Using genetic methods to get insight into human complex traits

Bochao Danae Lin

READING COMMITTEE:

Prof. CV Dolan, Vrije Universiteit Amsterdam
Prof. NG Martin, QIMR Berghofer Medical Research Institute, Brisbane, Australia
Prof. H Snieder, Universitair Medisch Centrum Groningen
Prof. M Kayser, Erasmus Universiteit Medisch Centrum Rotterdam
Dr. DJA Smit, Academisch Medisch Centrum Amsterdam
Dr. R Menezes, Vrije Universiteit Medisch Centrum Amsterdam

PARANYMPHS:

Jenny van Dongen
Junfeng Wang

ACKNOWLEDGEMENTS:

I would like to acknowledge the financial support I received from the China Scholarship Council (201206180099) for my study. I also acknowledge the financial support from the Faculty of Behavioural and Movement Sciences to visit Notre Dame University, United States. The data collection in the Netherlands Twin Register was supported by multiple grants from the Netherlands Organization for Scientific Research and Development Grant (ZonMW); the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL) [184.021.007]; the European Research Council, Genetics of Mental Illness [ERC-230374]; the National Institute for Mental Health (NIMH) [1RC2 MH089951-01: Integration of Genomics & Transcriptomics in Normal Twins & Major Depression, and 1RC2MH089995-01: Genomics of Developmental Trajectories in Twins]; and the Avera Institute for Human Genetics , Sioux Falls (USA).

I very warmly thank all twins and their family members for taking part in the research projects of the Netherlands Twin Register.

ISBN: 978-94-6295-615-5
Cover design and Layout: Saymand Alerachi
Printed by: Proefschriftmaken

VRIJE UNIVERSITEIT

Using genetic methods to get insight into human complex traits

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Gedrags- en Bewegingswetenschappen
op donderdag 20 april 2017 om 11.45 uur
in het auditorium van de universiteit,
De Boelelaan 1105

door

Bochao Lin

geboren te Jilin, China

promotor: prof.dr. D.I. Boomsma
copromotoren: dr. J.J. Hottenga
dr. A.H.M. Willemsen

Table of Contents

Chapter 1

General Introduction

We are all unique, even if we share certain characteristics with our family members and those around us. The individual differences that can be observed in the population are caused by a combination of genetic and environmental factors. In this PhD thesis I endeavor to add to our understanding of the way genetic factors explain individual differences in human complex traits by applying a variety of methodological tools to different sets of personal characteristics, which in genetics are commonly referred to as 'phenotypes'. In this first chapter I provide a general background, describe the methods I applied to answer the questions about the etiology of individual differences, and I introduce the traits of interest, in which variation is analyzed in this thesis.

1.1 General background.

Genetic studies of human complex traits aim to clarify the contribution of genetic factors to variation in the trait. The phrase 'complex trait' or 'complex phenotype' refers to traits, which result from variation at multiple genomic sites, and across multiple environmental factors. Complex traits do not follow a Mendelian pattern of inheritance and often show a continuous distribution in the population, either on the scale of measurement or on the underlying liability scale. Family studies, in particular twin studies, have proven to be useful tools to determine to which extent genetic factors influence individual variation in a traits, or stated differently, to provide us with estimates for the heritability of a trait [1]. Such studies, however, do not identify the source of the genetic contribution to individual differences, namely the variation in DNA sequence in human genomes. Until recently, there were two main approaches to gain information on the genetic variants influencing phenotypes of interest: candidate gene association studies and linkage studies. In a candidate gene association study the focus is on associations between the variation in the phenotype and variation in one or a few preselected genes. For example, my first genetic study conducted as part of my Master program investigated the association between rs16970495 (an intro variant on RASGRF1) and myopia in a sample of 557 Chinese adults. In this study I found that the A allele of rs16970495 was associated with an increased risk of myopia (OR=1.21, P=.003). However, such candidate gene studies are generally based on limited prior knowledge, especially in psychology or psychiatry, and often prove difficult to replicate in follow up studies [2]. Linkage studies do not focus on a specific gene but use data from family members to map individual differences in a trait to variation in genetic markers of a known chromosomal location. A location that correlates with the quantitative trait of interest is referred to as a quantitative trait locus (QTL), and is more likely to contain a causal genetic variant. However, linkage studies require family data, such as big pedigrees, or large samples of sibling pairs and have a low resolution mapping, with resulting QTLs thus referring to broad chromosomal regions, rather than to a specific base pair position. In addition, linkage studies are well suited for Mendelian traits with high penetrance but less so for complex traits [3].

Fortunately, advances in genotyping technology have now made it both time- and cost-wise possible to explore a large part of the variation in the human genome. Assessment of genomic variation in DNA can be done by typing samples on SNP (single nucleotide polymorphisms) arrays, or by sequencing the complete genome, which provides additional information of genetic variants such as insertions and deletions (indels) and copy number variants (CNV). These developments have further been supported by advances in computational support and methodologies for high-throughput data analysis. Imputation of missing genotypes in subjects with SNP array data, using a large reference panel of sequenced individuals such as the 1000 Genomes project [4], provides the possibilities for GWA studies to identify complex trait loci. The combination of arrays and sequence information results in a nearly complete number of studied genetic markers across the genome to map associations. Furthermore it allows to harmonize and combine datasets or results across research groups for meta-analysis [5]. As a result, the last decade has seen a large number of identified genes, and more scientific achievements obtained by genome-wide association studies (GWAS) for human complex traits. The results obtained by GWAS studies have provided the input for new analyses techniques, furthering our understanding of the way genetic influences are involved in individual variation, done by exploring the degree to which genes cluster in pathways, influence multiple traits or are differently expressed during the lifespan.

SNP data have also been analyzed to establish the extent to which all measured genetic variants are involved in the heritability for a trait. This form of analysis provides heritability estimates that are based on a genetic relatedness matrix (GRM), rather than employing the known genetic relatedness among relatives such as parents and offspring or between twins. A widely-used implementation of this method is in the software package “Genome-wide Complex Trait Analysis” (GCTA) [6]. Approaches such as implemented in GCTA can link the outcomes from twin-family studies to those from genotypic studies.

At other omics levels, it has become possible to study at a large scale the expression of genes, rather than the DNA variants itself, either through studies of RNA expression or through epigenetics approaches. As with genetic studies, here too the field has moved from small-scale studies to the more encompassing studies in the form of transcriptome-wide and epigenome-wide association studies (EWAS). Together, these various tools can provide us with a window on the way genetic and environmental factors interact and create individual variation in complex traits [7-8].

1.2 Genetic methodology used in this PhD thesis

1.2.1 Using structure equation models to estimate heritability with extended twin family design

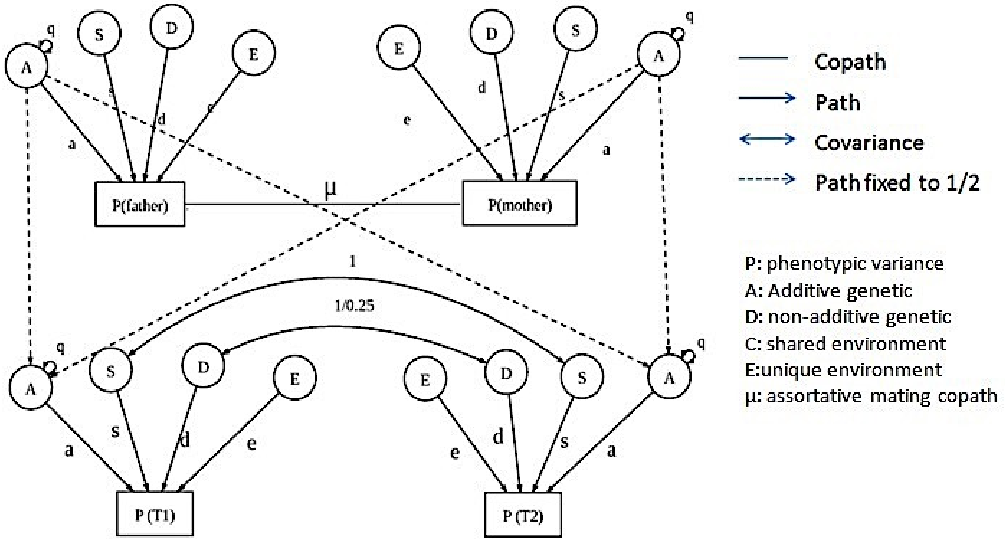
Heritability studies make use of genetically informative family data to quantify the magnitude of genetic and environmental influences on the variation in phenotypes of interest. The classical twin study has a long history in behavioral and social sciences and has been used to assess heritability for almost every conceivable trait [9-10]. Twin studies take advantage of the genetic differences between monozygotic and dizygotic twins to partition the variance of a trait into components attributable to genetic and environmental causes [11]. Monozygotic (MZ) twins, also known as identical twins, arise from one fertilized egg which, within a few days after fertilization, has split in two separately developing embryo's, who share ~100% of their segregating genes. Dizygotic (DZ) twins, also known as fraternal twins, result from two fertilized eggs within one pregnancy and share, like ordinary siblings, on average 50% of their segregating genes. Assuming there are no special twin environment effects, which indeed is corroborated by studies [11-12], differences between the resemblance of MZ twin pairs and the resemblance of DZ twin pairs (rMZ > rDZ, where r stands for correlation) must be due to genetic influences. Similarities in the resemblance for MZ and DZ twins that cannot be attributed to genetic resemblance (e.g. the situation where the correlation in MZ twins equals the correlation in DZ twins and both are unlike zero) reflect the influence of common environmental effects (C). All differences within MZ twin pairs can be contributed to unique environmental effects, e.g. life events unique for one twin but also measurement error (E), whereas differences within DZ pairs are due to both unshared genetic and unshared environmental influences.

Genetic effects can further be divided into additive genetic effects (A) and non-additive genetic effects (D). The additive genetic effects represent a linear effect of the genotypic contribution to a quantitative trait: any genotypic change in an additive variant has a similar effect. The non-additive effects include all nonlinear effects of genetic variation, including dominance (the effect is based on the specific allelic combinations at one locus) and in human data also epistasis (the effect of specific alleles is dependent on other genotypic variation). The variance in a phenotypic trait V_P can thus be decomposed into different components: A, D, C and E: $V_P = V_A + V_D + V_C + V_E$ [13].

The covariance for MZ twin pairs and DZ twin pairs equals, respectively, $V_A + V_D$ and $1/2\ V_A + 1/4\ V_D$. The proportion of the total phenotypic variance (V_P), which is explained by the additive genetic effects (V_A), is generally called the narrow-sense heritability as opposed to the broad-sense heritability (V_A+V_D/V_P), which includes all the genetic effects. In a classical twin design, however, the simultaneous estimation of D and C is not possible, and a choice is made based on the twin correlations to test either an ACE model, or an ADE model, which does not allow for any influence of the common environment [14].

Extending the twin design to include additional family members addresses some limitations in classical twin studies. When siblings of twins are added to the design, the power of the study is increased and a formal twin-singleton sibling comparison is possible. While there is very little evidence for twin-singleton sibling differences [15-17], it still is a question that is put forward by many non-twin researchers. The inclusion of parents and/or spouses in the study has multiple advantages [18]. Such an extended twin family design (see **Figure 1**) not only allows for the modelling of a full ACDE model, but it can also account for assortative mating. Assortative mating refers to the phenomenon that partner choice is not random but that individuals tend to choose partners who share certain similarities [13]. Evidence for a 'like with like' mating pattern comes from a positive correlation between the phenotypic values of spouses. In the case of a heritable trait, this phenomenon makes siblings and DZ twins genetically more similar than the expected average of 50%, leading to biased estimates of the genetic and environmental influences [1]. Despite the advantages, not all twin registers collect information on additional family members. In this thesis I made use of data from the Netherlands Twin Register, described in detail later in this chapter, which is one of the twin registers worldwide that collects data on twins as well as their family members.

Figure 1. ACDE path diagram for an extended-family twin design of twins and their biological parents.



In a classical twin study design the observed correlations for MZ and DZ twin pairs form the basis to estimate the percentage of the total variance (V_P) due to additive genetic (V_A), non-additive genetic (V_D), shared environmental (V_C), and unique environmental (V_E) effects. However, V_D and V_C cannot be estimated simultaneously as such a model is unidentified. To circumvent this problem, in classical twin either V_D (when rMZ > 2rDZ) or V_C (when rMZ< 2rDZ) is constrained at zero.

In the extended-family design, the observed correlations for parents-offspring, father-mother, and additional singleton sibling pairs are also included in the model, which allows a simultaneous estimation of V_C and V_D and the inclusion of assortative mating (μ) effects. In the path diagram above this is operationalized as: $V_P=V_A + V_D +V_C + V_E = a^2+d^2+c^2+e^2$, where

Covariance (spouse) = μ

Covariance (parents-offspring) = $0.5\ V_A = 0.5a^2(1+\mu)$

Covariance (DZ=siblings) $=0.5V_A + 0.25V_D +V_C = 0.5a^2(1+ \mu)+0.25d^2+c^2$

Covariance (MZ) $=V_A + V_D + V_C = a^2+d^2+c^2$

1.2.2 Genome-wide association study (GWAS)

Genome-wide association studies (GWAS) present an unbiased genomic screening of a population to establish whether any of the millions of measured genetic variants is associated with a trait. As indicated, compared with linkage and candidate gene association studies, GWA studies have larger statistical power to detect common gene variants and a higher resolution, enabling a detailed dissection of the genetic architecture of complex traits, although it should be recognized that hardly any of the linkage studies in human genetics ever achieved the sample sizes that are now seen in GWA studies.

GWAS is a hypothesis-free exploratory approach to detect associations between genetic variants and the individual differences in a trait. It makes use of genotypic information for millions of SNPs across the genome. Generally, the genotypic information refers to common SNPs which represent to the most common form of genetic variants, accounting for 90% of total genetic variation and representing base pair changes that occur in at least 1% of the population [19]. SNPs are variants, which may have a direct effect on the gene function, or they act as markers in linkage disequilibrium with the DNA variants that result in a studied trait, or disease. For example, most European individuals have two C alleles (CC genotype) at SNP rs1667394 while a minority of the European population has the T allele (CT or TT genotype). In a blue eye color GWAS, the frequency of all SNPs in the blue eye color population (the case group) is compared with the frequency of all SNPs in the non-blue eye color population (control group). The result of this comparison, which shows a significant higher T allele frequency in the blue eye color group, indicates the importance of SNP rs1667394 for eye color in these populations [20].

In this straightforward interrogation of the association between phenotype and genotype a very large number of SNPs is tested, so the alpha level is necessarily very small (generally significance is set at $p < 5\times10^{-8}$) [21] to correct for multiple testing. Also, an independent replication of the finding is usually needed, so multiple large sample studies are required. This has led to the formation of various consortia, in which research groups worldwide collaborate to find significant SNPs, or gene loci, with much success. For instance, we now have a number of confirmed loci for complex traits such as height [22], BMI [23] and schizophrenia[24]. These days GWAS is the standard approach to identify loci influencing complex traits. According to the GWAS catalogue website, from the first GWAS study till 8th May 2016, 21,750 unique SNP-trait associations were identified by 2,437 GWAS studies [25].

As increased numbers of GWAS are conducted, increased numbers of genetic variants are uncovered to be associated with complex traits of interest. Genetists are taking effort to meta-analyse and cluster the GWAS results to identify pleiotropy: the same genetic variants can be associated with multiple diseases and other complex traits. Pleiotropy studies unravel the genetic etiology links between complex diseases and traits, which leads to a deeper understanding of complex traits.

1.2.3 Genome-wide Complex Trait Analysis (GCTA)

When GWAS demonstrates significant SNPs for a trait, one question that comes up is to what extent these SNPs represent the heritability of the trait. The software program GCTA was developed to estimate SNP heritability (h_g^2), that is the percentage of genetic variance that can be explained by the pooled candidate SNPs or by all SNPs genome-wide [6]. Using data from unrelated individuals, the algorithms implemented in GCTA first estimate the genetic similarity among participants based on genotyped or imputed SNPs that have been appropriately clumped, and this can be genome-wide, represent a single chromosome or just a region of interest. The covariance matrix, referred to as the genetic relationship matrix (GRM) summarizes the genetic variance among persons. In the next step this GRM is used to predict the phenotypic similarity within the population. When including related individuals in the classical GCTA method would result in inflated SNP heritability estimates, or actually estimates of the narrow sense heritability when sib-pairs, parents and monozygotic twins are included [6], but when excluding related individuals when present has the disadvantage of decreasing the sample size and increasing the standard error of the estimated SNP heritability. To account for this limitation of the classical GCTA model, Zaitlen et al [26] introduced the two-component-covariance matrix model. Instead of one GRM matrix which only includes unrelated individual pairs (marker allele sharing Identity By State threshold < generally set at 0.025), now two covariate matrixes are used: One matrix focusing on closely related individuals (IBS > 0.025) to estimate h^2 and one matrix including all individuals to estimate the SNP h_g^2. The difference between h^2 and h_g^2 is called the missing heritability, that is the heritability that is currently not explained by the measured SNPs. In this way the two-component-covariance matrix option of GCTA results in a narrow sense heritability estimate under an AE model as well as a SNP heritability estimate.

1.2.4 Epigenome-wide association study (EWAS)

Although MZ twins have ~100% identical DNA sequence (pending relatively rare somatic mutations), discordance within MZ twin pairs is observed for a large range of complex traits, even when the trait is heritable. This has led to an interest in exploring non-genetic factors that could impact on complex trait variance. Epigenetic studies are defined as "the study of changes in gene function that are mitotically heritable and that do not entail a change in DNA sequence" [27]. Without altering the DNA sequence, epigenetic changes may be produced by DNA methylation or histone modification, under the influence of factors such as behavior, diet or smoking, or as a result of random events [28]. DNA methylation is the attachment of a methyl-group (CH_3) to a DNA molecule, which mostly occurs at specific sequences of DNA: a cytosine located next to guanine called CpG sites. Histone modification is a post-translational modification to the histone, the main component of chromatin proteins. These epigenetic variants regulate gene expression by influencing the chromatin structure and the binding of regulator proteins to the DNA [29]. Because DNA methylation is more stable and amenable to high-throughput analysis than histone modification, it is at the moment the most used epigenetic marker in large-scale human epidemiological studies. Epigenome-wide association studies (EWAS), uses a similar procedure as GWAS, relating genome-wide epigenetic differences among people (generally in the form of CpG sites) to differences in phenotype to understand the role of epigenetics in complex traits [30].

1.2.5 Expression quantitative trait loci studies (eQTLs)

eQTL analyses aim to detect DNA variants that affect gene expression levels in an aim to understand the basis of gene regulation and have a better interpretation for GWAS and EWAS results. Standard eQTL analysis detects the association between genetic variants with gene expression levels (transcriptome profile: amount of mRNA) in hundreds of samples. The principle is same as genome-wide association study, but gene expression levels are analyzed as the phenotype of interest [31]. The genetic variants may be associated with physical proximal gene expression with cis-effects, or distal gene expression with trans-effects. Incorporating eQTL analyses with GWAS or EWAS will provide understanding of complex trait genetic etiology, which may unravel causality: if one (epi) genetic variant is significantly associated with the phenotype of interest in genome-wide association study, and it also causes the alteration of proximal gene expression, it then implies the causal effect of this variant for the trait [32].

1.3 Phenotypes of interest

In my thesis, I applied the techniques briefly outlined in the previous sections to two different kinds of complex human traits: pigmentation traits and hematological parameters. These traits are of interest in their own right, but can also be thought of as serving as 'model phenotypes' for complex behavioral and complex traits.

1.3.1 Pigmentation traits

Pigmentation traits are among the most visible traits, drawing scientific attention from a wide range of disciplines, from biology and anthropology to psychology and dermatology and cosmetics. Pigmentation traits are highly heritable, with heritability estimates ranging from 61% to 98% for hair color [33-35], 80%-100% for eye color [34, 36-37] and 60%–90% for skin color [38-39]. The degree of pigmentation is primarily determined by the amount, the type, and the distribution of melanin, which protects the body against ultraviolet radiation damage. The variation in pigmentation traits results from a combination of gene mutations and natural selection.

About 1.2 million years ago our human ancestors (the early Hominids) did not have much pigmentation and had a lot of body hair. Over the time span of 1 million years hominids gradually lost their body hair and simultaneously pigmentation increased to ensure continued protection from sun exposure. When modern humans left Africa (about 11,000 to 19,000 years ago), our ancestors were likely dark skinned, with black hair and dark brown eyes. Up to this day, individuals from equatorial and tropical regions combine dark pigmentation traits with little body hair. In moving away from the equator, depigmentation may have occurred to account for the lower levels of ultraviolet (UV) radiation to ensure that the levels of folic acid and vitamin D levels, that are dependent on exposure to UV radiation, remained sufficient [40]. The reduction in the diversity of the daily diet , due to the emergence of agriculture, which also led to reduced vitamin D levels, may have further promoted depigmentation [41]. As a result diversity in pigmentation within the population increased [42] and nowadays, the diversity in hair, skin, and eye color reaches its maximum in the regions of the East Baltic regions, Northern Europe and Eastern Europe. The range, prevalence and distribution of human pigmentation traits continue to change slowly across the world. Assortative mating for pigmentation traits plays a role [34, 43] and sexual dimorphism also occurs. Women generally have relatively lighter pigmentation than men. A lighter skin facilitates vitamin

D synthesis needed to meet the higher calcium level requirements for females especially during pregnancy and lactation [44]. Cultural factors may also influence the preference for a specific skin color in a partner.

During the evolutionary process, positive selection occurred for several gene mutations relevant to pigmentation. More than 30 pigmentation genes have been identified, especially in the melanosome biogenesis or the melanin biosynthetic pathways [45], but the most studied gene is melanocortin 1 receptor (MC1R). The MC1R gene plays an important role in determining the type of melanin produced: pheomelanin (yellow-red color) or eumelanin (brown-black color) [46]. A comparison of the African human genome with the chimpanzee genome, revealed 17 genetic variants in the MC1R gene in the human population [47]. The authors of the article suggested that these genetic variants are associated with high levels of eumelanin, one of the two forms of melatonin, and were positive selected for along with loss of body hair in African population in the environments with high UV radiation exposure and high temperature.

Pigmentation traits are ideal candidates for studies on the genetic architecture of human complex traits. Hair and eye color in particular are very visible traits, easy to obtain with small measurement error and stable during the lifespan (in the case of hair color before graying or balding occurs). Compared with skin color, these traits are also influenced less by environmental variation.

In this thesis, I present outcomes of a GWAS for hair color and eye color in the Dutch population. I estimate the total heritability for these traits from twin-family data and estimate the SNP heritability and genetic correlation between hair color and eye color from SNP data.

1.3.2 The hematological profile

The hematological profile consists of high-dimensional information on the distribution and type of three different kinds of blood parameters; white blood cells, red blood cells and platelets. Red blood cells (also called erythrocytes) are responsible for the transport of oxygen to tissues and the removal of carbon dioxide from the body, white blood cells (also called leukocytes) coordinate immune responses to e.g. bacteria and virus infections, and platelets interact with clotting factors in plasma to prevent excessive bleeding and promote tissue repair. Abnormal values, that is values outside the normal reference range, in one or more components of the hematological profile may be indicative of disease or predict future disease development. For instance, a low blood cell count points to anaemia, high red blood cell count and high platelet count are related to an increased risk of cardiovascular disease [48], high or low white blood cell count may indicate the presence of an immune disorder or cancer [49], while high or low platelet counts may point to coagulation disorders [50]. Since the standard hematological profile is relatively easy to obtain and inexpensive, it is one of the most commonly diagnostic tests conducted in medical practice.

Though these hematological indices are tightly regulated and heritable, with heritability estimates ranging from 32% to 87% [51-53], they are also influenced by environmental factors. For example, the number of neutrophils, a specific type of white blood cell, fluctuates strongly in response to contact with bacteria or air pollutants and when the outside temperature changes [54], while dietary intake may be responsible for changes in the number of red blood cells [55-56]. People in high-altitude regions tend to have higher hemoglobin levels to maintain oxygen saturation levels, which may reflect a combination of genetic and environmental conditions [57-59].

Individual hematological indices have received a lot of attention within genetics and have been shown to be heritable. Linkage and candidate gene association studies have identified a number of QTLs for these individual parameters, and recently GWA and EWA studies showed evidence for the involvement of several (epi)genetic variants for these parameters [60-62].

However, hematological indices are closely interlinked and a multivariate approach may provide additional information. At present, the number of studies which use a multivariate approach is limited. One common strategy for high dimensional data is adding together items to sum scores and use these sum scores for subsequent statistics. One could argue that within hematology sum scores are used when total white or red blood cell count is analyzed and indeed, as indicated earlier, these scores have predictive values of their own for disease status. Recent studies however have used a different approach, examining the balance between two hematological indices rather than the individual blood counts. The myeloid-lymphoid ratios such as neutrophils to lymphocyte ratio (NLR), monocytes to lymphocyte ratio (MLR) and platelet to lymphocyte ratio (PLR) have been put forward as novel and useful predictive or prognostic biomarkers for cancers [63-65] and/or complex diseases such as immune disease [66-69]. So far these ratios received most attention within clinical studies, where cut-off values for each of these biomarkers were evaluated for their usefulness regarding early admission, stratification, classification, diagnosis, prognosis and response to therapy [70-73]. However, the genetic study of these three biomarkers in a healthy population has yet to be explored.

In this thesis, I present the results of a comprehensive series of genetic analyses for these biomarkers, including twin-family, GWA, GCTA and EWA studies. In addition, I examined the importance of demographic, environmental and lifestyle factors for individual variation in these hematological ratios.

Using the information from all hematological variables simultaneously may prove even more informative. To this aim, the possibility was explored of using the full hematological profile using multivariate distance matrix regression (MDMR). This analysis technique allows more than two variables to be analyzed at once, taking into account the relationship of dependent variables by using a so-called distance matrix. This matrix is constructed as an (n x n) matrix in which n is the number of participants and pair-wise distances represent the (dis)similarity of participants based on multiple variables. Using MDMR one can test the association between the full hematological profile and other variables of interest.

MDMR was used to investigate the effects of several demographic and lifestyle factors in relation to the full hematological profile consisting of 10 variables and to determine the potential of the profile thus obtained for future studies, including genetic studies.

1.4 My data source: the Netherlands Twin Register

The majority of the data reported on in this thesis are derived from the Netherlands Twin Register (NTR). The NTR was established in 1987 to study the way genetic and environmental factors influence individual variation in development, lifestyle, personality and health. The NTR consists of a young cohort (YNTR) and an adult cohort (ANTR), with all participants followed longitudinally. In the YNTR twins are followed from birth onwards at specific ages, first by parent report, later also by teacher report and from 14 years onwards by self-report. The main focus of the YNTR is the motoric and cognitive development as well as the development of emotional and behavioral problems (a detailed description of the YNTR is provided in [74]). The ANTR started as a longitudinal study in a cohort of adolescents and young adults in 1991 and since then has sent out surveys on lifestyle, personality and mental and physical health to the twins and their family members at two to four yearly intervals. At present, 12 surveys have been sent (for a detailed description of the ANTR survey study see [75]). The NTR contains around 86,000 young twins/multiples, 11,000 adult twins/multiples and 101,000 family members of twins, with a total N of more than 199,000 participants from ~ 50,000 families. The survey data provided the information used for the hair and eye color analyses. The seventh survey of the ANTR sent out in 2004 contained the question "What is your natural hair color?" with five answer categories (blond, red, light brown, dark brown and black) and the question "What color are your eyes?" with three answer categories (blue/gray, green/light and brown). The same questions on eye color and hair color were answered by adolescent (14-18 year old) YNTR twins when they completed the Dutch Health and Behavior Questionnaire in 2005 or 2006.

In 2004 the NTR started a large biobank project, asking adult participants whether they were willing to provide a blood sample. Between 2004 and 2008 ~9500 participants were visited at home for blood sampling and a short interview on lifestyle and health [76]. In

2011 a second, similar but smaller-scale project [77] was carried out bringing the total number of participants to ~10,000 individuals. The biobank project has been described in detail elsewhere [76]. The hematological data analyzed in this thesis were determined in whole blood sample using the Coulter system (Coulter Corporation, Miami, USA). One of the EDTA tubes collected in biobank participants was used for DNA isolation. In addition, DNA was isolated from buccal swab samples for twin participants within the biobank and for other NTR participants as well, as part of various projects. For a large number of these samples genome-wide data were obtained using one of several platforms. In chronological order the following platforms were used: Perlegen/Affymetrix 5.0 600K, Illumina 370, Illumina 660, Illumina Omni Express 1 M and Affymetrix 6.0. Currently genotyping is also being done on the Axiom array from Affymetrix (Ehli et al. under revision [78]). As some samples were included on multiple platforms it was possible to align the data from the different sources and then impute to 1000 Genomes reference set. The genotypic data thus generated were used in the GWAS and GCTA studies presented in this thesis. Regarding EWAS data, the DNA isolated from whole blood was bisulfate-treated using the ZymoResearch EZ DNA Methylation kit (ZymoResearchCorp, Irvine, CA, USA) following the standard protocol for Illumina 450K micro-arrays [79].

1.5 Thesis outline

This thesis applies multiple genetic methods to pigmentation traits and hematological profiles, to explore the way individual differences in these traits can be explained by genetic and environmental factors.

Chapter 2 and chapter 3 focus on pigmentation traits. In Chapter 2, first the heritability for hair color is determined, followed by a GWAS to identify genetic variants for this trait. Finally, GCTA is performed to estimate how much of the genetic variance can be explained by the measured SNPs. The appendix of chapter 2 contains the supplementary to the article on hair color. In chapter 3, the genetic correlation between eye color and hair color was estimated using European unrelated population. The appendix contains the supplementary to the article: GWAS results for eye color.

Chapters 4 to 9 focus on three hematological ratios: the neutrophil-lymphocyte ratio (NLR), the monocyte-lymphocyte ratio (MLR) and the platelet-lymphocyte ratios (PLR). Chapter 4 examines the genetic and environmental causes of variation in NLR and PLR. The heritability of NLR and PLR is estimated using a parent-offspring design, in which the majority of the offspring consists of mono- and dizygotic twin pairs. In addition, the associations of NLR and PLR with sex, age, smoking behavior, body mass index and seasonal conditions at the time of blood sampling were examined. Specific analyses regarding age and BMI interaction effects on NLR and PLR and the results of analyses in the unhealthy population and total population are presented in the two appendices of

Chapter 4. In chapter 5, GWAS and GCTA was applied to these ratios to identify the genetic variants and estimate the percentage of the variance in NLR and PLR explained by these genetic variants, as well as eQTL mapping to detect these genetic variants affected gene expression level. The detail of eQTL analysis is describe in Appendix V. Chapter 6 presents the heritability, GWAS, GCTA and eQTL results for MLR. Chapter 7 explores the multivariate hematological profile by comparing the results of a univariate and a multivariate analysis of hematological profile parameters. First, a standard linear univariate regression was conducted to determine for each individual hematological index whether it was associated with age, sex, smoking, body mass composition and its interactions. Next, MDMR was introduced to establish hematological profiles using all hematological indices simultaneously and the interactive effects of age, sex and lifestyle on the derived profiles was investigated.

In chapter 8 the association between the methylation profile and NLR and MLR is examined using EWAS meta-analysis. A EWAS study for PLR and a detail methylation profile investigation in the particular region which obtained from previous GWAS result was presented in chapter 9.

The thesis concludes in chapter 10 with a general summary of the results presented in this thesis and a general discussion.

Part I: Visible traits

Chapter 2

Heritability and Genome-Wide Association studies for hair color in a Dutch twin family based sample

This chapter is published as: Lin BD, Mbarek H, Willemsen G, Dolan CV, Fedko IO, Abdellaoui A, de Geus EJ, Boomsma DI, Hottenga JJ., *Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample.* Genes, 2015. **6**(3): p. 559-576.

Abstract

Hair color is one of the most visible and heritable traits in humans. Here, we estimated heritability by structural equation modeling (N = 20,142), and performed a genome wide association (GWA) analysis (N = 7091) and a GCTA study (N = 3340) on hair color within a large cohort of twins, their parents and siblings from the Netherlands Twin Register (NTR). Self-reported hair color was analyzed as five binary phenotypes, namely "blond *versus* non-blond", "red *versus* non-red", "brown *versus* non-brown", "black *versus* non-black", and "light *versus* dark". The broad-sense heritability of hair color was estimated between 73% and 99% and the genetic component included non-additive genetic variance. Assortative mating for hair color was significant, except for red and black hair color. From GCTA analyses, at most 24.6% of the additive genetic variance in hair color was explained by 1000G well-imputed SNPs. Genome-wide association analysis for each hair color showed that SNPs in the *MC1R* region were significantly associated with red, brown and black hair, and also with light *versus* dark hair color. Five other known genes (*HERC2*, *TPCN2*, *SLC24A4*, *IRF4*, and *KITLG*) gave genome-wide significant hits for blond, brown and light *versus* dark hair color. We did not find and replicate any new loci for hair color.

Keywords: hair color, twin-family based heritability, GRM based heritability, Genome wide association study

2.1 Introduction

Hair color is a genetic and physiologically complex phenotype, which represents one of the most visible variations within humans and between populations [80]. It influences many human interactions including spouse selection. Previous studies of the effect of hair color on social behavior and personality have shown that hair color is related to attractiveness, but it is unclear how strong this effect is [81-82]. In general, individuals with a lighter hair color, e.g., blond, have a higher probability to receive courtship solicitations, as compared to individuals with red hair. Dyeing the hair has been practiced for over four thousand years, and is still immensely popular today, with over 75 percent of American women dyeing their hair [83].

Pigmentation of skin and hair is important for protection against sunlight [84]. From a medical point of view, pigmentation is also relevant, as the mechanisms underlying pigmentation are involved in one of the most aggressive types of cancer, namely malignant melanomas. These tumors of melanocytes, cause about 75% of deaths related to skin cancer, with high rates of incidence in Caucasians, especially in northwestern Europeans [85]. This prevalence is associated with ultraviolet light (UV) exposure and the amount of skin pigmentation [86].

Hair pigmentation is a highly heritable trait. In Europeans, genetic factors explain a large part (61%–92%) of the variation in natural hair color, while the rest of the variation is due to environmental influences and measurement error [34, 87-88]. Previous studies have found several genes that are relevant in human pigmentation, especially in the melanosome biogenesis or the melanin biosynthetic pathways [45]. Differences in eye and hair color are mainly due to variation in the amount, type, and the packaging process of key pigment molecule melanin polymers produced by melanocytes secreted into keratinocytes. Melanin exists in two basic forms: brown-black eumelanin and yellow-red pheomelanin. Hair color mainly depends on the ratio of eumelanin and pheomelanin [89].

Studying the genetic background of a human pigmentation trait like hair color is useful to understand human evolution and biology and may have important applications to melanoma treatment and to forensics. Loss-of-function mutations at the melanocortin 1 receptor (*MC1R*) are known to be associated with a switch from eumelanin to pheomelanin production, resulting in a red or yellow coat color in animal models [90]. Over 30 *MC1R* variant alleles correlated with skin and hair color have been identified [91]. In addition to *MC1R*, the *HERC2* and *OCA2* genes, located close to the *MC1R* gene on chromosome 15, are also related to hair color. Studies have shown that *HERC2* regional variants function as enhancers regulating *OCA2* transcription [92]. Variants located in and around these genes determine the normal human hair pigment variation.

To identify human pigmentation genetic variants and to gain new knowledge of genetic background of hair color in the Dutch population, we estimated the broad-sense heritability of hair color with a liability threshold model on twins and their family members from the Netherlands Twin Register (NTR) [93]. In addition, we estimated SNP heritability based on the measured and imputed SNP data from unrelated NTR individuals, by associating their genetic relatedness to their phenotypic resemblances using GCTA [6]. Finally, to identify genetic variants underlying the heritability of hair color, we performed a series of genome-wide association studies (GWAS) of hair color in this population-based sample.

2.2 Materials and methods

2.2.1. Subjects

Hair color data were available for 25,201 NTR participants, clustered in a total of 7862 families, with ages ranging between 14 and 80 years old. Each participant completed one or more longitudinal surveys that included questions about age, sex, and natural hair color. The self-reported hair color was obtained from a question with five answer possibilities: “fair/blond”, “light brown”, “red/auburn”, “dark brown”, and “black”. Written informed consent was obtained from all participants. The Medical Ethics Committee of the VU University Medical Center approved the study protocols.

For the twin-family based heritability analyses, we selected families comprising at most 6 members (N = 20,142). The subjects included in these families were the twins (N = 15,359), a maximum of two siblings (N = 2008), and the biological father and mother (N = 2774) (**Table S1**). Of the twins, 2345 were monozygotic males (MZM), 4651 were monozygotic females (MZF), 1678 dizygotic males (DZM), 2710 were dizygotic females (DZF) and 3975 were from an opposite-sex (DOS) twin pair (T**able S2**). For the GWAS analyses, a total number of 7091 related subjects were available with genotype, phenotype and covariate data (**Table S3**). For GCTA, only unrelated people were selected from the available GWAS sample, resulting in a sample of 3340 individuals.

2.2.2 DNA sampling and genotyping

Buccal or blood DNA samples (N = 14,003) were collected for multiple NTR projects. DNA extraction and purification of these samples were performed at various points in time, following several manufacturer specific protocols to obtain the best quality and concentration prior to SNP platform genotyping [6]. Genotyping of several partly overlapping subsets was done on multiple platforms. Chronologically the following platforms have been 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 (BIRDSUITE, APT-GENOTYPER, BEADSTUDIO).

Quality control was done within and between platforms and subsets prior to imputation. For each platform, the individual SNP markers were lifted over to build 37 (HG19) of the Human reference genome, using the LiftOver tool. 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 GIANT phase1 release v3 20,101,123 SNPs INDELS SVS ALL panel. 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. From each platform, SNPs with a Minor Allele Frequency (MAF) <0.01 were removed, as well as SNPs that were out of Hardy–Weinberg Equilibrium (HWE) with p < 0.00001. 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.

After these steps, the data of the individual arrays were merged into a single dataset using PLINK 1.07 [94]. Within the merged set, identity by state (IBS) sharing was calculated between all possible individual pairs and compared to the known family structure of the NTR study. Samples were removed if the data did not match their expected IBS sharing. DNA samples, which were typed on multiple platforms, were tested to ascertain that the concordance rate among overlapping SNPs exceeded 99.0%. If the concordance rate was lower, we removed all data of these samples. Subsequently, from each MZ twin pair a single DNA sample was selected. The HWE-, MAF- and the reference allele frequency difference <0.20 filters were re-applied in the combined data. As a final step, SNPs with C/G and A/T allele combinations were removed when the MAF was between 0.35 and 0.50 to avoid incorrect strand alignment.

Phasing of all samples and imputing cross-missing platform SNPs was done with MACH 1 [19, 95]. The phased data were then imputed with MINIMAC [96] in batches of around 500 individuals for 561 chromosome chunks obtained by the CHUNKCHROMOSOME program [97]. After imputation, DNA confirmed MZ twins were re-duplicated back into the data. The format of the data was transformed to the basic three probabilities SNPTEST gen.gz format, as this is the most general applicable format for the subsequent genomic analyses tools. The mean imputation quality R2 metric is 0.38 (based on all
30,051,533 imputed autosomal SNPs).

After imputation, SNPs were filtered based on the Mendelian error rate in families. The Mendelian error rate was calculated on the best guess genotypes in families (trios and sib-pairs with parents) using first GTOOL to calculate best guess genotypes and then PLINK 1.07 to analyze the data. SNPs were removed if the Mendelian error rate >2%, if the imputed allele frequency differed more than 0.15 from the 1000G reference allele frequency, if MAF < 0.005 and if R2 < 0.30. HWE was calculated on the genotype probability counts for the full sample, and SNPs were removed if the p-value < 0.00001. This left 7,981,681 SNPs prior to the statistical analyses.

2.2.3 Statistical analyses

First, we created binary variables for each hair color, representing a given hair color *versus* the other hair colors. In the following analysis, we found a much lower heritability for the light brown and dark brown classification than for other classifications. By combining light and dark brown hair color into a single brown category, we obtained a more sensible heritability estimate as compared to the other colors. Assuming this discrepancy is due to confusion among participants concerning the distinction light *versus* dark brown, we decided to proceed with a simple category "brown". An additional binary variable, denoted "light *versus* dark", was created representing blond and red hair *versus* brown (light and dark) and black. We studied this binary variable to detect the genes involved in the switch from eumelanin to pheomelanin. Hence, in total we analyzed the following 5 binary classifications, namely "blond *versus* non-blond", "red *versus* non red", "brown *versus* non-brown", "black *versus* non-black", and "light *versus* dark".

2.2.4 Genetic covariance structure modeling of hair color

To obtain an indication of family clustering, we calculated (tetrachoric) correlations among family members for binary hair color variables in OPENMX [98]. We adopted a liability-threshold model, in which the probability of having a given hair color (say blond) is a function on the position on the continuous (standard normal) liability scale. The tetrachoric correlations express the association between family members at the level of the liability. The actual proportion of the given hair color in the sample is expressed in terms of the threshold, i.e., a point on the liability scale. For instance, a threshold of zero on the standard normal liability scale is associated with a proportion of 0.50 (e.g., 50% are blond). The more extreme the threshold is, the greater or smaller the proportion (threshold of 1 implies ~84% are blond; a threshold of −1 implies ~16% are blond). OPENMX was then used to estimate the heritability of hair color by fitting a genetic model to the liabilities of the parents and offspring (twins and sibs). The total liability variance was decomposed into genetic and environmental variance components [99]. We started with a model, which included the following effects: (1) the spousal correlation to account for phenotypic assortative mating (see supplemental material); (2) quantitative sex differences in the genetic and genetic influences on the liability; and (3) age effects on the threshold. Our visual inspection of the tetrachoric correlations strongly suggested the absence of common (or shared) environment influences for the hair colors blond, brown, and the dark *versus* light dichotomy [100] (i.e., the MZ correlation was higher than twice the DZ correlation). The twin correlations for the red and black hair color suggested the possible presence of common environmental influences (C). We focused on the model including additive genetic effects (A), which represent the sum of the additive effects of all loci relevant to the trait, genetic dominance effects (D), which represent interactions between alleles at the same locus, and unique (or unshared) environmental effects (E), which are not shared by family members. Note that the unique environmental factors may include genetic causes, such as personal genetic mutations, and also measurement errors. In the analyses of red and black hair, we also considered the model including C instead of D. We accommodated the phenotypic assortative mating by deriving the expected correlation between the family members taking into account phenotypic assortment (more detail in Supplementary Material 1).

Since the prevalence of hair color may vary with age, we included age (Z transformed) as covariates on the liability threshold. We fitted the full model as described, and tested various effects by dropping the effects and conducting a likelihood ratio test. In this fashion we tested whether the prevalence of hair color varied with age, whether the spousal correlation was zero (i.e., random mating for hair color), whether sex effects in genetic architecture were absent, and whether dominance effects (or common environmental effects in the cases of red and black hair) were absent. Given the power that our large sample size confers to detect minor effects, we tested the specific effects mentioned using a stringent alpha of 0.0001 [101].

Since hair color displays a geographical gradient in the Netherlands (e.g., blond is more prevalent in the northern provinces), we refitted the full twin model in the subsample of genotyped individuals (5777 individuals in 2225 families) while correcting for the gradient using their data on three Dutch genotype principal components (PCs). These PCs were calculated from the genotype data with the EIGENSOFT software [102]. First, 10 PCs were calculated by projecting the NTR data on the 1000 Genomes reference set and all individuals with non-Dutch ancestry were excluded. Then, three new only Dutch PCs within the NTR samples were calculated, which correlate highly with the geographic location [103]. We examined if these PCs affected the broad heritability with blond, brown and light *versus* dark hair colors (the rare black and red hair colors excluded) by including these PCs in the ADE model as covariates on the threshold.

2.2.5 Variance explained by common SNPs (GCTA)

The proportion of variance of the binary hair color traits that can be explained by the measured and imputed SNPs was estimated in GCTA (Genome-wide Complex Trait Analysis) using the Restricted maximum likelihood (REML) analysis procedure under a case-control design, where we report the proportion of genetic variance explained on the underlying liability scale untransformed for prevalence [6]. Sex and age were included as covariates and the three Dutch PCs were alternated. A single genetic relationship matrix

(GRM) was build to test all hair colors. From the 1000 genomes imputed and cleaned data, we selected SNPs with a minimum imputation R2 quality metric of 0.80 and MAF > 0.01. In order to avoid explained variance and artificially increased GRM differences due to differing platforms and subsamples, additional SNP Quality Control (QC) included an evaluation of the SNP platform effects. We tested the effect of different platforms and removed SNPs showing platform effects. This was done by defining individuals on a specific platform as cases 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-value < 0.00001. The selected 5,987,253 SNPs were transformed to best guess Plink binary format, and subsets were made for each of the 22 chromosomes. The GRM for all NTR samples was then calculated per chromosome and subsequently the 22 matrixes were merged into a single autosomal GRM using GCTA. Ethnic outliers based on the PCs were excluded and 3340 unrelated individuals with hair color data were selected using the standard GRM cut-off filter of 0.025.

2.2.6 GWA analyses

The GWA analyses were run for all binary hair color variables. The input SNPs used were all 7,981,681 that passed the initial imputation QC. However, post GWAS analyses we additionally filtered the SNPs, depending on the cases sample size for MAF and imputation quality in the sub-sample of individuals with a hair color phenotype. Re-filtering on MAF > 0.01 was done after the GWAS analyses for blond, brown, and light *versus* dark hair color. For red and black hair color a MAF > 0.05 was selected to account for the lower prevalence of these hair colors. For imputation quality we filtered on the Plink information criterion, which is similar to R2: the variance of the mean posterior genotype probabilities divided by the maximum expected variance given full HWE and complete known genotypes. In all hair colors we filtered with 0.40 < Plink Info < 1.02. In total 6,473,680 SNPs survived this QC leading to a mean original imputation R2 of 0.77 (0.22) for MAF 0.01–0.05 and a mean R2 of 0.97 (0.07) for MAF ≥ 0.05 for the selected SNPs.

Since, hair color is affected by population stratification; we used the three Dutch ancestry PCs as covariates [103] and excluded ethnic outliers, similar to the twin modeling. Other covariates that were included were binary dummies for genotyping platform, sex and age. Analysis was performed with the PLINK 1.07 software running a logistic regression on each SNP, taking genotype inaccuracy into account using dosage data in the analyses. Because the GWAS data includes family members, we included the family option in our analyses, which takes the familial structure into account using a sandwich estimator [104]. The assortative mating of parents was corrected with the same familial-based correction [105]. Only one of the two monozygotic twins was selected for the GWAS analyses in the case that there was hair color data for both. For the GWAS, we assume a p-value less than 5×10^{-8} to be statistically significant [21].

2.3 Results and discussion

2.3.1 Prevalences and phenotypic tetrachoric correlations

The prevalences and the familial tetrachoric correlations based on N = 20,142 (clustered in 7497 families) for blond, brown, red, black and light *versus* dark hair colors are presented in **Table 1**. The prevalence of the blond, brown, and light *versus* dark hair colors ranges from 39% to 53%, but the prevalence of red and black hair colors is appreciably lower (red: 4.5%; black: 3.4%).

The MZ correlations are consistently high, ranging from 0.93 to 0.99. The full sib correlations, including the DZ twins, are lower (range 0.14–0.86) and that also goes for the parent-offspring correlations (range 0.19–0.73). Overall, the pattern of correlations is consistent with the expected large genetic contribution to hair color variation. In all but the red and black hair colors, the correlations suggest the presence of additive genetic and dominance effects, as the MZ correlations (range 0.93–0.99) are appreciably higher than the full sibs (including the DZs, range 0.14–0.49) and the parents and offspring (range 0.33–0.58). The correlations among first-degree relatives for red and black hair color tend to be larger than for the other hair colors. However, given the lower prevalence of these hair colors, these correlations are subject to larger standard errors. The spousal correlations suggest weak assortative mating with respect to hair colors (except black hair where the correlation is −0.179 and red hair where the correlation is 0.528). We note that the spousal correlations may be due to direct phenotypic assortment for hair color, or may be related to the geographic population structure of the Netherlands, as was established previously in our data [103].

2.3.2 Genetic covariance structure modeling

The results of fitting all models are shown in **Tables S4** and **S5**. Briefly, the ADE model optimally fits for the blond, brown and light *versus* dark hair colors. Furthermore, the results unambiguously show that no effects could be dropped, i.e., age effects, phenotypic assortment, sex differences on the genetic architecture, and non-additive genetic effects are present (all p-values are <0.0001). The results pertaining to the red and black hair colors are mixed. In the ACE model, as fitted to the red hair color, there is no sex difference (**Table S5**; p-value = 0.05), but all other effects are present (**Table S5**, p-value < 0.0001). With respect to the black hair color, we find a simple AE with age effects, but no assortative mating (p-value = 0.07), no C (p-value = 0.21) and no sex differences (p = 0.56).

The absence of specific effects in the red and black hair colors may be due to low power as a consequence of the low prevalence of these hair colors.

Table 2 shows the parameter estimates as obtained in the best fitting model. With respect to the genetic influences, we find that the broad sense heritabilities in the ADE models are high (over 0.90). The narrow sense heritability of red hair color, as obtained in the ACE models, is lower (0.73), and the narrow sense heritability of black hair color is 0.96. Inclusion of the genotype based three Dutch PCs in the models, to account for Dutch genetic population stratification and the geographical gradient of hair color in the Netherlands, did not strongly alter the estimates of heritability, phenotypic assortment and variance decomposition (**Table S6**)**.** However, the PCs do significantly explain hair color variance in all modeled colors (all p-values are $<1.0 \times 10^{-9}$).

Table 1. Polychoric correlation of individuals of the same family for different classification groups.

	Prevalence	rSpouse	rFS	rFD	rMS	rMD	rMZM	rMZF	rDZM	rDZF	rDOS	rBB	rSS	rBS
Blond	0.394	0.226	0.364	0.361	0.454	0.472	0.959	0.972	0.357	0.337	0.454	0.417	0.375	0.442
Brown	0.527	0.144	0.190	0.238	0.351	0.413	0.935	0.967	0.375	0.373	0.416	0.375	0.383	0.421
Red	0.045	0.528	0.576	0.485	0.725	0.593	0.978	0.934	0.477	0.610	0.608	0.406	0.529	0.636
Black	0.034	−0.179	0.334	0.352	0.413	0.274	0.928	0.991	0.706	0.857	0.567	0.143	0.752	0.570
Light versus dark	0.439	0.229	0.387	0.380	0.472	0.476	0.957	0.971	0.375	0.398	0.478	0.481	0.407	0.490

Prevalence: the prevalence of the given hair color in our sample (N = 20,142); rSpouse: phenotypic spousal correlation; rFS: father-son correlation; rMS: mother-son correlation; rFD: father-daughter correlation; rMD: mother-daughter correlation; rMZM: monozygotic male twin correlation; rDZM: dizygotic male twin correlation; rMZF: monozygotic female twin correlation; rDZF: dizygotic female twin correlation; rDOS: dizygotic opposite-sex twins correlation; rBB: brothers correlation, rSS: sisters correlation, and rBS: brother-sister correlation.

Table 2. The heritability and other parameters estimated from A (D or C) E model for hair color (N=20,142 individuals).

Classification	M	Am	Dm/Cm	Em	Af	Df/Cf	Ef	h2m	h2f
Blond (ADE)	0.210	0.39	0.57	0.04	0.73	0.24	0.03	0.96	0.97
Brown (ADE)	0.143	0.22	0.71	0.07	0.68	0.29	0.03	0.93	0.97
Red (ACE)	0.704	0.73	0.26	0.01	0.73	0.26	0.01	0.73	0.73
Black (AE)	0	0.96		0.04	0.96		0.04	0.96	0.96
Light versus dark (ADE)	0.208	0.42	0.54	0.04	0.75	0.22	0.03	0.96	0.97

M: phenotypic assortative mating coefficient (as estimated in the model); Am, Af = total additive variance in males and females; Dm, Df = total non-additive variance in males and females (blond, brown and light versus dark); Em, Ef = total unique environment variance plus measurement error in males and females; Cm, Cf = shared environment variance in males and females (red and black hair color); h2m = heritability in males, h2f = heritability in females (broad sense h2 of blond, brown, and light versus dark; narrow sense h2 of red and black).

2.3.3 Variance of hair color explained by autosomal SNPs using GCTA

The GCTA analyses in unrelated individuals show that depending on the hair color, at most 24.6% of the hair color liability can be explained by the autosomal SNPs (**Table 3**). All SNPs of the top individual chromosomes explain between 1% and 16.3% of the liability. These SNPs are on chromosome 16 (where the *MC1R* gene is located) for red hair, on chromosome 15 (where the *HERC2* gene is located) for brown and light *versus* dark, and on chromosome 6 (where the RPS6KA2 gene is located) for black hair color. We also studied the explained liability given by the top SNPs that are already known hair color loci as reported in earlier studies. These SNPs explain between 0.5% and 6.9% of the hair color liability in our sample.

The estimated genetic liability explained by common SNPs is low (<30%), given the fact that the heritability of hair color as a trait is very high (>70%). There are several possible explanations for this: GCTA is less appropriate for binary data than for quantitative phenotypes; there are not many genes related to hair color; the distribution of effect sizes of the genes and a combination of assortative mating plus possible common environment and dominance for this trait are not accounted for by the modeling assumptions of GCTA. In addition, SNPs were filtered by MAF > 0.01 and R2 > 0.80, leading to a large reduction of all SNP variants present within the GCTA matrix, and the coverage of rare alleles and less well imputed SNPs might therefore not be optimal. Also the LD tagging of SNPs might not be good enough to detect all hair color variants. Finally, for red and black hair color the results may be difficult to interpret due to the lack of power.

When adding the three Dutch PCs as covariates in the model, the estimates of the hair color liability explained by the known hair color loci do not change much (**Table 4**). However, the estimates of the total autosomal variation, as well as the top chromosomes all drop to almost 0 for blond and brown hair (and therefore light *versus* dark). Although standard errors do not show significant differences between the estimates, it indicates there are still unknown variants that determine blond and brown hair color, which are captured by the PCs, or there are variants that determine traits related to hair color (population stratification).

2.3.4 GWA analysis

In total, genotype and hair color data were available for 7091 subjects from the NTR. We performed five case-control GWAS with logistic regression for all SNPs including age, sex, the three Dutch PCs and genotype platform as covariates. Familial structure was taken into account in the analysis using Plink and selecting a single monozygotic twin. The resulting QQ and Manhattan plots for all colors and light *versus* dark hair color are shown in **Figures S1–S5**. As shown with GCTA and with the twin heritability modeling, the PC's are significantly related to hair color in the Netherlands along the three Dutch major axes of genetic variation (**Figures S6** and **S7**). Using the PCs, we corrected for this population stratification, with post correction GWAS λs ranging between 1.004 and 1.027. However, as a consequence, we have likely also reduced the significance of SNPs, which are truly associated with hair color.

2.3.4.1 Known hair color variants in relation to the GWAS results

Table 5 and **Table 6** display the gene variants for hair color, which are known from previous studies and our most significant SNPs within these genes. The two genes *HERC2* (15q11.2-13), along with neighboring gene *OCA2*, are known as the most essential genes for determining human pigmentation traits including eye, hair and skin color [92, 105]. These genes also show strong signals on chromosome 15 in our study, with SNPs rs7495174 and rs79097182 for blond, brown and light hair color.

The solute carrier (SLC) gene family group is a large family gene group that consists of 458 genes in 52 families. Three loci have been found to be associated with human pigmentation: *SLC24A5*, *SLC45A2* and *SLC24A4* gene. Interactions between *HERC2* and *SLC24A4* play a role in determining blue eye color, but also light hair color and less tanning ability [106-107]. *SLC24A4* (14q32.12) encodes the sodium/potassium/calcium exchanger 4 protein (NCKX4). Alternative splicing of this gene results in multiple transcript variants. Variants in *SLC24A4* have been previously associated with eye and hair color, skin sensitivity to sun and cutaneous malignant melanoma [108]. We confirmed the associations for hair color. Within *SLC24A4*, rs8014907 was significantly associated with all hair colors, except red. The *SLC45A2* gene, which is in the same family as *SLC24A4*, encodes a transporter protein that mediates melanin synthesis. The protein of *SLC45A2* is expressed in a high percentage of melanoma cell lines. Mutations in this gene are a cause of oculocutaneous albinism type 4, and polymorphisms in this gene are associated with variations in skin and hair color [109-110]. Multiple transcript variants encoding different isoforms have been found for this gene. Our results do show a p-value = 9.1×10^{-5} for the gene. *SLC24A5* was found to be involved in skin pigmentation in European populations [107]. A 4-bp insertion (c.569_572insATTA rs1426654) in the *SLC24A5* gene, causing a frame shift and premature termination, was identified in a man with Indian ancestry [111]. Homozygosity of this insertion results in extreme hypopigmentation and pinkish-white skin, with dark brown hair and a brown iris [109]. However, we have not found any significant hits in this locus. For *SLC24A5* gene the lowest p-value is 0.7 and the question is whether we have people with this particular insertion present within our Dutch population.

MC1R (16q24.3) is an intron-less gene of the size less than 1 kb. Non-synonymous variants in *MC1R* are present in approximately 50% in the European population [112]. Its multiple variants were first found to be associated with human red hair color in 1995 [113]. A subsequent study found these variants to have the same effect on pigmentation at increased frequency with increasing latitude in humans [114-115]. We replicate this association in our study, as *MC1R* is associated with multiple hair colors, except blond.

KITLG (12q22) is known to regulate the number of melanocytes during development, melanin distribution and hyper/hypo pigmentation. Sequence variation is thought to affect expression of *KITLG* (184,745), which results in the blond hair color. In European populations rs12821256 T/C SNP is found to be associated with blond hair color [116-117], and this SNP explains 3%–6% of the variance [87]. Our study confirmed this result, rs12821256 showed a significant associations with blond, brown and light *versus* dark hair color. Recently, a functional study showed that this SNP alters a transcription factor binding site for lymphoid enhancer-binding factor 1 (*LEF1*), reducing *LEF1* responsiveness and enhancer activity in cultured human keratinocytes [118].

IRF4 (6p25.3) is associated with hair color and skin pigmentation [117]. There is a strong association of the A allele of a single-nucleotide polymorphism (SNP) on chromosome 6p25.3, rs1540771, with the presence of freckles in Icelandic and Dutch population samples (discovery OR = 1.40, p-value = 3.7×10^{-18}) [119]. In our study, the most significant SNP rs62389424 is a bit further away (34 kb) and is associated with all hair colors except red ($p < 1.3 \times 10^{-5}$).

Genetic variants in 3-prime-untranslated region of the *ASIP* result in skin/hair/eye pigmentation variation. 'The *ASIP* haplotype', rs1015362G and rs4911414T was shown to be associated with red hair color, freckling, and skin sensitivity to sun, in addition to burning and freckling that reached genome wide significance (max odds ratio = 1.60, p-value = 3.9×10^{-9}) in the Icelandic and Dutch populations [119]. In our study, however, we did not find any significant results for the *ASIP* locus. Minimum p-values for these genes were above 0.35 for the hair colors. Note that red hair color in our sample is not so frequent, so this could be related to detection power.

TPCN2 (11q13.3.) was found to be significantly associated with blond *versus* brown hair color in Icelandic Europeans [117]. In our study SNPs in this locus are significantly associated with blond, brown, black and light *versus* dark hair colors ($p < 4.7 \times 10^{-10}$).

