GENETIC ARCHITECTURE OF HEMATOLOGIC TRAITS AND HEALTHY AGING-RELATED ENDOPHENOTYPES IN THE LONG LIFE FAMILY STUDY

by

Jatinder Singh

B.Sc., Punjabi University, India, 2002

M.Sc., Thapar University, India, 2005

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Jatinder Singh

It was defended on

December 17th, 2013

and approved by

M. Michael Barmada, Ph.D.
Associate Professor, Department of Human Genetics
Graduate School of Pittsburgh, University of Pittsburgh

Daniel E. Weeks, Ph.D.
Professor, Department of Human Genetics
Graduate School of Pittsburgh, University of Pittsburgh

Anne B. Newman, M.D., M.P.H.
Professor and Chair, Department of Epidemiology
Graduate School of Pittsburgh, University of Pittsburgh

Dissertation Advisor: Candace M. Kammerer, Ph.D.
Associate Professor, Department of Human Genetics
Graduate School of Pittsburgh, University of Pittsburgh

GENETIC ARCHITECTURE OF HEMATOLOGIC TRAITS AND HEALTHY AGING-RELATED ENDOPHENOTYPES IN THE LONG LIFE FAMILY STUDY

Jatinder Singh, PhD

University of Pittsburgh, 2014

ABSTRACT

One of the goals of medicine and public health is to increase functional longevity. Anemia and other age-related blood cell trait abnormalities have been shown to be associated with adverse outcomes such as disability, hospitalization, morbidity and mortality in older adults. Results from the third National Health and Nutrition Survey (1988-1994) indicated that 11.0% of men and 10.2% of women ≥ 65 years of age were anemic. As the US and global populations age, the prevalence of hematologic disorders and all age-related disorders will increase. The identification of genes or novel biological pathways that regulate hematologic traits and healthy-aging phenotypes could lead to insights and possible future interventions to delay the onset of hematologic diseases, increase functional longevity, and concomitantly, decrease the burden of age-related diseases on public health. In the current study, data on from a unique population comprising long-lived siblings and their families (the Long Life Family Study) were used to identify genes that may influence age-related traits, such hematologic traits and healthy aging endophenotypes. Using family-based whole genome linkage and association analyses, I identified multiple loci that may affect hematologic traits and endophenotypes. The most promising results are as follows. I identified (and subsequently replicated) a locus on chromosome 11p15.2 near *SOX6* (a transcription factor gene) that influenced RBC count. I also used factor analyses to extend results of previously developed endophenotypes derived from five health domains (cognition, physical function, cardiovascular, metabolic and pulmonary). The

primary endophenotype (predominantly reflecting pulmonary and physical function traits) was significantly related to reduced mortality. In addition, this endophenotype and the relationship to mortality was replicated in an independent, population-based cohort. I also identified (and replicated) association of this endophenotype to a locus on chromosome 18q11.2 near *ZNF521*, a transcription factor gene. Intriguingly, both *SOX6* and *ZNF521* have been reported to play a role in erythropoiesis, consistent with the hypothesis that aging may result, in part, from fundamental biological processes that influence multiple disorders. These results also indicate that genetic studies in a unique set of families may reveal novel findings that will increase our understanding of the genetic regulation of aging.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to acknowledge the support and help from numerous people who have provided me with guidance, resources and assistance and made the completion of this thesis possible. First and foremost, I would like to thank my parents and my sister who have motivated and encouraged me throughout my academic career.

I would like to express my sincere gratitude to my Ph.D. advisor Dr. Candace Kammerer who has given me the opportunity and encouragement to start, develop and complete my doctoral work. Her practical approach, intellectual rigor and insightful comments were extremely helpful in guiding my research. I am really thankful to her for providing me the learning opportunity and the environment for my scientific development.

I would also like to thank my dissertation committee, Dr. Barmada, Dr. Weeks and Dr. Newmann for their guidance, critique, excellent feedback and for being so generous with their time for my dissertation.

I am greatly indebted to the faculty, staff and students of Department of Human Genetics and Department of Epidemiology who contributed to my doctoral education. I would like to thank my colleagues in Dr. Kammerer's lab: Dr. Ryan Minster, Dr. Allison Kuiper, and Kimberly Jacoby for the advice, help and encouragement.

I would like to thank the participants, staff, investigators, collaborators and funding sources of the Long Life Family Study; the Dynamics of Health, Aging and Body Composition Study for giving me the opportunity to be part of these projects.

Finally, a very special thanks goes to my partner, Jonida Cali, who has supported and encouraged me especially during the critical final months of my dissertation. I love you Jona and thank you for being there when I needed you the most.

