rs4845619	1	154405058	T	G	0.45	0.970	2.12E-11
308_15	rs59741504	1	154504601	T	C	0.21	0.991	2.13E-11
308_15	rs2229238	1	154437896	T	C	0.21	0.993	2.26E-11
201_497	rs11265613	1	154418415	T	C	0.39	0.983	2.33E-11
201_497	rs4845373	1	154417829	C	T	0.39	0.979	2.39E-11
308_15	rs11490956	1	154487258	G	T	0.21	0.997	2.50E-11
308_15	rs111600849	1	154487489	G	A	0.21	0.997	2.50E-11
308_15	rs60498985	1	154504172	C	T	0.21	0.995	2.51E-11
308_15	rs4379670	1	154439865	T	A	0.21	0.986	2.71E-11
308_15	rs12405637	1	154501285	C	T	0.21	0.995	2.75E-11
308_15	rs7514452	1	154438084	C	T	0.21	0.993	2.76E-11
308_15	rs73020232	1	154482669	C	T	0.21	0.997	2.80E-11
308_15	rs73020234	1	154482767	C	A	0.21	0.997	2.80E-11
308_15	rs73020246	1	154485640	A	G	0.21	0.997	2.80E-11
308_15	rs12403537	1	154487726	G	A	0.21	0.997	2.80E-11
308_15	rs77184252	1	154492533	A	G	0.21	0.997	2.80E-11
308_15	rs58348886	1	154446198	T	C	0.21	0.990	2.86E-11
201_497	rs4845622	1	154411419	A	C	0.39	0.970	3.03E-11
308_15	rs58710275	1	154505010	C	T	0.21	0.996	3.07E-11
308_15	rs7551873	1	154505857	C	T	0.21	0.995	3.07E-11
308_15	rs41308419	1	154507197	T	C	0.21	0.996	3.07E-11
308_15	rs112585956	1	154507767	T	C	0.21	0.996	3.07E-11
308_15	rs4633282	1	154507899	T	C	0.21	0.996	3.07E-11
308_15	rs73023331	1	154508210	G	A	0.21	0.996	3.07E-11
308_15	rs111742980	1	154508764	T	A	0.21	0.995	3.07E-11
308_15	rs112674677	1	154508792	C	T	0.21	0.996	3.07E-11
308_15	rs113915076	1	154508812	G	A	0.21	0.996	3.07E-11
308_15	rs4639752	1	154508974	T	A	0.21	0.996	3.07E-11
308_15	rs60368585	1	154509340	T	C	0.21	0.995	3.07E-11
308_15	rs112236096	1	154510790	C	T	0.21	0.995	3.07E-11
308_15	rs73023339	1	154510852	C	T	0.21	0.995	3.07E-11
308_15	rs6674171	1	154491683	A	G	0.21	0.997	3.34E-11

308_15	rs73018293	1	154465577	C	T	0.21	0.997	3.51E-11
308_15	rs59838898	1	154466301	C	T	0.21	0.997	3.51E-11
201_497	rs4129267	1	154426264	C	T	0.38	0.995	3.58E-11
308_15	rs4390168	1	154509096	A	G	0.21	0.992	3.64E-11
308_15	rs10047079	1	154468135	T	C	0.21	0.997	4.19E-11
201_497	rs12730935	1	154419892	G	A	0.38	0.985	4.41E-11
308_15	rs7556449	1	154505886	A	G	0.21	0.996	4.43E-11
308_15	rs12406822	1	154499484	C	T	0.21	0.983	4.48E-11
308_15	rs7526293	1	154444209	T	C	0.21	0.990	4.61E-11
308_15	rs6669229	1	154444591	A	G	0.21	0.990	5.03E-11
201_497	rs61812598	1	154420087	G	A	0.39	0.993	5.49E-11
308_15	rs6675472	1	154445503	T	C	0.22	0.972	5.53E-11
308_15	rs41269915	1	154521584	T	A	0.21	0.986	5.70E-11
308_15	rs57783436	1	154511709	A	G	0.21	0.995	6.30E-11
308_15	rs112394421	1	154512008	T	C	0.21	0.995	6.30E-11
308_15	rs73023346	1	154512077	C	T	0.21	0.995	6.30E-11
308_15	rs12408461	1	154512710	T	C	0.21	0.994	6.30E-11
308_15	rs73023348	1	154513418	G	A	0.21	0.994	6.30E-11
308_15	rs73023349	1	154513440	T	C	0.21	0.994	6.30E-11
308_15	rs72999415	1	154513712	C	T	0.21	0.994	6.30E-11
308_15	rs72999419	1	154514092	T	C	0.21	0.994	6.30E-11
308_15	rs41313910	1	154514203	C	G	0.21	0.994	6.30E-11
308_15	rs72999422	1	154514942	G	A	0.21	0.994	6.30E-11
308_15	rs58881140	1	154515193	C	T	0.21	0.994	6.30E-11
308_15	rs6672087	1	154516122	A	C	0.21	0.994	6.30E-11
308_15	rs6681207	1	154517217	T	C	0.21	0.993	6.30E-11
308_15	rs3811449	1	154517504	G	A	0.21	0.993	6.30E-11
308_15	rs59084843	1	154518286	A	G	0.21	0.993	6.30E-11
308_15	rs61559765	1	154518349	T	C	0.21	0.993	6.30E-11
308_15	rs60092776	1	154518496	A	G	0.21	0.992	6.30E-11
308_15	rs60931472	1	154518642	T	C	0.21	0.992	6.30E-11
201_497	rs12133641	1	154428283	A	G	0.39	0.993	6.65E-11
201_497	rs7518199	1	154407419	A	C	0.39	0.968	7.25E-11
201_497	rs7521458	1	154407713	T	C	0.39	0.969	7.70E-11
201_497	rs4845621	1	154409730	G	A	0.39	0.970	8.80E-11
308_15	rs3811448	1	154516578	G	A	0.21	0.994	9.09E-11
201_497	rs4845620	1	154406656	A	G	0.39	0.969	9.11E-11
201_497	rs56383622	1	154405024	A	G	0.39	0.969	9.97E-11
201_497	rs12126142	1	154425456	G	A	0.39	0.997	1.18E-10
308_15	rs4291493	1	154436920	T	C	0.27	0.919	1.18E-10
323_134	rs11265611	1	154395125	G	A	0.42	0.894	1.41E-10
201_497	rs2228145	1	154426970	A	C	0.39	0.967	1.77E-10
202_497	rs6683206	1	154418088	T	C	0.44	0.969	2.45E-10
202_497	rs11265612	1	154417044	A	G	0.44	0.975	3.03E-10
582_132	rs4845618	1	154400015	G	T	0.45	0.964	3.24E-10
308_15	rs7537291	1	154433407	G	A	0.25	0.990	3.80E-10
308_15	rs7546068	1	154433415	C	T	0.25	0.990	3.80E-10
308_15	rs7537316	1	154433466	G	A	0.25	0.990	3.80E-10
308_15	rs7546552	1	154433905	C	A	0.25	0.992	3.80E-10
308_15	rs7546555	1	154433911	C	T	0.25	0.992	3.80E-10
308_15	rs6687597	1	154434936	G	A	0.25	0.992	3.80E-10
308_15	rs4509570	1	154436384	G	C	0.25	0.994	3.80E-10
308_15	rs4341355	1	154436404	C	G	0.25	0.994	3.80E-10
308_15	rs6698040	1	154432948	T	C	0.25	0.990	4.04E-10
308_15	rs6427672	1	154435346	C	T	0.25	0.992	4.15E-10
308_15	rs4240872	1	154436195	C	T	0.25	0.971	4.22E-10
582_132	rs12117832	1	154397610	A	G	0.45	0.959	4.25E-10
202_497	rs6686750	1	154419843	A	G	0.45	0.990	4.26E-10

582_132	rs12118721	1	154397416	T	C	0.45	0.961	4.38E-10
582_132	rs6687726	1	154400320	A	G	0.45	0.969	4.38E-10
202_497	rs4845625	1	154422067	T	C	0.45	0.991	4.53E-10
308_15	rs10752641	1	154432042	G	C	0.25	0.973	4.60E-10
308_15	rs10908839	1	154430798	C	G	0.24	0.986	5.77E-10
202_497	rs7549338	1	154404380	C	G	0.44	0.970	6.61E-10
582_132	rs11265611	1	154395125	G	A	0.42	0.894	6.66E-10
582_132	rs35109459	1	154397933	G	A	0.47	0.917	6.88E-10
202_497	rs7553796	1	154404406	A	C	0.44	0.969	7.50E-10
202_497	rs6694258	1	154428505	C	A	0.45	0.991	7.96E-10
201_497	rs6684439	1	154395839	C	T	0.39	0.926	8.22E-10
202_497	rs7526131	1	154425135	G	A	0.45	0.993	8.52E-10
202_497	rs12129500	1	154423764	T	C	0.45	0.992	8.54E-10
202_497	rs7536152	1	154423909	A	G	0.45	0.992	8.54E-10
202_497	rs59632925	1	154406540	T	G	0.44	0.970	8.64E-10
202_497	rs4845371	1	154408340	T	C	0.44	0.969	9.26E-10
202_497	rs6689393	1	154426097	A	G	0.45	0.993	1.03E-09
201_497	rs4133213	1	154395212	C	A	0.42	0.893	1.27E-09
308_15	rs55668699	1	154435293	T	A	0.25	0.955	1.29E-09
202_497	rs6427658	1	154400799	T	C	0.45	0.965	1.69E-09
582_132	rs6690230	1	154432877	C	G	0.41	0.990	2.03E-09
582_132	rs6695045	1	154432957	A	G	0.41	0.990	2.03E-09
202_497	rs6694817	1	154401972	T	C	0.44	0.965	2.35E-09
202_497	rs6667434	1	154409100	A	G	0.44	0.969	2.50E-09
202_497	rs6689306	1	154395946	A	G	0.44	0.937	3.10E-09
202_497	rs7549250	1	154404336	C	T	0.45	0.970	3.61E-09
202_497	rs10908838	1	154397984	T	G	0.44	0.963	3.86E-09
202_497	rs12118770	1	154397589	T	C	0.44	0.959	4.31E-09
202_497	rs10908836	1	154397932	C	T	0.46	0.913	4.70E-09
202_497	rs4845619	1	154405058	T	G	0.45	0.970	4.94E-09
582_132	rs10908839	1	154430798	C	G	0.24	0.986	9.75E-09
202_497	rs4553185	1	154410955	C	T	0.45	0.966	1.67E-08
582_132	rs4845647	1	154514331	C	A	0.37	0.995	1.77E-08
202_497	rs12404927	1	154421438	T	C	0.48	0.931	2.20E-08
202_497	rs4845618	1	154400015	G	T	0.45	0.964	2.64E-08
582_132	rs11265632	1	154514067	G	A	0.37	0.995	2.67E-08
582_132	rs4345797	1	154495675	C	T	0.37	0.997	2.69E-08
582_132	rs4845642	1	154498028	G	A	0.37	0.996	2.69E-08
582_132	rs10908845	1	154500898	C	T	0.37	0.996	2.69E-08
582_132	rs11265621	1	154442960	G	A	0.37	0.991	2.80E-08
582_132	rs12119111	1	154478600	G	A	0.37	0.998	2.92E-08
582_132	rs9660786	1	154484017	A	T	0.37	0.998	2.92E-08
582_132	rs7518694	1	154484788	C	G	0.37	0.998	2.92E-08
582_132	rs7519499	1	154487926	G	A	0.37	0.998	2.92E-08
582_132	rs4845639	1	154490352	C	T	0.37	0.998	2.92E-08
582_132	rs4382717	1	154518610	C	T	0.37	0.989	2.95E-08
582_132	rs6698971	1	154475331	C	A	0.37	0.995	3.02E-08
582_132	rs4584384	1	154495697	T	C	0.38	0.996	3.04E-08
582_132	rs6658175	1	154475330	T	A	0.37	0.997	3.30E-08
202_497	rs35109459	1	154397933	G	A	0.47	0.917	3.39E-08
202_497	rs12118721	1	154397416	T	C	0.45	0.961	3.44E-08
202_497	rs6687726	1	154400320	A	G	0.45	0.969	3.44E-08
582_132	rs10908841	1	154487763	C	T	0.37	0.998	3.57E-08
582_132	rs4845637	1	154490178	A	G	0.37	0.998	3.57E-08
582_132	rs5018567	1	154480318	T	C	0.37	0.997	3.68E-08
582_132	rs10908843	1	154494480	C	G	0.38	0.997	4.03E-08
582_132	rs11265628	1	154496489	C	T	0.38	0.997	4.03E-08
582_132	rs12407048	1	154505270	C	T	0.38	0.997	4.03E-08

582_132	rs4638123	1	154505704	C	T	0.38	0.997	4.03E-08
582_132	rs12037271	1	154507888	G	A	0.38	0.996	4.03E-08
202_497	rs12117832	1	154397610	A	G	0.45	0.959	4.08E-08
582_132	rs6700296	1	154473660	T	C	0.37	0.998	4.23E-08
582_132	rs12118018	1	154477440	G	A	0.37	0.999	4.36E-08
582_132	rs6664608	1	154479670	C	T	0.37	0.999	4.36E-08
582_132	rs12740969	1	154487060	T	G	0.37	0.998	4.36E-08
582_132	rs12128408	1	154488533	A	G	0.37	0.998	4.36E-08
582_132	rs4845638	1	154490269	T	A	0.37	0.998	4.36E-08
582_132	rs10908842	1	154492702	G	A	0.37	0.998	4.36E-08
582_132	rs6684921	1	154464945	A	C	0.37	0.997	4.51E-08
582_132	rs6686276	1	154466188	C	A	0.37	0.998	4.51E-08
582_132	rs6689965	1	154470606	T	A	0.37	0.999	4.51E-08
582_132	rs12753680	1	154474900	G	A	0.38	0.998	4.51E-08
582_132	rs7514452	1	154438084	C	T	0.21	0.993	4.64E-08
582_132	rs10908847	1	154504954	G	A	0.37	0.995	4.80E-08
582_132	rs4478801	1	154464572	G	A	0.37	0.994	4.87E-08
582_132	rs12136771	1	154511502	T	C	0.37	0.994	5.02E-08
582_132	rs11265622	1	154451420	A	G	0.37	0.998	5.49E-08
582_132	rs12568083	1	154455949	T	C	0.37	0.998	5.49E-08
308_15	rs6686750	1	154419843	A	G	0.45	0.990	6.23E-08
308_15	rs6427658	1	154400799	T	C	0.45	0.965	6.52E-08
308_15	rs12129500	1	154423764	T	C	0.45	0.992	6.56E-08
308_15	rs7536152	1	154423909	A	G	0.45	0.992	6.56E-08
582_132	rs2229238	1	154437896	T	C	0.21	0.993	6.63E-08
308_15	rs6694258	1	154428505	C	A	0.45	0.991	6.95E-08
308_15	rs6689393	1	154426097	A	G	0.45	0.993	6.99E-08
308_15	rs7526131	1	154425135	G	A	0.45	0.993	7.09E-08
202_497	rs4382717	1	154518610	C	T	0.37	0.989	7.25E-08
202_497	rs12753666	1	154474875	G	A	0.38	0.986	7.75E-08
582_132	rs4379670	1	154439865	T	A	0.21	0.986	8.10E-08
308_15	rs6694817	1	154401972	T	C	0.44	0.965	8.32E-08
308_15	rs12118770	1	154397589	T	C	0.44	0.959	8.87E-08
202_497	rs4845647	1	154514331	C	A	0.37	0.995	8.94E-08
202_497	rs6698971	1	154475331	C	A	0.37	0.995	9.16E-08
202_497	rs5018567	1	154480318	T	C	0.37	0.997	9.51E-08
308_15	rs4845625	1	154422067	T	C	0.45	0.991	9.69E-08
202_497	rs11265621	1	154442960	G	A	0.37	0.991	9.82E-08
308_15	rs6683206	1	154418088	T	C	0.44	0.969	9.84E-08
202_497	rs11265611	1	154395125	G	A	0.42	0.894	9.86E-08
308_15	rs10908838	1	154397984	T	G	0.44	0.963	9.95E-08

[A]For complete Affymetrix probe IDs and further information on probes, see Table S5.
[B]Base pair position build 37/hg19.
[C]Info= Imputation quality metric from SNPtest

Table S6: Correlations between sIL-6R level and *IL6R* expression level for all expression probes.

***IL6R* probe**[A]	**Probe location**	**r / P value**	**Total sample (N=2727)**	**Total sample, correction for rs2228145 (N=2727)**	**Genotype AA (N=991)**	**Genotype AC (N=1278)**	**Genotype CC (N=458)**
545_104	3'-end	r	-0.008	-0.011	0.040	-0.020	-0.008
		P value	0.672	0.567	0.204	0.474	0.872
40_550	3'-end	r	-0.028	-0.019	0.022	0.007	-0.100
		P value	0.139	0.317	0.497	0.799	0.033

410_418	3'-end	r	-0.011	-0.008	0.026	-0.016	-0.075
		P value	0.563	0.668	0.421	0.570	0.111
308_15	3'-end	r	-0.055	-0.003	0.031	-0.043	-0.088
		P value	0.004	0.894	0.335	0.124	0.059
553_369	3'-end	r	-0.023	-0.026	0.011	-0.018	-0.032
		P value	0.234	0.172	0.741	0.514	0.495
372_285	3'-end	r	-0.040	-0.026	0.032	0.024	-0.109
		P value	0.036	0.176	0.319	0.383	0.019
495_603	3'-end	r	-0.057	-0.001	0.026	-0.072	-0.060
		P value	0.003	0.979	0.422	0.010	0.202
61_280	3'-end	r	-0.029	0.008	-0.033	-0.025	-0.074
		P value	0.133	0.681	0.300	0.367	0.113
551_225	3'-end	r	-0.017	0.017	0.041	-0.015	-0.028
		P value	0.386	0.369	0.199	0.596	0.556
390_362	3'-end	r	0.009	0.008	0.047	0.009	-0.012
		P value	0.645	0.680	0.139	0.738	0.802
323_134	Exon 7	r	-0.086	0.001	0.032	0.015	-0.023
		P value	7.48×10^{-6}	0.955	0.312	0.601	0.618
609_718	3'-end	r	-0.010	0.022	-0.004	0.013	-0.082
		P value	0.598	0.247	0.892	0.648	0.079
58_615	3'-end	r	-0.017	-0.011	0.012	0.010	-0.029
		P value	0.366	0.549	0.713	0.722	0.535
46_386	3'-end	r	0.021	0.014	0.050	0.036	0.001
		P value	0.270	0.449	0.118	0.199	0.981
484_717	Exon 8	r	-0.004	0.012	-0.018	0.018	-0.084
		P value	0.820	0.532	0.569	0.525	0.073
659_168	3'-end	r	-0.018	0.012	0.033	0.029	-0.010
		P value	0.345	0.541	0.298	0.297	0.826
201_497	Exon 7	r	-.054	0.011	0.047	0.009	0.002
		P value	0.005	0.577	0.136	0.755	0.974
435_552	3'-end	r	0.016	0.014	-0.004	0.041	-0.003
		P value	0.392	0.454	0.911	0.144	0.948
264_163	Exon 5	r	0.014	-0.004	0.060	0.009	-0.002
		P value	0.473	0.852	0.060	0.751	0.961
582_132	Exon 7	r	-0.092	0.019	0.024	0.005	-0.063
		P value	1.42×10^{-6}	0.314	0.449	0.857	0.178
610_718	3'-end	r	-0.029	0.003	0.008	-0.011	-0.004
		P value	0.133	0.884	0.803	0.694	0.925
483_717	Exon 8	r	0.019	0.006	0.029	0.023	-0.074
		P value	0.317	0.753	0.364	0.407	0.114
235_610	Exon 6	r	-0.010	0.021	0.056	-0.001	-0.052
		P value	0.585	0.276	0.080	0.970	0.263
202_497	Exon 7	r	-0.071	0.005	0.072	-0.016	-0.017
		P value	1.94×10^{-4}	0.791	0.023	0.569	0.724
705_156	3'-end	r	0.017	-0.016	0.021	0.059	-0.020
		P value	0.361	0.418	0.503	0.036	0.673
314_329	Exon 5	r	-0.006	-0.010	0.037	0.005	-0.024
		P value	0.735	0.602	0.249	0.870	0.606
505_200	3'-end	r	-0.014	0.003	0.013	-0.007	-0.106
		P value	0.463	0.880	0.674	0.800	0.023
361_694	3'-end	r	0.001	-0.011	0.038	-0.011	-0.102
		P value	0.947	0.553	0.237	0.691	0.030
586_417	3'-end	r	-0.006	0.017	-0.014	0.026	-0.012
		P value	0.754	0.381	0.669	0.354	0.798
735_334	3'-end	r	-0.027	-0.003	0.021	0.003	-0.135
		P value	0.152	0.875	0.516	0.921	0.004
109_540	3'-end	r	0.016	0.013	0.030	0.022	0.029
		P value	0.412	0.499	0.351	0.435	0.539

260_360	3'-end	r	0.003	-0.022	0.066	-0.014	-0.077
		P value	0.892	0.260	0.038	0.625	0.101
218_13	3'-end	r	0.003	-0.028	0.003	0.037	0.002
		P value	0.865	0.149	0.913	0.188	0.968
267_582	3'-end	r	-0.020	0.001	0.044	-0.033	-0.127
		P value	0.302	0.972	0.169	0.243	0.006
548_21	3'-end	r	0.001	-0.002	0.022	-0.003	-0.037
		P value	0.969	0.903	0.497	0.924	0.429
7_323	3'-end	r	-0.013	0.003	-0.007	0.027	-0.025
		P value	0.498	0.890	0.833	0.326	0.599

[A]For complete Affymetrix probe IDs and further information on probes, see Table S5. r= Pearson correlation between residualized expression level (corrected for technical covariates, age and sex). The last column only is based on levels that were additionally corrected for rs2228145 genotype) and residualized sIL-6R level (corrected for age and sex). P values are two-tailed.

Table S7: SNP information and conditional eQTL *P* values for all 80 significant eQTL associations after correction for the effect of rs2228145.

Probe Name[A]	SNP	SNP Chr	SNP position[B]	Alleles		MAF	Info[C]	eQTL *P* value
308_15	rs60255122	1	154504702	A	G	0.21	0.988	1.89E-09
308_15	rs60760897	1	154504849	C	T	0.21	0.988	1.89E-09
308_15	rs61698846	1	154504854	C	T	0.21	0.988	1.89E-09
308_15	rs61275241	1	154504887	T	G	0.21	0.988	1.89E-09
308_15	rs4072391	1	154438880	T	C	0.21	0.983	2.10E-09
308_15	rs112231452	1	154504900	G	A	0.21	0.988	2.14E-09
308_15	rs61403567	1	154504927	T	C	0.21	0.988	2.17E-09
308_15	rs59741504	1	154504601	T	C	0.21	0.991	2.20E-09
308_15	rs11490956	1	154487258	G	T	0.21	0.997	2.52E-09
308_15	rs111600849	1	154487489	G	A	0.21	0.997	2.52E-09
308_15	rs60498985	1	154504172	C	T	0.21	0.995	2.54E-09
308_15	rs73020232	1	154482669	C	T	0.21	0.997	2.75E-09
308_15	rs73020234	1	154482767	C	A	0.21	0.997	2.75E-09
308_15	rs73020246	1	154485640	A	G	0.21	0.997	2.75E-09
308_15	rs12403537	1	154487726	G	A	0.21	0.997	2.75E-09
308_15	rs77184252	1	154492533	A	G	0.21	0.997	2.75E-09
308_15	rs12405637	1	154501285	C	T	0.21	0.995	2.78E-09
308_15	rs58348886	1	154446198	T	C	0.21	0.990	2.91E-09
308_15	rs58710275	1	154505010	C	T	0.21	0.996	3.04E-09
308_15	rs7551873	1	154505857	C	T	0.21	0.995	3.04E-09
308_15	rs41308419	1	154507197	T	C	0.21	0.996	3.04E-09
308_15	rs112585956	1	154507767	T	C	0.21	0.996	3.04E-09
308_15	rs4633282	1	154507899	T	C	0.21	0.996	3.04E-09
308_15	rs73023331	1	154508210	G	A	0.21	0.996	3.04E-09
308_15	rs111742980	1	154508764	T	A	0.21	0.995	3.04E-09
308_15	rs112674677	1	154508792	C	T	0.21	0.996	3.04E-09
308_15	rs113915076	1	154508812	G	A	0.21	0.996	3.04E-09
308_15	rs4639752	1	154508974	T	A	0.21	0.996	3.04E-09
308_15	rs60368585	1	154509340	T	C	0.21	0.995	3.04E-09
308_15	rs112236096	1	154510790	C	T	0.21	0.995	3.04E-09
308_15	rs73023339	1	154510852	C	T	0.21	0.995	3.04E-09
308_15	rs59239860	1	154492107	G	T	0.21	0.994	3.08E-09
308_15	rs6674171	1	154491683	A	G	0.21	0.997	3.24E-09
308_15	rs73018293	1	154465577	C	T	0.21	0.997	3.39E-09
308_15	rs59838898	1	154466301	C	T	0.21	0.997	3.39E-09
308_15	rs4390168	1	154509096	A	G	0.21	0.992	3.57E-09
308_15	rs2229238	1	154437896	T	C	0.21	0.993	3.76E-09
308_15	rs10047079	1	154468135	T	C	0.21	0.997	3.99E-09

308_15	rs12406822	1	154499484	C	T	0.21	0.983	4.18E-09
308_15	rs7556449	1	154505886	A	G	0.21	0.996	4.23E-09
308_15	rs4379670	1	154439865	T	A	0.21	0.986	4.39E-09
308_15	rs7514452	1	154438084	C	T	0.21	0.993	4.47E-09
308_15	rs7526293	1	154444209	T	C	0.21	0.990	4.55E-09
308_15	rs6675472	1	154445503	T	C	0.22	0.972	4.62E-09
308_15	rs6669229	1	154444591	A	G	0.21	0.990	4.89E-09
308_15	rs41269915	1	154521584	T	A	0.21	0.986	5.22E-09
308_15	rs57783436	1	154511709	A	G	0.21	0.995	5.76E-09
308_15	rs112394421	1	154512008	T	C	0.21	0.995	5.76E-09
308_15	rs73023346	1	154512077	C	T	0.21	0.995	5.76E-09
308_15	rs12408461	1	154512710	T	C	0.21	0.994	5.76E-09
308_15	rs73023348	1	154513418	G	A	0.21	0.994	5.76E-09
308_15	rs73023349	1	154513440	T	C	0.21	0.994	5.76E-09
308_15	rs72999415	1	154513712	C	T	0.21	0.994	5.76E-09
308_15	rs72999419	1	154514092	T	C	0.21	0.994	5.76E-09
308_15	rs41313910	1	154514203	C	G	0.21	0.994	5.76E-09
308_15	rs72999422	1	154514942	G	A	0.21	0.994	5.76E-09
308_15	rs58881140	1	154515193	C	T	0.21	0.994	5.76E-09
308_15	rs6672087	1	154516122	A	C	0.21	0.994	5.76E-09
308_15	rs6681207	1	154517217	T	C	0.21	0.993	5.76E-09
308_15	rs3811449	1	154517504	G	A	0.21	0.993	5.76E-09
308_15	rs59084843	1	154518286	A	G	0.21	0.993	5.76E-09
308_15	rs61559765	1	154518349	T	C	0.21	0.993	5.76E-09
308_15	rs60092776	1	154518496	A	G	0.21	0.992	5.76E-09
308_15	rs60931472	1	154518642	T	C	0.21	0.992	5.76E-09
308_15	rs3811448	1	154516578	G	A	0.21	0.994	8.02E-09
308_15	rs4291493	1	154436920	T	C	0.27	0.919	1.24E-08
308_15	rs7537291	1	154433407	G	A	0.25	0.990	5.23E-08
308_15	rs7546068	1	154433415	C	T	0.25	0.990	5.23E-08
308_15	rs7537316	1	154433466	G	A	0.25	0.990	5.23E-08
308_15	rs7546552	1	154433905	C	A	0.25	0.992	5.23E-08
308_15	rs7546555	1	154433911	C	T	0.25	0.992	5.23E-08
308_15	rs6687597	1	154434936	G	A	0.25	0.992	5.23E-08
308_15	rs4509570	1	154436384	G	C	0.25	0.994	5.23E-08
308_15	rs4341355	1	154436404	C	G	0.25	0.994	5.23E-08
308_15	rs4240872	1	154436195	C	T	0.25	0.971	5.53E-08
308_15	rs6698040	1	154432948	T	C	0.25	0.990	5.53E-08
308_15	rs6427672	1	154435346	C	T	0.25	0.992	5.63E-08
308_15	rs10752641	1	154432042	G	C	0.25	0.973	6.45E-08
308_15	rs10908839	1	154430798	C	G	0.24	0.986	1.07E-07
308_15	rs55668699	1	154435293	T	A	0.25	0.955	1.66E-07

[A]For complete Affymetrix probe IDs and further information on probes, see Table S5.
[B]Base pair position build 37/hg19.
[C]Info= Imputation quality metric from SNPtest

Table S8: Results for the association with asthma for rs60760897 and SNPs in LD in the Australian Asthma Genetics consortium.