The variants of *TYR*P1 gene are known skin/hair/eye pigmentation variation loci. A suggestive association for blond *versus* brown hair was observed for rs1408799 in Iceland and Dutch populations, and functional data suggest that the *TYR*P1 gene encodes a melanosomal enzyme with a role in the eumelanin pathway [119]. The p-values in our study are about 10^{-4} implying a potential, but weak association with hair color.

Table 3. The explained genetic variance of the hair color liability scale for autosomal common SNPs in GCTA.

Phenotypes	N case	N control	Proportion of genetic variance explained by all common SNP (SE)	p	Top chromosome	Proportion of genetic variance explained by	p	Proportion of genetic variance explained by Known	P
Blond	1547	1793	0.165(0.081)	1.1E-02	15	0.014(0.017)	1.8E-01	0.058(0.022)	5.8E-40
Brown	1946	1394	0.095(0.079)	9.3E-02	15	0.011(0.016)	2.5E-01	0.059(0.022)	7.9E-39
Red	87	3253	0.246(0.087)	1.9E-03	16	0.163(0.025)	3.2E-14	0.069(0.026)	2.3E-55
Black	66	3274	<0.001(0.083)	5.0E-01	6	0.031(0.228)	7.7E-02	0.005(0.003)	1.2E-02
Light versus dark	1890	1450	0.140(0.080)	2.7E-02	15	0.014(0.017)	2.0E-01	0.069(0.026)	6.5E-46

SE: standard error. The given proportion of phenotypic variance explained by SNPs V(G)/V(P) without using a prevalence liability scale transformation.

Table 4. The explained genetic variance of the hair color liability scale for autosomal common SNPs in GCTA including the three Dutch PCs as covariates.

Phenotypes	N case	N control	Proportion of genetic variance explained by all common SNP (SE)	p	Top chromomsome	Proportion of genetic variance explained by top chromosome	p	Proportion of genetic variance explained by Known association	P
Blond	1547	1793	<0.001(0.084)	0.5	15	0.001(0.015)	0.48	0.054(0.024)	2.8E-37
Brown	1946	1394	<0.001(0.082)	0.5	15	<0.001(0.016)	0.5	0.054(0.021)	7.0E-39
Red	87	3253	0.165(0.084)	1.4E-03	16	0.255(0.083)	2.4E-14	0.053(0.020)	1.7E-55
Black	66	3274	<0.001(0.084)	0.5	6	0.027(0.228)	0.10	0.005(0.004)	2.4E-02
Light versus dark	1890	1450	<0.001(0.084)	0.5	15	0.001(0.016)	0.5	0.065(0.024)	1.5E-43

SE: standard error. The given proportion of phenotypic variance explained by SNPs V(G)/V(P) without using a prevalence liability scale transformation.

Table 5. SNP associations within our study for blond and brown hair against known hair color loci.

Locus	Chromosome location	The most significant SNP	MAF	OR Blond	SE Blond	P blond	OR brown	SE brown	P brown
SLC45A2	5p13.2	rs16891982*	0.050	0.5036	0.1752	9.1E-05	1.4506	0.1656	0.02472
IRF4	6p25.3	rs1540771	0.489	1.0253	0.1978	0.003	0.9027	0.0396	9.7E−03
IRF4	6p25.3	rs62389424*	0.087	2.4846	0.1188	1.3E-13	0.5551	0.1081	5.2E−08
TYRP1	9p23	rs1408799*	0.291	0.8522	0.0484	0.001	1.1846	0.0468	2.9E−03
TPCN2	11q13.3	rs72930659	0.109	0.6073	0.0657	3.2E-14	1.5544	0.0649	1.1E−11
KITLG	12q21.33	rs12821256	0.128	0.6715	0.0616	9.8E-11	1.4366	0.0596	1.2E−09
SLC24A4	14q32	rs8014907	0.178	1.4317	0.0554	3.0E-10	0.7262	0.0529	3.1E−09
SLC24A5	15q21.1	rs1834640	0.110	1.0163	0.1323	0.903	0.9858	0.1294	0.9121
HERC2	15q13	rs79097182	0.038	2.9822	0.1426	1.8E-14	0.4827	0.1235	3.7E−09
OCA2	15q13.1	rs7495174*	0.014	3.5485	0.2577	8.9E-07	0.3544	0.2168	1.7E−06
MC1R	16q24	rs2353688	0.028	1.8309	0.1432	2.4E-05	0.6028	0.1316	1.2E−04
MC1R	16q24	rs146972365	0.053	0.9765	0.0945	0.802	1.7771	0.0922	4.5E−10
MC1R	16q24	rs8063160	0.065	0.9119	0.0872	0.290	1.7837	0.0847	8.6E−12
MC1R	16q24	rs117322171	0.014	0.9469	0.1795	0.761	0.9039	0.1789	0.572
ASIP	20q11.22	rs1015362	0.273	1.0448	0.0464	0.344	0.9559	0.0452	0.3181
ASIP	20q11.22	rs4911414	0.347	0.9922	0.0428	0.855	1.0148	0.0417	0.724

* These SNPs failed the imposed quality control in our sample: rs62389424 (HWE p = 2.6E-21), rs7495174 (HWE p = 1.27203E-12), rs16891982 (HWE p = 1.05107E-05), rs1408799 (MAF difference with imputation reference set >0.15).

Table 6. SNP associations within our study for light *versus* dark, red and black hair against known hair color loci (same loci as **Table 5**).

Most significant SNP	OR red	SE red	p Red	OR Black	SE Black	p Black	OR Light versus Dark	SE Light versus Dark	p Light versus Dark
rs16891982*	0.8443	0.6410	0.792	4.9177	0.3258	1.0E−06	1.9867	0.1764	1.0E−04
rs1540771	0.6853	0.367	0.010	0.8300	0.1112	0.104	0.6822	0.307	0.002
rs62389424*	0.6897	0.3084	0.228	0.3985	0.2107	1.3E−05	0.4424	0.1171	3.4E−12
rs1408799*	0.9043	0.1467	0.493	0.9806	0.1355	0.885	1.1823	0.0478	4.6E−04
rs72930659	1.2327	0.2211	0.344	1.3688	0.1968	0.1106	1.6087	0.0653	3.4E−13
rs12821256	1.0088	0.1951	0.964	1.3023	0.1895	0.1634	1.4829	0.0612	1.2E−10
rs8014907	1.0271	0.1788	0.709	0.8070	0.1453	0.1400	0.7028	0.0547	2.6E−10
rs1834640	1.1650	0.4093	0.831	0.8708	0.3677	0.7067	0.9691	0.1315	0.811
rs79097182	0.5954	0.3135	0.098	0.4352	0.2254	2.2E−04	0.3900	0.1386	5.2E−14
rs7495174*	0.6939	0.5975	1.7E−06	0.9691	0.3839	0.9349	0.3216	0.2443	3.4E−06
rs2353688	0.5149	0.3862	0.086	1.4452	0.4024	0.3601	0.6148	0.1348	3.1E−04
rs146972365	0.0720	0.1674	1.1E−55	1.8403	0.3244	0.6007	1.8751	0.0924	1.0E−11
rs8063160	0.0770	0.1728	7.8E−50	1.9278	0.3041	0.3090	1.8891	0.0845	5.2E−14
rs117322171	1.7219	0.6122	0.375	452.3873	1.0414	4.3E−09	1.0119	0.1771	0.947
rs1015362	0.9583	0.1500	0.776	1.0513	0.1282	0.6967	0.9619	0.0464	0.402
rs4911414	0.7644	0.1371	0.725	1.1988	0.1200	0.1309	1.0366	0.0426	0.399

* These SNPs failed the imposed quality control in our sample: rs62389424 (HWE p = 2.6E-21), rs7495174 (HWE p = 1.27203E-12), rs16891982 (HWE p = 1.05107E-05), rs1408799 (MAF difference with imputation reference set >0.15).

2.3.4.2 Identification of new hair color variants from the NTR GWAS results

Within this study, no new SNPs were significantly associated with hair color after conservatively filtering on MAF > 0.01 for blond, brown and light *versus* dark hair color and MAF > 0.05 for red and black hair color. Initially we had some associations for black and red hair color when also filtering on MAF> 0.01. However, with this threshold of filtering the number of cases having the minor allele(s) is extremely small, which leads to inflated statistics. Subsequently, these were not confirmed as positive results as none of the red hair color findings (black hair color unavailable) were replicated by the Decode study, and permutation analyses showed that the findings were also likely under the hypothesis of no association (**Table S7**).

2.4 Conclusions

Our twin family analysis shows high heritability for hair color (70%–97%). Both additive and non-additive genetic models, as well as positive assortative mating and population stratification should be taken into consideration when conducting genetic studies of these traits in the Dutch population. In the GWA analysis we could confirm previously known associated variants in the *MC1R* and *HERC2*, *TPCN2*, *SLC24A4*, *IRF4* and *KITLG* genes. The GCTA analyses shows that common SNPs in these loci explain about 6% of the hair color liability in our population. In total, between 0% and 25% and, on average, roughly 13% of the hair color liability can be explained by common SNPs genome wide, and therefore new variants, either rare and tagged by the common SNPs or simply not identified yet, are likely still present within the genome. This study also shows the issues of current standard GWAS approaches. The modeling assumptions of GCTA assume additive effects with many genes influencing the trait. However, as made evident, there are non-additive effects, assortative mating, population stratification and the likelihood of involvement of rare gene variants and the question is thus whether estimates of the variance explained by these methods are optimal. To find the missing heritability, an investment in large sample sizes with meta-analysis and methodological innovation to deal with these other-than-additive circumstances, a stratified population and better rare allele detection is needed to improve locus detection, even for a highly heritable trait like hair color.

Appendix I.

Details on methodology

Variance decomposition including assortative mating within the applied twin model

The variance of the liability underlying the phenotype, V_{ph}, is standardized, *i.e.*, $V_{ph} = 1$. The variance is decomposed as follows $V_{ph} = V_A + V_D + V_E$, where, additive genetic (V_A), the dominance genetic (V_D) and the unshared environmental variance components (V_E) sum to one, and two parameters are estimated as free parameters. The broad-sense heritability equals $V_A + V_D$, and the narrow-sense heritability equals V_A. In the notation of Falconer and MacKay (1996) [13], we use r to denote the spousal phenotypic correlation, and m to denote the correlation between the parental breeding values, *i.e.*, $m = r \times V_A$. Given the standardization, we express the expected phenotypic covariances in terms of correlations. These are:

Spousal:	r
Parent-offspring covariance:	$½V_A (1 + r)$
Monozygotic twin correlation:	$V_A + V_D$
Dizygotic twin and full sib correlation:	$½V_A (1 + m) + ¼V_D = ½V_A (1 + r \times V_A) + ¼V_D$

The dizygotic twin, and full sib correlation, follows from $m = r \times V_A$ given the assumption of purely phenotypic assortment [13]. Note that the phenotype spousal correlation (r) may differ when it is estimated directly on the basis of the parental data as compared to the correlation estimated in the full model, as in the latter the dizygotic twin and full sib correlations also play a role.

We also considered an ACE model (in the case of red and black hair colors), where C stands for common environmental influences shared by family members. Letting VC denote the shared environmental variance, the expected correlations sib correlations are $V_A + V_C$ (MZs) and $½V_A (1 + r \times V_A) + V_C$ for full sibling and dizygotic twins.

Table S1. The sample of family members within the twin-family modeling study (N = 20,142).

Members	N	Age	Blond	Red	Light Brown	Dark Brown	Black
Fathers	1190	67.16 ± 6.52	458	39	227	349	117
Mothers	1584	64.40 ± 7.35	449	48	412	608	67
Twins	15,359	31.25 ± 14.19	6245	759	3840	4147	368
Brothers	370	31.27 ± 13.90	159	10	89	98	14
Sisters	1639	32.25 ± 13.72	772	55	410	384	18

Table S2. Zygosity information for the subsample of twins within the twin modeling study (N = 15,359).

Zygosity	N	Blond	Red	Light Brown	Dark Brown	Black
Monozygotic males	2345	904	101	534	709	97
Dizygotic males	1678	632	75	396	495	80
Monozygotic females	4651	1905	280	1196	1211	59
Dizygotic females	2710	1095	154	753	676	32
Opposite-sex twins	3975	1709	149	961	1056	100

Table S3. Characteristics for subjects in the GWA study (N = 7,091) and the sub-selection of unrelated individuals from these subjects for the GCTA study (N = 3,340).

Members	N	Blond	Red	Light Brown	Dark Brown	Black
Fathers	657	292	14	101	187	63
Mothers	1021	329	22	240	395	35
Twins	4320	1828	115	1103	1205	69
Siblings	882	353	25	246	235	23
Spouses	211	86	5	53	61	6
GCTA Unrelated	3340	1547	87	1019	927	66

Table S4. Results of the variance components model fitting for each hair color. The saturated model is compared to the ADE, ACE (where applicable) and AE models. Also the effects of including age as a covariate, sex limitations (quantitative and qualitative) and assortative mating are examined.

hair color	Model	P1	P2	−2LL	df	AIC	DLL	Ddf	*p*-Value
BLOND	SAT	15	15	23,837.1	20,127	−16,416.9			
	Full ADE	9	7	23,964.98	20,135	−16,105.02	127.88	8	1.74E−24
	ADE m = f	6	5	24,004.14	20,137	−16,271.86	167.04	10	1.14E−30
	Age = 0	8	6	24,135.06	20,136	−16,136.94	297.96	9	7.07E−59
	r = 0	8	6	23,983.35	20,136	−16,288.65	146.25	9	5.27E−27
	AE	7	5	24,123.37	20,137	−16,150.63	286.27	10	1.24E-55
BROWN	SAT	15	15	24,974.3	20,127	−15,279.7			
	Full ADE	9	7	25,027.81	20,135	−15,242.19	53.51	8	8.59E−09
	ADE m = f	6	5	25,080.32	20,137	−15,195.68	106.02	10	3.38E−18
	Age = 0	8	6	25,057.71	20,136	−15,214.29	83.42	9	3.37E−14
	r = 0	8	6	25,036.46	20,136	−15,235.54	62.16	9	5.13E−10
	AE	7	5	25,360.84	20,137	−14,917.16	386.54	10	6.88E−77
RED ACE	SAT	15	15	6250.44	20,127	−33,950.93			
	Full ACE	9	7	6336.7	20,135	−33,938.53	86.26	8	2.66E−15
	ACE m = f	6	5	6344.4	20,137	−33,939.55	93.95	10	8.79E−16
	Age = 0	8	6	6382.29	20,136	−33,895.88	131.85	9	4.94E−24
	r = 0	8	6	6419.43	20,136	−33,918.18	169.99	9	6.20E−32
	AE	7	5	6373.81	20,137	−33,915.86	123.37	10	1.05E−21

BLACK ACE	SAT	15	15	4597.64	20,127	−35,654.36			
	Full ACE	9	7	4618.18	30,135	−35,670.79	20.54	8	8.48E-03
	ACE m = f	6	5	4620.21	20,137	−35,665.82	22.57	10	0.012
	Age = 0	8	6	4795.28	20,136	−35,510.64	197.64	9	1.03E-37
	r = 0	8	6	4621.38	20,136	−35,645.17	23.74	9	4.7E-03
	AE	7	5	4621.3	20,139	−35,647.38	23.66	10	8.56E-03
LIGHT *vs.* DARK	SAT	15	15	24,356.28	20,127	−15,897.72			
	Full ADE	9	7	24,439.25	20,135	−15,830.75	82.97	8	1.23E-14
	ADE m = f	6	5	24,472.93	20,137	−15,803.07	116.65	10	2.42E-20
	Age = 0	8	6	24,533.94	20,136	−15,738.06	177.66	9	1.56E-33
	r = 0	8	6	24,456.93	20,136	−15,815.07	100.65	9	1.16E-17
	AE	7	5	24,578.55	20,137	−15,695.45	222.27	10	3.58E-42

P1: the number of parameters, P2: the number of independent parameters (taking into account the standardization of the phenotypic liability), −2LL: −2loglikelihood, df: total degrees of freedom, AIC: Akaike Information Criterion, DLL: difference in −2loglikelihood, Ddf: difference in degrees of freedom, compared to saturated model. SAT: saturated model, Full ADE: ADE model (including assortative mating (r), age as a covariate, and sex limitation), Full ACE: ACE model (including assortative mating (r), age as a covariate, and sex limitation), ADE/ACE m = f: ADE/ACE model without quantitative sex limitations, Age = 0: ADE/ACE model without age as covariate, r = 0: ADE/ACE model without assortative mating effect, AE: AE model without dominance genetic effects or common environment effects.

Table S5. Specific tests of including age as a covariate, sex differences and assortative mating as compared to the full ADE/ACE model for each hair color.

hair color	**Model Comparison**	**DLL**	**Ddf**	**P-value**
Blond	ADE m = f	39.16	2	3.14E−09
	Age = 0	170.08	1	7.11E−39
	r = 0	18.37	1	1.82E−05
	AE	158.39	2	4.04E−35
Brown	ADE m = f	52.51	2	4.28E−13
	Age = 0	29.91	1	4.53E−08
	r = 0	8.65	1	3.27E−03
	AE	333.03	2	4.82E−73
Red ACE	ACE m = f	7.7	2	0.05
	Age = 0	45.59	1	1.46E−11
	r = 0	82.73	1	9.41E−20
	AE	37.11	2	8.74E−09
Black ACE	ACE m = f	2.04	2	0.56
	Age = 0	177.1	1	2.08E−40
	r = 0	3.2	1	0.07
	AE	3.12	2	0.21
Light *vs.* Dark	ADE m = f	33.65	2	4.93E−08
	Age = 0	94.69	1	2.23E−22
	r = 0	17.68	1	2.61E−05
	AE	139.3	2	5.64E−31

DLL: Difference in −2 log likelihood, Ddf: difference in degrees freedom.

Table S6. Estimates from ADE model for hair color with—and without PC correction (compared with saturated model, N = 5777 individuals having genotype data and are fitting the twin modeling family characteristics).

Color	Model	Age	PC1	PC2	PC3	Rspouse	Am	Dm	Em	Af	Df	Ef	−2LL	P2	DLL	Ddf	P
Blond	M1	0.084	−11.33	−0.44	−1.81	0.281							6801.08	18			
	M2	0.083	−11.11	−0.43	−1.94	0.259	0.45	0.50	0.05	0.71	0.25	0.04	6830.85	10	29.77	8	2.32E−04
	M3	0.068				0.272	0.50	0.45	0.05	0.71	0.25	0.04	6876.9	7	75.83	3	2.41E−16
Brown	M1	−0.045	9.90	1.34	0.04	0.334							7049.92	18			
	M2	−0.048	9.44	1.11	0.06	0.225	0.23	0.71	0.06	0.65	0.30	0.05	7074.68	10	24.76	8	1.71E−03
	M3	−0.037				0.234	0.25	0.69	0.06	0.65	0.31	0.04	7110.94	7	61.02	3	3.56E−13
Light Dark	M1	−0.097	10.82	1.20	0.80	0.270							6868.74	18			
	M2	−0.098	10.56	1.14	0.83	0.248	0.46	0.49	0.05	0.74	0.22	0.04	6895.23	10	26.49	8	8.66E−04
	M3	−0.085				0.259	0.48	0.47	0.05	0.73	0.22	0.05	6937.31	7	68.57	3	8.64E−10

M1: saturated model with age and 3PC’s as covariates. M2: genetic models (ADE model for blond, brown and light *versus* dark hair), M3: genetics models without 3PC’s as covariates compared with full genetic model. Age, PC1, PC2, PC3: Beta coefficient of regression on age, and first to third principal component. rSpouse/r: tetrachoric correlation of spouse in M1, and assortative mating coefficient in M2 and M3, Am and Af = additive variance, Dm and Df = non-additive variance, Em and Ef = unique environment variance plus measurement error for males (m) and females (f), −2LL: −2loglikelihood, P2: the number of independent parameters (taking into account the standardization of the phenotypic liability), DLL: the difference in −2loglikelihood, Ddf: difference in degrees of freedom, compared to saturated model, P: *p*-Value of chi square test comparing different sub-models.

Table S7. Summary of results of standard association tests of rare SNPs (MAF 0.01–0.05) within the NTR discovery sample. Permutation tests indicate non-significance of the associations. Replication in the Decode cohort also indicate non-significance.

CHR	BP	Most Significant SNP	Hair Color	GENE	*p*-Value	MAF	Odds Ratio	SE	Permutation (N = 10,000) Empirical P	Decode Replication *p*-Value
1	199,471,603	rs74230273	black		1.90E−14	0.015	790	0.87	0.1487	NA
2	172,770,696	rs7563076	black		2.10E−10	0.013	>1000	1.11	0.2084	NA
3	5,984,546	rs149685327	black		1.40E−15	0.013	240	0.69	0.3316	NA
5	172,347,359	rs17074690	black		1.20E−08	0.024	61	0.72	0.0115	NA
6	166,860,270	rs2281057	black	RPS6KA2	1.40E−43	0.028	0.93	0.09	0.2237	NA
6	166,446,782	rs191122540	black	RPS6KA2	1.40E−43	0.028	>1000	13.35	0.3316	NA
7	122,530,292	rs183059797	black		4.50E−08	0.01	0.06	0.5	0.2754	NA
8	76,579,713	rs10993446	black		5.60E−09	0.012	160	0.87	0.2514	NA
9	97,759,920	rs12289701	black		9.00E−16	0.013	>1000	2.19	0.0106	NA
11	44,921,416	rs700019	red	TSPAN18	8.50E−17	0.015	320	0.69	0.0881	NA
1	216,416,300	rs4972217	red	USH2A	8.90E−22	0.012	>1000	1.83	0.3061	0.5765
2	88,842,075	rs7691567	red		5.70E−12	0.011	>1000	9.32	0.2213	0.7509
4	35,892,635	rs28407071	red		1.00E−27	0.015	>1000	1.61	0.0543	NA
7	111,917,427	rs16904010	red	ZNF277	8.80E−22	0.013	400	0.64	0.1537	NA
8	91,163,167	rs10506044	red		1.80E−16	0.014	1000	0.84	0.0514	0.7326
12	29,007,640	rs189573811	red	PDE3A	1.00E−19	0.019	>1000	10.34	0.0514	0.7665
18	6,083,752	rs7238024	red	MC4R	1.20E−22	0.016	>1000	1.8	0.0432	NA
19	4,513,559	rs4807597	red	PLIN4	2.40E−15	0.007	>1000	0.88	0.2391	0.4494
21	20,177,103	rs2825137	red		3.00E−19	0.012	>1000	1.97	0.0487	NA
22	26,842,941	rs112390186	red	HPS4	5.90E−12	0.015	64	0.61	0.3800	0.9565

Chr: chromosome, MAF: minor allele frequency, SE: standard error of odds ratio.

Some of these genes are biologically interesting candidates. *MC4R* belongs to the melanocortin receptor family, involved in a wide range of physiological functions, including pigmentation. Mutations in *HPS4* result in subtype 4 of Hermansky-Pudlak syndrome, a form of albinism as reported in Hutten *et al.* (2008) [120]. *USH2A* (Usher syndrome 2A) is previously found to be related to retinitis pigmentosa [121].

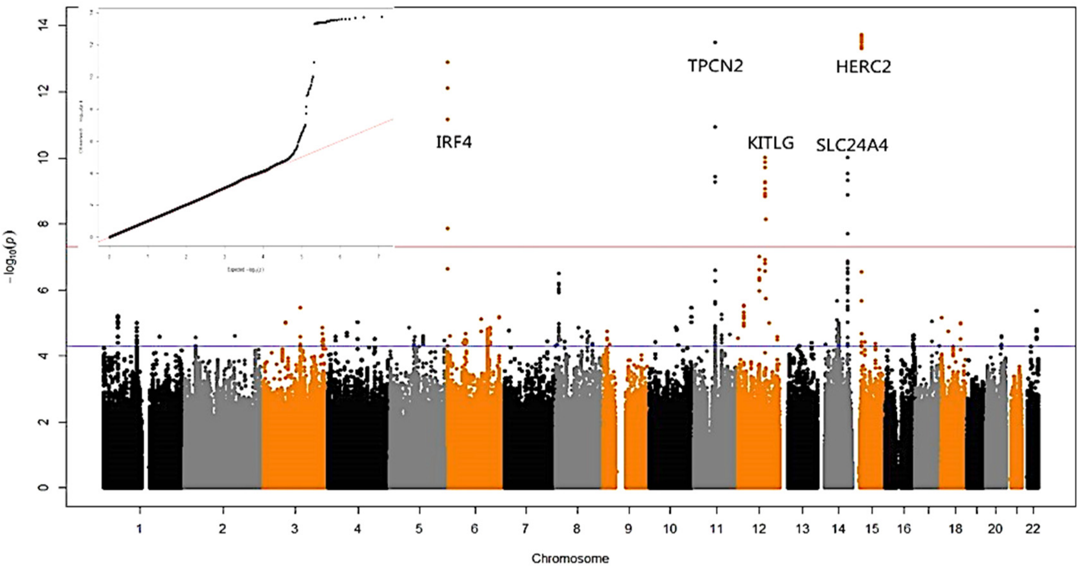


Figure S1. Manhattan plot and QQ plot for blond hair color (MAF > 0.01, λ = 1.004673).

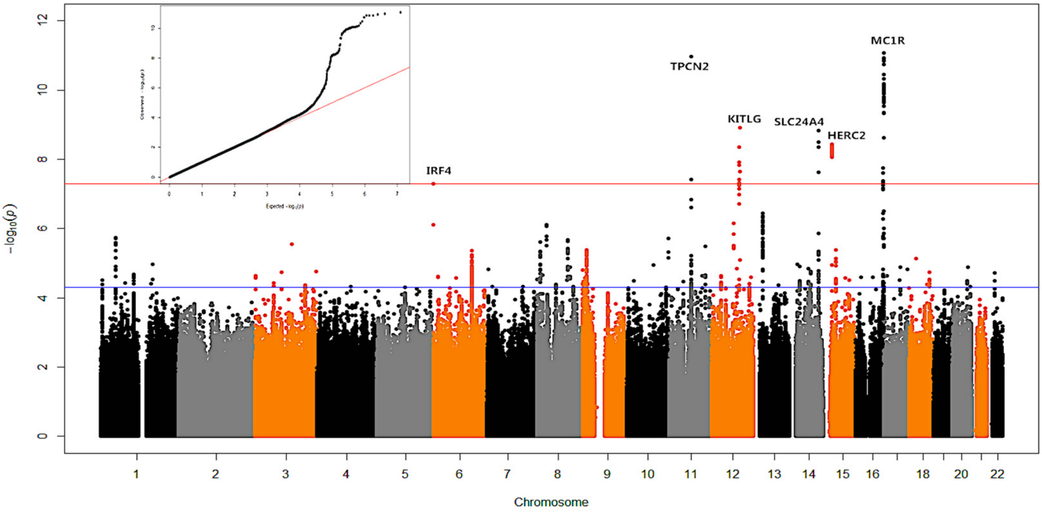


Figure S2. Manhattan plot and QQ plot for brown hair color (MAF > 0.01, λ = 1.003738).

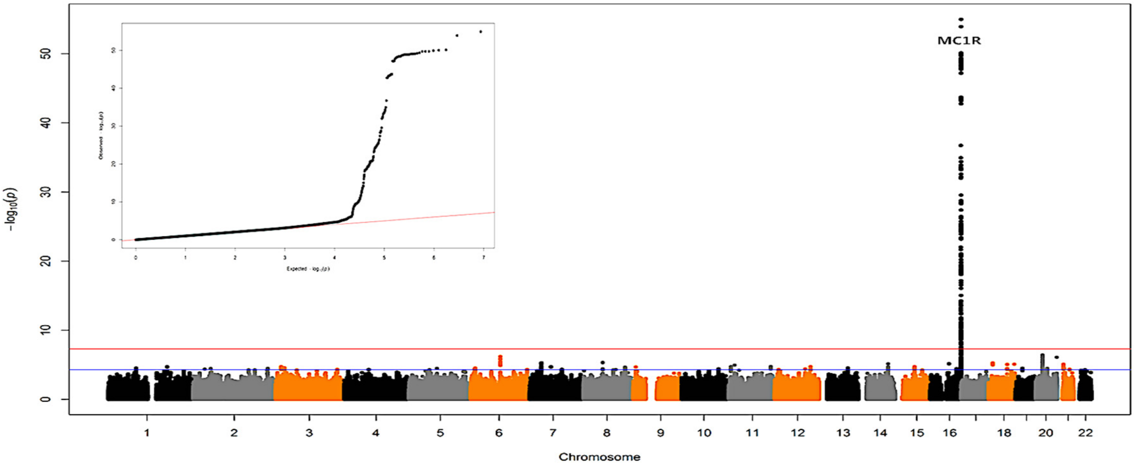


Figure S3. Manhattan plot and QQ plot for red hair color (MAF > 0.05, λ = 1.021156).

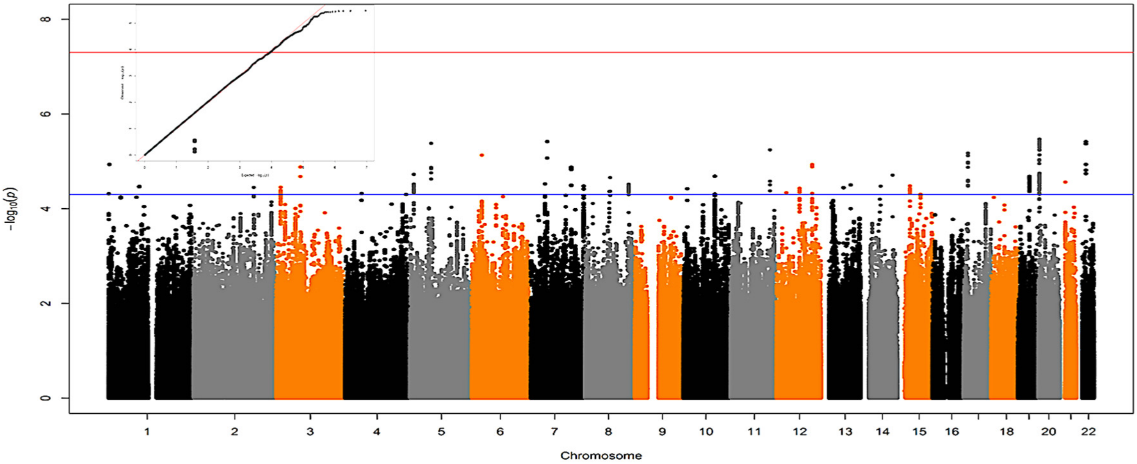


Figure S4. Manhattan plot and QQ plot for black hair color (MAF > 0.05, λ = 1.027328).

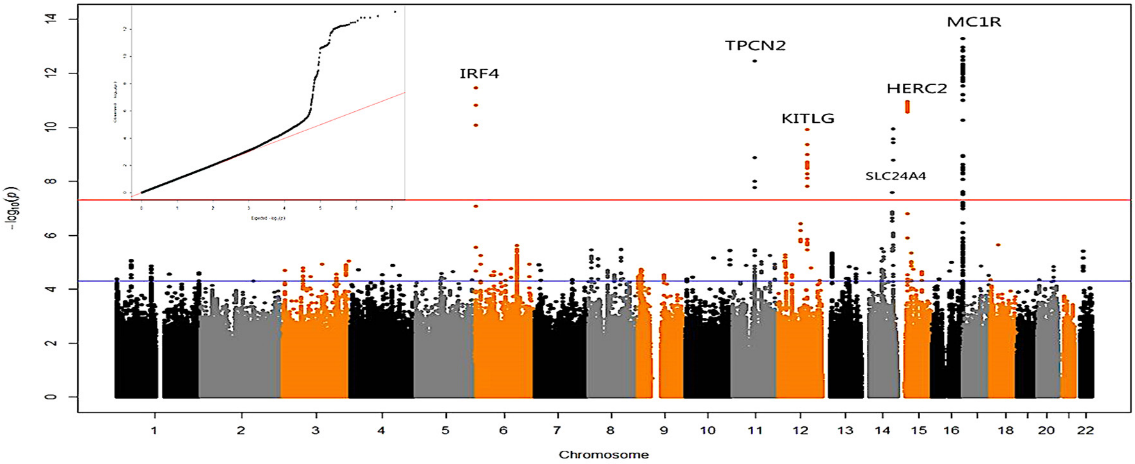


Figure S5. Manhattan plot and QQ plot for light vs. dark hair color (MAF > 0.01, λ =1.006079).

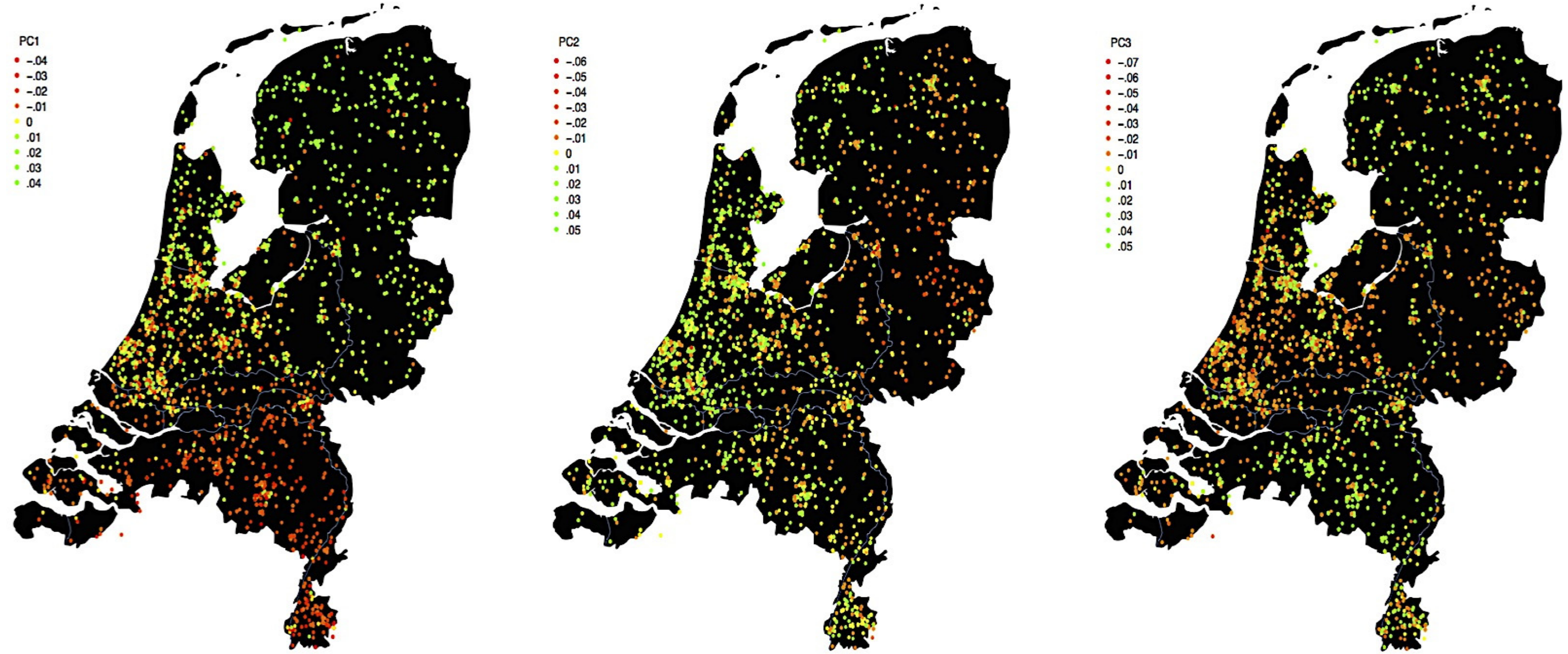


Figure S6. The three genomic PCs and their correlation with geography [103].

These figures show a map of the Netherlands, where the relation between geography and the genetic principal components 1 to 3 is plotted. The colors of the points indicate the mean value per postal code of PC1, PC2 and PC3 respectively. The plot is based on 7091 Dutch individuals with current address postal code information, hair color phenotype and genetic information [103].

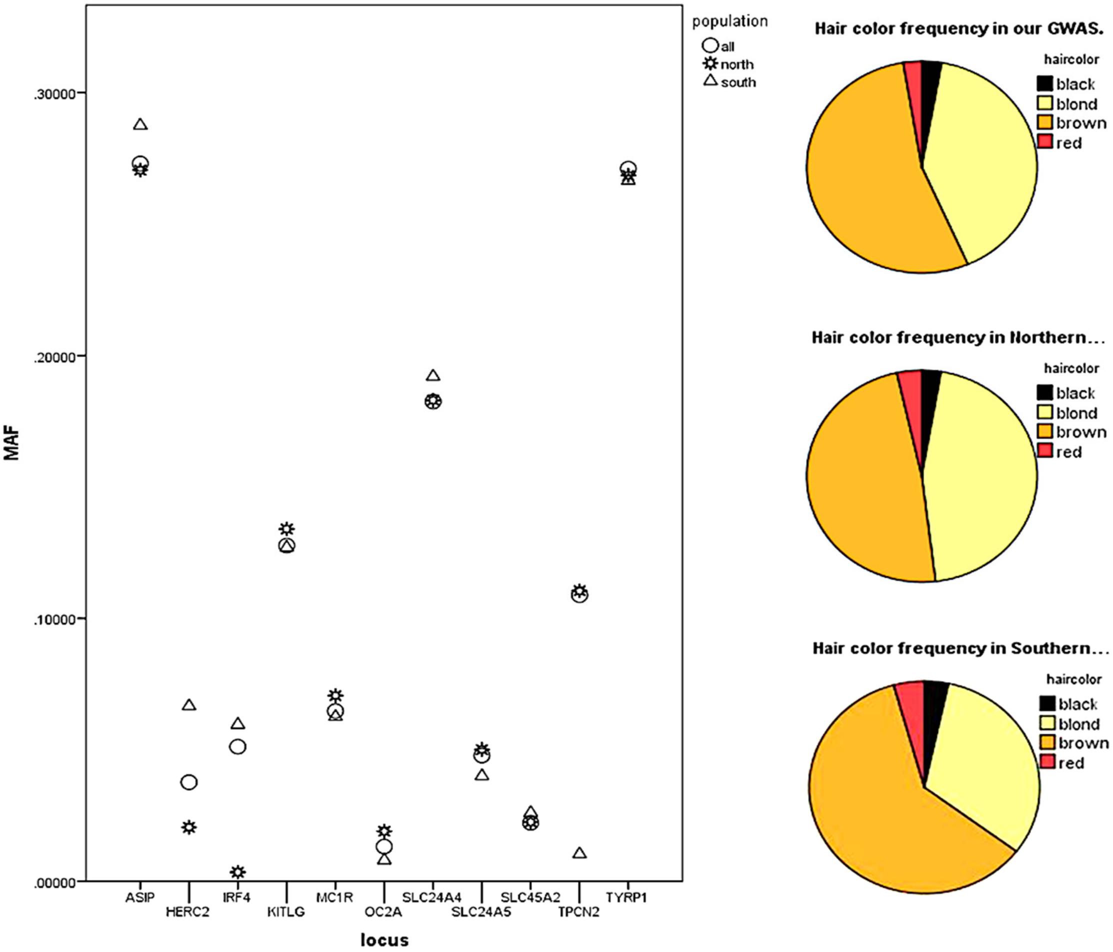


Figure S7. The minor allele frequency of the known hair color genes in relation to the hair color distribution, comparing the total GWAS sample to the 1000 PC1 based most northern individuals and 1000 PC1 based most southern individuals.

Chapter 3

The genetic overlap between hair and eye color

This chapter is published as : Lin BD, Willemsen G, Abdellaoui A, Bartels M, Ehli E, Davies G, Boomsma DI, Hottenga JJ., *The genetic overlap between hair and eye color*. Twin Res Hum Genet, 2016. **19**(6): p. 59.

Abstract

We identified the genetic variants for eye color by Genome-Wide Association Study (GWAS) in a European family-based population sample and examined the genetic correlation between hair and eye color using data from unrelated participants from the Netherlands Twin Register. With the GCTA software package, we found strong genetic correlations between various combinations of hair and eye colors. The strongest positive correlations were found for blue eyes with blond hair (0.87) and brown eyes with dark hair (0.71), whereas blue eyes with dark hair and brown eyes with blond hair showed the strongest negative correlations (-0.64 and -0.94, respectively). Red hair with green/hazel eyes showed the weakest correlation (-0.14). All analyses were corrected for age and sex, and we explored the effects of correcting for Principal Components (PCs) that represent ancestry and describe the genetic stratification of the Netherlands. When including the first three PCs as covariates, the genetic correlations between the phenotypes disappeared. This is not unexpected since hair and eye colors strongly indicate the ancestry of an individual. This makes it difficult to separate the effects of population stratification and the true genetic effects of variants on these particular phenotypes.

Key words: physical characteristics, population stratification, genetic correlation, hair color, eye color

3.1 Introduction

Hair and eye color are two major features in determining an individual’s appearance within a population. Both hair and eye color are highly heritable. Heritability estimates for hair color range from 61 to 100% and for eye color similar high estimates are obtained [34, 36, 122]. Linkage studies indicate quantitative trait loci (QTLs) on chromosomes 15q [37, 123], which contain the well-known pigment genes: *OCA2* and *HERC2*. Both hair and eye color are determined by gene variants present in the melanin pathway including *HERC2*/*OCA2*, *SLC24A4* and *TYR* [20, 87, 124-125], though these genes may not explain all genetic variance in hair and eye color. Here we focus on the question whether all genes affecting one of these visible traits also affect the other trait. To address this question, we estimated the genetic correlation between hair and eye color for common single nucleotide polymorphisms (SNP) within a sample from the Dutch population using GREML (genomic restricted maximum likelihood) estimation for bivariate analyses [126] as implemented in the GCTA (Genome-wide Complex Trait Analysis) software package [6]. The data came from unrelated individuals (N=3,619) registered with the Netherlands Twin Register, which includes participants from all regions of the Netherlands. We explored the effect of including Principal Components (PCs) on the genetic correlation between hair and eye color. The first 3 PCs in the Dutch population significantly correlate with participants' geographic location: PC1 with north-south, PC2 with east-west and PC3 with centre belt region of the Netherlands [103]. Peripheral pigmentation traits such as eye, hair, and skin color show a close correspondence with latitudes in national to world-wide geographic regions: low-pigmentation prevalence is found to be higher in high latitudes [103, 127]. Bolk [128] reported in a paper published in 1908, that early in the 20th century in the Netherlands, the northern part of the country was characterized by more blond hair and lighter eye color than the southern part of the country. As is evident from **Table 1** using data from the Netherlands Twin Register collected around 2004, this is still the case roughly a hundred years later.

Table 1. The distribution (in %) of hair color and eye color across Dutch provinces in the 1908 sample collected by Bolk (N=480,165) and for the 2004 survey in adult NTR participants (N=7,661). We report 3 classes as in the original 1908 paper, i.e. blond hair-blue eyes, black hair-brown eyes and red hair.

Birth	Blond hair + blue eye		Black hair + brown eye		Red hair	
Province	1908	2004	1908	2004	1908	2004
Friesland	43.2	40.0	1.7	1.3	2.5	2.1
Groningen	41.3	37.9	1.4	0.7	2.3	1.1
Drenthe	39.4	37.3	1.3	0.9	2.7	2.7
Flevoland**	-	35.7	-	5.7	-	2.9
North-Holland	31.2	35.6	1.8	1.7	2.5	3.3
Overijssel	35.5	34.2	1.6	1.4	2.2	4.1
South-Holland	31.4	32.7	2.5	1.8	2.4	1.7
Gelderland	34.4	32.1	2.8	1.4	2.5	1.8
Utrecht	20.1	30.8	2.4	0.3	2.5	2.3
Zeeland	28.4	22.0	4.0	0.5	1.8	2.0
North-Brabant	22.3	28.0	4.0	0.5	2.6	2.0
Limburg	21.8	24.2	4.6	0.6	2.2	3.0
Total	32.3	32.2	2.5	1.3	2.5	2.4

*This variable is Blond hair + blue eyes in the 1908 data and Blond hair + blue/ gray eyes in the 2004 data.** The province of Flevoland consists of reclaimed land and did not yet exist in 1908.

3.2 Methods

Participants in the Netherlands Twin Register [74-75] were included in this study based on the presence of self-reported data on natural hair and eye color and the presence of genotype data on an Illumina 370, 660, 1M or Affymetrix Perlegen-5.0, or 6.0 platform. There were 7063 genotyped Dutch-ancestry participants, clustered in 3407 families with data on eye color and 6965 genotyped individuals had data on both hair and eye color. For the genetic association analysis of eye color (see Supplementary Online Materials) all data were analyzed. For bivariate genetic analyses in GCTA all unrelated individuals were selected, based on a Genetic Relatedness Matrix (GRM) cut-off of 0.025 [6]. This left 3,619 individuals for the bivariate analyses with a genetic relatedness equivalent to less than third or fourth cousin.

Age, sex, natural hair and eye color were obtained from Adult NTR survey 7, which was collected in 2004 [75]. Adult participants reported their own natural hair color from one of five options: “fair/blond”, “hazel”, “red/auburn”, “dark brown”, and “black” and eye color with one of three options: "blue/ gray", "green/hazel" and " brown". The same questions on eye color and hair color were answered by adolescent (14-18 years old) twins when they completed the Dutch Health and Behavior Questionnaire in 2005 or 2006 [74].

For the statistical analyses we combined the black, light brown, and dark brown hair colors to “Dark”, as only very few people reported black hair color [122]. Written informed consent was obtained from all participants.

DNA extraction, purification and genotype calling of the samples were performed at various points in time following the manufacturer’s protocols and genotype calling programs [122]. For each platform, the individual single nucleotide polymorphisms (SNP) were remapped on the build 37 (HG19), ALL 1000 Genomes Phase 1 imputation reference dataset [129]. SNPs that failed unique mapping and SNPs with an allele frequency difference over 0.20 with the reference data were removed. Also, SNPs with a Minor Allele Frequency (MAF) <0.01 were removed, as well as SNPs that were out of Hardy–Weinberg Equilibrium (HWE) with $p < 10^{-5}$. The platform data were then merged into a single genotype set and the above SNP QC filters were re-applied. Samples were excluded from the data when their DNA was discordant with their expected sex or IBD status, the genotype missing rate was above 10%, the Plink F-inbreeding value was either larger than 0.10 or smaller than -0.10, or they were an ethnic outlier based on EIGENSTRAT Principal Components (PCs) calculated from the 1000G imputed data [129]. Phasing of the samples and imputing cross-missing platform SNPs was done with MACH 1 [95]. The phased data were then imputed with MINIMAC to the 1000G reference. After imputation, SNPs were filtered based on Mendelian error rate (> 2%), a R^2 imputation quality value of <0.80, MAF <0.01 and a difference of more than 0.15 between the allele frequency and the reference [96]. We tested the effect of different platforms and removed SNPs showing platform effects. This was done by defining individuals on a specific platform as cases and the others as controls. If the allelic association between the specific platform allele frequency and the other platforms allele frequency was significant ($p < 10^{-5}$) SNPs were removed. This left 5,987,253 SNPs, which were all used to construct a GRM.

A GRM based on autosomal SNPs was obtained from GCTA on the best-guess imputed data from Plink 1.07 [94]. The genetic correlation between the various dichotomous hair and eye color combinations (i.e., defined by whether a color was present or absent for the eyes and hair) was estimated using the GCTA bivariate analysis option [126]. Sex and age were used as covariates in the analyses. Next, we added three Dutch PCs calculated from the genetic data [103] as covariates, to explore the effect of ancestry-informative PCs on the analysis, since hair and eye color are likely related to the population diversity captured by these PCs.

3.3 Results

The GWAS for hair color in this sample was published previously [122] and the results from the GWAS for eye color are described in the supplementary online material. We replicated the known genetic variants for eye color including the *HERC2* region for brown eye color (top SNP: rs74940492, OR=0.09, p=5.4E-8) and for blue/ gray eye color (top SNP: rs2240202, OR=13.55, p=1.0E-47); *TYR* and *SLC24A4* for blue/ gray (top SNPs: rs4904871, OR=0.71, p=2.8E-13; rs67279079, OR=0.70, p=3.1E-11) and green/hazel eye color (top SNPs: rs4904871, OR=1.52, p=3.8E-20; rs67279079, OR=1.49, p=3.6E-10). Among these identified pigment genetic variants, we detected that HERC2 has pleiotropic effects on blond, brown, dark hair color and blue/ gray, brown eye color, and SLC24A4 has pleiotropic effects on blond, brown, dark hair color and blue/ gray and green/hazel eye color.

The phenotypic association of eye and hair colors confirms the two traits to be strongly related in our sample (χ2-test with 4 degrees of freedom gave a p-value < 2.2x10-16): people with blond and red hair are likely to have blue/ gray eyes while people with dark hair are more likely to have brown eyes. The counts and frequencies of the hair and eye color phenotypes for the 3619 individuals (1401 males; age: 41.04±19.81 and 2218 females; age: 39.13±17.17) are presented in **Table 2a** while **Table 2b** summarizes the information on the hair-eye color association from the much larger sample collected by Blok [128], confirming the strong association in the Dutch population.

Table 2a. Hair and eye color counts and percentages for unrelated genotyped individuals of the Netherlands Twin Register.

Color	Blond hair	Red hair	Dark hair
Brown eyes	91(2.5%)	10(0.3%)	566(15.6%)
Blue/ gray eyes	1165(32.2%)	58(1.6%)	1024(28.3%)
Green/hazel eyes	224(6.2%)	14(0.4%)	467(12.9%)

Table 2b. Hair and eye color counts and percentages from Blok 1908.

Color	Blond hair	Red hair	Brown hair	Black hair
Brown eyes	37102(7.8%)	1671(0.4%)	31791(6.6%)	11758(2.5%)
Blue eyes	155040(32.4%)	4595(1.0%)	21970(4.6%)	4428(0.9%)
gray eyes	121157(25.3%)	4000(0.8%)	22294(4.7%)	4653(1.0%)
Green/hazel eyes	35517(7.4%)	1493(0.3%)	16763(3.5%)	4682(1.0%)

The genetic correlations between the hair and eye colors are presented in **Table 3**. Here the same relation is shown as in the phenotypic description of the data, where the genes related to blond hair show a strong positive correlation with blue/ gray colored eyes and a negative correlation with brown and green colored eyes. Due to the low prevalence of red hair color in our population, we do not detect any significant genetic overlap with any eye color. Finally, there is a clear and strong genetic overlap for brown eyes and a dark hair color. When adding the first three genetic PCs that correlated with Dutch ancestry, the genetic correlations are reduced to zero (LRT=0, *P-value*=0.5). This indicates that the genetic PCs of the Dutch population are capturing the overlapping genetic variance of eye and hair colors.

Table 3. The genetic correlations (se) between hair and eye color estimated from common SNPs (MAF > 0.01) covering the full genome, corrected for age and sex, within unrelated individuals of the Netherlands Twin Register.

Color	Blond hair	Red hair	Dark hair
Brown eyes	-0.64(0.31)*	-0.18(0.40)	0.71(0.33)**
Blue/ gray eyes	0.87(0.35)**	0.21(0.40)	-0.94(0.37)***
Green/Hazel eyes	-0.61(0.47)	-0.14(0.63)	0.64(0.48)

*One-sided P-value < 0.05, **One-sided P-value < 0.01, ***One-sided P-value < 0.001, likelihood ratio χ^2-test, df = 1, with the correlation fixed at 0.

3.4 Discussion

Based on our analyses of genome-wide SNP data, there is a strong genetic overlap between eye and hair color within the Dutch population. This is in line with findings from previous molecular studies indicating that the same genes are involved in hair and eye color, for example, variants within the melanin producing pathway including *HERC2*, *OCA2*, *SLC24A4* and *TYR* [87, 117, 130]. We also conducted a GWAS for each of the two traits in the NTR population (see Lin et al., 2015 for hair color and supplementary online materials for eye color). The results confirmed the involvement of two genes, *HERC2* and *SLC24A4*, in both hair color and eye color.

It is important to realize when studying eye and hair color that these phenotypes can be highly correlated with the genetic constitution of the population. Although the overall pigmentation prevalence has changed during past 100 years (see **Table 1**: hyper-pigmentation traits are more prevalent in 2004), the distribution pattern of pigment traits following latitude is still the same. PCs representing Dutch ancestry and geographic location are likely to explain the largest part of the variability of human pigment traits. As shown here, the effect of population stratification and the true effects of genes on the two traits are closely linked, as PC1 to PC3 also explained the genetic overlap between the traits. In our study we only selected European Caucasian individuals based on the genetic PC projection and 1000 genomes. Subsequently, three Dutch PCs were calculated in the remaining individuals to account for the population stratification of regions where people live in the Netherlands. However, these PCs also capture multiple traits that likely underwent simultaneous genetic divergence between (sub)populations, such as eye and hair color. When conducting gene finding studies or GCTA analyses, researchers should therefore be aware of the effects of ancestral population differences on the relationship between stratified traits.

Appendix II.

GWAS results for eye color

Results of a genome-wide association study for eye color in the Netherlands Twin Register (NTR)

Participants

Within the NTR, there were 7063 Dutch ancestry participants, clustered in 3407 families (2641 men, age: 44.99±19.15; 4546 women, age: 45.15±16.72) with information on both genotype and eye color. Data on eye color were obtained from surveys. Participants reported their eye color by choosing from one of three answer possibilities: “blue/ gray", "green/hazel" and "brown". Written informed consent was obtained from all participants.

Genotyping data

Buccal or blood DNA samples were collected for multiple NTR projects. DNA extraction and purification of these samples were performed at various points in time [76, 131], following several manufacturer specific protocols to obtain the best quality and concentration prior to SNP platform genotyping. Quality control was done within and between platforms and subsets prior to imputation. For each platform, the individual SNP markers were lifted over to build 37 (HG19) of the Human reference genome, using the LiftOver tool. 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. 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. Quality control details were detailed in Lin et al. [122]. Phasing of all samples and imputing cross-missing platform SNPs was done with MACH 1

[132]. The phased data were then imputed with MINIMAC [96] against the 1000 Genomes Phase 1 Reference panel in batches of around 500 individuals for 561 chromosome chunks obtained by the program CHUNKCHROMOSOME [97]. After imputation, SNPs were filtered based on the Mendelian error rate (> 2%) in families. If the imputed allele frequency differed more than 0.15 from the 1000G reference allele frequency, SNPs were removed.

GWA analysis

We performed 3 case-control GWAS on binary eye color variables: brown versus non-brown eye color, blue/ gray versus non-blue/ gray eye color, green/hazel versus non-green/hazel eye color, with logistic regression, having age, sex, 3 Dutch PC's and corrections for genotype platform as covariates. Analyses were performed with the PLINK 1.07 software running a logistic regression on each SNP, taking genotype inaccuracy into account by using dosage data [94]. Familial structure was taken into account using "--family" option. Filtering on MAF > 0.01, imputation quality R^2> 0.80, and Hardy–Weinberg equilibrium (HWE) p-values >0.0001 were done after the GWA analyses within all eye color informative individuals. This left 5,834,593 SNPs for generating Manhattan and QQ plot.

Results

The eye color prevalence was 62.28% for blue/ gray eyes, 19.58% for green/light eyes and 18.14% for brown eyes. The top SNPs in a LD block for each eye color are shown in **Table S1**. The resulting Q-Q and Manhattan plots for all eye colors are shown in the supplemental **Figures (S1-S3)**.

For brown eye color, we found that the presence of the T allele at rs74940492, an intron variant for *HERC2*, significantly decreases the probability of brown eye color (OR=0.09, p=5.4E-8). This SNPs is in the same LD block with top SNP rs2240202 for blue eye color (OR=13.55, p=1.0E-47). This locus was also found to be associated with blond hair color, brown hair color and dark hair color in our study. *HERC2* which harbors this SNP has been identified as an eye iris color gene by multiple studies [123, 133].

The top SNP rs4904871, an intron genetic variant in *SLC24A4*, was significantly associated with both blue eye color (OR=0.71, p=2.8E-13) and green eye color (OR=1.52, p=3.8E-20). This SNP was has also been associated with hair color in GWA studies [122]. The rs12896399, which is located in the same LD block of rs4904871 (LD r^2=0.95 distance~22kb) was found to be associated with blond versus brown hair color, blue versus green eye color [87] and black versus blond hair color [117] in other GWA studies.

The T allele of rs67279079 at *TYR* has been found to significantly decrease the probability of blue eye color (OR=0.70, p=3.1E-11), and simultaneously to increase the probability of green/hazel eye color (OR=1.49, p=3.6E-10). The *TYR* gene codes tyrosinase located in melanocyte, which is responsible for the first step in melanin production. This gene is associated oculocutaneous albinism and skin tanning ability
[125, 130, 134]. Three known pigment genetic loci were thus confirmed, but no new genetic variants for eye color were identified.

Conclusion

In this study, we have replicated genetic variants for eye color: *HERC2* for brown eye color and blue/ gray eye color; *TYR* and *SLC24A4* for blue/ gray and green/hazel eye color.

Table S1. SNP associations for 3 eye colors

Locus	Chr location	Top SNPs	MAF	Brown eye color			Blue eye color			Green eye color		
				OR	SE	p	OR	SE	p	OR	SE	p
HERC2	15q13	rs2240202	0.036	0.09	0.123	5.4E-83	14.53	0.189	2.5E-45	0.89	0.128	0.343
HERC2	15q13	rs74940492	0.037	0.10	0.123	2.2E-81	13.55	0.179	1.0E-47	0.89	0.128	0.3655
SLC24A4	14q32.12	rs4904871	0.478	1.09	0.056	0.1218	0.71	0.046	1.3E-13	1.52	0.053	1.5E-15
TYR	11q14.3	rs67279079	0.236	1.17	0.066	0.0209	0.70	0.054	3.1E-11	1.49	0.064	3.6E-10

Figure S1. Manhattan and QQ plot for brown eye color (MAF > 0.01, λ =1.03114).

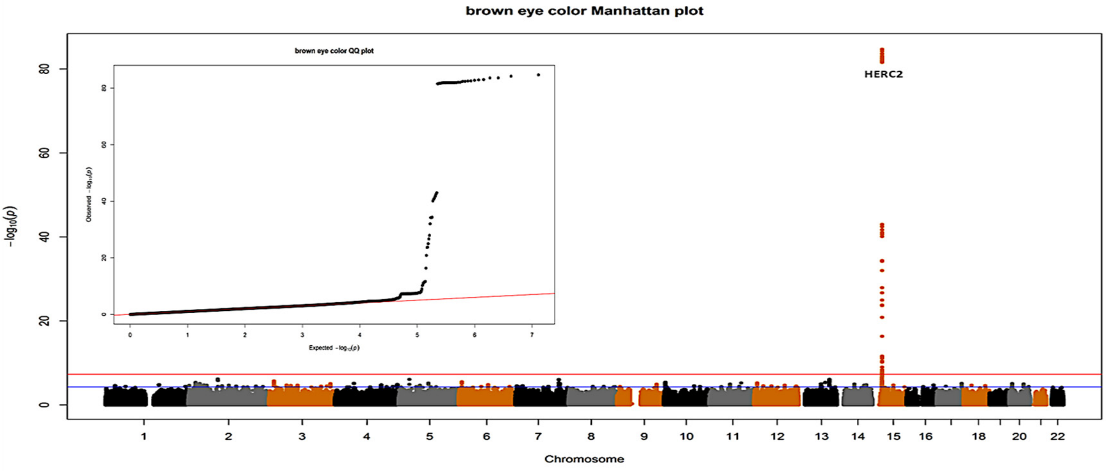


Figure S2. Manhattan and QQ plot for blue/gray eye color (MAF > 0.01, λ =1.0259).