1.0 INTRODUCTION

1.1 PUBLIC HEALTH SIGNIFICANCE

1.1.1 Hematologic traits

Hematologic traits such as counts of white blood cells (WBC), red blood cells (RBC) and platelets (PLT) and volume of red blood cells and platelets (MCV, MPV) and hemoglobin levels (HGB) are routinely used as important diagnostic markers in clinical practice because abnormalities in these traits are associated with a number of diseases including anemia, sickle cell disease, polycythemia etc.

Table 1.1: Blood Trait Abbreviations

RBC	Red Blood Cells
HCT	Hematocrit
MCV	Mean Corpuscular Volume
HGB	Hemoglobin
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MPV	Mean Platelet Volume
PLT	Platelets
WBC	White Blood Cells
ANEU	Absolute Neutrophil
ALYM	Absolute Lymphocyte

Anemia is defined by the World Health Organization (WHO) as hemoglobin levels less than 13 g/dL for men and less than 12 g/dL for non-pregnant women. MCH (hemoglobin amount per red blood cell) and MCV (average red blood cell volume) indices are used to define the types of anemia, such as microcytic, macrocytic and others. Macrocytic anemia may result from vitamin B12 or folate deficiency, liver disease, aplastic anemia etc., and presents with abnormally large red cells (MCV > 98 fL)[1]. Conversely, MCV and MCH are low in microcytic hypochromic anemia that may result from iron deficiency, sideroblastic anemia, thalassemia etc.[1] Hemoglobin disorders, such as sickle cell anemia and β-thalassemia, are among the most common inherited monogenic disorders in the world, especially in tropical regions of the world[2]. The high prevalence and geographic location of these hematologic diseases overlaps that of malaria, and our current understanding of the etiology and biology of malaria is consistent with the hypothesis that the prevalence of these genetic disorders is due to heterozygotes being more fit than both homozygotes[3]. According to one estimate, a minimum of 332,000 children are born each year with hemoglobin disorder[4]. Anemia is also common in the elderly population.

In the elderly population, anemia is often classified into three predominant types; chronic disease anemia, nutritional deficiency anemia and unexplained anemia[5]. Results from the Third National Health and Nutrition Examination Survey (NHANES III) data, a nationally representative sample of community dwelling persons, indicated that 11.0% of men and 10.2% of women ≥ 65 years of age were anemic (using the WHO definition of anemia). At older ages, the prevalence of anemia increases more rapidly in men than women; Skjelbakken *et al.* estimated that among individuals ≥ 85 years of age, 29.6% of men and 16.5% of women were anemic[6]. The prevalence of anemia also differs significantly by race. According to NHANES III data, elderly non-Hispanic blacks have three times the prevalence of anemia compared to elderly

non-Hispanic whites. Hemoglobin levels in non-Hispanic blacks are 4.0 – 10.0 g/L lower than in non-Hispanic whites and these differences persist even after adjusting for age, socio-economic status and iron intake[7,8]; thus it is not surprising that blacks show higher prevalence of anemia compared to whites.

Anemia and hemoglobin concentrations have been shown to be associated with adverse outcomes such as disability, hospitalization, morbidity and mortality in older adults [5,9,10,11]. For example, Izaks *et al.* investigated the association between hemoglobin concentration and mortality in a community-based study from the Netherlands, involving 755 individuals, age 85 or older. Risk of mortality was 1.60 (p-value < 0.001) in women with anemia and 2.29 (p-value < 0.001) for men with anemia, as compared with persons having normal hemoglobin concentration[12]. Patel *et al.* also reported an increased risk of mortality among white men and women with anemia, although the risk for women was higher than that of men (age-adjusted hazard ratio = 1.96 and 2.86 for men and women, respectively). However, for black men and women, no association was observed between mortality and anemia[13]. The results above indicate a need for race specific thresholds for defining anemia.

Elevated white blood cell count is a hallmark of acute or chronic systemic inflammation. Systemic inflammation, as measured by C-reactive protein, has been associated with mortality in a population-based sample of healthy older individuals[14]. In addition, many studies have implicated higher WBC counts as an independent risk factor for coronary artery disease (CAD) and myocardial infarction (MI)[15,16,17]. WBC counts have also been associated with all cancer mortality[18]. Furthermore, elevated platelet counts have been associated with coronary heart disease (CHD)[19] and insulin resistance in non-obese diabetic patients[20].

1.1.2 Healthy Aging-Related Endophenotypes and the Healthy Aging Index

One of the current challenges in medicine and public health is to enable individuals to achieve a long and healthy life. Modifications of some environmental and lifestyle factors, such as diet, are known to increase longevity and healthy aging at the population level. Studies of the genetic basis of healthy aging and longevity in humans and animal models may reveal additional mechanistic insights, and thus enable public health professionals to develop interventions to delay onset of age-related disorders.