SNP	r^2 with rs60760897	Position[A]	A1[B]	A2[C]	MAF	*P*	OR A1[D]	SE[E]
rs60760897	1	154504849	T	C	0.191	0.495	0.96	0.054
rs61275241	1	154504887	G	T	0.190	0.566	0.97	0.055
rs61403567	1	154504927	C	T	0.191	0.514	0.97	0.054
rs61698846	1	154504854	T	C	0.189	0.515	0.97	0.055
rs10047079	0.9787	154468135	C	T	0.191	0.415	0.96	0.054
rs11490956	0.9787	154487258	T	G	0.191	0.400	0.96	0.054
rs12405637	0.9787	154501285	T	C	0.191	0.495	0.96	0.054

rs41308419	0.9787	154507197	C	T	0.189	0.505	0.96	0.055
rs4633282	0.9787	154507899	C	T	0.191	0.511	0.96	0.054
rs4639752	0.9787	154508974	A	T	0.191	0.511	0.96	0.054
rs59838898	0.9787	154466301	T	C	0.191	0.415	0.96	0.054
rs60368585	0.9787	154509340	C	T	0.191	0.511	0.96	0.054
rs6674171	0.9787	154491683	G	A	0.191	0.429	0.96	0.054
rs7551873	0.9787	154505857	T	C	0.191	0.495	0.96	0.054
rs7556449	0.9787	154505886	G	A	0.191	0.496	0.96	0.054
rs4390168	0.9718	154509096	G	A	0.191	0.511	0.96	0.054
rs59741504	0.9652	154504601	C	T	0.191	0.514	0.97	0.054
rs12408461	0.9644	154512710	C	T	0.187	0.604	0.97	0.055
rs3811448	0.9644	154516578	A	G	0.190	0.492	0.96	0.054
rs3811449	0.9644	154517504	A	G	0.190	0.511	0.96	0.054
rs41313910	0.9644	154514203	G	C	0.191	0.510	0.96	0.054
rs57783436	0.9644	154511709	G	A	0.191	0.493	0.96	0.054
rs58881140	0.9644	154515193	T	C	0.190	0.510	0.96	0.054
rs61559765	0.9644	154518349	C	T	0.190	0.511	0.96	0.054
rs6672087	0.9644	154516122	C	A	0.190	0.511	0.96	0.054
rs6681207	0.9644	154517217	C	T	0.190	0.511	0.96	0.054
rs58348886	0.9642	154446198	C	T	0.191	0.383	0.95	0.054
rs6669229	0.9572	154444591	A	G	0.190	0.363	0.95	0.055
rs7526293	0.9572	154444209	T	C	0.191	0.401	0.96	0.054
rs41269915	0.9503	154521584	A	T	0.190	0.511	0.96	0.054
rs59239860	0.9223	154492107	T	G	0.191	0.400	0.96	0.054
rs4379670	0.8945	154439865	T	A	0.185	0.506	0.96	0.055
rs2229238	0.8874	154437896	T	C	0.179	0.602	0.97	0.056
rs4072391	0.8874	154438880	T	C	0.179	0.611	0.97	0.056
rs7514452	0.8803	154438084	C	T	0.180	0.514	0.96	0.056
rs10752641	0.7052	154432042	G	C	0.216	0.241	0.94	0.052
rs35699331	0.7052	154432904	T	C	0.222	0.122	0.92	0.052
rs6698040	0.7052	154432948	T	C	0.216	0.217	0.94	0.052
rs7537316	0.7052	154433466	G	A	0.217	0.246	0.94	0.052
rs7546068	0.7052	154433415	C	T	0.217	0.240	0.94	0.052
rs4240872	0.7029	154436195	C	T	0.220	0.266	0.94	0.052
rs4341355	0.7029	154436404	C	G	0.219	0.273	0.94	0.052
rs4509570	0.7029	154436384	G	C	0.219	0.273	0.94	0.052
rs7537291	0.7008	154433407	G	A	0.217	0.240	0.94	0.052
rs10908839	0.6589	154430798	C	G	0.214	0.267	0.94	0.052
rs55668699	0.5989	154435293	A	T	0.215	0.386	0.96	0.052
rs12128408	0.4703	154488533	A	G	0.357	0.165	0.94	0.045
rs11265622	0.4693	154451420	A	G	0.357	0.153	0.94	0.045
rs10908843	0.4669	154494480	C	G	0.357	0.210	0.95	0.045
rs12753680	0.4663	154474900	G	A	0.357	0.165	0.94	0.045
rs10908841	0.4605	154487763	C	T	0.357	0.165	0.94	0.045
rs10908842	0.4605	154492702	G	A	0.357	0.165	0.94	0.045
rs10908847	0.4605	154504954	G	A	0.357	0.226	0.95	0.045
rs11265628	0.4605	154496489	C	T	0.357	0.210	0.95	0.045
rs12037271	0.4605	154507888	G	A	0.357	0.226	0.95	0.045
rs12118018	0.4605	154477440	G	A	0.357	0.165	0.94	0.045
rs12119111	0.4605	154478600	G	A	0.357	0.165	0.94	0.045
rs12407048	0.4605	154505270	C	T	0.352	0.173	0.94	0.045
rs12568083	0.4605	154455949	T	C	0.357	0.153	0.94	0.045
rs12740969	0.4605	154487060	T	G	0.357	0.165	0.94	0.045
rs4345797	0.4605	154495675	C	T	0.357	0.210	0.95	0.045
rs4478801	0.4605	154464572	G	A	0.357	0.154	0.94	0.045
rs4584384	0.4605	154495697	T	C	0.357	0.229	0.95	0.045
rs4638123	0.4605	154505704	C	T	0.357	0.226	0.95	0.045
rs4845638	0.4605	154490269	T	A	0.357	0.165	0.94	0.045

rs4845639	0.4605	154490352	C	T	0.357	0.165	0.94	0.045
rs4845642	0.4605	154498028	G	A	0.357	0.226	0.95	0.045
rs5018567	0.4605	154480318	T	C	0.357	0.193	0.94	0.045
rs6664608	0.4605	154479670	C	T	0.357	0.165	0.94	0.045
rs6684921	0.4605	154464945	A	C	0.358	0.159	0.94	0.045
rs6686276	0.4605	154466188	C	A	0.357	0.165	0.94	0.045
rs6700296	0.4605	154473660	T	C	0.358	0.160	0.94	0.045
rs7518694	0.4605	154484788	C	G	0.356	0.216	0.95	0.045
rs7519499	0.4605	154487926	G	A	0.357	0.165	0.94	0.045
rs9660786	0.4605	154484017	A	T	0.357	0.165	0.94	0.045
rs6658175	0.4593	154475330	T	A	0.357	0.165	0.94	0.045
rs11265632	0.4483	154514067	G	A	0.357	0.226	0.95	0.045
rs12136771	0.4483	154511502	T	C	0.358	0.226	0.95	0.045
rs4845647	0.4483	154514331	C	A	0.357	0.226	0.95	0.045
rs10908845	0.4474	154500898	C	T	0.357	0.226	0.95	0.045
rs12753666	0.4448	154474875	G	A	0.356	0.172	0.94	0.045
rs11265621	0.4437	154442960	G	A	0.357	0.159	0.94	0.045
rs6690230	0.3532	154432877	C	G	0.383	0.165	0.94	0.044
rs6695045	0.3532	154432957	A	G	0.383	0.165	0.94	0.044
rs11265611	0.3095	154395125	G	A	0.411	0.184	0.94	0.044
rs4845625	0.3019	154422067	T	C	0.415	0.156	0.94	0.043
rs6689393	0.3019	154426097	A	G	0.413	0.173	0.94	0.043
rs4553185	0.3015	154410955	C	T	0.414	0.107	0.93	0.043
rs4845619	0.3015	154405058	T	G	0.414	0.124	0.94	0.043
rs7549250	0.3015	154404336	C	T	0.414	0.123	0.94	0.043

The Australian Asthma Genetics Consortium (AAGC) cohorts and methods are described in[1].
[A]Base pair position build 37/hg19. [B]Allele 1 (effect allele). [C]Allele 2. [D]Odds ratio associated with allele 1.[E]SE = Standard error of the odds ratio

Table S9: Results for the association with asthma for SNPs in LD with rs60760897 in the GABRIEL consortium.

SNP	r^2 with rs60760897	Position[A]	A1[B]	A2[C]	OR A1[D]	SE[E]	*P*
rs7556449	0.979	154505886	G	A	1.01	0.025	0.72
rs3811448	0.964	154516578	G	A	0.99	0.025	0.73
rs7514452	0.880	154438084	T	C	1.00	0.025	0.99
rs4240872	0.703	154436195	T	C	1.00	0.024	0.94
rs4584384	0.460	154495697	T	C	0.95	0.021	0.01
rs4478801	0.460	154464572	G	A	0.95	0.021	0.01
rs4553185	0.302	154410955	T	C	1.04	0.020	0.07

For a description of the GABRIEL consortium cohorts and methods, see[2].
[A]Base pair position build 37/hg19
[B]Allele 1 (effect allele) [C]Allele 2 [D]Odds ratio associated with allele 1
[E]SE = Standard error of the odds ratio

Supplementary Note

Subjects

The twin family and biometrical model analyses were based on data from 4980 individuals from 2360 NTR families, including 3083 twins, 236 of their non-twin brothers (up to two per family), 453 non-twin sisters (up to 2 per family), 657 of their mothers and 551 fathers (non-biological siblings, non-biological parents, spouses of twins and other types of family members were excluded). The GWAS of sIL-6R was performed on data from 4846 NTR participants (including all types of familial relations except for MZ twin pairs). The eQTL analysis included subjects with genome-wide SNP and expression data from the NTR (N=2977) and from the Netherlands Study of Depression and Anxiety (NESDA[3], N=1966). Correlations between sIL-6R level and *IL6R* expression were

examined in a subset of NTR subjects with data on sIL-6R, expression data and genome-wide SNP data (N=2727). The combined analysis of linkage and association was performed using data from 355 nuclear families from the NTR (total N subjects=1254), consisting of two parents and one child (213 families), two children (98 families), three children (43 families) or five children (1 family). The GCTA analysis of variance explained by chromosome-wide SNPs was performed on data from NTR participants including all types of familial relations except for MZ twin pairs (N=4846, same subjects as included in GWAS), and on data from a subset of unrelated NTR subjects (N=2875). Individuals using anti-inflammatory medication or medication influencing the HPA (Hypothalamic-Pituitary-Adrenal)-axis (NTR: N = 426/4.8%, NESDA: N=538/26%), or with a sIL-6R level > 100.000 pg/mL (N=6/0.07% of NTR subjects) at the time of blood sampling were excluded from all analyses. Individuals of non-Dutch origin identified based on genotype data[4] (N= 455, 3.8 % of the entire sample of 12,133 genotyped individuals from NTR and NESDA), were excluded from all SNP-based analyses (GWA, GCTA , eQTL and combined linkage and association analysis) to rule out potential effects of population stratification. Characteristics of the subjects included in each analysis are described in Table S1.

sIL-6R level

sIL-6R levels were measured in pg/mL. For the extended twin-family model and linkage analyses, sIL-6R level (pg/mL) was divided by 10.000 (corresponding to 10^{-8} g/mL) while retaining all decimals to meet the computational demands of the software used. For practical reasons, the variance due to rs2228145 and biometrical model estimates (Table 2) are also displayed in 10^{-8} g/mL. The original sIL-6R values in pg/mL were used as input for all other analyses.

Medication use

Following the ATC-coding system, anti-inflammatory medication was defined as all L-class (antineoplastic and immunomodulating) medication and M01 class medication (anti-inflammatory and antirheumatic medication). Medication influencing the HPA-axis was defined as medication in class H01 (pituitary and hypothalamic hormones and analogues) and class H02 (corticosteroids for systemic use). Following the ATC-coding system, anti-inflammatory medication was defined as all L-class (antineoplastic and immunomodulating) medication and M01 class medication (anti-inflammatory and antirheumatic medication). Medication influencing the HPA-axis was defined as medication in class H01 (pituitary and hypothalamic hormones and analogues) and class H02 (corticosteroids for systemic use).

Pre-imputation QC

Quality control was done within and between chip platforms. For each platform the individual SNP markers were lifted over to build 37 (HG19) of the Human reference genome, using the LiftOver tool (“http://genome.sph.umich.edu/wiki/LiftOver”). SNPs that were not mapped at all, SNPs that had ambiguous locations, and SNPs that did not have matching - or strand opposite alleles were removed. Subsequently, the data were strand aligned with the 1000 Genomes phase I Interim release ALL panel of 23 November 2010 (sequence) and June 2011 (haplotypes) prepared by J. Marchini & B. Howie for IMPUTEv2.1
(https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_interim.html).

SNPs from each platform were removed if they still had mismatching alleles with this imputation reference set, if the allele frequencies differed more than 0.20 with the reference set, if the MAF was < 1%, if the HWE p-value was < 0.00001 or if the call rate was <95%. All samples were excluded from the data if their expected sex did not match their genotyped sex, if the genotype missing rate was above 10% or if the Plink F inbreeding value was either > 0.10 or < -0.10 (heterozygosity). After these steps the data of the individual chips were merged into a single dataset using the Plink 1.07 software.
Within the merged set IBD was calculated between all possible individual pairs and compared to the expected family structure of the NTR and NESDA studies. Samples were removed if the data did not match the expected IBD sharing, or if potentially consistent with biographic data, corrections were made to the family structure. DNA samples that were typed on multiple platforms were tested if the overlapping SNPs had a concordance rate above 99.0% and if not, all data of these samples were removed. On the merged data, the HWE and MAF SNP filters were re-applied, as well as the reference allele frequency difference <0.20 checks. As a final prior step to imputation SNPs with C/G and A/T allele combinations were removed if the MAF was between 0.35 and 0.50 to avoid wrong strand alignment for these SNPs.

Imputation
Genome wide SNP imputation was done with the IMPUTE 2.1.2 program for the autosomal genome using the above phase I integrated reference panel. A total number of 10,726 unique samples (for MZ twin pairs data imputation was done including one twin) were imputed in batches of around 500 individuals for 5 million base blocks, a 250 kb buffer and the NE parameter set to 20.000. Monomorphic SNPs present in the 500 individual subsets were removed. To avoid issues having SNPs from different platforms partly imputed and partly genotyped we re-imputed all genotyped SNPs. After imputation of these SNPs, we generally find a high concordance between re-imputed SNPs and the original genotype (0.9868), if the SNP info or R2 is above 30%.

Post-imputation QC
After imputation, additional QC included evaluation of the SNP platforms effect, the Mendelian error rate in families and filters for HWE, imputation quality and MAF. First we tested the effect of having different platforms imputed, and we removed SNPs showing platform effects. This was done, defining an individual with a specific platform as a case and the remaining individuals as controls. Allelic association was then calculated and SNPs were removed if the specific platform allele frequencies were significantly different from the remaining platforms with p < 0.00001. Subsequently, HWE was calculated on the SNPTEST allele probability counts for the full sample and SNPs were removed if the p-value < 0.000001. The Mendelian error rate was calculated on the best guess genotypes in families (trios and sib-pairs with parents) using first Gtool and then Plink 1.07. SNPs were removed if the Mendelian error rate was above 2%. SNPs were subsequently filtered from the dataset if the imputation quality R2 < 0.30 (mean is 0.85) and if the MAF in the total sample was less than 0.004. Finally, all SNPs were removed that were by then known to show problems in the reference sequence data of the phase I Interim set (159k).

Statistical analysis

Analyses of genotype data

All analyses were performed on imputed data from autosomal SNPs. The heritability analysis in twin families, biometrical model, GCTA analyses, linkage analyses and eQTL analyses were performed using best guess genotypes, and the GWAS was performed on genotype dosage data (using only SNPs with no missing data). SNPs with a mendelian error rate > 2 % were removed during post-imputation QC. For linkage analysis, all mendelian errors were removed from the data to meet software demands, using the option --me 1 1 --set-me-missing in PLINK[5], which sets both parental and offspring genotype to missing when a mendelian error is encountered (rather than removing the entire family or SNP). In all analyses of genome-wide SNP data (GWA, eQTL, combined linkage and association, and GCTA analyses), covariates for 6 principle components (PCs) were included: 3 PCs to correct for population structure within the Dutch population, 1 PC for effects of the plate and source of DNA (blood vs buccal) and 2 PCs reflecting the effects of genotype array[4].

Observed variance due to rs2228145 genotype

The confidence interval around the observed variance due to rs2228145 genotype (V_{SNP}, Table 2) was estimated as follows: Confidence interval (CI) lower value = V_{SNP} *((N-1) /lower critical value)= 0.71 * ((4980-1)/ 5176.472= 0.68, and upper value= V_{SNP} *((N-1) /upper critical value)= 0.71 * ((4980-1)/ 4785.317=0.74, where lower critical value corresponds to (1- alpha)/2 from a chi-squared distribution with df =N-1=4979, and upper critical value corresponds to alpha/2, from a chi-squared distribution with df =N-1=4979.

Heritability of sIL-6R level based on twin family data

Models were fitted in Mx using raw-data maximum likelihood[6]. We used the model described by Neale *et al*[7], for mono- and dizygotic twins, siblings and parents. For each model, Mx computes a goodness-of-fit statistic based on minus twice the logarithm of the likelihood (-2LL). Comparison of sub-models was done by means of likelihood-ratio tests; the ratio of likelihoods of sub-models follows a X^2 distribution with degrees of freedom (df) equal to the difference in the number of parameters of the two models. If a more restricted model fits the data significantly worse, the more general model is preferred.

In a so-called saturated model, familial correlations of sIL-6R level were estimated while taking age and sex effects on sIL-6R level into account. Next, the covariance structure among family members was modeled as a function of additive genetic variance (variance due to A, or V_A), non-additive genetic variance (variance due to D, or V_D), sibling-shared environment environmental variance (variance due to C, or V_C) and unique environmental variance (variance due to E, or V_E)[8, 9]. MZ twins derive from a single fertilized egg cell (zygote) and therefore share (nearly) 100% of their genetic material. DZ twins (who are derived from two zygotes) and non-twin sibling pairs share on average 50% of their segregating genetic material. Parents and offspring always share half of their segregating genetic material. Siblings and both types of twins share 100% of V_C. V_E comprises all variation that is not shared among family members. Thus, familial correlations in sIL-6R level may be represented as:

$r = a^2 + d^2 + c^2$ for MZ twins, $r = ½\ a^2 + ¼\ d^2 + c^2$ for DZ twins and non-twin sib pairs, and $r = ½\ a^2$ for Parent-offspring;
Where a^2= proportion of the variance due to additive genetic effects (V_A / V_P, also called narrow-sense heritability or h^2), d^2= proportion of the variance due to non-additive genetic effects (V_D / V_P), c^2= proportion of the variance due to sibling-shared environment (V_C / V_P), and V_P=Total (phenotypic) variance of sIL-6R level = $V_A + V_D + V_C + V_E$. We additionally fitted models that incorporated assortative mating (results not shown), which gave highly similar estimates of variance components.

Variance explained by chromosome-wide SNPs
The variance of sIL-6R level that can be explained by all imputed SNPs was estimated using GCTA (Genome-wide Complex Trait Analysis[10, 11]). GCTA uses the genotype data to estimate the fraction of total DNA that two individuals share, which is summarized in a genetic relationship matrix (GRM). Next, a linear model is fitted in which the measure of genetic relatedness is used to predict phenotypic similarity of individuals, giving an estimate of the variance explained by all SNPs. The analyses in GCTA were performed using unrelated subjects only (excluding individuals from pairs with an estimated genetic relatedness > 0.025 (corresponding to third or fourth cousins) (1) and with inclusion of close relatives (including DZ twins, siblings and parent-offspring pairs) (2). An advantage compared to traditional methods of estimating heritability (e.g. the twin-family model) is that the SNP-based method in GCTA allows to assess the contribution of specific genomic regions (e.g. chromosomes) to the total heritability and does not require assumptions about IBD-sharing among family members.

GRMs were estimated separately for each chromosome (and for chromosome 1 separately for the *IL6R* gene +/-10MB vs all other SNPs), and GRMs for all chromosomes were fitted simultaneously in the regression model to estimate the variance attributable to each chromosome. SNPs with a MAF lower than 0.001 were excluded. To estimate the variance explained by the *IL6R* region after taking out the effect of rs2228145, a regression model was fitted that included the GRM for the *IL6R* region plus a covariate for rs2228145 (with genotype coded as 0, 1, 2 – corresponding to the number of minor alleles).

Supplementary Reference List
1. Ferreira,M.A. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006-1014 (2011).
2. Moffatt,M.F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211-1221 (2010).
3. Penninx,B.W. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17**, 121-140 (2008).
4. Abdellaoui,A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.*doi: 10.1038/ejhg.2013.48 (2013).
5. Purcell,S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).
6. Neale,M.C., Boker,S.M., Xie,G., & Maes,H.H. *Mx: Statistical Modeling.*(Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, 2006).
7. Neale,M.C., Walters,E.E., Eaves,L.J., Maes,H.H., & Kendler,K.S. Multivariate genetic analysis of twin-family data on fears: Mx models. *Behav. Genet* **24**, 119-139 (1994).

8. Distel,M.A. *et al.* Familial resemblance of borderline personality disorder features: genetic or cultural transmission? *PLoS One* **4**, e5334 (2009).
9. Falconer,D.S. *Introduction to quantitative genetics*(Ronald Press Co, New York, 1960).
10. Yang,J., Lee,S.H., Goddard,M.E., & Visscher,P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).
11. Yang,J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519-525 (2011).

Appendix 2

Supplement to Chapter 4

Supplemental Methods

MZ twins are derived from a single fertilized egg cell (zygote) and therefore share (nearly) 100% of their segregating genes. DZ twins (who are derived from two zygotes) and non-twin siblings share on average 50% of their segregating genes. Parents share 50% of their genetic material with their children. Common environmental influences may be shared by siblings ("sibling-shared environmental effects") or by individuals who share a household ("shared household effects"). In our study, the parents of twins who live together may share household effects, but household effects are not likely to contribute largely to the similarity of siblings and of parents with their offspring because the offspring in our study are adults, thus most live on their own. All family members share 0% of V_E by definition. Assuming that the variation of a trait is caused by A, D, C (sibling-shared) and E ($V_P = V_A + V_{D.} + V_C + V_E$), the phenotypic covariance of family members can be expressed as a function of V_A, V_D, V_C and V_E (i.e. covariance MZ twins= $V_A + V_D + V_{C,}$ covariance DZ twins and sibs=1/2 V_A + 1/4 V_D+ V_C, and covariance parent-offspring=1/2 V_A). Correlations, which represent a standardized measure of the shared variance between variables, are also frequently used to summarize the similarity of family members. Following from the equations for the covariances explained above, the phenotypic correlations among relatives can be written as:

Phenotypic correlation MZ twins = $h^2 + d^2 + c^2$

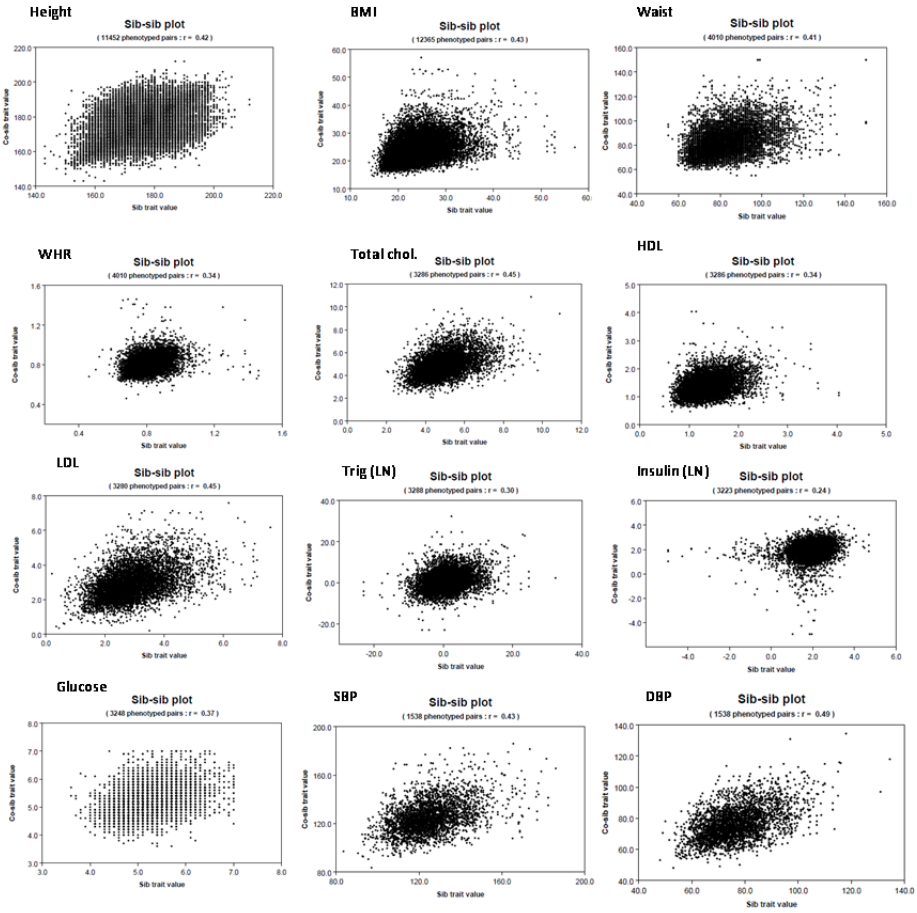
Phenotypic correlation Parent-offspring = ½ h^2

Phenotypic correlation DZ twin / siblings = ½ h^2 + ¼ d^2 + c^2

Where h^2= proportion of variance due to additive genetic effects (V_A / V_P), d^2= proportion of variance due to non-additive genetic effects (V_D/ V_P), c^2 =proportion of variance due to sibling shared environment (V_C / V_P).

Prior exploratory analyses of the twin family data in Mx indicated the absence of sibling-shared environmental effects for the vast majority of traits in our study; these effects were therefore not included in further analyses. Sex or age differences in V_A , $V_{D.}$,V_C or V_E are indicated by differences in the pattern of phenotypic correlations in male versus female relatives ("quantitative sex difference") or young versus old relatives ("quantitative age difference"). A smaller correlation in opposite-sex relatives compared to same-sex relatives indicates that different factors (e.g. different genes or different environmental factors) contribute to variation in males and females ("qualitative sex difference"). Similarly, if the phenotypic correlation decreases with increasing age difference between relatives, this is an indication that different factors may play a role at different ages ("qualitative age difference").

Supplemental Figure 1: Sib-pair resemblance for metabolic syndrome traits. The figure shows scatterplots of trait values for all sib-sib relations in the sample (including DZ twins).