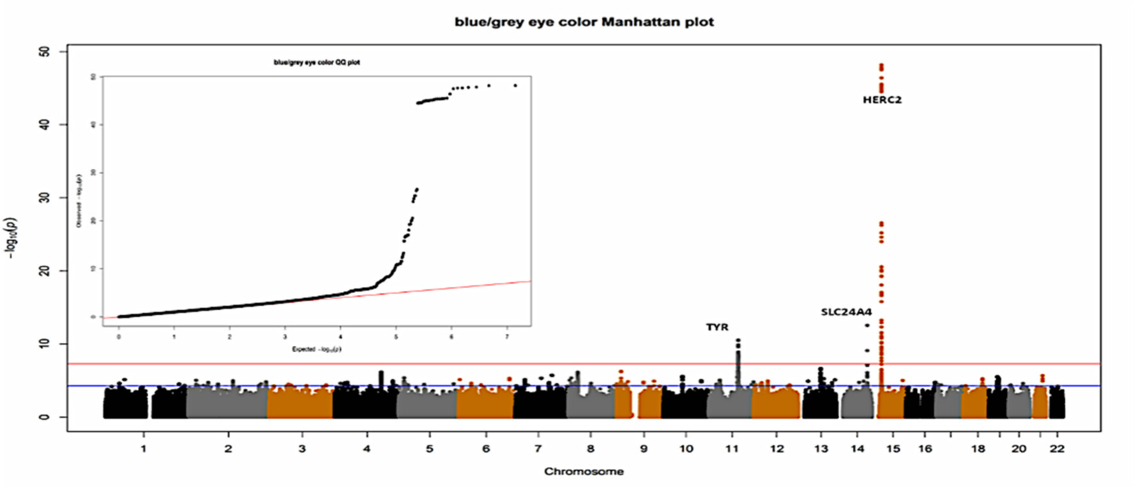
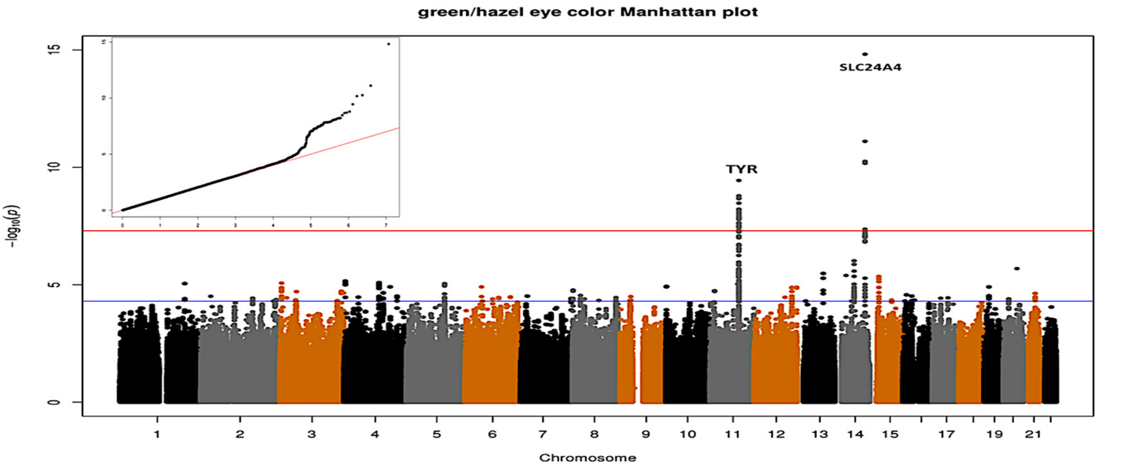


Figure S3. Manhattan and QQ plot for green/hazel eye color (MAF > 0.01, λ = 1.02163).



Part II: Hematological profiles

Chapter 4

Causes of variation in the neutrophil-lymphocyte and platelet-lymphocyte ratios: a twin-family study

This chapter is published as: Lin BD, Hottenga JJ, Abdellaoui A, Dolan CV, de Geus EJ, Kluft C, Boomsma DI, Willemsen G., Causes of variation in the neutrophil–lymphocyte and platelet–lymphocyte ratios: a twin-family study. *Biomark Med*, 2016. **10**(10):p. 1061-1072.

Abstract

Aim: Neutrophil–lymphocyte ratio (NLR) and platelet–lymphocyte ratio (PLR) are biomarkers for disease development, for whom little is known about causes of variation in the general population. Materials & methods: We estimated the heritability of PLR and NLR and examined their association with gender, demographic, lifestyle and environmental factors in a Dutch nonpatient twin family population (n = 8108). Results: Heritability was estimated at 64% for PLR and 36% for NLR. Men had on average higher NLR, but lower PLR levels than women. PLR and NLR increased significantly with age, decreased in colder months and showed small but significant sex- and age-specific associations with body composition and smoking. **Conclusion**: NLR and PLR levels are heritable and influenced by age, sex and environmental factors, such as seasonal conditions and lifestyle.

Keywords: heritability, PLR, NLR, BMI, smoking, age, sex differences, weather conditions

4.1 Introduction

Hematological biomarkers in peripheral blood are indicators of physiological function and their levels may direct clinical decisions regarding disease status and treatment of patients. The two largest sets of immune cells, as reported in clinical hematological profiles, are neutrophils and lymphocytes. While both cell types play a key role in human inflammation and disease response, recent clinical studies suggest that their ratio may serve as a useful biomarker of disease. The neutrophil to lymphocyte ratio (NLR) has prognostic value for cancer progression [135-136], inflammatory disease [66-67], and cardiovascular disease [137]. A second hematological ratio of interest is the platelet to lymphocyte ratio (PLR), which has also been related to cancer progression [70], cardiovascular disease and inflammation [138].

To understand the role of NLR and PLR in disease processes, it is important to gain insight into the degree of variation in these ratios within non-patient populations and the extent to which variation is due to genetic and non-genetic causes. Normal variation in immune function may be due to inherent factors such as age, sex, and genetic constitution, environmental factors such as season, and lifestyle factors such as smoking and diet. To date, few studies examined the factors influencing the variation of NLR and PLR in non-patient populations and most of those focused on NLR. Sex and age effects on NLR in the general population were examined in two studies [139-140], with similar results. No evidence was seen for sex differences in NLR but NLR did increase with increasing age. Li et al. [139] suggested that this age-related increase may reflect a higher prevalence of, often undetected, chronic infectious disease and cancer development in the older population. Genetic epidemiological studies of NLR and PLR are, to the best of our knowledge, lacking. However, genetic factors have been shown to contribute substantially to the phenotypic variation in neutrophil, lymphocyte, and platelet counts, with heritability estimates of 67%, 48-71%, and 57-86% respectively [51-53, 141]. In addition to the contribution of inherent factors to variation among individuals, immune function may also be influenced by external factors. Seasonality is thought to be an important source of variation in the hematological profile [142]. Lymphocyte subset counts as well as platelet levels have been found to be lower in the summer season [143-144] and month-to-month changes in leukocyte and platelet levels were observed in a study of trained and untrained men [145]. Buckley et al. [144] estimated that seasonal factors accounted for 2% of the overall variation in platelet count, but not all studies show evidence for seasonal effects on platelet count [146]. Lifestyle may also contribute to variation in immunological function. Positive associations between body mass index (BMI) and NLR were observed in two non-patient populations [139-140], but a third study did not find BMI to be related to NLR nor to PLR [147] . With respect to their subcomponents, larger waist circumference has been related to higher levels of lymphocytes, neutrophils and platelets [148] and these cell counts were also increased in obese women compared to non-obese women [149-150]. Smoking has also been related to increased NLR in two studies in the general population [140, 151] and to increased neutrophil [151] and lymphocyte counts [152]. PLR, however, was not related to smoking [151] in the general population, and neither was platelet count in this study. Lack of evidence for an association between platelet count and smoking has been reported more often [153-154], though lower platelet levels in smokers have also been observed [155-156].

The present study analyzed data collected in over 8,000 adult participants from the Netherlands Twin Register, including adult twins and their family members, who were very well characterized with respect to demographic and lifestyle traits and for whom information on date and time of blood sampling was available. Also, the study collected blood samples in women at a fixed moment of the menstrual cycle. We have two aims: Firstly, to estimate the contribution of the genome (heritability) and of non-genetic factors to variation in NLR and PLR and their subcomponents. Secondly, to further study non-genetic factors by examining the associations of the two ratios with age, sex, weather conditions at the day of sampling, indicators of inflammation, i.e. C-reactive protein (CRP) and Interleukin 6 (IL6) levels, and the influence of smoking behavior and BMI, although it should be recognized that some of these traits, e.g. BMI or smoking, are themselves influenced by genes.

4.2 Materials and Methods

4.2.1 Participants

Data for the present study came from participants in the Netherlands Twin Register (NTR) Biobank projects, which took place between 2004 and 2008, and in 2011 [75-77]. After excluding outliers (i.e. absolute values exceeding mean ±5×SD), NLR and PLR data were available for 9,434 participants, clustered in 3,411 families. In a next step, data were excluded in case of: 1) illness in the week prior to blood sampling (N=539); 2) CRP ≥ 15 (N=287); 3) basophil count > 0.02×10^9/L (N=151); 4) blood related disease or cancer (N=83); and 5) use of anti-inflammatory medication (N=437), glucocorticoids (N=143) or iron supplements (N=28). This resulted in data for 8,108 participants from 3,411 families. The study protocol was approved by the Medical Ethics Committee of the VU University Medical Center Amsterdam, (the Netherlands), and all participants provided informed consent.

4.2.2 General biobank procedure

Participants were visited at home, or in some cases at work, between 7 a.m. and 10:00 a.m. They were instructed to fast overnight and to refrain from smoking, heavy physical exertion and from medication use (if possible) in the morning prior to the visit. Fertile women without hormonal birth control were, if possible, seen on the 2nd to the 4th day of the menstrual cycle and women taking hormonal birth control were visited in their pill-free week. During the home visit, a brief interview was conducted concerning general health status, any chronic diseases, medication use and smoking history. Measures of height, weight, waist circumference and hip circumference were obtained. Peripheral venous blood samples were drawn by safety-lock butterfly needles in EDTA, lithium and sodium heparin, CTAD and PAX tubes. Immediately after blood collection, tubes were inverted several times to prevent clotting and subjected to initial processing in a mobile laboratory. Within 3 to 6 hours after the blood draw all samples were transported to the laboratory facility in Leiden, the Netherlands (for details see [76-77]).

4.2.3 Blood parameters

Hematological profile. The 2 ml EDTA tubes were transported at room temperature to the laboratory, where the hematological profile was obtained using the Coulter system (Coulter Corporation, Miami, USA). The profile consisted of total white blood cell count, percentages and numbers of neutrophils, lymphocytes, monocytes, eosinophils and basophils, red blood cell count, hemoglobin, hematocrit, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, red cell distribution width, platelet count and mean platelet volume.

*NLR and PL*R levels. NLR was calculated as absolute neutrophil count (109/ L) divided by absolute lymphocyte count (109/ L), and PLR was calculated as absolute platelet count (109/ L) divided by absolute lymphocyte count (109/ L).

CRP level. Plasma heparin was collected from a 9 ml heparin blood tube that was transported in melting ice to the laboratory. The plasma subsamples were snap-frozen and stored at -30^{o}C. One heparin plasma subsample was used to determine C-reactive protein (CRP) by the 1000 CRP assay (Diagnostic Product Corporation) [157].

IL6. EDTA plasma was obtained from the 9ml EDTA tubes, which were stored in melting ice during transport. Upon arrival at the laboratory, the tubes were centrifuged for 20 min at 2000×g at 4 °C, and the plasma subsamples were snap-frozen and stored at -30^{o}C. IL-6 level was subsequently measured in one EDTA plasma subsample using the Quantikine Elisa Human IL-6 sR assay of R&D systems. Data were made missing if they exceeded mean ± 5×SD (1.02% in total sample size) [158].

4.2.4 Health status, seasonal effects, BMI and smoking behavior

Health status. Participants were asked to report any chronic diseases and when they were last ill (i.e., less than 1 week ago, less than 1 month ago, or more than 1 month ago). For any medication use, the dosage, brand and name were recorded.

BMI. BMI was calculated as weight (kg) divided by height squared (m^2).

Smoking behavior. Participants indicated whether they currently smoked or ever had smoked. If so, they were asked to provide information on the number of cigarettes smoked and how long they (had) smoked. Based on this information, participants were divided into 5 categories: nonsmoker, ex-smoker, light smoker (currently smoking less than 10 cigarettes a day), average smoker (currently smoking 10 to 19 cigarettes a day), and heavy smoker (currently smoking 20 or more cigarettes a day).

Seasonal effects. The information on daily weather conditions was obtained from the website of the Royal Netherlands Meteorological Institute (KNMI). We analyzed the daily data on temperature, wind speed, mean sea level, sunshine duration, global radiation and mean relative atmospheric humidity and potential evapotranspiration [159].

4.2.5 Analyses

For NLR and PLR, the contribution of genetic factors (heritability) was estimated based on the resemblance between relatives including mono- and dizygotic twins. First, we summarized familial resemblance with respect to NLR and PLR, corrected for age, sex, and age x sex effects, by means of correlations. Next, genetic and non-genetic variance components were estimated by raw-data maximum likelihood in OpenMx [98]. The total variance in each phenotype was decomposed into four sources of variation: additive genetic (A), non-additive genetic or dominance (D), common environmental (C) and unique environmental (E) variation. Common environmental variance was considered as the variance shared between siblings and twins (V_s) who grow up in the same family. The resemblance among family members was modeled as a function of A, D and C, making use of well-established genetic relatedness among family members. As monozygotic (MZ) twins derive from a single fertilized egg (zygote), they share ~100% of their genetic material, and consequently share all genetic (additive and dominance) variance. Dizygotic (DZ) twins, like full siblings, derive from two zygotes and share on average 50% of their segregating genes. Consequently, they share 50% of additive (V_A) and 25% of dominance genetic variance (V_D) [13]. Parents and offspring share exactly half of their genetic material, and share 50% of V_A, but no V_D. Our model allowed siblings and twins to share variance attributable to shared environment (V_S). Unshared influences (environmental, measurement error and personal mutations) (V_E) contribute to total variance, but not to familial resemblance. We allowed for a correlation in phenotype between spouses (μ). Genetic analyses were done in twin families with at most one twin pair per family, and two brothers and two sisters, father and mother. The sample of 3,251 families included 7,481 participants (238 MZM, 99 DZM, 530 MZF, 215 DZF and 221 DOS complete twin pairs). Nested sub-models were compared to the full model by log-likelihood ratio test (-2LL), at a significance level of 0.05.

The association of NLR and PLR with IL6 and CRP was quantified by Pearson correlations in sex and age-corrected data. To test the effect of sex, we performed a T-test with sex as the independent factor on age-corrected NLR and PLR. The effects of age, temperature, smoking and BMI on NLR and PLR were tested by linear regression in STATA [160], separately for men and women. All analyses were corrected for familial clustering using the option of robust cluster. All beta values presented below represent raw values and are evaluated at a significance level of 0.05.

4.3 Results

We carried out a series of analyses of the twin family data to gain insight into the heritability of PLR and NLR and their association with demographic factors, indicators of inflammation, seasonal conditions and lifestyle. **Table 1** provides the descriptive statistics for NLR and PLR, their subcomponents neutrophil, lymphocyte, and platelet count, and CRP and IL6 levels, separately for men and women. **Table 2** contains the familial correlations for NLR and PLR. We found that the NLR and PLR familial correlations did not depend on sex (i.e., correlations in MZ males and MZ female twin pairs were equal, as were the correlations for male and female first-degree relatives; p = 0.23). For NLR, the MZ correlation was 0.36 (CI is 0.30-.42) and the DZ correlation was 0.19 (.16-.22), which indicates an additive genetic model. For PLR, the MZ correlation was 0.64 (0.60-0.68), but the DZ correlation was less than half the MZ correlation. i.e. 0.24 (0.21-0.27), suggesting the presence of non-additive genetic effects. Spousal correlations were significant at 0.14 (0.07-0.21) for NLR and 0.17 (0.10-0.23) for PLR. The most parsimonious genetic models showed no evidence for common environmental influences on NLR (p = 0.47) and PLR (p = 0.99). The narrow sense heritability (proportion of total variance explained by additive genetic factors) of NLR was estimated at 35.8%, with no evidence for non-additive effects. For PLR, the narrow sense heritability was 38.3%, with non-additive effects accounting for an additional 25.9% of the total variance. The broad sense heritability for PLR was thus 64.2%. The remainder of the variance (64.2% in NLR and 35.8% in PLR) was explained by environmental factors. We also estimated the heritability for the three subcomponents of the ratios. The broad-sense heritability for neutrophil count was estimated at 41.1% (no non-additive effects), for lymphocyte at 57.6% (22.4% due to non-additive effects), and for platelet numbers at 70.5% (with 21.9% due to non-additive effects). There was no evidence for common environmental effects for neutrophil count (p = 0.87), platelet count (p = 0.32) and lymphocyte count (p = 0.99).

For age- and sex-corrected values, the correlation between NLR and PLR was 0.49 (p < 0.001). We further determined the correlation of the two ratios with two established markers of inflammation, namely CRP and IL6. PLR correlated neither with CRP nor with IL6 (p > 0.05), but NLR correlated significantly with CRP (0.15, p<0.001) and with IL6 (.08, p < 0.001). NLR and PLR levels were affected by both sex and age. For age-corrected values, men had higher mean NLR levels than women (men: Mean $_{NLR}$=1.667, SE$_{NLR}$ =0.012; women Mean $_{NLR}$=1.626, SE$_{NLR}$ =0.010; t(8106)=2.2602, P=0.009) and lower PLR levels than women (men: Mean $_{PLR}$=116.944, SE$_{PLR}$ =0.753; women: Mean $_{PLR}$=125.156, SD$_{PLR}$ =0.587; t(8106)=20.073, P<0.001). NLR increased with age in men but not in women (**Table 3**, model 1), while PLR increased with age in both men and women (see **Table 3**, model 1).

Table 1. Mean (SD) levels of NLR and PLR and their constituents for men and women in the twin family population.

	Men	Women
N	3068	5040
Age	44.13 (15.89)	43.07 (14.53)
NLR	1.67 (.66)	1.62 (.70)
PLR	117.11 (40.27)	125.05 (42.81)
Neutrophil	3.41 (1.16)	3.45 (1.28)
Lymphocyte	2.17 (.634)	2.27 (.71)
Platelet	236.48 (53.36)	263.13(6.84)
CRP*	2.01 (2.36)	2.66 (2.92)
IL6*	1.69 (3.07)	1.637 (3.80)
BMI	25.47(3.67)	24.84(4.37)
Current Smoker y/n (%)	14.1	11.3

* The sample size for CPR: N=3045 for men, N=4980 for women; IL6: N=2929 for men, N=4867 for women.

Table 2. Familial correlations and confidence intervals for NLR and PLR.

Pairs	NLR		PLR	
	R	95% CI	R	95% CI
MZ twins	***0.361***	***0.296-0.420***	***0.644***	***0.603-0.680***
MZ male	0.396	0.277-0.496	0.607	0.518-0.675
MZ female	0.348	0.270-0.418	0.658	0.610-0.699
Male first-degree relatives	***0.186***	***0.111-0.258***	***0.223***	***0.142-0.299***
DZ male	0.160	-0.078-0.392	0.295	0.085-0.461
Brother-male twin	0.331	0.199-0.439	0.342	0.028-0.557
Brother-brother	0.036	-0.225-0.309	0.308	0.122-0.450
Father-son	0.132	0.032-0.226	0.191	0.092-0.282
Female first- degree relatives	***0.172***	***0.127-0.216***	***0.240***	***0.199-0.279***
DZ female	0.293	0.152-0.405	0.355	0.228-0.462
Sister-female twin	0.205	0.101-0.296	0.337	0.201-0.447
Sister-sister	0.179	0.083-0.266	0.241	0.150-0.327
Mother-daughter	0.141	0.079-0.198	0.221	0.165-0.275
Female-male first degree relatives	***0.205***	***0.165-0.244***	***0.240***	***0.197-0.282***
DZ opposite sex	0.172	0.037-0.297	0.257	0.129-0.371
Brother-female twin	0.180	0.049-0.296	0.211	0.165-0.275
Sister-male twin	0.183	0.061-0.293	0.342	0.028-0.557
Sister-brother	0.127	0.006-0.240	0.217	0.102-0.322
Mother-son	0.237	0.149-0.317	0.261	0.173-0.340
Father-daughter	0.235	0.175-0.296	0.233	0.172-0.291
Parents (father-mother)	***0.137***	***0.066-0.207***	***0.166***	***0.101-0.230***
Heritability	***0.358***	***0.304-0.421***	***0.642***	***0.598-0.683***

Correlations in bold italic were obtained from sub-models, in which all matching correlations of the tested subgroup of family relations were set to be equal.

Next, we explored the influence of seasonal conditions on variation in the ratios. **Figure 1** illustrates the association between daily temperature and age-corrected NLR and PLR for men and women. To avoid outliers due to periods with very few observations, we restricted the entries in the graph to the months with more than 75 data points between August 2004 and December 2007. We note a similar pattern for NLR and PLR from year to year: Overall, NLR and PLR ratios increase with decreasing temperature. This pattern seems more evident in the female group than in the male group. To formally test for the effect of temperature, we included this variable in a regression analysis conducted separately by sex and taking age into account. The results, shown in **Table 3** (model 2) demonstrate that both NLR and PLR are negatively significantly associated with daily temperature in women, but not in men. There was no evidence for significant age x temperature interactions for NLR and PLR.

We also explored the associations of NLR and PLR levels with the other weather-related information available. Although sunshine duration, global radiation, atmospheric humidity and evapotranspiration were related to NLR and PLR, these associations were rendered insignificant by the addition of temperature. One exception was the effect of global radiation on NLR: as the daily global radiation level increased, NLR levels decreased ($\beta = 2.01E-5$, $p < 0.001$).

Table 3. Results of the linear regression modeling for NLR and PLR, separate for men and women.

Variables	Men						Women					
	NLR			PLR			NLR			PLR		
	Model1	Model2	Model3	Model1	Model2	Model3	Model1	Model2	Model3	Model1	Model2	Model3
Age	.0105***	.0092***	.0009	.2824***	.2723*	1.3813***	-.0000	-.0008	.0083	.2455***	.1062	1.4980***
Temperature		-.0000	-.0002		-.0481	-.0352		-.0123*	-.0014*		-.1083***	-.0991**
Age*temp		.0000	.0000		.0000	.000		.000	.0000		.0012	.0000
BMI			-.0023			.496			.0362***			2.1353***
Age*BMI			.0000			-.0350**			-.0004*			-.0524***
Smoke			.0501			-3.7760*			-.0091			-5.5355***
Age*smoke			.0000			-.0640			.0000			-.0471
N	3068	3068	3040	3068	3068	3040	5040	5040	4980	5040	5040	4980
R2	.0580	.0582	.0608	.0124	.0168	.0611	.0001	.0048	.0162	.0069	.0137	.0571

* P<0.05, **P<0.01, ***P<0.001.

Figure 1. The relationship between monthly temperature (grey dotted line) and the average NLR and PLR for men (blue line) and women (red line).

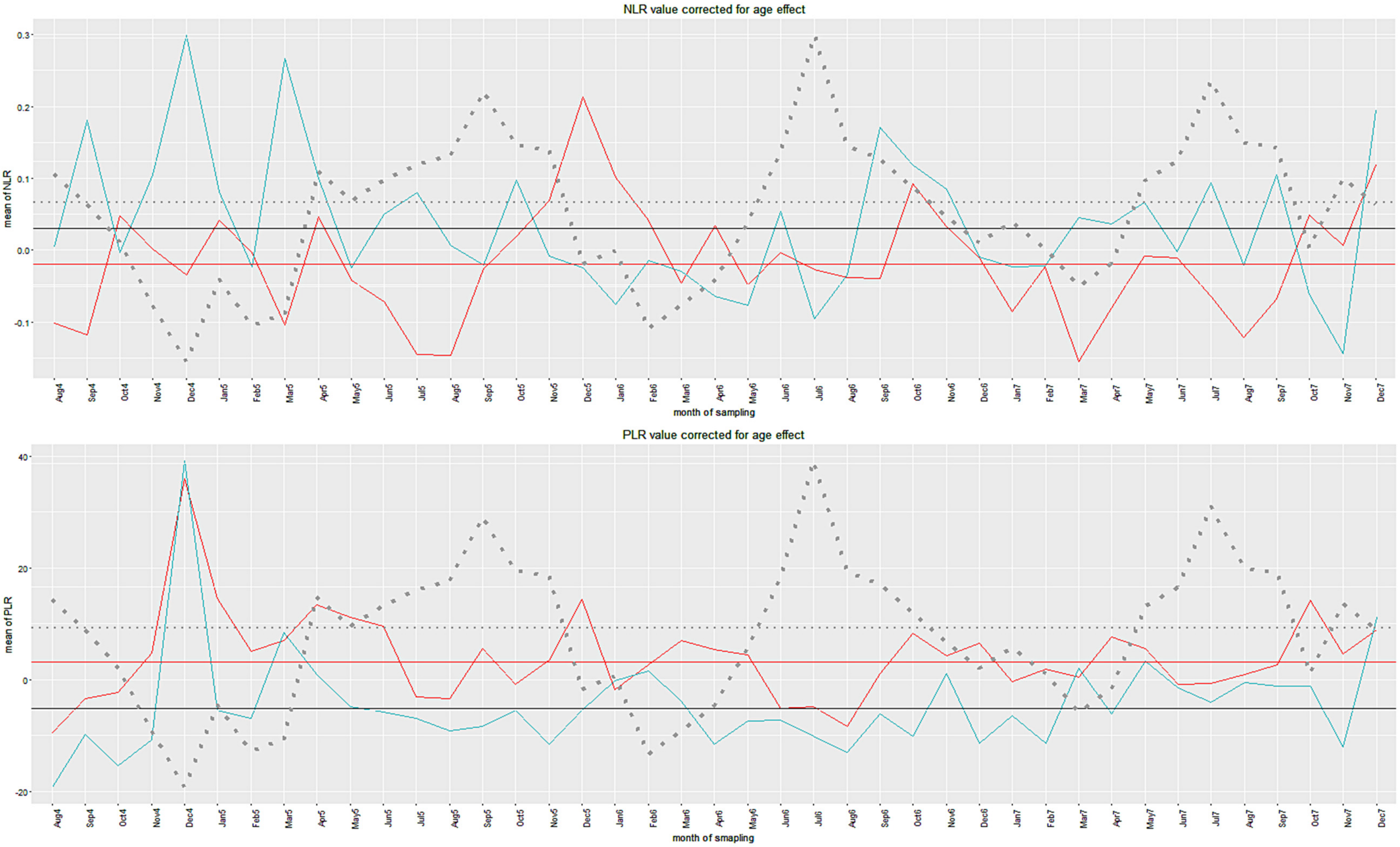


Table 4 includes the average NLR and PLR values as a function of smoking, BMI and sex, while **Table 3** includes the results of the linear regression modeling (see model 3 in **Table 3**). Smoking was not significantly associated with NLR in either men or women. BMI was not associated with NLR in men, but it was related to NLR in women. In women, NLR increased with increasing BMI and there was a significant age x BMI interaction, due to an alleviation of the BMI association with increased age. PLR was more strongly affected by smoking and BMI. In women, there was a significant BMI main effect as well as an age x BMI interaction: the positive association was reduced at older age. Though we had limited numbers of participants at older ages, an exploration of the data seems to suggest the direction of event may be even reversed at old age. A similar pattern, though less strong, was seen for the men. Unexpectedly, smoking was associated with a decrease in PLR in both men and women, while age x smoking interaction effects were not present. To explore the mechanisms underlying the association with smoking, we also examined the relation between the subcomponents and smoking. Smoking was related to an increase in neutrophils ($\beta = 0.305$, $p < 0.001$) and lymphocytes ($\beta = 0.260$, $p < 0.001$), but had no significant effect on platelets ($\beta = 0.017$,$p = 0.133$). There was no evidence for smoking x age interactions for the subcomponents.

The full model (**Table 3**, model 3) including age, temperature, BMI, smoking and their interactions with age, explained about 6% of the variance in PLR in both men and women. In men, this model also explained around 6% of the variance for NLR, but in women only 1.6% of the variance in NLR was explained by the factors included in the model.

Table 4. Age-corrected mean (SE) NLR and PLR as a function of BMI and smoking category, for men and women separately.

	Men			Women		
	N	NLR	PLR	N	NLR	PLR
Underweight	63	1.764(0.082)	126.569(5.097)	188	1.555(0.051)	125.943(4.120)
Normal	1404	1.660(0.017)	119.264(1.089)	2791	1.569(0.013)	125.418(0.817)
Overweight	1372	1.676(0.018)	115.797(1.100)	1594	1.689(0.018)	124.749(1.085)
Obesity	216	1.667(0.044)	108.952(2.735)	436	1.763(0.034)	123.169(2.058)
Never-smoker	1268	1.655(0.019)	121.610 (1.147)	2669	1.630(0.014)	129.469(0.813)
Ex-smoker	1021	1.626(0.021)	12.297(1.310)	1377	1.576(0.019)	127.758(1.143)
Light smoker	330	1.737(0.035)	114.615 (2.160)	392	1.656(0.035)	118.092(2.111)
Average smoker	239	1.743(0.041)	97.758(2.543)	378	1.668(0.036)	104.433(2.149)
Heavy smoker	193	1.785(0.046)	97.890(2.825)	193	1.724(0.051)	98.335(3.007)

For the purpose of the overview, BMI was classified in 4 categories: underweight (BMI < 18.5), normal (18.5-24.9), overweight (25.0-29.9) and obese (BMI>=30).

4.4 Discussion

The current study examined causes of variation in NLR and PLR to provide an insight into individual differences in these biomarkers in the non-patient population. We examined the effects of genetics, demographics, seasonal conditions and lifestyle and described for the first time the importance of genetic factors for variation in PLR and NLR. Especially PLR is influenced to a large extent (64%) by additive and non-additive genetic influences. This high heritability is in accordance with the heritability estimates reports for the individual platelets and lymphocytes components, which ranged from 48% to 86% in previous studies [51-53, 141] and which we here estimated to be 71% for platelets and 58% for lymphocytes. Genetic factors also explain the variation in NLR with heritability estimated at 36%. A lower heritability was also observed for neutrophil count (38%). The genetic architecture underlying NLR and PLR was similar in men and women. Also, there were no differences between the generations in the genetic architecture of NLR and PLR as indicated by similar correlations for parents and offspring as for siblings. Our data showed significant spousal associations, which is in line with previous reports of assortative mating for immune parameters [161] but may also reflect a shared spousal environment leading to a similar immune response.

In addition to being significantly influenced by our genome, normal variation in NLR and PLR levels is also explained by differences in gender, age, and environmental and lifestyle traits. There were sex differences in mean levels, with higher NLR and lower PLR in men compared to women, and an older age was related to an increase in PLR and, to a lesser extent, to an increase in NLR. As suggested by Li at el. [139], the age effect could reflect underlying diseases in the older population, even though we selected relatively healthy individuals as determined by immunological data, medication use and disease reports. It is possible that some diseases were present at sub-threshold level and the higher prevalence of autoimmune disease in women, especially after age 50, is well established [162].

Non-genetic causes of variance in NLR and PLR included the effect of weather conditions. For both ratios, average levels were higher in colder months, indicating seasonal influences on immune parameters. This is consistent with previous work, which found higher levels of inflammation during the winter in European countries [163] and in line with previous studies on seasonal effects on cell counts in humans [142-145]. One likely explanation for the higher levels during the winter is that there is a higher prevalence of viral infections during the cold season [146], though many other factors are likely involved in the seasonal effects on immune response and further study is needed [163]. Women may be more sensitive to the seasonal changes, as an effect of temperature was mainly visible in women.

Lifestyle factors were related to individual differences in the two ratios. In women, a higher BMI was related to higher NLR and PLR levels and there was also evidence for significant BMI x age interactions for both ratios. In men, NLR was not influenced by BMI but there was a significant age x BMI interaction for PLR. The interactions with age were due to the fact that the influence of BMI became less strong at an older age. Our data suggested there may even be a reduction in the ratios at old age, and studies including more participants in the old-age range are needed to confirm this. A positive association between BMI and NLR has been found before [139-140] and obesity is often considered to be associated with a chronic state of inflammation [164]. Dietary habits may also influence both platelet and leukocyte counts [165]. The greater influence of BMI at younger ages points to the importance of weight control in early life.

Smoking increased both neutrophil and lymphocyte count, but whereas we observed decreased PLR levels in men and women who smoked, there was no effect of smoking on NLR. Tulgar et al. [151] found no effect of smoking on PLR or its subcomponents, but this may be due to a small sample size, as the descriptive data do suggest a lower average PLR in smokers compared to non-smokers. This study also reported NLR to be increased in smokers, as did a larger study [140]. The mechanisms underlying the association of NLR and PLR with disease are not fully understood. Increases in NLR and PLR may be indicative of a decreased ability to detect and destroy infected cells, and of increased tumor-promoting activities. A higher NLR indicates a shift in the balance between neutrophil and lymphocytes, which in our sample was due to both a decrease in lymphocyte count and an increase in neutrophil count. Lower lymphocyte counts are associated with poorer survival in different types of cancer [166-167], while high lymphocyte counts are related to better responses to cytotoxic treatment and to better prognosis in cancer patients [168]. Neutrophils have also been reported to secrete tumor growth promoting factors, including vascular endothelial growth factor, hepatocyte growth factor, multiple interleukins and matrix metalloproteinases, and may thus contribute to a tumor stimulating microenvironment [169-170]. A high BMI seems to be related to an increased imbalance between lymphocytes and neutrophils, resulting in an increased NLR, especially in women. With respect to PLR, overall, higher PLR in our study was related to lower lymphocyte and higher platelet numbers. Platelets play an important role in angiogenesis, thrombosis and hemostasis and increased platelet numbers have been implicated in the development of cardiovascular disease [171] and cancer progression [172]. Further study of the relationship of the two ratios with smoking and BMI in a longitudinal sample, with attention to sex differences and interactions, may provide important information about the way lifestyle influences our health.

Several studies have suggested that NLR and PLR may also be used as indicators of inflammation and provide a cheap and easily-obtainable alternative to the currently used CRP and cytokines, such as IL6 [173]. However, the low correlations we observed between the ratios and these two inflammatory markers argue against this. Correlations may have been low because of exclusion criteria in our study, which included high CRP levels. Upon exploring the correlations in the total sample, the correlations for NLR with CRP and IL6 were not much higher (0.214 and 0.121 respectively) while PLR remained unrelated to CRP and IL6. Our results agree with those of Oh et al. [174], in that NLR and PLR are no replacements for CRP and IL6 but should be used in addition to each other.

The correlation between NLR and PLR in our healthy population was moderate (r=0.49). The presented differences in heritability, in the effects of lifestyle and in the association with IL6 and CRP indicate that mechanism underlying individual differences in the two rations are not the same for NLR and PLR. This is in line with studies showing that the two ratios do not predict disease progress to the same extent [175] and may act as independent disease predictors [176].

The combination of demographic and seasonal factors, smoking and BMI explained around 6% of the variation in NLR and PLR. This is substantially smaller than the part of the variance explained by genetic factors; 36% in NLR and 64% in PLR. Thus, it is of importance to realize that variation in NLR and PLR to a large extent can represent genetic variation, and that high levels in these ratios also may occur independent of disease status. While a further search for additional environmental factors influencing variation in these immune parameters is warranted, more insight into the genes and genetic mechanisms underlying the high heritability is needed and gene finding studies form an important next step in characterizing the DNA polymorphisms causing variation in NLR and PLR.

In conclusion, variation in basal NLR and PLR in a general population sample is influenced by the genome, by age and sex, by lifestyle factors and by environmental factors, such as seasonal weather conditions.

4.5 Conclusion

This first study on the heritability of NLR and PLR showed that genetic factors influence variation in NLR, and to an even larger extent, in PLR. To provide more insight into the genetic variation in NLR and PLR gene finding studies are needed. Non-genetic factors are more relevant to NLR than to PLR and while sex, age, seasonal conditions and lifestyle play a role, these factors explain only a small part of the variation. For NLR in particular, studies are warranted to identify additional environmental influences.

Appendix III.

Age and BMI interaction effects on NLR and PLR

In regression **Table 3** model 3 of chapter 4, we found age × BMI interaction effects on PLR in women and men, and on NLR in women. To further investigate these interaction effects, we studied BMI effects on NLR and PLR 5 age groups (see **Table S1**). The age range between 18 to 84 in our NTR data. We categorize the age into 5 levels: 18-30 years old, 30-40 years old, 40-50 years old,50-60 years old, and older than 60 years old.

The linear regression model shows that (**Table S2**): BMI increases NLR level in all women except women who are more than 60 years old. BMI decrease PLR level in women and men older than 50, and men between 30-40 years. The results show that BMI have constant direction effects on NLR or PLR level. For NLR level in women, BMI have positive effects effect on NLR, and this effect was alleviated with increased age. For PLR, BMI does not have effect in younger male and female. However, BMI has negative effect in both old males and females, and this effect was enhanced with increased ages.

Then I did the ANOVA test of NLR and PLR level on 5 age groups and 4 BMI groups (4 categories same as the paper) as fixed factors (**Table S2**).

The post hoc test for age and BMI results shows that: For NLR, there are significant standard deviation differences in almost all different age groups; and there are significant standard deviation differences between normal and overweight group and between normal and obese group in men, between underweight and obese group, normal and overweight group and between normal and obese group in women. For PLR, there are significant standard deviation differences between 18-30 and other age groups in men, 18-30 and other age groups and 40-50 and other groups in women; and there are marginal standard deviation differences between normal and obese group and between overweight and obese group in men, there are no significant standard deviation differences among BMI groups in women. We further check the standard deviations in different age groups (**TableS3**): although older groups have smaller sample size, they have larger deviation for NLR and PLR.

In conclusion, BMI age interactions are exist in female NLR level and both male and female PLR level. Higher BMI increase NLR level in female, but this effect was alleviated with increased age. BMI do not have effects in younger male and female. However, BMI have negative effects in both old male and female, and this effect was enhanced with increased age. In addition, older people have larger variance for NLR and PLR levels.

Table S1. Linear model of NLR or PLR level on BMI effects separately in age groups.

Age group	men					women				
	N	β(for NLR)	p (for NLR)	β(for PLR)	p(for PLR)	N	β(for NLR)	p (for NLR)	β(for PLR)	p(for PLR)
18-30 years old	744	0.005	0.451	0.141	0.717	1130	**0.019**	**0.000**	0.415	0.174
30-40 years old	804	-0.007	0.288	**-1.412**	**0.001**	1360	**0.021**	**0.000**	0.185	0.486
40-50 years old	235	0.021	0.058	-0.715	0.270	827	**0.012**	**0.045**	0.085	0.822
50-60 years old	686	0.001	0.941	**-1.559**	**0.002**	1032	**0.013**	**0.004**	**-0.696**	**0.019**
older than 60	577	-0.010	0.252	**-1.581**	**0.002**	650	0.007	0.227	**-1.182**	**0.004**

Table S2. ANOVA test of NLR or PLR level on BMI and age effects.

Models	Age groups (p-value)	BMI groups (p-value)	Age groups*BMI groups (p-value)
NLR in men	**<0.001**	0.591	0.182
NLR in women	**<0.001**	**<0.001**	0.322
PLR in men	**<0.001**	**0.002**	0.096
PLR in women	**0.006**	0.363	**0.001**

Table S3. Standard deviations in different age and BMI groups.

Age group	N(men/women)	NLR(men)	NLR(women)	PLR(men)	PLR(women)
18-30 years old	744/1130	.577	.667	34.26	36.62
30-40 years old	804/1360	.609	.710	39.51	41.69
40-50 years old	235/827	.652	.754	37.98	46.60
50-60 years old	686/1032	.662	.670	43.48	43.41
older than 60	577/650	.727	.668	43.85	45.90

Appendix IV.

Results for NLR and PLR in total and unhealthy population

In this section, we mainly focus on the comparison of NLR, PLR and its subcomponents' characters between healthy population, unhealthy population and total population. Based upon disease and immune parameter specifications, we first divided the population into a healthy and unhealthy subpopulation from total population (N=9434). Being unhealthy was based on the possibility of having a compromised immune system, and referred to several indicators: 1) illness reported in the week (N=539); 2) CRP ≥ 15 (N=287); 3) basophil count > 0.02×109/L (N=151); 4) report of chronic immune disease or cancer (N=83); and 5) use of anti-inflammatory medication (N=437), glucocorticoids (N=143) or iron supplements (N=28). All individuals who met one or more of these criteria were classified as unhealthy, leading to 1326 individuals within the unhealthy population and 8108 within the healthy population.

Healthy versus unhealthy population.

Table SS1 provides the descriptive statistics for NLR and PLR, their subcomponents neutrophil, lymphocyte, and platelet count, CRP and IL6 level in the unhealthy population. To test for differences in NLR and PLR characteristics across two subpopulations, one classified as healthy and the other as unhealthy, we conducted an ANOVA using SPSS (version 21.0) [177], including age and sex as covariates. To explore the relationship of NLR and PLR with two inflammatory markers, IL6 and CRP, we computed their correlations in the unhealthy population and total population.

The unhealthy population was significantly older (t(9432)=3.002, p=0.003) and more likely to be female (t(9432)=3.837, p<0.001) compared to healthy population. Using age- and sex- corrected data, the unhealthy population has a higher mean NLR (t(9432)=12.460, p<0.001) and PLR (t(9432)=6.451, p<0.001). This difference was related to higher neutrophil (t(9432)=10.925, p<0.001) and platelet counts (t(9432)=4.465, p<0.001) and lower lymphocyte counts (t(9432)=2.126, p<0.034). As one would expect, since CRP was one of the criteria used to defined the unhealthy population, the unhealthy population had a higher mean CRP (t(9432)=35.064, p<0.001) but also higher IL6 values (t(9432)=10.014, p<0.001) than the healthy population. To further explore the association between the ratios and the two established markers of inflammation, CRP and IL6, we also examined the correlation in the total sample, and separately for the healthy and unhealthy population (**Table SS2**). Expect the correlation between NLR, PLR with lymphocyte count, all association are stronger in unhealthy population than the healthy population.

Subjective health was more often judged as less than good in the unhealthy population and, as can be seen in **Table SS3**, NLR and PLR increased with a decrease in subjective health, more obviously so in the unhealthy population.

Table SS1. Mean (sd) for NLR and PLR and their constituents in unhealthy population.

	Male	Female
N	429	897
age	46.66(17.38)	43.95(15.65)
NLR	2.05(0.96)	1.85(0.87)
PLR	127.59(49.31)	132.69(49.08)
Neutrophil	3.94(1.57)	3.82(1.58)
Lymphocyte	2.09(0.73)	2.23(0.74)
Platelet	242.96(59.42)	271.24(66.27)
CRP*	8.09(16.27)	9.00(13.43)
IL6*	3.08(5.30)	3.09(9.24)

* The sample size for CRP and IL6 is somewhat smaller due to missing data, CPR: N=423 for men, N=890 for women; IL6: N=399 for men, N=849 for women.

Table SS2. Correlation of NLR, PLR and other immune biomarkers (corrected age and sex) in unhealthy population and total population.

	r with NLR		r with PLR	
variables	unhealthy	all	unhealthy	all
NLR	1	1	0.495**	0.497**
PLR	0.495**	0.497**	1	1
Neutrophils	0.699**	0.669**	-0.029	-0.053**
lymphocytes	-0.471**	-0.472**	-0.676**	-0.681**
Platelet	0.055*	0.040**	0.464**	0.474**
CRP	0.291**	0.214**	0.049	0.040
IL-6	0.176**	0.121**	0.022	0.004

Table SS3. Means and standard errors of immune biomarkers in different health status categories.

	Healthy population					Unhealthy population				
	%	NLR	PLR	CRP	IL-6	%	NLR	PLR	CRP	IL-6
Bad	5.3	1.711	122.619	3.148	2.223	17.8	2.024	135.265	14.106	3.676
		(0.033)	(2.018)	(0.132)	(0.173)		(0.059)	(3.194)	(0.937)	(0.548)
Reasonable	9.4	1.666	122.305	2.834	1.702	20.0	1.896	130.730	7.423	3.349
		(0.025)	(1.517)	(0.099)	(0.130)		(0.056)	(3.019)	(0.889)	(0.530)
Good	75.4	1.641	121.827	2.355	1.654	58.2	1.890	130.113	7.787	2.941
		(0.009)	(0.535)	(0.035)	(0.046)		(0.033)	(1.767)	(0.519)	(0.305)
Excellent	9.9	1.584	122.788	2.100	1.362	4.1	1.811	132.149	5.712	1.598
		(0.024)	(1.475)	(0.096)	(0.129)		(0.125)	(6.710)	(1.983)	(1.159)

Heritability in total population

We restricted the dataset for the total population to subset of twins within a family, a maximum of two brothers and two sisters, father and mother. This resulted in 8689 total participants with 3388 families, including 299 MZM, 131 DZM, 700 MZF, 294 DZF and 287 DOS twin pairs. We took age, sex, age x sex interaction on the mean into account in twins study. The genetic model for heritability study in total population is same as genetic model for healthy population.

We were allowed to equate the MZ correlation for males and females and to equate all first-relative correlations for both ratios. Familiar correlations of NLR and PLR are shown in **Table SS4**. For NLR, the resulting MZ correlation was 0.34 (.28-.39) and the DZ correlation was .15 (.09-.22), which suggests non-additive genetic effects are absent. For PLR, the MZ correlation was 0.60 (.56-.68) and DZ was .23 (.16-.30), suggesting the presence of non-additive genetic effect. Spouse correlations could not be dropped from the model: the spouse correlation was 0.12 for NLR and 0.18 for PLR. Correlations in total populations have similar pattern as correlations of healthy population, but estimations are slightly lower.

The most parsimonious models showed no evidence for a common environmental influence for both NLR and PLR. The heritability for NLR was estimated at 32.1% in total population, with no evidence for non-additive effects. For PLR the broad sense heritability of PLR was 60.4% in total population, with non-additive effects accounted for 22.1% in total population. Unique environmental effects thus counted for 67.9% in total population of the variance in NLR and 39.6% of the variance in PLR. Compared to healthy population, both genetic effects on NLR and PLR are smaller in total population.

With regards of three sub components for these two biomarker (**Table SS5**), for neutrophils count, the correlations of MZ, other first degree relative and spouse pairs are 0.421 and 0.195 and 0.154. The heritability is 37.3%, the else part of variance due to unique environmental effects. For lymphocyte count, the correlations of MZ, other first degree relative and spouse pairs are 0.579 and 0.209 and 0.178. The narrow sense and broad sense heritability's are and 33.7% and 54.8%, the else part of variance due to unique environmental effects. For platelets count, the correlations of MZ , other first degree relative and spouse pairs are 0.639 and 0.279 and 0.130. The narrow sense and broad sense heritability are 45.2% and 61.3%. The common environmental effect is small (5.6%) but significant (p=0.03), and unique environmental effect is 33.2%. Compared the results between healthy population and total population, the correlations of family membership in healthy population are always smaller than that of total population, which give bigger heritability's in healthy population.

Table SS4. Familial correlations of PRL and NLR in total population. (N=8689, family=3388).

pairs	NLR		PLR	
	R	95% CI	R	95% CI
MZ twins	**.341**	**.284-.394**	**.603**	**.563-.637**
MZ male	.341	.237-.432	.581	.503-.644
MZ female	.341	.271-.404	.613	.567-.653
Male first-degree relatives	**.151**	**.087-.215**	**.230**	**.155-.300**
DZ male	.288	.111-.435	.327	.129-.480
Brother-male twin	.213	.096-.317	.265	.094-.404
Brother-brother	.078	-.192-.336	.405	.096-.596
Father-son	.108	.024-.189	.200	.114-.280
Female first- degree relatives	**.189**	**.152-.224**	**.247**	**.210-.283**
DZ female	.245	.131-.348	.387	.282-.477
Sister-female twin	.209	.118-.291	.263	.178-.341
Sister-sister	.128	.033-.219	.167	.084-.248
Mother-daughter	.124	.073-.173	.217	.169-.263
Female-male first degree relatives	**.152**	**.113-.190**	**.229**	**.191-.266**
DZ opposite sex	.210	.008-.237	.286	.171-.388
Brother-female twin	.112	-.007-.223	.239	.124-.341
Sister-male twin	.141	.027-.248	.362	.247-.453
Sister-brother	.126	.008-.238	.265	.165-.354
Mother-son	.273	.193-.344	.255	.173-.329
Father-daughter	.188	.132-.241	.227	.175-.278
Parents (father-mother)	**.117**	**.058-.174**	**.178**	**.123-.232**
Heritability	**.321**	**.293-.419**	**.604**	**.586-.679**

Correlations in bold italic were obtained from sub-models in which all matching correlations of the tested subgroup of family relations were set to be equal.

Table SS5. Familial correlations of neutrophil, lymphocyte and platelet count in healthy population (N=7481, family=3251) and total population. (N=8689, family=3388).

	Neutrophil R (95%)		Lymphocyte R (95%)		Platelet R (95%)	
	Healthy population	Total population	Healthy population	Total population	Healthy population	Total population
MZ twins	***.457*** ***(.403-.506)***	***.421*** ***(.371-(.468)***	***.582*** ***(.537-.621)***	***.579*** ***(.541-.614)***	***.681*** ***(.651-.704)***	***.639*** ***(.628-.697)***
MZ male	.529 (.432-.607)	.505 (.413-.580)	.576 (.487-.646)	.576 (.498-.640)	.757 (.689-.968)	.744 (.697-.782)
MZ female	.431 (.366-.490)	.392 (.332-.448)	.584 (.531-.630)	.581 (.536-.621)	.845 (.811-.930)	.642 (.602-.677)
Malefirstdegreerelatives	***.238*** ***(.155-.316)***	***.193*** ***(.122-.262)***	***.234*** ***(.156-.307)***	***.226*** ***(.157-.292)***	***.540*** ***(.492-.794)***	***.307*** ***(.240-.371)***
DZ male	.174 (-.122-.412)	.070 (-.156-.289)	.390 (.153-.551)	.447 (.267-.575)	.497 (-.673-.612)	.331 (.094-.504)
Brother-male twin	.206 (.012-.369)	.173 (.041-.293)	.181 (.007-.336)	.130 (-.027-.278)	.356 (-.623-.386)	.464 (.340-.558)
Brother-brother	.155 (-.103-.398)	.081 (-.144-.307)	.341 (.007-.580)	.439 (.124-.622)	.266 (.244-.604)	.505 (.293-.649)
Father-son	.262 (.165-.348)	.231 (.146-.309)	.219 (.122-.307)	.216 (.131-.294)	.609 (.585-.619)	.254 (.167-.333)
Femalefirst- degreerelatives	***.203*** ***(.159-.247)***	***.177*** ***(.138-.215)***	***.225*** ***(.183-.266)***	***.187*** ***(.150-.224)***	***.552*** ***(.539-.732)***	***.245*** ***(.209-.279)***
DZ female	.214 (.091-.327)	.222 (.113-.321)	.286 (.181-.382)	.219 (.125-.308)	.506 (-.831-.955)	.341 (.243-.428)
Sister-female twin	.250 (.153-.335)	.196 (.108-.276)	.175 (.008-.265)	.189 (.101-.271)	.129 (.072-.148)	.285 (.208-.356)
Sister-sister	.216	.211	.151	.117	.336	.239

	(.122-.305)	(.117-.300)	(.055-.243)	(.022-.208)	(.203-.385)	(.157-.316)
Mother-daughter	.178	.153	.210	.196	.369	.223
	(.118-.235)	(.102-.201)	(.157-.260)	(.150-.240)	(-.457-.640)	(.178-.172)
Female-male first degree relative	***.216***	***.213***	***.203***	***.229***	***.451***	***.312***
	(.173-.257)	***(.176-.250)***	***(.162-.244)***	***(.192-.265)***	***(.448-.464)***	***(.277-.346)***
DZ opposite sex	.111	.127	.216	.171	.267	.366
	(-.018-.235)	(.016-.235)	(.086-.333)	(.022-.208)	(-.623-.267)	(.264-.455)
Brother-female twin	.256	.227	.182	.174	.397	.398
	(.151-.350)	(.128-.317)	(.007-.336)	(.065-.276)	(.394-.409)	(.297-.483)
Sister-male twin	.136	.122	.172	.179	.285	.384
	(.010-.255)	(.007-.231)	(.023-.307)	(.040-.303)	(.259-.330)	(.284-.470)
Sister-brother	.246	.230	.312	.356	.211	.356
	(.099-.365)	(.095-.342)	(.178-.419)	(.238-.447)	(.200-.243)	(.251-.443)
Mother-son	.277	.293	.198	.227	.064	.346
	(.194-.353)	(.222-.358)	(.115-.276)	(.153-.296)	(.049-.561)	(.273-.412)
Father-daughter	.210	.207	.245	.239	.306	.246
	(.143-.273)	(.148-.263)	(.183-.260)	(.187-.289)	(-.204-.582)	(.189-.299)
Parents (father-mother)	***.201***	***.156***	***.166***	***.178***	***.147***	***.130***
	(.121-.277)	***(.088-.221)***	***(.089-.241)***	***(.112-.241)***	***(-.163-.308)***	***(.077-.194)***
Heritability	***.378***	***.373***	***.576***	***.548***	***.709***	***.613***
	(.325-.411)	***(.295-.409)***	**(.461-.653)**	***(.412-.637)***	***(.534-.718)***	***(.483-.709)***

Correlations in bold italic were obtained from sub-models in which all matching correlations of the tested subgroup of family relations were set to be equal.

Conclusion

Unhealthy individuals, classified on disease history and biomarker profiles, had higher NLR and PLR levels than healthy individuals. The familial correlations of NLR and PLR are higher in unhealthy population than correlations in healthy population. Same result pattern has been observed in correlations of NLR with CRP, NLR with IL-6, NLR with neutrophil count, NLR with platelet count, PLR with platelet. The variance of blood ratios and blood counts caused by less genetic effects if we including unhealthy population into model.

Discussion

Our comparison of healthy individuals with unhealthy individuals who had possibly a compromised immune system showed that higher NLR and PLR levels in the unhealthy individuals. NLR thus seems to be the most promising marker for inflammation, which also suggest from other study [178], though more study is needed to establish its use in clinical practice.

Heritability studies show the familial correlations of NLR and PLR to become smaller after adding unhealthy individuals into account, which results smaller heritabilities in total population than healthy population. Health status, which is also tightly regulated by genetic effects, can become a confounding factor for twins study.

Chapter 5

SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio

This paper is to be submitted as: BD Lin, EC Montoro, JT Bell, DI Boomsma, EJCN de Geus, R Jansen, C Kluft, M Mangino, BWJH Penninx, TD Spector, G Willemsen, JJ Hottenga. SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio.

Abstract

Neutrophil-to-lymphocyte ratio (NLR) and platelet-to-lymphocyte ratio (PLR) are important biomarkers for disease development and progression. To gain insight into the genetic causes of variance in NLR and PLR in the general population, we conducted genome-wide association (GWA) analyses and estimated SNP heritability in a sample of 5901 related Dutch individuals. GWA analyses identified a new genome-wide significant locus on the *HBS1L-MYB* intergenic region for PLR, which replicated in a sample of 2538 British twins. For platelet count, we replicated three known genome-wide significant loci in our cohort (at *CCDC71L-PIK3CG*, *BAK1* and *ARHGEF3*). For neutrophil count, we also replicated the *PSMD3* locus. For the identified top SNPs, we found significant cis and trans eQTL effects for several loci involved in hematological and immunological pathways. Linkage Disequilibrium score (LD) regression analyses for PLR and NLR confirmed that both traits are heritable, with a polygenetic SNP-heritability for PLR of 14.1%, and for NLR of 2.4%. Genetic correlations were present between ratios and the constituent counts, with the genetic correlation (r=0.45) of PLR with platelet count reaching statistical significance. In conclusion, we established that two important biomarkers have a significant heritable SNP component, and identified the first genome-wide locus for PLR.

Keywords: NLR, PLR, GWAS, eQTL, SNP heritability.

5.1 Introduction

Both neutrophil-to-lymphocyte ratio (NLR) and platelet-to-lymphocyte ratio (PLR) have been suggested as novel and useful biomarkers for the diagnosis or prognostic prediction of diseases [179-180]. A high NLR level was shown to be an independent predictor of mortality in patients undergoing cardiac revascularization [181] and in patients with myocardial infarction [182]. Elevated NLR levels were also related to a poor prognosis of various cancers, such as esophageal, pancreatic, lung, ovarian and hepatocellular cancer [183-185]. Similar to NLR, PLR was also reported as an index for diagnosis or prognostic prediction of oncologic disorders and inflammatory diseases [70, 186]. NLR and PLR thus may serve as biomarkers in patient populations. However, studies of variation in these biomarkers within healthy populations are scarce. Recently, we showed that variation in NLR and PLR levels is due to genetic influences, with a broad sense heritability of 36% for NLR and 64% for PLR, using a twin-family epidemiological design [187]. Here, we investigate if the significant heritability estimates can be explained by common SNPs (single nucleotide polymorphisms) and if we can identify the genes that play a role in these two blood ratios. We also investigate if our findings are unique to the two ratios or whether their count-components (i.e. lymphocyte, neutrophil and platelet counts) show similar results.

No genome-wide association study (GWAS) has yet been published for NLR and PLR. However, GWASs on their subcomponents, the neutrophil, lymphocyte and platelet counts were carried out in different populations including European [188-191], African-American [192-195], Korean [196-197] and Japanese populations [60, 198]. These GWASs for blood cell count in different cohorts have identified multiple genetic loci for blood cell components. For neutrophil count, the *DARC* gene promoter at 1q23.3 was identified in African-American populations [199] and loci at 20p12 (*PLCB4* gene) [198] and 7q21.2 (*CDK6* gene) [60] were found in the Japanese population. The chromosomal region nearby *PSMD3* on 17q21 was associated in a GWAS meta-analysis in both Japanese and European ancestry cohorts, but not in African-American cohorts. The variants at *AK123889* on 6p21.33 were novel findings in a European ancestry cohort [192], and were also confirmed by meta-analysis [200]. For lymphocyte count, two genetic variants nearby *EPS15L1* gene on 6p21 and *LOC101929772* on 19p13 were identified [195]. For platelet count, many loci were identified: *SH2B3* on 12q24, *ARHGEF3* on 3p14.3, *ZBTB9-BAK1* on 6p21.31, *KIAA0232* on 4p16.1, *EGF* on 4q25, *PNPLA3* on 22q13.31 in the Korean population [196-197]. *ARHGEF3* on 3p14.3, *PEAR1* on 1q23.1, *BMPR1A* on 10q23.2, loci on 6p22, 7q11, 10q21, 11q13, 20q13 were detected in the African-American population [193-194, 201] and over 55 loci including *CCDC71L-PIK3CG*, *ARHGEF3*, *BAK1* and *HBS1L-MYB* in the European population [188-189, 191, 202].

Some blood cell count loci show pleiotropy: they influence multiple hematological indices [60, 202-205]. For example, the genetic region nearby *AK123889* on 6p21.33 was associated with neutrophil count, lymphocyte count and total white blood cell count [195, 200] and the DARC promoter on 1q23.2 was associated with neutrophil count, monocyte count and total number of white blood cells [192, 206]. The intergenic *HBS1L-MYB* variants were associated with total white blood cell count and also with number of neutrophils, lymphocytes, erythrocytes, eosinophils, monocytes, and platelets [60, 207]. Therefore, we also examined genetic effects across the ratios and constituent cell counts. We conducted five GWASs to identify genetic variants associated with NLR, PLR and neutrophil, lymphocyte and platelet counts. The discovery cohort consisted of 5901 participants from the Netherlands Twin Register (NTR) and replication of top results was sought in a TwinsUK cohort consisting of 2538 participants. Furthermore, all top SNPs, which showed a significant association with our phenotypes of interest, were selected for an eQTL analysis to test whether these variants have an effect on the gene expression level. For the ratios, we estimated the proportion of trait variance explained by significant SNPs from the GWAS and the variance explained by SNPs that were associated with lymphocyte, platelet and neutrophil counts [6, 26]. Using the summary statistics of the GWAS results, we applied LD regression to determine the variance explained by all autosomal SNPs to examine polygenetic effects between NLR and PLR and to determine the genetic correlation between variants affecting the two ratios and their subcomponents [208-209].