Longevity and healthy aging are complex phenotypes. Although longevity is easy to measure (age at death), it does not measure individual functionality and it is not an ideal phenotype for genetic studies, because long wait times are required. On the other hand, healthy aging could be measured at many ages, but there is no single "healthy aging phenotype" or definition of "disease-free survival." Recently, my colleagues in the Long Life Family Study derived a Healthy Aging Index, HAI, to measure subclinical disease[21,22] as well as an endophenotype[23] in an effort to increase our ability to detect loci that influence longevity and function. Endophenotypes have been defined as underlying traits that influence development of disease and may be estimated by factor analyses of correlated physiologic measures. The HAI and derived endophenotypes may better characterize a long and highly functional life without cognitive decline than do single trait measures[24,25]. Furthermore, such indices and endophenotypes may improve detection of genes associated with high physical and cognitive functions. A detailed description of these traits is provided below.

1.1.3 Summary of Public Health Impact

Because the number of older adults is increasing both in the US and globally, elucidation of the mechanisms and biological pathways that regulate hematologic traits and their relationship to age-related outcomes could provide new insights into additional measures of prophylaxis that may delay or mitigate onset of hematologic disorders and their sequelae. Similarly, identification of genes and/or biological pathways that contribute to healthy aging could lead to insights and possible future interventions to increase functional longevity and concomitantly decrease the burden of age-related diseases on public health.

1.2 GENETIC EPIDEMIOLOGY OF HEMATOLOGIC TRAITS AND ENDOPHENOTYPES

1.2.1 Hematologic Traits

Environmental covariates and heritability: Quantitative variation in hematologic traits is highly heritable and is under the influence of both genetic and environmental factors. Several studies have reported the effect of environmental factors such as age, sex, obesity, smoking, alcohol and oral contraception on the levels of blood traits[26,27,28,29]. A study by Fisch *et al.* involving 14,961 healthy women identified smoking, oral contraceptive use and obesity as important factors influencing white blood cell counts[30]. In the Framingham Heart Study, a set of environmental covariates (including age, sex, height, weight, HDL (High Density Lipoprotein) levels, triglyceride levels, total serum protein, diabetes, smoking and alcohol) explained 47%, 14%, 9%,

40% and 49% of the total variation in RBC count, MCV, MCH, HCT and HGB respectively[31,32] (Table 1.1 for abbreviations). Distinct ethnic groups also show significant differences in hematologic traits. As mentioned previously, hemoglobin levels in non-Hispanic blacks are 4.0 – 10.0 g/L lower than in non-Hispanic whites. Similarly, non-Hispanic blacks are known to have lower WBC and neutrophil counts than non-Hispanic whites[33].

Several studies with monozygotic and dizygotic twins have reported that genetic factors account for 40 to 90% of the observed variation in the blood traits. Residual heritability estimates for RBC count, MCV, MCH, HCT, HGB were 56%, 52%, 52%, 41% and 45% respectively in the Framingham Heart Study[31,32] (Residual heritability is the proportion of total phenotypic variation after removing effects of measured covariates). The reported moderate to high heritabilities of hematologic traits indicate that performing genomewide linkage (GWL) and genomewide association (GWA) studies to detect and identify genetic factors should be successful.

Statistical genetic analysis of RBC-related traits: Linkage studies have identified significant evidence for quantitative trait loci (QTL) influencing several hematologic traits including: hematocrit (HCT) on chromosome 6q23[32], RBC count on chromosomes 19p13, 12p13, 11p15.2, 18p11.32, MCH on chromosome 11p15.5, and MCV on chromosome 11p15.5[31,34].

In particular, the *HBS1L-MYB* region on chromosome 6q23 has been identified by multiple studies as a key regulator of blood traits. Via linkage analysis, this region was initially identified in an Asian-Indian kindred to contain a genetic determinant for fetal hemoglobin (HbF) production[35]. Fine mapping of this region identified genetic variants associated with HbF levels residing in the *HBS1L* and *HBS1L-MYB* intergenic region (HMIR). The most significant

common SNP in this region was rs9399137, but was not reported to be a functional SNP[36]. Subsequently, other SNPs within this region have also been shown to be associated with MCV, RBC, MCH, MCHC, HCT[37,38,39], WBC[40] and PLT[41]. Recently, Farrell *et al.* reported an association between HbF expression and a three base pair deletion in HMIR. This deletion is in complete LD with the rs9399137 and encompasses a region having enhancer-like activity[42]. However, there may be additional functional variants within this region.