Supplemental Table 1: Numbers and age of twins and family members

Descriptives total sample				*Descriptives of twins by zygosity*			
	N	Age Men (SD)	Age women (SD)		N twins	Total N	Age (SD)
BMI							
Twins	12877 (*M*: 4666, *F*:8211)	28.79 (11.93)	30.71 (12.84)	MZM	1660	1901	29.83 (12.51)
Sisters	2065 (*first*: 1759, *second*: 306)		35.32 (12.87)	DZM	1088	1326	28.04 (11.87)
Brothers	1550 (*first*: 1333, *second*: 217)	33.86 (12.88)		MZF	3656	4060	32.60 (13.68)
Parents	10529 (*Mo*: 6036; *Fa*:4493)	54.81 (8.37)	52.24 (8.86)	DZF	2020	2345	29.90 (12.09)
Total	27021 (*M:* 10709, *F:* 16312)	40.44 (16.35)	39.26 (15.30)	DOS	2462	3245	27.77 (11.02)
Waist, WHR							
Twins	4305 (*M*: 1496 , *F:* 2809)	33.95 (11.62)	35.45 (11.58)	MZM	568	684	35.17 (12.66)
Sisters	869 (*first*: 725, *second*: 144)		41.35 (12.88)	DZM	286	395	32.99 (11.20)
Brothers	575 (*first:* 500, *second*: 75)	40.03 (13.72)		MZF	1278	1514	36.52 (11.92)
Parents	3216 (*Mo*: 2052, *Fa*: 1164)	61.57 (7.34)	55.87 (10.60)	DZF	670	818	34. 89 (11.51)
Total	8965 (*M*: 3235, *F:* 5730)	44.97 (16.58)	43.66 (14.78)	DOS	630	894	32.94 (10.00)
Glucose and insulin							
Twins	3724(*M*: 1262 , *F:* 2462)	33.95 (10.99)	35.84 (11.44)	MZM	398	579	35.07 (11.93)
Sisters	1143 (*first*: 818, *second*: 325)		42.12(10.58)	DZM	206	325	33.29(11.05)
Brothers	470 (*first:* 416 *second*: 54)	39.69 (13.12)		MZF	1018	1366	36.96 (11.80)
Parents	2226 (*Mo*: 1263, *Fa*: 963)	61.88 (7.05)	60.15(7.57)	DZF	526	695	35.12 (11.34)
Total	7563 (*M*: 2697, *F:* 4866)	44.94 (16.38)	43.61 (14.47)	DOS	498	759	33.01 (9.40)
Lipids							
Twins	3721 (*M*: 1273 , *F*: 2448)	33.92 (11.02)	35.56 (11.16)	MZM	394	583	34.87 (11.89)
Sisters	1185 (*first*: 842, *second*: 343)		41.93 (10.47)	DZM	212	329	33.49 (11.32)
Brothers	480 (*first*: 424, *second*: 56))	39.46(12.66)		MZF	1004	1352	36.61 (11.57)
Parents	2083 (*Mo*: 1210, *Fa*: 873)	61.62 (6.98)	60.26 (7.70)	DZF	526	694	32.91 (11.00)
Total	7469 (*M*:2628 , *F*: 4841)	44.15 (16.14)	43.28 (14.39)	DOS	502	763	32.95 (9.23)
Blood pressure							
Twins	2002 (*M*: 765 , *F*: 1237)	32.65 (10.46)	32.83 (10.72)	MZM	388	416	32.53 (9.83)
Sisters	243 (*first*: 213, *second*: 30)		35.23 (11.81)	DZM	196	213	32.38 (11.42)
Brothers	185 (*first*: 166, *second*: 19)	33.85 (12.02)		MZF	704	760	33.24 (10.72)
Parents	362 (*Mo*: 185, *Fa*: 177)	47.68 (6.33)	45.28 (5.70)	DZF	284	324	31.88 (10.66)
Total	2792 (*M*: 1127 , *F*: 1665)	35.21 (11.54)	34.57 (11.15)	DOS	238	289	33.10 (10.78)

M= Male, F=Female, Mo=Mother, Fa=Father, N=Number of subjects

Supplemental Table 2: Twin correlations for metabolic syndrome traits stratified by age and sex.

	Height	**BMI**	**Waist**	**WHR**	**TC**	**HDL**	**LDL**	**Trig**	**Insulin**	**Glu-cose**	**SBP**	**DBP**
Young twins												
MZ twins												
MZ males	0.91	0.73	0.65	0.48	0.78	0.79	0.74	0.72	0.40	0.54	0.59	0.63
MZ females	0.90	0.72	0.66	0.41	0.70	0.70	0.74	0.57	0.39	0.53	0.57	0.54
DZ twins												
DZ males	0.55	0.58	0.47	0.50	0.43	0.36	0.43	0.32	0.26	0.42	0.48	0.25
DZ females	0.54	0.33	0.20	0.26	0.41	0.38	0.43	0.12	0.09	0.35	0.56	0.49
DZ opposite	0.39	0.39	0.40	0.24	0.07	0.14	0.07	0.26	0.34	0.39	0.41	0.40
Older twins												
MZ twins												
MZ males	0.89	0.75	0.76	0.57	0.67	0.58	0.64	0.69	0.56	0.43	0.67	0.69
MZ females	0.88	0.70	0.68	0.51	0.71	0.58	0.71	0.57	0.44	0.50	0.64	0.68
DZ twins												
DZ males	0.42	0.38	0.40	0.36	0.17	0.44	0.11	0.25	0.15	0.20	0.10	0.17
DZ females	0.55	0.32	0.42	0.38	0.47	0.34	0.40	0.30	0.22	0.35	0.50	0.48
DZ opposite	0.57	0.19	0.27	0.34	0.47	0.34	0.41	0.2	0.25	0.25	0.22	0.26

Age categories were defined based on the median age of twins (young twins=twins with an age below the median and older twins=twins with an age above the median). The median age of twins was 32 years for all metabolic syndrome variables. For height only, birth year (median=1973) instead of age was used to define categories. TC= Total cholesterol. Trig= Triglycerides.

Supplemental Table 3: Age-specific heritability estimates of metabolic syndrome traits and height.

	Height	BMI	Waist	WHR	TC	HDL	LDL	Trig	Insu-lin	Glu-cose	SBP	DBP
Young												
V_P	42.4	11.92	104.68	0.004	0.81	0.12	0.65	20.5	0.51	0.25	132.6	92.72
a^2	0.84	0.33	0.37	0.29	0.39	0.36	0.42	0.25	0.23	0.58	0.61	0.62
d^2	0.07	0.47	0.38	0.16	0.28	0.27	0.27	0.32	0.25	0.00	0.00	0.00
H^2	0.91	0.79	0.75	0.45	0.67	0.63	0.69	0.57	0.48	0.58	0.61	0.62
SE	0.03	0.04	0.06	0.10	0.06	0.08	0.06	0.06	0.06	0.06	0.13	0.12
Old												
V_P	43.4	16.15	135.02	0.01	1.07	0.15	0.90	22.0	0.50	0.33	211.2	120.7
a^2	0.83	0.34	0.41	0.43	0.40	0.53	0.41	0.34	0.23	0.48	0.61	0.64
d^2	0.07	0.34	0.30	0.13	0.21	0.21	0.19	0.30	0.26	0.00	0.00	0.00
H^2	0.90	0.68	0.70	0.56	0.61	0.74	0.60	0.63	0.49	0.48	0.61	0.64
SE	0.03	0.03	0.05	0.10	0.06	0.07	0.05	0.06	0.07	0.05	0.08	0.10
r_A Young-Old	0.96	1.00	1.00	0.85	1.00	0.90	1.00	1.00	1.00	0.64	0.24	0.56

The estimates were obtained from a model in which V_P, V_A and V_E were allowed to differ between young and old subjects, whereas V_D was constant. Age categories were defined based on the median age of subjects (young=subjects with an age below the median and old=subjects with an age above the median). The median age was 35 years for SBP and DBP, 39 years for BMI, 40 years for HDL, LDL, total cholesterol and triglycerides, and 41 for insulin, glucose, waist and WHR. For height only, birth year (median=1966) instead of age was used to define categories. TC= Total cholesterol. Trig= Triglycerides. SE=Robust standard error of the broad heritability estimate from poly. r_A Young-Old=Correlation between additive genetic effects in young and old subjects.

Appendix 3

Supplement to chapter 5

Supplemental Methods

1. Subjects

MZ twins were registered with the Netherlands Twin Register (NTR). The majority of twins was recruited through City Councils between 1990 and 1993 when they were adolescents or young adults. Zygosity assessment was based on DNA markers for 55% of all MZ pairs, for 91.4 % of MZ pairs included in the analysis of blood biomarker profiles and for 96.6% of MZ pairs included in the analysis of gene expression profiles. For the remaining pairs, zygosity was based on validated survey items or blood groups.

2. NTR survey studies

2.1 Quality checks BMI data

For subjects who completed multiple surveys at age 18 or older, height values were averaged to obtain one measure of adult body height if the difference in height reported by an individual across time did not exceed 2 cm. If different self-reports differed by 3 cm or more, the most deviating report was removed and the remaining values were averaged if the difference between remaining values was less than 3 cm. If the difference in height reported by an individual at different surveys was 3 or 4 cm after removal of two outlier values, remaining values were averaged to obtain one measure of adult height. Height data were considered unreliable and therefore not used if the difference in height reported by an individual at different surveys still exceeded 4 cm after removal of two outlier values. Height and weight assessed in lab-based studies were used to validate survey-based data on height and weight.

2.2 Quality checks birth weight data

Data on birth weight were collected as part of multiple NTR surveys and projects. Data reported by the twins themselves and / or by their parents were combined and consistency across family members and time was checked. When multiple data points differed, the average was taken, but only if the difference was smaller than 200 grams.

2.3 Lifestyle characteristics

Data on eating habits and leisure time exercise activities were obtained from the eighth NTR survey study (2009). Eating habits were assessed by the following questions: 1. Did you ever go on a diet to lose weight or to avoid gaining weight? 2. How scared are you to gain weight or become fat? 3. How fast do you normally eat? 4. Do you normally eat until you feel full? 5. Who of you eats most (you and your co-twin)? Exercise activities were assessed with the following questions: 1. Do you participate in leisure time exercise activities regularly? (yes/no) 2. How often do you participate in these activities (4 answer categories, ranging from “almost never” to > 10 hours per week)?

Current smoking status was assessed in all NTR surveys and as part of the NTR biobank project. To assess whether changes in smoking status may contribute to changes in BMI discordance over time, individuals were grouped into four categories based on their smoking status at two time points (the time point of first discordance and at the first next time point, ≥ 1 year later): 1.“non-smoker” (non-smoker at both time points), 2.“started smoking” (non-smoker at first time point, smoker at the second time

point), 3.“smoker” (smoker at both time points), and quit smoking (“smoker at first time point and non-smoker at second time point).

For further information about lifestyle data obtained through surveys, we refer the reader to previous papers that reported on these variables: eating [1-4], smoking [5-7], leisure time exercise participation [8, 9].

3. NTR biobank project

3.1 Study protocol and measures

In total, 9992 subjects participated in the NTR biobank project, including monozygotic and dizygotic twins and their family members. For a detailed description of the biobank study protocol, see Willemsen et al[10]. In short, venous blood samples were drawn in the morning after an overnight fast and usually in the subjects' home, to measure blood biomarker profiles and gene expression. Additionally, information on height, weight and a range of health-related variables such as medication use and smoking status were obtained. In the current study, data on lipid-lowering medication (user/non-user) and medication for diabetes (user/non-user) are analyzed. Female twins were asked if they were pregnant or if they had entered menopause at the moment of blood draw. In the current analyses, cigarette smoking status at blood draw was assessed by two variables: current-smoking (current smoker/ non-smoker) and smoking ever (ever-smoked/never-smoked).

3.2 Blood biomarkers

Total cholesterol, high-density lipoprotein cholesterol (HDL) and triglyceride levels were measured in heparin plasma and low-density lipoprotein cholesterol (LDL) was calculated with the Friedewald Equation [11]. Glucose and insulin were measured in heparin plasma. TNF-α, IL-6, and sIL-6R were measured in EDTA plasma, CRP level was measured in heparin plasma, and fibrinogen level was measured in CTAD plasma. The liver enzymes Gamma glutamyl transferase (GGT),
alanine aminotransferase (ALT), and aspartate aminotransferase (AST) were determined in heparin plasma. For further details on the measurements of blood biomarker profiles, see Willemsen *et al* [10]. Because the distributions of insulin, triglycerides, CRP, TNF-alpha, IL-6, AST, ALT and GGT were skewed, an LN-transformation was applied to these variables to obtain a normal distribution. For insulin, TNF-α, and GGT extreme outlier values identified by inspection of the distribution of the data from discordant pairs were adjusted: For insulin, 1 value of 300.8 μIU/ml was changed to 109.9 μIU/ml, which corresponds to 3 times the SD above the mean of the LN-transformed distribution. For TNF-alpha, 1 value of 266.80 pg/mL was changed to 2.5 pg/mL, which corresponds to 4 times the SD above the mean of the LN-transformed distribution. For GGT, 4 values (5.06 U/L, 5.11 U/L, 5.26 U/L and 4.87 U/L), were changed to 4.85 U/L, which corresponds to 3 times the SD above the mean of the LN-transformed distribution. Following the same procedure, 1 extreme outlier value in the analysis of BMI concordant twins who became discordant after blood draw was identified and adjusted: for insulin, 1 value of 109 μIU/ml was changed to 34.8 μIU/ml, which corresponds to 3 times the SD above the mean of the LN-transformed distribution in this group of twins.

3.3 Gene Expression

RNA extraction [10], expression profiling and expression QC procedures have been described in detail previously [12]. In short, PAXgene tubes were shipped to the Rutgers University Cell and DNA Repository (RUCDR), USA. RNA was extracted using Qiagen Universal liquid handling system (PAXgene extraction kits as per the manufacturer's protocol). Total RNA was determined using spectroscopy (Trinean DropSense) to determine purity and concentration (total yield is calculated from these values) while RNA fidelity was measured by the Agilent Bioanalyzer analysis. Samples were hybridized to Affymetrix U219 array plates (GeneTitan), which contain 530,467 probes for 49,293 transcripts, each with a length of 25 bases. Array hybridization, washing, staining, and scanning were carried out in an Affymetrix GeneTitan System per the manufacturer's protocol. Probes were excluded if they did not pass standard Affymetrix QC metrics (Affymetrix expression console), if they did not map uniquely to Human Genome build 19 (hg19), or if they contained a polymorphic SNP based on snp137 (UCSC). Expression values were obtained using Robust Multi-array Average (RMA) normalization implemented in Affymetrix Power Tools (APT, v 1.12.0). The data from MZ twin pairs used in the current study are part of a larger dataset in which expression was measured and normalized jointly [12]. After quality control, gene expression data were available for 634 complete MZ pairs with data on BMI and all covariates for which the expression data were corrected (e.g. blood cell counts).

4. Analyses

4.1 Correction for multiple testing of biomarkers, lifestyle and health variables

The number of dimensions in the data was determined using Matrix Spectral Decomposition (matSpD, "http://gump.qimr.edu.au/general/daleN/matSpD/"). MatSpD was applied to all data from all MZ twin pairs, except for gene expression data and the question about which of the twins eats most. Although this variable is informative in the within-twin pair comparison, it is not informative when computing correlations between variables across individuals, because the variable does not give insight into the actual amount a person eats (it only gives insight into the food intake relative to the co-twin). Among the total number of 32 variables used as input for MatSpD, 26 effective independent variables were identified based on the formula by Li et al [13]. Including the question about which twin eats most, we thus tested 27 independent variables, resulting in a significance threshold of 0.002 at a Type I error rate at 5%.

4.2 Gene Expression Analysis - Gene Ontology analysis

To test for enrichment of Gene Ontology (GO) terms among probe sets with stronger expression differences between discordant twins, all probe sets were ranked by P-value from the paired t-test and the resulting ranked gene list was supplied to the online software tool GOrilla (47). This tool performs GO enrichment analysis in gene lists based on rank, using information from all genes tested in discordant twins (i.e. no p-value cut-off is required to define the input list of genes). The analysis was run with default settings.

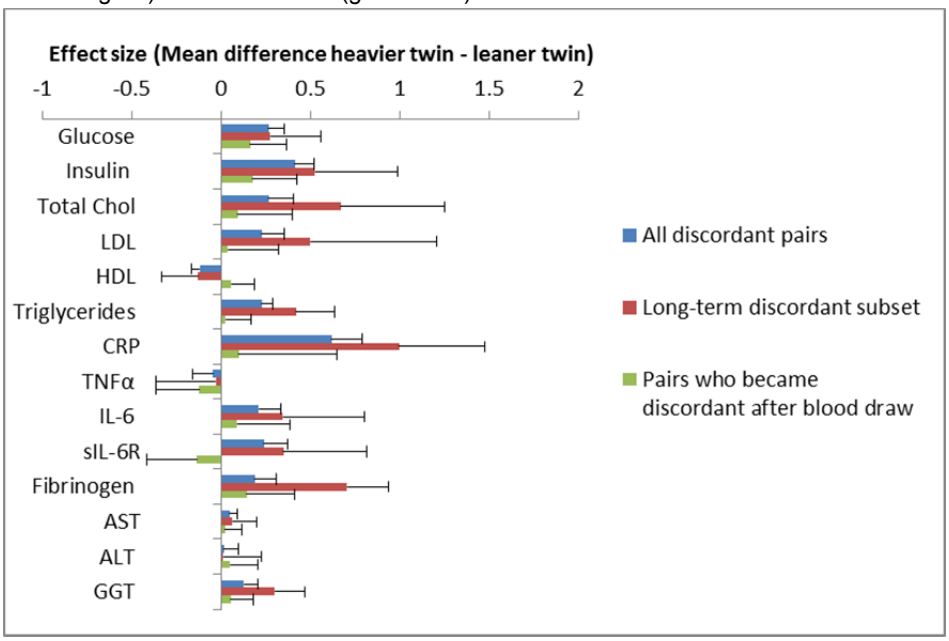
4.3 Gene Expression Analysis - Gene Set Enrichment Analysis of BMI loci from GWAS

Gene set enrichment analysis (GSEA) of loci identified in a genome-wide association analysis of BMI [14] was conducted with the software package GSEA, using default

settings [15, 16]. To obtain a set of 'candidate genes for BMI' we selected the 28 nearest genes from genome-wide significant SNPs as described by Speliotes *et al* [14]. Our expression dataset included 23 of these genes, targeted by 67 probe sets in total. In our final set of candidate BMI loci from GWAS, we only included probe sets that were expressed in blood (N=34 probe sets, N=18 genes, see Supplemental Table 2). Expressed probe sets were defined as probe sets of which the average expression level across all MZ twins was ≥ the average expression level in females of probe sets targeting chromosome Y. According to this criterion, 30 609 probe sets in total had an average expression level above the threshold (69% of all probe sets), and 13 632 had an average below the threshold.

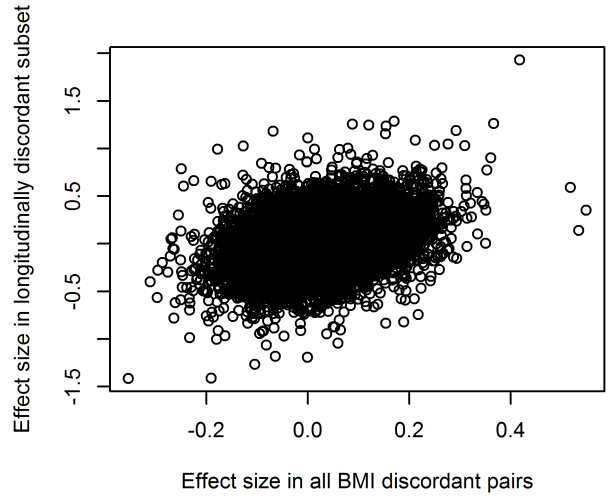
Supplemental Figures

Supplemental Figure 1: Effect sizes from the comparison of biomarkers (heavier twin - leaner twin) in all pairs who were BMI discordant at blood draw (blue bars), the longitudinally discordant subset (red bars), and MZ pairs who became discordant (Δ BMI ≥3 kg/m^2) after blood draw (green bars).



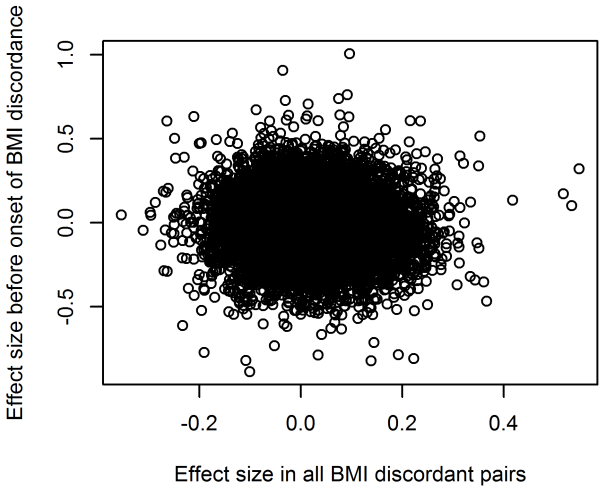
Error bars depict 95% confidence intervals.

Supplemental Figure 2: Effect sizes (heavier twin - leaner twin) from the comparison of gene expression level at each probe set in all pairs who were BMI discordant at blood draw (x-axis), and the longitudinally discordant subset (y-axis).



.

Supplemental Figure 3: Effect sizes (heavier twin - leaner twin) from the comparison of gene expression level at each probe set in all pairs who were BMI discordant at blood draw (x-axis), and MZ pairs who became discordant (Δ BMI ≥3 kg/m^2) after blood draw (y-axis).



Supplemental Tables

Supplemental Table 1: Number and frequency of MZ pairs with a BMI difference ≥2 or ≥3 at 0, 1 or multiple time points

	Never	1 time point	2 time points	>=3 time points	Total N pairs[A]
BMI difference >= 2	529 (45.8%)	290 (25.1%)	154 (13.3%)	181 (15.7%)	1154
BMI difference >= 3	797 (69.1%)	201 (17.4%)	74 (6.4%)	82 (7.1%)	1154

[A] The table includes only pairs who participated in at least three NTR projects (N total=1154 pairs).

Supplemental Table 2: Characteristics of BMI discordant MZ twin pairs followed over time

	Heavier twin at baseline	**Leaner twin at baseline**
All pairs	Mean (SD) or N (%)	Mean (SD) or N (%)
Number of pairs	305	
Age (years) baseline	33.6 (14.4), range 13-80	
BMI (kg/m^2) baseline	26.2 (3.7)	21.8 (3.2)
BMI difference (kg/m^2) baseline	4.3 (1.5), range=3.0-14.0	
Years between baseline & follow-up	3.4 (2.2), range=1-16	
BMI difference (kg/m^2) follow-up	3.1 (2.3), range: 0-12.5	
Converging pairs		
Number of pairs	169 (55.4%)	
BMI difference 2-3 at follow-up	64%	
BMI difference 1-2 at follow-up	49%	
BMI difference <1 at follow-up	56%	
Age (years) baseline	32.0 (13.7)	
BMI difference (kg/m^2) baseline	4.08 (1.4)	
BMI (kg/m^2) baseline	25.4 (3.3)	21.3 (3.2)
BMI (kg/m^2) follow-up	23.6 (3.1)	23.1 (3.0)
Weight change (kg)	-2.9 (6.8)	5.2 (6.1)
Smoking status[A]	69	71
Non-smoker at both time points	34 (49.3%)	30 (42.2%)
Started smoking	10 (14.5%)	2 (2.8%)
Smoker at both time points	22 (31.9%)	29 (40.8%)
Quit smoking	3 (4.3%)	10 (14.8%)
Pairs who are still discordant at follow-up		
Number of pairs	136 (44.6%)	
Age (years) base line	35.6 (14.9)	
BMI difference (kg/m^2) baseline	4.7 (1.6)	
BMI (kg/m^2) baseline	27.2 (3.9)	22.5 (3.2)
BMI (kg/m^2) follow-up	28.0 (4.3)	23.2 (3.6)
Weight change (kg)	2.6 (7.0)	2.2 (5.6)
Smoking status[B]	65	53
Non-smoker at both time points	29 (54.7%)	25 (38.5%)
Started smoking	5 (9.4%)	3 (4.6%)
Smoker at both time points	17 (32.1%)	28 (43.1%)
Quit smoking	2 (3.8%)	9 (13.8%)

[A] Wilcoxon signed rank test in converging pairs: p-value=0.012

[B] Wilcoxon signed rank test in pairs who are still discordant at follow-up: p-value=0.086

Supplemental Table 3: Gene expression results MZ pairs discordant at blood draw.