5.2 Methods

5.2.1 Participants

All participants were registered with the Netherlands Twin Register (NTR) and had taken part in one of two biobanking projects with similar procedures conducted between 2004 and 2008 and in 2011 [76-77]. After removing outliers (defined as values outside mean ±5×SD for NLR, PLR or their subcomponents), the initial sample size for PLR and NLR was 9434 individuals clustered in 3411 families. We further excluded individuals who met one or more of these criteria: 1) illness in the sampling collection week (N=539); 2) values of CRP ≥ 15 mg/L (N=287); 3) basophil count > 0.02×10^9/L (N=151); 4) report of chronic immune disease or cancer (N=83); and 5) anti-inflammatory medication, glucocorticoids or iron supplements (N=537). The sample size reduced to 8018 with these criteria applied. When linking these data to the genetic data, 6112 individuals had both phenotype and genotype data. After exclusion of 211 individuals with non-Dutch ancestry, the final sample size for GWAS, GCTA and LD score regression was 5901 individuals. Written informed consent was obtained from all participants and the Medical Ethics Committee of the VU Medical Centre approved the study protocols.

5.2.2 Cell Counts

Blood samples were obtained during a home visit, or sometimes a work visit, between 7 and 10 a.m. Participants were instructed to fast overnight and to refrain from heavy physical exertion and medication use (if possible) in the morning before the visit. Smokers were asked to abstain from smoking at least one hour prior to the home visit. For fertile women without hormonal birth control, when possible, an appointment was made within the 2nd to 4th day of the menstrual cycle and women taking hormonal birth control were visited during the pill-free week. Peripheral venous blood samples were collected into multiple anticoagulant vacuum tubes. Within 3 to 6 hours upon blood withdraw all tubes were transported to the laboratory in Leiden. During the visit to the participants, phenotypic data were also collected on body composition, the presence of chronic diseases, medication use, and smoking history [76].

The hematological profile, including the number of neutrophils, lymphocytes and platelets, was obtained from 2 ml EDTA tubes using the Coulter system (Coulter Corporation, Miami, USA). NLR was calculated as the absolute neutrophil count (10^9/L) divided by the absolute lymphocyte count (10^9/L), and PLR was calculated as the absolute platelet count (10^9/L) divided by the absolute lymphocyte count (10^9/L).

5.2.3 Genotype Data

For DNA isolation, we used the GENTRA Puregene DNA isolation kit. All procedures were performed according to the manufacturer's protocols [131]. Genotyping was done on multiple chip platforms, with a number of overlapping participants. Chronologically the following platforms were used: Affymetrix Perlegen 5.0 (N=1,718), Illumina 370 (N=424), Illumina 660 (N=1,103), Illumina Omni Express 1 M (N=346) and Affymetrix 6.0 (N=3602). Genotype calls were made with the platform specific software (Birdsuite, APT-Genotyper, Beadstudio) for each specific array.

Quality control was done within and between platforms and subsets. 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”). The data were then strand aligned with the 1000 Genomes GIANT phase1 release v3 20,101,123 SNPs INDELS SVS ALL panel. SNPs from each platform were removed if they had ambiguous locations, mismatching alleles with this imputation reference set or the allele frequencies differed more than 0.20 compared to the reference. From each platform, SNPs were also excluded if meeting the following criteria: a Minor Allele Frequency (MAF) <1%, Hardy–Weinberg Equilibrium (HWE) with $p < 0.00001$, and call rate <95%. Samples were excluded from the analysis when their expected sex did not match their genotyped sex, the genotype missing rate was above 10% or the PLINK1.07 F inbreeding value was either >0.10 or <−0.10.

After these steps, the data of the individual arrays were merged into a single dataset using PLINK 1.07 [94]. Within the merged set, identity by state (IBS) sharing was calculated between all possible pairs of participants and compared to the known NTR family structures. Samples were removed if the data did not match their expected IBS sharing. The concordance rate of DNA samples on multiple platforms for overlapping SNPs generally exceeded 99.0% after data cleaning. The HWE-, MAF- and the reference allele frequency difference <0.20 filters were re-applied in the combined data. As a final step, SNPs with C/G and A/T allele combinations were removed when the MAF was between 0.35 and 0.50 to avoid incorrect strand alignment. Phasing of all samples and imputing cross-missing platform SNPs was done with MACH 1 [132]. The phased data were then imputed with MINIMAC [210] in batches of around 500 individuals for the autosomal genome using the above 1000G Phase I integrated reference panel for 561 chromosome chunks obtained by the CHUNKCHROMOSOME program [97]. To avoid issues having SNPs from different platforms partly imputed and partly genotyped we took the re-imputed calls for all genotyped SNPs. After imputation of these SNPs, we generally find a high concordance between re-imputed SNPs and the original genotype (0.9868). The mean imputation quality R^2 metric is 0.38 (based on all 30,051,533 imputed autosomal SNPs).
After imputation, SNPs were filtered based on the Mendelian error rate in families. The Mendelian error rate was calculated on the best guess genotypes in families (trios or sib-pairs with parents) using first GTOOL to calculate best guess genotypes and then PLINK 1.07 to analyze the data. SNPs were removed if the Mendelian error rate >0.02, the imputed allele frequency differed more than 0.15 from the 1000G reference allele frequency, MAF < 0.01 or $R^2 < 0.80$. HWE was calculated on the genotype probability counts for the full sample, and SNPs were removed if the p-value < 0.00001. This left 6,010,458 SNPs for the GWAS analyses.

5.3 Analyses

5.3.1 Generation of Genetic Relatedness Matrices

Genetic Relatedness Matrices (GRMs) with the values of the identity by state (IBS) allele sharing for a given set of SNP markers between all possible pairs of individuals were calculated with the GCTA software [211]. Since we combined genotype data from different genotyping platforms, which cannot be corrected for when calculating the GRMs, we first removed SNPs that showed significant genotyping differences between platforms (p < 0.0001). A total of 6,009,498 SNPs were retained, which is sufficient for GRM estimation [211]. The SNP data were transformed to best guess Plink binary format, and subsets were made for each of the 22 chromosomes. We generated 25 GRMs: one GRM containing only the significant GWAS SNPs for PLR from our own study, and one GRM containing the SNPs known to be involved in the cell counts. A third GRM was constructed for only closely related individuals (IBS> 0.05), pairwise relationship estimates smaller than 0.05 were set to 0 in this matrix [26]. This matrix is used as second covariate matrix in the GWAS and GCTA studies to account for the family structure and it provides an estimate of the total heritability [26]. Finally, 22 GRM matrixes were made that include all autosomal SNPs, except for the one chromosome on which the SNP is present that is tested in the GWAS: the Leave One Chromosome Out (LOCO) strategy [212]. These matrixes are used in the GWAS study as covariates to remove any remaining statistical inflation due to subsample stratification.

5.3.2 GWAS

Three Dutch Principal Components (PCs) were generated with the EIGENSOFT software as described earlier by Abdellaoui et al. [213-214] to be used as covariates in the GWAS. Additional covariates were age, sex and genotype platform. For NLR and PLR as well as for the three sub-component counts we modelled the phenotypes as being influenced by SNP and the six covariates. Analyses were performed with the GCTA software running a mixed linear association model (MLMA) including the LOCO GRMs for chromosome 1 to 22, and the close-related GRM [26, 215]. For the GWAs, the significance threshold was p-value < $5×10^{-8}$ [21].

5.3.3 GWAS Replication

Replication of significant GWAS hits for NLR, PLR or individual blood cell counts, which were not previously found, was examined in TwinsUK. TwinsUK is an United Kingdom based twin registry with a focus on the genetic and environmental etiology of age related complex traits and diseases [216]. Samples from TwinsUK were genotyped using the Illumina Hap317K and Hap610K assays (Illumina, San Diego, USA) following standard procedures. Normalised intensity data were pooled and genotypes called on the basis of the Illumina’s algorithm [217]. No calls were assigned if the most likely call was less than a posterior probability of 0.95. SNPs that had a low call rate. Subjects were also removed if the sample call rate was less than 95%, autosomal heterozygosity was outside the expected range, genotype concordance was over 97% with another sample and the sample was of lesser call rate. Imputation of genotypes was carried out using the software IMPUTE [96]. The best guess Plink binary format data was used to conduct the replication analysis. The sample size of the TwinsUK dataset was 2538 subjects with genetic and phenotypic information, after values outside mean±5SD in the phenotype of interest were removed. We tested the association with the SNPs using a linear mixed model, in which the traits were regressed on the SNPs, while correcting for age and sex as fixed effects variables.

5.3.4 eQTL Analysis

To determine the effects of the GWAS located genetic variants for both ratios as well as the constituent counts, we conducted eQTL analysis, using the NESDA-NTR Conditional eQTL Catalog (online accessible: https://eqtl.onderzoek.io). The details of the eQTL analysis are described in the supplementary material. In brief, eQTL effects were examined with a linear model approach using MatrixeQTL [218] with expression level as dependent variable and SNP genotype values as independent variable. eQTL effects were defined as cis when probe set–SNP pairs were at distance < 1M base pairs (Mb), and as trans when the SNP and the probe set were separated by more than 1 Mb on the genome according to the Human reference genome HG19. To determine whether the observed cis and trans effects may reflect causal mechanism we checked the LD of our top SNPs with the top SNPS identified for gene expression in the implicated genes. Since gene expression is related to blood composition we repeated the analysis with and without correction for blood composition components (specifically mean corpuscular volume, red cell distribution width, and neutrophil, lymphocyte, monocyte, eosinophil, basophil and platelet counts).

5.3.5 SNP Heritability and Genetic Correlation

The variance explained by the significant SNPs in our GWAS for PLR was estimated with the GCTA software [211]. The variance explained in NLR and PLR was estimated with GCTA for the known loci from literature for neutrophil, platelet and lymphocyte blood cell counts. For each analysis we included family members and therefore included the closely-related GRM under the Restricted maximum likelihood (REML) analysis procedure within GCTA [26]. Sex, age, genotype platform and three Dutch PCs were used as covariates. The variance explained by all SNPs was estimated by Linkage Disequilibrium (LD) regression between our computed GWAS summary statistics effect sizes and the expected Hapmap 3 LD [208]. In order to do this, we used the HapMap3 LD scores (NSNPs= 1,293,150), computed for each SNP based on the LD observed in European ancestry individuals from 1000 Genomes project (http://github.com/bulik/ldsc). The criteria of passing quality control for SNPs were the default by LD regression: imputation quality info > 0.90, MAF > 0.01. SNPs with invalid P values (P >= 1 or P < 0) were excluded. In addition, variants that are not SNPs (e.g., insertion-deletions), strand ambiguous SNPs, and SNPs with duplicated RS numbers were also excluded. After quality control, the number of SNPs for these analyses reduced to 951,097.

Genetic correlations among the phenotypes of interest were also estimated using LD regression. The principle of this technique is that the genetic correlation of two traits can be calculated by the slope from the LD regression on the product of effect sizes (z-score) for two phenotypes. Pearson correlations between PLR, NLR and the constituent cell counts were calculated with the R program [219].

5.4 Results

5.4.1 GWAS

Summary statistics for the phenotypes of interest are given in **Table 1** and GWAS results for NLR and PLR are summarized in the Manhattan and QQ plots in **Figures 1** and **2** respectively. The GWAS inflation factors (λ) were 0.9963005 for NLR and 1.020995 for PLR, indicating that there is no hidden stratification left in the studied GWAS sample. For NLR, no loci were found that reached genome-wide significance levels. For PLR, there were 20 SNPs located between *HBS1L* and *MYB* genes on chromosome 6q23.3 in the *HBS1L-MYB* region, which were significantly associated with the phenotype (in **Figure 2** Manhattan, **Table 2** descriptive and **Figure 3** locus zoom). The top SNP rs9376092 of this locus has a C allele which significantly increases PLR level (β=5.48, p =2.75 × 10^{-9}). This SNP was also significantly associated with platelet count (β=6.98, p =4.05 × 10^{-8}), but not with lymphocyte count (β=-.039, p =0.008). In the TwinsUK sample, rs9376092 replicated with a similar effect for PLR (β=4.766, p=0.004) as well as platelet count (β=6.053, p =0.002). Here again, the SNP was not associated with lymphocyte count (β=0.014, p = 0.49) (**Table 3**).

Table 1. Summary statistics of neutrophil-lymphocyte ratio (NLR) and platelet-lymphocyte ratio (PLR), the constituent blood cell count phenotypes and age in males and females.

	mean (SD) males (N=2250)	mean (SD) females (N=3651)	mean (SD) all (N=5091)
NLR	1.662 (.653)	1.615 (.690)	1.633 (.676)
PLR	116.354 (39.457)	124.770 (42.626)	121.561 (41.643)
Neutrophil	3.404 (1.146)	3.429 (1.256)	3.419 (1.215)
Lymphocyte	2.170 (.622)	2.264 (.699)	2.228 (.673)
Platelet	235.464 (52.147)	262.006 (60.187)	251.886 (58.684)
Age	43.60 (15.88)	42.26 (14.27)	42.77 (14.91)

Table 2. The significant SNPs associated in our study for PLR and the p-values for the platelet and lymphocyte counts.

RSNUMBER	CHR	BP	MAF	B	SE	P (PLR)	P(platelet count)	P(lymphocyte)
rs9376092	6	135427144	0.270	5.483	0.922	2.75E-09	4.05E-08	0.008
rs4895440	6	135426558	0.271	5.478	0.922	2.83E-09	4.45E-08	0.007
rs35959442	6	135424179	0.272	5.476	0.922	2.9E-09	2.30E-08	0.009
rs4895441	6	135426573	0.270	5.472	0.922	2.95E-09	4.37E-08	0.007
rs11759553	6	135422296	0.273	5.446	0.919	3.19E-09	2.64E-08	0.010
rs7776054	6	135418916	0.258	5.492	0.934	4.21E-09	3.10E-08	0.014
NA	6	135418632	0.257	5.474	0.934	4.74E-09	3.10E-08	0.016
rs9399137	6	135419018	0.257	5.468	0.934	4.89E-09	1.99E-08	0.016
rs35786788	6	135419042	0.257	5.468	0.934	4.89E-09	1.99E-08	0.016
rs9373124	6	135423209	0.274	5.365	0.918	5.24E-09	2.64E-08	0.011
rs9389268	6	135419631	0.258	5.449	0.934	5.42E-09	2.64E-08	0.014
rs9376091	6	135419636	0.258	5.449	0.934	5.42E-09	2.52E-08	0.014
rs34164109	6	135421176	0.258	5.449	0.934	5.42E-09	2.52E-08	0.014
rs9402685	6	135419688	0.258	5.407	0.933	6.93E-09	3.30E-08	0.015
rs7758845	6	135428537	0.263	5.362	0.929	7.86E-09	5.06E-08	0.015
rs9376090	6	135411228	0.252	5.402	0.937	8.29E-09	1.69E-08	0.021
rs9389269	6	135427159	0.263	5.325	0.928	9.64E-09	4.51E-08	0.016
rs9402686	6	135427817	0.263	5.325	0.928	9.64E-09	4.51E-08	0.029
rs1331309	6	135406178	0.252	5.221	0.936	2.45E-08	3.93E-08	0.029
rs9399136	6	135402339	0.250	5.179	0.936	3.23E-08	3.73E-08	0.033

Table 3. Top SNP rs9376092 GWAS statistics results in NTR data and TwinsUK data.

Dataset	Alleles(A1/A2)	Frequency A1		Beta	se	P
NTR (N=5091)	C/A	0.72	PLR	5.484	0.923	**2.75E-9**
			Platelet	6.984	1.273	**4.05E-8**
			Lymphocytes	-0.039	0.015	0.008
TwinsUK (N=2538)	C/A	0.73	PLR	4.766	1.642	**0.004**
			Platelets	6.053	1.942	**0.002**
			Lymphocytes	0.014	0.020	0.49

Figure 1. A) Manhattan and B) QQ plot for neutrophil-lymphocyte ratio (NLR) GWAS results with SNPs' MAF> 0.01.

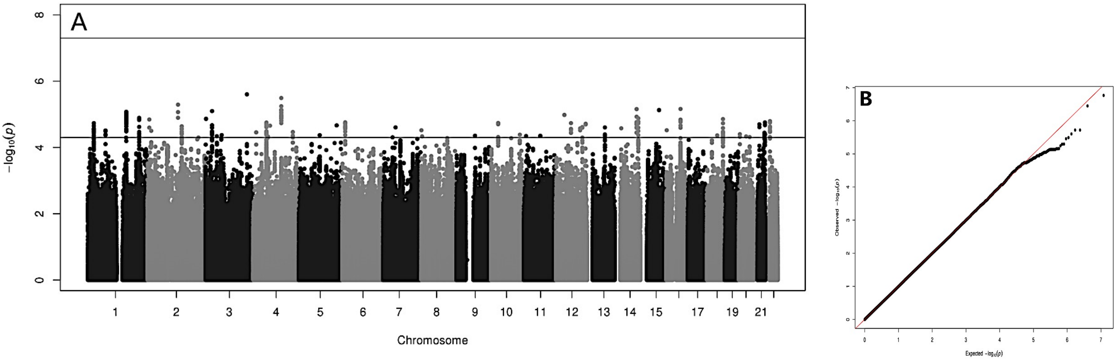


Figure 2. A) Manhattan and B) QQ plot for platelet-lymphocyte ratio (PLR) GWAS results with SNPs' MAF> 0.01.

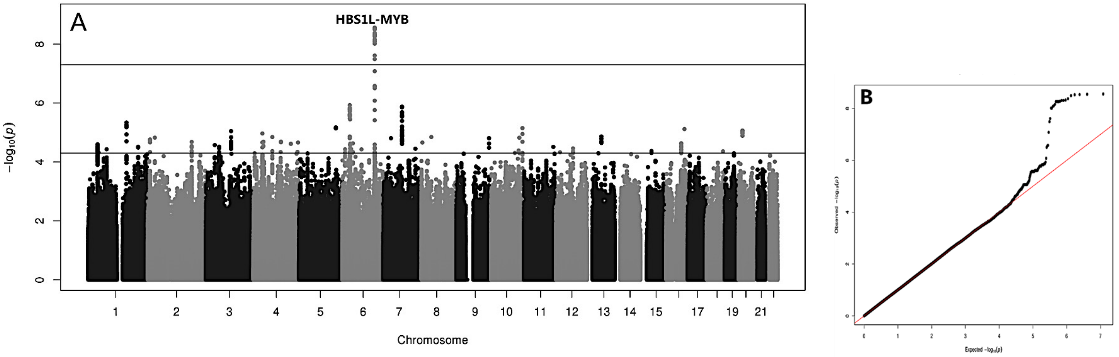
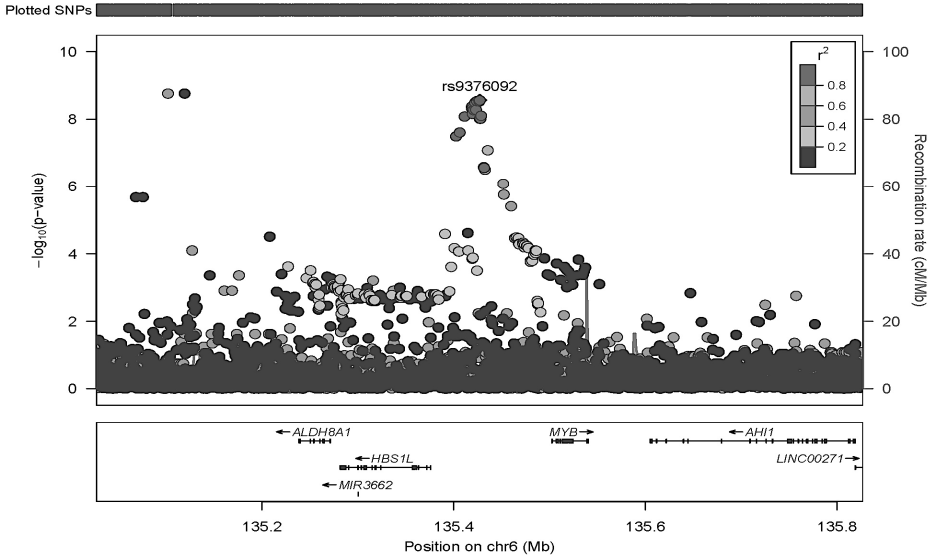


Figure 3. Regional plot for the rs9376092 association with PLR level.



Manhattan and QQ plots for the GWAS of neutrophil, lymphocyte and platelet counts are given in **Figure 4 to 6**. For neutrophil counts we found significant associations ($P < 5x10^{-8}$) for 65 SNPs in LD in the *PSMD3* locus (**Table 4**). For lymphocyte count we did not detect any significant genetic associations. For platelet count, a locus in *CCDC71L-PIK3CG* on 7q22.3 showed the strongest signal in our study ($p = 3.45\times 10^{-10}$). We also detected genetic variants for platelet count *ARHGEF3*, *BAK1* and *HBS1L-MYB*.

Table 4. Significant loci associated with neutrophil and platelet blood cell count within the NTR study.

Count	RSNUMBER	CHR	BP	GENE	MAF	β	SE	P
Neutrophil	rs8081692	17	38154595	PSMD3	0.370	0.157	0.024	1.77E-10
Platelet	rs342213	7	106324612	CCDC71L-PIK3CG	0.433	-6.981	1.112	3.45E-10
Platelet	rs169738	6	33537546	BAK1	0.412	-6.824	1.151	3.06E-09
Platelet	rs11925835	3	56865445	ARHGEF3	0.387	-6.457	1.147	1.84E-08
Platelet	rs9376090	6	135411228	HBS1L-MYB	0.254	7.296	1.293	1.69E-08

In **Table 5** we report the known genetic variants from literature for the three blood cell counts of interest and their significance level as reported previously, together with the p-values obtained from our GWAS study. For neutrophil count we replicated the *PSMD3* locus, which also showed an indication of association with PLR ($p < 1.0x10^{-3}$). The *AK123889* locus showed a similar pattern for PLR ($p = 3.67x10^{-4}$), and this locus also had a somewhat lower p-value for lymphocyte count (p = 0.001). For lymphocyte count, the known locus rs25224079 was marginally significant ($p = 3.02x10^{-4}$), while this locus showed a stronger association with PLR ($p = 6.56x10^{-5}$). We did not detect an association for lymphocyte count with the other known locus *ESP15L1* for lymphocyte count (p = 0.107). For platelet count, our top hit *CCDC71L-PIK3CG* was a replication of earlier studies and also associated with mean platelet volume [188, 202]. We also replicated the loci at *ARHGEF3, BAK1* and *HBS1L-MYB*, with the latter being associated with PLR as well. Furthermore five loci showed some signal at ($p < 1.0x10^{-3}$) for platelet counts: PDIA5, MEF2C, *JMD1C*, rs7149242 and *TAOK1*. Other platelet count loci showed some association ($p < 1.0x10^{-3}$) with related phenotypes: *RCL1, JMD1C*, rs7149242 and *SNORD7-AP2B1* with PLR, and *MICA* with lymphocytes.

Figure 4. A) Manhattan and B) QQ plot for neutrophil count GWAS results with SNPs' MAF> 0.01(λ=1.011742).

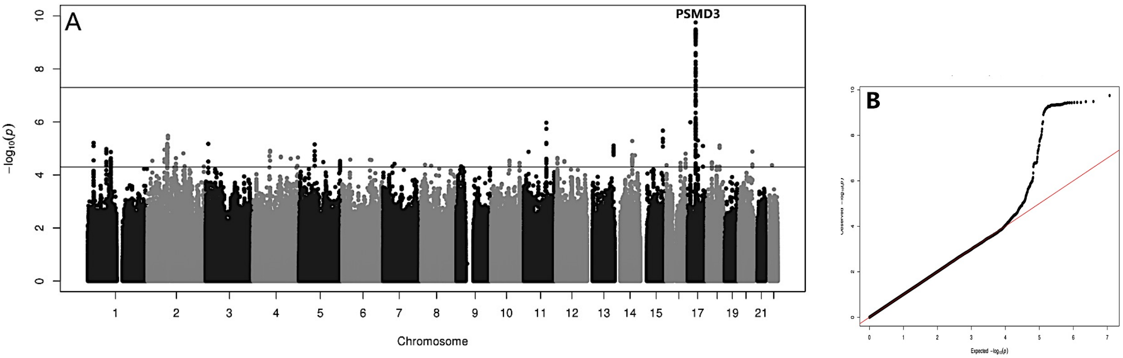


Figure 5. A) Manhattan and B) QQ plot for lymphocyte count GWAS results with SNPs' MAF> 0.01 (λ=1.022341).

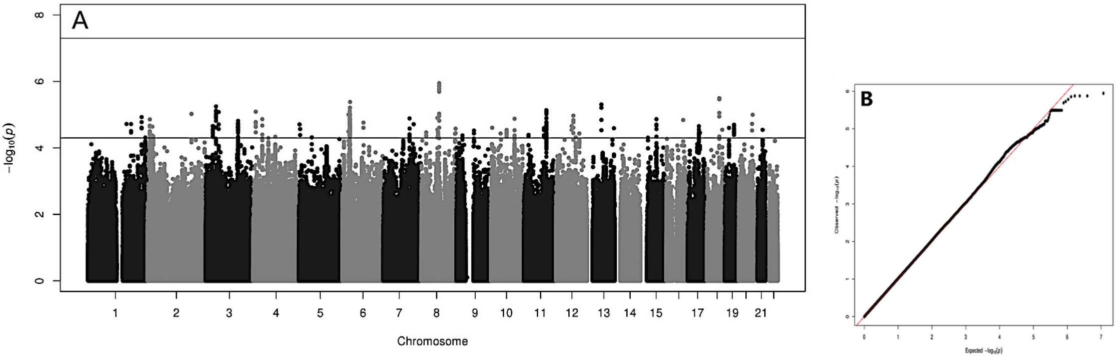


Figure 6. A) Manhattan and B) QQ plot for platelet count GWAS results with SNPs' MAF> 0.01 (λ=1.018586).

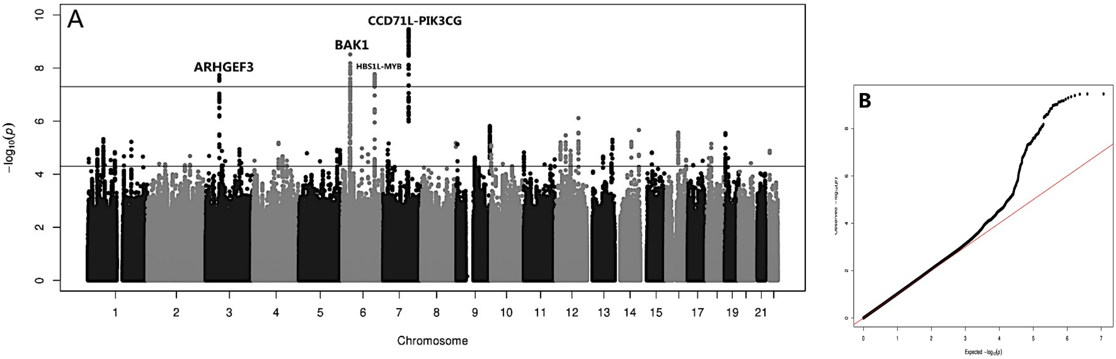


Table 5. Known blood cell count loci and their significance in our GWAS study.

Phenotype	Known SNPs	Gene	CHR	BP	Min P in reference	P in NLR	P in PLR	P in neutrophils	P in lymphocytes	P in platelets
neutrophil	rs2814778	DARC	1	1.59E+08	1.0E-524	0.904	0.388	0.56	0.615	0.038
	rs6936204	AK123889	6	32217092	5.83E-06	0.600	**3.67E-04**	0.016	0.001	0.156
	rs445	CDK6	7	92408370	4.50E-11	0.682	0.196	0.751	0.529	0.689
	rs2305482	PSMD3	17	38140927	3.05E-07	0.002	0.003	**1.61E-07**	0.12	0.354
	rs4794822	PSMD3	17	38156712	6.30E-10	0.001	**5.01E-04**	**3.68E-10**	0.018	0.266
	rs2072910	PLCB4	20	9365303	3.10E-10	0.907	0.497	0.273	0.207	0.953
lymphocyte	rs2524079	LOC101929772	6	31242174	1.85E-08	0.698	**6.56E-05**	0.005	**3.02E-04**	0.288
	rs11878602	EPS15L1	19	16555153	3.42E-09	0.337	0.007	0.462	0.107	0.359
platelet	rs2336384	MFN2	1	12046063	1.25E-08	0.288	0.938	0.824	0.900	0.368
	rs3091242	TMEM50A	1	25674785	3.85E-08	0.945	0.023	0.365	0.204	0.205
	rs12566888	PEAR1	1	156869047	1.17E-09	0.853	0.515	0.882	0.780	0.333
	rs10914144	DNM3	1	171949750	2.22E-12	0.593	0.017	0.611	0.108	0.023
	rs1668871	TMCC2	1	205237137	2.59E-14	0.691	0.004	0.771	0.162	0.264
	rs7550918	GCSAML	1	247675559	2.91E-11	0.716	0.022	0.313	0.721	0.023
	rs3811444	TRIM581	1	248039451	5.60E-11	0.275	0.02	0.43	0.58	0.068
	rs1260326	GCKR	2	27730940	9.12E-10	0.303	0.559	0.931	0.305	0.036
	rs625132	EHD3	2	31482300	1.17E-12	0.983	0.006	0.761	0.826	0.036
	rs6734238	IL1F10/IL1RN	2	113841030	3.77E-09	0.042	0.329	0.049	0.228	0.485
	rs78446341	LY75-CD302	2	160690656	1.97E-13	0.891	0.996	0.823	0.659	0.729

	rs7616006	SYN2	3	12267648	4.86E-08	0.548	0.361	0.619	0.166	0.271
	rs7641175	SATB1	3	18311412	3.37E-11	0.365	0.337	0.210	0.838	0.047
	rs17030845	THADA	3	43687879	1.27E-10	0.169	0.293	0.005	0.005	0.229
	rs1354034	ARHGEF3	3	56849749	2.86E-54	0.618	0.001	0.238	0.233	**3.34E-13**
	rs140286696	THPO	3	84090217	6.18E-08	0.600	0.326	0.481	0.631	0.121
	rs3792366	PDIA5	3	122839876	3.60E-09	0.678	0.010	0.693	0.764	**8.10E-04**
	rs56106611	KALRN	3	124377326	2.14E-09	0.249	0.167	0.035	0.395	0.002
	rs1470579	IGF2BP2	3	185529080	1.08E-07	0.040	0.279	0.279	0.211	0.145
	rs7694379	HSD17B13	4	88186509	8.70E-09	0.914	0.743	0.914	0.061	0.014
	rs1126673	LOC100507053	4	100045616	1.87E-08	0.790	0.995	0.163	0.168	0.043
	rs700585	MEF2C	5	76046939	9.86E-10	0.890	0.016	0.755	0.939	**4.10E-04**
	rs17568628	F2R	5	76058509	9.65E-16	0.231	0.446	0.258	0.939	0.721
	rs2070729	IRF1	5	131819921	1.13E-10	0.551	0.505	0.484	0.301	0.012
	rs1473247	RNF145	5	158603571	7.66E-10	0.613	0.046	0.162	0.393	0.048
	rs441460	LRRC16	6	25548288	8.70E-18	0.449	0.099	0.115	0.925	0.048
	rs3819299	HLA-B	6	31322367	8.80E-10	0.648	0.569	0.423	0.464	0.054
	rs2256183	MICA	6	31380529	3.20E-07	0.642	0.13	0.015	**4.51E-04**	0.037
	rs399604	HLA-DOA	6	32975014	1.30E-10	0.738	0.773	0.098	0.101	0.089
	rs210134	BAK1	6	33540209	1.30E-10	0.453	0.007	0.08	0.517	**1.66E-08**
	rs9399137	HBS1L-MYB	6	135419018	5.04E-47	0.262	**4.89E-09**	0.378	0.001	**1.99-08**
	rs1050331	ZMIZ2	7	44808091	3.28E-18	0.664	0.445	0.978	0.172	0.622
	rs4731120	FLH36031-PIK3CG	7	123411223	5.57E-25	0.258	0.203	0.543	0.376	0.006
	rs6993770	ZFPM2	8	106581528	4.30E-17	0.773	0.385	0.057	0.038	0.269
	rs6995402	PKEC1	8	145005561	5.09E-10	0.66	0.358	0.885	0.704	0.028
	rs409801	AK3	9	4744743	2.59E-49	0.086	0.007	0.156	0.777	0.094
	rs385893	AK3	9	4763176	5.60E-07	0.006	0.035	0.046	0.490	0.005
	rs13300663	RCL1	9	4814948	9.83E-30	0.710	**4.00E-04**	0.259	0.095	0.002

	rs3731211	CDKN2A	9	21986847	6.43E-14	0.654	0.013	0.234	0.310	0.094
	rs755109	HEMGN	9	100696203	2.59E-14	0.698	0.466	0.913	0.815	0.376
	rs11789898	BRD3	9	136925663	2.39E-10	0.142	0.04	0.196	0.789	0.073
	rs10761731	JMD1C	10	65027610	2.39E-44	0.279	**2.20E-04**	0.852	0.240	**1.40E-04**
	rs2068888	EXOC6	10	94839642	8.61E-08	0.172	0.889	0.269	0.695	0.076
	rs505404	PSMD13-NLRP6	11	243268	7.44E-25	0.875	0.015	0.406	0.676	0.195
	rs3794153	ST5	11	8751889	1.74E-10	0.005	0.018	0.612	0.008	0.548
	rs4246215	FEN1	11	61564299	3.31E-10	0.096	0.363	0.069	0.944	0.023
	rs174583	FADS2	11	61609750	4.42E-12	0.145	0.363	0.21	0.908	0.029
	rs45535039	CCDC153	11	119060963	4.02E-10	0.079	0.009	0.982	0.008	0.491
	rs4938642	CBL	11	119099906	7.66E-11	0.855	0.133	0.312	0.430	0.164
	rs7342306	LOC105369623	12	6291093	1.55E-21	0.828	0.41	0.702	0.947	0.036
	rs11616188	LTBR	12	6502742	1.26E-08	0.286	0.312	0.34	0.728	0.378
	rs10506328	NFE2	12	54687232	5.63E-11	0.285	0.035	0.354	0.267	0.009
	rs941207	BAZ2A	12	57023284	1.74E-10	0.297	0.04	0.053	0.328	0.059
	rs2279574	DUSP6	12	89745477	1.57E-07	0.95	0.38	0.846	0.521	0.369
	rs61745424	CUX2	12	111785515	6.30E-09	0.746	0.361	0.113	0.321	0.956
	rs3184504	SH2B3	12	111884608	1.22E-26	0.354	0.609	0.147	0.004	0.014
	rs11065987	ATXN2	12	112072424	8.30E-09	0.175	0.695	0.743	0.095	0.438
	rs11066301	PTPN11	12	112871372	2.30E-09	0.093	0.36	0.202	0.526	0.19
	rs17824620	RPH3A-PTPN11	12	113100994	9.67E-09	0.698	0.436	0.501	0.018	0.064
	rs7961894	WDR66	12	122365583	1.22E-10	0.811	0.37	0.572	0.332	0.036
	rs4148441	ABCC4	13	95898207	6.76E-12	0.926	0.789	0.617	0.445	0.032
	rs2784521	DDHD1	14	53657823	2.92E-08	0.699	0.16	0.475	0.95	0.002
	rs8022206	RAD51L1	14	68520906	1.55E-10	0.335	0.58	0.475	0.964	0.048
	rs8006385	ITPK1	14	93501026	1.24E-10	0.403	0.065	0.834	0.277	0.262

	rs7149242	C14orf70DLK1	14	101159416	2.68E-08	0.067	**4.90E-04**	0.058	0.715	**1.93E-05**
	rs11628318	RCOR1	14	103040087	2.04E-10	0.163	0.904	0.432	0.979	0.067
	rs2297067	C14orf73	14	103566785	1.58E-10	0.514	0.596	0.035	0.113	0.023
	rs55707100	MAP1A	15	43820717	3.77E-15	0.533	0.643	0.97	0.689	0.028
	rs3809566	TPM1	15	63333724	3.65E-10	0.727	0.486	0.089	0.103	0.158
	rs1719271	ANKDD1A	15	65183801	1.05E-11	0.934	0.401	0.606	0.695	0.934
	rs10852932	SMG6	17	2143460	2.15E-10	0.933	0.236	0.699	0.929	0.206
	rs6065	GP1BA	17	4836381	2.92E-11	0.791	0.295	0.689	0.456	0.791
	rs397969	AKAP10	17	19804247	2.32E-09	0.816	0.669	0.707	0.234	0.816
	rs559972	TAOK1	17	27814496	3.30E-218	0.260	0.002	0.098	0.808	**6.30E-06**
	rs10512472	SNORD7-AP2B1	17	33884804	2.40E-14	0.857	**4.00E-04**	0.236	0.038	0.011
	rs708382	FAM171A2-ITGA2B	17	42442344	1.51E-08	0.670	0.49	0.313	0.899	0.323
	rs76066357	ITGA2B	17	42463054	5.78E-19	0.177	0.407	0.376	0.549	0.958
	rs1801689	APOH	17	64210580	1.57E-18	0.197	0.198	0.807	0.29	0.911
	rs11082304	CABLES1	18	120720973	5.27E-11	0.509	0.355	0.493	0.775	0.351
	rs8109288	TPM4	19	16185559	2.75E-10	0.712	0.586	0.774	0.222	0.431
	rs892055	RASGRP4	19	38912764	5.30E-10	0.41	0.88	0.246	0.842	0.808
	rs17356664	EXOC3L2	19	45740771	3.60E-10	0.645	0.288	0.199	0.321	0.033
	rs3865444	CD33	19	51727962	2.59E-09	0.462	0.855	0.994	0.265	0.065
	rs6136489	SIRPA	20	1923734	8.78E-14	0.275	0.964	0.662	0.405	0.256
	rs1034566	ARVCF	22	19984277	3.06E-08	0.412	0.463	0.060	0.296	0.209
	rs855791	TMPRSS6	22	37462936	2.97E-12	0.562	0.581	0.540	0.943	0.324
	rs1018448	ARFGAP3	22	43206950	4.02E-10	0.562	0.581	0.539	0.943	0.324
	rs738409	PNPLA3	22	44324727	9.73E-21	0.536	0.644	0.537	0.899	0.508

5.4.2 eQTL effects for significant SNPs

Whole blood cis and trans eQTL analysis was performed for the top significant SNPs per region identified in the GWAS for PLR (in **Table 3**) and blood cell counts (in **Table 4**), with and without correcting for blood composition. The eQTL results are shown in **Table 6**. Information on the function of the genes and the involved pathways was retrieved from the GeneCards website (http://www.genecards.org).

Cis effects were found for rs8081692: it increases *GSDMB, MSL1* and *KRT23* gene expression and decreases *GSDMA* expression. However, after blood components correction, only *GSDMA* gene expression was upregulated by rs8081692. The locus rs169738 was found to increase *HLA-DPB1* and decrease *TAPBP* and *HLA-DPA1* expression, also after correcting for blood composition. For rs9376090, we detected a significant negative association with ALDH8A1 gene expression, but this SNP is not in LD with the top SNP for *ALDH8A1* gene rs4646871.

Trans effects for both rs9376090 and rs9376092 were found to increase *TMEM158* and *HBE1* gene expression, and while the trans effects were alleviated when correcting for blood composition, they remained significant. In addition, some eQTLs for genes involved in platelet activation, signaling and aggregation pathways, were present for the uncorrected expression results but disappeared when correcting for blood composition: *GNAS* (for rs9376090), *AQP9* and *CREB5* (for rs8081692). The top SNP rs11925835 nearby *ARHGEF3* gene was found to regulate several sets of genes involved in: 1) platelet activation, signaling and aggregation (*ITGB3, PPBP, ITGA2B, PF4, GP1BA, PRKAR2B, C6orf25, SELP, THBS1, GNG11, CLU, SPARC, F13A1, VCL, EHD3, CD9, PDGFA, MGLL, GUCY1A3, TBXA2R, MMRN1)*; 2) immune system (*TREML1, CD9, CD226)* ; and 3) metabolism (*PTGS1, VS1G2, EVOVL7, MGLL, ALOX12, MFAP3L,* and *NDUFAF3)*. In addition to these genes, there were several eQTLs for genes that regulate cell division, proliferation, and differentiation such as *ABL1M3, LMSM1, c7orf41, FHL1, MAX, RSU1, TSPAN9* and *MTPN.* Furthermore, some genes play a key role in hematopoietic stem cell differentiation pathways and lineage-specific markers, such as *PEAR1* and *CD226*. For the majority of these genes the effect was alleviated after correction for blood composition. Some trans effects were no longer present after the correction such as the effects for *TPM1, EHD3, PDLIM1, MGLL, LMNA, SLA2,ELOVL7, MGLL , TBXA2R, RSU1, MFAP3L, NEXN, CMTM5, ALOX12, PGRMC1, SEPT5, CDK2AP1, CD226, NDUFAF3, MMRN1, TSPAN9,* and *MTP.*

Table 6. The eQTL analysis results: the association between genetic variants of interest with uncorrected gene expression level and corrected gene expression level in blood.

Table 6A. Cis effects.

GWA SNP	Gene	β	FDR	Top SNP associated with the gene expression	Top SNP FDR	LD r^2	β after correction cell counts	FDR after correction cell counts
rs9376090	ALDH8A1	-.181	1.34e-5	rs4646871	1.34e-5	.7	-.173	1.34e-5
rs8081692	**GSDMB**	.126	1.34e-5	rs11557467	1.34e-5	1	.152	1.34e-5
rs8081692	MSL1	.094	.0015	rs17678928	1.34e-5	1		NS
rs8081692	GSDMA	-.101	.001	rs60701125	1.34e-5	1		NS
rs8081692	KRT23	.083	.0087	rs2051808	1.34e-5	1		NS
rs169738	**HLA-DPB1**	.204	1.34e-5	rs115378869	1.34e-5	1	.199	1.34e-5
rs169738	**TAPBP**	-.078	.0008	rs114742850	1.34e-5	1	-.076	.0013
rs169738	**HLA-DPA1**	-.079	.0159	rs115162296	1.34e-5	1	-.089	6.53E-05

Table 6B. Trans effects.

GWA SNP	Gene	β	FDR	Top SNP associated with the gene expression	Top SNP FDR	LD r^2	β after correction cell counts	FDR after correction cell counts
rs9376092	**TMEM158**	.278	<1.2e-4	rs7776054	<1.2e-4	.92	.245	<1.34e-5
rs9376092	**HBE1***	.15	.0003	rs9399136	<1.2e-4	.96	.145	.0023
rs9376090	**TMEM158**	.289	<1.2e-4	rs7776054	<1.2e-4	.95	.257	<1.34e-5
rs9376090	**HBE1***	.155	.0003	rs9399136	<1.2e-4	.98	.148	.0023
rs9376090	GNAS*	.099	.0223	rs79007502	<1.2e-4	1		NS
rs8081692	AQP9*	.122	.0205	rs7221894	.0048	.98		NS
rs8081692	CREB5*	.118	.0442	rs12941811	.0182	.82		NS
rs11925835	**ITGB3***	-.332	<1.2e-4	rs1354034	<1.2e-4	1	-.147	<1.34e-5
rs11925835	**PPBP***	-.335	<1.2e-4	rs1354034	<1.2e-4	1	-.283	<1.34e-5
rs11925835	**ITGA2B***	-.287	<1.2e-4	rs1354034	<1.2e-4	1	-.247	<1.34e-5
rs11925835	**CALD1**	-.281	<1.2e-4	rs1354034	<1.2e-4	1	-.241	<1.34e-5
rs11925835	**PF4***	-.271	<1.2e-4	rs1354034	<1.2e-4	1	-.224	<1.34e-5
rs11925835	**GP1BA***	-.264	<1.2e-4	rs1354034	<1.2e-4	1	-.234	<1.34e-5
rs11925835	**LTBP1**	-.260	<1.2e-4	rs1354034	<1.2e-4	1	-.223	<1.34e-5
rs11925835	**PTGS1**	-.255	<1.2e-4	rs1354034	<1.2e-4	1	-.208	<1.34e-5
rs11925835	**ABLIM3**	-.271	<1.2e-4	rs1354034	<1.2e-4	1	-.224	<1.34e-5
rs11925835	**PRKAR2B***	-.241	<1.2e-4	rs1354034	<1.2e-4	1	-.196	<1.34e-5
rs11925835	**C6orf25***	-.251	<1.2e-4	rs1354034	<1.2e-4	1	-.200	<1.34e-5
rs11925835	**MYL9**	-.247	<1.2e-4	rs1354034	<1.2e-4	1	-.203	<1.34e-5
rs11925835	**SELP***	-.250	<1.2e-4	rs1354034	<1.2e-4	1	-.210	<1.34e-5
rs11925835	**SH3BGRL2**	-.237	<1.2e-4	rs1354034	<1.2e-4	1	-.188	<1.34e-5
rs11925835	**THBS1***	-.223	<1.2e-4	rs1354034	<1.2e-4	1	-.181	<1.34e-5

rs11925835	**LIMS1**	-.191	<1.2e-4	rs1354034	<1.2e-4	1	-.160	<1.34e-5
rs11925835	**CTTN**	-.230	<1.2e-4	rs1354034	<1.2e-4	1	-.187	<1.34e-5
rs11925835	**GNG11**	-.233	<1.2e-4	rs1354034	<1.2e-4	1	-.186	<1.34e-5
rs11925835	**TSPAN33**	-.210	<1.2e-4	rs1354034	<1.2e-4	1	-.159	<1.34e-5
rs11925835	**TUBB1***	-.218	<1.2e-4	rs1354034	<1.2e-4	1	-.170	<1.34e-5
rs11925835	**CLU***	-.214	<1.2e-4	rs1354034	<1.2e-4	1	-.164	<1.34e-5
rs11925835	**SPARC***	-.214	<1.2e-4	rs1354034	<1.2e-4	1	-.133	.0001
rs11925835	**F13A1***	-.202	<1.2e-4	rs1354034	<1.2e-4	1	-.167	<1.34e-5
rs11925835	**JAM3**	-.223	<1.2e-4	rs3811444	<1.2e-4	1	-.173	<1.34e-5
rs11925835	**TREML1**	-.195	<1.2e-4	rs1354034	<1.2e-4	1	-.153	<1.34e-5
rs11925835	**VSIG2**	-.207	<1.2e-4	rs1354034	<1.2e-4	1	-.171	<1.34e-5
rs11925835	**LGALSL**	-.200	<1.2e-4	rs1354034	<1.2e-4	1	-.173	<1.34e-5
rs11925835	**VCL***	-.191	<1.2e-4	rs1354034	<1.2e-4	1	-.121	<1.34e-5
rs11925835	**NRGN**	-.181	<1.2e-4	rs1354034	<1.2e-4	1	-.144	<1.34e-5
rs11925835	**RAB27B**	-.204	<1.2e-4	rs1354034	<1.2e-4	1	-.156	<1.34e-5
rs11925835	**CTDSPL**	-.213	<1.2e-4	rs1354034	<1.2e-4	1	-.166	<1.34e-5
rs11925835	**SDPR**	-.175	<1.2e-4	rs1354034	<1.2e-4	1	-.133	<1.34e-5
rs11925835	**GP1BB***	-.173	<1.2e-4	rs1354034	<1.2e-4	1	-.129	<1.34e-5
rs11925835	**TTC7B**	-.181	<1.2e-4	rs1354034	<1.2e-4	1	-.161	<1.34e-5
rs11925835	**GNAZ**	-.178	<1.2e-4	rs1354034	<1.2e-4	1	-.132	.0003
rs11925835	**TUBA1C**	-.143	<1.2e-4	rs1354034	<1.2e-4	1	-.122	<1.34e-5
rs11925835	**C7orf41**	-.158	<1.2e-4	rs1354034	<1.2e-4	1	-.127	.0009
rs11925835	**FHL1**	-.171	<1.2e-4	rs1354034	<1.2e-4	1	-.126	.0073
rs11925835	TPM1	-.157	<1.2e-4	rs1354034	<1.2e-4	1		NS
rs11925835	**PCSK6**	-.171	<1.2e-4	rs1354034	<1.2e-4	1	-.135	.0006
rs11925835	**C12orf39**	-.163	<1.2e-4	rs1354034	<1.2e-4	1	-.135	.0009
rs11925835	PDLIM1	-.147	<1.2e-4	rs1354034	<1.2e-4	1		NS

rs11925835	**PEAR1**	-.159	<1.2e-4	rs1354034	<1.2e-4	1	-.123	.0137
rs11925835	**ILK**	-.138	<1.2e-4	rs1354034	<1.2e-4	1	-.123	<1.34e-5
rs11925835	EHD3*	-.133	<1.2e-4	rs1354034	<1.2e-4	1		NS
rs11925835	**CD9***	-.144	<1.2e-4	rs1354034	<1.2e-4	1	-.143	.0001
rs11925835	**PDGFA***	-.152	<1.2e-4	rs1354034	<1.2e-4	1	-.122	.0185
rs11925835	**MAX**	-.107	<1.2e-4	rs1354034	<1.2e-4	1	-.089	.0205
rs11925835	LMNA	-.134	.0004	rs1354034	<1.2e-4	1		NS
rs11925835	SLA2	-.139	.0004	rs1354034	<1.2e-4	1		NS
rs11925835	ELOVL7*	-.138	.0004	rs1354034	<1.2e-4	1		NS
rs11925835	MGLL*	-.137	.0004	rs1354034	<1.2e-4	1		NS
rs11925835	**GUCY1A3***	-.142	.0004	rs1354034	<1.2e-4	1	-.124	.0246
rs11925835	TBXA2R	-.134	.0004	rs1354034	<1.2e-4	1		NS
rs11925835	RSU1	-.096	.0016	rs1354034	<1.2e-4	1		NS
rs11925835	MFAP3L	-.129	.0019	rs1354034	<1.2e-4	1		NS
rs11925835	NEXN	-.106	.0021	rs1354034	<1.2e-4	1		NS
rs11925835	CMTM5	-.122	.0023	rs1354034	<1.2e-4	1		NS
rs11925835	ALOX12	-.13	.0029	rs1354034	<1.2e-4	1		NS
rs11925835	PGRMC1	-.101	.0052	rs1354034	<1.2e-4	1		NS
rs11925835	SEPT5	-.116	.0055	rs1354034	<1.2e-4	1		NS
rs11925835	CDK2AP1	-.111	.0057	rs1354034	<1.2e-4	1		NS
rs11925835	CD226	-.111	.0078	rs1354034	<1.2e-4	1		NS
rs11925835	NDUFAF3	-.116	.0127	rs1354034	.0028	1		NS
rs11925835	MMRN1*	-.122	.0217	rs1354034	.0003	1		NS
rs11925835	TSPAN9	-.120	.0321	rs1354034	.0003	1		NS
rs11925835	MTPN	-.103	.0442	rs1037590	.0047	1		NS

*: Platelet activation, signaling and aggregation. LD r^2: LD between SNP and top SNP. FDR: False Discovery Rate. NS: Not Significant.

5.4.3 SNP heritability and correlations among phenotypes

The SNP heritability of NLR and PLR was estimated at 2.4% (se = 0.082) and 14.1% (se = 0.084) respectively using LD regression (**Table 7**). With GCTA the estimated variance explained by the known loci from literature was 0.5% (se = 0.300) for NLR and 3.28% (se = 0.700) for PLR within our study. Finally, the significant SNPs for PLR, the single *SHB1L-MYB* region found in our study, explained 0.50% (se = 0.600) of variance.

Table 7. LD regression results of NLR and PLR and the blood cell counts.

	NLR	PLR	Neutrophil	Lymphocyt	Platelet
Median of Signed_sumstatistic	-3.66E-5	-0.006	-2.12E-5	-6.58E-5	-0.003
Mean of X^2	0.997	1.014	1.009	1.021	1.034
λ GC	0.996	1.024	1.017	1.018	1.027
H^2 (se)	0.024 (0.082)	0.141 (0.084)	0.154 (0.101)	0.191 (0.090)	0.362 (0.088)
Intercept (se)	0.994 (0.007)	0.999 (0.006)	0.993 (0.008)	1.001 (0.007)	0.997 (0.007)

Median of Signed_sumstatistic: median value of beta values from GWAS.

Significant positive phenotypic correlations were observed between NLR and PLR (r=0.491, p < 0.0001), between neutrophil count and NLR (r=0.651, p < 0.0001), and between platelet count and PLR (r=0.478, p < 0.0001). Significant negative phenotypic correlations were observed between NLR and lymphocyte count (r=-0.481, p < 0.0001), between neutrophil count and PLR (r=-0.073, p < 0.0001), and between PLR and lymphocyte count (r=-0.678, p < 0.0001). Significant and nearly significant genetic correlations were found between PLR and platelet count (r=0.457, p = 0.031) and between PLR and lymphocyte count (r=-0.486, p=0.070) (**Table 8**). The other genetic correlations were non-significant: NLR with neutrophil count (r=0.223, p=0.850) and NLR with lymphocyte count (r=0.153, p=0.882).

Table 8. Phenotypic correlations and genetic correlations of NLR and PLR levels and blood cell counts.

Phenotypes	Phenotype correlation estimated in R				Genetic correlation estimated by LD regression			
	NLR		PLR		NLR		PLR	
	r	p	r	p	r_g (SE)	p	r_g (SE)	p
PLR	.491	<.0001			0	1		
Neutrophil	.651	<.0001	-.073	<.0001	.223(1.184)	.850	0	1
Lymphocyte	-.481	<.0001	-.678	<.0001	-.153(1.024)	.882	-.486(.268)	.070
Platelet	.018	.174	.478	<.0001	0	1	.457(.212)	.031

5.5 Discussion

We studied the genetic architecture of NLR and PLR as well as the genetic relationship between NLR, PLR, and the corresponding immune cell counts. The intergenic *HBS1L-MYB* region is a well-known locus for hematological parameters such as red blood cell count [220], platelet count, hemoglobin level [221], MCHC level [222], and blood related diseases such as myeloproliferative neoplasms [223], beta-thalassemia [224] and sickle cell anaemia [207]. We found this intergenic *HBS1L-MYB* region to be significantly associated with PLR. *HBS1L-MYB* intergenic variants reduce the transcription factor binding and affect long-range interactions with *MYB* and *MYB* expression levels [225]. This region was first identified as a quantitative trait locus (QTL) controlling fetal hemoglobin level and is associated with iron deficiency anemia, beta-thalassemia, and sickle cell disease [226-227]. The *MYB* gene encodes a protein with three HTH DNA-binding domains that functions as a transcription regulator. This protein plays an essential role in the regulation of hematopoiesis and lymphocyte differentiation. This gene can be aberrantly expressed, rearranged or undergo translocation in leukemia's and lymphomas, and is thus considered to be a (proto-)oncogene [228-230]. The *HBSIL* (Hsp70 subfamily B suppressor 1-like) gene encodes a member of the GTP-binding elongation factor family. A single nucleotide polymorphism in exon 1 of *HBS1L* gene is significantly associated with severity in beta-thalassemia/hemoglobin E as found in a sequencing study [231] and verified in several other studies [232-233]. Recently, this gene has been associated with several traits, including erythrocyte and platelet count [188, 207, 234] and cholesterol level [235]. A pleiotropic association study on a wide number of hematological traits found that rs9373124, also in the *HBS1L-MYB* region, was significantly associated with all of the evaluated hematological traits ($p<0.005$) including white blood cell count, red blood cell count and platelet count [60].

The eQTL results show that some of the GWAS top SNPs for PLR and blood cell counts regulate the expression of genes which are mainly involved in immune system pathways: platelet activation, signaling and aggregation; metabolism; cell division, proliferation, and differentiation; and hematopoietic stem cell differentiation pathways and lineage-specific markers. These results provide new genetic targets for immune biomarkers and inform future functional studies. In our GWAS study, SNPs with significant associations for NLR were not identified, which is consistent with the small SNP heritability found with LD regression analyses. Compared to PLR, NLR shows more phenotypic plasticity, because neutrophils are part of the immune response to viral infections, autoimmune diseases, acute-phase reactions and several drugs [236]. Furthermore, compared to the longer lifespan of platelets (8-9 days), the life span of neutrophils is shorter (a few hours to max 5 days) [237-238]. The phenotype is therefore much more dependent on environmental effects, e.g. the time of measurement and health state of the individual also indicated by our own heritability findings [187]. By selecting only healthy individuals, we may have reduced the genetic variation in neutrophil count as well as lymphocyte count. We examined the heritability of NLR in the full population, not excluding anybody based on potential immune disease. There was no significant difference in heritability compared to our healthy sample. The point estimates were somewhat higher for the healthy population compared to the total population.

For both NLR and PLR a large part of the heritability is not explained by common SNPs or genetic variants in LD. This may suggest that other genetic variants, such as rare variants and copy number variants need to be studied. Furthermore, the missing heritability might be high because of non-additive effects and genetic interactions, which are not taken into account with the current applied statistical models. Epistatic effects of genetic variants for hematological indexes are already found [201, 239]. We thus assume that, especially for immune system phenotypes, gene-gene and gene-environment interactions need to be studied further.

The LD regression results show that polygenic effects, rather than confounding factors explain NLR and PLR variance in our study. We also demonstrated significant genetic correlation between PLR and platelet count, but none of the other correlations between ratios and cell counts were large enough to be significant. Since we found no SNP effects on NLR, it is not surprising that no genetic overlap between NLR and PLR is detected, although the genetic background of the lymphocyte count is expected to be affecting both ratios.

In summary, our study found the *HBS1L-MYB* locus to be associated with PLR level and with platelet count. In addition, we verified 3 additional known loci for platelet count(rs342213 in *CCDC71L-PIK3CG*, rs169738 nearby *BAK1* and rs11925835 nearby *ARHGEF3)* and one locus for neutrophil count (rs8081692 nearby *PSMD3)*. We did not identify any locus or any significant SNP heritability for NLR. Although NLR and PLR are both utilized as predictive or prognostic biomarkers for the same diseases, and phenotypic correlations are present, there seems to be no genetic overlap between the two biomarkers in our healthy population. The NLR and PLR responses associated with same disorders, thus likely represent the simultaneous influence of separate and multiple immune genetic pathways.

Appendix V.

Details on eQTL analysis

Expression Data

The two parent projects that supplied data for the eQTL analysis are large-scale longitudinal studies: the Netherlands Study of Depression and Anxiety (NESDA) [240] and the Netherlands Twin Registry (NTR) [93]. NESDA and NTR studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam (institutional review board [241] number IRB-2991 under Federal wide Assurance 3703; IRB/institute codes: NESDA 03-183 and NTR 03-180). All participants provided written informed consent.

Subjects for eQTL analysis

The sample used for eQTL analysis consisted of 4,896 subjects with European ancestry (1,880 unrelated subjects from NESDA, 559 MZ twin pairs, 102 siblings of MZ twins (one per MZ twin pair), 594 DZ twin pairs , 111 siblings of DZ twins (one per DZ twin pair), 51 parent-sibling trios and 344 unrelated subjects from NTR). The age of the participants ranged from 17 to 88 years (mean=38, SD=13); 65% of the sample was female.