Iron is a key component of red blood cells; therefore, it is not surprising that genetic variation in genes involved in iron homeostasis (*HFE, TMPRSS6, TFR2*) have been reported to be associated with red blood cell related traits (HGB, MCH, MCV, HCT)[37,39,43]. *TMPRSS6* is a type II plasma membrane serine protease and plays an important role in iron hemostasis[44]. Chambers *et al.* reported the association of *TMPRSS6* with hemoglobin levels in individuals of European and Indian ancestry. The most significantly associated SNP (rs855791) is likely to be a causal variant as it results in nonsynonymous (V736A) change in the functional domain of the enzyme TMPRSS6 that alters its activity[45]. The nonsynonymous mutations (C282Y and H63D) in the *HFE* (High Iron Fe) gene are used routinely to confirm the clinical diagnosis of hereditary hemochromatosis[46].

In addition to linkage and candidate gene studies, several large consortia have performed genomewide association studies on hematologic traits. HaemGen[47], the first large consortium on hematological parameters, analyzed data on 13,943 individuals and has identified a total of 6 loci (22q12.3, 6p21.1, 6p21.3, 22q12, 6q23 and 7q22) that influence variation in red blood cell traits (RBC, MCV and MCH). The HaemGen consortium is comprised of six European population-based studies having average age ranging from 41.4 to 61.2 years. Another large consortium, the CHARGE (Cohorts for Heart and Aging Research in Genetic Epidemiology) Consortium[39],

analyzed data on 24,167 individuals of European ancestry for six red blood cell traits (RBC, HCT, MCH, MCHC, MCV and HGB) and identified 23 loci associated with at least one of the red blood cell traits, 17 of which were novel. The largest of the meta-analysis for red blood cell related traits was reported by *van der Harst et al.* in 2012[48], which included 135,367 individuals of European and South Asian ancestry. In total, 75 loci showed evidence of association, 43 of which were novel. However, similar to the results of meta-analyses of many other phenotypes and diseases[49], these identified variants explain little of the observed inter-individual variation in RBC count. For example, the CHARGE Consortium reported that variation at the two loci associated with RBC count, i.e., *HBS1L-MYB* and *EPO*, explained only 0.85% of variation in RBC count[39]. Furthermore, the majority of the SNPs associated with the hematologic traits are not known to be functional variants.

Statistical genetic analysis of WBC-related traits: Multiple quantitative trait loci (QTLs) associated with WBC counts and WBC subtypes have also been reported by many studies[50,51]. Many of these loci are associated with both neutrophil count and WBC count, which is not surprising because neutrophils are the most predominant type of WBC. One such QTL is the *PSMD3-CSF3* region on chromosome 17q21.1. This QTL region was significantly associated with WBC count and neutrophil count in the European and Japanese populations, respectively[37,52]. A priori, *CSF3* (Colony Stimulating Factor 3) was the most likely candidate gene in this region because it encodes a cytokine that regulates granulocyte production. However, Okada *et al.*[52] reported that the most significant SNP in the region is also associated with expression levels of proteasome 26S subunits non-ATPase 3 (*PSMD3*) and not with *CSF3*. Further studies are required to identify the functional genetic variants within this region, which regulate the counts of WBC and its subtypes.

Genetic association studies have also identified several variants that explain part of the difference in the total WBC counts between European and African populations. Using admixture mapping methods[53], the Duffy antigen receptor for chemokines (*DARC*) gene at 1q23 was associated with lower WBC and neutrophil counts in African Americans[54]. The Duffy "null" polymorphism (rs2814778) explains approximately 20% of inter-individual variance in baseline WBC count among African Americans and the frequency of this variant is estimated as 99.8 ± 0.1% in Africans and 0.7 ± 0.4% in Europeans. The "null" form of this variant abolishes the expression of the "Duffy antigen receptor for chemokines" (*DARC*) on RBC; and cellular expression studies have demonstrated that individuals with "Duffy negative" phenotype are resistant to invasion by *P. Vivax*[55]. Thus, this variant would be advantageous in regions in which malaria was present.

Genetic effects on RBC and WBC-related traits: In addition to the above loci, several candidate genes influence both RBC and WBC-related traits, particularly loci involved in cell division. For example, neutrophil count[50] and WBC[40] have been associated with common variants in *CDK6*, a gene located on 7q21 that encodes a cyclin dependent kinase that plays an important role in cell cycle progression. Additionally, common variants in members of the cyclin-D family, *CCND2* (12p13) and *CCND3* (6p21.1), that regulate CDKs (cyclin-dependent kinases) have been reported to be associated with MCH, MCV and RBC[37,39,40]. The above reports indicate that variants in some genes may influence RBC, WBC, and platelet traits individually, and particular genes or biological pathways may have pleiotropic effects on multiple hematologic traits.