Paired t-test results for the Top 100 probe sets from the comparison in all MZ pairs who were discordant at blood draw

Probeset ID	Gene	chr	bp start	bp end	Mean difference (Heavy twin-lean twin)	P-value	q-value
X11722037_at	CRYBA2	chr2	219854911	219858135	-0.1559575	2.22E-05	0.29495
X11738909_a_at	H1FOO	chr3	129267844	129269652	-0.266928667	4.49E-05	0.29495
X11720432_s_at	TFEB	chr6	41651716	41652816	-0.22880975	4.52E-05	0.29495
X11753998_a_at	CD58	chr1	117064409	117078850	0.311937083	5.19E-05	0.29495
X11721835_s_at	TMEM14B	chr6	10756700	10757214	0.206527917	5.76E-05	0.29495
X11744544_s_at	KDELR1	chr19	48885827	48886584	-0.141909833	6.40E-05	0.29495
X11761957_x_at		chr4	419224	432421	0.3348195	6.62E-05	0.29495
X11748304_a_at	ENTPD6	chr20	25202428	25207365	-0.150988917	6.90E-05	0.29495
X11723536_a_at	SPATA8	chr15	97328226	97328845	-0.1734235	7.11E-05	0.29495
X11715818_at	ZFP36L2	chr2	43449541	43452891	-0.139368583	0.000101767	0.380147
X11744580_x_at	ZNF271	chr18	32885940	32890730	0.230477667	0.000119316	0.405181
X11743279_x_at	WNK1	chr12	1017641	1020618	0.22331075	0.000130267	0.405506
X11746677_x_at	REXO1	chr19	1815248	1816568	-0.167292	0.000151961	0.407011
X11762707_a_at	C14orf182	chr14	50469467	50474238	-0.149497583	0.000152542	0.407011
X11747719_a_at	RP11-173D9.3	chr14	35739607	35743497	0.1403315	0.000187525	0.435791
X11743066_s_at	PTDSS1	chr8	97345685	97349223	0.18558875	0.00019626	0.435791
X11718381_a_at	PSD4	chr2	113958735	113960814	-0.186774333	0.000226454	0.435791
X11718609_a_at	VPS72	chr1	151146867	151149540	-0.1965665	0.0002282	0.435791
X11733403_x_at	CLEC7A	chr12	10269376	10271189	0.232779667	0.000232062	0.435791
X11751906_a_at	PTPRM	chr18	8406107	8406859	0.263364417	0.000237758	0.435791
X11731446_s_at	SNCA	chr4	90645250	90647811	0.322955083	0.000251742	0.435791
X11720682_at	TIFA	chr4	113196783	113199590	0.20228125	0.000261257	0.435791
X11757341_s_at	OPTN	chr10	13178745	13180291	0.23637175	0.000268326	0.435791
X11715810_a_at	TRIP12	chr2	230631930	230633435	0.172217833	0.00028818	0.448535
X11737820_a_at	MAP3K3	chr17	61769093	61773663	-0.261552	0.000330068	0.473931
X11716095_s_at	KLF6	chr10	3818188	3821782	0.110515583	0.000330793	0.473931
X11739687_a_at	FAM107B	chr10	14560556	14563305	0.203113917	0.000350644	0.473931
X11750531_a_at	PYCARD	chr16	31212806	31214771	-0.152709417	0.000355246	0.473931
X11739095_a_at	CPSF7	chr11	61170121	61179409	-0.183803833	0.000378095	0.475697
X11739339_s_at	FAM46C	chr1	118165464	118170994	0.312222417	0.000404466	0.475697
X11754960_a_at	PRRT2	chr16	29824311	29827201	-0.174264917	0.000404628	0.475697
X11731132_s_at	JRK	chr8	143738874	143747746	0.193752917	0.000413249	0.475697
X11720466_a_at	ARL6IP1	chr16	18802991	18804692	-0.286802333	0.000435947	0.475697
X11740534_x_at	PTGES3L	chr17	41120105	41121203	-0.08969375	0.00043676	0.475697
X11726022_a_at	FAM177A1	chr14	35522406	35548230	0.21662525	0.00045696	0.475697
X11732358_x_at	ZNRF1	chr16	75127470	75144611	-0.175354	0.000472791	0.475697
X11748248_a_at	CKAP2	chr13	53049030	53050485	0.222333	0.000472939	0.475697
X11749102_a_at	SECISBP2	chr9	91972914	91974557	0.177691417	0.000492534	0.475697
X11716233_a_at	CTNNBIP1	chr1	9908334	9910834	-0.16573925	0.000520889	0.475697
X11759560_s_at	SLC25A37	chr8	23428848	23432976	0.138722083	0.000548822	0.475697
X11723368_a_at	ARHGEF6	chrX	135747706	135750328	0.111911167	0.000583576	0.475697
X11719076_a_at	REEP1	chr2	86441116	86444233	-0.123607667	0.000584063	0.475697
X11746767_a_at	ITGA4	chr2	182347241	182351036	0.194040833	0.000590018	0.475697
X11763763_a_at	DGKA	chr12	56333199	56334682	-0.1891795	0.000598581	0.475697
X11737437_a_at	CNTLN	chr9	17502549	17503921	-0.097775667	0.000603245	0.475697
X11723078_s_at	CXorf26	chrX	75397451	75398039	0.26485675	0.000603365	0.475697
X11761337_at	TRMT2A	chr22	20102488	20103303	-0.166182583	0.000620511	0.475697
X11725142_a_at	TGS1	chr8	56717452	56738007	0.226291833	0.000621967	0.475697
X11717189_x_at	DNAJA1	chr9	33036572	33039905	0.18519575	0.000659125	0.475697
X11729260_a_at	ZNF644	chr1	91403829	91406866	0.287202417	0.0006716	0.475697
X200010_PM_at	RPL11	chr1	24018796	24020403	0.19060025	0.000679053	0.475697
X11728932_x_at	FXYD7	chr19	35642157	35645204	-0.138288833	0.000732285	0.475697
X11746235_a_at	LIN37	chr19	36243077	36244721	-0.195913333	0.000759223	0.475697
X11761810_at	FUK	chr16	70500785	70501869	-0.187282917	0.00076224	0.475697
X11760760_x_at	UCHL5	chr1	192989017	192989586	0.141950417	0.000765789	0.475697

X11731369_x_at		chr19	12460185	12462132	-0.19750175	0.000769483	0.475697
X11757971_s_at	THAP5	chr7	108202959	108205549	0.51778925	0.000776868	0.475697
X11744359_s_at	MBD3	chr19	1576677	1578537	-0.275785	0.000788131	0.475697
X11734263_at	ZNF780A	chr19	40575059	40582116	0.205208583	0.000816413	0.475697
X11743512_x_at	KDM4B	chr19	5151346	5153606	-0.186464583	0.000832775	0.475697
X11742518_at	OR7G2	chr19	9212945	9213982	-0.229504333	0.000839003	0.475697
X11746510_x_at	PSMD6	chr3	63996225	64004383	0.245186417	0.000852269	0.475697
X11750306_x_at	RP11-770J1.4	chr11	118303836	118305921	-0.065328417	0.000863184	0.475697
X11745492_a_at	CCNO	chr5	54528595	54529508	-0.112790667	0.000875228	0.475697
X11726606_at	MTF2	chr1	93602227	93604638	0.2138585	0.000888109	0.475697
X11741451_a_at	RDM1	chr17	34245070	34245447	-0.2009395	0.000891303	0.475697
X11762406_s_at	GBP2	chr1	89571815	89572735	0.194151167	0.000900771	0.475697
X11737758_x_at	CARS	chr11	3022152	3033506	0.162189167	0.000914648	0.475697
X11717992_a_at	DNAJC7	chr17	40129279	40134421	0.1758085	0.000930425	0.475697
X11732675_at	RNF180	chr5	63665442	63668696	-0.0680505	0.00093238	0.475697
X11761082_at	HINFP	chr11	119000570	119003283	-0.122547	0.000972815	0.475697
X11717432_a_at	ECHDC1	chr6	127609855	127611422	0.2918125	0.000973988	0.475697
X11743450_a_at	ELOVL6	chr4	110967002	110972918	-0.1433735	0.001002042	0.475697
X11744744_a_at	SLITRK3	chr3	164904508	164908639	-0.114370917	0.00102978	0.475697
X11756604_a_at	PCDHB2	chr5	140474227	140476962	-0.095217333	0.001032168	0.475697
X11727706_a_at	BBC3	chr19	47724079	47725175	-0.154623667	0.001050753	0.475697
X11735117_at	VAMP4	chr1	171669300	171673674	0.183243	0.001054273	0.475697
X11739386_at	CDC73	chr1	193219806	193223031	0.1948555	0.001091936	0.475697
X11741283_a_at	AKAP7	chr6	131602670	131604675	0.160234167	0.001112374	0.475697
X11754982_x_at		chr22	29694723	29696515	0.111470333	0.001118688	0.475697
X11756383_a_at	FARSB	chr2	223478530	223478647	0.227243833	0.00112136	0.475697
X11760039_a_at	LRP1	chr12	57542764	57543842	-0.16667475	0.001139725	0.475697
X11721136_a_at	CMC2	chr16	81009698	81010076	0.19487475	0.001142803	0.475697
X11755873_a_at	ACSS1	chr20	24986868	24988577	0.131453083	0.001156619	0.475697
X11719998_a_at	SYNJ2BP	chr14	70838148	70883778	-0.2325685	0.001165786	0.475697
X11729846_a_at	BST1	chr4	15733363	15734410	0.2108295	0.001178086	0.475697
X11728358_a_at	FAM13B	chr5	137273649	137276037	0.15703675	0.001203324	0.475697
X11751373_a_at	PDGFC	chr4	157681606	157689142	0.1944855	0.001206933	0.475697
X11736605_a_at	XPO7	chr8	21862506	21864096	0.159236083	0.001207148	0.475697
X11739000_a_at	SLC38A2	chr12	46751972	46754992	0.174002	0.0012289	0.475697
X11719604_a_at	PTPN9	chr15	75759462	75761324	-0.182001917	0.00126991	0.475697
X11760832_at		chr6	170870879	170871321	-0.161687917	0.001276276	0.475697
X11726511_a_at	TCHP	chr12	110340832	110344445	-0.156983583	0.001305642	0.475697
X11757873_x_at	ATP5B	chr12	57031959	57033976	0.126153417	0.001325367	0.475697
X11718016_a_at	ADARB1	chr21	46641933	46646478	0.187034917	0.001328605	0.475697
X11746565_a_at	CHSY1	chr15	101715928	101719185	-0.2250135	0.001330175	0.475697
X11755199_a_at	C11orf35	chr11	554855	555930	-0.179232667	0.001344629	0.475697
X11758700_s_at		chr10	112363989	112364394	0.548986833	0.001381421	0.475697
X11748417_a_at	PROX1	chr1	214178508	214214853	-0.113066667	0.001393676	0.475697
X11759563_a_at	EPC1	chr10	32556679	32558063	0.21392525	0.001408086	0.475697

Supplemental Table 3 Continued

Results of the enrichment analysis of GO processes in the expression data from all MZ pairs who were discordant at blood draw

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0043170	macromolecule metabolic process	3.84E-06	4.66E-02	1.16	16602	5774	1786	720
GO:0044260	cellular macromolecule metabolic process	5.00E-06	3.04E-02	1.17	16602	5299	1757	656
GO:0044238	primary metabolic process	5.68E-06	2.30E-02	1.13	16602	7088	1772	857
GO:0008152	metabolic process	7.50E-06	2.28E-02	1.12	16602	7813	1772	934
GO:0080090	regulation of primary metabolic process	9.69E-06	2.35E-02	1.21	16602	4536	1374	455
GO:0071704	organic substance metabolic process	1.81E-05	3.67E-02	1.12	16602	7354	1772	881
GO:2000737	negative regulation of stem cell differentiation	1.83E-05	3.17E-02	33.69	16602	9	219	4
GO:0044237	cellular metabolic process	2.30E-05	3.50E-02	1.12	16602	7139	1768	855
GO:0070365	hepatocyte differentiation	2.57E-05	3.46E-02	31	16602	9	238	4
GO:0010638	positive regulation of organelle organization	3.19E-05	3.88E-02	3.22	16602	274	376	20
GO:0019222	regulation of metabolic process	3.57E-05	3.95E-02	1.19	16602	5044	1374	495
GO:0031323	regulation of cellular metabolic process	4.36E-05	4.42E-02	1.2	16602	4549	1374	451
GO:0032480	negative regulation of type I interferon production	4.82E-05	4.50E-02	6.62	16602	34	664	9
GO:0043280	positive regulation of cysteine-type endopeptidase activity involved in apoptotic process	4.86E-05	4.21E-02	5.14	16602	104	373	12
GO:0045963	negative regulation of dopamine metabolic process	6.21E-05	5.03E-02	178.52	16602	2	93	2
GO:0045914	negative regulation of catecholamine metabolic process	6.21E-05	4.71E-02	178.52	16602	2	93	2
GO:0019538	protein metabolic process	7.79E-05	5.57E-02	1.23	16602	2829	1786	373
GO:0006974	cellular response to DNA damage stimulus	9.32E-05	6.29E-02	1.77	16602	597	941	60
GO:2001056	positive regulation of cysteine-type endopeptidase activity	9.42E-05	6.02E-02	4.81	16602	111	373	12

GO:0006793 phosphorus metabolic process	1.02E-04	6.22E-02	1.56	16602	1639	610	94
GO:0009892 negative regulation of metabolic process	1.06E-04	6.10E-02	1.31	16602	1563	1778	220
GO:0032880 regulation of protein localization	1.09E-04	6.04E-02	1.69	16602	460	1435	67
GO:0010950 positive regulation of endopeptidase activity	1.49E-04	7.85E-02	4.6	16602	116	373	12
GO:0097193 intrinsic apoptotic signaling pathway	1.51E-04	7.63E-02	2.22	16602	154	1410	29
GO:2001242 regulation of intrinsic apoptotic signaling pathway	1.53E-04	7.44E-02	3.16	16602	106	891	18
GO:0009987 cellular process	1.55E-04	7.24E-02	1.07	16602	11666	1626	1217
GO:0090200 positive regulation of release of cytochrome c from mitochondria	1.62E-04	7.28E-02	6.21	16602	24	891	8
GO:0031325 positive regulation of cellular metabolic process	1.77E-04	7.68E-02	1.28	16602	2049	1526	242
GO:0006796 phosphate-containing compound metabolic process	1.88E-04	7.86E-02	1.55	16602	1599	610	91
GO:0060255 regulation of macromolecule metabolic process	1.95E-04	7.89E-02	1.16	16602	4300	1781	537
GO:2000484 positive regulation of interleukin-8 secretion	1.99E-04	7.81E-02	188.66	16602	8	22	2
GO:0010952 positive regulation of peptidase activity	2.03E-04	7.70E-02	4.45	16602	120	373	12
GO:0009893 positive regulation of metabolic process	2.12E-04	7.80E-02	1.25	16602	2164	1780	290
GO:0044267 cellular protein metabolic process	2.30E-04	8.23E-02	1.24	16602	2353	1786	313
GO:0045321 leukocyte activation	2.35E-04	8.16E-02	1.68	16602	333	1777	60
GO:0022603 regulation of anatomical structure morphogenesis	2.69E-04	9.07E-02	1.52	16602	673	1456	90
GO:2000482 regulation of interleukin-8 secretion	3.32E-04	1.09E-01	150.93	16602	10	22	2
GO:0008219 cell death	3.37E-04	1.08E-01	1.47	16602	831	1410	104
GO:1901990 regulation of mitotic cell cycle phase transition	3.38E-04	1.05E-01	2.39	16602	221	754	24
GO:0032026 response to magnesium ion	3.43E-04	1.04E-01	159.63	16602	13	16	2
GO:0010604 positive regulation of macromolecule metabolic process	3.57E-04	1.06E-01	1.26	16602	1959	1780	264
GO:1901360 organic cyclic compound metabolic process	3.77E-04	1.09E-01	1.17	16602	4164	1625	477
GO:0016265 death	3.91E-04	1.11E-01	1.47	16602	835	1410	104
GO:0010467 gene expression	3.91E-04	1.08E-01	1.51	16602	588	1641	88
GO:0008150 biological_process	3.98E-04	1.07E-01	1.04	16602	13924	1778	1548
GO:0048522 positive regulation of cellular process	4.63E-04	1.22E-01	1.2	16602	3381	1540	376
GO:0031324 negative regulation of cellular metabolic process	5.20E-04	1.35E-01	1.3	16602	1448	1778	201
GO:0031057 negative regulation of histone modification	5.28E-04	1.34E-01	35.47	16602	26	54	3
GO:1901987 regulation of cell cycle phase transition	5.45E-04	1.35E-01	2.32	16602	228	754	24
GO:0051246 regulation of protein metabolic process	5.61E-04	1.36E-01	1.27	16602	1704	1781	232
GO:0015937 coenzyme A biosynthetic process	5.77E-04	1.37E-01	8.75	16602	11	862	5
GO:0050808 synapse organization	5.83E-04	1.36E-01	2.5	16602	107	1367	22
GO:0090199 regulation of release of cytochrome c from mitochondria	5.90E-04	1.35E-01	5.48	16602	36	673	8
GO:0006919 activation of cysteine-type endopeptidase activity involved in apoptotic process	6.09E-04	1.37E-01	5.14	16602	78	373	9
GO:0006725 cellular aromatic compound metabolic process	6.40E-04	1.41E-01	1.17	16602	3955	1625	453
GO:0070265 necrotic cell death	6.65E-04	1.44E-01	8.47	16602	11	891	5
GO:0033554 cellular response to stress	6.87E-04	1.46E-01	1.45	16602	1078	1118	105
GO:0048518 positive regulation of biological process	7.01E-04	1.47E-01	1.18	16602	3774	1540	413
GO:0001775 cell activation	7.27E-04	1.50E-01	6.24	16602	532	35	7
GO:0060216 definitive hemopoiesis	7.28E-04	1.47E-01	5.04	16602	13	1774	7
GO:0051171 regulation of nitrogen compound metabolic process	7.95E-04	1.58E-01	1.21	16602	3559	1285	332
GO:0046483 heterocycle metabolic process	8.01E-04	1.57E-01	1.16	16602	3942	1720	475
GO:0043412 macromolecule modification	8.15E-04	1.57E-01	1.24	16602	1986	1786	265
GO:0046649 lymphocyte activation	8.35E-04	1.59E-01	1.69	16602	271	1777	49
GO:0001879 detection of yeast	8.43E-04	1.58E-01	1,185.86	16602	1	14	1
GO:0032469 endoplasmic reticulum calcium ion homeostasis	8.57E-04	1.58E-01	12.69	16602	12	436	4
GO:0090161 Golgi ribbon formation	8.66E-04	1.57E-01	14.09	16602	3	1178	3
GO:0070201 regulation of establishment of protein localization	8.83E-04	1.58E-01	1.65	16602	393	1435	56
GO:0051945 negative regulation of catecholamine uptake involved in synaptic transmission	9.64E-04	1.70E-01	1,037.62	16602	1	16	1
GO:0070494 regulation of thrombin receptor signaling pathway	9.64E-04	1.67E-01	1,037.62	16602	1	16	1
GO:0070495 negative regulation of thrombin receptor signaling pathway	9.64E-04	1.65E-01	1,037.62	16602	1	16	1
GO:0051585 negative regulation of dopamine uptake involved in synaptic transmission	9.64E-04	1.63E-01	1,037.62	16602	1	16	1
GO:0051622 negative regulation of norepinephrine uptake	9.64E-04	1.60E-01	1,037.62	16602	1	16	1
GO:0051621 regulation of norepinephrine uptake	9.64E-04	1.58E-01	1,037.62	16602	1	16	1
GO:0033043 regulation of organelle organization	9.69E-04	1.57E-01	2.08	16602	636	376	30
GO:0043001 Golgi to plasma membrane protein transport	9.89E-04	1.58E-01	8.98	16602	20	462	5

Results of the enrichment analysis of GO functions in the expression data from all MZ pairs who were discordant at blood draw

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0005515	protein binding	1.09E-08	4.32E-05	1.15	16602	7785	1791	964
GO:0044822	poly(A) RNA binding	1.76E-06	3.49E-03	1.58	16602	1003	1309	125
GO:0005488	binding	2.39E-06	3.17E-03	1.08	16602	11600	1788	1343
GO:0097159	organic cyclic compound binding	3.62E-06	3.59E-03	1.21	16602	5318	1223	474
GO:1901363	heterocyclic compound binding	4.03E-06	3.20E-03	1.21	16602	5245	1249	477
GO:0003723	RNA binding	8.06E-05	5.34E-02	1.35	16602	1342	1730	189
GO:0003676	nucleic acid binding	1.71E-04	9.71E-02	1.2	16602	3481	1625	409
GO:0036094	small molecule binding	2.04E-04	1.01E-01	1.29	16602	2417	1223	230
GO:0008894	guanosine-5'-triphosphate,3'-diphosphate diphosphatase activity	4.22E-04	1.86E-01	2,371.71	16602	1	7	1
GO:1901265	nucleoside phosphate binding	5.21E-04	2.07E-01	1.29	16602	2161	1223	206
GO:0000166	nucleotide binding	5.21E-04	1.88E-01	1.29	16602	2161	1223	206
GO:0004844	uracil DNA N-glycosylase activity	5.50E-04	1.82E-01	80.59	16602	4	103	2
GO:0097506	deaminated base DNA N-glycosylase activity	5.50E-04	1.68E-01	80.59	16602	4	103	2
GO:0001874	(1->3)-beta-D-glucan receptor activity	8.43E-04	2.39E-01	1,185.86	16602	1	14	1
GO:0001873	polysaccharide receptor activity	8.43E-04	2.23E-01	1,185.86	16602	1	14	1

Supplemental Table 3 Continued
Results of the enrichment analysis of GO components in the expression data from all MZ pairs who were discordant at blood draw

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0043231	intracellular membrane-bounded organelle	2.26E-08	3.16E-05	1.21	16602	6234	1285	585
GO:0044428	nuclear part	5.57E-08	3.91E-05	1.36	16602	2142	1719	302
GO:0043229	intracellular organelle	1.10E-07	5.16E-05	1.18	16602	7227	1240	638
GO:0044424	intracellular part	2.50E-07	8.77E-05	1.08	16602	11466	1756	1314
GO:0043227	membrane-bounded organelle	3.96E-07	1.11E-04	1.17	16602	7202	1265	644
GO:0044446	intracellular organelle part	9.50E-07	2.22E-04	1.17	16602	5875	1769	730
GO:0032991	macromolecular complex	1.36E-06	2.73E-04	1.22	16602	3985	1781	520
GO:0043226	organelle	2.36E-06	4.14E-04	1.15	16602	8058	1265	704
GO:0031968	organelle outer membrane	3.13E-06	4.88E-04	3.6	16602	129	714	20
GO:0019867	outer membrane	5.42E-06	7.60E-04	3.47	16602	134	714	20
GO:0043234	protein complex	1.29E-05	1.64E-03	1.22	16602	3506	1781	457
GO:0005794	Golgi apparatus	1.34E-05	1.57E-03	1.71	16602	522	1471	79
GO:0044422	organelle part	1.52E-05	1.64E-03	1.15	16602	6053	1769	739
GO:0005634	nucleus	5.15E-05	5.16E-03	1.22	16602	3981	1285	377
GO:0072559	NLRP3 inflammasome complex	1.64E-04	1.54E-02	13.79	16602	5	963	4
GO:0005768	endosome	1.69E-04	1.48E-02	1.74	16602	370	1471	57
GO:0005654	nucleoplasm	2.22E-04	1.83E-02	1.39	16602	1028	1656	143
GO:0031010	ISWI-type complex	2.25E-04	1.75E-02	6.93	16602	9	1596	6
GO:0031090	organelle membrane	3.64E-04	2.69E-02	1.29	16602	2263	1214	214
GO:0005737	cytoplasm	5.28E-04	3.71E-02	1.19	16602	3139	1730	390
GO:0005741	mitochondrial outer membrane	5.49E-04	3.67E-02	3.29	16602	106	714	15
GO:0005681	spliceosomal complex	5.94E-04	3.79E-02	2.03	16602	136	1746	29
GO:0098588	bounding membrane of organelle	6.70E-04	4.08E-02	1.45	16602	1795	663	104
GO:0044444	cytoplasmic part	7.41E-04	4.33E-02	1.11	16602	6524	1769	773
GO:1902494	catalytic complex	9.76E-04	5.48E-02	1.42	16602	713	1776	108

N - is the total number of genes
B - is the total number of genes associated with a specific GO term
n - is the number of genes in the top of the input list
b - is the number of genes in the intersection
Enrichment = (b/n) / (B/N)

Supplemental Table 3 Continued BMI candidate genes from GWAS

Probeset ID	Gene	chr	bp start	bp end	Mean difference (Heavy twin- lean twin)	P-value
X11757158_x_at	RPL27A	chr11	8705553	8706439	-0.265600833	0.01726931
X11722222_at	TMEM160	chr19	47549168	47551882	-0.139438083	0.042584018
X11728331_a_at	KCTD15	chr19	34302696	34306666	-0.072732167	0.112544529
X11722804_x_at	PTBP2	chr1	97278829	97280349	0.10804825	0.122102125
X11755163_a_at	FANCL	chr2	58386378	58386935	0.120043833	0.14734394
X11763807_at	MAP2K5	chr15	67951109	67951694	-0.059615333	0.154978677
X11757157_at	RPL27A	chr11	8705553	8706439	-0.152119917	0.167965257
X11749270_a_at	GPRC5B	chr16	19868616	19871866	0.058514417	0.178945818
X11723064_a_at	TMEM18	chr2	667335	670975	-0.037379833	0.192874871
X11748421_a_at	TMEM18	chr2	667335	669675	0.062703667	0.215433744
X11758115_s_at	PTBP2	chr1	97278829	97280349	0.059483167	0.283516582
X11735168_a_at	QPCTL	chr19	46205056	46207247	0.048152833	0.289610822
X11757335_s_at	MTIF3	chr13	28009776	28014586	0.057331	0.332608659
X11735738_x_at	SH2B1	chr16	28884768	28885526	-0.036996167	0.371866712
X11719312_a_at	MAP2K5	chr15	68061940	68099461	-0.053008667	0.384309664
X11758687_s_at	TMEM18	chr2	667335	669675	0.057649333	0.388020616
X11723450_a_at	FTO	chr16	54145674	54155853	0.05670125	0.444320049
X11728061_a_at	ZNF608	chr5	123982372	123984826	-0.054238417	0.447142328
X11721075_a_at	MTCH2	chr11	47638868	47650907	0.03643175	0.480440294
X11748888_a_at	FTO	chr16	53967897	54155853	-0.034729667	0.600806646
X11723451_x_at	FTO	chr16	54145674	54155853	0.03448675	0.601559716
X11722380_at	NUDT3	chr6	34255997	34256708	0.033521417	0.603998358
X11748612_a_at	ETV5	chr3	185764097	185775033	-0.019385083	0.64291969
X11724249_a_at	SLC39A8	chr4	103188647	103189236	0.026031333	0.652173182
X11762571_a_at	GNPDA2	chr4	44703885	44705159	0.0190655	0.652820119
X11721866_a_at	GNPDA2	chr4	44703885	44705159	0.020161917	0.655448171
X11719175_a_at	GPRC5B	chr16	19868616	19873662	0.00949025	0.745942093
X11749835_x_at	SH2B1	chr16	28883855	28885526	-0.014989	0.761435581
X11724252_s_at	SLC39A8	chr4	103182823	103184350	-0.017443917	0.775788688
X11719176_a_at	GPRC5B	chr16	19868616	19871866	0.016042667	0.789953565

X11721076_at	MTCH2	chr11	47638867	47640471	0.00940325	0.884434048
X11721867_a_at	GNPDA2	chr4	44703885	44705159	-0.006933333	0.91246732
X11740764_at	SLC39A8	chr4	103172198	103174587	-0.005699167	0.91704398
X11720353_s_at	ETV5	chr3	185764097	185766649	-0.002964667	0.943630725

Supplemental Table 3 Continued

SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
18	0.288449	0.7147	0.923076	0.923076	0.912	7970	tags=44%, list=44%,

Supplemental Table 4: Gene expression results MZ pairs discordant in longitudinally discordant MZ pairs.