Blood sampling, RNA extraction, and RNA expression measurement

Study protocols and biological sample collection methods were harmonized between NTR and NESDA. RNA processing and measurements have been described in detail previously [242-243]. Venous blood samples were drawn in the morning after an overnight fast. Heparinized whole blood samples were transferred within 20 minutes of sampling into PAXgene Blood RNA tubes (Qiagen, Valencia, California, USA) and stored at −20°C. Gene expression assays were conducted at the Rutgers University Cell and DNA Repository. Samples were hybridized to Affymetrix U219 arrays (Affymetrix, Santa Clara, CA) containing 530,467 probes summarized in 49,293 probe sets. Array hybridization, washing, staining, and scanning were carried out in an Affymetrix GeneTitan System per the manufacturer's protocol. Gene expression data were required to pass standard Affymetrix QC metrics (Affymetrix expression console) before further analysis. We excluded from further analysis probes that did not map uniquely to the hg19 (Genome Reference Consortium Human Build 37) reference genome sequence, as well as probes targeting a messenger RNA (mRNA) molecule resulting from transcription of a DNA sequence containing a single nucleotide polymorphism (based on the dbSNP137 common database). After this filtering step, data for analysis remained for 423,201 probes, which could be summarized into 44,241 probe sets targeting 18,238 genes. Normalized probe set expression values were obtained using Robust Multi array Average (RMA) normalization as implemented in the Affymetrix Power Tools software (APT, version 1.12.0, Affymetrix). Data for samples that displayed a low average Pearson correlation with the probe set expression values of other samples, and samples with incorrect sex-chromosome expression were removed, leaving 4,896 subjects for analysis.

Gene expression normalization

Inverse quantile normal transformation was applied for each expression probe set to obtain normal distributions. The transformed probeset data were then residualized by multiple linear regression with respect to the covariates sex, age, body mass index (kg/m2), blood hemoglobin level, smoking status, several technical covariates (plate, well, hour of blood sampling, lab, days between blood sampling and RNA extraction and average correlation with other samples) and the scores on three principal components (PCs) as estimated from the imputed SNP genotype data [103] using the EIGENSOFT package. The residuals resulting from the linear regression analysis of the probe set intensity values onto the covariates listed above were subjected to a principal component analysis, with the aim to further filter out environmental variation from the data [244]. For each principal component a genome wide association study was performed, and the first 50 principal components without genome wide significant SNP associations were removed from the residualized probeset data before eQTL analysis.

DNA extraction and SNP genotyping and imputation

DNA was extracted from peripheral blood or buccal swabs as has described previously [131]. SNP genotype pre imputation quality control, haplotype phasing and 1000 Genomes imputation were performed as described previously [245]. Imputed SNP genotypes were coded into reference allele dosage format, and filtered at MAF>0.01 and HW P>1E−04 resulting in 8,158,830 remaining SNPs for eQTL analysis.

eQTL analysis and FDR based on permutations accounting for relatedness

eQTL effects were detected with a linear model approach using MatrixeQTL [218] with expression level as dependent variable and SNP genotype values as independent variable. To account for relatedness of the NTR subjects, permutations were performed where in each permutation the relatedness was preserved (i.e, in each permutation the genotypes of the MZ twin pairs were assigned the expression of a random MZ twin pair, the genotypes of the DZ twin pairs were assigned the expression of a random DZ twin pair, the genotypes of the MZ twin pairs with sibling were assigned the expression of a random MZ twin pair with sibling, the genotypes of the parent-sibling trios were assigned the expression of a random parent-sibling trios and the genotypes of the unrelated subjects were assigned the expression of a random subject from the group of unrelated subjects). For each permutation the complete cis or trans eQTL analysis was repeated, and after each permutation the P-value threshold for rejecting at FDR<0.05 was computed. This can be done in 2 ways: 1) divide the total number of significant eQTLs in the permuted data by the total number of significant eQTLs in the unpermuted data (=false positives/true positives) or 2) divide the total number of probesets with a significant eQTL in the permuted data by the total number of probesets with a significant eQTLs in the unpermuted data. We used the the second method which is more conservative and was proposed by [244] to account for large LD blocks with strong eQTL effects that inflate the FDR when using the first method. Similar as what was observed in Fehrman et al, only 10 permutations were needed to have the P-value threshold corresponding to FDR<5% converging. Of note, the eQTL P-values reported in this manuscript are based on the complete sample with related subject and thus are too liberal: however the FDR takes into account the family structure and should be used to draw conclusions. The reported betas from the linear models can be correctly estimated from samples containing related subjects.

eQTL effects were defined as cis when probe set–SNP pairs were at distance < 1M base pairs (Mb), and as trans when the SNP and the probe set were separated by more than 1 Mb on the genome according to hg19. For each probe set that displayed a statistically significant association with at least one SNP in the cis region, we identified the most significantly associated SNP (E1 SNP). Conditional eQTL analysis was carried out by first residualizing probeset expression using the corresponding E1 SNP and than repeating the eQTL analysis using the residualized data. Then, for each probe set the most significant SNP was selected (E2 SNP) and each probeset was residualized using the E1 and E2 SNPs, and eQTL analysis was repeated using the residualized expression. This was repeated until no more significant associations were found between residualized expression and SNP data (after up to 12 rounds of conditional analysis). For the trans eQTL analysis the expression was first residualized using the E1,.., and E12 cis SNPs. For each probeset with significant trans-eQTL a similar conditional analysis was performed by conditioning the gene expression on the strongest corresponding trans eQTL.

Blood composition trait estimates

For 2958 subjects (60% of the complete sample), mean corpuscular volume, red cell distribution width, and neutrophile, lymphocytre, monocyte, eosinophile, basophil and platelet counts were available. Using partial least squares regression, estimates of these blood trait were computed for the samples without blood composition measurements (using the R function plsr (50 components) from the R library pls) based on the gene expression measures. To verify how well the estimates are, first two third of the samples having these blood cell traits were used to compute the estimates for the other third of the sample having the blood cell traits, in order to compute the correlation between the estimates and the actual blood composition traits. Spearman rank correlations were >0.75 for neutrophile, lymphocyte, monocyte and eosinophile counts, and >0.3 for the other traits except for basophil count (rho=0.129). Basophil counts were mostly zero, not associated with gene expression and therefore excluded from further analysis. The final predictor was computed using all the samples for which the blood composition were measured. Gene expression was corrected for these blood composition traits and eQTL analysis was repeated.

Chapter 6

Heritability and GWA studies for the monocyte-lymphocyte ratio

This chapter is published as: Lin BD, Willemsen G, Fedko IO, R Jansen, BWJH Penninx, de Geus EJ, C Kluft, Hottenga JJ, Boomsma DI. Heritability and GWA studies for the monocyte-lymphocyte ratio. Twin Res Hum Genet. doi: 10.1017/thg.2017.3.

Abstract

Aim. The monocyte–lymphocyte ratio (MLR) may be a useful biomarker for disease development, but little is known about the way genetic and environmental factors influence MLR variation. Here we study the genetic architecture of MLR and determine the influence of demographic, lifestyle and environmental factors on MLR using data from a Dutch non-patient twin-family population. **Methods**. Using data obtained in 9501 individuals from the Netherlands Twin Register, regression analyses were applied to determine the effects of age, sex, weather conditions, smoking and body mass index (BMI) on MLR and its subcomponents. Data on twins, siblings and parents (N=7513) were used in structural equation modelling to establish heritability and whole genome wide analyses were conducted in a genotyped subsample (N=5892) to identify the genes involved. GCTA, LD regression and eQTL analyses were performed to further explore the significance and nature of the genetic findings. **Results.** Age, sex and age x sex interaction effects were present for MLR and its subcomponents. Taking these effects into account, heritability was estimated at 39.7% for MLR and at 58.3% for monocyte and 57.6% for lymphocyte count. GWAS identified a locus on *ITGA4* which was associated with both MLR and monocyte count. For monocyte count, additional genetic variants were identified on *ITPR3*, *LPAP1* and *IRF8*. For lymphocyte count, GWAS provided no significant findings. Taking all measured SNPs together, their effects accounted for 13% of the heritability of MLR, while all known and identified genetic loci explained 1.3% of variation in MLR. Additional eQTL analyses showed that these genetic variants are unlikely to be eQTLs. Variation in MLR was not related to weather conditions, nor to BMI, but smoking was positively associated with MLR.**Conclusion**. Variation in MLR level in the general population is heritable and influenced by age, sex and smoking. Though we identified one gene variant associated with variation in NLR, more research is needed to fully elucidate the genetic pathways.

Keywords: Heritability, MLR, GWAS, Age, Sex differences, Weather conditions, Lifestyle.

6.1 Introduction

As early as the 1960s the relevance of the monocyte-lymphocyte ratio (MLR) for disease prediction was pointed out. The first studies focused on infectious diseases, suggesting MLR to reflect the balance between effector and host [246] and found MLR to predict the development and progress of tuberculosis [247]. In later studies the association between MLR and other diseases was studied and the MLR emerged as a predictor for cancer [248]. While MLR is examined in relation to disease, studies examining the causes of variance in MLR in the general population are lacking. To fully understand the role of MLR in disease, it is however necessary to understand the factors underlying variation in MLR in the general non-patient population.

We recently showed heritability to play a role in individual differences in two other lymphoid ratios, the neutrophil-lymphocyte ratio (NLR) and the platelet-lymphocyte ratio (PLR). For NLR, the heritability was moderate (35%), but for PLR heritability was high (64%), with evidence for the presence of non-additive genetic effects [187]. A first GWAS study for these two rations identified a genome-wide locus on the *HBS1L-MYB* intergenic region associated with PLR (chapter 5), which has been be associated with multiple blood parameters, including platelet count [187].

Although the heritability of MLR has not been studied, genetic studies have been conducted for its subcomponents, the monocyte and lymphocyte levels, showing heritability estimates of 56-73% for monocyte level and 35-66% for lymphocyte level [51, 53]. In addition, linkage and GWAS have pointed to the genetic variants partly responsible for the individual variation in monocyte and lymphocyte levels. GWAS studies have identified eight loci associated with monocyte level: *ITGA4* at 2q31.3, *HLA-DRB1* at 6p21.32, *CCBP2* at 3p22.1, *RPN1* at 3q21.3, *LPAR1* at 9q31.3, intergenic regions at 8q24 and 3q21, and *IRF8* at 16q24.1 [192, 195, 200, 249]. GWAS also identified two loci for lymphocyte level: 6p21 (*EPS15L1* gene) and 19p13 [195, 249]. Taken together, genetic factors are likely to play a role in normal variation in MLR, but the nature of the involvement remains to be determined.

Environmental and lifestyle factors may also influence MLR levels. Here too studies on MLR itself are lacking, but our own study on PLR and NLR [187] showed that seasonal conditions influence PLR and NLR levels, and agreed with other studies [139-140] that smoking and BMI may also affect these parameters. In addition, its subcomponents, monocyte an lymphocyte level have been found to be influenced by BMI [250-251] and smoking [252], though these effects are not found in all studies [253-254].

In this paper we examine several potential causes of variance in MLR in the general population. First, we present a series of genetic studies on MLR to provide more insight into its genetic architecture. We started by estimating the heritability of MLR and it subcomponents by extended twin family modelling. Next, we used GWAS to identify genetic variants associated with MLR variation and GCTA to determine the percentage of variance of MLR that is explained by significant versus all measured genetic variants (single nucleotide polymorphisms, SNPs). Subsequently, we performed an eQTL analysis using all the top SNPs, which were significantly associated with MLR. We repeated the GWAS, GCTA and eQTL analyses for monocyte level, while referring for lymphocyte level to the results presented previously (chapter 5). Finally, LD-regression was performed using the summary statistics of the GWAS results to determine the polygenetic effects and genetic correlations between MLR and subcomponents.

After our examination of the role of genetic factors in the variance in MLR, we also investigated additional causes of individual differences in MLR. We first studied the effect of seasonal conditions such as outside temperature and next investigated the effect of lifestyle factors, specifically smoking and BMI, on MLR.

6.2 Methods

6.2.1 Participants

All participants were adults registered with the Netherlands Twin Register (NTR) who took part in a longitudinal study on health and lifestyle in twins and their family members [75]. Data were obtained as part of NTR biobanking projects conducted in 2004-2008 [76] and 2011 [77]. After removing outliers (i.e. absolute values exceeding mean ±5SD), data on MLR, monocyte count and lymphocyte count were available for 9501 participants clustered in 3412 families.

The following exclusion criteria were used to identify individuals who may have had a compromised immune system at the time of blood sampling: 1) illness reported in the week prior to sampling (N=552); 2) CRP ≥ 15 (N=307); 3) basophil count > $.02\times10^9$/L (N=154); 4) report of blood related disease or cancer (N=84); and 5) use of anti-inflammatory medication (N=423), glucocorticoids (N=143) or iron supplements (N=29). Participants meeting one or more of these criteria were labelled as unhealthy (N=1362), leaving 8139 individuals from 3280 families as the population which we will here refer to as the healthy population. Genetic twin-family modelling was conducted using data from twin families limited to at most one twin pair per family and at most two brothers and two sisters and father and mother. This resulted in a sample of 7513 participants from 3252 families, including 240 monozygotic male (MZM), 98 dizygotic same-sex male (DZM), 536 monozygotic female (MZF), 219 dizygotic same-sex female (DZF) and 222 dizygotic opposite-sex (DOS) twin pairs.

The study protocol was approved by the Medical Ethics Committee of the VU University Medical Center Amsterdam and all participants provided informed consent.

6.2.2 Procedure and measurements

Participants were visited at home, or when preferred at work, to obtain blood samples and conduct a brief health-related interview. Visits took place in the morning between 7 a.m. and 10:00 a.m. and women were seen, when possible, between the 2nd to the 4th day of the menstrual cycle or, when on hormonal birth control, were visited in their pill-free week. Participants were asked to remain fasting as of the evening before and to refrain from smoking or physical exercise one hour before blood sampling (for more details see [76]).

Health status

Participants were asked to indicate when they were last ill and the nature of the illness. In the case of medication use the dosage, brand and name were recorded. In addition, participants indicated on the presence and nature of any chronic disease.

Blood parameters.

Procedures have been described in detail elsewhere ([76] and chapter 4 in this thesis). In short, peripheral blood was collected in anticoagulant vacuum tubes, which were inverted 8-10 times immediately after the blood draw. All samples were transported to the laboratory facility in Leiden, the Netherlands, within 3 to 6 hours after blood sampling. The blood samples were then directly used or stored to measure parameters of interest or extract DNA or RNA at a later moment.

The haematological profile was obtained from blood collected in an EDTA tube using the Coulter system (Coulter Corporation Miami USA). This profile consisted of total white blood cell count, percentages and numbers of neutrophils, lymphocytes, monocytes, eosinophils, and basophils, and indicators of red blood cell types and platelets. We calculated MLR as the absolute monocyte count (10^9/L) divided by the absolute lymphocyte count (10^9/L). C-reactive protein (CRP) was determined from a heparin plasma sample using the 1000 CRP assay (Diagnostic Product Corporation).

Seasonal effects.

The information on daily weather conditions was obtained from the website of the Royal Netherlands Meteorological Institute (KNMI) [159]. We analyzed the daily data on temperature, wind speed, mean sea level, sunshine duration, global radiation and mean relative atmospheric humidity and potential evapotranspiration.

Lifestyle

During the interview conducted at the time of the home visit, height and weight were obtained. BMI was calculated as weight (kg) divided by height squared (m^2). Participants reported whether they currently smoked or had smoked. If so, they were asked for the number of cigarettes smoked per day and how long they (had) smoked. Based on their answers, participants were divided into 5 categories: non-smoker, ex-smoker, light smoker (currently smoking less than 10 cigarettes a day), average smoker (currently smoking 10 to 19 cigarettes a day), and heavy smoker (currently smoking 20 or more cigarettes a day).

6.2.3 Genotype Data

For DNA isolation we used the GENTRA Puregene DNA isolation kit on frozen whole blood samples, which were collected in EDTA tubes. All procedures were performed according to the manufacturer's protocols [131]. Genotyping was done on multiple chip platforms, including a number of partly overlapping subsets of participants. The following platforms were used: Affymetrix Perlegen 5.0, Illumina 370, Illumina 660, Illumina Omni Express 1 M and Affymetrix 6.0 (for details see chapter 4). 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"). Genotype calls were made with the platform specific software (BIRDSUITE APT-Genotyper Beadstudio) for each specific array. Phasing of all samples and imputing cross-missing platform SNPs was done with MACH 1 [132]. The phased data were then imputed with MINIMAC [96] in batches of around 500 individuals for the autosomal genome using the above 1000G Phase I integrated reference panel for 561 chromosome chunks obtained by the CHUNKCHROMOSOME program [97]. SNPs were removed if the Mendelian error rate >0.02, if the imputed allele frequency differed more than 0.15 from the 1000G reference allele frequency, if MAF < 0.01 and if R2 < 0.80. Hardy-Weinberg Equilibrium was calculated on the genotype probability counts for the full sample and SNPs were removed if the p-value < 0.00001. After imputation, the DNA confirmed MZ twins were re-duplicated back into the data. This left 6,010,458 SNPs in the GWAS analyses.

As several different platforms were used, additional SNP Quality Control (QC) included an evaluation of the SNP platform effects and SNPs showing platform effects were removed. This was done by defining individuals on a specific platform as cases 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-value < 0.00001. In total 5,987,253 SNPs survived this QC and these SNPs were then used to build the Genetic Relationship Matrix (GRM)s for all individuals. The selected SNPs were transformed to best guess Plink binary format and subsets were made for each of the 22 chromosomes. The GRMS for all NTR samples were then calculated using GCTA [6].

We generated 24 GRMs in total. A first autosomal GRM reflects an IBS matrix for all individuals. This GRM matrix is determined from all autosomal SNPs and is used to estimate the SNP heritability (h_g^2). A second autosomal GRM represents closely related individuals (IBS> 0.05), any remaining pairwise relationship estimates smaller than 0.05 were set to 0 in this matrix. This matrix is used as second covariate matrix in the GWAS and GCTA studies to account for the family structure of individuals and to estimate the narrow-sense heritability (h^2) applying an additive model. Finally 22 GRM matrixes were created that include all autosomal SNPs except for those on the one chromosome that is tested in the GWAS (the leave one chromosome out or LOCO strategy). These matrices were used in the GWAS study as a covariate matrix to remove artificial inflation due to all kinds of subsample stratification.

6.3 Analyses

6.3.1 Influence of health status, age and sex.

First, using age- and sex corrected values, we tested for differences in MRL, monocyte and lymphocyte level between the healthy and unhealthy population using a T-test. Next, within the healthy population, we explored the age and sex effects by linear regression. All these analyses were performed using STATA [160], using the cluster option to correct for the inclusion of family members.

6.3.2 Heritability estimation

Using structural equation modelling in OpenMx [98], the heritability of MLR, monocyte count and lymphocyte count was estimated in the healthy population. MLR, monocyte count, lymphocyte count, age were standardized using Z-scores. Parameters were estimated by maximum likelihood. First, we summarized the family resemblance with respect to MLR by means of correlations corrected for age, sex, and age x sex effects. Next, we fitted a series of genetic models. The total phenotypic variance was decomposed into four sources of variation: additive genetic (A) non-additive genetic (D), common environmental (C) and unique environmental (E) variation. The common environmental variance reflects the variance shared between siblings and twins (Vs). The resemblance among family members was modelled as a function of A, D and C. We allowed for a correlation in phenotype between spouses (μ). In fitting the genetic models we included as covariates age, sex, sex x age. We fitted the full model as described and tested the presence of assortative mating (i.e. the correlation between phenotypes of spouses) and the presence of shared environment and non-additive influences. The nested sub-models were compared to the full model by log likelihood ratio test (-2LL) using a significance level of 0.05.

6.3.4 GWAS

We performed 2 GWASs: MLR level and monocyte count on the quality controlled imputed SNPs including age, sex, three Dutch PCs generated with the EIGENSOFT software and genotype platform as covariates (N=5892) [103]. As we already conducted a GWAS for lymphocyte count, using a largely overlapping sample (N=5901, overlap of 5890 individuals) we did not rerun this analysis but instead refer here to the results published in chapter 5. Analyses were performed with the GCTA software running a mixed linear model association (MLMA) model [215]. To avoid inflated test statistics in datasets with related individuals and other remaining cryptic stratification we used two covariate GRM matrixes: the matrix for all individuals excluding the chromosome under analysis (LOCO analysis) and the matrix only focusing related individuals with IBS>0.05 [215]. For the GWAS we assume the statistically significant threshold as p-value less than 5 x 10^{-8} [21], and we refer to marginally significant when p-values exceed this threshold but remain below 10^{-4}.

6.3.5 GCTA

We performed GCTA analyses to estimate narrow sense heritability, the fraction of genetic variance explained by the significant SNPs detected in the GWA and the fraction of genetic variance explained by the known significant SNPs from the published literature. These analyses were done for MLR level, monocyte count and lymphocyte count. A restricted maximum likelihood (REML) analysis procedure was used under a linear design [6]. Sex, age, genotype platform and three Dutch PCs were included as covariates. We used two covariance matrixes to estimate narrow sense heritability (h^2), and GWAS and known loci heritability. The first GRM matrix is the full autosomal GRM as described previously. The second GRM matrix is the closely related (IBS> 0.05) matrix.

6.3.6 Linkage Disequilibrium score regression

First, overall Pearson correlations between the phenotypes of interest were calculated in R [219]. Then polygenetic effects [209] the SNP heritability [255] of MLR, monocyte count and lymphocyte count; and genetic correlations [208] among the phenotypes of interest were determined using Linkage Disequilibrium (LD) regression on our computed GWAS summary statistics. The genetic correlation of two traits can be calculated by the slope from the LD regression on the product of effect sizes (z-score) for two phenotypes of interest. In order to do this we used the HapMap3 LD scores (NSNPs= 1293150) computed for each SNP based on the LD observed in European ancestry individuals from 1000 Genomes project (online accessible: http://github.com/bulik/ldsc). Quality control for genetic data is the default setting in the program.

6.3.7 eQTL analysis

To detect the casual effects for the genetic variants for phenotypes of interest, we conducted an eQTL analysis. Details are described elsewhere (chapter 5). In short, eQTL effects were detected with a linear model approach using MatrixeQTL [218] with expression level as dependent variable and SNP genotype values as independent variable. eQTL effects were defined as cis when the distance between probe set–SNP pairs was smaller than 1M base pairs (Mb), and as trans when the SNP and the probe set were separated by more than 1 Mb on the genome according to hg19.

Weather conditions and lifestyle.

To detect the influence of seasonal conditions on variation in the MLR level, we included mean temperature and other weather parameters in a regression analysis conducted separately by sex and taking age into account. Analyses were conducted in STATA [160], again using the cluster option to correct for the family structure within the data. In a similar manner, the association with BMI and smoking behavior was examined.

6.4 Results

6.4.1 Health status, sex and age

Table 1 provides the descriptive statistics for MLR and its subcomponents, the monocyte and lymphocyte count, for the healthy and for the unhealthy part of the population. The comparison of the healthy and unhealthy population, taking sex and age into account as well as family structure, showed as expected that individuals in the unhealthy population had on average a higher MLR ratio (t(9499)=-7.95, p < 0.001) and monocyte count (t(9499)=-5.06, p < 0.001) and a lower lymphocyte count (t(9499)=-2.57, P=0.01).

We continued our investigation in the healthy population, examining the influence of age and sex. Men had higher MLR levels than women and MLR increased with higher age in both men and women. There was also evidence for an age x sex interaction: the age effects were alleviated in the women. With respect to the subcomponents, monocyte and lymphocyte levels were higher in men than in women and increased with age. These age effects were similar in men and women.

Table 1. Average level (SD) for MLR and its subcomponents for the healthy and unhealthy population, separately for men and women.

	Healthy population		Unhealthy population	
	Men	Women	Men	Women
N	3074	5065	444	918
Age	44.12(15.88)	42.98(14.48)	46.75(17.37)	43.75(15.48)
MLR	0.29(0.09)	0.24(0.08)	0.32(0.14)	0.26(0.12)
Monocyte	0.58(0.17)	0.51(0.16)	0.61(0.19)	0.54(0.20)
Lymphocyte	2.17(0.64)	2.27(0.71)	2.07(0.75)	2.22(0.75)

Table 2. Age, sex and age x sex interaction effects (β value) on MLR and its subcomponents in the healthy population.

variable	MLR	monocyte count	lymphocyte count
Sex	**-0.0176****	**-0.0503*****	**0.0986***
Age	**0.0013*****	**0.0006****	**-0.0057*****
Age x Sex	**-0.0006*****	-0.0005	5.7E-5

* P<0.05, **P<0.01, ***P<0.001.

6.4.2 Heritability

Next, the known genetic relations among mono- and dizygotic twins and their family members were used to model familial resemblance in MLR, monocyte and lymphocyte count as a function of genetic and environmental parameters. These models included sex, age and sex x age effects as fixed effects. **Table 3** contains the familial correlations as obtained for MLR, monocyte and lymphocyte count. For MLR, twin pair correlations did not depend on sex, and the correlations did not differ across DZ twin and sibling relations. The correlations in MZ males and MZ female twin pairs were equal as were the other male and female first-degree relative correlations. The resulting MZ correlation was 0.43 (CI is 0.33-0.46) and the DZ correlation was 0.22 (0.14-0.24), with spousal correlations significant at 0.104 (0.002-0.135). The pattern of twin correlations showed no evidence for non-additive or common environmental effects. This was confirmed by model fitting in which the heritability of MLR was estimated at 40% (0.34-0.43).

We also conducted these series of genetic modelling analyses for monocyte and lymphocyte count. For monocyte count, there were no significant spousal correlations and the MZ correlation was 0.58 (0.54-0.62) while the DZ correlation was 0.27 (0.21-0.31). In line with the pattern of the correlations, genetic modelling estimated the broad sense heritability at 58% with non-additive effects accounting for 12% and no evidence for the influence of common environmental factors. For lymphocyte count, we estimated the heritability in the current set (N=5892, with > 99% overlap with the set described in chapter 4) and as to be expected results were similar to those published in chapter 4 with a broad sense heritability at 58% and non-additive effects accounting for 22%.

Table 3. Familial correlations (confidence interval) for MLR monocyte and lymphocyte count within the healthy population.

Pairs	MLR		Monocyte		Lymphocyte	
	R	95% CI	R	95% CI	R	95% CI
MZ twins	**0.431**	**0.330-0.463**	**0.583**	**0.539-0.622**	**0.582**	**0.537-0.621**
MZ male	0.340	0.329-0.344	0.515	0.432-0.585	0.576	0.487-0.646
MZ female	0.489	0.478-0.492	0.620	0.568-0.663	0.584	0.531-0.630
Male first-degree relatives	**0.182**	**0.082-0.205**	**0.247**	**0.183-0.308**	**0.234**	**0.156-0.307**
DZ male	0.244	0.121-0.248	0.170	-0.034-0.354	0.390	0.153-0.551
Brother-male twin	0.128	0.116-0.132	0.182	0.054-0.301	0.181	0.007-0.336
Brother-brother	0.242	0.231-0.246	0.346	0.082-0.545	0.341	0.007-0.580
Father-son	0.186	0.174-0.190	0.269	0.194-0.339	0.219	0.122-0.307
Female first- degree relatives	**0.224**	**0.139-0.239**	**0.228**	**0.186-0.268**	**0.225**	**0.183-0.266**
DZ female	0.279	0.267-0.282	0.385	0.154-0.279	0.286	0.181-0.382
Sister-female twin	0.169	0.157-0.173	0.285	0.156-0.390	0.175	0.008-0.265
Sister-sister	0.241	0.230-0.247	0.315	0.214-0.406	0.151	0.055-0.243
Mother-daughter	0.228	0.216-0.232	0.247	0.183-0.307	0.210	0.157-0.260
Female-male first degree relatives	**0.225**	**0.152-0.235**	**0.285**	**0.237-0.331**	**0.203**	**0.162-0.244**
DZ opposite sex	0.098	0.086-0.103	0.181	0.054-0.301	0.216	0.086-0.333
Brother-female twin	0.232	0.220-0.236	0.335	0.026-0.443	0.182	0.007-0.336
Sister-male twin	0.193	0.181-0.197	0.189	0.060-0.307	0.172	0.023-0.307
Sister-brother	0.200	0.188-0.204	0.227	0.080-0.352	0.312	0.178-0.419
Mother-son	0.217	0.205-0.221	0.240	0.168-0.308	0.198	0.115-0.276
Father-daughter	0.262	0.250-0.265	0.218	0.154-0.278	0.245	0.183-0.260
Parents (father-mother)	**0.104**	**0.002-0.135**	**0.061**	**-0.013-0.135**	**0.166**	**0.089-0.241**
Narrow-sense Heritability(V_A)	**0.397**	**0.341-0.429**	**0.468**	**0.404-0.530**	**0.353**	**0.294-0.402**
Broad-sense Heritability(V_A+V_D)			**0.583**	**0.446-0.720**	**0.576**	**0.461-0.653**

Correlations in bold italic were obtained from sub-models in which all matching correlations of the tested subgroup of family relations were set to be equal.

6.4.5 GWAS

Figures 1 and 2 show the results of the GWAS in the form of the QQ and Manhattan plots for MLR and monocytes. After adjusting for age, sex, genotype platform, PCs and using the LOCO and family based GRM matrix correction, the GWAS λs were 0.9965 for MLR and 1.0166 for monocyte count.

For MLR, associations were found with 11 SNPs situated on the *ITGA4* (VLA-4 Subunit Alpha) genes on chromosome 2q31 (**Figure 1** and **Table 4**). The top SNP rs3755021 T allele was linked to a decrease in MLR level (β=-0.012, p=2.21 × 10^{-8}). This SNP was not associated with lymphocyte count, but was in our study marginally significantly associated with monocyte count (β=-0.018, p= 6.34× 10^{-6}), and has also been associated with monocyte count in a linkage study [256] and two previous GWA studies [195, 249]. The genetic variant rs6740847 G allele in this region has been linked to decreased *ITGA4* expression levels in the blood, which increases the number of circulating monocytes and may indicate this is a causal gene [256].

For monocyte count, the four top hits were rs13029501 at *ITGA4,* rs55929401 located at a region nearby *LPAR1* at 9q31.3, rs391855 at *IRF8* and rs9469532 at 6p21. The most significant locus rs13029501 at 9q31 has been previously associated with monocyte count in European and Japanese populations [195, 256-258]. It is located in a region 163kb downstream of lysophosphatidic acid receptor 1 gene (*LPAR1*, also known as *EDG2*) and increases *LPAR1* expression, which is linked to an increased number of monocytes [256]. As indicated previously, genetic variants nearby the *ITGA4* region are involved in the down regulation of *ITGA4* expression, which increases the number of monocytes circulating in the peripheral blood. The *IRF8* gene has also been associated before with monocyte count and has been identified as multiple sclerosis susceptibility loci [259]. Animal model studies showed that *IRF8* as transcription factor plays an essential role in the regulation of lineage commitment during monocyte differentiation [260-262]. The top SNP at 6p21 rs9469532 is an intergenic genetic variant nearby *ITPR3*, *LOC101929188* and LOC105375023. The *HLA-DRB1* region 1,043kb upstream of this SNP has previously been associated with monocyte count [60].

As published in chapter 5, there were no significant hits when conducting the GWAS for lymphocyte count. However, it is of interest to note that the locus on chromosome 6p21, which was associated with monocyte level was also marginally associated with MLR (for rs9469532, β=-0.069, p= 7.69× 10^{-5}) and lymphocyte count (for rs114641912, β= -.059, p=6.19× 10^{-6}). This region harbours candidate genes like *ITPR3* [263], and *HLA-DRB1* [264] which have been previously implicated in immunological diseases. In addition, other loci with "potential association peaks", meaning p-values are low but do not reach the required significance level, have been found to be associated with immune disease such as *ERAP1* at 5q15 [265] and *CNTN5* at 11q22 [266].

Figure 1. Manhattan and QQ plot for MLR level with SNPs having a minor allele frequency above 0.01(λ=0.996503).

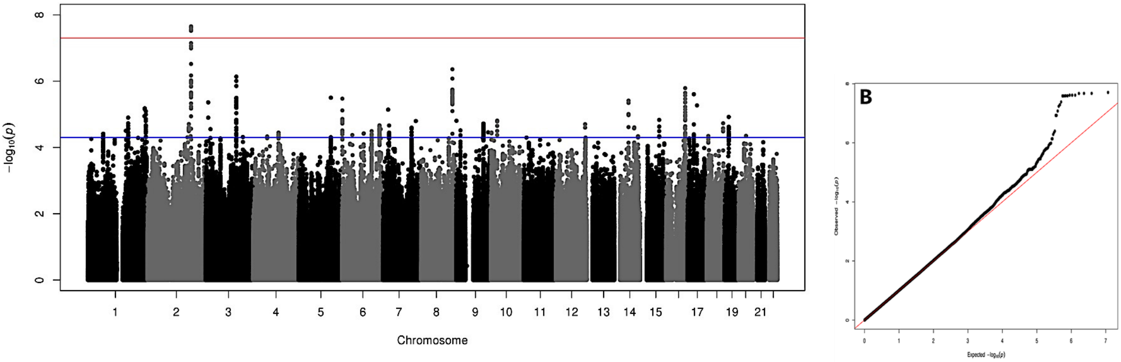


Figure 2. Manhattan and QQ plot for monocyte count with SNPs having a minor allele frequency above 0.01(λ=1.0166495).

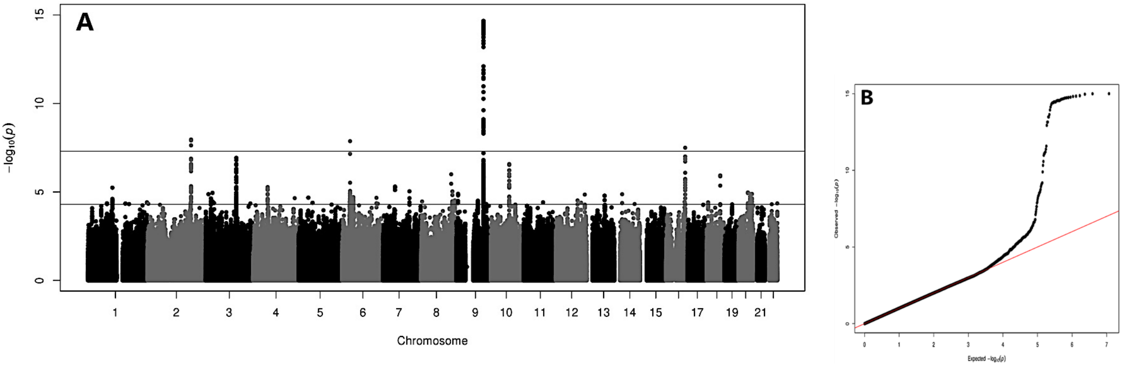


Figure 3. Manhattan and QQ plot for lymphocyte count with SNPs having a minor allele frequency above 0.01(λ=1.022341).

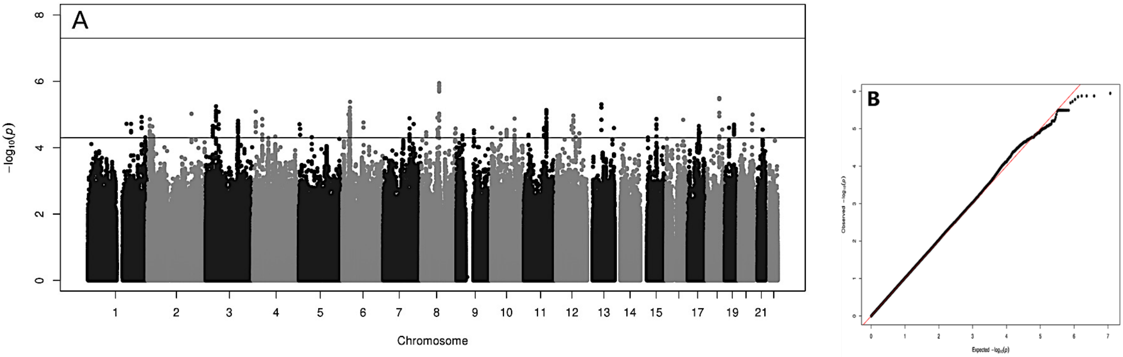


Table 4. Significant SNPs associations within our study for MLR, also including the p-values for the monocyte and lymphocyte counts.

RSNUMBER	CHR	BP	MAF	β (MLR)	SE (MLR)	P (MLR)	P(monocyte count)	P(lymphocyte count)
rs3755021	2	182349409	0.186439	-0.0119024	0.00212769	2.21E-08	6.34E-6	.067
rs17224699	2	182348554	0.186524	-0.0118630	0.00212596	2.40E-08	6.55E-6	.069
rs17290693	2	182350489	0.186354	-0.0118593	0.00212614	2.43E-08	5.63E-6	.074
rs79965377	2	182351172	0.186354	-0.0118593	0.00212614	2.43E-08	5.63E-6	.074
rs17290351	2	182345875	0.187712	-0.0117619	0.00211863	2.83E-08	8.91E-6	.067
rs17224524	2	182346146	0.187712	-0.0117619	0.00211863	2.83E-08	8.91E-6	.067
rs2305588	2	182347072	0.187712	-0.0117619	0.00211863	2.83E-08	8.91E-6	.067
NA	2	182348141	0.187797	-0.0117540	0.00212030	2.96E-08	1.05E-5	.064
rs12479308	2	182348342	0.187797	-0.0117540	0.00212030	2.96E-08	1.06E-5	.064
rs2305590	2	182346544	0.187627	-0.0117502	0.00212048	3.00E-08	9.12E-6	.068
rs2305589	2	182346937	0.187627	-0.0117502	0.00212048	3.00E-08	9.12E-6	.068

Table 5. Known blood cell count loci and their significance in our GWAS study.

Phenotype	Gene	SNP	chr	location	P	P(mlr)	P(mono)	P(lymp)
monocyte	ITGA4	**rs2124440**	2	182328214	4.63E-10	**2.21E-6**	**1.53E-7**	0.838
	ITGA4	**rs1449263**	2	182319301	6.71E-14	**2.40E-6**	**2.84E-7**	0.776
	RPN1	**rs2712381**	3	128338600	1.94E-10	0.005	**4.01E-4**	0.620
	C3orf27	**rs9880192**	3	128297569	1.35E-8	**5.97E-7**	**1.39E-7**	0.430
	CCBP2	rs2228467	3	42906116	1.57E-10	0.769	0.770	0.804
	CCBP2	rs2228468	3	42907112	5.15E-14	0.128	0.146	0.588
	Intergenic	rs2047076	5	76058509	1.64E-8	0.492	0.296	0.956
	HLA-DRB1	rs3095254	6	31221668	8.27E-9	0.910	0.198	0.276
	Intergenic	**rs1991866**	8	130624105	4.58E-11	**2.44E-6**	**2.08E-4**	0.171
	EDG2	**rs10980800**	9	113915905	1.1E-14	**2.23E-5**	**1.28E-14**	0.011
	PTGR1	rs2273788	9	114348617	4.50E-10	0.579	0.024	0.159
	LPAR1	**rs7023923**	9	113925534	8.9E-6	**4.26E-4**	**3.81E-12**	0.001
	IRF8	**rs424971**	16	85946450	3.16E-10	**2.1E-4**	**8.16E-6**	0.736
lymphocyte	LOC101929772	**rs2524079**	6	31242174	1.85E-8	0.0870	0.056	**3.02E-4**
	EPS15L1	rs11878602	19	16555153	3.42E-9	0.0735	0.857	0.107

Table 5 shows the loci for monocyte and lymphocyte count found in previous studies and their significance levels for MLR and its subcomponents in the current study. For some loci, p-values were low, indicating a "potential" for association, even though they did not reach the required significance level. For example, rs9880192 located in the intergenic region between *C3orf27* and rs1991866, a intergenic variant at 8q24.21, show p-values < 10^{-6} for monocyte level and < 10^{-3} for MLR.

6.4.6 eQTL results

Among the significant GWA loci for MLR and blood cell counts, there were a number of associations between the SNPs of interest and nearby gene expression (**Table 6**). However, the SNPs identified in our GWAS have low LD (r2 <0.8) with the top SNPs associated with gene expression, which suggest the GWAS SNPs are not part of the functional eQTL locus. Furthermore, no eQTLs with trans-effects were identified. In conclusion, we did not detect any cis- or trans- effects for the SNPs of interest.

Table 6. Overview of eQTL results: the association between genetic variants of interest (Beta) with gene expression level, uncorrected for blood composition.

GWA SNP of interest	Top SNP in eQTL analysis	gene	LD r2	Beta	FDR
rs3755021	rs2305591	ITGA4	0.32	0.191	1.4e-05
rs13029501	rs2305591	ITGA4	0.03	0.203	1.34e-05
rs13029501	rs16867443	CERKL	0.43	0.132	1.34e-05
rs9469532	rs115378869	HLA-DPB1	0.04	0.106	3.07e-04
rs391855	rs1568391	IRF8	0.45	0.13	1.34e-05
rs55929401	rs7023923	LPAR1	0.24	0.42	1.34e-05

LD r^2: LD between GWAS SNP and top SNP in eQTL analysis. Beta= eQTL beta of GWA SNP, FDR= eQTL FDR GWA SNP.

6.4.7 GCTA

The results of the GCTA analyses are shown in **Table 7**. From GCTA we found a narrow sense heritability of 43.3% for MLR, 54.1% for monocyte count and 51.7% for lymphocyte count. The significant SNPs obtained in the GWAS for MLR explained 0.6% of the variance in MLR and the significant SNPS obtained in the GWAS for monocyte count explained 4.4% of the variance in monocyte count. All known loci from published literature together explained 1.3% of MLR variance, 2.4% of monocyte count variance and 0.3% of lymphocyte count variance.

Table 7. Narrow sense heritability and the proportion of genetic variance explained by known and significant SNPs according to GCTA analyses for MLR and its subcomponents.

Phenotype	Narrow sense heritability (SE)	P	Proportion of Genetic Variance Explained by Significant SNPs (SE)	P	Proportion of Genetic Variance Explained by Known loci (SE)	P
MLR	.4336(.025)	3.0E-8	.0058(.008)	9.6E-9	.0132(.006)	4.1E-13
monocyte	.5408(.022)	2.0E-12	.0437(.024)	8.5E-5	.0234(.009)	5.9E-11
lymphocyte	.5174(.023)	4.9E-8	NA	NA	.0027(.002)	.014

6.4.8 LD regression

In LD regression all λ values were larger than the LD score regression intercept and intercepts were close to 1, indicating that the inflation of the P value distribution from the GWAS results is caused by polygenetic effects, rather than population stratification. The SNP heritability of MLR, monocyte count and lymphocyte count, when applying LD regression was 13% and 17% and 19% respectively (see **Table 8**).

Table 8. LD regression results for MLR, monocyte count and lymphocyte count.

	MLR	Monocyte	Lymphocyte
Median of SIGNED_SUMSTATS	-9.23E-05	9.03E-06	-6.58E-05
Mean of X^2	1.007	1.011	1.021
λ GC	1.002	1.017	1.018
H^2 (se)	0.1302(.0733)	0.1702(0.0854)	0.1912(0.0895)
Intercept (se)	0.9915(.0063)	0.9912(0.0073)	1.0011(0.0069)

* P<0.05, **P<0.01, ***P<0.001. Median of Signed_sumstatistic: median value of beta values from GWAS.

In addition, we observed positive phenotypic correlations between MLR and monocyte count (r=0.550, p<.0001) and between monocyte count and lymphocyte count (r=0.386, p<.0001), and a negative phenotypic correlation between MLR and lymphocyte count (r=-0.494, p<.0001). However, despite the presence of phenotypic associations, no significant genetic correlations were detected between any pair of variables.

6.4.9 Exploring the effect of seasonal conditions and lifestyle.

Next, we explored the influence of seasonal conditions and lifestyle on variation of MLR in men and women (**Table 9**). Seasonal conditions did not influence MLR levels in either group. **Figure 4** illustrates the association between daily temperature and age-corrected MLR for men and women between August 2004 and December 2007. We note that MLR ratio seems do not influenced by temperature in both male and female groups. With respect to the subcomponents, a clear pattern of seasonal influence was only seen for monocyte count in women, with lower temperature being related to higher monocyte levels (β=-0.0005, p <0.004). Significant associations were also found between monocyte level and sunshine duration, global radiation, mean relative atmospheric humidity and potential evapotranspiration, but these associations did not survive correction for mean temperature.

Table 9. Results of the linear regression modeling for MLR and its subcomponents, separate for men and women.

Dependent variable	Independent variable	Men		Women	
		model 1	model 2	model 1	model 2
MLR	Age	**0.0013*** **	**0.0016** **	**0.0008*** **	**0.0014*** **
	BMI	-0.0041	-0.0011	-0.0034	0.0075
	Smoking	**-0.0029* **	0.0016	**-0.0044*** **	0.0046
	Age*BMI		0.0007		**-0.0002* **
	Age*Smoking		-0.0001		2.8E-6
Monocyte count	Age	0.00016	0.0005	-0.0001	-0.0011
	BMI	**0.0193*** **	**0.0304* **	**0.0151*** **	-0.0003
	Smoking	**0.0343*** **	**0.0245** **	**0.0297*** **	**0.00262*** **
	Age*BMI		-0.0002		0.0004
	Age*Smoking		0.0002		0.0009
Lymphocyte count	Age	**-0.0078*** **	**-0.0086** **	**-0.0078*** **	**-0.0154*** **
	BMI	**0.0947*** **	**0.0972* **	**0.0902*** **	-0.0487
	Smoking	**0.1512*** **	**0.1163*** **	**0.1765*** **	**0.1836*** **
	Age*BMI		-0.0003		**0.0031** **
	Age*Smoking		0.0009		-0.0001

* P<0.05, **P<0.01, ***P<0.001.

To test the effects of BMI and smoking, we included this variable in a regression analysis conducted separately by sex and taking age into account. The results, shown in Table 9 (model 1), indicate that smoking is related to a decrease in MLR level in both men and women. BMI was not associated with MLR in both men and women. However, an age x BMI interaction is seen for MLR in women (model 2): the age effects were alleviated by increased BMI level. The BMI and smoking effects were also examined in the MLR subcomponents: Higher BMI and being a smoker were related to higher monocyte and lymphocyte levels. For lymphocyte count in women, there was evidence for an age x BMI interaction, again indicating a reduction of the BMI effect at an older age.

6.5 Discussion

In this paper, we present a detailed examination of the causes of variance in MLR in the general population. Health status was as expected an important determinant of MLR level: individuals who were thought to have a compromised immune system, our so-called unhealthy group, had on average a higher MLR than our healthy participants. We continued with our healthy population and showed sex and age and their interaction to be important determinants of variation in MLR and its subcomponents.

Next, genetic factors were shown to be play a role in MLR variation in the general population. Heritability for MLR was estimated at 40% and MLR level was associated with a locus near *ITGA4*. Previous studies have shown this locus to be associated with monocyte level. Heritability estimates were higher for its subcomponents (58 % for both lymphocyte and monocyte count) and, in contrast to MLR, evidence was found for the presence of non-additive effects. Monocyte level was also associated with *ITGA4* and four more genes were related to monocyte level in our analyses, replicating findings in previous GWAs studies. For lymphocyte level no significant genetic variants emerged.

From our results it is clear that the genetic variants associated with blood cell counts may also influence their balance as reflected in their ratios. In addition to the genetic variant *ITGA4* which was significantly associated with both MLR and monocyte count in our study, there were a number of loci which were significantly associated with monocyte count and may have pleiotropic effects: The loci nearby *LPAR1*, *IRF8* and *ITPR3* were marginally significant associated with MLR level. Also, a locus nearby *C3orf27* was marginally significantly associated with both MLR and monocyte count. We did not see any suggestive evidence for pleiotropic genetic variants associated with both MLR and lymphocyte count.

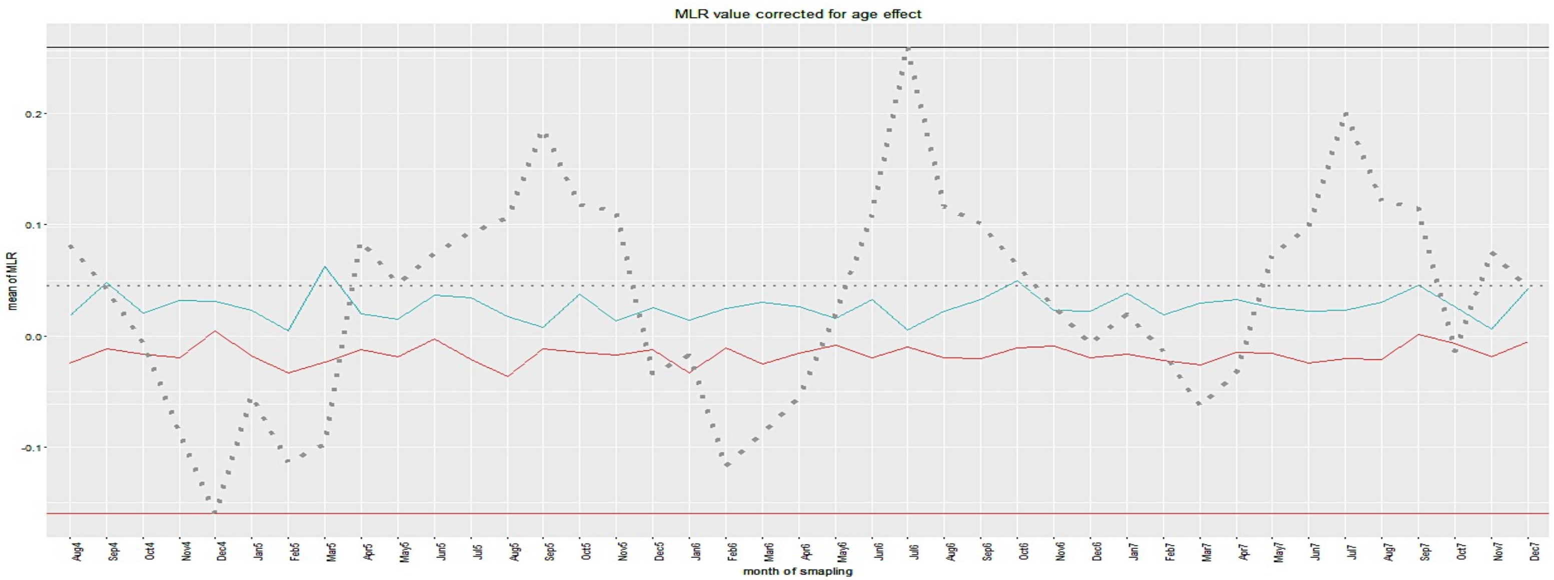
To understand more about the role of the genetic variants in MLR variation, we investigated what is known about the role the identified genetic variants play in regulating gene expression. However, we did not find any evidence for cis-effects or trans-effects by these genetic variants.

Among three phenotypes of interest, the narrow sense heritability h^2 of lymphocyte (35.3% in the healthy population) was the lowest, but its SNP heritability was the highest (19.12% from LD regression). These results suggest more common autosomal SNPs may be associated with lymphocyte count. The LD regression results show that polygenetic effects, rather than confounding factors explain both ratio and counts variance in our study. Although there are significant overall correlations and an overlap in associated genetic variants has been detected between MLR and monocyte count, no significant genetic correlations were detected between any pair of variables. Overall, our results suggest that the polygenetic effects are too small to be detected with the current sample size.

We also examined the impact of weather conditions and lifestyle on MLR variation. In contrast to what we found in an earlier study for PLR and NLR, weather conditions did not influence MLR, and neither did BMI. Smoking however was associated with a higher MLR. Note that this is in line with the higher MLR in the individuals with a compromised immune system in our study and the higher MLR seen in cancer patients.

Overall, this series of studies provided more insight into the causes of variation in MLR within the general population. While the genetic pathways as well as non-genetic causes of variance still need more clarification, it is clear that these factors need to be taken into consideration when studying the relationship between MLR and disease development.

Figure 4. The relationship between monthly temperature (grey dotted line) and the average MLR for men (blue line) and women (red line).



Chapter 7

The interactive effects of age, sex, and lifestyle on the hematological profile

This chapter was under review as: McArtor DB*, Lin BD*, Hottenga JJ, Boomsma D, Willemsen G, Lubke G., *The interactive effects of age, sex, and lifestyle on the hematological profile*. Biomark Med, 2016.

* Both authors contributed equally

Abstract

Aim. The hematological indices obtained in a standard blood test are closely interlinked. In this study, we exploit these interrelations in an investigation of the effects of age, sex, and lifestyle. We establish subjects' hematological profiles (i.e., scores on 10 indices, including measures of hemoglobin, corpuscular volume, platelets, and red blood cell distribution) and study the effects of demographic and lifestyle differences on these profiles. Results from this multivariate approach are compared to results based on the traditional approach of modeling each hematological index in isolation, which implicitly ignores the relationships among the indices.

Method. The sample consists of 3278 unrelated individuals from the Netherlands Twin Register biobank who provided hematological data. We use standard linear regression to examine the association between the individual hematological indices and the following predictors: age, sex, BMI, smoking, and their two-way interactions. Next, we use Multivariate Distance Matrix Regression (MDMR) to investigate the effects of the same predictors on the hematelogical profiles as a whole.

Results. Several main effects were identified in univariate analyses, but there was little evidence for two-way interaction between the predictors. The multivariate analyses of the hematological profiles, however, highlight the interactions of age with sex, BMI, and smoking, as well as the main effects of all predictors.

Conclusion. The multivariate approach increases the power to detect important interaction effects involving age and other predictors, and may help identify subgroups who benefit from different treatment or prevention measures. Implications for personalized medicine and gene-finding studies are discussed.

7.1 Introduction

Hematological indices are complex, heritable [51-53] and tightly regulated human phenotypes [267]. The set of blood cells targeted in a standard laboratory blood test provides information on a wide range of functions, including immune response, hormone regulation, osmotic balance and coagulation regulation [268-269]. Abnormal values on hematological indices that fall outside the reference range may be indicative of current or future disease [270]. Because the standard hematological profile is relatively easy and inexpensive to obtain, it provides the basis for many commonly used tests in diagnoses.

Many hematological variables are related to demographics and lifestyles. For instance, red blood cell count, hematocrit and hemoglobin have been shown to be associated with age, sex, smoking, and BMI [271]. Age and sex have also been found to be strongly related to platelet count, and age- and sex-specific reference ranges have even been proposed [272]. White blood cell count and platelet numbers are increased in obese participants [149], and in fact most hematological parameters show an association with BMI [148]. Smoking has also been associated with increases in white blood cell count, and changes in smoking behavior result in changes in the number of white blood cells [273].

Most studies concerning associations with hematological variables have taken an approach of investigating one hematological variable at a time. This approach is appropriate if a researcher is interested in the effects on a specific individual blood characteristic. If, however, the goal is to identify predictors associated with multiple blood characteristics, then the strategy of modeling each hematological variable in isolation is suboptimal. A more efficient strategy to accomplishing this goal is to test the association between a set of covariates and subjects' *hematological profiles*. Here, a hematological profile is defined as a set of scores on multiple observed hematological variables. Analyzing blood characteristics jointly rather than individually is theoretically appealing because it facilitates the identification of predictors (and interactions between predictors) that influence multiple blood traits jointly. Furthermore, considering blood counts as a multivariate outcome is statistically beneficial because it removes the necessity to correct for multiple testing, potentially resulting in more powerful tests.

An important benefit of the multivariate approach is that it facilitates the identification of characteristic profiles for subgroups, for example, characteristic profiles of subjects with high versus normal BMI levels in males and females. Differences among these profiles can be highly informative and useful in the distillation of personalized treatments. For example, they can characterize risk for maladaptive levels of particular blood counts in subgroups of the population that may otherwise tend to appear normal on many other hematological indices. The first step in this process is to identify predictors that are relevant in explaining differences in the hematological profiles. Once established, profiles can be compared across subgroups.

In this study, we establish hematological profiles and investigate whether hematological profiles are associated with age, sex, BMI, smoking, and assess the importance of moderation of main effects by age. This question is addressed using the standard univariate approach, as well as a multivariate approach that employs Multivariate Distance Matrix Regression (MDMR) [274-275] to test the association of the predictors with individual differences between the blood profiles as a whole.

7.2 Methods

7.2.1 Participants

The participants in this study are registered with the Netherlands Twin Register (NTR) and took part in NTR biobank projects [75-77]. In these projects, blood samples were collected during a home visit, and a brief interview was conducted to collect information on health status, lifestyle and body composition. All participants provided informed consent and the project was approved by The Medical Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam.

Within the group of individuals with hematological data (*n* = 9672), several exclusion criteria were applied. First, we excluded subjects with (1) blood C-reactive protein greater than or equal to 15 (mg/L), (2) basophil count ≥ 0.3 (10^9/ L), (3) illness within one week of measurement, (4) cancer treatment, (5) use of anti-inflammatory medication, (6) use of iron supplementation. The resulting sample was comprised of 8176 subjects. Next, subjects who had at least one blood cell score beyond ±5 standard deviations from that variable's mean were excluded. The resulting sample was comprised of 7999 subjects from 3278 families. The final dataset (*n* = 3278) was formed by randomly sampling one member from each family to ensure independence of observations because MDMR cannot currently be adjusted to account for familial clustering.

7.2.2 Data collection

Blood sampling and hematological indices. Participants were visited at home to obtain blood samples and a brief interview. These visits occurred between 7:00 a.m. and 10:00 a.m. following an overnight fast. Participants were asked to refrain from strenuous exercise and, if possible, medication as of the evening before the visit, and smokers were instructed to refrain from smoking one hour prior to the home visit. Fertile women without oral contraceptives were, when feasible, visited on the 2nd to 4th day of the menstrual cycle, and women taking oral contraceptives were visited in the pill-free week. During the home visit, peripheral venous blood samples were drawn by safety-lock butterfly needles into anticoagulant vacuum tubes in the following sequence: 2×9 ml EDTA, 2 x 9 ml lithium heparin (only one tube in a subset), 1 x 9 ml sodium heparin (in a subset only), 1×4.5 ml CTAD, 1 x 2.5 ml PAX (in a subset only), 1×4.5 ml serum and 1×2 ml

EDTA tube. After collection, all tubes were inverted about 10 times to prevent clotting and then transported to the laboratory in Leiden.

Hematological parameters. The 2 ml EDTA tubes were transported at room temperature and upon arrival in the laboratory used to determine the hematological parameters using the Coulter system (Coulter Corporation, Miami, USA). These parameters consisted of the total white blood cell count, percentage and absolute cell counts of five subtypes of white blood cells (neutrophils, lymphocytes, monocytes, eosinophils and basophils), red blood cell count, hemoglobin, hematocrit, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, red blood cell distribution width, platelet count and mean platelet volume.

C-creactive protein. C-reactive protein (CRP) level was obtained from a plasma subsample that came from a 9 (mL) heparin tube that was transported in melting ice to the laboratory. The plasma subsample was snap-frozen and stored at -30^{o}C. Upon defrosting one of these subsamples, CRP was determined by the 1000 CRP assay (Diagnostic Product Corporation).

Health, BMI and smoking. During the visit, a brief interview was conducted. Participants provided information on their current medication use and disease status and were asked about their smoking history. Height was reported and weight was measured. Body Mass Index (kg/m^2) was calculated from weight (kg) divided by the square of height (m^2). Based on their current smoking behavior, participants were divided non-smokers and current smokers.