1.2.2 Endophenotypes Derived from Five Health-Related Domains (Five-Domain Endophenotypes)

Previously, my colleagues in the Long Life Family Study (LLFS) had derived five heritable endophenotypes to assess exceptional survival. For the development of the endophenotypes, 28 measures from five domains were chosen based on availability and on hypothesized physiologic significance to exceptional survival[23]. The five domains and their continuous, physiological measures included (1) cognitive function: immediate memory, delayed memory, category fluency, and digit substitution forward and backward; (2) cardiovascular health domain: presence of hypertension, total cholesterol (milligrams per deciliter), high-density lipoprotein cholesterol (milligrams per deciliter), low-density lipoprotein cholesterol (milligrams per deciliter), triglycerides (milligrams per deciliter), systolic blood pressure (millimeter of mercury), diastolic blood pressure (millimeter of mercury), and pulse pressure (millimeter of mercury); (3) metabolic health: presence of diabetes, blood glucose (milligrams per deciliter), glycosylated hemoglobin, creatinine, body mass index (kg/m^2), and waist circumference; (4) pulmonary health: presence of lung disease, forced expiratory volumes (FEV1 and FEV6, milliliters), and FEV1/FEV6 ratio; and (5) physical functioning: average and maximal grip strength (kilograms), walking speed (meter per second), and total physical activity.

Using these 28 measures from five domains in LLFS, Matteini and colleagues[23] derived five endophenotypes by factor analysis. The first factor was predominantly comprised of pulmonary and physical function measures, and accounted for 14.4% of the variation, and was moderately heritable ($h^2 = 0.39$). These two domains are highly associated among older-aged individuals, although the underlying causes of this relationship are unclear (see Matteini *et al.*, 2010)[23]. The second factor consisted of metabolic and cholesterol-related traits, accounted for

11.9% of the variance and had modest heritability (h^2 = 0.25). Metabolic phenotypes, such as low insulin resistance, have been associated with longevity in multiple species[56]. The third factor was related to global cognition, accounting for 8.9% of the underlying variance with heritability of 0.36. The fourth factor was mainly characterized by blood pressure measures, accounting for 8.3% of the variance with an estimated heritability of 0.25. Finally, factor 5 was predominately comprised of total and LDL cholesterol, and accounted for 6.2% of the variation. The relationship between blood pressure and lipid traits with longevity is well known.

These endophenotypes may indicate the presence of pleiotropic effects on sets of genes on seemingly disparate traits and domains, e.g. pulmonary function and physical activity. Identification of genes that influence multiple aging-related traits may reveal pathways that could be exploited to develop novel interventions. Additional research is needed however. Although many of the individual components of the endophenotypes are associated with mortality, there is no evidence that the endophenotypes are related to mortality. In addition, these endophenotypes need to be validated in other populations.

1.2.3 Summary

GWL and GWA studies, including large consortia, such as CHARGE and HaemGen, have analyzed data on thousands of individuals and have identified many loci that influence variation in hematologic traits. However, similar to results of meta-analyses of many other phenotypes and diseases, the identified variants explain little of the observed inter-individual variation in hematologic traits. Identified loci, in the CHARGE consortium, explained 1.14% of HGB variation, 1.16% of HCT variation, 4.53% of MCH variation, 0.63% of MCHC variation, 5.98% of MCV variation and 0.85% of variation in RBC[39]. Furthermore, the majority of the variants

associated with these traits are not known to be functional variants. Finally, the overall genetic and biological architecture of hematologic traits remains unclear, as does the genetic relationship between these traits and age-related endophenotypes, as well as their relationship to age-related morbidity and mortality.

Longevity and healthy aging are complex traits. Numerous epidemiologic and genetic studies have been performed on measures of longevity, however, few genes have been identified that influence longevity in humans[57]. Furthermore, the genetic and environmental determinants of healthy aging, and the relationship of measure of healthy aging to mortality, are mostly unknown[58].

1.3 STUDY APPROACH AND SPECIFIC AIMS

Hematological phenotypes (e.g., counts of white blood cells, red blood cells and platelets) are heritable, play important roles in immune response, oxygen carrying and blood clotting, and are associated with age-related diseases, such as anemia. To date, genetic studies have identified multiple variants that are associated with hematologic traits, however, they account for little of the heritable variation. Furthermore, the relationship between these variants and susceptibility to age-related health outcomes is unclear. In addition, one of the fundamental goals of the LLFS is to identify genetic and environmental factors that influence healthy aging. Toward this goal, several endophenotypes that correlate with healthy aging have been constructed. These endophenotypes are heritable, but the specific genes that may affect these traits are unknown. The overall goals of my study were to (1) characterize the genetic architecture of hematologic traits by detecting and statistically characterizing possible quantitative trait loci (QTLs)

influencing these traits, (2) detect and identify QTLs that influence specific healthy aging endophenotypes, and then (3) assess the relationship of these traits and/or QTLs for these traits to age-related health outcomes. To achieve these goals, I employed a variety of statistical genetic and bioinformatic methods on phenotypic and genetic data that are available on a unique population of long-lived individuals and their families, the LLFS. I also used phenotypic and genotypic data from the Health Aging and Body Composition (HABC) Study to replicate my results from the LLFS.