Paired t-test results for the Top 100 probe sets from the comparison in longitudinally discordant MZ pairs

Probeset ID	Gene	chr	bp start	bp end	Mean difference (Heavy twin- lean twin)	P-value	q-value
X11727799_at	CHRM1	chr11	62676151	62678650	-0.330721111	5.17E-06	0.22888
X11721227_at	LEMD3	chr12	65639080	65642107	-0.395771111	5.12E-05	0.999482
X11737263_at	HRK	chr12	117293949	117319246	-0.769828889	0.000107	0.999482
X11736257_at	PRKCE	chr2	46411874	46415129	-0.482706667	0.000113	0.999482
X11757638_s_at	CD93	chr20	23059986	23064611	-0.246394444	0.000134	0.999482
X11716725_a_at	ADCY6	chr12	49159975	49162499	-0.33289	0.000161	0.999482
X11746884_a_at	RAD52	chr12	1021243	1022618	0.418235556	0.000235	0.999482
X11739853_a_at	SNX16	chr8	82711816	82736175	0.406994444	0.000253	0.999482
X11732927_x_at	KLRC1	chr12	10598627	10599261	1.247322222	0.000255	0.999482
X11760342_a_at	PPP3CB	chr10	75203538	75204580	0.620455556	0.000267	0.999482
X11733165_s_at	YIPF5	chr5	143537723	143540123	0.563431111	0.000283	0.999482
X11761097_at	ARFGAP2	chr11	47194649	47195033	-0.509937778	0.000316	0.999482
X11747816_a_at	GDAP1	chr8	75276220	75279111	0.395358889	0.000377	0.999482
X11727636_at	RERGL	chr12	18233803	18234407	-0.170217778	0.000381	0.999482
X11763763_a_at	DGKA	chr12	56333199	56334682	-0.720315556	0.000468	0.999482
X11750678_a_at	ALK	chr2	29415640	29416788	-0.283086667	0.000481	0.999482
X11744971_at	PSORS1C3	chr6	31141512	31145676	0.26571	0.000545	0.999482
X11734853_x_at	SCEL	chr13	78208514	78219398	-0.321234444	0.000554	0.999482
X11742211_x_at	APP	chr21	27252861	27264180	-0.25373	0.000558	0.999482
X11748086_a_at	PIK3CG	chr7	106545554	106547590	-0.46727	0.000569	0.999482
X11747177_x_at	AQP1	chr7	30963065	30965131	-0.339784444	0.000611	0.999482
X11754790_a_at	FAM193A	chr4	2701407	2702271	-0.467374444	0.000633	0.999482
X11729436_a_at	LY6H	chr8	144239331	144239902	-0.358746667	0.000643	0.999482
X11718198_at	LHFP	chr13	39917029	39918191	-0.324992222	0.000701	0.999482
X11756615_x_at	SHMT2	chr12	57627786	57628718	0.569896667	0.000825	0.999482
X11738272_a_at	PSMG4	chr6	3259128	3259430	-0.677882222	0.000854	0.999482
X11717611_a_at	MAST2	chr1	46500210	46501896	-0.328534444	0.000863	0.999482
X11733161_a_at	TNFAIP8	chr5	118728511	118730294	0.592585556	0.000866	0.999482
X11748265_a_at	LDLRAD3	chr11	36250710	36253686	-0.472254444	0.000899	0.999482
X11733199_a_at	CCDC14	chr3	123632300	123634565	-0.322354444	0.000912	0.999482
X11764211_at	RP11-71J4.2	chr12	68738444	68743739	0.303654444	0.000928	0.999482
X11742468_at	RASSF10	chr11	13031083	13032647	-0.430362222	0.000932	0.999482
X11739911_at	PCBD2	chr5	134246025	134296967	-0.414174444	0.000989	0.999482
X11725151_at	MYBPC3	chr11	47352957	47354253	-0.348562222	0.001034	0.999482
X11724825_a_at	DOLK	chr9	131707809	131709898	0.622142222	0.001047	0.999482
X11731978_s_at	PCSK7	chr11	117075053	117077072	-0.195322222	0.001049	0.999482
X11731694_s_at	HIST1H2AE	chr6	26217165	26217711	0.302743333	0.001095	0.999482
X11746570_s_at	FRYL	chr4	48499378	48502237	0.339261111	0.00122	0.999482
X11729288_at	MSL3	chrX	11783586	11786096	-0.402761111	0.001311	0.999482
X11756760_a_at	GCDH	chr19	13010282	13010953	0.385726667	0.001334	0.999482
X11731795_at	EPHX4	chr1	92518067	92529093	0.267196667	0.001334	0.999482
X11752620_s_at	YPEL5	chr2	30381485	30383399	0.291173333	0.001364	0.999482
X11761010_a_at	CKMT1B	chr15	43889574	43890525	-0.305822222	0.001379	0.999482
X11736644_a_at	G6PC2	chr2	169764078	169766505	0.099615556	0.001387	0.999482
X11732685_a_at	POGZ	chr1	151375200	151378940	0.57982	0.001405	0.999482
X11742589_at	OR4L1	chr14	20528204	20529142	0.30026	0.001464	0.999482
X11741689_s_at	SUMF2	chr7	56147221	56148363	1.003967778	0.001498	0.999482
X11737235_x_at	USP6	chr17	5076100	5078329	-0.56039	0.001501	0.999482
X11718800_at	PTPRU	chr1	29652111	29653325	-0.405303333	0.001595	0.999482
X11763585_s_at	TMPO	chr12	98941351	98944157	0.521868889	0.001596	0.999482
X11740546_a_at	ADAM22	chr7	87825786	87832204	-0.191864444	0.001633	0.999482
X11731577_a_at	SLC22A1	chr6	160560685	160579750	-0.488166667	0.001701	0.999482

X11738116_a_at	GABRQ	chrX	151821004	151825999	-0.177167778	0.001705	0.999482
X11745014_at	WT1-AS	chr11	32459391	32462950	0.615003333	0.001745	0.999482
X11733302_at	ZNF790	chr19	37308330	37311016	-0.45587	0.001745	0.999482
X11741702_a_at	ABHD11	chr7	73152000	73152472	-0.341222222	0.001746	0.999482
X11754466_a_at	SH3BP4	chr2	235962236	235964358	-0.422815556	0.001761	0.999482
X11733195_a_at	FMNL3	chr12	50031724	50039686	-0.568922222	0.001828	0.999482
X11724990_s_at	CLTCL1	chr22	19166986	19167482	0.361827778	0.001868	0.999482
X11744490_x_at	C20orf24	chr20	35236118	35240960	0.539621111	0.001914	0.999482
X11752864_a_at	FER	chr5	108207066	108295048	-0.503152222	0.001972	0.999482
X11741281_a_at	AKAP7	chr6	131571299	131571756	-0.551327778	0.002016	0.999482
X11732809_a_at	PROK2	chr3	71820807	71821979	0.256727778	0.002052	0.999482
X11736562_at	ZNF518A	chr10	97915954	97923517	0.577573333	0.002063	0.999482
X11727800_a_at	POP4	chr19	30104732	30108144	0.336887778	0.002087	0.999482
X11730156_a_at	WWOX	chr16	78143675	78149051	-0.670443333	0.002088	0.999482
X11749252_a_at	STOM	chr9	124101355	124103686	0.41952	0.002109	0.999482
X11760847_at	KCNMB1	chr5	169809656	169810854	-0.288943333	0.002133	0.999482
X11761341_at	SLC9A1	chr1	27431942	27432578	-0.353345556	0.002145	0.999482
X11734260_a_at	TRIML1	chr4	189067976	189068897	-0.334801111	0.002149	0.999482
X11732738_a_at	CECR2	chr22	18032535	18033845	-0.538021111	0.002157	0.999482
X11735074_at	C15orf37	chr15	80215113	80217194	0.856636667	0.002197	0.999482
X11738946_at	MRAP	chr21	33686862	33687095	0.529556667	0.002201	0.999482
X11748052_x_at		chr11	125449989	125454575	-0.487018889	0.002213	0.999482
X11758819_x_at	RRP7A	chr22	42905974	42909001	-0.789261111	0.002215	0.999482
X11733411_at	TCF15	chr20	584441	591042	-0.384284444	0.002219	0.999482
X11746084_a_at	PCMT1	chr6	150131752	150132556	0.560476667	0.002228	0.999482
X11740650_a_at	SSH1	chr12	109176466	109183020	0.476375556	0.002286	0.999482
X11741495_a_at	KCNQ3	chr8	133139193	133142243	-0.244285556	0.002292	0.999482
X11731222_a_at	SHOX2	chr3	157813743	157816109	-0.186572222	0.002302	0.999482
X11722261_at	EDEM1	chr3	5255004	5261642	-0.493644444	0.002454	0.999482
X11717832_a_at	MIIP	chr1	12089280	12092102	-0.414705556	0.002458	0.999482
X11746675_a_at	ZNF232	chr17	5012221	5013163	-0.322361111	0.002498	0.999482
X11746844_a_at	TMEM143	chr19	48835613	48836690	-0.58645	0.002502	0.999482
X11727289_a_at	LIM2	chr19	51883164	51885821	-0.399387778	0.002517	0.999482
X11749189_s_at	HOXB3	chr17	46626232	46628543	0.36586	0.002517	0.999482
X11735404_a_at	ANKS1B	chr12	99145133	99166957	-0.264331111	0.002575	0.999482
X11752970_a_at	RAD52	chr12	1023058	1023698	0.504576667	0.002583	0.999482
X11717118_a_at	P4HA2	chr5	131527531	131530951	-0.452391111	0.00259	0.999482
X11760670_x_at	FMO3	chr1	171068946	171069876	0.223623333	0.002654	0.999482
X11724307_x_at	ORAOV1	chr11	69480331	69482355	-0.710774444	0.002668	0.999482
X11734860_a_at	WDR37	chr10	1175153	1178237	-0.46481	0.002671	0.999482
X11741548_a_at	MBNL1	chr3	152163071	152163328	0.249118889	0.002681	0.999482
X11735461_at	SPRR2B	chr1	153042704	153043334	0.239678889	0.002704	0.999482
X11748173_a_at	RIMS1	chr6	73108657	73112845	-0.252658889	0.002719	0.999482
X11740281_at	TSEN2	chr3	12571261	12573158	0.352466667	0.002734	0.999482
X11740495_x_at	SLC6A12	chr12	301644	305427	0.417687778	0.00275	0.999482
X11715678_s_at	NAP1L4	chr11	2965667	2966876	0.388807778	0.00276	0.999482
X11743251_s_at	MMP2	chr16	55539251	55540603	0.243546667	0.002763	0.999482
X11749532_a_at	RASGRF1	chr15	79282531	79291167	-0.414716667	0.002777	0.999482

Results of the enrichment analysis of GO processes in the expression data from longitudinally discordant pairs

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
	transmembrane receptor protein tyrosine kinase							
GO:0007169	signaling pathway	5.38E-05	6.53E-01	1.67	16602	528	1410	75
GO:0016192	vesicle-mediated transport	1.03E-04	6.24E-01	8.25	16602	741	19	7
GO:0001775	cell activation	1.57E-04	6.37E-01	9.85	16602	532	19	6
GO:0007167	enzyme linked receptor protein signaling pathway	1.65E-04	5.00E-01	1.53	16602	706	1442	94
GO:0042116	macrophage activation	1.76E-04	4.27E-01	255.42	16602	26	5	2
GO:0044093	positive regulation of molecular function	2.32E-04	4.70E-01	1.5	16602	1165	970	102
GO:2001031	positive regulation of cellular glucuronidation	2.41E-04	4.18E-01	4,150.50	16602	1	4	1
GO:0050878	regulation of body fluid levels	2.46E-04	3.73E-01	9.06	16602	550	20	6
GO:0065007	biological regulation	2.89E-04	3.90E-01	1.09	16602	8689	1786	1019
GO:0030168	platelet activation	3.30E-04	4.01E-01	17.56	16602	199	19	4
GO:0045760	positive regulation of action potential	3.41E-04	3.77E-01	13.73	16602	7	691	4
GO:0050789	regulation of biological process	3.43E-04	3.48E-01	1.09	16602	8229	1786	969
GO:0006810	transport	3.66E-04	3.42E-01	3.44	16602	2895	20	12
GO:0051234	establishment of localization	4.56E-04	3.96E-01	3.36	16602	2964	20	12
GO:0007207	phospholipase C-activating G-protein coupled acetylcholine receptor signaling pathway	5.27E-04	4.27E-01	5,534.00	16602	3	1	1
GO:0021546	rhombomere development	5.93E-04	4.50E-01	24.97	16602	7	285	3
GO:0038031	non-canonical Wnt signaling pathway via JNK cascade	6.78E-04	4.85E-01	20.37	16602	5	489	3
GO:0009887	organ morphogenesis	6.83E-04	4.61E-01	3.25	16602	387	185	14
GO:0019220	regulation of phosphate metabolic process	7.62E-04	4.87E-01	1.41	16602	1513	927	119
GO:0065008	regulation of biological quality	8.32E-04	5.05E-01	1.84	16602	2343	158	41
GO:2001029	regulation of cellular glucuronidation	8.33E-04	4.82E-01	2,075.25	16602	2	4	1
GO:0035668	TRAM-dependent toll-like receptor signaling pathway	8.33E-04	4.60E-01	2,075.25	16602	2	4	1
GO:0035669	TRAM-dependent toll-like receptor 4 signaling pathway	8.33E-04	4.40E-01	2,075.25	16602	2	4	1

GO:0061178	regulation of insulin secretion involved in cellular response to glucose stimulus	8.76E-04	4.43E-01	114.5	16602	29	10	2
GO:0002274	myeloid leukocyte activation	9.05E-04	4.40E-01	33.61	16602	78	19	3
GO:0050679	positive regulation of epithelial cell proliferation	9.06E-04	4.23E-01	3.02	16602	135	570	14
GO:0050794	regulation of cellular process	9.08E-04	4.09E-01	1.09	16602	7793	1786	917
GO:0051642	centrosome localization	9.16E-04	3.97E-01	5.59	16602	10	1783	6
GO:0006897	endocytosis	9.31E-04	3.90E-01	7.09	16602	270	52	6
GO:0006468	protein phosphorylation	9.80E-04	3.97E-01	1.51	16602	483	1779	78

Results of the enrichment analysis of GO functions in the expression data from longitudinally discordant pairs

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0005488	binding	7.25E-05	2.88E-01	1.06	16602	11600	1786	1329
GO:0008047	enzyme activator activity	1.51E-04	3.01E-01	1.89	16602	391	1058	47
GO:0004714	transmembrane receptor protein tyrosine kinase activity	3.61E-04	4.78E-01	2.94	16602	64	1502	17
GO:0016709	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen	7.16E-04	7.11E-01	6.3	16602	35	527	7
GO:0005326	neurotransmitter transporter activity	7.73E-04	6.15E-01	15.2	16602	24	182	4

Results of the enrichment analysis of GO components in the expression data from longitudinally discordant pairs

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0016020	membrane	6.63E-05	9.30E-02	2.45	16602	5357	24	19
GO:0044459	plasma membrane part	3.60E-04	2.53E-01	1.3	16602	2050	1296	208

N - is the total number of genes

B - is the total number of genes associated with a specific GO term

n - is the number of genes in the top of the input list

b - is the number of genes in the intersection

Enrichment = (b/n) / (B/N)

Supplemental Table 4 continued: BMI candidate genes from GWAS

Probeset ID	Gene	chr	bp start	bp end	Mean difference	P-value
X11763807_at	MAP2K5	chr15	67951109	67951694	0.261961111	0.00874709
X11721075_a_a	MTCH2	chr11	47638868	47650907	-0.413622222	0.01915122
X11748612_a_a	ETV5	chr3	185764097	185775033	0.270972222	0.06993190
X11740764_at	SLC39A8	chr4	103172198	103174587	-0.325171111	0.10781581
X11748888_a_a	FTO	chr16	53967897	54155853	-0.21378	0.12423122
X11719312_a_a	MAP2K5	chr15	68061940	68099461	0.348952222	0.12493022
X11721866_a_a	GNPDA2	chr4	44703885	44705159	-0.234468889	0.19605749
X11758687_s_a	TMEM18	chr2	667335	669675	-0.164804444	0.25353003
X11723451_x_a	FTO	chr16	54145674	54155853	0.316087778	0.26542667
X11720353_s_a	ETV5	chr3	185764097	185766649	0.111271111	0.2677416
X11728331_a_a	KCTD15	chr19	34302696	34306666	-0.113144444	0.28830922
X11762571_a_a	GNPDA2	chr4	44703885	44705159	0.180142222	0.36566345
X11719176_a_a	GPRC5B	chr16	19868616	19871866	-0.198771111	0.37758752
X11721076_at	MTCH2	chr11	47638867	47640471	-0.141058889	0.39127769
X11719175_a_a	GPRC5B	chr16	19868616	19873662	0.116055556	0.41869169
X11722380_at	NUDT3	chr6	34255997	34256708	0.147534444	0.46638071
X11722804_x_a	PTBP2	chr1	97278829	97280349	0.155097778	0.46840526
X11735168_a_a	QPCTL	chr19	46205056	46207247	-0.124088889	0.48173707
X11749835_x_a	SH2B1	chr16	28883855	28885526	0.100666667	0.53555615
X11749270_a_a	GPRC5B	chr16	19868616	19871866	0.179831111	0.55105368
X11735738_x_a	SH2B1	chr16	28884768	28885526	0.064416667	0.64409108
X11721867_a_a	GNPDA2	chr4	44703885	44705159	-0.095638889	0.66097926
X11728061_a_a	ZNF608	chr5	123982372	123984826	-0.079452222	0.66260341
X11758115_s_a	PTBP2	chr1	97278829	97280349	0.059338889	0.68194526
X11757335_s_a	MTIF3	chr13	28009776	28014586	0.060194444	0.72914422
X11722222_at	TMEM16	chr19	47549168	47551882	0.077915556	0.74948506
X11723450_a_a	FTO	chr16	54145674	54155853	0.045583333	0.79499697
X11748421_a_a	TMEM18	chr2	667335	669675	0.044701111	0.79669509
X11755163_a_a	FANCL	chr2	58386378	58386935	-0.03592	0.80997778
X11724249_a_a	SLC39A8	chr4	103188647	103189236	-0.059984444	0.82537923
X11757157_at	RPL27A	chr11	8705553	8706439	-0.07958	0.82813435
X11757158_x_a	RPL27A	chr11	8705553	8706439	-0.053602222	0.88054561
X11724252_s_a	SLC39A8	chr4	103182823	103184350	-0.005823333	0.97251022
X11723064_a_a	TMEM18	chr2	667335	670975	-0.004955556	0.97453698

Enrichment analysis output GSEA for BMI GWAS gene set

SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
18	0.288449	0.71753	0.909274	0.909274	0.902	7970	tags=44%, list=44%,

Supplemental Table 5 Gene expression results MZ pairs who became discordant after blood draw
Paired t-test results for the Top 100 probe sets from the comparison in MZ pairs who became discordant after blood draw

Probeset ID	Gene	chr	bp start	bp end	Mean difference (Heavy twin-lean twin)	P-value	q-value
X11743518_s_at	NNT	chr5	43704357	43707507	-0.365688571	1.31E-05	0.573441
X11737062_a_at	MTMR3	chr22	30421619	30426855	0.32376381	5.98E-05	0.810097
X11725284_a_at	ZFP90	chr16	68595964	68601039	-0.288890476	7.61E-05	0.810097
X11722188_at	MTDH	chr8	98736828	98740998	-0.593707619	0.000100749	0.810097
X11753045_a_at	MBNL2	chr13	98043576	98046374	-0.292324286	0.000108778	0.810097
X11753133_x_at	MCM8	chr20	5935748	5975852	-0.487959048	0.000110915	0.810097
X11741698_a_at	LCN12	chr9	139847918	139849949	0.372813333	0.000199421	0.946831
X11730425_at	ZNF628	chr19	55992568	55995854	0.390548095	0.00020386	0.946831
X11734708_at	C5orf43	chr5	60453536	60456026	-0.384932857	0.000235318	0.946831
X11723606_a_at	AGFG1	chr2	228419152	228421384	-0.520450476	0.000269666	0.946831
X11752039_a_at	PHC3	chr3	169813009	169815172	-0.391388571	0.000273923	0.946831
X11721110_s_at	NAP1L3	chrX	92925929	92928567	-0.178418571	0.000288322	0.946831
X11748088_a_at	AKAP1	chr17	55191820	55198710	0.360475238	0.000334695	0.946831
X11734685_at	PABPC3	chr13	25670006	25673392	-0.208024762	0.000346304	0.946831
X11735822_at	AC093063.1	chr19	40174127	40176361	0.295154286	0.000354759	0.946831
X11750892_a_at	SLC5A7	chr2	108626688	108630450	-0.159283333	0.000367085	0.946831
X11748102_x_at	RUNX1T1	chr8	92967203	92983388	0.241191905	0.000408501	0.946831
X11740942_at		chr18	76756891	76757631	-0.182959048	0.000445227	0.946831
X11726105_s_at	PPP2R3A	chr3	135863885	135866733	-0.451051905	0.000449082	0.946831
X11742470_a_at	SLC35F4	chr14	58030640	58033282	-0.185888095	0.000534403	0.946831
X11754230_at	DBIL5P	chr17	655900	658576	0.205343333	0.000538816	0.946831
X11722929_a_at	CNKSR3	chr6	154726433	154727786	-0.138506667	0.000560886	0.946831
X11752244_a_at	KIRREL3	chr11	126300837	126301421	0.249024762	0.00062776	0.946831
X11721822_a_at	FSD1	chr19	4322983	4323843	0.363334762	0.000633987	0.946831
X11759002_at	CACNA1E	chr1	181767428	181777219	0.213613333	0.000645416	0.946831
X11719977_a_at	MMP15	chr16	58078911	58080805	-0.162205238	0.000703666	0.946831
X11758282_s_at	TUBA1C	chr12	49666036	49667114	0.242599048	0.000808715	0.946831
X11762207_at	PCOLCE-AS1	chr7	100193491	100194274	0.44369	0.000818373	0.946831
X11728618_a_at	SLC1A7	chr1	53552855	53554651	0.296800476	0.000830727	0.946831
X11725050_a_at	FAIM2	chr12	50260679	50284505	0.341319048	0.000836513	0.946831
X11732704_a_at	FGFRL1	chr4	1018693	1020685	0.483508571	0.000868256	0.946831
X11719570_at	RCHY1	chr4	76404247	76407876	-0.329786667	0.000870498	0.946831
X11723676_at	CASQ1	chr1	160171035	160171676	-0.17496	0.000914138	0.946831
X11753130_at	TM4SF1	chr3	149086809	149089654	-0.227473333	0.000936619	0.946831
X11736930_at	G6PC	chr17	41062932	41065386	0.274805238	0.000958321	0.946831
X11720580_a_at	TCF12	chr15	57578355	57582051	-0.237150952	0.000964217	0.946831
X11741102_a_at	PTPN22	chr1	114375569	114381029	-0.504869524	0.000976091	0.946831
X11718075_at	IL4R	chr16	27373573	27376099	-0.349058095	0.001004522	0.946831
X11740977_a_at	ZNF574	chr19	42582739	42585717	0.423873333	0.00106909	0.946831
X11729671_s_at	FAM200B	chr4	15687859	15690724	-0.400123333	0.00108091	0.946831
X11715169_at	PLEKHG2	chr19	39914373	39919054	0.271612381	0.001086907	0.946831
X11762044_a_at		chr19	50476300	50476379	0.726696667	0.001104771	0.946831
X11738424_at	BMP10	chr2	69092613	69093703	0.328448095	0.001110323	0.946831
X11715806_a_at	SEPW1	chr19	48284365	48284645	-0.633579048	0.001126817	0.946831
X11718950_s_at	MPP5	chr14	67799497	67802536	-0.277761429	0.001147089	0.946831
X11718157_s_at	PTX3	chr3	157160155	157161417	0.26869	0.001151166	0.946831
X11728707_a_at	EMP2	chr16	10622279	10626949	0.222689524	0.001163692	0.946831
X11743328_a_at	RBBP6	chr16	24580066	24581820	0.47306381	0.001167839	0.946831
X11763461_x_at	PLCB3	chr11	64034825	64035395	0.322246667	0.001172763	0.946831
X11717730_s_at	SQLE	chr8	126033995	126034525	-0.208751429	0.001177493	0.946831
X11748418_a_at	RSPO1	chr1	38076951	38078593	0.324177143	0.001209425	0.946831
X11727339_a_at	ZBTB2	chr6	151685252	151688027	-0.436278095	0.001232832	0.946831
X11725511_a_at	SLC38A4	chr12	47158546	47168956	0.259348095	0.001235417	0.946831
X11748634_a_at	ZCCHC24	chr10	81142081	81146214	0.463185714	0.001237709	0.946831
X11731296_a_at	SPATA6L	chr9	4598326	4618110	0.30629619	0.001269753	0.946831
X11763571_at		chr1	53580248	53581904	0.353449524	0.001269863	0.946831
X11721596_s_at	CDYL	chr6	4954132	4955785	0.907460476	0.001272092	0.946831
X11744274_at	NDC80	chr18	2608699	2616634	-0.292049524	0.001292978	0.946831
X11752195_s_at	TAX1BP1	chr7	27868247	27868880	-0.262581429	0.001294817	0.946831
X11718638_a_at	ELP6	chr3	47537130	47537687	0.445229524	0.001296358	0.946831
X11744911_a_at	RIC3	chr11	8160601	8161740	0.298277143	0.00135359	0.972425
X11719535_a_at	AP3M2	chr8	42026479	42028701	0.33133	0.001394001	0.972612
X11721611_a_at	IPO11	chr5	61847050	61887499	-0.557774762	0.001445969	0.972612
X11734983_a_at	GPRASP1	chrX	101908508	101914008	-0.303210476	0.001466821	0.972612
X11719601_a_at	MOXD1	chr6	132617194	132619094	0.25399619	0.001527923	0.972612

X11729134_s_at	STK24	chr13	99102455	99105463	0.281950952	0.001584065	0.972612
X11718713_a_at	SFSWAP	chr12	132249001	132262875	0.174039048	0.001613215	0.972612
X11742631_at		chr11	48366900	48367505	0.299267619	0.001634535	0.972612
X11719608_a_at	BCAR3	chr1	94027347	94027976	0.267894762	0.00164268	0.972612
X11732165_x_at	LMF1	chr16	903634	904706	0.481082857	0.001683827	0.972612
X11757625_s_at	CD200	chr3	112080390	112081659	-0.212637619	0.001700372	0.972612
X11722492_x_at	SEZ6L2	chr16	29882480	29883049	-0.318754762	0.001707191	0.972612
X11756341_a_at	RBM23	chr14	23369854	23370942	0.27126381	0.001755072	0.972612
X11734511_at	ASCL4	chr12	108168162	108170421	0.231417143	0.001759277	0.972612
X11741941_at	GML	chr8	143921832	143928262	0.449686667	0.001765515	0.972612
X11743531_at	AF238380.3	chrX	49644470	49647166	0.45235381	0.001800042	0.972612
X11734136_at	MTBP	chr8	121530943	121535875	-0.229165238	0.001803633	0.972612
X11735470_a_at	LGALS14	chr19	40197237	40200087	0.187409048	0.001813547	0.972612
X11753393_a_at	CA6	chr1	9017196	9031040	-0.315138095	0.001915303	0.972612
X11720046_x_at	DNAJC10	chr2	183641848	183644334	-0.292651429	0.001925568	0.972612
X11736458_x_at	KCNK6	chr19	38817820	38819660	0.382955714	0.001957909	0.972612
X11731374_a_at	IL22RA2	chr6	137464968	137466910	0.211147143	0.001987976	0.972612
X11737532_a_at	FMO5	chr1	146672734	146673086	-0.234817143	0.002169705	0.972612
X11725137_x_at	SYNGR1	chr22	39770321	39774386	0.367552857	0.002179337	0.972612
X11723902_at	PLEKHG1	chr6	151160969	151164799	-0.264667143	0.002186768	0.972612
X11752327_a_at	ZNF238	chr1	244217090	244220778	-0.392979048	0.002228575	0.972612
X11756913_s_at	NOD1	chr7	30464143	30465326	0.275504762	0.002252491	0.972612
X11743721_at	LGALS1	chr22	38071615	38074661	-0.405049048	0.002260195	0.972612
X11724938_at	DLL1	chr6	170591294	170591967	0.30033	0.00227028	0.972612
X11729725_at	ZBTB6	chr9	125670335	125674360	-0.314026667	0.002303202	0.972612
X11727058_at	MRPS36	chr5	68525035	68525956	-0.272352381	0.002320279	0.972612
X11716456_at	RHOBTB3	chr5	95128763	95132071	0.324785714	0.002337871	0.972612
X11758157_s_at	COQ10B	chr2	198338481	198340032	-0.255466667	0.002351496	0.972612
X11747465_a_at	NDST1	chr5	149931315	149937773	-0.405833333	0.002356536	0.972612
X11756032_a_at	CBR3	chr21	37518374	37518864	0.333822381	0.002361204	0.972612
X11720273_at	SFT2D3	chr2	128458597	128461385	0.340614286	0.002362251	0.972612
X11739636_s_at	CDK12	chr17	37686857	37690797	-0.38296	0.002447918	0.972612
X11746040_at	ZFHX2	chr14	23990069	23999645	0.284611429	0.002448079	0.972612
X11758715_s_at	DEFB126	chr20	126056	126392	0.273578571	0.002484351	0.972612
X11721849_x_at	REEP2	chr5	137781372	137782658	0.356548571	0.002498579	0.972612

Supplemental Table 5 Continued

Results of the enrichment analysis of GO processes in the expression data from MZ pairs who became discordant after blood draw

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0044802	single-organism membrane organization	6.46E-05	7.85E-01	2.09	16602	302	1024	39
GO:0061024	membrane organization	1.38E-04	8.39E-01	1.8	16602	462	1078	54
GO:0030534	adult behavior	1.50E-04	6.07E-01	6.31	16602	134	157	8
GO:0014809	regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion	2.51E-04	7.62E-01	86.92	16602	2	191	2
GO:2001287	negative regulation of caveolin-mediated endocytosis	2.55E-04	6.19E-01	85.58	16602	2	194	2
GO:0008150	biological_process	2.61E-04	5.29E-01	1.05	16602	13924	1079	951
GO:0035265	organ growth	3.17E-04	5.49E-01	7.88	16602	15	843	6
GO:0021680	cerebellar Purkinje cell layer development	4.14E-04	6.28E-01	33.74	16602	12	123	3
GO:0022612	gland morphogenesis	4.40E-04	5.94E-01	7.36	16602	15	902	6
GO:0007006	mitochondrial membrane organization	4.45E-04	5.41E-01	3.76	16602	47	1126	12
GO:0006883	cellular sodium ion homeostasis	5.00E-04	5.52E-01	10.19	16602	5	1304	4
GO:0033554	cellular response to stress	7.36E-04	7.46E-01	1.34	16602	1078	1731	151
GO:0051205	protein insertion into membrane	7.86E-04	7.35E-01	8.65	16602	13	738	5
GO:0006974	cellular response to DNA damage stimulus	8.57E-04	7.44E-01	1.47	16602	597	1726	91
GO:0061053	somite development	8.62E-04	6.98E-01	85.8	16602	9	43	2
GO:0071451	cellular response to superoxide	8.63E-04	6.55E-01	5.82	16602	11	1557	6
GO:0019430	removal of superoxide radicals	8.63E-04	6.17E-01	5.82	16602	11	1557	6
GO:0051716	cellular response to stimulus	8.78E-04	5.93E-01	1.16	16602	4545	1473	468
GO:0051336	regulation of hydrolase activity	8.81E-04	5.64E-01	1.47	16602	901	1180	94
GO:1902667	regulation of axon guidance	8.92E-04	5.42E-01	6.49	16602	16	959	6