7.2.3 Statistical analyses

Selecting outcome variables. To avoid the possibility of analyzing highly collinear variabels that measure extremely similar traits, we excluded several hematological variables that displayed large (>0.70) correlations with other candidate outcome variables. Specifically, we removed white blood cell count, red blood cell count, mean corpuscular hemoglobin and hematocrit ratio. In addition, basophil level was not included because variation in the basophil numbers was limited. This resulted in ten hematological outcome variables in total: neutrophil count (#*nneut*), lymphocyte count (#*nlymp*), monocyte count (#*nmono*), eosinophil count (#*neos*), hemoglobin level (*hgb*), mean corpuscular volume (*mcv*), mean corpuscular hemoglobin concentration (*mchc*), red cell distribution width in percent (*rdw%*), platelet count (#*plt*), and mean platelet volume (*mpv*).

Univariate association tests. Standard multiple regression with a univariate hematological outcome was used to investigate the effects of age, sex (62.6% female), smoking (77.6% non-smoker), BMI, and their two-way interactions on ten hematological variables (see **Table I** for descriptive statistics). The main interest of the analyses was to investigate interactions in order to explore potential risk groups.

The statistical significance of each predictor on each outcome was evaluated using two different univariate significance criteria, each of which use a Bonferroni correction to account for multiple testing. First the criterion α_f^u = 0.05/10 = 0.005 controls the family-wise Type-I error rate at 0.05. That is, α_f^u sets the probability of committing a Type-I error on a particular predictor to 0.05, which is accomplished by dividing 0.05 by the number of fitted regression models. Second, α_{pc}^u = 0.05/(10 × 10) = 0.0005 controls the per-comparison Type-I error rate at 0.05. This stricter criterion controls the probability of committing any Type-I errors across all ten models that each use ten predictors.

Table 1. Descriptive statistics for the hematological outcome variables and the numeric predictors.

	Mean	*SD*	*Min*	*1st Quantile*	*Median*	*3rd Quantile*	*Max*
Age	42.279	15.075	13.000	30.000	39.000	55.000	90.000
BMI	24.951	4.140	14.906	22.018	24.403	27.166	49.071
nneut(10^9/ L)	3.473	1.268	0.300	2.600	3.200	4.100	9.700
nlymp(10^9/ L)	2.230	0.682	0.300	1.798	2.100	2.600	5.900
nmono(10^9/ L)	0.534	0.171	0.000	0.400	0.500	0.600	1.400
neos(10^9/ L)	0.200	0.128	0.000	0.100	0.200	0.300	0.900
hgb(mmol/ L)	8.798	0.769	6.100	8.300	8.700	9.400	11.100
mcv (fL)	91.536	4.534	69.300	88.800	91.600	94.400	113.500
mchc(g/dL)	20.721	0.549	16.600	20.400	20.700	21.000	23.000
rdw(%)	12.364	0.743	10.700	11.900	12.200	12.700	16.600
plt(10^9/ L)	253.807	59.702	51.000	212.000	248.000	287.000	537.000
mpv(fL)	8.889	1.069	6.300	8.200	8.700	9.400	14.000

Multivariate association tests. Multivariate distance matrix regression (MDMR) [274-275] is a procedure that permits testing the association of hematological profiles based on multiple blood cell indices with predictor variables. More specifically, differences between each pair of subjects' profiles are collected in symmetric $n \times n$ "distance matrix". Distance matrices are often subjected to cluster analysis, but MDMR utilizes them in a regression framework instead in order to test the effects of covariates on the profiles. This is done by partitioning the sums of squares of the distance matrix into a portion due to regression and a portion due to error. This decomposition is analogous to the partitioning of the sums of squares of a univariate outcome in standard linear regression. Importantly, differences between profiles on multiple variables can be quantified using different measures of dissimilarity (i.e., distance). In this study, two different distance metrics were computed to characterize the dissimilarity between subjects' blood profiles, and the two resulting distance matrices were regressed onto the set of predictors using MDMR. The first metric considered was the Euclidean distance. If $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ denote vectors of scores along q outcome variables for subjects *i* and *j*, the Euclidean distance between these two subjects' response profiles is defined as,

$$d_e(i,j) = \sqrt{\sum_{k=1}^{q} \left(y_{ik} - y_{jk}\right)^2}$$

It can be shown that Euclidean-MDMR is the same model as multivariate multiple regression, so this approach also represents the natural multivariate extension to the standard linear regression used in the univariate analyses described above. Second, Manhattan distances were considered. The Manhattan distance between subjects *i* and *j* is defined as the sum of their absolute item-wise differences:

$$d_m(i,j) = \sum_{k=1}^{q} |y_{ik} - y_{jk}|$$

These distances are less sensitive to outliers and therefore more robust than Euclidean distances because they are based on absolute rather than squared differences. That is, the use of Euclidean distances (and standard linear regression) can result in spuriously significant effects due to outlying observations, but Manhattan distances are less prone to this phenomenon. When conducting MDMR, one model is fit to all ten outcome variables jointly. This approach therefore requires a less stringent correction for multiple testing than the univariate approach. More specifically, the criterion α_f^m= 0.05 controls the family-wise Type-I error rate of MDMR at 0.05 because only one model is fit to all outcome items jointly. Similarly, α_{pc}^m = 0.05/10 = 0.005 controls the probability of committing a Type-I error on any of the ten predictors at 0.05.⊠All analyses were conducted in R [219] using the MDMR package, which is available on CRAN (https://cran.r-project.org). This package is the software companion to McArtor et al. [276], where the reader can also find a more detailed discussion of MDMR.

7.3 Results

7.3.1 Univariate association tests

Table 2 reports the result of the univariate analyses in the form of *p*-values, and Table 3 gives the standardized regression coefficients and variance explained. One or more main effects were significant for all hematological variables except mean corpuscular hemoglobin concentration (*mchc*). Results indicated that smoking and high BMI are associated with elevated levels of most hematological variables, while sex and age were both found to be related to roughly half of outcomes. Unlike smoking and BMI, the direction of the main effects of sex and age differed across hematological variables. The majority of the two-way interactions assessed with the univariate regression models were not marked as significant using either of the univariate significance criteria. Only hemoglobin (*hgb*) was significantly predicted by multiple interaction effects (age:sex, age:smoker, sex:smoker), and neutrophil count (*nneut*) was found to be significantly associated with the interaction of age and sex. **Figure 1** illustrates the nature of these four interaction effects on their corresponding univariate outcome. None of the other hematological variables were found to be significantly predicted by any interaction effects.

Table 2. Results of models predicting individual hematological indices: *p*-values for the overall models, and for each test of main effects and interactions.

	nneut	*nlymp*	*nmono*	*neos*	*hgb*	*mcv*	*mchc*	*rdw*	*plt*	*mpv*
Full Model	**<1e-16**	**<1e-16**	**<1e-16**	**<1e-16**	**<1e-16**	**<1e-16**	0.019	**<1e-16**	**<1e-16**	**0.00026**
Age	0.00055	**3.9e-13**	0.015	0.13	0.23	**1.4e-23**	0.04	**1.2e-19**	**8.0e-06**	0.0055
Sex	**0.00013**	**3.9e-07**	**1.3E-23**	**1.9e-05**	**2.3e-16**	0.021	0.38	0.01	**6.1E-35**	0.011
BMI	**1.1e-29**	**1.2e-10**	**5.9e-07**	*0.0027*	**5.2e-09**	**2.0e-10**	0.048	*0.0016*	**1.3e-10**	0.97
smoker	**5.3e-60**	**2.9e-51**	**2.8e-39**	**1.8e-13**	**5.2e-15**	**4.8e-38**	*0.0017*	0.22	0.10	*0.0023*
age:sex	**2.8e-08**	0.29	0.5	0.051	**1.4e-17**	0.13	0.92	0.0082	0.27	0.12
age:bmi	0.62	0.011	0.62	0.057	0.13	0.044	0.73	0.39	0.031	0.12
age:smoker	0.56	0.74	0.38	0.25	**3.0e-07**	0.0065	0.25	0.043	0.41	0.26
sex:bmi	0.03	0.26	0.76	0.95	0.17	0.62	0.18	0.24	0.012	0.10
sex:smoker	0.71	0.011	0.43	0.84	**3.5e-06**	0.069	0.64	0.24	0.63	0.52
bmi:smoker	0.86	0.54	0.47	0.01	0.59	0.56	0.53	0.64	0.40	0.42

Each column corresponds to a model fit to one of the outcome variables. The first row corresponds to the *p*-value of the model as a whole, rows 2-5 correspond to the *p*-value of a main effect, and rows 6-11 report the *p*-values of each interaction effect. Values smaller than the Bonferroni-adjusted significance criterion to ensure that each predictor has a Type-I error rate of 0.05 (i.e. α_f^u= 0.05/10) are emphasized with italic font. Values smaller than the Bonferroni-adjusted significance criterion to ensure that the probability of any Type-I error is 0.05 (i.e. α_{pc}^u= 0.05/100) are emphasized with bold font.

Table 3. Results of models predicting individual hematological indices: R^2 of the different full models including all main and interaction effects, and standardized regression coefficients of age, sex, BMI, smoking status and their two-way interactions.

	nneut	*nlymp*	*nmono*	*neos*	*hgb*	*mcv*	*mchc*	*rdw*	*plt*	*mpv*
R^2	**0.136**	**0.106**	**0.11**	**0.041**	**0.472**	**0.097**	0.007	**0.048**	**0.071**	**0.011**
Age	*-0.066*	**-0.145**	0.048	0.031	-0.018	**0.189**	-0.04	**0.185**	**-0.09**	-0.048
Sex	**0.068**	**0.093**	**-0.181**	**-0.08**	**-0.662**	-0.04	-0.016	0.048	**0.232**	0.041
BMI	**0.215**	**0.126**	**0.096**	*0.06*	**0.086**	**-0.117**	0.038	*0.062*	**0.128**	0.001
smoker	**0.287**	**0.272**	**0.232**	**0.134**	**0.105**	**0.219**	*0.055*	0.022	0.029	*0.048*
age:sex	**-0.106**	0.021	0.013	-0.039	**0.127**	-0.028	0.002	-0.053	-0.022	-0.027
age:bmi	-0.01	0.051	0.01	0.039	0.023	0.038	-0.007	0.017	-0.043	0.027
age:smoker	0.011	-0.007	0.017	-0.023	**0.078**	0.051	0.023	0.041	-0.017	0.02
sex:bmi	0.042	-0.023	-0.006	0.001	-0.021	-0.009	-0.026	0.024	0.051	-0.029
sex:smoker	-0.006	0.044	-0.013	-0.004	**0.061**	0.03	0.008	-0.021	-0.008	0.01
bmi:smoker	-0.003	-0.012	-0.013	-0.05	-0.008	-0.01	-0.012	-0.009	0.016	-0.013

R^2 is shown in the first row, standardized regression coefficients of each main effect in rows 2-5, and all two-way interaction effects in rows 6-11. Values smaller than the Bonferroni-adjusted significance criterion to ensure that each predictor has a Type-I error rate of 0.05 (i.e. α_f^u= 0.05/10) are emphasized with italic font. Values smaller than the Bonferroni-adjusted significance criterion to ensure that the probability of any Type-I error is 0.05 (i.e. α_{pc}^u= 0.05/100) are emphasized with bold font.

Figure 1. Age sex interaction, age somker interaction and sex smoker interaction on the blood parameters.

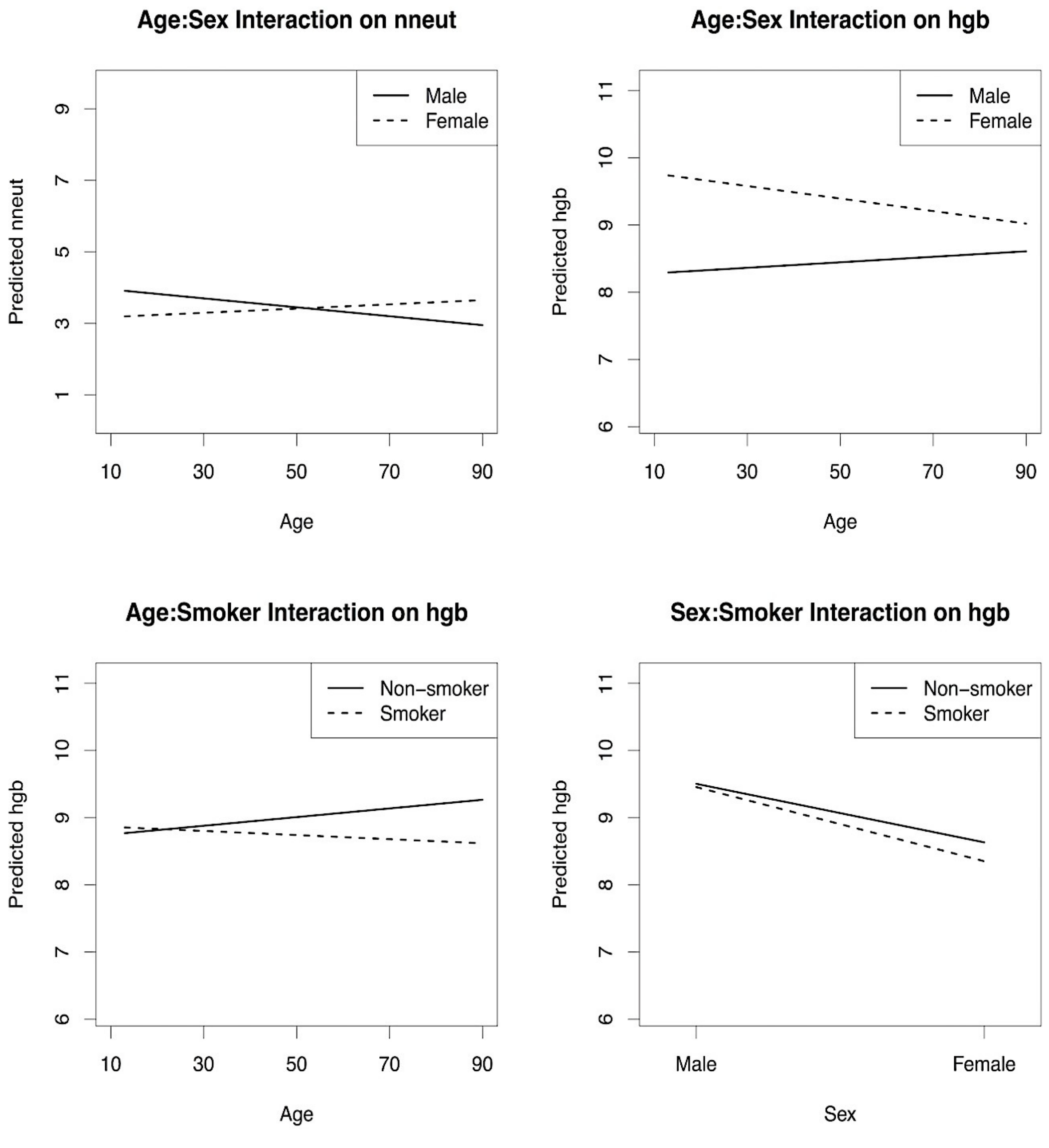


Illustration of significant interaction effects found in the univariate linear regression models. Each panel illustrates predicted values (vertical axis) across two variables (horizontal axis and line type) that interacted to predict a hematological variable. All predictors not involved in an illustrated interaction were kept fixed at their sample means.

7.3.2 Multivariate association tests

Both Euclidean- and Manhattan-MDMR resulted in extremely small *p*-values for all four main effects. However, Euclidean-MDMR also detected three interactions with age (age:sex, age:bmi, age:smoker) that were significant at α_{pc}^{m}, and two more that were significant at α_{f}^{m} (sex:bmi, sex:smoker). Manhattan-MDMR found the same three highly significant interaction effects involving age, and one more that was significant at α_{f}^{m} (sex:bmi). BMI and smoking were the only two predictors that did not combine to yield a statistically significant interaction effect from at least one of the MDMR models. See **Table 4** for all MDMR *p*-values.

Significant effects imply substantial differences between participants in their hematological profiles based on the ten observed blood variables. To visualize these effects, we conducted a median split on each of the two predictors comprising a significant interaction and plotted the average blood profiles in each of the resulting four groups (high/low × high/low on each pair of predictors). **Figure 2** displays these "prototypical" or "average" hematological profiles for each resulting group.

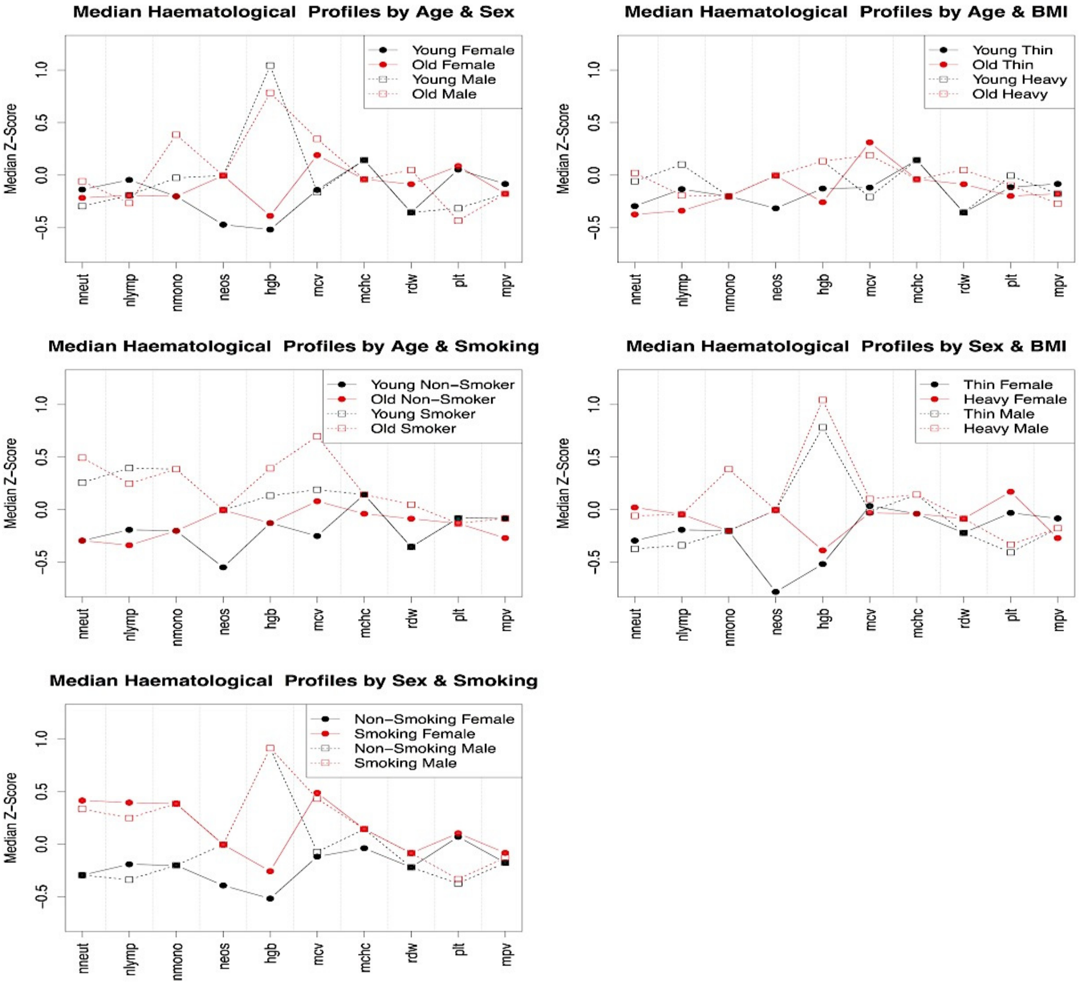
The five sub-plots comprising Figure 2 elucidate the five significant two-way interaction effects and the main effects of each predictor by illustrating how differences in the predictor variables relate to differences in the blood profiles. For example, the top-left subplot illustrates the effects of age, sex, and their interaction on the multivariate outcome. There are clear differences in hematological profiles among younger females, older females, younger males, and older males, but the differences between groups are not constant among the ten indices defining the profile. That is, it is not the case that one group tends to score uniformly higher or lower than another on all ten outcomes. The multivariate effects have complex patterns that allow for potentially different effects on each variable comprising the outcome. For example, both age and sex seem to have comparatively small effects on mean platelet volume *(mpv)* (all four groups tend to score similarly), age seems to have a main effect on mean corpuscular hemoglobin concentration *(mchc)* (young people tend to score higher regardless of sex), sex seems to have a strong main effect on hemoglobin level *(hgb)* (males tend to score higher regardless of age), and the interaction between these two predictors is important in predicting *neos* (younger females tend to score lower than the other three groups). The differential effects of age and sex on the remaining six hematological variables are also illustrated in the top-left sub-plot of **Figure 2**, and the other four sub-plots illustrate the differential effects of the other pairs of predictors that comprise a significant interaction effect.

Table 4. Results of models predicting hematological profiles: MDMR *p*-values for each overall model, and for each test of main effects and interactions.

	Euclidean	*Manhattan*
Full Model	**<1e-16**	**<1e-16**
age	**<1e-16**	**<1e-16**
sex	**<1e-16**	**<1e-16**
bmi	**<1e-16**	**<1e-16**
smoker	**<1e-16**	**<1e-16**
age:sex	**4.3e-10**	**4.4e-09**
age:bmi	**0.00034**	**0.0044**
age:smoker	**0.0022**	**0.00034**
sex:bmi	*0.0099*	*0.0094*
sex:smoker	*0.0075*	0.33
bmi:smoker	0.33	0.062

Rows correspond to predictors, columns correspond to metrics used to define the dissimiliarty between pairs of haematological profiles. Values smaller than the Bonferroni-adjusted significance criterion to ensure that each predictor has a Type-I error rate of 0.05 (i.e. α_f^u= 0.05/10) are emphasized with italic font. Values smaller than the Bonferroni-adjusted significance criterion to ensure that the probability of any Type-I error is 0.05 (i.e. α_{pc}^u= 0.05/100) are emphasized with bold font. The *p*-value corresponding to the joint effect of all predictors is found in the upper panel, the main effects of each predictor in the middle panel, and all two-way interaction effects in the lower panel.

Figure 2. Median values for each hematological outcome variable for subgroups.



Median standardized scores (vertical axes) on each hematological outcome variable (horizontal axes) for five sets of subgroups (each plot) to illustrate the effects of the five interactions identified as significantly associated with subjects' hematological profiles. Each interaction is illustrated by plotting the "average hematological profile" of four subgroups that characterize the two-way interaction. These groups are defined by, (a) a median-split on age and by sex, (b) a median split on age and on BMI, (c) a median-split on age and by smoking, (d) sex and a median split on BMI, and (e) sex and smoking. The average profiles of each subgroup within each plot are differentiated by color, point type, and line type, as indicated in each figure legend. Connecting lines were added to allow for an easier visual comparison of the groups'profiles. These visualizations illustrate the differential covariate effects on the hematological profiles as a whole. For example, the sub-plot concerning the effects of age and sex illustrates the comparatively small effects of both predictors on mean platelet volume (all four groups tend to score similarly), the main effect of age on mean corpuscular hemoglobin concentration (young tending to score higher regardless of sex), the main effect of sex on on hemoglobin level (males tend to score higher regardless of age), and the effect of the interaction between these two predictors on eosinophil count (younger females tend to score lower than the other three groups).

7.4 Discussion

In line with previous research, our univariate analyses confirmed that age, sex, BMI and smoking are related to individual hematological parameters. This set of analyses, however, did not provide strong evidence for interactive effects of these predictors. On the other hand, focusing on differences among subjects' hematological profiles rather than differences in individual hematological indices was shown to yield sufficient power to detect interactions among predictors in the model. For example, the interaction of age and BMI was not marked as significant in any of the univariate analyses, but this interaction was found to be significantly related to the hematological profiles as a whole. This phenomenon can be understood by examining the upper-rightmost plot of **Figure 2**, which illustrates the effects of age and BMI on the hematological profiles. While no single outcome variable is characterized by a large interaction effect, the interaction of age and BMI clearly has a modest effect on many of the blood variables (e.g., *nneut, neos, hgb, rdw*). By using information from all of the outcomes jointly, the multivariate approach can detect these smaller, but still meaningful, effects more efficiently than the traditional univariate approach, which can only consider each effect in isolation.

Beyond facilitating higher statistical power than the standard univariate approach, using the multivariate approach to identify "prototypical profiles", such as those illustrated in **Figure 2**, may be useful for clinicians in the future. These profiles could be used to formulate expectations about patient groups in a more fine-grained manner than could be achieved based on analyses of individual outcome variables.

Furthermore, the benefits of this multivariate approach invites future research on personalized treatment that directly utilizes multivariate association tests. Jointly modeling several outcomes facilitates the simultaneous study of multiple biological responses to a treatment. The multivariate approach can therefore be used to test the effectiveness of a treatment on several target variables while also considering potential treatment interactions with demographic variables. For example, the multivariate approach could be used to model phenotypes that are known to be impacted as a side effect of a treatment in conjunction with the variable(s) that are targeted by the treatment. This allows the identification of subgroups of individuals within a population who respond well to the treatment while also uncovering subgroups who are particularly succeptible to its side effects.

Importantly, the multivariate approach may also be useful in the context of genetic association studies. The effects of individual genetic variants on complex human traits are usually small [277]. Genome wide association studies for hematological parameters have now implicated several loci in the regulation of hematological indices, but the power is currently insufficient to detect all loci involved [188, 200, 220, 258]. To attain sufficient power to detect these effects, consortia currently focus on increasing sample sizes.

However, an alternative approach to increasing power involves improving the way that the phenotypes are operationalized and analyzed. When a researcher has multiple variables that measure a trait of interest, the multivariate approach can be used to test their joint association with individual genetic variants. The results presented here suggest that this approach could lead to increased power relative to analyzing each variable on its own, and this approach can also yield higher power than analyzing an aggregate-score computed from all of the variables measuring the trait [278]. MDMR facilitates the inclusion of an arbitrary number of outcome variables. It can even be used when there are more outcomes than observations, so these benefits can still be capitalized upon when the outcome is extremely high-dimensional.

In the analyses presented here, both Euclidean- and Manhattan-MDMR marked the interactions of age with sex, BMI, and smoking, as well as the interaction of sex and BMI, as significantly related to the hematological profiles. The use of Euclidean and Manhattan distances, however, yielded inconsistent results with respect to the interaction of sex and smoking. The use of Euclidean distances to define the dissimilarity between pairs of response profiles resulted in a significant sex by smoking interaction, but the use of Manhattan distances did not. Manhattan distances are less sensitive to outlying observations, and are therefore preferable if analyses are conducted in small samples in order to avoid potentially spurious results. This robustness, however, comes at the expense of potentially suboptimal power to detect genuine effects in larger samples. Researchers should therefore consider their sample size in addition to the relative cost of false positives and false negatives when choosing between Euclidean and Manhattan distances.

In conclusion, a multivariate approach to hematological analysis increases the power to detect important interactions within predictors relative to standard univariate analyses. In the future, multivariate methods, including MDMR, have the potential to help identify subgroups of patients who benefit from different treatment or prevention measures.

Part III: Epigenome-wide studies

Chapter 8

Blood hypomethylation is associated with elevated myeloid-lymphoid ratios in cell-specific active genomic regions

This chapter is ready to submitted as: Carnero-Montoro E*, Lin BD*, van Dongen J, Fairfax BP, Boomsma DI, Spector TD, Naranbhai V, Bell J. *Blood hypomethylation is associated with elevated myeloid-lymphoid ratios in cell-specific active genomic regions*. Blood.

*These authors contributed equally to this work.

Abstract

Peripheral blood myeloid-lymphoid ratios are increasingly being recognized as better biomarkers of immunological, cancer and cardiovascular diseases than single leukocyte counts on their own. Epigenetic mechanisms play an important role in cell fate and differentiation during hematopoiesis, and ultimately cell function. We hypothesized that differentially methylated genomic regions in whole blood associated with myeloid-lymphoid ratios may reveal genes involved in the process of hematopoietic cell differentiation and lineage specification. In this study we performed epigenome-wide analyses of myeloid-lymphoid ratios (monocyte- lymphocyte ratio: MLR and neutrophil-lymphocyte ratio: NLR) in two population-based cohorts; TwinsUK (N=844) and the Netherlands Twin Register (N= 2876). Through meta-analysis we identified thousands of sites where hypomethylation is associated with elevated myeloid-lymphoid ratios independently of single cell proportions. By contrasting our results with available data on genome functionality for different cell types, we find that differentially methylated sites are enriched in myeloid-specific active regulatory regions. The relevance of the identified differentially methylated sites (DMS) for disease biology is illustrated by enrichment of DMS for proximity to genes implicated in haematological malignancies and proximity to GWAS loci associated with autoimmune, inflammatory and endocrine diseases. Our results illustrate that epigenome-wide association study approaches utilizing peripheral whole blood DNA methylation in human population-based cohorts may provide fundamental insights into hematopoietic cell biology.

8.1 Introduction

Myeloid and lymphoid lineages of blood cells are formed from distinct hematopoietic stem cell-derived progenitor cells through the process of hematopoiesis [279]. Lymphoid cells (including T cells, B cells, and natural killer cells) are thought to be the major effector cells in pathogen immunity while myeloid cells (namely erythrocytes, megakaryocytes, monocytes, and granulocytes (which include neutrophils, basophils, and eosinophils) are the major host cells for infection. The relative abundance of myeloid and lymphoid cells reflects a balance between effectors and target cells in the immune response. Elevation of the ratios NLR and MLR are associated with impaired health outcomes due to cardiovascular diseases [181, 280], type 2 diabetes [281], hematological malignancy [282-283], solid-organ malignancy [184, 284-288] and some infectious diseases such as tuberculosis [288]. Recent studies have shown that peripheral blood MLR and NLR are better predictors of impaired health outcomes than the individual peripheral blood leukocyte subset counts on their own
[289-291]. The adverse physiological effects and health outcomes connected to elevated MLR and NLR may be mediated by epigenetic mechanisms such as DNA methylation. No study to date has performed a methylome-wide association analysis of myeloid-lymphoid ratios.

Epigenetic mechanisms are a key player in hematopoiesis [292]: the differentiation process of hematopoietic stem cells is accompanied by extensive changes in their methylation pattern [292-293]. Specifically, hypomethylation is essential to differentiate stem cells into myeloid progeny, but not into lymphocytes [292]. As lifelong replenishment of hematopoietic cells is sustained by differentiation of hematopoietic stem cells (HSCs), methylome-wide association analysis of myeloid-lymphoid ratios may reveal molecular mechanisms regulating stem cell function, such as self-renewal, cell differentiation and cell fate decision. Methylation analysis may also reveal the functional impact of changes in stem cell fate programs, which are still poorly understood.

We performed a genome-wide DNA methylation analysis to identify methylation sites associated with peripheral blood MLR and NLR based on data from two large twin registers, including 2876 participants from the Netherlands Twin Register (NTR) and 877 participants from TwinsUK. With this analysis, we aimed to gain insight into 1) the epigenetic mechanisms driving changes in myeloid-lymphoid ratios and 2) the epigenetic consequences of elevated ratios, which may be connected to the adverse physiological effects and health outcomes. We report 4185 methylation sites associated with MLR and/or NLR independently of single leukocyte counts. Functional annotation analysis reveals that these sites are enriched in myeloid-specific active enhancer regions and mQTL analysis shows that methylation at many of these sites is influenced by SNPs.

8.2 Methods

8.2.1 Study populations

This study includes participants of the Netherlands Twins Register (NTR) and participants of the Twins UK Registry (TwinsUK). The study-specific characteristics and methodology is described in previous work [75-77, 294-295]. The NTR includes more than 175,000 participants who are members of twin families: twins, parents, sibling, spouses, etc. from all regions of the Netherlands. The TwinsUK cohort includes more than 13,000 monozygotic and dizygotic twin volunteers from all regions across the United Kingdom. Participants were not selected based on any disease or phenotypic characteristic. For this work, information on blood traits and DNA methylation was available for 3,089 subjects in NTR and for 877 subjects in TwinsUK. The study protocol for NTR was approved by the Medical Ethics Committee of the VU University Medical Center Amsterdam, and all participants provided informed consent. For TwinsUK, Guy's and St Thomas' Hospital NHS Trust Research Ethics Committee approved the study, and all twins provided informed written consent.

8.2.2 Blood sample collection and cellular composition assessment

Blood samples from participants were collected during visits following well established protocols described somewhere else for NTR [76-77] and for TwinsUK [295]. In NTR, the hematological profile of the blood samples was obtained, including white blood cell type counts that were measured using the Coulter system (Coulter Corporation, Miami, USA). The total white blood cell counts, numbers and percentages of sub types of white blood cells namely neutrophils, lymphocytes, monocytes, eosinophils and basophils, were obtained from hematological profile results. NLR was calculated by dividing the percentage of neutrophil counts among all white blood cell counts by the percentage of absolute lymphocyte count among all white blood cell count. Likewise, MLR was calculated by dividing the percentage of monocyte counts among all white blood cell counts by the percentage of absolute lymphocyte count among all white blood cell count. The blood samples used to characterize the blood composition for NTR were the same as those used to assess the genome-wide DNA methylation levels. In TwinsUK, measured blood cell counts were not available for the blood samples in which DNA methylation was measured. Therefore, white blood cell proportions were predicted in TwinsUK based on the genome-wide methylation data, by means of the widely used method described by Houseman *et al* [296], which makes use of a reference dataset consisting of DNA methylation profiles in purified white blood cells. We used the Houseman reference-based method as implemented within the R package meffil, using blood gse35069 as the reference [297]. With this method we predicted the proportions of monocytes, neutrophils, basophils, B cells, CD8T, CD4T and NK cells. We calculated MLR and NLR based on predicted leukocyte cell proportions, in which the lymphocyte denominator was formed by the sum of B cells, CD8T, CD4T and NK cells. To allow for comparison of EWAS results obtained with measured versus predicted white blood cell counts, we also obtained predicted white blood cell traits in NTR with the Houseman *et al* algorithm following exactly the same methods.

8.2.3 Genome-wide DNA methylation assessment

The Infinium HumanMethylation450 BeadChip (Illumina Inc, San Diego, CA, USA) was used to measure DNA methylation in 500ng of genomic DNA obtained from whole blood. Details of experimental approaches and quality control steps have been previously described for both cohorts [298-299]. Briefly, in TwinsUK, probes with detection of P value > 0.05 in more than 5% of the samples, mapping to multiple locations in the genome, or non-autosomal were excluded. In NTR, probes were set to missing in a sample if they had an intensity value of exactly zero, or a detection p > 0.01, or a bead count of < 3. After these steps, probes that failed based on the above criteria in > 5% of the NTR samples were excluded from all NTR samples (only probes with a success rate > 0.95 were retained). In total, 453,288 and 452,874 probes passing QC filters in NTR and in TwinsUK were analysed. The methylation data were normalized using functional normalization in NTR [300], and using the BMIQ method in TwinsUK [301]. For subsequent analyses, we used methylation β-values, which represent the methylation level at a site, ranging from 0 to 1, and is calculated as: β=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.

8.2.4 Exclusion criteria

We excluded from the analyses individuals that were outliers for MLR and/or NLR as determined by values deviating more than 5 times the SD from the mean. In TwinsUK, we excluded non-Caucasian samples. For NTR, we also filtered out individuals either with hpa-axis related medication, anti-immune medication, or cancer treatment. After applying exclusion criteria, we ended up with 2,876 subjects for NTR and 844 subjects for TwinsUK that were used for our statistical analyses.

8.2.5 Statistical analyses

The complete analysis workflow is illustrated in **Figure 1**. All statistical analyses were performed using R programming language [302]. In the NTR cohort, Pearson's correlation was computed to evaluate how strongly the measured cell ratios (MLR and NLR) correlate with the prediction-based cell ratios.

Epigenome-wide association studies (EWAS) were performed to test, at each CpG site, if the methylation β value is associated with MLR and NLR in each cohort separately (see model 1). For NTR, we ran EWASs on measured and predicted ratios, while in TwinsUK we only ran the EWAS based on predicted ratios. In NTR, we used generalized estimation equation (GEE) models [303], with DNA methylation β value as outcome and the following predictors: NLR or MLR, sex, age at blood sampling, smoking status (3 categories: current smoker, former smoker and non- smoker), HM450k array row, sample plate, and single cell proportions (monocytes, neutrophils and lymphocytes). GEE models were fitted with the R package GEE, with the following specifications: Gaussian link function (for continuous data), 100 iterations, and the 'exchangeable' option to account for the correlation structure within families and within persons. In TwinsUK, we used a linear mixed model framework with methylation β value as outcome, MLR or NLR as the predictors, the following fixed effect covariates: age, gender, smoking, predicted single cell proportions (monocytes, neutrophils and lymphocytes) and batch effects (plate ID and plate position within 450K array); and the following random effects variables: zygosity and family. Linear mixed models were fitted using the function "lmer" within the R package "lme4" [304].

To combine association results from both cohorts, a random effects meta-analysis of the EWAS results based on the predicted white blood cell traits was performed on a total of 437,677 shared probes using the "metafor" R package and the REML method [305]. Next, we considered robust differentially methylated sites for MLR and NLR (DMS_{MLR} and DMS_{NLR}) as those CpGs showing associations that passed a Bonferroni threshold $P<1.14x10^{-7}$ in the meta-analyses and that showed a nominal significance level (P<0.05) in the EWAS of observed (measured) ratios in NTR.

Based on the same methodology, we also performed EWASs and meta-analyses to test the association between methylation and monocyte, neutrophil and lymphocyte proportions separately, while correcting for other single leukocyte proportions (see model 2,3,4).

Model 1: **β value** ~ **MLR or NLR** + % Monocytes + % Neutrophils + % Lymphocytes + Age + Sex + Smoking + Family/Population structure + Batch effects.

Model 2: **β value** ~ **%Monocytes** + % Neutrophils + % Lymphocytes + Age + Sex + Smoking + Family/Population structure + Batch effects.

Model 3: **β value** ~ **% Neutrophils** + % Lymphocytes + %Monocytes + Age + Sex + Smoking + Family/Population structure + Batch effects.

Model 4: **β value** ~ **% Lymphocytes** + %Monocytes + % Neutrophils + Sex + Age + Smoking + Family/Population structure + Batch effects.

Figure 1. Overview of study design and main results.

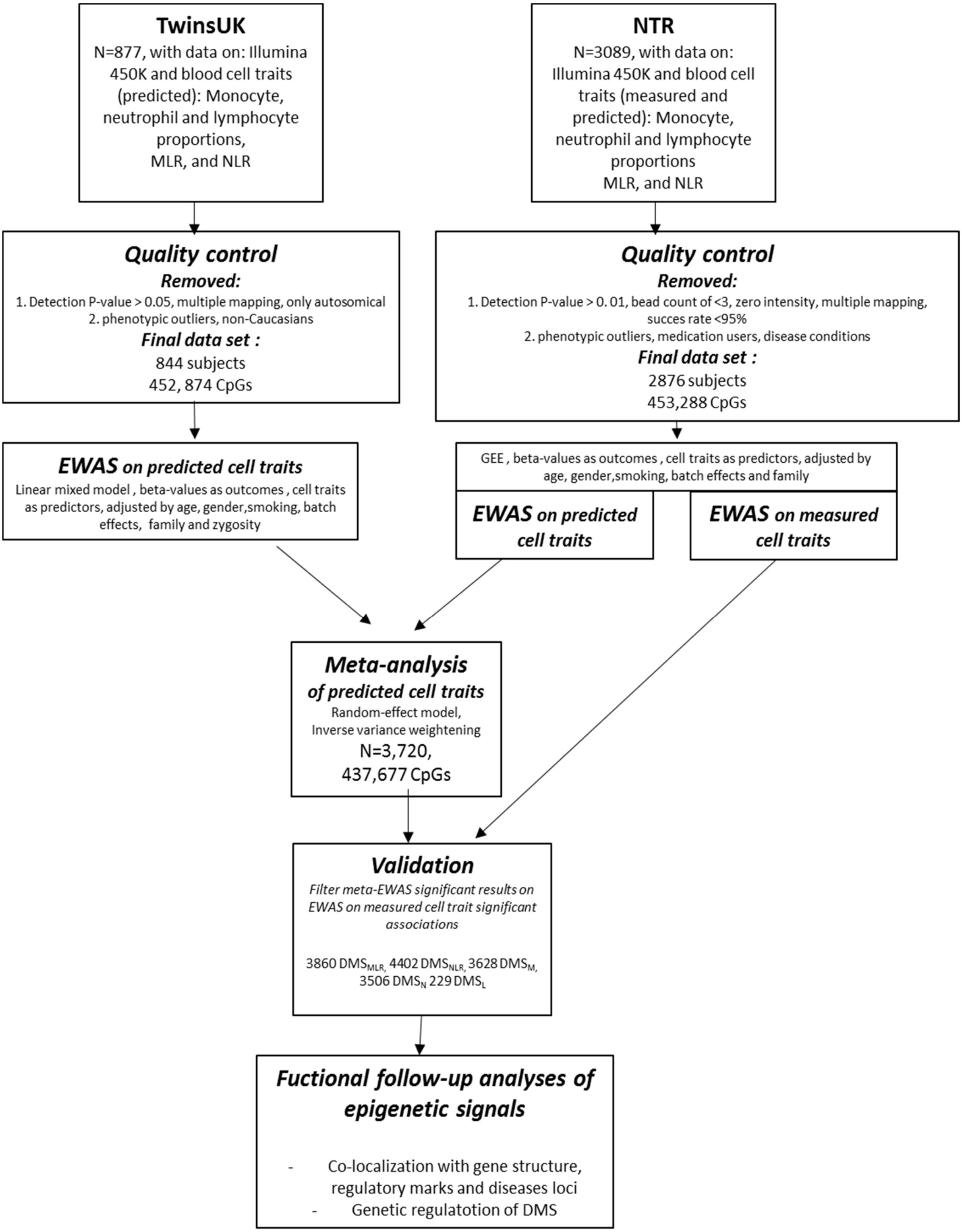

TwinsUK
N=877, with data on: Illumina 450K and blood cell traits (predicted): Monocyte, neutrophil and lymphocyte proportions, MLR, and NLR
NTR
N=3089, with data on: Illumina 450K and blood cell traits (measured and predicted): Monocyte, neutrophil and lymphocyte proportions MLR, and NLR
Quality control
Removed:
1. Detection P-value > 0.05, multiple mapping, only autosomical
2. phenotypic outliers, non-Caucasians
Final data set :
844 subjects
452, 874 CpGs
Quality control
Removed:
1. Detection P-value > 0. 01, bead count of <3, zero intensity, multiple mapping, succes rate <95%
2. phenotypic outliers, medication users, disease conditions
Final data set :
2876 subjects
453,288 CpGs
EWAS on predicted cell traits
Linear mixed model , beta-values as outcomes , cell traits as predictors, adjusted by age, gender,smoking, batch effects, family and zygosity
GEE , beta-values as outcomes , cell traits as predictors, adjusted by age, gender,smoking, batch effects and family
EWAS on predicted cell traits
EWAS on measured cell traits
Meta-analysis of predicted cell traits
Random-effect model, Inverse variance weightening
N=3,720, 437,677 CpGs
Validation
Filter meta-EWAS significant results on EWAS on measured cell trait significant associations
3860 $DMS_{MLR,}$ 4402 $DMS_{NLR,}$ 3628 $DMS_{M,}$ 3506 DMS_{N} 229 DMS_{L}
Fuctional follow-up analyses of epigenetic signals
- Co-localization with gene structure, regulatory marks and diseases loci
- Genetic regulatotion of DMS

Table 1. Study population characteristics.

	NTR	TwinsUK
N	2876	844
%Females	65.6	96.6
Age	36.71 (12.84)	57.99 (10.34)
% Current Smokers	20.58	11.25
% Neutrophils measured	52.35 (8.74)	NA
% Monocyte measured	8.35 (2.08)	NA
% Lymphocyte measured	35.71 (8.08)	NA
NLR measured	1.61 (0.69)	NA
MLR measured	0.25 (0.09)	NA
% Neutrophils predicted	53.45 (9.15)	56.29 (11.57)
% Monocyte predicted	8.27 (2.13)	8.29 (2.35)
% Lymphocyte predicted	39.42 (9.66)	36.06 (12.03)
NLR predicted	1.52 (0.75)	1.86 (1.02)
MLR predicted	0.23 (0.10)	0.26 (0.13)

NTR: Netherlands Twin Register. Quantitative values are represented as mean (standard deviations). N=Sample size after quality control and exclusion criteria. NLR: Neutrophil to lymphocyte ratios. MLR: Monocyte to lymphocyte ratios. NA: Non-available information. Measured: data obtained directly from experimental procedures. Predicted: Data obtained indirectly by applying algorithm for predictions.

8.3 Results

8.3.1 Characteristics of study populations

We included in this study 2,876 subjects from the NTR cohort and 844 subjects from TwinsUK with information on genome-wide methylation and white blood parameters that passed our quality control and exclusion criteria. The characteristics of the samples can be seen in **Table 1**. In both cohorts we estimated leucocyte subset proportions using a widely-adopted reference-based approach [296]. For NTR, we had information on white blood cell ratios (MLR and NLR) available based on both measured and predicted data. We observed a very high and significant correlation between these two type of data (Pearson correlations cor_{MLR}=0.86, P<2.2e-16 ; cor_{NLR}=0.89, P<2.2e-16) , (**Figure 2**).

8.3.2 Association between DNA methylation and blood ratios.

The complete analysis workflow is illustrated in **Figure 1**. In a total of 3,720 individuals we identified differentially methylated sites (DMS) associated with MLR (DMS_{MLR}) and NLR (DMS_{NLR}), and with single leukocyte proportions: neutrophils (DMS_N), monocyte (DMS_M) and lymphocyte (DMS_L). We assessed CpG methylation on the HumanMethylation450 Beadchip, and performed epigenome-wide association analyses (EWAS), comparing methylation levels at each CpG site in a regression framework to the predicted myeloid-lymphoid ratio, adjusting for population and/or family structure, age, sex, smoking status, batch effects, and leucocyte subset proportions (lymphocyte, monocyte and neutrophil proportions) separately in each sample. We combined effect estimates from TwinsUK and

NTR in a random-effects meta-analysis and applied a Bonferroni-correction to determine significant DMS (P <1.14x10^{-07}). To exclude potential bias due to the white blood cell proportion prediction method, we filtered meta-analysis results further, and report only those validated in an EWAS of measured proportions. A distribution of P values and direction of effects obtained in each meta-analyses EWAS can be seen in **Figure 3A.**

The level of methylation at 9749 CpG sites (2.2% of tested) was associated with NLR, MLR, neutrophil proportion, monocyte proportion or lymphocyte proportion. The largest number of DMS were observed for NLR (DMS$_{NLR}$=4402) and MLR (DMS$_{MLR}$ =3860) and for the neutrophil (DMS$_{N}$=3506) and monocyte (DMS$_{M}$=3628) proportions while few DMS were associated with lymphocyte proportion(DMS$_{L}$=229). Examining the patterns of overlap demonstrates a striking overlap of DMS$_{NLR}$ and DMS$_{MRL}$. Although some overlap was seen between DMS$_{N}$ and DMS$_{M}$, there was smaller overlap between DMS$_{NLR}$ and DMS$_{N}$ or DMS$_{L.}$ Similarly, there was also smaller overlap between DMS$_{MLR}$ and DMS$_{M}$ or DMS$_{L}$ (**Figure 3B**). These data suggest profound epigenetic correlation between MLR and NLR, which is more marked than the epigenetic correlation between the ratios and their constituents.

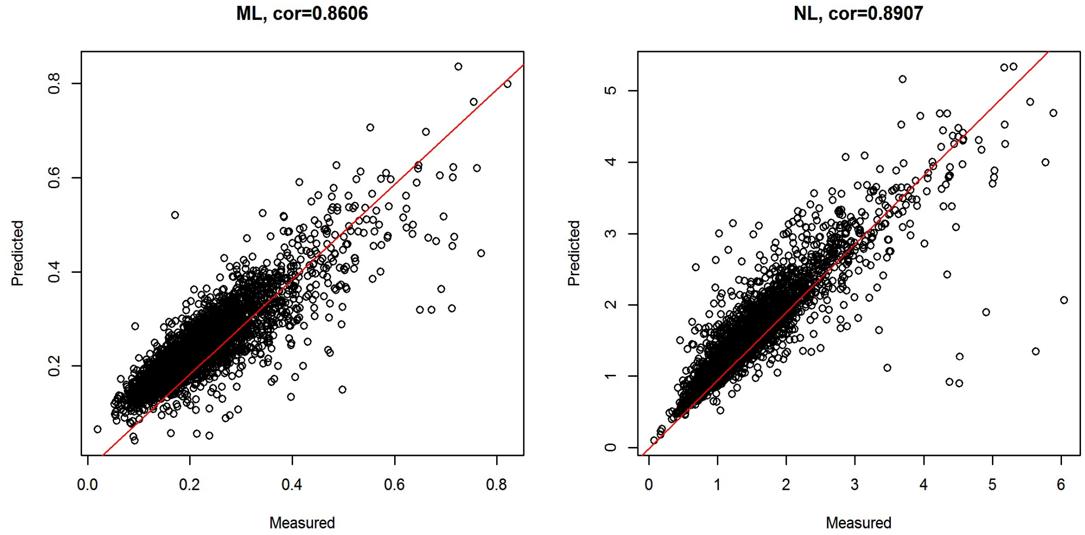
A second noteworthy observation is that the majority of DMS were inversely associated with the outcome of interest, i.e. that hypomethylation was associated with elevated ratios or proportions and this characteristic was more marked for the ratios than for their constituents. This pattern was true for NLR (95.56% of DMS$_{NLR}$ have $\beta<0$), MLR (94.09% of DMS$_{MLR}$ have $\beta<0$), neutrophil proportion (79.41% of DMS$_{N}$ have $\beta<0$), monocyte proportion (67.19% of DMS$_{M}$ have $\beta<0$) and lymphocyte proportion (60.26% of DMS$_{L}$ have $\beta<0$). For DMS associated with both MLR and NLR (3170 sites), all DMS shared the same direction of association (3021 $\beta<0$, 149 $\beta>0$). The DMS$_{NLR}$ and DMS$_{MLR}$ tend to be more likely to overlap with methylation sites that are hypomethylated in neutrophils or monocytes than expected, and less likely to overlap sites that are hypermethylated (data not shown).

The marked overlap in the sites of (mostly) hypomethylation and their directionally similar effect on MLR and NLR is consistent with a model in which long-term hematopoietic stem cells (LT-HSC), which have constitutive DNA methylation, actively undergo demethylation to give rise to myeloid-biased progeny leading to higher ML and NL ratios. Moreover, the regions of hypomethylation likely remain hypomethylated in the myeloid progeny.

8.3.3 Functional characterization of blood ratios epigenetic signals

To characterize potential mechanisms through which DMS$_{MLR}$ and DMS$_{NLR}$ may associate with myeloid-lymphoid ratio we examined their location relative to gene structures, a number of regulatory marks (as histone modification, TFBS, enhancer regions), the overlap with differentially expressed regions, and the influence of genetic variants on differential methylation and/or association.

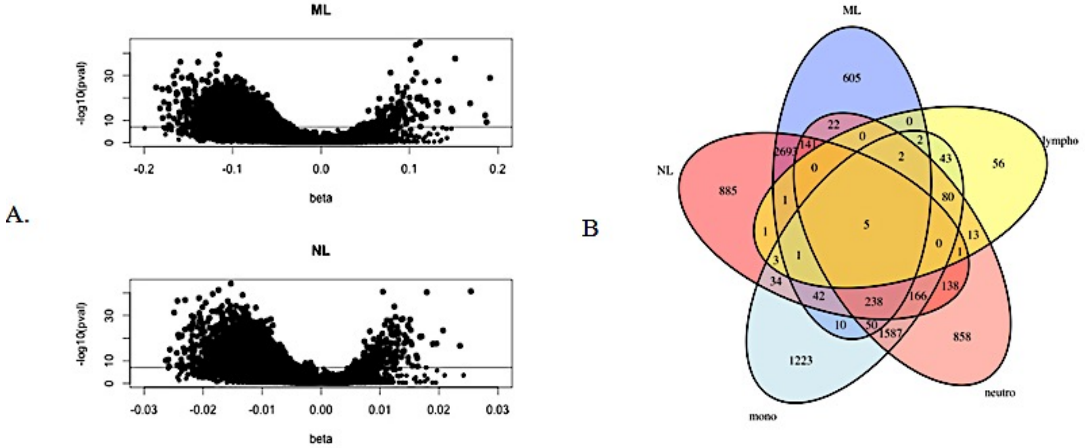
Figure 2. Comparison between predicted and measured white blood ratios



Observed (measured) values are plotted on the x-axis, predicted values are plotted on the y-axis. ML=Monocyte--lymphocyte ratio. NL=Neutrophil- lymphocyte ratio

Figure 3. Genome-wide DNA methylation meta-analyses results

A. Volcano plots for MLR and NL BVenn diagram showing the overlap of significant hits for all 5 blood traits analyzed.



8.3.4 Genomic co-localization

We examined the genomic distribution of $\mathrm{DMS_{NLR}}$ and $\mathrm{DMS_{MLR}}$ relative to gene structures and CpG island regions and compared this to all sites evaluated (background). We observed a depletion of $\mathrm{DMS_{NLR}}$ and $\mathrm{DMS_{MLR}}$ within 200bp of the TSS or within the first exon and reciprocally observed a substantial enrichment in the body of genes (**Figure 4A**). DMS_{NLR} and DMS_{MLR} were relatively depleted for locating precisely within a CpG Island, but enriched in the shores (2-4KB up/downstream from an island) (**Figure 4B**).

8.3.5 Enrichment of epigenetic signals in cell-specific regulatory histone marks and enhancers

We evaluated the overlap between DMS and regulatory histone marks reasoning that co-localisation may indicate evidence for multiple regulatory mechanisms being involved in regulation of gene expression. Both DMS_{NLR} and DMS_{MLR} are markedly enriched for overlapping with activating histone marks (H3K27Ac, H3K36me3 or H3K4me3), or histone marks reflective of active (H3K27Ac) or poised enhancers (H3K4me1) expressed in mature myeloid cells (monocytes or neutrophil ChIPSeq data from BLUEPRINT) [306] and are depleted for overlap with regions bound by repressive histone marks (H3K27me3 or H3K9me3), (**Figure 4C**). An orthogonal analysis of expressed enhancers (from FANTOM5) [307] confirms this observation and demonstrates relatively greater enrichment for myeloid (basophil, neutrophil, eosinophil, monocyte or dendritic cell) than lymphoid (B, T or NK cell) expressed enhancers (**Figure 4D**). These data collectively suggest that the DMS_{MLR} and DMS_{NLR} occur more frequently in the gene body and the shores of CpG island regions, in regions bound by activating histone marks and in enhancers.

8.3.6 Overlap between epigenetic and transcriptional signals

We previously demonstrated that an elevated myeloid-lymphoid cell ratio is associated with a consistently altered signature of transcript in myeloid cells. We overlapped the epigenome maps generated here to transcriptome maps in order to test whether the regions of hypomethylation may explain transcription levels in monocytes and neutrophils. Overall, 136 DMS_{MLR} were within genes that we previously identified as differentially expressed in monocytes dependent on the MLR, which represents a 1.65-fold enrichment over background (95%CI 1.38-1.96, p=1.22x10^{-7}). There was modest evidence for localization of DMS to within 1000kb of reported eQTL in neutrophils, monocytes, DC, T and B-cells (Enrichment OR all <2, data not shown).

8.3.7 Genetic determination of epigenetic signals

We investigated how the interplay between genetic and epigenetic variation could influence MLR and NLR levels. We used a genome-wide catalog of genetic variants influencing DNA methylation (mQTLs) produced in TwinsUK and interrogated whether MLR and NLR-associated CpGs were under genetic control. We found that a proportion of 0.49 DMS_{MLR} (1910 CpGs) and DMS_{NLR} (2161 CpGs) showed at least one SNP regulating the methylation level at a significance level of P<1x10^{-05} and FDR<0.05, which represents a significant enrichment of mQTL among significant cpgs compared to the number of mQTLs among all CpGs included in the meta-analysis (Fold enrichment = 1.43 for DMS_{MLR} and 1.41for DMS_{NLR}, P< 0.0001). This enrichment is slightly higher among mQTLs acting in cis, than among mQTLs acting in trans (those SNP-CpG pairs located in different chromosomes or further than 1Mb) (data not shown). Our results reveal that at least partially, the epigenetic signals associated with elevated white blood ratios could be the consequence of genetic variations influencing blood ratios via methylation changes.

8.3.8 Disease relevance

The myeloid-lymphoid ratio is implicated in a range of human disorders in epidemiological studies. To further explore this observation leveraging the EWAS findings we pursued two additional approaches: we tested whether DMS_{NLR} and DMS_{MLR} are enriched for proximity to genes involved in haematological malignancies and whether they are enriched within 100kb of a GWAS associated genetic variant according to disease category. We found that the DMS_{NLR} and DMS_{MLR} were enriched for proximity to genes implicated in haematological malignancies in particular. They were also enriched for proximity to GWAS loci associated with autoimmune, inflammatory and endocrine diseases.

8.4 Discussion

We performed an epigenome-wide association meta-analysis of the myeloid-lymphoid ratios MLR and NLR in two population-based cohorts; TwinsUK and the Netherlands Twin Register, including a total number of 3,720 subjects. This is the first EWAS to date specifically focused on MLR and NLR. The meta-analysis identified thousands of sites where methylation level is associated with myeloid-lymphoid ratios independently of single cell proportions. For the large majority (> 94%) of these differentially methylated sites (DMS_{MLR} and DMS_{NLR}), a higher MLR or a higher NLR was associated with hypomethylation. There was very large overlap between sites significantly associated with MLR and sites significantly associated with NLR, and between their direction of effect. Functional annotation analyses indicated that DMS_{MLR} and DMS_{NLR} occur preferentially in myeloid-specific active regulatory regions including enhancers and mQTL results indicated that approximately 50% DMS_{MLR} and DMS_{NLR} are also associated with one or more SNPs. This pattern is consistent with a model in which long-term hematopoietic stem cells (LT-HSC), which have constitutive DNA methylation, actively undergo demethylation at myeloid-specific regulatory regions to give rise to myeloid-biased progeny leading to higher MLR and NLR, and this process may be influenced at least to some extent by SNPs that affect the methylation levels of these locations .

Peripheral blood myeloid-lymphoid ratios have repeatedly been reported to be better biomarkers of immunological, cancer and cardiovascular diseases than single leukocyte counts on their own. Our study showed that myeloid-lymphoid ratio- associated methylation sites are enriched for proximity to genes implicated in haematological malignancies and proximity to GWAS loci associated with autoimmune, inflammatory and endocrine diseases. This suggests that differential epigenetic regulation of disease-related genes may be a key mode of action linking elevated myeloid-lymphoid ratios to disease.

Figure 4. Functional annotation results.

Figure 4 A. Enrichment of DMS_{NLR} and DMS_{MLR} with respect to gene-centric annotations.