Specifically, I completed the following general aims and answered the following questions.

Aim 1: Characterize the phenotypic and genetic architecture of hematologic traits and their relationship to measures of healthy aging (Chapter 2).

What is the heritability of each trait and the genetic correlations between them?

Are they genetically correlated to measures of healthy aging?

Aim 2: Detect and statistically characterize QTLs involved in the regulation of the hematologic traits and their endophenotypes (Chapters 2 and 3).

Do previously identified variants (from other studies) influence the hematologic traits in LLFS? Do novel QTLs (identified by genomewide linkage and association analyses) influence the hematologic traits (and endophenotypes) in the LLFS cohort?

Do these variants replicate in the HABC cohort?

Aim 3: Characterize the relationship of the healthy aging-related endophenotypes (five-domain endophenotype) to mortality (Chapter 4).

Are any of the five-domain endophenotypes associated with mortality?

Are these relationships replicated in another population (HABC cohort)?

Aim 4: Detect and statistically characterize quantitative trait loci (QTLs) involved in the regulation of novel healthy aging endophenotypes, especially the five-domain endophenotype (Chapter 5).

Do novel QTLs (identified by genomewide linkage and association analyses) influence the healthy aging endophenotypes in the LLFS cohort?

Do these variants replicate in the HABC cohort?

1.4 STUDY POPULATIONS

1.4.1 Long Life Family Study (LLFS)

LLFS comprises 4,535 individuals in 574 two-generation families: 1,515 in the older generation (mean age = 89.4 years), 2,255 in the offspring generation (mean age = 60.5 years) and 765 spousal controls (mean age = 60.8 years). These families were recruited by four sites, three in the US and one in Denmark, based on a measure of exceptional longevity[59]. Family eligibility and ascertainment criteria have been described previously[60]. Briefly, probands of age 89 years and older were identified and their families were selected based on Family Longevity Selection Score (FLoSS)[59], which ranks sibships by age of the siblings, the size of sibship and the number of individuals available for the study. These families were also required to meet the criteria of having a minimum family size of three (proband, at least one living sibling and one of their living sibling).

1.4.2 Replication Population – Health Aging and Body Composition Study (HABC)

Health Aging and Body Composition Study (HABC) is a longitudinal study of African American and European American men and women. Phenotypic and genotypic data are available for 2,802 individuals—1,139 African Americans and 1,663 European Americans—between the ages of 68 and 80. The individuals were drawn equally from two sites, Pittsburgh, Pennsylvania, and Memphis, Tennessee. For this study, I used phenotypic and genotypic data obtained from the European American cohort.

1.5 STUDY DATA

1.5.1 Phenotypes

The following ten hematological traits for 4,535 individuals belonging to 574 families of European ancestry were determined in EDTA whole blood using a Sysmex XE10 2100 instrument (Sysmex, Kobe, Japan): (1) Red blood cell (RBC) count (10^{12}/L); (2) Hemoglobin (HGB) level (g/dL); (3) Hematocrit (HCT; %): the volume percentage of the RBCs in blood; (4) Platelet (PLT) count (10^{9}/L); (5) White blood cell count (WBC) (10^{9}/L); (6) Absolute neutrophil (ANEU) count (10^{9}/L); (7) Absolute lymphocyte (ALYM) count (10^{9}/L); (8) Mean red blood cell hemoglobin, (MCH; pg), calculated as Hemoglobin(g/dL)/RBC(10^{12}/L) × 10; (9) Mean red blood cell hemoglobin concentration (MCHC) calculated as Hemoglobin(g/dL)/HCT(%) × 100; (10) Mean red blood cell volume (MCV) calculated as HCT(%)/RBC(10^{12}/L) × 10. Absolute

numbers for the WBC subtypes were obtained by multiplying each subtype's proportion with the total WBC count.

In addition, the following traits were used in the construction of the five-domain endophenotype[23]: (1) Cognitive domain: animal recall, vegetable recall, digit substitution forward and backward, immediate memory, delayed memory; (2) Cardiovascular health domain: presence of hypertension, systolic blood pressure, diastolic blood pressure, pulse pressure, total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides; (3) Metabolic health domain: presence of diabetes, BMI, creatinine, glucose, glycosylated hemoglobin, waist circumference; (4) Physical activity: average grip strength, maximum grip strength, gait speed, total physical activity; (5) Pulmonary domain: presence of lung disease, forced expiratory volume 1 (FEV1), forced expiratory volume 6 (FEV6), FEV1/FEV6 ratio.