Results of the enrichment analysis of GO functions in the expression data from MZ pairs who became discordant after blood draw

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0008750	NAD(P)+ transhydrogenase (AB-specific) activity	6.02E-05	2.39E-01	16,602.00	16602	1	1	1
GO:0003957	NAD(P)+ transhydrogenase (B-specific) activity	6.02E-05	1.20E-01	16,602.00	16602	1	1	1
GO:0008746	NAD(P)+ transhydrogenase activity	2.39E-04	3.16E-01	8,301.00	16602	2	1	1
GO:0016652	oxidoreductase activity, acting on NAD(P)H, NAD(P) as acceptor	2.39E-04	2.37E-01	8,301.00	16602	2	1	1
GO:0004714	transmembrane receptor protein tyrosine kinase activity	2.99E-04	2.38E-01	4.07	16602	64	764	12
GO:0008331	high voltage-gated calcium channel activity	3.72E-04	2.46E-01	15.3	16602	10	434	4
GO:0016936	galactoside binding	4.73E-04	2.69E-01	89.26	16602	4	93	2
GO:0035662	Toll-like receptor 4 binding	6.75E-04	3.35E-01	15.69	16602	3	1058	3
GO:0045499	chemorepellent activity	6.96E-04	3.07E-01	20.16	16602	5	494	3
GO:0005307	choline:sodium symporter activity	9.04E-04	3.59E-01	1,106.80	16602	1	15	1

Supplemental Table 5 Continued
Results of the enrichment analysis of GO components in the expression data from MZ pairs who became discordant after blood draw

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0016020	membrane	5.87E-04	8.23E-01	1.15	16602	5357	1473	545
GO:0005634	nucleus	5.90E-04	4.14E-01	1.18	16602	3981	1466	415

N - is the total number of genes
B - is the total number of genes associated with a specific GO term
n - is the number of genes in the top of the input list
b - is the number of genes in the intersection
Enrichment = (b/n) / (B/N)

Supplemental Table 5 Continued, BMI candidate genes from GWAS

Probeset ID	Gene	chr	bp start	bp end	Mean difference (Heavy twin-lean twin)	P-value
X11758687_s_at	TMEM18	chr2	667335	669675	-0.224002381	0.083281052
X11722804_x_at	PTBP2	chr1	97278829	97280349	-0.264306667	0.103622433
X11757335_s_at	MTIF3	chr13	28009776	28014586	0.328060476	0.11241314
X11748421_a_at	TMEM18	chr2	667335	669675	-0.212365238	0.1512073
X11719312_a_at	MAP2K5	chr15	68061940	68099461	0.164670476	0.163720407
X11735738_x_at	SH2B1	chr16	28884768	28885526	0.139932857	0.216469668
X11724249_a_at	SLC39A8	chr4	103188647	103189236	0.146988571	0.220316839
X11721075_a_at	MTCH2	chr11	47638868	47650907	0.136092857	0.22039422
X11757158_x_at	RPL27A	chr11	8705553	8706439	0.181799048	0.235287258
X11721866_a_at	GNPDA2	chr4	44703885	44705159	0.118775238	0.237641028
X11757157_at	RPL27A	chr11	8705553	8706439	0.184607619	0.321757416
X11723451_x_at	FTO	chr16	54145674	54155853	0.19219	0.32607976
X11722222_at	TMEM160	chr19	47549168	47551882	0.128854762	0.327178263
X11721076_at	MTCH2	chr11	47638867	47640471	-0.083596667	0.377271287
X11723450_a_at	FTO	chr16	54145674	54155853	0.136567143	0.406841576
X11755163_a_at	FANCL	chr2	58386378	58386935	-0.16898381	0.417659357
X11723064_a_at	TMEM18	chr2	667335	670975	-0.041884762	0.447702709
X11719176_a_at	GPRC5B	chr16	19868616	19871866	0.10887	0.504990939
X11762571_a_at	GNPDA2	chr4	44703885	44705159	-0.048939524	0.565483882
X11719175_a_at	GPRC5B	chr16	19868616	19873662	0.048178571	0.576411552
X11749270_a_at	GPRC5B	chr16	19868616	19871866	0.068614286	0.583801443
X11724252_s_at	SLC39A8	chr4	103182823	103184350	-0.053428571	0.605300353
X11749835_x_at	SH2B1	chr16	28883855	28885526	-0.043771905	0.644212625
X11735168_a_at	QPCTL	chr19	46205056	46207247	0.042580952	0.649132758
X11728061_a_at	ZNF608	chr5	123982372	123984826	-0.048458571	0.698498727
X11758115_s_at	PTBP2	chr1	97278829	97280349	-0.049395714	0.707704728
X11728331_a_at	KCTD15	chr19	34302696	34306666	0.027834286	0.748845008
X11740764_at	SLC39A8	chr4	103172198	103174587	0.020517143	0.818440028
X11722380_at	NUDT3	chr6	34255997	34256708	0.032593333	0.839427629
X11748612_a_at	ETV5	chr3	185764097	185775033	-0.015240952	0.862888922
X11721867_a_at	GNPDA2	chr4	44703885	44705159	0.021470476	0.894685166
X11720353_s_at	ETV5	chr3	185764097	185766649	-0.009119524	0.927578069
X11763807_at	MAP2K5	chr15	67951109	67951694	0.007815238	0.941270573
X11748888_a_at	FTO	chr16	53967897	54155853	0.006940952	0.969651971

Enrichment analysis output GSEA for BMI GWAS gene set

SIZE	ES	NES	NOM p-val	FDR q-val	FWER	RANK AT	LEADING EDGE
18	0.23349632	0.580054	0.9848024	0.9848024	0.972	12650	tags=67%, list=69%,

Supplemental Reference List

1. Hsieh,S.D., Muto,T., Murase,T., Tsuji,H., & Arase,Y. Eating until feeling full and rapid eating both increase metabolic risk factors in Japanese men and women. *Public Health Nutr.* **14**, 1266-1269 (2011).

2. Maruyama,K. *et al.* The joint impact on being overweight of self reported behaviours of eating quickly and eating until full: cross sectional survey. *BMJ* **337**, a2002 (2008).

3. Ochiai,H. *et al.* Eating behavior and childhood overweight among population-based elementary schoolchildren in Japan. *Int. J. Environ. Res. Public Health* **9**, 1398-1410 (2012).
4. Ochiai,H. *et al.* The impact of eating quickly on anthropometric variables among schoolgirls: a prospective cohort study in Japan. *Eur. J. Public Health* **24**, 691-695 (2014).
5. Connor,G.S., Schofield-Hurwitz,S., Hardt,J., Levasseur,G., & Tremblay,M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine. Tob. Res.* **11**, 12-24 (2009).
6. Kentala,J., Utriainen,P., Pahkala,K., & Mattila,K. Verification of adolescent self-reported smoking. *Addict. Behav.* **29**, 405-411 (2004).
7. Taylor,A.E. *et al.* Investigating the possible causal association of smoking with depression and anxiety using Mendelian randomisation meta-analysis: the CARTA consortium. *BMJ Open.* **4**, e006141 (2014).
8. de Geus,E.J., Bartels,M., Kaprio,J., Lightfoot,J.T., & Thomis,M. Genetics of regular exercise and sedentary behaviors. *Twin. Res. Hum. Genet.* **17**, 262-271 (2014).
9. de Moor,M.H.M. & de Geus,E.J.C. Genetic influences on exercise behavior. in *Lifestyle Medicine* (ed. Rippe,J.M.) 1367-1378 (Taylor & Francis, Boca Raton, FL, USA, 2013).
10. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* **13**, 231-245 (2010).
11. Friedewald,W.T., Levy,R.I., & Fredrickson,D.S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18**, 499-502 (1972).
12. Jansen,R. *et al.* Sex differences in the human peripheral blood transcriptome. *BMC. Genomics* **15**, 33 (2014).
13. Li,J. & Ji,L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb.)* **95**, 221-227 (2005).
14. Speliotes,E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937-948 (2010).
15. Mootha,V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267-273 (2003).
16. Subramanian,A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A* **102**, 15545-15550 (2005).

Appendix 4

Genome-wide DNA Methylation data NTR

Subjects and samples

The subjects in this study participated in the Netherlands Twin Register (NTR)[1, 2] biobank project[3]. Most subjects were twin pairs. A small number of subjects were family members of twins who were enrolled in the Genome of the Netherlands project (GONL[4]; which included parent-offspring trios where the offspring where twins or siblings of twins), or in the Complete Genomics project (CG; which included a few spouses of twins). Venous blood samples were drawn in the morning after an overnight fast, and separate EDTA tubes were collected for isolation of DNA and assessment of haematological profiles. Blood sampling and buccal sample collection procedures were described in detail previously[3].

For the genome-wide methylation study 3264 blood samples from 3221 subjects were assessed: For 3178 subjects, a single blood sample was assessed and two blood samples were assessed for 43 subjects who participated twice in the NTR biobank (Biobank 1 and Biobank 2). After quality control (QC) of the methylation data, genome-wide methylation data were available for 3089 peripheral blood samples (mean age at blood draw=36.9, SD 12.9, range 17-79) from 3057 subjects (65.7% female; 34.3% male; GONL=331, CG=16) of which 32 subjects had methylation data for two samples (mean follow-up time=5 years, range=2.8-7.0). Subjects came from 1601 NTR families (Table 1). The post-QC whole blood DNA methylation dataset included 1250 complete twin pairs (MZ:803 (male= 239, female= 564), DZ:447 (male= 97,female=193, opposite-sex=157).

Table 1: Composition of families included in the whole blood genome-wide DNA methylation dataset (after DNA methylation Quality Control).

Description	Family size*					
	1	2	3	4	5	Total N families
1 twin	268	0	0	0	0	268
2 twins or 2 sibs	0	1217	0	0	0	1217
1 parent	7	0	0	0	0	7
1 parent + 1 child	0	2	0	0	0	2
1 parent + 2 children	0	0	3	0	0	3
2 parents, no children	0	8	0	0	0	8
2 parents, 1 child	0	0	64	0	0	64
2 parents, 2 children	0	0	0	25	0	25
2 parents, 3 children	0	0	0	0	4	4
1 spouse	1	0	0	0	0	1
1 spouse (of twin) + 2 twins	0	0	2	0	0	2
Total N families	276	1227	69	25	4	1601
Total N subjects	276	2454	207	100	20	3057

* Number of individuals per family

Genotype data used for imputation against the GONL reference panel for the BIOS consortium

Of the 3221 subjects for whom peripheral blood methylation samples were assessed with the illumina 450k array, 349 subjects were part of GONL, of whom 347 subjects were successfully sequenced (two subjects were excluded during quality control of the sequence data). Of the remaining 2872 subjects who were not part of the GONL study, 2665 subjects had been previously genotyped or had a MZ co-twin who had been genotyped one or multiple times on any of the following genotype arrays: Affymetrix6, Affymetrix 5, and Illumina660. The SNP data from the Affymetrix6, Affymetrix 5, and Illumina660 array were used for imputation against the GONL reference panel by the Biobank-based Integrative Omics Study (BIOS) consortium, in which a subset of the NTR DNA methylation samples takes part. Prior to imputation, one set of genotypes was selected (the one with the best quality) for MZ twins if both twins were genotyped and for individuals who had been genotyped on multiple platforms. Because the majority of subjects had been genotyped with Affymetrix 6, while only a small proportion had been genotyped with Affymetrix 5 or Illumina660, the individuals genotyped on Affymetrix 6 were separately imputed. The Affymetrix5 and Illumina660 data plus a small subset of subjects for whom only raw affymetrix6 genotypes were available that had not yet been cleaned (in December 2013) were merged and this combined dataset was imputed separately, after selecting the overlapping SNPs from Affymetrix5 and Illumina660 and (raw) Affymetrix 6 data. The following number of subjects were included in each set (including 1 subject from MZ pairs): affymetrix6; N=1621 subjects, merged SNP data set from other chips; N= 249 subjects. In total, there were 795 subjects whose own genotypes were not imputed but who were part of an MZ twin pair of which the co-twin was genotyped on Affymetrix6, Affymetrix5 or Illumina660.

Reference List

1. Boomsma,D.I. *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* **9**, 849-857 (2006).
2. Willemsen,G. *et al.* The adult Netherlands twin register: twenty-five years of survey and biological data collection. *Twin. Res. Hum. Genet.* **16**, 271-281 (2013).
3. Willemsen,G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* **13**, 231-245 (2010).
4. Boomsma,D.I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22**, 221-227 (2014).

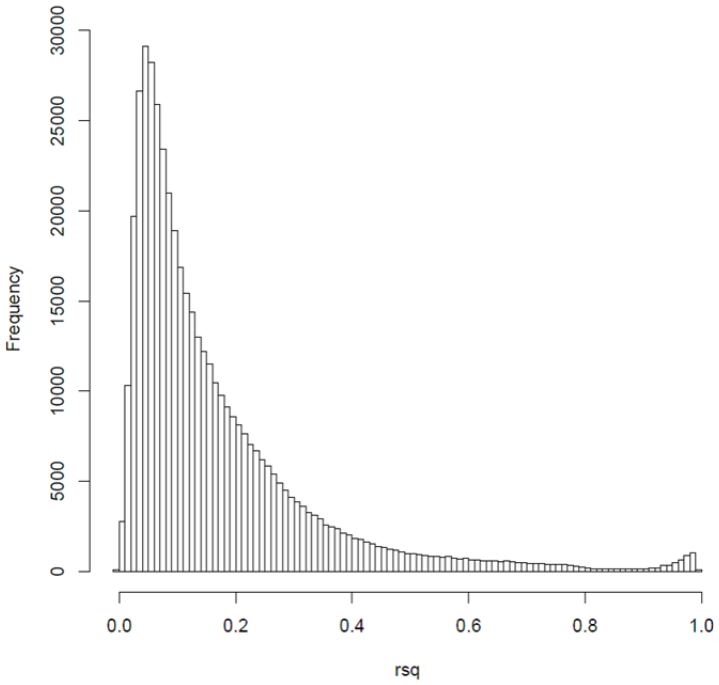
Supplement to Chapter 6

Supplementary Table 1: Twin correlations, heritability and longitudinal correlation, stratified by the amount of variation in DNA methylation between individuals.

		***r* MZ twins**			***r* DZ twins**			**Classical Twin Heritability**			**Longitudinal Correlation**		
SD of the β-value	N CpGs	mean	SD	median	mean	SD	median	mean	SD	median	mean	SD	median
>= 0 (All)	411169	0.20	0.21	0.13	0.09	0.11	0.06	0.22	0.27	0.16	0.21	0.30	0.16
>= 0-0.01	49599	0.04	0.05	0.04	0.02	0.05	0.02	0.05	0.13	0.04	0.01	0.19	0.01
>= 0.01-0.02	171461	0.09	0.09	0.08	0.04	0.07	0.04	0.10	0.17	0.09	0.07	0.20	0.07
>= 0.02-0.03	85003	0.23	0.15	0.23	0.11	0.09	0.10	0.25	0.22	0.24	0.24	0.26	0.24
>= 0.03-0.04	49918	0.35	0.19	0.36	0.15	0.11	0.15	0.38	0.25	0.38	0.39	0.28	0.43
>= 0.04-0.05	27110	0.44	0.20	0.44	0.19	0.12	0.19	0.49	0.28	0.52	0.51	0.29	0.57
>= 0.05	28078	0.57	0.25	0.60	0.25	0.13	0.25	0.64	0.33	0.73	0.64	0.31	0.77

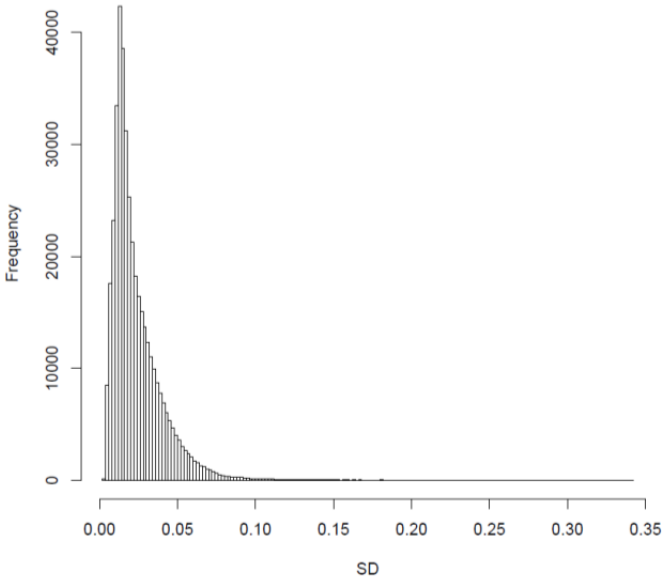
SD= Standard deviation. β-value= Methylation beta-value, which represents the proportion of DNA Methylation. The first row summarizes the results for all analyzed CpGs and all other rows summarize the results for CpGs grouped based on the standard deviation of methylation level across all subjects.

Figure S1: Histogram of the variance of DNA methylation level explained by covariates at individual genome-wide CpGs.



Rsq=Adjusted r-squared from a linear regression model with DNA methylation level at one CpG site as outcome and the following predictors: sex, age, array row, 96-wells plate (dummy coded), white blood cell percentages (neutrophils, monocytes and eosinophils; assessed at sample collection), and the first ten PCs derived from the genotype data.

Figure S2: Histogram of the standard deviation of the methylation β-value for genome-wide CpGs.



SD=Standard Deviation of DNA methylation level (β-value).

Figure S3: Histogram of the difference in heritability between females and males (h^2 in females minus h^2 in males) for 2654 CpGs with significant interaction between sex and genetic variance or between sex and unique environmental variance.

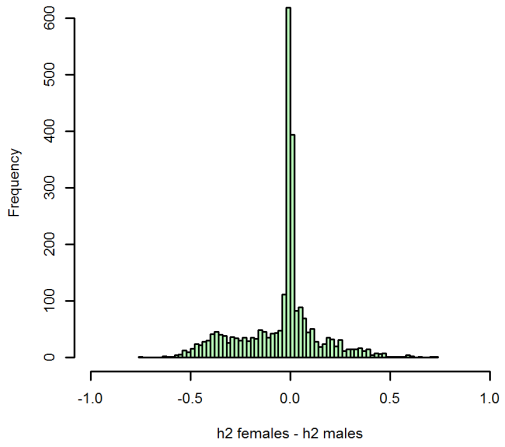
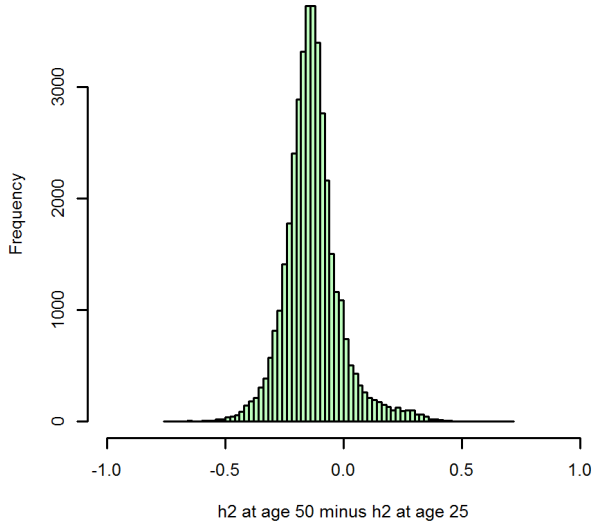


Figure S4: Histogram of the difference in (IBD-based) heritability between age 50 and age 25 (h^2 at age 50 minus h^2 at age 25) for 39194 CpGs with significant interaction between age and genetic variance or between age and unique environmental variance.



Supplementary Methods

Genome-wide SNP data

Three distinct genotype datasets were available. The first consisted of previously collected genome-wide SNP data that were only used as part of the Quality Control (QC) procedure of the DNA methylation data. The second previously collected genome-wide SNP data were used only as part of the statistical analyses of the DNA methylation data. The third SNP dataset consisted of 65 common SNPs targeted by the Illumina 450k array that were only used as part of QC procedure of the DNA methylation data.

Genotype data used during QC of the DNA methylation data

Of the 3221 subjects for whom peripheral blood methylation samples were assessed with the illumina 450k array, 2665 subjects had been previously genotyped or had a MZ co-twin who had been genotyped one or multiple times on any of the following genotype arrays: Affymetrix6, Affymetrix5, and Illumina660. One set of genotypes was selected (the one with the best quality) for MZ twins if both twins were genotyped and for individuals who had been genotyped on multiple platforms. In total, 1870 genome-wide SNP data sets were available, which were informative 2665 individuals (including 795 MZ co-twins). For the DNA methylation data QC, the overlapping SNPs from the Affymetrix6, Affymetrix5, and Illumina660 arrays were selected. Because of the small overlap of SNPs on these three arrays, this dataset was not used for the heritability analyses of DNA methylation

Genotype data used in the heritability analyses of DNA methylation

The analyses of DNA methylation heritability were performed using genome-wide SNP data collected with the Affymetrix6 array and SNP data that were extracted from whole genome sequence data that were available for a small subset of subjects (described previously)[1]. Of the 2975 subjects with good quality DNA methylation data and data on white blood cell counts, Affy6 genotype data were available for 2289 subjects and sequence data for 341 individuals (numbers include both MZ twins). Only SNPs present on the Affy6 platform were extracted from the sequence data. For a subset of 84 subjects for whom sequence data and Affy6 data were available, the sequence data were selected. SNPs with an allele frequency difference between individuals genotyped on Affy6 and individuals who were sequenced were removed (based on a p-value < $1*10^{-5}$ in a case-control genome-wide association analysis, where case-control status reflected whether a person was genotyped on Affy6 or whole-genome sequenced). The genome-wide SNP data were used to construct a genetic relatedness matrix (GRM), which summarizes overall genetic relatedness between all subjects (N=2603) based on all genotyped autosomal SNPs (MAF > 0.01) with Genome-wide Complex Trait Analysis (GCTA)[2].

DNA methylation Quality Control and filtering of methylation probes

Quality control and processing of the DNA methylation data from buccal samples has been previously described[3]. The following text describes the quality control and processing of the DNA methylation data from blood samples. The raw intensity files (idat) were imported into the R environment[4], where further processing, quality control and normalization took place using a protocol developed by the LUMC Molecular Epidemiology department.

First, the methylation data were examined with the R package MethylAid[5], which marks outlier samples for a number of quality metrics that are computed based on sample dependent and sample independent quality metrics. The performance of the 3264 samples is plotted for each of five quality metric in Supplementary Figure 5-9. Only samples that passed all five quality criteria (using the default MethylAid thresholds) were kept for further analyses. In total, 70 low-performing samples were excluded (2.1%), the majority of which failed based on multiple criteria (Supplemental Table 2). Only the 3194 samples showing good overall quality were taken on to further processing steps.

Several probe-level QC steps were performed to filter out probes with low performance. For all samples, ambiguously-mapped probes were excluded, based on the definition of an overlap of at least 47 bases per probe from Chen et al[6], and all probes containing a SNP, identified in the Dutch population[1], within the CpG site (at the C or G position) were excluded, irrespective of minor allele frequency. For each sample individually, probes with an intensity value of zero (not present on the array of a particular sample), probes with a detection P value > 0.01 (calculated using the function detectionP from the minfi package[7]), and probes with a bead count < 3 were excluded. After these steps, probes with a success rate < 0.95 across samples were removed from all samples and the success rate across probes for each sample was computed (Mean per sample success rate=99.89%, range=97.86%-99.96%). The total number of CpGs after these filtering steps was 421119. Only autosomal sites were kept in the current analyses (N=411169).

We performed several checks to confirm sample identity, by making use of previously collected genotype data, 65 SNP (control) probes targeted by the 450k array,

and differential methylation patterns in males versus females. Previously collected raw genotype data was used as input for the program MixupMapper, which computes the probability that a DNA methylation sample matches supplied genotype information based on mQTLs estimated from the dataset[8]. To confirm sex, we clustered samples based on their methylation data, by calculating the Euclidean distance from the pair-wise correlations between samples followed by hierarchical clustering (cluster method=complete linkage). Clustering based on all probes and based on probes on the sex chromosomes only yielded similar results. We computed the correlation between samples for 65 SNP (control) probes targeted by the 450k array to confirm zygosity of twins, and to confirm that longitudinal samples indeed belonged to the same person. Finally, we used the 65 SNP probes to examine potential contamination of samples with foreign DNA, by computing the number of SNPs per sample with an unclear genotype (which we defined as SNPs where the proportion of signal from each allele lay between 0.2 and 0.4 or between 0.6 and 0.8, on a scale from o to 1, i.e. a pattern not clearly supporting membership to any of the three genotype classes. The number of 'unclear genotypes' showed a mean of 3.3 across all samples (median=2, SD=3.5, Supplemental Figure 10). We excluded samples with >= 15 unclear genotypes (99th percentile). The genome-wide methylation distribution of these excluded samples showed relatively more intermediate methylation levels (Supplemental Figure 11). An example scatterplot of the 65 SNP probes in MZ twin samples illustrating DNA contamination of the sample of one of the twins, as detected by this method, is given in Supplementary Figure 12.

In total, 132 samples were involved in at least one of the following issues: genotype mismatch, sex-mismatch, DNA contamination, inconsistent SNP probe correlation (either between twins or between longitudinal samples from the same person). After solving a swap between 2 methylation samples identified by MixupMapper (and confirmed by the other checks) by re-swapping methylation data IDs (leaving 128 samples with issues), 67 samples were excluded based on the following grounds: only sex mismatch (22 samples), only genotype mismatch (10 samples), only DNA contamination (27 samples), genotype + sex mismatch (6 samples), DNA contamination + sex mismatch (2 samples). After removal of these samples, there were still 38 samples with an inconsistent SNP probe correlation (involving i.e. a zygosity mismatch or mismatch between longitudinal samples), which were all excluded, giving a total of 105 samples (3.3%) excluded based on failed identity or contamination, on top of the 70 samples excluded based on bad quality of the methylation data.

Finally, for 22 persons with 450k methylation data available from blood and buccal samples, the 65 SNP probes confirmed that blood and buccal samples indeed belonged to the same individual.

Exploration of technical and biological confounding

To get an impression of the impact of technical and biological effects on overall variation in methylation, Principle Component Analysis (PCA) was performed on the raw genome-wide methylation data (Supplemental Table 3, Supplemental Figure 13), and the correlation between PC scores and several known technical batches and biological outcomes were computed. PC1 related to sex (r=0.92), PC2 was strongly correlated with position on the array (in particular, array row, r=0.50), PC3 with several white blood cell counts (e.g. lymphocytes: r=0.45), and PC4 with age (r=-0.59). Other batch variables (e.g. 96-wells plate, array, scanner) correlated to a smaller degree with multiple components.

Normalization of the methylation data and correction for covariates

To reduce technical variability between samples while retaining as much biological variation in DNA methylation as possible, the data were normalized with Functional Normalization, a between-sample normalization method that normalizes the data using PCs (the number of which is user-specified) estimated from control probes that are specifically designed not to measure biological variation in samples[9]. We chose to perform Functional Normalization with the first 4 PCs, because PCA based on the data from control probes showed that in our data only the first four PCs had an eigen value > 1 (Supplemental Table 4).