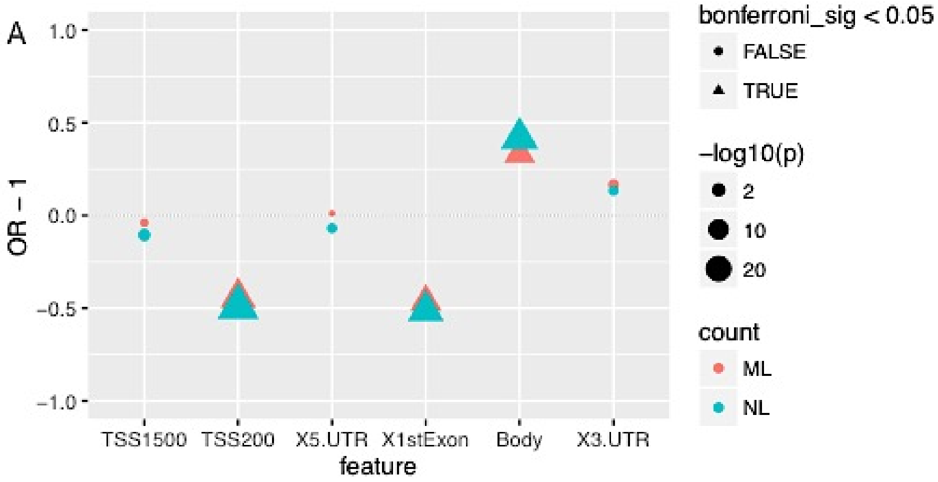


Figure 4B. Enrichment of DMS_{NLR} and DMS_{MLR} with respect to CpG island-centric annotations.

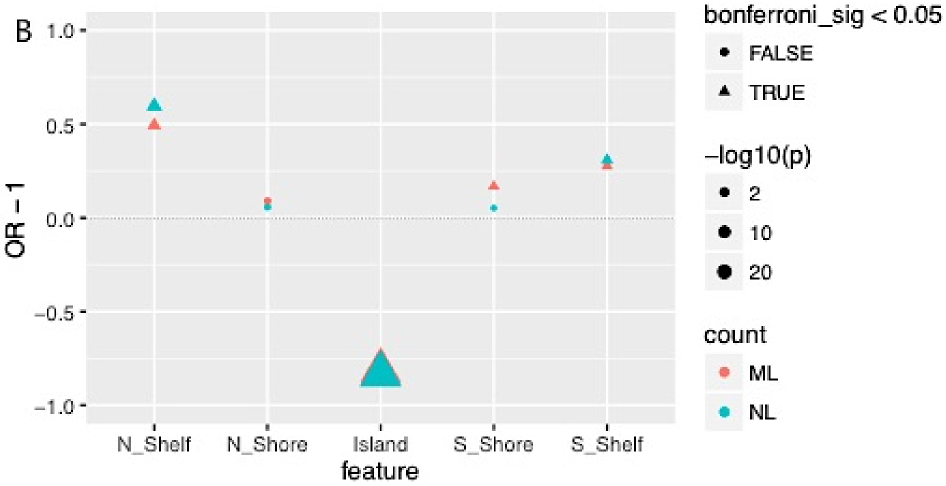
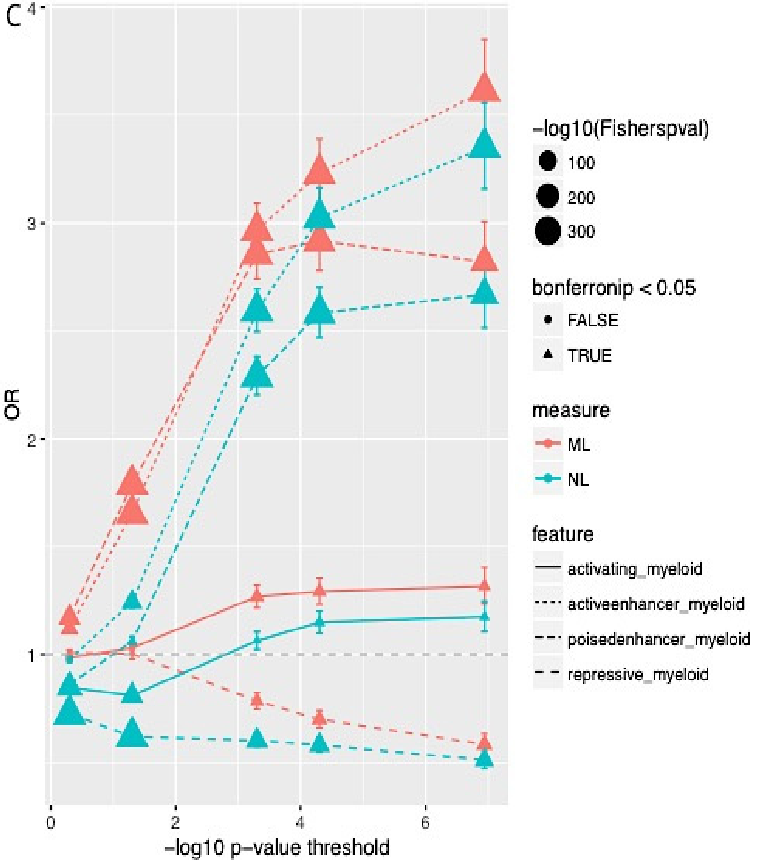


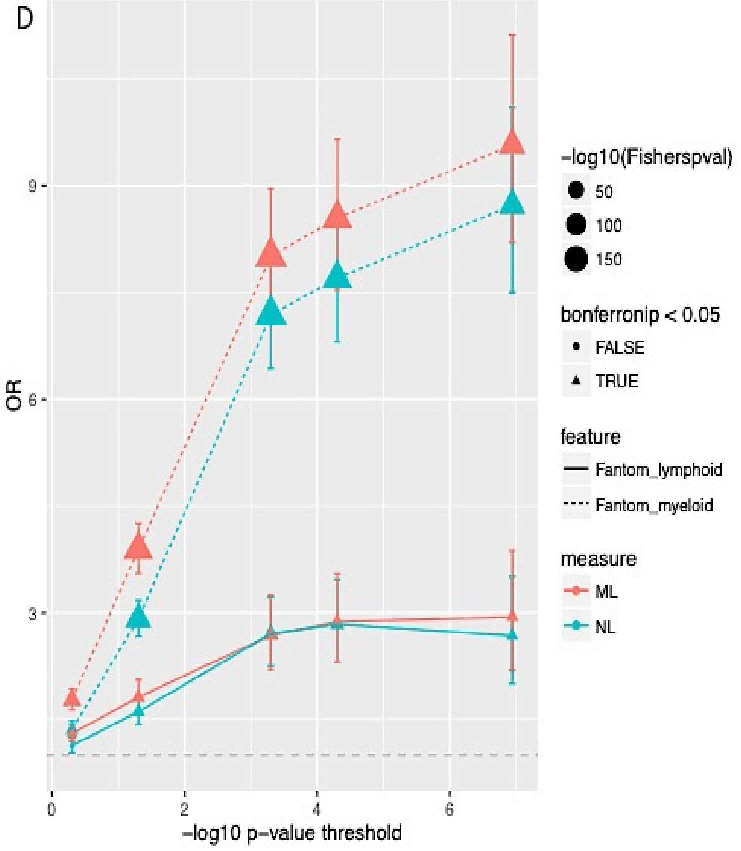
Figure A and **Figure B**. Gene-centric annotations:

TSS1500: from 1.5kb upstream of RefSeq transcription start sites. TSS2000: from 2.0kb upstream of RefSeq transcription start sites. X5'UTR: five prime untranslated region of X gene (the region of an mRNA that is directly upstream from the initiation codon. X 1st Exon: first exon of X gene. Body: the entire gene from the transcription start site to the end of the transcript. 3'UTR: three prime untranslated region of X gene (the region of an mRNA from the 3' end to the position of the last codon used in translation). N_Shelf: North shelf of CpG island. S_Shelf: South shelf of CpG island. N_Shore: North shore of CpG island. S_Shore: South shore of CpG island.Island: a region with a high frequency of CpG sites. It is a region with at least 200 bp, and a GC percentage that is greater than 50%, and with an observed-to-expected CpG ratio that is greater than 60%.

Figure 4C. Enrichment of DMS$_{NLR}$ and DMS$_{MLR}$ with respect to activating histone marks (H3K27Ac, H3K36me3 or H3K4me3), histone marks reflective of active (H3K27Ac) or poised enhancers (H3K4me1) and repressive histone marks (H3K27me3 or H3K9me3) expressed in mature myeloid cells (based on monocyte and neutrophil ChIPSeq data from BLUEPRINT).

Figure 4D. Enrichment of DMS$_{NLR}$ and DMS$_{MLR}$ with respect to expressed enhancers (FANTOM5) in myeloid versus lymphoid cells.





Chapter 9

Epigenome-wide association study for platelet-lymphocyte ratio (PLR) level.

Lin BD, van Dongen J, Boomsma DI, Willemsen G, Abdellaoui A, de Geus EJ, Hottenga JJ.

9.1 Introduction

Platelet–lymphocyte ratio (PLR), derived from the complete blood count, is a novel biomarker, reflecting systemic inflammatory status, which may predict the outcome of cancer, cardiovascular and infectious disease [70, 308-309]. Recently, PLR received increased attention by epidemiologists. The variation in PLR level was found to be influenced by sex and age, but also with environmental conditions and lifestyle, with the effect of lifestyle being dependent on age and sex [147, 151, 187].

The genetic component of PLR was also investigated. We previously reported that PLR is a highly heritable trait with heritability estimated at 64% [187]. A genome-wide association study identified genetic variants in the *HBS1L-MYB* intergenic region that were significantly associated with PLR. The SNP-heritability for PLR was estimated at 14.1%.

A meta-analysis of data from the Netherlands Twin Register (NTR) and TwinsUK identified thousands of methylation sites associated with two other myeloid-lymphoid ratios (MLR and/or NLR, chapter 8 of this thesis). We hypothesize that elevated PLR is also mediated by epigenetic mechanisms such as DNA methylation. However, no study to date has performed a methylome-wide association analysis of PLR. In this study, we performed a genome-wide DNA methylation analysis to identify methylation sites associated with PLR using the DNA methylation data from a Dutch non-patient population. Additionally, we focus on the *HBS1L-MYB* region identified in our GWAS of PLR (chapter 5 in this thesis).

9.2 Methods

9.2.1 Participants

Good quality DNA methylation data obtained from peripheral blood samples were available for 3057 participants in the Netherlands Twin Register [75] (NTR) biobank project [76-77]. We excluded individuals with extreme values for platelet-lymphocyte ratio (PLR) (> mean ± 5 × standard deviation (SD)), individuals with HPA-axis related medication or anti-inflammatory medication, and individuals who underwent cancer treatment, leading to a final dataset of 2876 individuals. The study protocol was approved by the Medical Ethics Committee of the VU University Medical Center Amsterdam, and all participants provided informed consent.

9.2.2 Blood sampling, blood cell counts, and smoking status

Blood samples were collected during home visits following well-established protocols described previously [76-77]. The hematological profile was obtained in EDTA whole blood samples using the Coulter system (Coulter Corporation, Miami, USA). PLR level was calculated as platelet count divided by lymphocyte count. **Table 1** shows the mean and SD of the blood cell traits studied. Information on smoking was obtained during the home visit. Smoking status at the moment of blood draw, coded into three categories (0=non-smoker,1= former smoker, 2=current smoker), was included as a covariate in the analyses.

Table 1. Descriptive statistics for blood parameters of interest.

Population	N	Age	PLR	Platelet count	Lymphocyte count	Neutrophil count	Monocyte count
Total	2876	36.71 (12.84)	119.34 (41.36)	255.4 (61.46)	2.30 (0.76)	3.47 (1.28)	0.54 (0.17)
Male	988	39.96 (13.56)	114.15 (38.54)	239.2 (54.97)	2.24 (0.73)	3.42 (1.46)	0.58 (0.17)
female	1888	36.58 (12.45)	122.06 (42.53)	263.8 (62.97)	2.34 (0.77)	3.50 (1.35)	0.51 (0.16)

9.2.3 DNA methylation

The Infinium HumanMethylation450 BeadChip (Illumina Inc, San Diego, CA, USA) was used to measure DNA methylation in 500ng of genomic DNA obtained from whole blood following the manufacturer's protocol. The quality control steps and normalization of the data have been previously described [298] . The analysis included all autosomal methylation sites (N=411,169). The analyses were performed on methylation β-values, which represent the methylation level at a site, ranging from 0 to 1, and are calculated as: β=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.

9.2.4 Statistical analysis

We performed an epigenome-wide association study (EWAS) of platelet- lymphocyte ratio (PLR) level, based on measured white blood cell and platelet count data. All statistical analyses were performed using R programming language. In the EWAS, we tested for each CpG site, whether the methylation β value was associated with PLR. We used generalized estimation equation (GEE) models, with DNA methylation β value as outcome and the following predictors: PLR, sex, age at blood sampling, smoking status, HM450k array row, sample plate and, added at a second step, blood cell counts (platelets, monocytes, neutrophils and lymphocytes). GEE models were fitted with the R package GEE, with the following specifications: Gaussian link function (for continuous data), 100 iterations, and the 'exchangeable' option to account for the correlation structure within families and within persons. We applied two methylation association study models for PLR:

model 1: methylation **β value** ~ **PLR** + Sex + Age + Smoking + sample plate + 450k array row.

model 2: methylation **β value** ~ **PLR** + #Platelet + #Neutrophils + #Lymphocytes + #Monocytes + Sex + Age + Smoking + sample plate + 450k array row.

9.3 Results

9.3.1 EWAS

In model 1, which did not correct for individual blood cell counts, 92060 methylation sites were genome-wide significantly associated with PLR level following Bonferroni correction. The QQ-plot is provided in **Figure 1**. In model 2, which corrected for white blood cell counts and platelet count, 3629 methylation sites were significantly associated with PLR (DMS_{PLR}). At 83% (3029) of DMS_{PLR}, an increase in PLR level was associated with hypomethylation. The Manhattan plot and QQ plot for PLR level, platelet count and lymphocyte count are shown in **Figure 2** to **4**.The number of methylation sites significantly associated with individual counts are 3840, 11189, 6496 and 38071 for platelet count (DMS_{PLT}), neutrophil count (DMS_{NEUT}), monocyte count (DMS_{MONO}), and lymphocyte count (DMS_{LYMPH}), respectively. The overlap of $DMS_{PLR,}$ DMS_{PLT}, and DMS_{LYMPH} is shown in **Table 2**. The overlap between DMS_{PLR} with DMS_{PLT} (1,971 CpGs) and between DMS_{PLR} with DMS_{LYMP} (1,496 CpGs) was large, and the overlap among the three parameters of interest was 965 CpGs. In total, 1,112 CpG sites were significantly associated with PLR but not with the individual platelet and lymphocyte counts.

Table 2. The number and overlap of significant methylation sites for PLR, platelet count and lymphocyte count.

	Methylation sites
DMSPLR(raw)	92,060
DMSPLR(corrected)	3,467
DMSPLT	3,840
DMSNEUT	11,189
DMSMONO	6,496
DMSLYMPH	38,071
DMSPLR&PLT	1,971
DMSPLR&LYMP	1,496
DMSPLR&PLT&LYMP	965

9.3.2 Co-localization of methylation signal and GWAS hit for PLR

In our GWAS study of PLR (chapter 5) we identified one locus that was significantly associated with PLR level: rs9376092 (location: chr6: 135427144) in the intergenic region between the *HBS1L* and *MYB* gene. To examine co-localization of SNPs and methylation sites associated with PLR, we zoomed in on the EWAS results of all methylation sites within a range of +/-200kb around this SNP. Our dataset contained 53 methylation sites in this region. The regional plot, which was generated by coMET package in R [310], is shown in Figure 5. Following Bonferroni-correction for 53 tests, methylation at 5 CpGs in the region are significantly ($p < 0.05/53$) associated with PLR. The nearest significant CpG, cg14743594, is 51kb upstream of rs9376092 and located in HBS1L. The other 4 CpGs are located in MYB; 12-13kb downstream of rs9376092. At all 5 CpG sites, a higher PLR level was associated with hypomethylation.

9.4 Conclusion

In this EWAS study, we have identified 3,629 CpGs sites associated with PLR levels, of which 1,112 CpG sites are uniquely associated with PLR level rather than with individual blood cell counts. In the region harboring the significant SNP from our GWAS of PLR level, we observed 5 CpGs for which hypomethylation was significantly associated with a higher PLR level, when applying a Bonferroni correction for the number of interrogated methylation sites in the region surrounding the SNP. This suggests that both genetic variation and variation in methylation level in this region are associated with PLR levels. It is important to note that the methylation sites that are most strongly associated with PLR (reaching genome-wide significance) do not overlap with the GWAS region.

Figure 1. Manhattan plot and QQ plot for PLR without individual blood cell count correction (model 1).

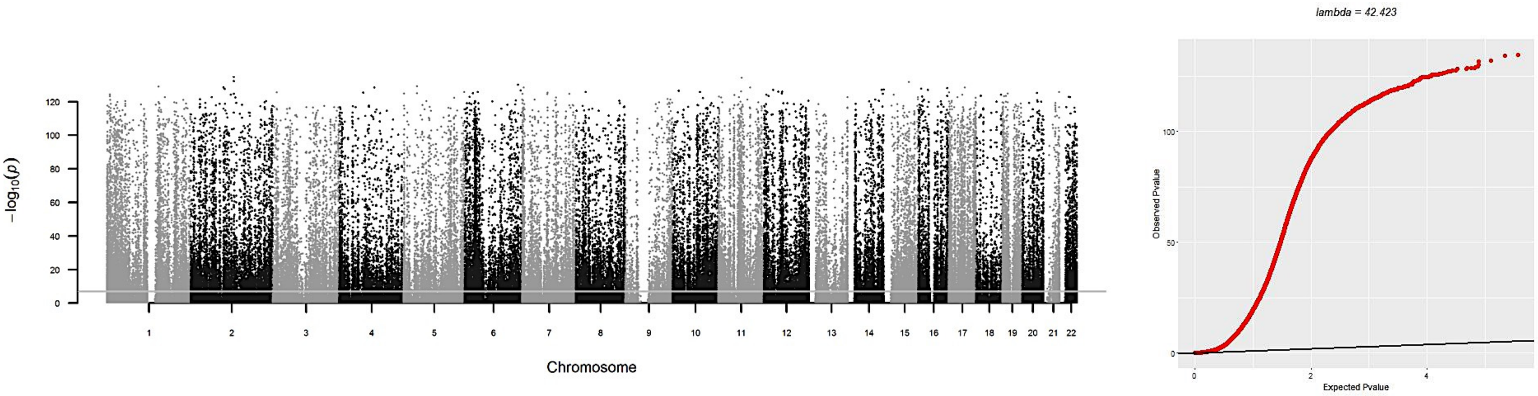


Figure 2. Manhattan plot and QQ plot for PLR with individual blood cell count correction (model 2).

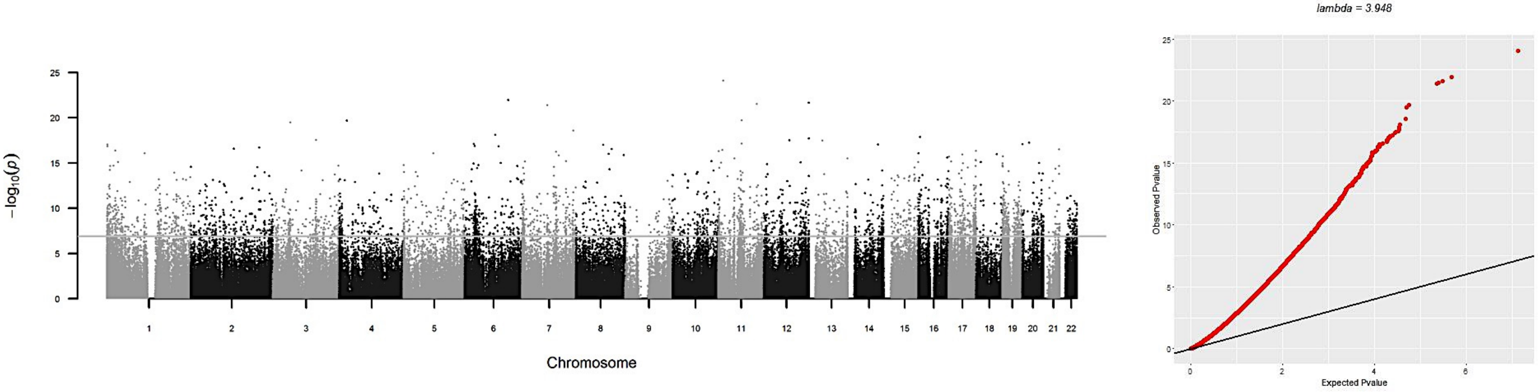


Figure 3. Manhattan plot and QQ plot for platelet count in model 2.

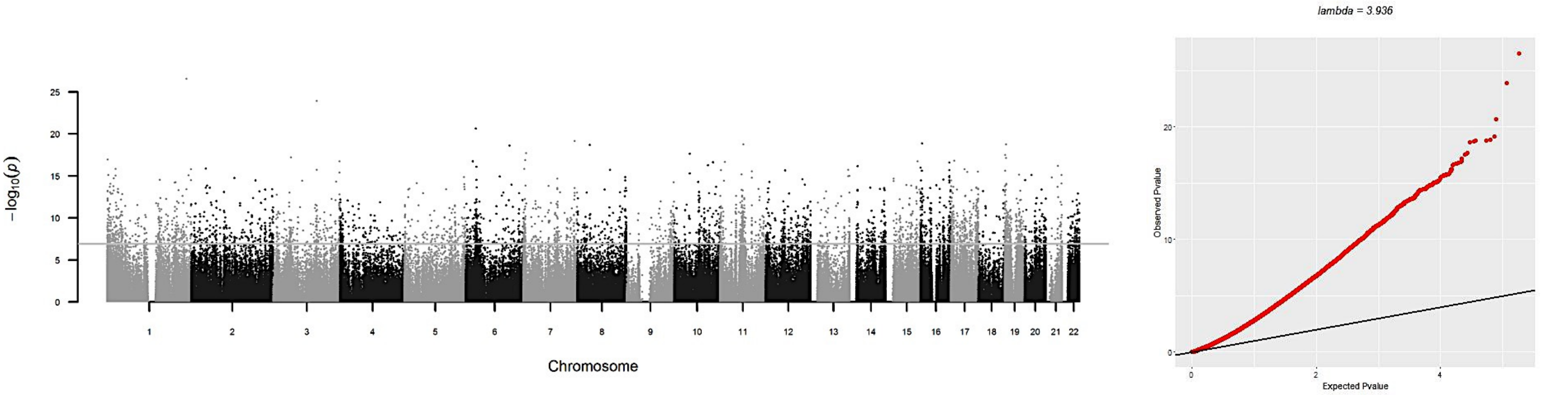


Figure 4. Manhattan plot and QQ plot for lymphocyte count in model 2.

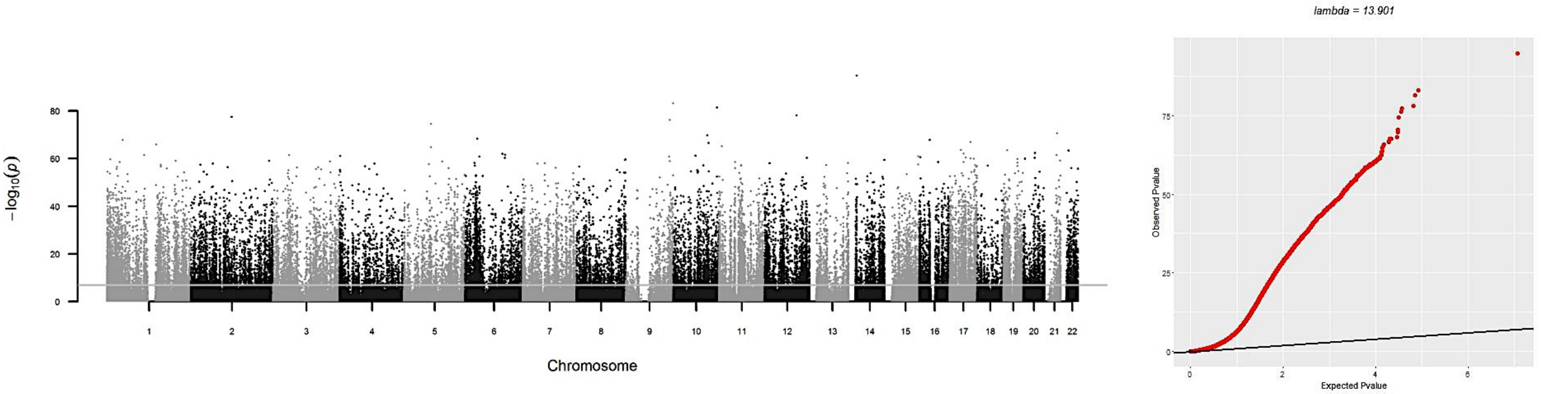
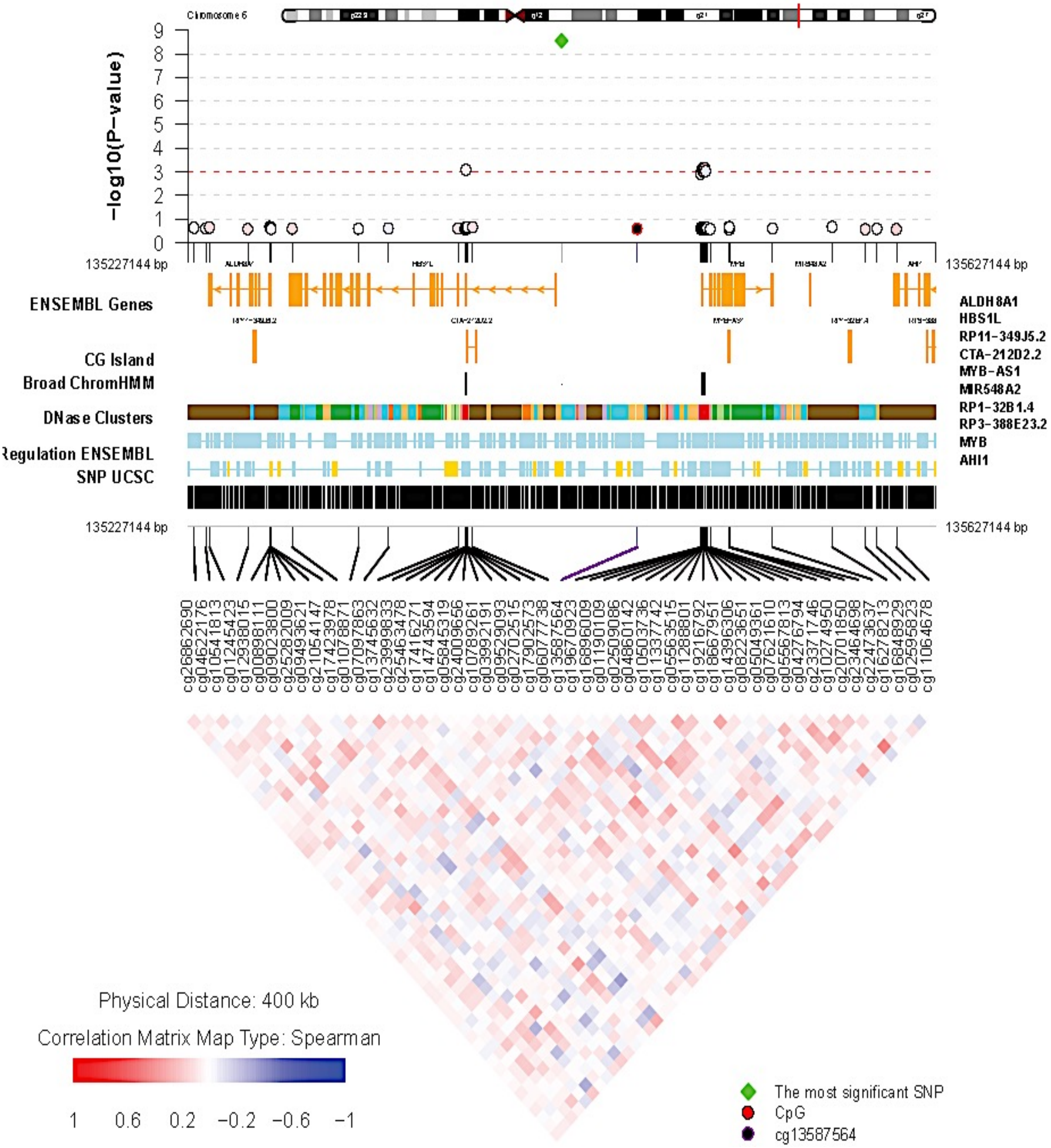


Figure 5. Regional plot showing EWAS results for the region harboring the most significant SNP for PLR level.



The figure shows Chr 6: 135,227,144 - 135,627,144 which is 200kb around rs9376092 (indicated in the green color). rs9376092 is the most significant SNP (β=5.483, p=2.75E-9) associated with PLR level from Chapter 5. The closest significant CpG is cg14743594 (p= 0.00087) indicated in purple.

Chapter 10

General Summary and Discussion

10.1 General Summary

The present dissertation focused on the genetic influence of human complex traits with two main research topics: the genetics of pigmentation traits and the ontology of hematological traits. In this chapter, I will first summarize the most important results. In the second part of the chapter, I will discuss these findings as well as reflect upon the current state and future direction of complex trait genetics.

Chapter 2 investigates the genetic architecture of hair color. The broad-sense heritability of hair color was estimated to be between 73% and 99% and the genetic component included non-additive genetic variance. This is an estimate of pigmentation heritability obtained in the study with largest statistic power so far (sample size =22,000 subjects). Assortative mating was present for most pigmentation traits, reflecting that mate choice is to a fairly large extent governed by visible trait characteristics. Of course, the pigmentation spectrum distribution correlates with latitude and this contributes to the correlation of spouses. Known pigmentation loci were confirmed in the Genome-Wide Association (GWA) analysis for each phenotype of interest: MC1R region for red, brown and black, and light versus dark hair color; *TPCN2*, *IRF4*, and *KITLG* for blond, brown and light versus dark hair color; *SLC24A4* for blond, brown and light versus dark hair color, green eye color, and blue eye color; *HERC2* for blond, brown, light versus dark hair color, blue and brown eye color; and PTPRT for green eye color. At most 24.6% of the additive genetic variance in hair color (specifically red hair color) was explained by the 1000G well-imputed SNPs in GCTA analysis.

Chapter 3 presents a bivariate analysis of eye and hair color. I detected strong genetic correlations between various combinations of hair and eye colors with GCTA: a significant positive correlation between blue eyes with blond hair (0.87) and brown eyes with dark hair (0.71), and a significant negative correlation between blue eyes with dark hair (-0.64) and brown eyes with blond hair (-0.94). In addition, GWAS results for hair color and eye color also indicate a large genetic overlap between eye color and hair color: *HERC2* is significantly associated with blue eyes, brown eyes, blond hair, brown hair and dark hair color; *SLC24A4* is significantly associated with blue eyes, green eyes, blond hair, brown hair, and dark hair color. Like hair color, eye color prevalence has correlated tightly with latitude over the past centuries. Consequently, eye color also correlates with Principal Components (PCs) which represent ancestry and describe the genetic stratification of the

Netherlands. Including PCs into genetic studies of such phenotypes, which underwent simultaneous genetic divergence between (sub)populations could underestimate the genetic association. Therefore, I suggested that when conducting gene finding studies or GCTA analyses, the effects of ancestral population differences on the relationship between stratified traits should be carefully considered.

Chapter 4 investigates the cause of variation in neutrophil-lymphocyte ratio (NLR) and platelet-lymphocyte ratio (PLR) in the Dutch population. These immune biomarkers are moderately (NLR h^2=35%) to highly heritable (PLR h^2=64%). The correlation between NLR and PLR is 0.49. PLR correlated neither with CRP nor with IL6. However, NLR correlated positively with CRP (0.157, $p< .001$) and with IL6 (0.083, $p < .001$). Compared to women, men had higher average NLR levels, but lower average PLR levels. PLR, and to a lesser extent NLR, increased significantly with age. Sex specific effects also were observed for seasonal differences: in colder months NLR and PLR were on average higher in women, but not in men. In addition, small but significant, age and sex specific associations of NLR and PLR with BMI and smoking behavior were observed.
Our comparison of healthy individuals with subjects who had a possibly compromised immune system showed higher NLR and PLR levels along with higher CRP and IL-6 levels in the unhealthy individuals. When including the unhealthy individuals into the analyses, heritability estimates were lower than in the healthy population, but since the confidence intervals overlapped, the difference between the total and the healthy population was not significant.

Chapter 5 investigates the genetic variants for NLR and PLR. I identified genetic variants in the intergenic *HBS1L -MYB* region which were significantly associated with PLR level, and also affected platelet count. Other loci like *PSMD3* for neutrophil count or *CCDC71L-PIK3CG, ARHGEF3* and *BAK1* for platelet count were not associated with the blood ratios derived from them. These loci likely regulate specific types of white blood cells count rather than the balance of subtype blood cell counts. For most of the top SNPs obtained in our GWAS analyses, we found significant cis and trans eQTL effects related to the expression of genes involved in hematological and immunological pathways. Although the overall correlation between NLR and PLR was 0.49, no significant genetic correlation between the two ratios was found. The results point to the complexity of the genetic underpinnings of immune response regulation. Since NLR and PLR serve as biomarkers for the development of some of the same diseases, their correlation might be mainly due to environmental stimulation and genetic environmental interaction.

Chapter 6 focuses on the phenotype of monocyte-lymphocyte ratio (MLR) and its subcomponents: monocyte count and lymphocyte count. The approaches that were used were similar to those in chapters 4 and chapter 5. The heritability and fraction of heritability explained by associated genetic variants were estimated. The heritability was

40% for MLR, 58% for monocyte count (12% non-additive genetic effects) and 58% for lymphocyte count (21% non-additive genetic effects). The genetic correlation between MLR and monocyte count was 0.48 by LD regression (se=0.3162, p=0.123). For MLR, I identified a locus nearby *ITGA4* which is a well-known locus for monocyte count. This *ITGA4* locus was also associated with monocyte count in my study, as were three other loci ((*LPAR1, IRF8* and *ITPR3)*. All the genetic variants together explained a small part of phenotypic variance, which motivated us conduct an epigenetic study for these immune biomarkers in Chapter 8.

Chapter 7 focuses on the simultaneous analysis of multiple variables, which is a main challenge of high dimension data analysis to avoid loss of information and gain statistic power. The haematological profile was examined as a function of age, sex, smoking, BMI, and their two-way interactions, in both univariate and multivariate analyses for the same dataset of hematological profile variables (including neutrophil count, lymphocyte count, monocyte count, eosinophil count, hemoglobin level, mean corpuscular volume, mean corpuscular hemoglobin concentration, red cell distribution width, platelet count, and mean platelet volume). Compared to univariate regression results, multivariate distance matrix regression (MDMR) analysis resulted in relatively smaller p-values for all main effects and provided evidence for interaction effects (age × sex, age × smoker and age × BMI interactions). The results show that MDMR increases the power to detect important interactions within predictors and may help identify subgroups who benefit from different treatment or prevention measures.

Chapter 8 investigates epigenetic variants for NLR and MLR. I examined the prediction of the distribution of blood cell counts using the whole methylation profile according to the Houseman method. The correlation between predicted indices and measured indices was larger than 0.8, demonstrating that the Houseman method is a valid way to estimate the blood cell count distribution in samples where blood cell count measurements are not available. This method was then used to obtain blood cell count distribution in the TwinsUK sample. Next, a meta-analysis EWAS of NLR and MLR, which combined datasets from NTR and TwinsUK was conducted. I reported 4185 methylation sites associated with MLR and/or NLR that were not associated with individual cell count components. Many of the sites were overlapping between NLR and MLR and the pattern of results found is consistent with a model in which long-term hematopoietic stem cells (LT-HSC), which have constitutive DNA methylation, actively undergo demethylation at myeloid-specific regulatory regions to give rise to myeloid-biased progeny leading to higher MLR and NLR.

In **Chapter 9**, epigenetic variants for PLR level were investigated. A genome-wide DNA methylation analysis was conducted to identify methylation sites associated with PLR using the DNA methylation data from a Dutch non-patient population. I identified 3,629 CpGs sites associated with PLR levels, of which 1,112 CpG sites were uniquely associated with PLR level rather than with individual blood cell counts. Similar to the results for the NLR and MLR EWAS , hypomethylation of the majority of these methylation sites (83%) was associated with high PLR level. Additionally, I focused on the *HBS1L-MYB* region identified in our GWAS of PLR. I found both genetic variation and variation in methylation level in this region to be associated with PLR levels: there were 5 CpGs for which hypomethylation was significantly (Bonferroni correction threshold: p < .00094) associated with a higher PLR level. However, the methylation sites that were most strongly associated with PLR (reaching genome-wide significance) did not overlap with *HBS1L-MYB* region from GWAS study in Chapter 5.

10.2 Discussion

In this dissertation, I introduced several approaches currently utilized in human genetic studies and applied them to the main characteristics of two traits of interest: pigmentation and hematology. Different statistical-genetic approaches are appropriate for different kinds of questions and phenotypes based on features of the trait and hypotheses underlying its etiology. Both the pigmentation and hematology trait investigations started with extended family twin designs to estimate heritability. Different results were obtained for the two traits: the two pigmentation traits hair and eye color are highly heritable (>73%), while the heritability of the studied immune biomarkers varied between 35% and 60%.

For highly heritable and possibly less polygenetic traits, linkage studies have successfully mapped QTLs for hair color and eye color, even if the DNA-marker resolution was not high [106]. However, it is clear that more loci remained to be discovered and the current genome wide association studies (GWAS) are well suited to do this.

Hematological traits are expected to be more polygenic compared to pigmentation traits, involving many genes with small effects. It is difficult to detect loci for such polygenic traits by linkage analysis; other approaches are needed to identify the individual genetic variants involved. As we have seen in the past few years, such traits may successfully be analyzed in GWA studies. Compared to pigmentation traits, immune biomarkers, which show more phenotypic plasticity, are influenced more strongly by environmental factors. Therefore EWASs, examining the association of epigenetic variants and measurable phenotypes, are valuable, given that epigenetic variants may be influenced by non-genetic effects such as environment, in addition to stochastic variation and measurement error.

Twin studies provide an upper-bound limitation for genetic studies of a trait

For a geneticist, the most basic question one would ask about a trait is whether it is heritable in a human population, In other words, whether the observed variation in the phenotype can be explained by genetic variation. It is important to note that this is not the same as asking whether genes play a role in the trait. Gene-mediated developmental processes lie at the basis of all traits (including behavioral traits), but phenotypic variation among individuals is not necessarily the result of genetic variation. In addition, heritability is a population-based concept, which is not informative about the individual. A heritability of 0.96 for blond hair color indicates that, on average, 96% of the observed individual differences are attributable to genetic differences. It does not mean that 96% of any person's hair color is due to his/her genes and the other 4% is due to his/her environment. In addition, heritability estimates do not reveal anything about the specific genes that contribute to a trait. Similarly, a numerical estimate of environmental effects could not provide any information about the important environmental parameters that influence a trait. However, heritability studies will provide an upper bound of genetic variation contributing to a trait, which can determinate the necessity of follow-up genetic studies.

Twin studies often provide the first estimates of total heritability; there are few heritable traits for which a twin study has not been carried out [11]. It has been suggested that heritability estimates may be overestimated [311], for example due to genetic interactions (e.g. epistasis). The evidence for epitasis, which is a hidden complexity of genetic regulation in complex traits, can be examined in experimentally amenable organisms, or with statistical models that take interaction of genetic variants into account [312]. One such model is the ADE model, which can be fitted to data from MZ and DZ twins, where dominance (D) variance components represent the genetic interaction term. In human data, epistatic effects are then included in the dominance variance component and contribute to the broad-sense of heritability of a trait.

Estimates of heritability in twin studies may be biased if shared environmental factors are not properly accounted for in the model. In small studies, the statistical power to estimate the influence of common environmental factors (C) is often low, which may mistakenly lead to conclusions that the genetic model included additive genetic factors (A) and unique environment (E) only. The estimation of heritability also depends on whether we consider the covariance and interactions between genotype and environments (shared environment C, or unique environment E) [313], which give rise to an additional source of phenotypic variance. Ignoring the interaction or correlation of genotype and environments will result in biased parameter estimates. However we know which parts of the model are affected: interaction between A and C acts like A; interaction between A and E acts like E, whereas correlation between A and C acts like C; and correlation between A and E acts like A. We can thus interpret the results from twin studies in light of this knowledge. If no C is found, a correlation between A and C is unlikely to be present, if there is A x C then ignoring interaction between A and C will overestimate the heritability. The adoption-twin design in which twins reared in different environments are included can be utilized to explore interaction between genotype and environment. The variance of a phenotype can be decomposed into the variance between genotypes, the variance between specific environments and the variance attributed to the interaction of genotype and environment. Individuals with particular genotypes may seek out particular environments. Such genotype environment correlations where subjects choose environments driven by genotypes can be also regards as a part of heritability. Results from the famous adoption-twin studies from Minneapolis are consistent with results from classical twin design, which suggest that the similarities between twins are due to genes, rather than environment or gene environment interaction [314].

Just as means and variances are population parameters, heritability is specific to a particular population

in a particular environment, and when the environment changes, the heritability can change as well. Heritability estimates cannot be used to determine the causes of phenotypic variance between populations, in other words, to distinguish whether the differences between populations are determined by genes or by environment, unless genotype data, or environmental exposures are available for both populations. Heritability estimates may differ depending on how the sample is defined and drawn from the population. For example, when the hematological profile heritability study in this dissertation, included a random sample that also contained unhealthy sample of participants, an additional source of variation was introduced, which resulted in an somewhat lower point estimate for heritability compared to the estimate based on a healthy sample only, though the confidence intervals overlapped.

GWAS unravels genetic variants associated with a trait

There is no doubt that the GWAS design has become a very useful and common method to understand the genetic basis of complex traits and to understand the etiology of heritable diseases. It provides a very efficient way to identify representative genes and pathways. These identified genetic variants are expected to play a role in clinical applications such as early diagnosis of diseases, identifying drug targets, and personalized medication. As I stated in my introduction, due to the virtues of GWAS, we have better understanding of the genetic architectures of many complex human traits such as height, BMI, pigmentation, hematological profile and also human diseases. Many significant loci have been identified by GWAS, which would never have shown up in candidate genes studies, such as the genetic variants in non-coding regions for type 2 diabetes, rheumatoid arthritis, obesity, cancer and coronary heart disease [315]. Our GWAS for pigmentation traits and hematological traits replicated some known genetic variants (*PSMD3* for neutrophil count; *HBS1L-MYB, CCDC71L-PIK3CG*, *BAK1* and *ARHGEF3* for platelet count) and identified genetic variants for PLR (in *HBS1L-MYB* intergenic region), which contributed to new genetic knowledge of these traits. A high and significant genetic overlap between eye color and hair color was identified by both GWAS and GCTA. The genetic variants that were detected in the HBS1L-MYB region for the immune biomarker PLR in the healthy population suggest that this region is possibly involved in the differentiation direction of hematological stem cells .

The genome-wide genetic association studies require robust findings , with results that cross a threshold of significance level of 5×10^{-8} in order to avoid type I errors. To achieve this statistical rigor, identification of individual genetic variants generally requires big sample sizes. The larger the sample size of the GWAS conducted, the more genes tend to be identified with higher confidence. In recent years, scientist in genetics have collaborated in their efforts at an unprecedented scale and set up several consortia to conduct comprehensive GWAS meta-analyses. Consequently, many QTLs and related trait genetic variants were identified. Such an effort is currently also ongoing for hair color genes. However, none can guarantee that GWASs will reach sufficient power to detect all genetic variants related to a trait given the expected small effect sizes of many still undetected SNPs [316]. The narrow-sense heritability estimated by twin studies provides an upper-limit of the variation due to genetic effects. GWASs for eye color and hair color have identified and verified multiple pigmentation genes such as *MC1R*, *TYR*, *OCA2*, *HERC2* and *SLC24A4.* The genetic variants on these genes have explained a large part of the heritability of hair color (24.6%) and eye color (88.5%).

With the ambition of increasing the sample size, larger meta-analyses of GWASs containing multiple cohorts are being conducted more frequently. The latent structure of populations became one of the main issues in association studies, because population stratification may yield spurious associations if not properly addressed [317]. As the basic assumption of a GWAS is the analysis of independent individuals from homogenous populations, any kind of shared ancestry that is not accounted for can introduce biased estimates [318]. Shared ancestry can have two main sources: cryptic or unknown relatedness and population stratification.

Family structure (including cryptic relatedness) refers to kinship within a sample that needs to be taken into account in association studies, especially when family based data are utilized [319]. The most straightforward way to address this issue is to select unrelated individuals, which results in compromised statistical power (if the genotyped sample was larger than the selected sample of unrelated people). Another method is using the pedigree information, clustering the family memberships, and fitting the data into a model as I did in chapter 2 and chapter 3 for the pigmentation trait GWAS. This method is, however, inappropriate in the presence of cryptic relatedness, where the family structure is unknown to the investigator. To control for unknown family structure, we can infer the relationship of any pair of individuals in the population based on genome-wide SNP data in a genetic relationship matrix (GRM), which we can include in a mixture model [215]. The GRM can be estimated by all autosomal SNPs or by the majority of autosomal

SNPs (all SNPs except the chromosome where the SNP that is being tested resides: Leave-One-Chromosome-Out: LOCO method) [212]. The virtue of this method is that pedigree information is not needed and relatives do not have to be excluded, which retains the number of subjects at a maximum to keep the statistic power (it is more computationally intense however). In the GWASs in chapters 5 and 6 I utilized a LOCO method to correct for family structure, which gives a non-inflated P-value distribution, successfully correcting for the shared ancestry in our dataset.

Another potential problem can arise when data are included from more than one subpopulation, and a false positive result can arise by different allele frequencies across the subpopulations rather than true different allele frequencies between case and control groups [320]. The approaches to avoid this issue are multiple: optimally defining the population of interest (excluding ethic outliers), correcting for principle components, and using mixed models or genomic control (LOCO method). Population stratification needs to be considered even if the population of interest is assumed to be relatively homogeneous and localized in a small geographic area like the Dutch population. The main pattern of population structure can be summarized by uncorrelated principle components (PCs) obtained from a principle component analysis (PCA) of genome-wide genetic data. Based on the Dutch demographic history and geographic landscape, the Dutch population can be assumed to be relatively homogenous with relatively low migration rates in recent Dutch history. Regarding the NTR dataset, Abdellaoui et al [103] used the 1000 Genome dataset to exclude individuals with a non-Dutch ancestry, and using stringent quality control and stringent LD pruning criteria, computed principal components that captured Dutch ancestry differences [321]. The first three Dutch PCs highly correlated with geography (PC1: North-South, PC2: East-West, and PC3: Middle-Band distribution) which indicates that these PCs indeed reflect ancestry differences [103].

In our hematological traits studies, I took the first three Dutch PCs into account in GWAS and GCTA analysis, thus successfully correcting for population stratification. However, because Dutch PCs are partly representing the phenotypic variability in pigmentation traits, the significance of the association might be alleviated when including the first three Dutch PCs as covariates in the pigmentation analyses. Therefore, ancestral population differences for such stratified traits should be carefully considered when conducting genetic studies.

Pleiotropy

Besides conducting the GWAS for each individual trait, I also investigated the genetic relationship between different phenotypes. Pleiotropy describes the phenomenon of genetic effects of a single gene, or set of genes, on multiple phenotypic traits. Pleiotropy was first observed by Gregor Mendel who hypothesized that three distinct traits of pea plants seemed to be inherited together: brown seed coat, violet flowers, and axial spots

[322]. However, based on the techniques of that time, it was unknown if this inheritance manner was due to physical linkage of multiple distinct genes or to a single gene affecting multiple traits (genuine pleiotropy). Molecular studies have shed light on pleiotropic mechanisms: a single locus may produce different products by alternative splicing, alternative start/stop codons, and modifying protein structure after translation. A classic example of pleiotropy in humans is phenylketonuria (PKU) which leads to mental disorders and pigmentation changes: the mutations in the PAH gene can cause mental retardation and depigmentation of hair and skin. The molecular mechanism is clear: the PAH gene codes for the rate-limiting enzyme phenylalanine hydroxylase, which converts the amino acid phenylalanine to tyrosine. The PAH mutation leads to reduced activity of the enzyme, which results in an abnormal high phenylalanine concentration that is toxic to the developing nervous system and a low tyrosine concentration that is a key substrate of the melanogenic pathway [323].

The genetic correlation between traits can be estimated by multiple methods such as bivariate and multivariate twin studies [13], bivariate data in subjects with genome-wide SNP data in GCTA [126], and a more recent method LD regression [208]. These methods require different types of information to explore the genetic overlap between two traits. Twin studies do not require genotype information, but need explicit relatedness information to decompose phenotypic covariance into genetic and non-genetic components. GCTA requires genome-wide SNP information to estimate a genotypic relationship matrix (GRM), which is then used to model the phenotypic covariance. LD score regression can be done using only GWAS summary statistics. In combination with information on the correlations between the SNPs (LD score), the genetic covariance is estimated by the product of two slopes obtained in regression for both phenotypes to detect if SNPs have concordant effects on both phenotypes. However, these methods fail to answer which genetic variants have pleiotropic effects, whether these genetic variants have concordant or discordant effects across traits, and whether these effects act in a common or different biological pathways [324].

Recently, Pickrell et al [325] analyzed results from large scale GWAS on 42 traits including hematalogical traits such as platelet count and MPV, physical traits, behaviors, immune diseases and psychiatric disorders. More than 300 loci were associated with more than one phenotype. The genetic etiologic links between distinct traits were identified and the genetic correlations among these traits estimated, which showed several tight clusters of related traits such as hematological traits.

Some of the genes identified in our studies are also associated with multiple traits. For example, *MC1R* was associated in our study with red hair color [326], but has also been associated with higher Parkinson's disease risk [327]. I also identified and confirmed that *HERC2* and *SLC24A4* are associated with both hair color and eye color. Such findings demonstrate etiological links between traits and knowledge about genetic variants that are associated with multiple phenotypes may help unravel complex issues such as comorbidity of disorders. This is even more complex than we may realize: pleiotropic variants may not only have concordant effects on two or more diseases, some variants may also have discordant effects (risk factor for one disease but protective factor for the other) as has been shown for immune disorders [328]. Pleiotropy is also more common than was expected: a protein network study has shown that each gene influences on average four to five traits [329] and acomparative study of protein sequencing and gene expression suggests that on average, one gene affects six or seven traits [330]. From an evolutionary point of view, the ubiquity of pleiotropy can be a consequence of fitness for rapidly adapting to dynamic anthropogenic environments [329]. Insight into the complexity of pleiotropy can facilitate our understanding of the wide range of trait combinations within current human populations, and also shed light on the etiology of genetic diseases. However, it may also make targeting specific gene mutation disorders with medicines much more difficult.

The gap between twin studies and GWAS

Although GWASs have discovered thousands of complex traits related variants, these variants often explain only a minor part of the heritability. The SNP heritability can be estimated by multiple methods: such as GCTA [331], LD Score regression [208], and GEMMA (Genome-wide Efficient Mixed Model Association) [332]. GCTA can estimate to what extent phenotypic variation is explained by selected genetic variants. This analysis can cover candidate genes, but also findings at chromosome or whole genome-wide level to answer the question which percentage of variance is explained by measured genetic variants, the genetic profile on the chromosome of interest, and the associated SNPs based on GWAS data or GWA summary statistics. The classical GCTA estimates the extent to which the phenotypic similarity across pairs of individuals in a sample is explained by their genotypic similarity at common variants by restricted maximum likelihood (REML). However, SNP-heritability can be underestimated by GCTA, because of several reasons, including incomplete genomic coverage and not including genetic interaction in the model. A novel SNP heritability estimation approach based on GCTA called GREML-LDMS method (LD- and MAF-stratified GREML) accounts for both linkage disequilibrium and rare variants. The strategy is similar to classical GCTA, but the REML analysis is performed using multiple LD and/or MAF stratified GRMs. SNP heritability estimated by GREML-LDMS method has been proven to be robust, unbiased and independent of the MAF and LD properties of causal variants, which account for the majority of narrow-sense heritability from twins study [333].

Beyond genetic variation: Epigenome-Wide Association Studies

Epigenetic variation has come to the attention of geneticists since a substantial proportion of variation in complex traits remains unexplained by GWAS, and this may be particularly relevant for highly dynamic and environmentally-sensitive traits such as immunological biomarkers. By interrogating the epigenome of a population, Epigenome-Wide Association Studies (EWASs) may provide a better understanding of the (epi)genetic architecture of complex traits and the etiology of diseases. Family-based data especially twin data is very valuable for EWAS [334]. For example, monozygotic twin pairs have identical genetics and very similar prenatal environments, but the twins may still differ, sometimes to a large extent, in their phenotypes. Comparison of discordant MZ twin pairs in EWAS studies can exclude genetic differences, which may point to disease-associated epigenetic marks. Compared to the high MZ twin correlations (near 100%) for SNP data, the MZ twin correlations for DNA methylation are low (on average 0.20 in whole blood [298]and 0.31 in buccal [335] across all CpGs), resulting in enough epigenetic variation within MZ pairs to make an investigation of these differences informative.

EWAS may yield new insights into the causes of complex traits. Using GWAS, we determine which genetic variants are associated with a phenotype, but additional studies are needed to explain other sources of variation. The direction of causality in GWAS is from DNA to phenotype, but in EWAS studies causality may be in both directions. The causality of epigenetic variants may be detected by longitudinal EWAS studies. Because epigenetic variants are dynamic and dependent on environments and conditions, tracing the changes in the epigenetic profile of an individual across time, using longitudinally collected samples, or in different environments can provide useful information on how epigenetic variants mediate disease development or trait variation [336].

While the DNA profile is uniform across the whole body, epigenetic profiles can differ across tissues [337]. For example, distinctive DNA methylation patterns contribute to distinct cell functions and cellular identity. Given that the methylation of DNA from whole blood constitutes a heterogeneous profile and different cells are characterized by their own differentially methylated regions (DMRs), knowledge about cell counts is important in EWA studies. When cell counts are not known for a sample, the proportions of white blood cell components may be predicted from the methylation profile of whole blood samples, by using a reference dataset [296]. In our study, we compared measured white blood cell proportions to predicted proportions in NTR and identified more than 4000 methylation sites associated with NLR and MLR through EWAS meta-analysis based on predicted cell counts from two cohorts (NTR and Twins UK), one of which did not have information on measured cell counts (TwinsUK). Using measured platelet and lymphocyte counts, we identified 1,112 methylation sites for PLR in NTR.

Functional studies of identified genetic or epigenetic variants

GWASs have identified genetic variants for human complex traits, which tend to be located outside coding regions in the genome [338-339]. Genome function analysis efforts therefore also focus on genetic variants in non-coding regions by carrying out expression quantitative trait loci (eQTLs) analyses [340].

To detect the causal effects for the genetic variants for phenotypes of interest, we have combined our GWAS and EWAS results with RNA expression databases. The significant association between the genetic variants with quantitative amount of transcripts can unravel whether SNPs play a regulating role on the gene expression level. In our study, I identified eQTL cis and trans effects on gene expression involved in multiple pathways such as platelet activation, signaling and aggregation; immune system; metabolism; cell division, proliferation, and differentiation; and genes playing a key role in hematopoietic stem cell differentiation pathways and lineage-specific markers. In addition, I tested if these effects remain, when correcting for cell counts. A number of eQTL effects, but not all, which were present in uncorrected gene expression data, disappeared in the corrected data, possibly because the genetic variants directly influence the variation of blood components and influence the plasma gene expression level.

The emergence of eQTLs may provide an easily accessible and interpretable link to understand the gap between genetic variation and phenotypes. eQTLs analysis, as a new biological function study, provides insight into human genetics, which cannot be gained by animal knock-down experiments. Utilizing eQTLs will largely strengthen the interpretation of the genetic variants mechanisms from genome-wide studies and will add to our understanding of the biological pathways involved in human complex trait variability.

Phenotyping

In all genetics-related research, a careful definition of the phenotype is critically important, as it helps to define the biological pathway of interest. A well-defined phenotype, which perfectly reflects the research question, is one of the first and foremost factors to successfully conduct genetic research. Phenotype decisions include whether the variable should be binary, categorical or continuous but also whether variables should be considered independently or jointly in a multivariate approach. For example, regarding the hematological profile, a single parameter fails to represent the whole profile, so the multivariate analysis becomes more informative as was shown. The virtues of multivariate analysis compared to univariate analysis in our study were clear: by retaining the information between variables we increased statistical power. Such a strategy of combining phenotypes in different ways should be encouraged when new research questions are considered by a geneticist.

Next generation sequencing

Despite the extensive discovery of common (epi)genetic variants (SNPs and CpGs), rare variants and copy number variants are other main components to explain additional disease risk or trait variability. Rare variants, which often are defined as MAF < 0.01, are known to play an important role in human complex traits [341]. For example, highly penetrance rare variants have been found to contribute to many Mendelian disorders and rare forms of common traits (including red hair color [342]) [343]. In addition, low-frequency and rare variants are found to be associated with complex diseases [344-345]. Copy number variation (CNV) is a type of structural variation with short nucleotide repeats varying in copy number between individuals and studies have shown that CNVs influence human complex traits including susceptibility to diseases [346], for example CNVs effects for cancer [347] and immune disorders [348-349]. Sequencing enables the detection of low-frequency, rare genetic variants and copy number variants. With newer advanced next-generation sequencing (NGS) technologies this detection will become more efficient and may lead to the sequencing of the methylome.

In conclusion

It is obvious that a single approach will not elucidate human complex trait etiology. However, the combination of genetic approaches I have used in my thesis and which I discussed above, and other methods such as computational biology, which provide multi-faceted findings in different angles, will lead to a better understanding of the processes and genetic architecture of human complex traits.

Bibliography

1. Rice, T.K. and I.B. Borecki, *Familial resemblance and heritability.* Adv Genet, 2001. **42**: p. 35-44.
2. Daly, A.K. and C.P. Day, *Candidate gene case-control association studies: advantages and potential pitfalls.* Br J Clin Pharmacol, 2001. **52**(5): p. 489-99.
3. Boehnke, M., *Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes.* Am J Hum Genet, 1994. **55**(2): p. 379-90.
4. Moore, J.H., F.W. Asselbergs, and S.M. Williams, *Bioinformatics challenges for genome-wide association studies.* Bioinformatics, 2010. **26**(4): p. 445-455.
5. Browning, S.R., *Missing data imputation and haplotype phase inference for genome-wide association studies.* Hum Genet, 2008. **124**(5): p. 439-450.
6. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
7. Fu, W.Q., T.D. O'Connor, and J.M. Akey, *Genetic architecture of quantitative traits and complex diseases.* Curr Opin Genet Dev, 2013. **23**(6): p. 678-683.
8. Tsai, P.C., T.D. Spector, and J.T. Bell, *Using epigenome-wide association scans of DNA methylation in age-related complex human traits.* Epigenomics, 2012. **4**(5): p. 511-526.
9. Polderman, T.J.C., et al., *Meta-analysis of the heritability of human traits based on fifty years of twin studies.* Nat Genet, 2015. **47**(7): p. 702-702.
10. van Dongen, J., et al., *The continuing value of twin studies in the omics era.* Nat Rev Genet, 2012. **13**(9): p. 640-53.
11. Martin, N., D. Boomsma, and G. Machin, *A twin-pronged attack on complex traits.* Nat Genet, 1997. **17**(4): p. 387-92.
12. Medland, S.E., et al., *Special twin environments, genetic influences and their effects on the handedness of twins and their siblings.* Twin Res, 2003. **6**(2): p. 119-130.
13. Falconer, D.S. and T.F.C. Mackay, *Introduction to Quantitative Genetics*, ed. 4. 1996, UK: Longman: Essex.
14. Boomsma, D.I., *Twin, association and current "omics" studies.* J Matern-Fetal Neo M, 2013. **26**: p. 9-12.
15. Levy, F., et al., *Twin-sibling differences in parental reports of ADHD, speech, reading and behaviour problems.* J Child Psychol Psyc, 1996. **37**(5): p. 569-578.
16. Connolly, K., T. Bouchard, and P. Proppin, *Twins as a tool of behavioral genetics.* Brit J Dev Psychol, 1996. **14**: p. 111-112.
17. Knickmeyer, R.C., et al., *Twin-Singleton Differences in Neonatal Brain Structure.* Twin Res Hum Genet, 2011. **14**(3): p. 268-276.
18. Keller, M.C., et al., *Modeling Extended Twin Family Data I: Description of the Cascade Model.* Twin Res Hum Genet, 2009. **12**(1): p. 8-18.
19. Gibson, G., *Rare and common variants: twenty arguments.* Nat Rev Genet, 2012. **13**(2): p. 135-145.
20. Candille, S.I., et al., *Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations.* PLoS One, 2012. **7**(10): p. e48294.