1.5.2 Genotypes, Imputation, and Admixture Principle Components

Details of general genotyping, imputation and admixture principal components used for controlling population structure in LLFS and HABC are given below.

1.5.2.1 Long Life Family Study

The following two paragraphs have been provided by the LLFS Coordinating Center as a description of the general genotyping, imputation, and quality control methods and are recommended for use in all LLFS proposals and publications.

"The Center for Inherited Disease Research (CIDR) assayed all LLFS subjects using the Illumina Human Omni 2.5 v1 chip. Quality control was performed by CIDR and the LLFS Coordinating Center. We excluded 83,774 markers with < 98% call rate and 3,647 SNPs with a

high Mendelian error rate. In addition, we excluded 18 subjects who had < 97% genotype call rate. Finally, 153,363 Mendelian errors were set to missing in the families in which they occurred. After these quality control measures were applied, there were 4,693 subjects genotyped at 2,225,478 markers available for analysis. Principal components (PCs), for controlling for population structure, were produced with EIGENSTRAT (Price et al., 2006) using 116,867 tag SNPs on 1,522 unrelated LLFS individuals. These SNPs had MAF < 5% and HWE $p > 10^{-6}$. We also excluded SNPs from some chromosomal regions that may bias the PC analysis, including 2q21, 2q21.1, HLA1 and HLA (chromosome 6), 8p23.1, 8p23, and 17q21.31. PCs produced from unrelated subjects were expanded, within EIGENSTRAT framework, to all members of LLFS.

Additional imputed genotypes were generated based on the cosmopolitan phased haplotypes of 1000 Human Genome (1000HG, version 2010-11 data freeze, 2012-03-04 haplotypes). Programs used for imputation were MACH (version 1.0.16, for pre-phasing of LLFS data) and MINIMACH (version of May 2012) for performing imputations and ChunkChromosome script for splitting the LLFS data into smaller blocks to speed the process of imputation (Li et al., 2009, 2010). Imputations were performed in chunks with 5,000 SNPs blocks and 1,000 SNPs overlap from our data. Filters before imputing were: removing markers that had MAF < 1%, HWE $p > 10^{-6}$, if LLFS SNPs alleles mismatched with those of 1000HG, and not present in the 1000HG panel, as well as flipping any SNP when appropriate to the forward strand. A total of 38.05 million SNPs were imputed. Monomorphic SNPs and those with an imputation quality score of $r^2 < 0.3$ were removed. This reduced the potential variants for analyses to 18.3 million."

Note: Data on imputed SNPs were not available until late in my study (Summer 2013). Therefore, I did not have time to perform GWA analyses using data on all 18.3 million of the assayed and imputed SNPs. For my dissertation research, I performed genomewide family-based association analyses using data on 2.2 million assayed SNPs. Then, for each suggestive GWA signal, I performed family–based association analyses using data on all imputed and assayed SNPs within a 2 Mb window around the assayed SNP that had the lowest *p-value* (i.e., the 'lead' SNP). Before submitting manuscripts for publication, I will perform genomewide association analyses on all 18.3 million SNPs.

1.5.2.2 HABC Study

The following text was provided by the Wake Forest team of investigators for all researchers who use the HABC genotype data. "For all subjects in the HABC study, genotyping of genetic markers was performed by the Center for Inherited Disease Research (CIDR) using the Illumina Human1M-Duo BeadChip system. Samples were removed from the data if the sample failed overall (< 97% SNPs genotyped), if the chromosome sex did not match the reported sex or if first-degree relatedness was detected using the SNP data. SNPs were removed if the SNP had a minor allele frequency (MAF) < 1%, was called with < 97% success, or had a Hardy-Weinberg equilibrium (HWE) test p value < 10^{-6}. A total of 1,151,215 autosomal SNPs were successfully genotyped in 1,663 European American individuals and were carried forward to imputation. Principle components of ancestry were derived by the investigators at Wake Forest using EIGENSTRAT. They determined that two ancestry PCs were sufficient to account for genetic admixture in European Americans. Imputation was performed using MACH 1.0.16 and the HapMap II phased haplotypes as the reference. Genotypes were available for 914,263 SNPs based on the HapMap CEPH reference panel (rel. 22, b36). A total of 2,543,887 genotyped and

imputed autosomal SNPs were ultimately available for analysis as part of the 'genotyped and HapMap-imputed SNPs' set. A total of 40,949 Chromosome X SNPs were successfully genotyped in all European Americans subjects. An additional 40,818 SNPs were imputed using a method similar to that used for the autosomes for a total of 81,767 X chromosome SNPs. The chromosome X SNPs were included in the "genotyped and HapMap-imputed SNP" set for a total of 2,625,654 SNPs.