Normalized intensity values were converted into beta-values (β) and M-values[10]; β -values were used for descriptive purposes only because of their biological interpretability, while M-values were used as input for all analyses. The β -value, which represents the methylation level at a CpG for an individual and ranges from 0 to 1, is calculated as:

$\beta = M/(M+U+\alpha)$

where M=Methylated signal, U=Unmethylated signal, and α represents a correction term (100 by default) to control the β –value of probes with very low overall signal intensity (i.e. probes for which M+U~0 after normalization).

The M-value is equivalent to a log2 logistic transformation of β:

$M=\log2(M+\alpha/U+\alpha) = \log2(\beta/1- \beta)$.

Supplemental Table 2: Number of bad quality DNA methylation samples that failed sample quality checks.

Quality Metric	**N outliers[A]**
MU	47
BS	52
NP	44
HC	22
DP	48
Combinations of failure	**N outliers[B]**
BS	1
HC	14
HC + BS	5
HC + BS + MU	2
DP	3
DP + DS + MU	1
DP + NP + MU	1
DP + NP + MU + BS + MU	42
DP + NP + MU + BS + MU + HC	1
Total N bad quality samples[C]	70

The following five quality metrics were computed with the R package MethylAid: **MU=** median Methylated versus Unmethylated signal intensity, **BS=**Efficiency of bisulphite conversion, **NP=**overall quality based on sample-dependent control probes (non-polymorphic quality control probes), **HC=** overall quality based on sample-independent hybridization control probes. **DP=** Fraction of probes per sample where the signal exceeds the background signal, as assessed with the detection p-value, which uses the negative control probes to assess background signal.

[A]N outliers= Number of samples that failed based on each quality metric.

[B]N outliers= Number of samples that failed based on a particular combination of multiple quality metrics.

[C]All samples that failed based on one or more of the five quality metrics were discarded (70 samples).

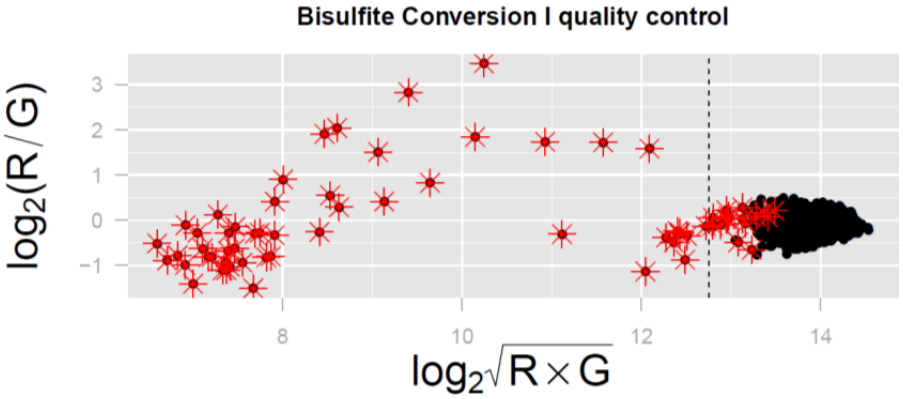
Supplemental Table 3: Eigen values and proportion of variance explained by Principle Components 1 to 15, calculated based on the raw genome-wide methylation data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
Eigen value	54.71	37.13	14.09	5.65	3.95	2.71	1.51	1.23	0.98	0.95	0.68	0.65	0.55	0.49	0.46
Proportion of Variance	0.2211	0.1501	0.0569	0.0228	0.0160	0.0110	0.0061	0.0050	0.0040	0.0038	0.0028	0.0026	0.0022	0.0020	0.0019
Cumulative Proportion	0.2211	0.3712	0.4281	0.4510	0.4669	0.4779	0.4840	0.4890	0.4929	0.4968	0.4995	0.5021	0.5044	0.5063	0.5082

Supplemental Table 4: Eigen values and proportion of variance explained by Principle Components 1 to 10, calculated based on control probes from the methylation array.

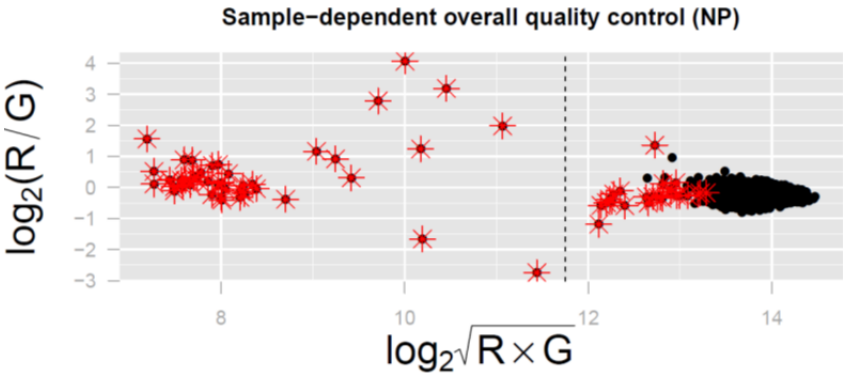
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Eigen value	25.37	7.09	3.61	1.36	0.73	0.48	0.45	0.36	0.30	0.25
Proportion of Variance	0.6039	0.1687	0.0860	0.0323	0.0175	0.0114	0.0108	0.0087	0.0071	0.0060
Cumulative Proportion	0.6039	0.7727	0.8587	0.8910	0.9085	0.9199	0.9307	0.9393	0.9465	0.9524

Figure S5: Quality control plot of bisulfite conversion.



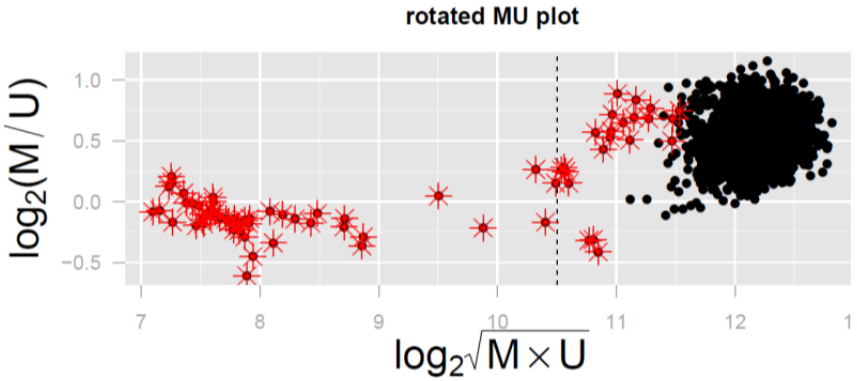
The performance of bisulfite conversion quality control probes is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics. R=Red Channel. G=Green Channel.

Figure S6: Quality control plot of overall sample quality based on sample-dependent control probes (Non-Polymorphic quality control probes)



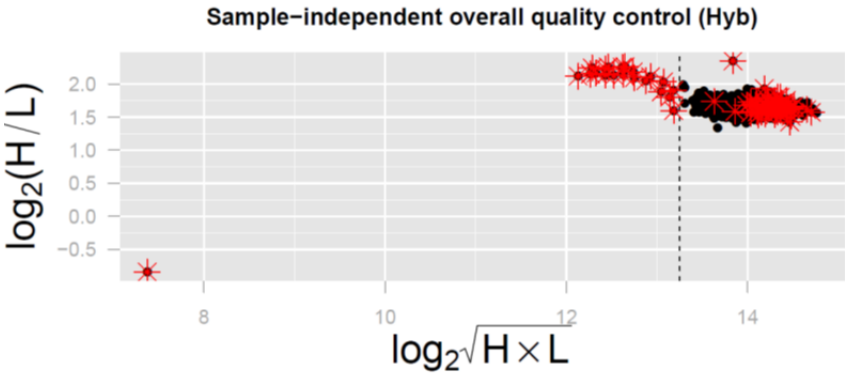
The performance of Non-Polymorphic quality control probes is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics. R=Red Channel. G=Green Channel.

Figure S7: Quality control plot of the median Methylated versus Unmethylated signal intensity.



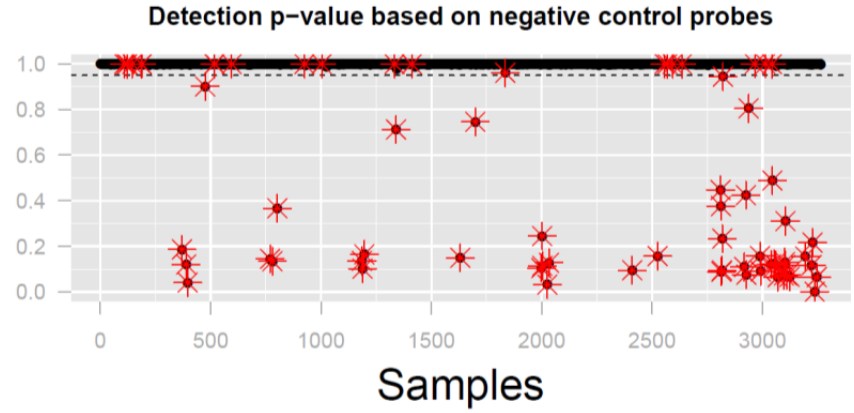
The relationship between the Median Methylated (M) and Unmethylated (U) signal intensity is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics.

Figure S8: Quality control plot based on sample-independent hybridization control probes.



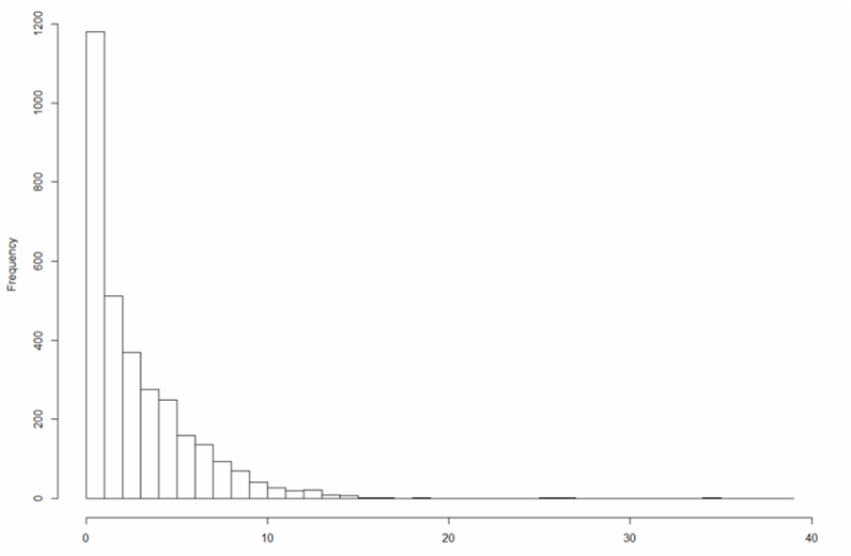
The performance of sample-independent hybridization control probes is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics.

Figure S9: Quality control plot showing the proportion of probes with a detection p-value < 0.01 within samples.



For all methylation samples, the proportion of probes per sample with a detection p-value < 0.01 is plotted (y-axis). The detection p-value indicates whether the probe signal exceeds the background signal, where the background signal is calculated using the negative control probes. Red stars denote samples that failed on the basis of any of the five quality metrics.

Figure S10: Histogram of the number of Illumina 450k SNP probes per sample displaying an unclear genotype.



X-axis= the number of unclear genotype per sample: SNPs where the proportion of signal from each allele lay between 0.2 and 0.4 or between 0.6 and 0.8, on a scale from o to 1, i.e. a pattern not clearly supporting membership to any of the three genotype classes. In total 65 common SNPs from the Illumina 450k array were assessed. Y-axis=Number of methylation samples. Methylation samples with >= 15 unclear genotypes (99th percentile) were excluded from analyses.

Figure S11: DNA methylation density plot showing samples excluded based on suspected DNA contamination (1/orange) and all other samples (0/green).

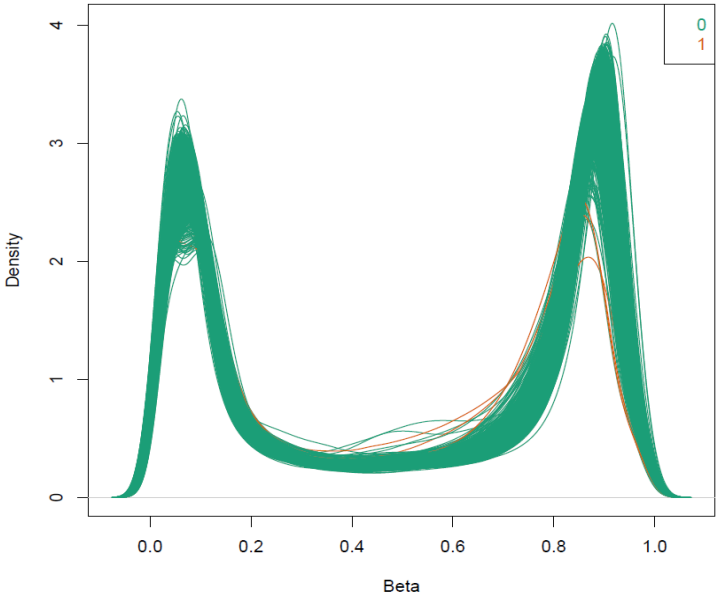


Figure S12: Example scatterplot of the 65 Illumina 450k SNP probes in one pair of MZ twins, of which one twin (on the y-axis) was excluded based on suspected DNA contamination.

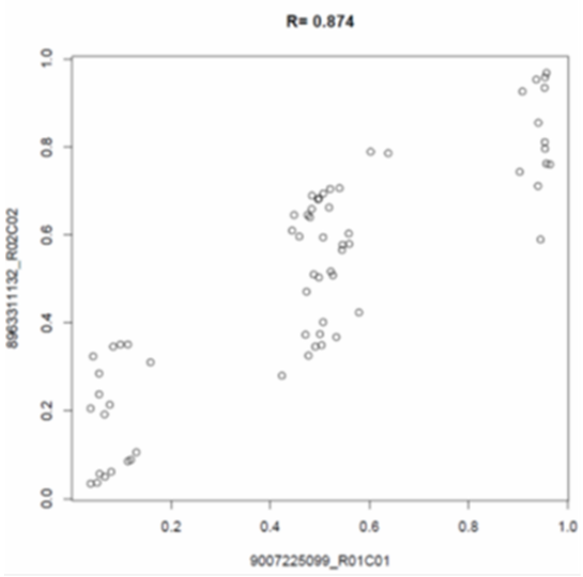
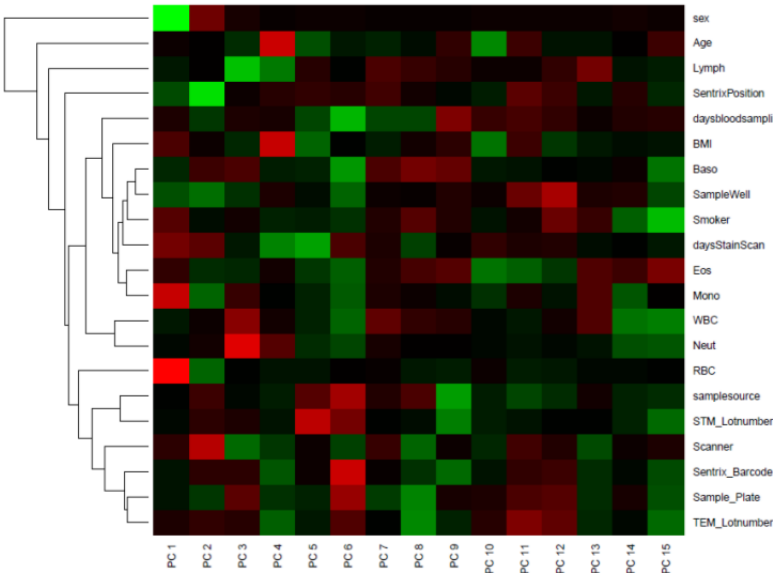


Figure S13: Heatmap depicting the correlation between the first 15 Principal Components from the raw genome-wide methylation data (x-axis) and technical batches and biological effects (y-axis).



Stronger green=larger positive correlation. Stronger red= larger negative correlation. Lymph=lymphocyte counts. Daysbloodsmpli=Days between blood sampling and hybridization. BMI=Body Mass Index. Baso=Basophil count. daysStainScan=Days

between staining and scanning. Eos=Eosinophil count. Mono=Monocyte count. WBC=Total white blood cell count. Neut=Neutrophil count. RBC=Red Blood cell count. Samplesource= 1 or 2, for individuals with 2 longitudinal samples (both from blood).

Reference List

1. Genome of the Netherlands Consortium Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**, 818-825 (2014).
2. Yang,J., Lee,S.H., Goddard,M.E., & Visscher,P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).
3. van Dongen,J. *et al.* Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. *Genes (Basel)* **5**, 347-365 (2014).
4. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 2013.
5. van Iterson M. *et al.* MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics.*(2014).
6. Chen,Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* **8**, 203-209 (2013).
7. Aryee,M.J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics.* **30**, 1363-1369 (2014).
8. Westra,H.J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics.* **27**, 2104-2111 (2011).
9. Fortin,J.P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *bioRxiv*(2014).
10. Du,P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC. Bioinformatics.* **11**, 587 (2010).

Supplement to chapter 7

Supplemental Methods

To gain insight into the degree to which our data may be influenced by variation in the cellular composition of buccal swab samples, we first looked at the beta-values of our twin samples at cg18384097; a CpG site located in the *PTPN7* (protein tyrosine phosphatase non-receptor type 7) gene. Based on a comparison of publicly available DNA methylation data from blood samples and buccal swab samples [1-3], cg18384097 was found to be one of the most differentially methylated CpGs between buccal swabs samples and blood (beta-value buccal = 0.82 and beta-value blood = 0.05), suggesting that this CpG could be a suitable marker to obtain insight into the relative proportions of buccal *versus* blood cells within a sample [4]. In our buccal data from twins, the average beta-value of this CpG was 0.89. Beta-values for this CpG for all twin samples are presented in Table S3. Table S3 indicates that the majority of twins had similar methylation levels at this CpG, although some variation is evident. MZ twin 3.2 had the most deviant methylation value, which suggests that the sample of this individual may have contained relatively more blood cells compared to the other samples.

To examine this further, we next turned to another reference dataset consisting of 450k methylation data on multiple tissue types [5] and selected from this dataset all CpGs with an average beta-value difference >.6 between buccal swab samples (N = 5) and blood samples (N = 5), in order to obtain a set of CpGs of which the methylation value measured in buccal swab samples is presumably reflective of the amount of buccal epithelial cells *versus* blood cells present in buccal swab samples. This selection yielded 881 CpGs (*p*-value range: $2.58 \times 10^{-5} - 7.15 \times 10^{-4}$), after excluding probes containing SNPs in the CpG site. For this set of 881 CpGs, we plotted the methylation beta-values of our buccal samples from twins together with beta-values from the reference set [5] for buccal swab samples (N = 5), blood samples (N = 5) and saliva samples (N = 5) in a heatmap (Figure S2). Figure S2 illustrates that some variation in methylation level at these CpGs is present between the buccal swab samples: some buccal samples show relatively more intermediate methylation, a pattern that is more similar to the saliva reference samples, and is suggestive of a higher proportion of blood cells in the sample. Based on methylation levels at this set of CpGs that differentiates strongly between buccal and blood samples, two smaller clusters were identified among buccal swab samples, containing twin Samples 3.2, 6.2, 7.1, 10.1 and 10.2 and reference buccal Samples 3 and 5 (Figure S2), which demonstrated more intermediate methylation values compared to the other buccal samples, suggesting that the samples in the two small clusters (twin Samples, 3.2, 6.2, 7.1, 10.1 and 10.2) contained higher proportions of leukocytes compared to the other buccal samples. Twin Samples, 3.2, 6.2, 7.1, 10.1 and 10.2 also showed the lowest methylation beta-values at cg18384097 (Table S3). To examine the extent to which our analyses may be affected by heterogeneity across samples related to cell type proportions, we repeated our analyses with twin Pairs 3, 6, 7 and 10 excluded, thus keeping only the most homogenous samples that seemed to have the highest buccal epithelial cell content (based on the approach illustrated in Figure S2) in the analyses, which yielded highly similar results (see Table S1 and Figure S1).

Table S1. Correlations between methylation values of twins, with twin Pairs 3, 6, 7 and 10 excluded.

Category	N CpGs	Mean rho	Median rho	Min rho	Max rho
All CpGs	59,041	0.57	0.60	−1	1
Gene-centric annotations	N CpGs (%)	Mean rho	Median rho	Min rho	Max rho
Intergenic (>10 kb from TSS)	11,430 (19.4%)	0.56	0.60	−1	1
Distal Promoter (−10 kb to −1.5 kb from TSS)	3193 (5.4%)	0.57	0.60	−0.89	1
Proximal Promoter (−1.5 kb to +500 bp from TSS)	17,880 (30.3%)	0.60	0.66	−0.94	1
Gene Body (+500 bp to 3' end)	25,163 (42.6%)	0.56	0.60	−1	1
Downstream region (3' end to +5 kb from 3' end)	1,375 (2.3%)	0.58	0.66	−0.94	1
CGI annotations	N CpGs (%)	Mean rho	Median rho	Min rho	Max rho
CGI	10,576 (17.9%)	0.66	0.77	−1	1
CGI shore	14,803 (25.1%)	0.57	0.60	−0.89	1
CGI shelf	6,001 (10.2%)	0.55	0.60	−1	1
Non-CGI	27,661 (46.9%)	0.55	0.60	−1	1
Methylation level	N CpGs (%)	Mean rho	Median rho	Min rho	Max rho
Hypomethylated (average beta <0.3)	17,581 (29.8)	0.58	0.60	−0.94	1
Intermediately methylated (average beta ≥0.3–0.7)	29,519 (50.0)	0.58	0.60	−1	1
Hypermethylated (average beta ≥0.7)	11,941 (20.2)	0.54	0.60	−1	1

Table S2. Names and number of analyzed CpGs of imprinted genes from Yuen *et al*. [6]

Gene No.	Gene Name	N CpGs analyzed [A]	Gene No.	Gene Name	N CpGs analyzed [A]
1	*ABCA1*	1	24	*NNAT*	3
2	*ANKRD11*	1	25	*OSBPL5*	1
3	*ATP10A*	4	26	*PEG10*	5
4	*CALCR*	5	27	*PEG3*	1
5	*CDKN1C*	1	28	*PHLDA2*	7
6	*COPG2*	1	29	*PLAGL1*	3
7	*DDC*	2	30	*PPP1R9A*	2
8	*DLX5*	4	31	*PRIM2A*	3
9	*GNAS*	17	32	*RBP5*	1
10	*GRB10*	7	33	*SGCE*	2
11	*H19*	4	34	*SLC22A18*	1
12	*IGF2*	1	35	*SLC22A18AS*	1
13	*IGF2AS*	1	36	*SLC22A2*	4
14	*IGF2R*	2	37	*SLC22A3*	1
15	*INPP5F*	1	38	*SNRPN*	2
16	*KCNQ1*	1	39	*SNURF*	6
17	*KCNQ1DN*	1	40	*TCEB3C*	1
18	*KLF14*	9	41	*TFPI2*	5
19	*L3MBTL*	2	42	*TP73*	5
20	*MAGEL2*	1	43	*UBE3A*	4
21	*MEG3*	3	44	*WT1*	3
22	*MEST*	2	45	*ZIM2*	6
23	*NDN*	2	46	*ZNF264*	4
				Total:	144

[A] The number of CpGs in each of the imprinted genes described by Yuen *et al.*, which showed an intermediate methylation level in our buccal data from twins (mean β ≥ 0.3–0.7 across subjects).

Figure S1. MZ twin correlations for individual CpGs grouped by genomic regions and average methylation level, with twin Pairs 3, 6, 7 and 10 excluded. Hypo = hypomethylated. Inter = intermediate methylation. Hyper = hypermethylated.

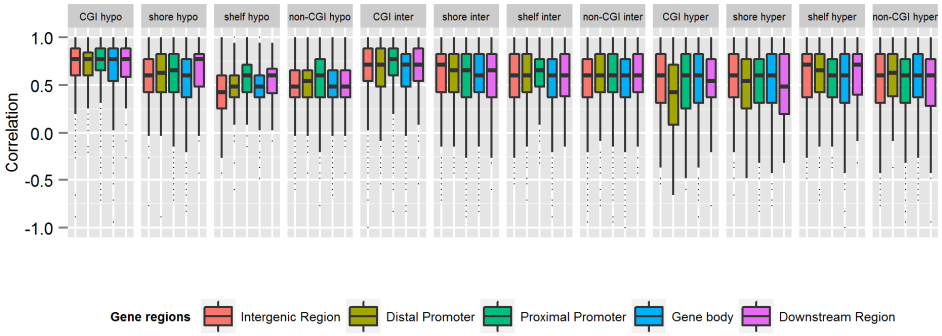


Figure S2. Beta-values of twin samples (buccal swabs) and reference data from five blood samples, five saliva samples and five buccal samples for CpGs with an average (absolute) beta-value difference >0.6 between blood and buccal in the reference data. Brightest yellow: beta-value = 1. Brightest blue: beta-value = 0. The clustering of CpGs (N = 881,
y-axis) and samples (x-axis) was performed using complete linkage based on the Euclidian distance between beta-values.

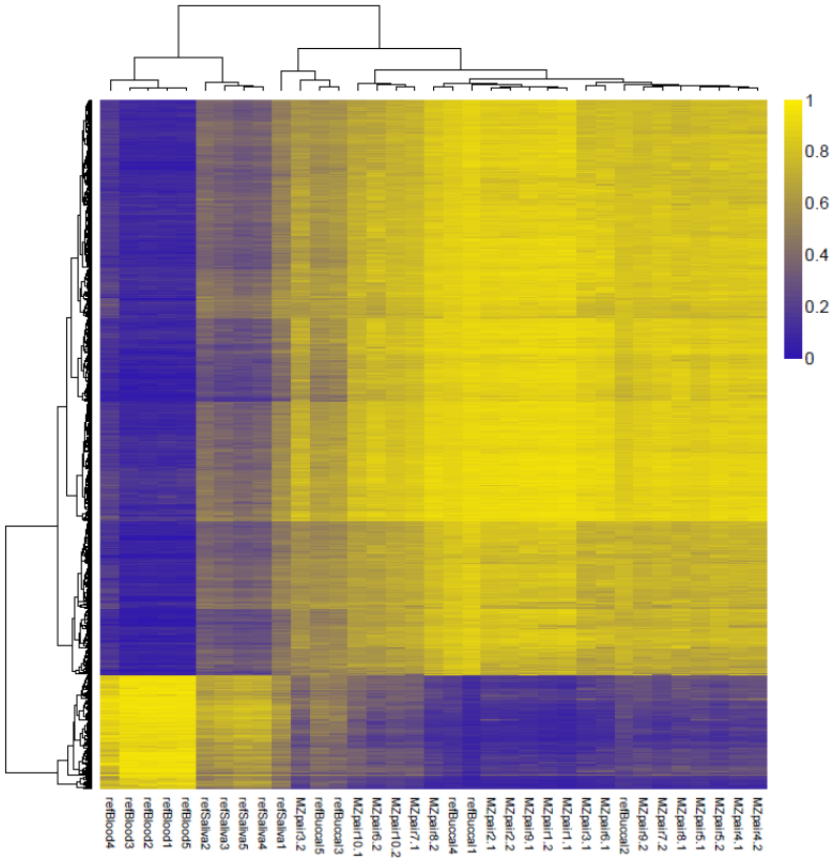


Table S3. Beta-values of twin samples for cg18384097 in the in *PTPN7* gene; a CpG that was previously reported to be highly discriminative between blood samples and buccal swab samples[1-4].