21. Pe'er, I., et al., *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genet Epidemiol, 2008. **32**(4): p. 381-5.
22. Visscher, P.M., B. McEvoy, and J. Yang, *From Galton to GWAS: quantitative genetics of human height.* Genet Res (Camb), 2010. **92**(5-6): p. 371-9.
23. Sandholt, C.H., et al., *The effect of GWAS identified BMI loci on changes in body weight among middle-aged Danes during a five-year period.* Obesity (Silver Spring), 2014. **22**(3): p. 901-8.
24. Grigoroiu-Serbanescu, M., et al., *Association of age-of-onset groups with GWAS significant schizophrenia and bipolar disorder loci in Romanian bipolar I patients.* Psychiatry Res, 2015. **230**(3): p. 964-7.
25. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.
26. Zaitlen, N., et al., *Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits.* PloS Genet, 2013. **9**(5).
27. Dupont, C., D.R. Armant, and C.A. Brenner, *Epigenetics: Definition, Mechanisms and Clinical Perspective.* Semin Reprod Med, 2009. **27**(5): p. 351-357.
28. Fuks, F., *DNA methylation and histone modifications: teaming up to silence genes.* Curr Opin Genet Dev, 2005. **15**(5): p. 490-495.
29. Minarovits, J., et al., *Epigenetic Regulation.* Adv Exp Med Biol, 2016. **879**: p. 1-25.
30. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-541.
31. Ellis, S.E., et al., *RNA-Seq optimization with eQTL gold standards.* BMC Genom, 2013. **14**: p. 892.
32. Nica, A.C. and E.T. Dermitzakis, *Expression quantitative trait loci: present and future.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120362.
33. Binkley, C.J., et al., *Genetic variations associated with red hair color and fear of dental pain, anxiety regarding dental care and avoidance of dental care.* J Am Dent Assoc, 2009. **140**(7): p. 896-905.
34. Brauer, G. and V.P. Chopra, *Estimating the Heritability of Hair Color and Eye Color.* J Hum Evol, 1980. **9**(8): p. 625-630.
35. Matheny, A.P., Jr. and A.B. Dolan, *Sex and genetic differences in hair color changes during early childhood.* Am J Phys Anthropol, 1975. **42**(1): p. 53-6.
36. Zhu, G., et al., *A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q.* Twin Res, 2004. **7**(2): p. 197-210.
37. Posthuma, D., et al., *Replicated linkage for eye color on 15q using comparative ratings of sibling pairs.* Behav Genet, 2006. **36**(1): p. 12-7.
38. Relethford, J.H., *Apportionment of global human genetic diversity based on craniometrics and skin color.* Am J Phys Anthropol, 2002. **118**(4): p. 393-398.
39. Post, P.W. and D.C. Rao, *Genetic and environmental determinants of skin color.* Am J Phys Anthropol, 1977. **47**(3): p. 399-402.
40. Jablonski, N.G. and G. Chaplin, *The evolution of human skin coloration.* J Hum Evol, 2000. **39**(1): p. 57-106.
41. Khan, R. and B.S.R. Khan, *Diet, disease and pigment variation in humans.* Med Hypotheses, 2010. **75**(4): p. 363-367.
42. Parra, E.J., *Human Pigmentation Variation: Evolution, Genetic Basis, and Implications for Public Health.* Yearb Phys Anthropol, 2007. **50**: p. 85-105.

43. Aoki, K., *Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited.* Ann Hum Biol, 2002. **29**(6): p. 589-608.

44. Zemelman, V., et al., *Sexual dimorphism in skin, eye and hair color and the presence of freckles in Chilean teenagers from two socioeconomic strata.* Revista Medica De Chile, 2002. **130**(8): p. 879-884.

45. Schiaffino, M.V., *Signaling pathways in melanosome biogenesis and pathology.* Int J Biochem Cell B, 2010. **42**(7): p. 1094-1104.

46. Rogers, A.R., D. Iltis, and S. Wooding, *Genetic variation at the MCIR locus and the time since loss of human body hair.* Curr Anthropol, 2004. **45**(1): p. 105-108.

47. Harding, R.M., et al., *Evidence for variable selective pressures at MC1R.* Am J Hum Genet, 2000. **66**(4): p. 1351-1361.

48. Puddu, P.E., et al., *Red blood cell count in short-term prediction of cardiovascular disease incidence in the Gubbio population Study.* Acta Cardiol, 2002. **57**(3): p. 177-185.

49. Wasserman, M., et al., *The Utility of Temperature, White Blood-Cell Count, and the Peripheral-Blood Smear in Diagnosing Bacterial-Infection in the Elderly.* J Am Geriatr Soc, 1988. **36**(6): p. 589-589.

50. Samama, C.M. and L. Simon, *Detecting coagulation disorders of pregnancy: bleeding time or platelet count?* Can J Anaesth, 2001. **48**(6): p. 515-518.

51. Evans, D.M., I.H. Frazer, and N.G. Martin, *Genetic and environmental causes of variation in basal levels of blood cells.* Twin Res, 1999. **2**(4): p. 250-7.

52. Garner, C., et al., *Genetic influences on F cells and other hematologic variables: a twin heritability study.* Blood, 2000. **95**(1): p. 342-6.

53. Hall, M.A., et al., *Genetic influence on peripheral blood T lymphocyte levels.* Genes Immun, 2000. **1**(7): p. 423-7.

54. Lewis, S.L. and D.E. Van Epps, *Neutrophil and monocyte alterations in chronic dialysis patients.* Am J Kidney Dis, 1987. **9**(5): p. 381-95.

55. Verjee, Z.H. and R. Behal, *Protein-Calorie Malnutrition - Study of Red Blood-Cell and Serum Enzymes during and after Crisis.* Clin Chim Acta, 1976. **70**(1): p. 139-147.

56. Catalan, U., et al., *Biomarkers of food intake and metabolite differences between plasma and red blood cell matrices; a human metabolomic profile approach.* Mol Biosyst, 2013. **9**(6): p. 1411-1422.

57. Beall, C.M., *Origins: Human adaptation to high-altitude hypoxia.* Am J Phys Anthropol, 2007: p. 70-70.

58. Yi, X., et al., *Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude.* Science, 2010. **329**(5987): p. 75-78.

59. Roach, R.C., et al., *Transcriptomic and Epigenomic Reponses During Human Adaptation to High-Altitude Hypoxia.* Faseb Journal, 2015. **29**: p. 1051-1504.

60. Okada, Y., et al., *Identification of nine novel loci associated with white blood cell subtypes in a Japanese population.* PloS Genet, 2011. **7**(6): p. e1002067.

61. Vasquez, L.J., et al., *From GWAS to function: lessons from blood cells.* ISBT Sci Ser, 2016. **11**(Suppl Suppl 1): p. 211-219.

62. Yousefi, P., et al., *Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies.* Environ Mol Mutagen, 2015. **56**(9): p. 751-8.

63. Ferroni, P., et al., *Venous thromboembolism risk prediction in ambulatory cancer patients: clinical significance of neutrophil/lymphocyte ratio and platelet/lymphocyte ratio.* Int J Cancer, 2015. **136**(5): p. 1234-40.

64. Grenader, T., et al., *Prognostic value of neutrophil-to-lymphocyte ratio in advanced oesophago-gastric cancer: exploratory analysis of the REAL-2 trial.* Ann Oncol, 2016. **27**(4): p. 687-92.

65. Absenger, G., et al., *Preoperative neutrophil-to-lymphocyte ratio predicts clinical outcome in patients with stage II and III colon cancer.* Anticancer Res, 2013. **33**(10): p. 4591-4.

66. Sarikaya, M., et al., *Neutrophil-to-lymphocyte ratio as a sensitive marker in diagnosis of celiac disease.* Ann Gastroenterol, 2014. **27**(4): p. 431-432.

67. Acarturk, G., et al., *Neutrophil-to-lymphocyte ratio in inflammatory bowel disease - as a new predictor of disease severity.* Bratisl Lek Listy, 2015. **116**(4): p. 213-7.

68. Fowler, A.J. and R.A. Agha, *Neutrophil/lymphocyte ratio is related to the severity of coronary artery disease and clinical outcome in patients undergoing angiography--the growing versatility of NLR.* Atherosclerosis, 2013. **228**(1): p. 44-5.

69. Ayhan, H., et al., *Relationship of Neutrophil-to-Lymphocyte Ratio with Aortic Stiffness in Type 1 Diabetes Mellitus.* Can J Diabetes, 2015. **39**(4): p. 317-21.

70. Zhou, X., et al., *Prognostic value of PLR in various cancers: a meta-analysis.* PLoS One, 2014. **9**(6): p. e0101119.

71. Ying, H.Q., et al., *The prognostic value of preoperative NLR, d-NLR, PLR and LMR for predicting clinical outcome in surgical colorectal cancer patients.* Med Oncol, 2014. **31**(12): p. e3248.

72. Gu, X.B., et al., *Prognostic significance of neutrophil-to-lymphocyte ratio in non-small cell lung cancer: a meta-analysis.* Sci Rep-UK, 2015. **5**: p. e12493.

73. Feng, F., et al., *Combination of PLR, MLR, MWR, and Tumor Size Could Significantly Increase the Prognostic Value for Gastrointestinal Stromal Tumors.* Medicine, 2016. **95**(14): p. e3248.

74. van Beijsterveldt, C.E.M., et al., *The Young Netherlands Twin Register (YNTR): Longitudinal Twin and Family Studies in Over 70,000 Children.* Twin Res Hum Genet, 2013. **16**(1): p. 252-267.

75. Willemsen, G., et al., *The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection.* Twin Res Hum Genet, 2013. **16**(1): p. 271-81.

76. Willemsen, G., et al., *The Netherlands Twin Register biobank: a resource for genetic epidemiological studies.* Twin Res Hum Genet, 2010. **13**(3): p. 231-45.

77. Sirota, M., et al., *Effect of genome and environment on metabolic and inflammatory profiles.* Plos One, 2015. **10**(4): p. e0120898.

78. Ehli, *Customizing population-specific arrays for imputation-based genome-wide association testing: the Axiom-NL genotyping array.* European journal of human genetics, 2016. **Manuscript submitted for publication**.

79. van Dongen, J., et al., *Epigenome-Wide Association Study of Aggressive Behavior.* Twin Res Hum Genet, 2015. **18**(6): p. 686-698.

80. Liu, F., B. Wen, and M. Kayser, *Colorful DNA polymorphisms in humans.* Semin Cell Dev Biol. **24**(6-7): p. 562-75.

81. Juni, S.R., M. R, *The influence of hair color on soliciting help.* J Soc Behav Pers, 1985(13): p. 8.

82. Guéguen, N., *Hair Color and Courtship: Blond Women Received More Courtship Solicitations and Redhead Men Received More Refusals.* Psychol Stud, 2012. **57**(4).

83. Ames, B.N., K.H. O., and E. Yamasaki, *Hair dyes are mutagenic: identification of a variety of mutagenic ingredients.* Proc Natl Acad Sci U.S.A., 1975. **72**(6).

84. Boniol, M., P. Autier, and J.F. Dore, *Re: A prospective study of pigmentation, sun exposure, and risk of cutaneous malignant melanoma in women.* J Natl Cancer Inst, 2004. **96**(4): p. 335-6; author reply 336-8.

85. Lens, M.B. and M. Dawes, *Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma.* Br J Dermatol, 2004. **150**(2): p. 179-85.

86. Diffey, B.L., *Solar ultraviolet radiation effects on biological systems.* Phys Med Biol, 1991. **36**(3): p. 299-328.

87. Sulem, P., et al., *Genetic determinants of hair, eye and skin pigmentation in Europeans.* Nat Genet, 2007. **39**(12): p. 1443-52.

88. Branicki, W., et al., *Model-based prediction of human hair color using DNA variants.* Hum Genet, 2011. **129**(4): p. 443-54.

89. Mitra, D., et al., *An ultraviolet-radiation-independent pathway to melanoma carcinogenesis in the red hair/fair skin background.* Nature, 2012. **491**(7424): p. 449-53.

90. Rompler, H., et al., *Nuclear gene indicates coat-color polymorphism in mammoths.* Science, 2006. **313**(5783): p. 62.

91. Raimondi, S., et al., *MC1R variants, melanoma and red hair color phenotype: a meta-analysis.* Int J Cancer, 2008. **122**(12): p. 2753-60.

92. Visser, M., M. Kayser, and R.JPalstra, *HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter.* Genome Res, 2012. **22**(3): p. 446-55.

93. Boomsma, D.I., et al., *Netherlands Twin Register: from twins to twin families.* Twin Res Hum Genet, 2006. **9**(6): p. 849-57.

94. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

95. Li, Y. and G.R. Abecasis, *Mach 1.0:Rapid haplotype reconstruction and missing genotype inference.* Am J Hum Genet, 2006. **79**(S2290).

96. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nat Genet, 2012. **44**(8): p. 955-9.

97. Liu, E.Y., et al., *MaCH-admix: genotype imputation for admixed populations.* Genet Epidemiol, 2013. **37**(1): p. 25-37.

98. Boker, S., et al., *OpenMx: An Open Source Extended Structural Equation Modeling Framework.* Psychometrika, 2011. **76**(2): p. 306-317.

99. Sham, P.C., et al., *Logistic regression analysis of twin data: estimation of parameters of the multifactorial liability-threshold model.* Behav Genet, 1994. **24**(3): p. 229-38.

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

101. Rogers, W.H., *Regression standard errors in clustered samples.* Stata Technical Bulletin, 1993. **13**: p. 19–23.

102. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.

103. Abdellaoui, A., et al., *Population structure, migration, and diversifying selection in the Netherlands.* Eur J Hum Genet, 2013. **21**(11): p. 1277-85.

104. Minica, C.C., et al., *Sandwich corrected standard errors in family-based genome-wide association studies.* Eur J Hum Genet, 2015. **23**(3): p. 388-94.

105. Redden, D.T. and D.B. Allison, *The effect of assortative mating upon genetic association studies: Spurious associations and population substructure in the absence of admixture.* Behav Genet, 2006. **36**(5): p. 678-686.

106. Shekar, S.N., et al., *Linkage and association analysis of spectrophotometrically quantified hair color in Australian adolescents: the effect of OCA2 and HERC2.* J Invest Dermatol, 2008. **128**(12): p. 2807-14.

107. Sturm, R.A., *Human 'coat colour' genetics.* Pigm Cell Melanoma Res, 2008. **21**(2): p. 115-116.

108. Li, X.F., A.S. Kraev, and J. Lytton, *Molecular cloning of a fourth member of the potassium-dependent sodium-calcium exchanger gene family, NCKX4.* J Biol Chem, 2002. **277**(50): p. 48410-7.

109. Graf, J., et al., *Promoter polymorphisms in the MATP (SLC45A2) gene are associated with normal human skin color variation.* Hum Mutat, 2007. **28**(7): p. 710-7.

110. Sabeti, P.C., et al., *Genome-wide detection and characterization of positive selection in human populations.* Nature, 2007. **449**(7164): p. 913-8.

111. Mondal, M., et al., *Molecular basis of albinism in India: evaluation of seven potential candidate genes and some new findings.* Gene, 2012. **511**(2): p. 470-4.

112. Makova, K.D., et al., *Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter.* Genetics, 2001. **158**(3): p. 1253-68.

113. Woodworth, S.H., et al., *A prospective study on the association between red hair color and endometriosis in infertile patients.* Fertil Steril, 1995. **64**(3): p. 651-2.

114. Sturm, R.A., et al., *Genetic association and cellular function of MC1R variant alleles in human pigmentation.* Ann N Y Acad Sci, 2003. **994**: p. 348-58.

115. Nakayama, K., et al., *Identification of novel functional variants of the melanocortin 1 receptor gene originated from Asians.* Hum Genet, 2006. **119**(3): p. 322-30.

116. Eriksson, N., et al., *Web-based, participant-driven studies yield novel genetic associations for common traits.* PLoS Genet. **6**(6): p. e1000993.

117. Han, J., et al., *A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation.* PLoS Genet, 2008. **4**(5): p. e1000074.

118. Guenther, C.A., et al., *A molecular basis for classic blond hair color in Europeans.* Nat Genet, 2014.

119. Sulem, P., et al., *Two newly identified genetic determinants of pigmentation in Europeans.* Nat Genet, 2008. **40**(7): p. 835-7.

120. Hutton, S.M. and R.A. Spritz, *Comprehensive analysis of oculocutaneous albinism among non-Hispanic Caucasians shows that OCA1 is the most prevalent OCA type.* J Invest Dermatol, 2008. **128**(10): p. 2442-2450.

121. Sandberg, M.A., et al., *Disease expression in patients with USH2A mutations.* Invest Ophthalmol Vis Sci, 2004. **45**: p. U575-U575.

122. Lin, B.D., et al., *Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample.* Genes, 2015. **6**(3): p. 559-76.

123. Kayser, M., et al., *Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene.* Am J Hum Genet, 2008. **82**(2): p. 411-23.

124. Liu, F., et al., *Eye color and the prediction of complex phenotypes from genotypes.* Curr Biol, 2009. **19**(5): p. R192-R193.

125. Zhang, M.F., et al., *Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans.* Hum Mol Genet, 2013. **22**(14): p. 2948-2959.

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

127. Beleza, S., et al., *The timing of pigmentation lightening in Europeans.* Mol Biol Evol, 2013. **30**(1): p. 24-35.

128. Bolk, L., *Heeft roodharigheid de beteekenis van nuance of varieteit? .* Ned Tijdschr Geneeskd, 1908. **52**.

129. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

130. Liu, J., et al., *Tyrosinase gene (TYR) mutations in Chinese patients with oculocutaneous albinism type 1.* Clin Exp Ophthalmol, 2010. **38**(1): p. 37-42.

131. 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, 2008. **16**(3): p. 335-42.

132. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.

133. Sturm, R.A., et al., *A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color.* Am J Hum Genet, 2008. **82**(2): p. 424-431.

134. Morice-Picard, F., et al., *High-resolution array-CGH in patients with oculocutaneous albinism identifies new deletions of the TYR, OCA2, and SLC45A2 genes and a complex rearrangement of the OCA2 gene.* Pigm Cell Melanoma R, 2014. **27**(1).

135. Hu, K., et al., *Prognostic role of the neutrophil-lymphocyte ratio in renal cell carcinoma: a meta-analysis.* BMJ Open, 2015. **5**(4): p. e006404.

136. Hu, Z.D., et al., *Prognostic value of neutrophil to lymphocyte ratio for gastric cancer.* Ann Transl Med, 2015. **3**(4): p. 50-58.

137. Shah, N., et al., *Neutrophil lymphocyte ratio significantly improves the Framingham risk score in prediction of coronary heart disease mortality: Insights from the National Health and Nutrition Examination Survey-III.* Int J Cardiol, 2014. **171**(3): p. 390-397.

138. Ozturk, C., et al., *Neutrophil-lymphocyte ratio and carotid-intima media thickness in patients with Behcet disease without cardiovascular involvement.* Angiology, 2015. **66**(3): p. 291-6.

139. Li, J., et al., *Neutrophil-to-Lymphocyte Ratio Positively Correlates to Age in Healthy Population.* J Clin Lab Anal, 2015. **29**(6): p. 437-443.

140. Azab, B., M. Camacho-Rivera, and E. Taioli, *Average values and racial differences of neutrophil lymphocyte ratio among a nationally representative sample of United States subjects.* Plos One, 2014. **9**(11): p. e112361.

141. Whitfield, J.B. and N.G. Martin, *Genetic and environmental influences on the size and number of cells in the blood.* Genet Epidemiol, 1985. **2**(2): p. 133-44.

142. Maes, M., et al., *Seasonal variation in peripheral blood leukocyte subsets and in serum interleukin-6, and soluble interleukin-2 and -6 receptor concentrations in normal volunteers.* Experientia, 1994. **50**(9): p. 821-9.

143. Paglieroni, T.G. and P.V. Holland, *Circannual variation in lymphocyte subsets, revisited.* Transfusion, 1994. **34**(6): p. 512-6.

144. Buckley, M.F., et al., *A novel approach to the assessment of variations in the human platelet count.* Thromb Haemost, 2000. **83**(3): p. 480-4.

145. Broadbent, S., *Seasonal changes in haematology, lymphocyte transferrin receptors and intracellular iron in Ironman triathletes and untrained men.* Eur J Appl Physiol, 2011. **111**(1): p. 93-100.

146. Crawford, V.L., S.E. McNerlan, and R.W. Stout, *Seasonal changes in platelets, fibrinogen and factor VII in elderly people.* Age Ageing, 2003. **32**(6): p. 661-5.

147. Furuncuoglu, Y., et al., *How obesity affects the neutrophil/lymphocyte and platelet/lymphocyte ratio, systemic immune-inflammatory index and platelet indices: a retrospective study.* Eur Rev Med Pharmacol Sci, 2016. **20**(7): p. 1300-1306.

148. Vuong, J., et al., *Reference intervals of complete blood count constituents are highly correlated to waist circumference: should obese patients have their own "normal values?".* Am J Hematol, 2014. **89**(7): p. 671-7.

149. Farhangi, M.A., et al., *White blood cell count in women: relation to inflammatory biomarkers, haematological profiles, visceral adiposity, and other cardiovascular risk factors.* J Health Popul Nutr, 2013. **31**(1): p. 58-64.

150. Samocha-Bonet, D., et al., *Platelet counts and platelet activation markers in obese subjects.* Mediators Inflamm, 2008. **2008**: p. 834153.

151. Tulgar, Y.K., et al., *The effect of smoking on neutrophil/lymphocyte and platelet/lymphocyte ratio and platelet indices: a retrospective study.* Eur Rev Med Pharmacol Sci, 2016. **20**(14): p. 3112-8.

152. Andreoli, C., et al., *Effects of cigarette smoking on circulating leukocytes and plasma cytokines in monozygotic twins.* Clin Chem Lab Med, 2015. **53**(1): p. 57-64.

153. Suwansaksri, J., V. Wiwanitkit, and S. Soogarun, *Effect of smoking on platelet count and platelet parameters: an observation.* Clin Appl Thromb Hemost, 2004. **10**(3): p. 287-8.

154. Mercelina-Roumans, P.E., J.M. Ubachs, and J.W. van Wersch, *Platelet count and platelet indices at various stages of normal pregnancy in smoking and non-smoking women.* Eur J Clin Chem Clin Biochem, 1995. **33**(5): p. 267-9.

155. Varol, E., et al., *Effect of smoking cessation on mean platelet volume.* Clin Appl Thromb Hemost, 2013. **19**(3): p. 315-9.

156. Green, M.S., I. Peled, and T. Najenson, *Gender differences in platelet count and its association with cigarette smoking in a large cohort in Israel.* J Clin Epidemiol, 1992. **45**(1): p. 77-84.

157. Neijts, M., et al., *Genetic architecture of the pro-inflammatory state in an extended twin-family design.* Twin Res Hum Genet, 2013. **16**(5): p. 931-40.

158. van Dongen, J., et al., *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): p. 368-82.

159. KNMI, *Daily data on the weather in Netherlands.* 2004-2008.

160. *Stata Corp.Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.* 2013.

161. Havlicek, J. and S.C. Roberts, *MHC-correlated mate choice in humans: a review.* Psychoneuroendocrinology, 2009. **34**(4): p. 497-512.

162. Fairweather, D., S. Frisancho-Kiss, and N.R. Rose, *Sex differences in autoimmune disease from a pathological perspective.* Am J Pathol, 2008. **173**(3): p. 600-9.

163. Dopico, X.C., et al., *Widespread seasonal gene expression reveals annual differences in human immunity and physiology.* Nat Commun, 2015. **6**: p. 7000.

164. Yudkin, J.S., et al., *Inflammation, obesity, stress and coronary heart disease: is interleukin-6 the link?* Atherosclerosis, 2000. **148**(2): p. 209-14.

165. Bonaccio, M., et al., *Adherence to the Mediterranean diet is associated with lower platelet and leukocyte counts: results from the Moli-sani study.* Blood, 2014. **123**(19): p. 3037-3044.
166. Fogar, P., et al., *Decreased total lymphocyte counts in pancreatic cancer: an index of adverse outcome.* Pancreas, 2006. **32**(1): p. 22-8.
167. Liang, L., et al., *Predictive value of pretreatment lymphocyte count in stage II colorectal cancer and in high-risk patients treated with adjuvant chemotherapy.* Oncotarget, 2016. **7**(1): p. 1014-1028.
168. Di Caro, G., et al., *Occurrence of Tertiary Lymphoid Tissue Is Associated with T-Cell Infiltration and Predicts Better Prognosis in Early-Stage Colorectal Cancers.* Clin Cancer Res, 2014. **20**(8): p. 2147-2158.
169. McCourt, M., et al., *Proinflammatory mediators stimulate neutrophil-directed angiogenesis.* Arch Surg, 1999. **134**(12): p. 1325-31; discussion 1331-2.
170. Di Carlo, E., G. Forni, and P. Musiani, *Neutrophils in the antitumoral immune response.* Chem Immunol Allergy, 2003. **83**: p. 182-203.
171. Thaulow, E., et al., *Blood-Platelet Count and Function Are Related to Total and Cardiovascular Death in Apparently Healthy-Men.* Circulation, 1991. **84**(2): p. 613-617.
172. Franco, A.T., A. Corken, and J. Ware, *Platelets at the interface of thrombosis, inflammation, and cancer.* Blood, 2015. **126**(5): p. 582-588.
173. Imtiaz, F., et al., *Neutrophil lymphocyte ratio as a measure of systemic inflammation in prevalent chronic diseases in Asian population.* Int Arch Med, 2012. **5**(1): p. 2.
174. Oh, B.S., et al., *Prognostic value of C-reactive protein and neutrophil-to-lymphocyte ratio in patients with hepatocellular carcinoma.* Bmc Cancer, 2013. **13**.
175. Xia, W.K., et al., *Prognostic performance of pre-treatment NLR and PLR in patients suffering from osteosarcoma.* World J Surg Oncol, 2016. **14**.
176. Koh, C.H., et al., *Utility of pre-treatment neutrophil-lymphocyte ratio and platelet-lymphocyte ratio as prognostic factors in breast cancer.* Brit J Cancer, 2015. **113**(1): p. 150-158.
177. Jansen, R., et al., *Conditional eQTL Analysis Reveals Allelic Heterogeneity of Gene Expression. Human Molecular Genetics.* in press.
178. Forget, P., V. Dinant, and M. De Kock, *Is the Neutrophil-to-Lymphocyte Ratio more correlated than C-reactive protein with postoperative complications after major abdominal surgery?* PeerJ, 2015. **3**.
179. Durmus, E., et al., *Neutrophil-to-Lymphocyte Ratio and Platelet-to-Lymphocyte Ratio are predictors of heart failure.* Arq Bras Cardiol, 2015.
180. Lee, S., et al., *Prognostic significance of neutrophil lymphocyte ratio and platelet lymphocyte ratio in advanced gastric cancer patients treated with FOLFOX chemotherapy.* BMC Cancer, 2013. **13**: p. 350.
181. Wang, X.D., et al., *Neutrophil to lymphocyte ratio in relation to risk of all-cause mortality and cardiovascular events among patients undergoing angiography or cardiac revascularization: A meta-analysis of observational studies.* Atherosclerosis, 2014. **234**(1): p. 206-213.
182. He, J.Y., et al., *Neutrophil-to-lymphocyte ratio (NLR) predicts mortality and adverse-outcomes after ST-segment elevation myocardial infarction in Chinese people.* Int J Clin Exp Pathol, 2014. **7**(7): p. 4045-4056.
183. Kang, M.H., et al., *The prognostic impact of the neutrophil-to-lymphocyte ratio in patients with small-cell lung cancer.* J Clin Oncol, 2014. **32**(15).

184. Templeton, A.J., et al., *Prognostic role of neutrophil to lymphocyte ratio (NLR) in solid tumors: A systematic review and meta-analysis.* Eur J Cancer, 2013. **49**: p. S211-S211.

185. Yodying, H., et al., *Prognostic significance of Neutrophil-to-Lymphocyte Ratio and Platelet-to-Lymphocyte Ratio in oncologic outcomes of esophageal cancer: A systematic review and meta-analysis.* Ann Surg Oncol, 2016. **23**(2): p. 646-654.

186. Turkmen, K., et al., *Platelet-to-lymphocyte ratio better predicts inflammation than neutrophil-to-lymphocyte ratio in end-stage renal disease patients.* Hemodialysis International, 2013. **17**(3): p. 391-396.

187. Lin, B.D., et al., *Causes of variation in the neutrophil-lymphocyte and platelet-lymphocyte ratios: a twin-family study.* Biomark Med, 2016: p. 1061-1072.

188. Soranzo, N., et al., *A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium.* Nat Genet, 2009. **41**(11): p. 1182-90.

189. Gieger, C., et al., *New gene functions in megakaryopoiesis and platelet formation.* Nature, 2011. **480**(7376): p. 201-8.

190. Shameer, K., et al., *A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects.* Hum Genet, 2014. **133**(1): p. 95-109.

191. Schick, U.M., et al., *Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans.* Am J Hum Genet, 2016. **98**(2): p. 229-42.

192. Reiner, A.P., et al., *Genome-Wide Association Study of White Blood Cell Count in 16,388 African Americans: the Continental Origins and Genetic Epidemiology Network (COGENT).* PloS Genet, 2011. **7**(6): p. e1002108.

193. Qayyum, R., et al., *A meta-analysis and genome-wide association study of platelet count and mean platelet volume in African Americans.* PLoS Genet, 2012. **8**(3): p. e1002491.

194. Qayyum, R., et al., *Genome-wide association study of platelet aggregation in African Americans.* BMC Genet, 2015. **16**: p. 58.

195. Nalls, M.A., et al., *Multiple loci are associated with white blood cell phenotypes.* PloS Genet, 2011. **7**(6): p. e1002113.

196. Kim, Y.K., et al., *Influence of genetic variants in EGF and other genes on hematological traits in Korean populations by a genome-wide approach.* Biomed Res Int, 2015. **2015**: p. 914965.

197. Oh, J.H., et al., *Genome-wide association study identifies candidate Loci associated with platelet count in Koreans.* Genomics Inform, 2014. **12**(4): p. 225-30.

198. Okada, Y., et al., *Common variations in PSMD3-CSF3 and PLCB4 are associated with neutrophil count.* Hum Mol Genet, 2010. **19**(10): p. 2079-2085.

199. Nalls, M.A., et al., *Admixture mapping of white cell count: Genetic locus responsible for lower white blood cell count in the health ABC and Jackson Heart Studies.* Am J Hum Genet, 2008. **82**(1): p. 81-87.

200. Keller, M.F., et al., *Trans-ethnic meta-analysis of white blood cell phenotypes.* Hum Mol Genet, 2014. **23**(25): p. 6944-60.

201. Li, J., et al., *GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children.* Hum Mol Genet, 2013. **22**(7): p. 1457-64.

202. Soranzo, N., et al., *A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function.* Blood, 2009. **113**(16): p. 3831-3837.

203. Ganesh, S.K., et al., *Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium.* Nat Genet, 2009. **41**(11): p. 1191-U48.

204. Kamatani, Y., et al., *Genome-wide association study of hematological and biochemical traits in a Japanese population.* Nat Genet, 2010. **42**(3): p. 210-215.
205. Gudbjartsson, D.F., et al., *Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction.* Nat Genet, 2009. **41**(3): p. 342-347.
206. Lo, K.S., et al., *Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans.* Human Genetics, 2011. **129**(3): p. 307-317.
207. Menzel, S., et al., *Experimental generation of SNP haplotype signatures in patients with sickle cell anaemia.* PLoS One, 2010. **5**(9).
208. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015. **47**(11): p. 1236-1241.
209. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nat Genet, 2015. **47**(3): p. 291-295.
210. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6).
211. Yang, J.A., et al., *GCTA: A tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
212. Yang, J., et al., *Advantages and pitfalls in the application of mixed-model association methods.* Nature Genetics, 2014. **46**(2): p. 100-6.
213. Abdellaoui, A., et al., *Population structure, migration, and diversifying selection in the Netherlands.* Eur J Hum Genet, 2013. **21**(11): p. 1277-1285.
214. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
215. Tucker, G., et al., *Two-Variance-Component Model Improves Genetic Prediction in Family Datasets.* Am J Hum Genet, 2015. **97**(5): p. 677-690.
216. Andrew, T., et al., *Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women.* Twin Res, 2001. **4**(6): p. 464-77.
217. Teo, Y.Y., et al., *A genotype calling algorithm for the Illumina BeadArray platform.* Bioinformatics, 2007. **23**(20): p. 2741-6.
218. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations.* Bioinformatics, 2012. **28**(10): p. 1353-8.
219. *R Core Team. R: a language and environment for statistical computing.* 2014: Vienna, Austria.
220. van der Harst, P., et al., *Seventy-five genetic loci influencing the human red blood cell.* Nature, 2012. **492**(7429): p. 369-+.
221. Lettre, G., et al., *DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease.* P Natl Acad Sci USA, 2008. **105**(33): p. 11869-11874.
222. Pistis, G., et al., *Genome Wide Association Analysis of a Founder Population Identified TAF3 as a Gene for MCHC in Humans.* PLoS One, 2013. **8**(7).
223. Tapper, W., et al., *Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms.* Nat Commun, 2015. **6**.
224. He, Y.Y., et al., *Analysis of the rs35959442 polymorphism in Hb E/beta-thalassemia in Guangxi province of the Republic of China.* Hemoglobin, 2012. **36**(2): p. 166-169.
225. Stadhouders, R., et al., *HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers.* J Clin Invest, 2014. **124**(4): p. 1699-710.

226. Thein, S.L., et al., *Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults.* P Natl Acad Sci USA, 2007. **104**(27): p. 11346-51.
227. Roy, P., et al., *Influence of BCL11A, HBS1L-MYB, HBBP1 single nucleotide polymorphisms and the HBG2 XMNI polymorphism on Hb F levels.* Hemoglobin, 2012. **36**(6): p. 592-599.
228. Le Guen, T., et al., *An in vivo genetic reversion highlights the crucial role of MYB-Like, SWIRM, and MPN domains 1 (MYSM1) in human hematopoiesis and lymphocyte differentiation.* J Allergy Clin Immunol, 2015. **136**(6): p. 1619-26.
229. Mets, E., et al., *MicroRNA-193b-3p acts as a tumor suppressor by targeting the MYB oncogene in T-cell acute lymphoblastic leukemia.* Leukemia, 2015. **29**(4): p. 798-806.
230. Srivastava, S.K., et al., *MYB is a novel regulator of pancreatic tumour growth and metastasis.* Brit J Cancer, 2015. **113**(12): p. 1694-1703.
231. Sripichai, O., et al., *Genetic analysis of candidate modifier polymorphisms in Hb E-beta 0-thalassemia patients.* Ann N Y Acad Sci, 2005. **1054**: p. 433-8.
232. Danjou, F., et al., *A genetic score for the prediction of beta-thalassemia severity.* Haematologica, 2015. **100**(4): p. 452-7.
233. Pandit, R.A., et al., *Association of SNP in exon 1 of HBS1L with hemoglobin F level in beta0-thalassemia/hemoglobin E.* Int J Hematol, 2008. **88**(4): p. 357-61.
234. Farrell, J.J., et al., *A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression.* Blood, 2011. **117**(18): p. 4935-4945.
235. Willer, C.J., et al., *Discovery and refinement of loci associated with lipid levels.* Nat Genet, 2013. **45**(11): p. 1274-83.
236. Wright, H.L., et al., *Neutrophil function in inflammation and inflammatory diseases.* Rheumatology (Oxford), 2010. **49**(9): p. 1618-31.
237. Pillay, J., et al., *In vivo labeling with 2H2O reveals a human neutrophil lifespan of 5.4 days.* Blood, 2010. **116**(4): p. 625-7.
238. Tak, T., et al., *What's your age again? Determination of human neutrophil half-lives revisited.* J Leukoc Biol, 2013. **94**(4): p. 595-601.
239. Dorababu, P., et al., *Epistatic interactions between thiopurine methyltransferase (TPMT) and inosine triphosphate pyrophosphatase (ITPA) variations determine 6-mercaptopurine toxicity in Indian children with acute lymphoblastic leukemia.* Eur J Clin Pharmacol, 2012. **68**(4): p. 379-87.
240. Penninx, B.W., et al., *The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods.* Int J Methods Psychiatr Res, 2008. **17**(3): p. 121-40.
241. Rivas, M.A., et al., *Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.* Nature Genetics, 2011. **43**(11): p. 1066-U50.
242. Jansen, R., et al., *Sex differences in the human peripheral blood transcriptome.* BMC Genomics, 2014. **15**: p. 33.
243. Wright, F.A., et al., *Heritability and genomics of gene expression in peripheral blood.* Nat Genet, 2014. **46**(5): p. 430-7.
244. Fehrmann, R.S.N., 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, 2011. **7**(8).
245. Nivard, M.G., et al., *Further confirmation of the association between anxiety and CTNND2: replication in humans.* Genes Brain Behav, 2014. **13**(2): p. 195-201.
246. Hanifin, J.M. and M.J. Cline, *Human monocytes and macrophages. Interaction with antigen and lymphocytes.* Journal of Cell Biology, 1970. **46**(1): p. 97-105.

247. Iqbal, S., A. Umbreen, and S.B.H. Zaidi, *Monocyte Lymphocyte Ratio as a Possible Prognostic Marker in Antituberculous Therapy* JRMC, 2014. **18**(2): p. 178-181.

248. Nishijima, T.F., et al., *Prognostic value of lymphocyte-to-monocyte ratio in patients with solid tumors: A systematic review and meta-analysis.* Cancer Treat Rev, 2015. **41**(10): p. 971-978.

249. Crosslin, D.R., et al., *Genetic variation associated with circulating monocyte count in the eMERGE Network.* Hum Mol Genet, 2013. **22**(10): p. 2119-2127.

250. Tenorio, T.R., et al., *Relation between leukocyte count, adiposity, and cardiorespiratory fitness in pubertal adolescents.* Einstein, 2014. **12**(4): p. 420-4.

251. Zaldivar, F., et al., *Body fat and circulating leukocytes in children.* Int J Obes (Lond), 2006. **30**(6): p. 906-11.

252. Schwartz, J. and S.T. Weiss, *Cigarette smoking and peripheral blood leukocyte differentials.* Ann Epidemiol, 1994. **4**(3): p. 236-42.

253. Perez-de-Heredia, F., et al., *Influence of sex, age, pubertal maturation and body mass index on circulating white blood cell counts in healthy European adolescents-the HELENA study.* Eur J Pediatr, 2015. **174**(8): p. 999-1014.

254. Al-Sufyani, A.A. and S.H. Mahassni, *Obesity and immune cells in Saudi females.* Innate Immun, 2011. **17**(5): p. 439-50.

255. Finucane, H.K., et al., *Partitioning heritability by functional annotation using genome-wide association summary statistics.* Nat Genet, 2015. **47**(11): p. 1228-1235.

256. Maugeri, N., et al., *LPAR1 and ITGA4 regulate peripheral blood monocyte counts.* Hum Mutat, 2011. **32**(8): p. 873-6.

257. Kamatani, Y., et al., *Genome-wide association study of hematological and biochemical traits in a Japanese population.* Nat Genet, 2010. **42**(3): p. 210-5.

258. Ferreira, M.A., et al., *Sequence variants in three loci influence monocyte counts and erythrocyte volume.* Am J Hum Genet, 2009. **85**(5): p. 745-9.

259. De Jager, P.L., et al., *Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci.* Nat Genet, 2009. **41**(7): p. 776-82.

260. Yanez, A., et al., *IRF8 acts in lineage-committed rather than oligopotent progenitors to control neutrophil vs monocyte production.* Blood, 2015. **125**(9): p. 1452-9.

261. Terry, R.L., et al., *Defective inflammatory monocyte development in IRF8-deficient mice abrogates migration to the West Nile virus-infected brain.* J Innate Immun, 2015. **7**(1): p. 102-12.

262. Kurotaki, D., et al., *Essential role of the IRF8-KLF4 transcription factor cascade in murine monocyte differentiation.* Blood, 2013. **121**(10): p. 1839-49.

263. Oishi, T., et al., *A functional SNP in the NKX2.5-binding site of ITPR3 promoter is associated with susceptibility to systemic lupus erythematosus in Japanese population.* J Hum Genet, 2008. **53**(2): p. 151-162.

264. Farragher, T.M., et al., *Association of the HLA-DRB1 gene with premature death, particularly from cardiovascular disease, in patients with rheumatoid arthritis and inflammatory polyarthritis.* Arthritis Rheum, 2008. **58**(2): p. 359-369.

265. Alvarez-Navarro, C., et al., *Endoplasmic Reticulum Aminopeptidase 1 (ERAP1) Polymorphism Relevant to Inflammatory Disease Shapes the Peptidome of the Birdshot Chorioretinopathy-Associated HLA-A*29:02 Antigen.* Mol Cell Proteomics, 2015. **14**(7): p. 1770-1780.

266. Thomas, D., et al., *Association of rs1568885, rs1813443 and rs4411591 polymorphisms with anti-TNF medication response in Greek patients with Crohn's disease.* World J Gastroentero, 2014. **20**(13): p. 3609-3614.

267. Okada, Y. and Y. Kamatani, *Common genetic factors for hematological traits in humans.* J Hum Genet, 2012. **57**(3): p. 161-9.

268. Parichehreh, V., et al., *Exploiting osmosis for blood cell sorting.* Biomed Microdevices, 2011. **13**(3): p. 453-62.

269. Karpman, D., et al., *Complement Interactions with Blood Cells, Endothelial Cells and Microvesicles in Thrombotic and Inflammatory Conditions.* Adv Exp Med Biol, 2015. **865**: p. 19-42.

270. Ambayya, A., et al., *Haematological reference intervals in a multiethnic population.* PLoS One, 2014. **9**(3): p. e91968.

271. Barazzoni, R., et al., *The Association between Hematological Parameters and Insulin Resistance Is Modified by Body Mass Index - Results from the North-East Italy MoMa Population Study.* PLoS One, 2014. **9**(7).

272. Biino, G., et al., *Age- And Sex-Related Variations in Platelet Count in Italy: A Proposal of Reference Ranges Based on 40987 Subjects' Data.* PLoS One, 2013. **8**(1).

273. Sunyer, J., et al., *Longitudinal Relation between Changes in Smoking and Changes in White Blood-Cells.* Am J Epidemiol, 1995. **141**(11): p. S52-S52.

274. Anderson, M.J., *A new method for non-parametric multivariate analysis of variance.* Austral Ecology, 2001. **26**(1): p. 32-46.

275. McArdle, B.H. and M.J. Anderson, *Fitting multivariate models to community data: A comment on distance-based redundancy analysis.* Ecology, 2001. **82**(1): p. 290-297.

276. McArtor, D.B., Lubke, G. H., Bergeman, C. S., *Extending multivariate distance matrix regression with an effect size measure and the asymptotic null distribution of the test statistic.* Manuscript submitted for publication.

277. Lee, J.J., S. Vattikuti, and C.C. Chow, *Uncovering the Genetic Architectures of Quantitative Traits.* Comput Struct Biotechnol J, 2016. **14**: p. 28-34.

278. Lubke, G. and D. McArtor, *Multivariate genetic analyses in heterogeneous populations.* Behav Genet, 2014. **44**(3): p. 232-9.

279. Kondo, M., *Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors.* Immunol Rev, 2010. **238**: p. 37-46.

280. Bhat, T., et al., *Neutrophil to lymphocyte ratio and cardiovascular diseases: a review.* Expert Rev Cardiovasc Ther, 2013. **11**(1): p. 55-9.

281. Shiny, A., et al., *Association of Neutrophil-Lymphocyte Ratio with Glucose Intolerance: An Indicator of Systemic Inflammation in Patients with Type 2 Diabetes.* Diabetes Technol Ther, 2014. **16**(8): p. 524-530.

282. Porrata, L.F., et al., *Peripheral blood lymphocyte/monocyte ratio at diagnosis and survival in classical Hodgkin's lymphoma.* Haematol Hematol J, 2012. **97**(2): p. 262-269.

283. Beltran, B.E., et al., *The neutrophil-to-lymphocyte ratio is an independent prognostic factor in patients with peripheral T-cell lymphoma, unspecified.* Leukemia Lymphoma, 2016. **57**(1): p. 58-62.

284. Sarraf, K.M., et al., *Neutrophil/lymphocyte ratio and its association with survival after complete resection in non-small cell lung cancer.* J Thorac Cardiov Sur, 2009. **137**(2): p. 425-428.

285. Stotz, M., et al., *The preoperative lymphocyte to monocyte ratio predicts clinical outcome in patients with stage III colon cancer.* Brit J Cancer, 2014. **110**(2): p. 435-440.

286. Shimada, H., et al., *High preoperative neutrophil-lymphocyte ratio predicts poor survival in patients with gastric cancer.* Gastric Cancer, 2010. **13**(3): p. 170-176.

287. Azab, B., et al., *Usefulness of the Neutrophil-to-Lymphocyte Ratio in Predicting Short- and Long-Term Mortality in Breast Cancer Patients.* Ann Surg Oncol, 2012. **19**(1): p. 217-224.

288. Wang, J., et al., *Ratio of monocytes to lymphocytes in peripheral blood in patients diagnosed with active tuberculosis.* Braz J Infect Dis, 2015. **19**(2): p. 125-131.

289. Lee, S.M., A. Russell, and G. Hellawell, *Predictive value of pretreatment inflammation-based prognostic scores (neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, and lymphocyte-to-monocyte ratio) for invasive bladder carcinoma.* Korean J Urol, 2015. **56**(11): p. 749-755.

290. Papa, A., et al., *Predictive value of elevated neutrophil-lymphocyte ratio on cardiac mortality in patients with stable coronary artery disease.* Clin Chim Acta, 2008. **395**(1-2): p. 27-31.

291. Ghaffari, S., et al., *The predictive Value of Total Neutrophil Count and Neutrophil/Lymphocyte Ratio in Predicting In-hospital Mortality and Complications after STEMI.* J Cardiovasc Thorac Res, 2014. **6**(1): p. 35-41.

292. Broske, A.M., et al., *DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction.* Nat Genet, 2009. **41**(11): p. 1207-15.

293. Unterberg, M., et al., *NFKB1 Promoter DNA from nt+402 to nt+99 Is Hypomethylated in Different Human Immune Cells.* Plos One, 2016. **11**(6): p. e0156702.

294. Spector, T.D. and F.M.K. Williams, *The UK Adult Twin Registry (TwinsUK).* Twin Res Hum Genet, 2006. **9**(6): p. 899-906.

295. Moayyeri, A., et al., *The UK Adult Twin Registry (TwinsUK Resource).* Twin Res Hum Genet, 2013. **16**(1): p. 144-9.

296. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* Bmc Bioinformatics, 2012. **13**.

297. Reinius, L.E., et al., *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.* PloS One, 2012. **7**(7): p. e41361.

298. van Dongen, J., et al., *Genetic and environmental influences interact with age and sex in shaping the human methylome.* Nat Commun, 2016. **7**: p. 11115.

299. Tsaprouni, L.G., et al., *Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation.* Epigenetics, 2014. **9**(10): p. 1382-96.

300. Fortin, J.P., et al., *Functional normalization of 450k methylation array data improves replication in large cancer studies.* Genome Biol, 2014. **15**(12): p. 503.

301. 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, 2013. **29**(2): p. 189-96.

302. Team, R.C., *R: A language and environment for statistical computing.*, in *R Foundation for Statistical Computing*. 2016: Vienna, Austria.

303. Liang, K.Y. and S.L. Zeger, *Longitudinal Data-Analysis Using Generalized Linear-Models.* Biometrika, 1986. **73**(1): p. 13-22.

304. Bates, D., et al., *Fitting Linear Mixed-Effects Models Using lme4.* J Stat Softw, 2015. **67**(1): p. 1-48.

305. Viechtbauer, W., *Conducting Meta-Analyses in R with the metafor Package.* J Stat Softw, 2010. **36**(3): p. 1-48.

306. Martens, J.H. and H.G. Stunnenberg, *BLUEPRINT: mapping human blood cell epigenomes.* Haematologica, 2013. **98**(10): p. 1487-9.

307. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-61.

308. Azab, B., et al., *Value of platelet/lymphocyte ratio as a predictor of all-cause mortality after non-ST-elevation myocardial infarction.* J Thromb Thrombolys, 2012. **34**(3): p. 326-334.

309. Meng, X.C., et al., *The platelet-to-lymphocyte ratio, superior to the neutrophil-tolymphocyte ratio, correlates with hepatitis C virus infection.* Int J Infect Dis, 2016. **45**: p. 72-77.

310. Martin, T.C., et al., *coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns.* Bmc Bioinformatics, 2015. **16**: p. 131.

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

312. Phillips, P.C., *Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems.* Nat Rev Genet, 2008. **9**(11): p. 855-67.

313. Purcell, S., *Variance components models for gene-environment interaction in twin analysis.* Twin Res, 2002. **5**(6): p. 554-571.

314. Bouchard, T.J., Jr. and M. McGue, *Genetic and rearing environmental influences on adult personality: an analysis of adopted twins reared apart.* J Pers, 1990. **58**(1): p. 263-92.

315. Stranger, B.E., E.A. Stahl, and T. Raj, *Progress and promise of genome-wide association studies for human complex trait genetics.* Genetics, 2011. **187**(2): p. 367-83.

316. Hong, E.P. and J.W. Park, *Sample size and statistical power calculation in genetic association studies.* Genomics Inform, 2012. **10**(2): p. 117-22.

317. Hoffman, G.E., *Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions.* PLoS One, 2013. **8**(10).

318. Price, A.L., et al., *New approaches to population stratification in genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 459-63.

319. Manichaikul, A., et al., *Analysis of family- and population-based samples in cohort genome-wide association studies.* Hum Genet, 2012. **131**(2): p. 275-87.

320. Wu, C., et al., *A comparison of association methods correcting for population stratification in case-control studies.* Am Hum Genet, 2011. **75**(3): p. 418-27.

321. Altshuler, D.M., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-+.

322. Pennazio, S., P. Roggero, and M. Conti, *A history of plant virology. Mendelian genetics and resistance of plants to viruses.* New Microbiol 2001. **24**(4): p. 409-424.

323. Oh, H.J., et al., *Long-term enzymatic and phenotypic correction in the phenylketonuria mouse model by adeno-associated virus vector-mediated gene transfer.* Pediatr Res, 2004. **56**(2): p. 278-84.

324. Gratten, J. and P.M. Visscher, *Genetic pleiotropy in complex traits and diseases: implications for genomic medicine.* Genome Med, 2016. **8**(1): p. 78.

325. Pickrell, J.K., et al., *Detection and interpretation of shared genetic influences on 42 human traits.* Nat Genet, 2016. **48**(7): p. 709-17.

326. Lin, B.D., et al., *Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample.* Genes, 2015. **6**(3): p. 559-576.

327. Gao, X., et al., *Genetic determinants of hair color and Parkinson's disease risk.* Ann Neurol, 2009. **65**(1): p. 76-82.

328. Parkes, M., et al., *Genetic insights into common pathways and complex relationships among immune-mediated diseases.* Nat Rev Genet, 2013. **14**(9): p. 661-673.

329. Li, R., et al., *Structural model analysis of multiple quantitative traits.* PLoS Genet, 2006. **2**(7): p. e114.

330. Su, Z., Y. Zeng, and X. Gu, *A preliminary analysis of gene pleiotropy estimated from protein sequences.* J Exp Zool B Mol Dev Evol, 2010. **314**(2): p. 115-22.

331. Yang, J.A., et al., *GCTA: A Tool for Genome-wide Complex Trait Analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.

332. Zhou, X., *A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-wide Association Studies.* bioRxiv, 2016.

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

334. Bell, J.T. and T.D. Spector, *A twin approach to unraveling epigenetics.* Trends Genet, 2011. **27**(3): p. 116-25.

335. van Dongen, J., et al., *Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells.* Genes, 2014. **5**(2): p. 347-65.

336. Ng, J.W., et al., *The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities.* Genome Biol, 2012. **13**(6): p. 246.

337. Barrero, M.J., S. Boue, and J.C.I. Belmonte, *Epigenetic Mechanisms that Regulate Cell Identity.* Cell Stem Cell, 2010. **7**(5): p. 565-570.

338. 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, 2009. **106**(23): p. 9362-7.

339. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.

340. Ackermann, M., W. Sikora-Wohlfeld, and A. Beyer, *Impact of natural genetic variation on gene expression dynamics.* PLoS Genet, 2013. **9**(6): p. e1003514.

341. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.

342. Liu, F., et al., *Detecting Low Frequent Loss-of-Function Alleles in Genome Wide Association Studies with Red Hair Color as Example.* PLoS One, 2011. **6**(11).

343. Lee, S., et al., *Rare-Variant Association Analysis: Study Designs and Statistical Tests.* American Journal of Human Genetics, 2014. **95**(1): p. 5-23.

344. Rivas, M.A., et al., *Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.* Nat Genet, 2011. **43**(11): p. 1066-U50.

345. Gudmundsson, J., et al., *A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer.* Nat Genet, 2012. **44**(12): p. 1326-1329.

346. Usher, C.L. and S.A. McCarroll, *Complex and multi-allelic copy number variation in human disease.* Brief Funct Genomics, 2015. **14**(5): p. 329-338.

347. Bi, H.R., et al., *Copy number variation of E3 ubiquitin ligase genes in peripheral blood leukocyte and colorectal cancer.* Sci Rep-Uk, 2016. **6**.

348. Machado, L.R. and B. Ottolini, *An evolutionary history of defensins: a role for copy number variation in maximizing host innate and adaptive immune responses.* Front Immunol, 2015. **6**.

349. Traherne, J.A., et al., *Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex.* Hum Mol Genet, 2010. **19**(5): p. 737-751.

List of Pulications

Lin BD, et al., Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample. Genes, 2015. 6(3). 559-576.

Lin BD, et al., The Genetic Overlap Between Hair and Eye Color. Twin Res Hum Genet, 2016. 19(6). 595-599.

Lin BD, et al., Causes of variation in the neutrophil-lymphocyte and platelet-lymphocyte ratios: a twin-family study. Biomark Med, 2016. 1061-1072.

Lin BD, et al., SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio. Hum Mol Genet, Manuscript submitted for publication.

Lin BD, et al., Genetic and environmental causes of variance in monocyte-lymphocyte ratio level. Twin Res Hum Genet, 2017 Feb 14:1-11. doi: 10.1017/thg.2017.3.

McArtor DB***, Lin BD***, et al., Establising the haematological profile: the interactive effects of age, sex and lifestyle. Biomark Med, under review.

Carnero-Montoro E*, **Lin BD***, et al., Blood hympomethylation is associated with elevated myeloid: lymphoid ratios in cell-specific active genomic regions. Manuscript is submitted Blood journal.

Baselmans BM, van Dongen J, Nivard MG, **Lin BD**, BIOS Consortium, Zilhão NR, Boomsma DI, Bartels M., Epigenome-Wide Association Study of Wellbeing. Twin Res Hum Genet, 2015. 18(6): p. 710-9.

Korporaal SJA, Estourgie-van Burk GF, **Lin BD**, Jones CI, Bartel M, Ouwehand WH, Goodall AH, Boomsma DI, de Groot PG. Inheritance of the platelet response measured in a population of twins. Manuscript in preparation.

Hysi P, et al. A GWAS meta-analysis of two large populations of European ancestry identifies numerous new genetic loci explaining significant portions of hair color heritability. Nat Genet, under review.

*Both are first co-authers and contribute equally.

Acknowledgments

My greatest gratitude goes first to my supervisors Prof. Dorret Boomsma, Dr. Jouke-Jan Hottenga and Dr. Gonneke Willemsen, who have guided me with great expertise through my graduate curriculum and who have shared the excitement of four years of scientific discovery with me. Both their unwavering enthusiasm for the field belonging to my research topics and their personal generosity helped to offer me an enjoyable time in Amsterdam. Meanwhile, my deep gratitude goes to Chinese Scholarship Council, who has financial supported my PhD study for four years.

My appreciation also extends to my visiting study supervisor Prof. Gitta Lubke from Notre Dame University in U.S.. Thanks to her patient and demanding supervising for one semester, I learned much more than expected. That semester was an unforgettable and special period in my life. The gratitude also goes to the colleagues in Gitta's quantitative psychology group: Dan, Ian, Justin and Prick. I also want thank other colleagues I met in Gitta's department: Ge, Haiyan, Han, Mengyu, Megan, Miao, Qian and Yujiao for your support and friendly company in ND.

I thank the reading committee for devoting the significant amount of time to this thesis: Prof. CV Dolan, Prof. NG Martin, Prof. H Snieder, Prof. M Kayser, Dr. DJA Smit, Dr. R Menezes. A special acknowledgement goes to the Chairman: Conor Dolan, for the patient help regarding my twins modeling and the daily greeting at the office door.

I cordially thank my paranymphs: Jenny and Junfeng for all your significant advice, support and help for my work and my social life.

I thank my colleagues at BioPsy: Abdel, Anne, Anouk, Bart, Camelia, Charlotte, Christel, Connor, Dennis, Dirk, Eco, Elsje, Eveline, Fiona, Gonneke, Hamdi, Harmen, Harold, Ineke, Iryna, Jacqueline, Jenny, Kees-Jan, Klaasjan, Lannie, Lot, Maria, Mathijs, Meike, Michel, Michiel, Natascha, Nienke, Nuno, Quinta, Rene, Sabine, Sara, Sanja, Suzanne, Toos and Yayouk. It was really a pleasure to study, to work and to have fun with you. I will never forget the time that we did homework together, solved statistical problems, did 'stukjes', een borrel hadden, and learned a little bit Chinese or Dutch.

I am grateful that I have many friends in Netherlands, who mostly came from the Chinese community. They supported me and made my life abroad much more fun: Anna Swierta, Aniket Mishra, Enqi He, Fangfang Cai, Gang Wu, Guangquan Chen, Guanyi Sun, Jiangyu Ji, Jimmy Vuong, Jin He, Junfeng Wang, Junhui Wu, Kai Liu, Liang Yu, Liang Zhao, Lintao Li, Weiluan Chen, Matthew Aburn, Ma Yuan, Meichen Yu, Ning Chen, Ouyang Xiyu, Pengqi Xu, Qiongjie Chen, Sunyun Chen, Tie Liu, Xiangyu Rao, Xiaolin Zhang, Xilin Sun, Xingnan Lin, Yue Li, Yiyi Yang, Yuanjie Yu, Yuanyuan Zhang and Yun Jiang.

I would like to express my special gratitude to my dearest friends Marion Krings and Prof. Antoine Braet. Thanks to your so many hours of patient English and Dutch teaching.

Above all, I am indebted to my family, especially to my passed away grandpa Guoqin Zhao, whose value to me only grows with age. And my cover picture designer Saymand, who blessed me with a life of joy in the hours when the work PC was off.

Finally, I thank you, my dear reader, for your interest in my PhD work.