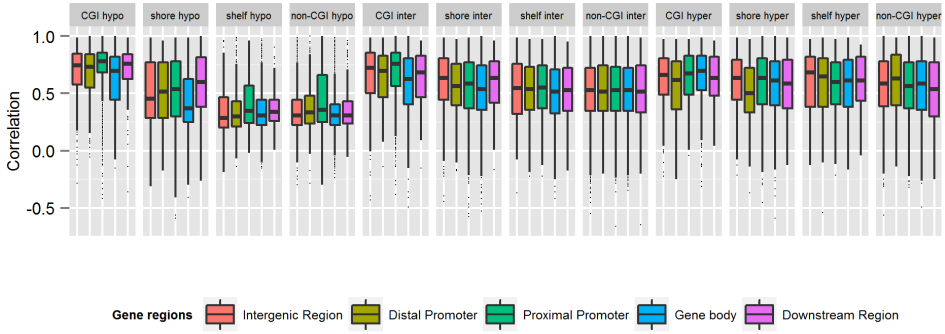
Sample	Beta-value cg18384097
MZpair3.2	0.7434
MZpair10.2	0.8360
MZpair6.2	0.8513
MZpair7.1	0.8527
MZpair10.1	0.8583
MZpair9.1	0.8723
MZpair7.2	0.8751
MZpair4.2	0.8929
MZpair5.1	0.8949
MZpair9.2	0.9002
MZpair8.2	0.9037
MZpair2.1	0.9055
MZpair1.2	0.9057
MZpair6.1	0.9076
MZpair5.2	0.9129
MZpair3.1	0.9130
MZpair4.1	0.9165
MZpair2.2	0.9187
MZpair8.1	0.9230
MZpair1.1	0.9312
Average	**0.8857**

Highlighted in blue are the twin samples that clustered separately from the other twin samples in Figure S2. The samples in the table are sorted by beta-value (lowest to highest).

Table S4. Spearman correlation between the methylation level of MZ twins at individual CpGs based on M-values.

Genome-wide	N CpGs	Mean r	Median r	Min r	Max r
All variable CpGs	59,041	0.54	0.54	−0.66	1

Figure S3. MZ twin correlations for individual CpGs grouped by genomic region and average methylation level, based on M-values. Hypo = hypomethylated. Inter = intermediate methylation. Hyper = hypermethylated.



Supplemental Reference List

1. Calvanese,V. *et al.* A promoter DNA demethylation landscape of human hematopoietic differentiation. *Nucleic Acids Res.* 40, 116-131 (2012).
2. Essex,M.J. *et al.* Epigenetic vestiges of early developmental adversity: childhood stress exposure and DNA methylation in adolescence. *Child Dev.* 84, 58-75 (2013).
3. Teschendorff,A.E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* 20, 440-446 (2010).
4. Souren,N.Y. *et al.* Adult monozygotic twins discordant for intra-uterine growth have indistinguishable genome-wide DNA methylation profiles. *Genome Biol.* 14, R44 (2013).
5. Slieker,R.C. *et al.* Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics. Chromatin.* 6, 26 (2013).
6. Chen,Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* 8, 203-209 (2013).

Appendix 7

Supplement to chapter 9

Supplemental Table 1 | Twin registries worldwide

Country	Twin Registry Name	Registry Characteristics	Age	Website	N twins/ subjects (approx.)[A]	N twins/subjectswith DNA available (approx.)[A]	Biospecimens (available for at least subset of the sample)
Africa							
Guinea-Bissau	Bandim Health Project Twin Registry	Population-based with ongoing longitudinal data collection	0-30	www.bandim.org	2,500 (twins and singleton controls)	200 twin pairs	Whole blood, plasma
Nigeria	Nigerian Twin and Sibling Registry (NTSR)	Cohort of school children + young adult volunteers	7-24	-	1,300 (twins and sibs)	1,200 (twins and sibs)	Buccal, saliva
Asia and Australia							
Australia	Australian Twin Registry	Population-based with ongoing longitudinal data collection	0-90	www.twins.org.au	66,000	12,000 (twins and other family members)	Serum, plasma, buccal
Australia	Western Australian Twin Register (WATR)	Population-based with ongoing longitudinal data collection	5-30	in transition	11,000 (plus other family members)	600 (twins and other family members)	Whole blood
Australia	Queensland Twin Registry	Systematic ascertainment through schools at age 12 and longitudinal follow-up	12-32	www.qtwin.org.au	2,000 twin pairs	2800	Blood (including Paxgene tube), saliva; processed to DNA, RNA, serum, plasma, red cells
China	Chinese National Twin Registry (CNTR)	Population-based with ongoing longitudinal data collection	All	cntr.bjmu.edu.cn	35,000 twin pairs	3,200	Serum, DNA
China	Qingdao Twin Register	Enrollment through medical registration systems, schools, and media publicity	All	-	8,000 twin pairs	1,500 twin pairs	Serum
Korea	The Healthy Twin Study, Korea	Ongoing in-depth survey based on NKTR	20-73	www.twinkorea.org	783 twin pairs	3,480 (twins and first degree relatives)	Serum, plasma, buffy coat, urine, stool, cell-line (lymphocytes)
Korea	The National Korean Twin Registry (NKTR)	No direct person contact. Zygosity information based on sex only. Basic data in adults on health and disease outcome.	8-80	-	187,130 twin pairs	-	None
Korea	South Korean Twin Registry (SKTR)	Volunteer preschoolers, cohort of school children, volunteer young adults	0-30	www.ktrc.org	10,000 twin pairs	800 twin pairs	Hair, saliva
Israel	The Longitudinal Israeli Study of Twins (LIST)	Population-based with ongoing longitudinal data collection	3-8	www.teomim.com	1,500 twin pairs	1,391 (twins and parents)	Buccal
Japan	Keio Twin Registry	Adult and adolescent twins from the general population in the Tokyo area	14-30	totcop.keio.ac.jp ; kts.keio.ac.jp ; kotrec.keio.ac.jp	4,000 twin pairs (plus other family members)	600 twin pairs	Buccal, blood
Japan	Twin Database of the Secondary School Attached to the Faculty of education of the	3 cohorts: applicants (11-12 years), students (12-18 years), and graduates (18-72 years). Follow-up of graduates 3 times.	0-80	-	2,200 (plus other family members)	-	None

	University of Tokyo						
Japan	Japanese Database of Multiples in Childhood	Main focus of the cohort is on child development	0-12	-	5,000 (plus other family members)	-	None
Japan	The Osaka University Aged Twin Registry	Population-based, cohort-study	50-105	http://www.twin.med.osaka-u.ac.jp/en/	12,000 twin pairs	280 twin pairs	Whole blood, serum, plasma
Sri Lanka	Sri Lankan Twin Registry	Voluntary twin registry component and a population-based database with ongoing data collection	6-94	www.ird.lk/Twin%20Registry.php	35,000	Plans to collect DNA from 4,000	Buccal
Europe							
Belgium	East Flanders Prospective Twin Survey (EFPTS)	Population-based at birth, phenotypic characterization in subsets	0-47	www.twinsvzw.com	8,700 twin pairs, 237 triplet sets	1,600 (twins and triplets)	Whole blood, saliva, buccal, placental tissue
Denmark	The Danish Twin Registry (DTR)	Population- based with ongoing longitudinal data collection	0-107	www.sdu.dk/dtr	170,000	20,000	Serum, plasma, buffy coat, saliva, buccal, urine
Finland	Finnish Twin Cohort study	Population-based with ongoing longitudinal data collection	11-100+	www.twinstudy.helsinki.fi	87,500 (plus family members)	14,600 (twins and family members)	Whole blood, serum, plasma, saliva, urine, fat and muscle by biopsy
Germany	HealthTwiSt: The Berlin Twin Registry	Registry operated as a CRO. Focus on health, functional genomics and nutraceuticals	16-80	www.zwillingsstudie.de	2,000 twin pairs	550 twin pairs	Whole blood
Germany	CoSMoS (Twin study on Cognitive Ability, Self-Reported Motivation, and School Achievement)	Longitudinal study of voluntarily enrolled twins with 3 measurement occasions.	7-11	www.uni-saarland.de/fak5/diff	808	-	None
Germany	TwinPaW (Twin study on Personality and Well-being)	Cross-sectional study of voluntarily enrolled twins	15-71	www.uni-saarland.de/fak5/diff	608	-	None
Germany	SOEP- twin project	Cross-sectional representative multi-group design, including twins, full sibs, mother-child and grandparent-child pairs. Subset with longitudinal data.	17-70	www.uni-saarland.de/fak5/diff	860 (plus other relatives)	-	None
Germany	Chronotyp-twin project (ChronoS)	Cross-sectional study of voluntarily enrolled twins.	17-70	www.uni-saarland.de/fak5/diff	755	-	None
Germany	Bielefeld Longitudinal Study of Adult Twins (BiLSAT)	Voluntary twin sample with longitudinal follow-up 5 times over 15 years	16-88	www.uni-bielefeld.de/psychologie/ae/AE04/Forschung/index.html	2,404	-	None
Germany	Jena Twin Study of Social Attitudes (JeTSSA)	Cross-sectional study of voluntarily registered twins and their parents and spouses	17-82	www.uni-bielefeld.de/psychologie/ae/AE04/Forschung/index.html	875	-	None
Italy	Italian Twin	Phenotypic	0-99	www.iss.it	25,000	2400	Whole blood,

	Register	characterization in subsets of the sample and longitudinal data collection in newborns		/gemelli			serum, plasma, buffy coat, saliva
Netherlands	Netherlands Twin Register (NTR)	Population-based with ongoing longitudinal data collection	0-100	http://www.tweelingenregister.org/en/	87,500 (plus family members)	18,000	DNA, RNA, cell lines, serum, plasma, buccal, urine
Norway	Norwegian Twin Registry (NTR)	Population-based with ongoing longitudinal data collection	18+	www.fhi.no/twins	40,000	4,800	Whole blood, buccal swabs, plasma
Spain	Murcia Twin Register (MTR)	Population-based with ongoing longitudinal data collection	45-71	http://www.um.es/registrogemelos/ing/index.html	2,281	743 (female twins)	Whole blood, saliva (gDNA, plasma & PBMC)
Sweden	Swedish Twin Register (STR)	Population-based with ongoing longitudinal data collection	5-100+	http://ki.se/ki/jsp/polopoly.jsp;jsessionid=acR0ziTHzWEclO_cNC?l=en&d=9610	194,000	44,600	Whole blood, serum, saliva
UK	TwinsUK registry	Population-based with ongoing longitudinal data collection	18-90	www.twinsuk.ac.uk	12,000	7,000	Whole blood, serum, plasma, buffy coat, saliva, buccal, urine, skin, fat, muscle
UK	The Cardiff Study of All Wales and North West of England Twins (CaStANET)	Population-based		http://medicine.cf.ac.uk/psychological-medicine-neuroscience/areas-research/child-and-adolescent-psychiatry/castanet-twin-register/	6000 twin families	300 twin pairs	None
UK	Twins Early Development Study (TEDS)	Population-based		http://www.teds.ac.uk/	10,000 twin pairs	10,000	None
UK	The Twins and Multiple Births Association Heritability Study (TAMBAHS)	Volunteer-based cohort of twins - closed to participation	0-5	www.birmingham.ac.uk/research/activity/mds/projects/HaPS/PHEB/UROGENEPID/projects/TAMBAhs.aspx	1,600 (plus parents)	-	None
UK	Birmingham Registry for Twin and Heritability Studies (BiRTHS)	Population-based cohort with ongoing follow-up	prenatal-4	www.birmingham.ac.uk/research/activity/mds/projects/HaPS/PHEB/UROGENEPID/pr	250 (including mothers)	50	Cord blood, saliva

				ojects/BiRTHS			
North America							
Canada	University of British Columbia Twin Project	National cross-section sample of general adult population	18-93	-	2,000 twin pairs	-	None
USA	California Twin Program	Population-based native residents	cohort born 1908-1982	twins.usc.edu	115,000 located, 37,000 twin pairs with complete phenotypic characterization	3,4000	Extracted DNA, frozen white cells, serum, plasma salivary DNA, DNA from buccal scrapings
USA	Colorado Twin Registry	Population-based with ongoing longitudinal data collection	1-41	ibg.colorado.edu/research/ctr.html	14,000 (plus family members)	5,500 (twins and family members)	Extracted saliva-derived DNA
USA	Florida State Twin Registry (FSTR)	Twins and other multiples enrolled in Florida schools (grades K-5)	5-13	www.psy.fsu.edu/~taylorlab/TwinRsch.htm	2,591 twin pairs (and other multiples)	-	None
USA	Georgia Carolina Cardiovascular Twin Study	Longitudinal study, mixed sample of American twins of various ancestries recruited at public middle and high schools.		-	1,664	1,607	
USA	International Twin Study	Continent-wide volunteers with chronic disease	All ages	twins.usc.edu	17,000 twin pairs	1,000	Extracted DNA, Frozen white cells, salivary DNA, serum, stool
USA	Michigan State University Twin Registry	Population-based and high-risk samples, some of which are longitudinal	3-30	www.msutwinstudies.com	20,000 (plus family)	4,100 (plus family)	Saliva, buccal, blood spots
USA	Mid-Atlantic Twin Registry (MATR)	Population-based, ascertained at birth	0-94	www.matr.vcu.edu	56,000	1,500	Whole blood, serum, plasma, buffy coat, saliva, buccal
USA	Minnesota Twin Family Study	Ongoing population - based longitudinal study	11-47	mctfr.psych.umn.edu	5,000 (plus family)	10,000 (twins and family members)	Blood-derived or saliva-derived DNA
USA	Minnesota Twin Registry	Cross-sectional sample of twins born between 1936 and 1964. Subsamples followed up through surveys	28-56	www.mntwin.umn.edu	15,000 (plus family)	-	None
USA	Missouri Family Registry	Vital record ascertained	1.5-44	-	25,979	2,575	Blood, saliva, fecal
USA	NAS-NRC (National Academy of Sciences-National Research Council) Twin Registry of World War II Male Veterans	Male twins born between 1917-27, both of whom served in the military, mostly during World War II	85-95	iom.edu/Activities/Veterans/TwinsStudy.aspx	31,848	700+	Blood and other materials collected for various investigations
USA	NAS-NRC (National Academy of Sciences-National Research Council) Twin	Longitudinal data collection from 1990 – 2006, primary aims to investigate cognitive decline and dementia in later life.	85-95	aging-memory.duhs.duke.edu/twins.html	13,000	770	Post-mortem brain tissue, DNA from buccal or blood

	Registry of World War II Male Veterans – Duke Twins Study Component						
USA	Pittsburgh Registry of Infant Multiplets (PRIM)	Population-based database of twins ascertained at birth – currently no follow-up data	-	www.pitt.edu/~mmv/prim/prim.htm	2,956	-	None
USA	University of Washington Twin Registry	Population-based with ongoing longitudinal data collection	18-97	depts.washington.edu/uwccer/	6,688 twin pairs	8,7000	Whole blood, saliva, serum, plasma, urine, buccal
USA	Vietnam Era Twin Registry (VET Registry)	Closed cohort of male twins whom served in the military during the Vietnam conflict (1964-1975)	54-72	www.seattle.eric.research.va.gov/VETR/Home.asp	14,760 (plus offspring and spouses of twins)	1,685	Whole blood, serum, plasma
USA	Wisconsin Twin Panel (WTP)	Population based, longitudinal data, extensive phenotypic characterization, follow up of selected samples	0-23	www.waisman.wisc.edu/twinresearch	19,638 twins (plus parents and sibs)	3,489 (twins, parents, sibs)	Saliva, buccal
South America							
Chile	The University of Chile School-Age Children Twin Registry (REMEUCHI)	Cohort of school children graduated from high school	18	-	327 twin pairs	-	None
Cuba	Cuban Twin Registry	Population-based with ongoing longitudinal data collection	All	-	55.400 twin pairs	250 twin pairs	Blood-derived DNA

List of Publications

- Polderman TJ, van **Dongen J**., Boomsma DI. The relation between ADHD symptoms and fine motor control: a genetic study. Child Neuropsychol 2011; 17/2: 138-150.
- **van Dongen J**., Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nat Rev Genet 2012; 13/9: 640-653.
- Tobi EW, Slagboom PE, **van Dongen J**., Kremer D, Stein AD, Putter H et al. Prenatal famine and genetic variation are independently and additively associated with DNA methylation at regulatory loci within IGF2/H19. PLoS One 2012; 7/5: e37933.
- Vink JM, Bartels M, van Beijsterveldt TC, **van Dongen J**., van Beek JH, Distel MA, de Moor MH, Smit DJ, Minica CC, Ligthart L, Geels LM, Abdellaoui A, Middeldorp CM, Hottenga JJ, Willemsen G, de Geus EJ, Boomsma DI. Sex differences in genetic architecture of complex phenotypes? PLoS One. 2012; 7:e47371
- **van Dongen J**., Willemsen G, Chen WM, de Geus EJ, Boomsma DI. Heritability of metabolic syndrome traits in a large population-based sample. J Lipid Res 2013; 54/10: 2914-2923.
- 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; 1-10.
- **van Dongen J**., Boomsma DI. The evolutionary paradox and the missing heritability of schizophrenia. Am J Med Genet B Neuropsychiatr Genet 2013; 162B/2: 122-136.
- Willemsen G, Vink JM, Abdellaoui A, den BA, van Beek JH, Draisma HH, **van Dongen J**, van 't Ent D, Geels LM, van LR, Ligthart L, Kattenberg M, Mbarek H, de Moor MH, Neijts M, Pool R, Stroo N, Kluft C, Suchiman HE, Slagboom PE, de Geus EJ, Boomsma DI. The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. Twin Res.Hum.Genet. 2013; 16:271-281
- Sabater-Lleal M, Huang J, Chasman D, Naitza S, Dehghan A, Johnson AD, Teumer A, Reiner AP, Folkersen L, Basu S, Rudnicka AR, Trompet S, Malarstig A, Baumert J, Bis JC, Guo X, Hottenga JJ, Shin SY, Lopez LM, Lahti J, Tanaka T, Yanek LR, Oudot-Mellakh T, Wilson JF, Navarro P, Huffman JE, Zemunik T, Redline S, Mehra R, Pulanic D, Rudan I, Wright AF, Kolcic I, Polasek O, Wild SH, Campbell H, Curb JD, Wallace R, Liu S, Eaton CB, Becker DM, Becker LC, Bandinelli S, Raikkonen K, Widen E, Palotie A, Fornage M, Green D, Gross M, Davies G, Harris SE, Liewald DC, Starr JM, Williams FM, Grant PJ, Spector TD, Strawbridge RJ, Silveira A, Sennblad B, Rivadeneira F, Uitterlinden AG, Franco OH, Hofman A, **van Dongen J**, Willemsen G, Boomsma DI, Yao J, Swords JN, Haritunians T, McKnight B, Lumley T, Taylor KD, Rotter JI, Psaty BM, Peters A, Gieger C, Illig T, Grotevendt A, Homuth G, Volzke H, Kocher T, Goel A, Franzosi MG, Seedorf U, Clarke R, Steri M, Tarasov KV, Sanna S, Schlessinger D, Stott DJ, Sattar N, Buckley BM, Rumley A, Lowe GD, McArdle WL, Chen MH, Tofler GH, Song J, Boerwinkle E, Folsom AR, Rose LM, Franco-Cereceda A, Teichert M, Ikram MA, Mosley TH, Bevan S, Dichgans M, Rothwell PM, Sudlow CL, Hopewell JC, Chambers JC, Saleheen D, Kooner JS, Danesh J, Nelson CP, Erdmann J, Reilly MP, Kathiresan S, Schunkert H, Morange PE, Ferrucci L, Eriksson JG, Jacobs D, Deary IJ, Soranzo N, Witteman JC, de Geus EJ, Tracy RP, Hayward C, Koenig W, Cucca F, Jukema JW, Eriksson P, Seshadri S, Markus HS, Watkins H, Samani NJ, Wallaschofski H, Smith NL, Tregouet D, Ridker PM, Tang W, Strachan DP, Hamsten A, O'Donnell CJ. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation 2013; 128:1310-1324
- **van Dongen J**, Ehli EA, Slieker RC, Bartels M, Weber ZM, Davies GE et al. Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. Genes (Basel) 2014; 5/2: 347-365.
- **van Dongen J,** Jansen R, Smit D, Hottenga JJ, Mbarek H, Willemsen G, Kluft C, AAGC collaborators, Penninx BW, Ferreira MA, Boomsma DI, de Geus EJ. The Contribution of the Functional IL6R Polymorphism rs2228145, eQTLs and Other Genome-Wide SNPs to the Heritability of Plasma sIL-6R Levels. Behav Genet 2014; 44/4: 368-382.

- Baumert J, Huang J, McKnight B, Sabater-Lleal M, Steri M, Chu AY, Trompet S, Lopez LM, Fornage M, Teumer A, Tang W, Rudnicka AR, Mälarstig A, Hottenga JJ, Kavousi M, Lahti J, Tanaka T, Hayward C, Huffman JE, Morange PE, Rose LM, Basu S, Rumley A, Stott DJ, Buckley BM, de Craen AJ, Sanna S, Masala M, Biffar R, Homuth G, Silveira A, Sennblad B, Goel A, Watkins H, Müller-Nurasyid M, Rückerl R, Taylor K, Chen MH, de Geus EJ, Hofman A, Witteman JC, de Maat MP, Palotie A, Davies G, Siscovick DS, Kolcic I, Wild SH, Song J, McArdle WL, Ford I, Sattar N, Schlessinger D, Grotevendt A, Franzosi MG, Illig T, Waldenberger M, Lumley T, Tofler GH, Willemsen G, Uitterlinden AG, Rivadeneira F, Räikkönen K, Chasman DI, Folsom AR, Lowe GD, Westendorp RG, Slagboom PE, Cucca F, Wallaschofski H, Strawbridge RJ, Seedorf U, Koenig W, Bis JC, Mukamal KJ, **van Dongen J**, Widen E, Franco OH, Starr JM, Liu K, Ferrucci L, Polasek O, Wilson JF, Oudot-Mellakh T, Campbell H, Navarro P, Bandinelli S, Eriksson J, Boomsma DI, Dehghan A, Clarke R, Hamsten A, Boerwinkle E, Jukema JW, Naitza S, Ridker PM, Völzke H, Deary IJ, Reiner AP, Trégouët DA, O'Donnell CJ, Strachan DP, Peters A, Smith NL. (2014) No Evidence for Genome-Wide Interactions on Plasma Fibrinogen by Smoking, Alcohol Consumption and Body Mass Index: Results from Meta-Analyses of 80,607 Subjects. PLoS ONE 9(12): e111156. doi:10.1371/journal.pone.0111156
- **Jenny van Dongen**, Gonneke Willemsen, Bastiaan T. Heijmans, Jacoline Neuteboom, Cornelis Kluft, Rick Jansen, Brenda Penninx, P Slagboom, Eco de Geus, and Dorret Boomsma (2015). Longitudinal weight differences, gene expression, and blood biomarkers in BMI discordant identical twins. International Journal of Obesity (accepted for publication).
- Jana Strohmaier, **Jenny van Dongen**, Gonneke Willemsen, Dale R. Nyholt, Gu Zhu,Veryan Codd, Boris Novakovic, Narelle Hansell, Margaret J Wright, Liz Rietschel, Fabian Streit, Anjali K. Henders, Grant W. Montgomery, Nilesh J. Samani , Nathan A. Gillespie, Ian B. Hickie, Jeffrey M. Craig, Richard Saffery, Dorret I. Boomsma, Marcella Rietschel, Nicholas G. Martin (2015). Low birth weight in MZ twins discordant for birth weight is associated with shorter telomere length and lower IQ, but not anxiety/depression in later life. Twin Research (accepted for publication).
- **Jenny van Dongen** , Bastiaan T. Heijmans, Michel G. Nivard, Gonneke Willemsen, Jouke-Jan Hottenga, Quinta Helmer, Conor V. Dolan, Erik A. Ehli, Gareth Davies, BIOS Consortium, H. Eka Suchiman, Rick Jansen, Joyce B. van Meurs, P. Eline Slagboom, Dorret I. Boomsma. The heritability of DNA methylation in peripheral blood: influences of common SNPs and variability of genetic and environmental variance with age and sex. (manuscript in preparation)

Dankwoord

Op deze pagina wil ik graag alle personen bedanken die dit proefschrift mogelijk hebben gemaakt.

Allereerst mijn dank aan de tweelingen en familieleden voor hun deelname aan het onderzoek van het Nederlands Tweelingen Register.

Ik wil mijn promotoren Dorret en Eline en co-promotoren Gonneke en Bas bedanken voor de prettige samenwerking en begeleiding. Ik had het geluk dat ik op twee afdelingen kennis en ervaring kon opdoen. In Amsterdam op het gebied van de genetica en het tweelingonderzoek en in Leiden op het gebied van de epigenetica. Dorret en Eline, van jullie heb ik veel geleerd en het was fijn om twee begeleiders te hebben die met een verschillende blik mijn papers konden lezen. Gonneke, bedankt voor je hulp en in het bijzonder voor de ondersteuning bij alle zaken omtrent samples, data en alles wat daarbij komt kijken, waarvoor je altijd voor de hele afdeling klaarstaat. Ook op het laatste moment voordat ik mijn proefschrift ging opsturen naar de leescommissie, op de middag voor kerst, kon je het laatste hoofdstuk van mijn proefschrift nog even doorlezen. Bedankt daarvoor! Bas, bedankt voor de prettige begeleiding, adviezen en de motiverende feedback (“Goed werk!”). Ik ben blij dat ik mijn werk op het gebied van epigenetica en tweelingen bij Biologische Psychologie kan voortzetten en dat wij ook in de toekomst blijven samenwerken.

I would like to thank all members of the reading committee; Dr. Marian. Beekman, Dr. Erik Ehli, Dr. Lude Franke, Prof.dr. Eco de Geus, Prof.dr. Jaakko Kaprio, Dr. Joyce van Meurs, and Prof.dr Carlo Schuengel for reading my thesis.

Ik wil graag mijn collega’s van de afdeling BioPsy en mijn collega’s van MolEpi in Leiden bedanken, mijn mede-AIO’s, en mijn huidige en oude kamergenootjes: Lot, Jenny, Fiona, Jorien en alle bewoners van de Apenrots.

Een aantal collega’s wil ik in het bijzonder bedanken voor hun hulp bij het verzamelen van de data die ik in mijn proefschrift heb gebruikt. Onze samenwerking met Leiden, Rotterdam en het BIOS consortium maakte het mogelijk om DNA methylatie te meten in een groot aantal samples van het NTR. Ik wil graag Eka Suchiman en Joyce van Meurs bedanken voor de coördinatie en goede zorg voor de samples. I would like to thank our colleagues from the Avera Institute in South Dakota for their assistance with measuring DNA methylation of the NTR samples. Erik, it was nice working together on the first pilot project and I look forward to our collaboration on future projects. Graag bedank ik het secretariaat van de afdeling Biologische Psychologie en het Nederlands Tweelingen Register voor hun hulp bij alle

praktische zaken en voor hun bijdrage aan de verzameling van vragenlijstdata die ik in mijn proefschrift heb gebruikt.

Ik wil de volgende personen bedanken voor de onmisbare hulp, inzichten en tips die ik van hen gekregen heb bij de analyses van genotype-, genexpressie- en DNA methylatie data. Jouke-Jan, bedankt voor alle hulp in de afgelopen jaren met de genotype data. Rick, bedankt voor je hulp met de genexpressie data. In Leiden ging ik voor het eerst aan de slag met DNA methylatie. Gelukkig kreeg ik daarbij hulp van de “450k taskforce”: Elmar, Roderick, René, Maarten, Wouter, bedankt! Michel, veel dank voor alle hulp bij de analyses voor mijn laatste hoofdstuk.

Jorien en Sue, ontzettend fijn dat jullie mijn paranimfen willen zijn! Rosie en Ratna, bedankt voor jullie hulp met de omslag van mijn proefschrift, en voor de gezelligheid, waarbij ik natuurlijk ook Sophie wil betrekken. Tot slot, mijn familie en Philip, bedankt voor jullie steun en interesse.