A second set of genotyped and imputed SNPs was prepared from the 1.2 million successfully genotyped SNPs and 1,663 subjects using the 1000 Genomes reference haplotypes (June 2010 release). A total of 6,858,264 genotyped and imputed autosomal SNPs were available as part of the 'genotyped and 1000 Genomes–imputed SNPs set.' The HapMap imputation was performed by Yongmei Liu and Kurt Lohman of Wake Forest University. The 1000 Genomes imputation was performed by Michael Nalls of the National Institutes of Health."

1.5.3 Genotypes/Haplotypes for Linkage Analyses in LLFS

Because many of the LLFS families are relatively complex and comprise three generations, the LLFS group needed to reduce the numbers of SNPs used to create Multipoint Identity By Descent (MIBD) matrices for performing linkage analysis. Initially, three different SNP sets were chosen from 'cleaned' genotyped SNPs: one set by the LLFS Coordinating Center (stLouis SNP set) and two sets (PittA and PittB) by Dr. Ryan Minster. SNPs were chosen to have MAF close to 0.5 to be maximally informative and at intervals ~ 1 cM as failure to model LD between SNPs can erroneously increase the sharing estimates. I used Loki[61] to estimate multipoint Identity By Descent (IBD) probabilities every 1 cM for the SNP sets chosen by Dr. Ryan Minster. Because individual SNPs are not as informative as microsatellite markers, I initially

performed linkage analysis using MIBD matrices derived from multiple SNP sets and assessed the results for consistency.

Subsequently, the LLFS Coordinating Center developed multiallelic haplotypes for use in the linkage analyses. The following text has been provided by the Long Life Family Study Coordinating Center as a description of the generation of haplotypes and MIBD matrices for linkage analyses. "The Long Life Family Study Coordinating Center generated multiallelic haplotypes across the LLFS genomes that would be more informative of identity-by-descent than biallelic markers alone. Haplotypes were constructed using ZAPLO[62]. To select SNPs for haplotypes within small regions, we divided the genome into 0.5 cM intervals; the cM positions of SNPs were approximated by linear interpolation from the deCODE map and base-pair positions of the SNPs. We removed all SNPs that had Mendel inconsistencies and an average pedigree heterozygosity ≤ 0.1. Within each 0.5 cM interval we used the first five such SNPs to construct a haplotype and if there were fewer than five SNPs, we used all SNPs in the interval. For a few individuals, no zero-recombination haplotype configurations within a specific 0.5 cM region were possible. These haplotype estimates were designated as missing, however, because of the density of the intervals and the high information of the haplotypes, very little information was lost. Multipoint IBD estimates from the haplotype data were calculated using Loki with a mean spacing of 0.5 cM."

1.6 STATISTICAL METHODS

1.6.1 Development of Endophenotypes for Hematologic Traits and Healthy Aging-Related Endophenotypes

The methods described below were developed for use on unrelated individuals (samples), however, LLFS is comprised of families. In general, the parental generation (and married-ins) would be a logical set on which to perform the clustering and principle components analyses. However, the probands for these families were long-lived *siblings*, and the parents of these siblings (that is, the founders) are deceased. Therefore, I used an iterative, random sampling procedure to obtain an unbiased estimate of the correlation matrix as follows. In general, one person was randomly selected from each family and correlations among variables were calculated. This procedure was done 1,000 times to generate a matrix of average correlations across all iterations, and this average correlation matrix was used for hierarchical clustering and to calculate eigenvectors.

(a) Hierarchical clustering: Hierarchical clustering is a statistical method that organizes the data points/samples in the form of a cluster tree or dendrogram based on pairwise distance/similarity between them. I first determined whether phenotypic correlations between hematologic traits differed among related family members and spousal controls, by performing hierarchical clustering separately for each group. For the spousal controls, all individuals were used for clustering. For related family members, pairwise correlations were calculated by using the iterative process, as described earlier. Pairwise correlations between hematologic traits were used as the distance metric and cluster trees were generated using *hclust* method as implemented in R suite of statistical packages (R Foundation for Statistical Computing, Vienna, Austria).

Along with dendrograms, heatmaps (distance scores displayed as colors) were generated using the *pheatmap* function in R for better visual representation of the data structure.

*(b) Principal Components Analysis and Factor Analysis***:** Principal components analysis (PCA) and factor analysis (FA) was conducted to develop composite traits (or endophenotypes) from the set of hematologic traits and the five health-related domains, respectively. The composite traits are linear combinations of correlated components. Composite phenotypes may better capture underlying genetic variation than the individual components that comprise them. Analyses were performed in R using the *princomp* function using the average correlations matrix, calculated from related family members using the iterative process as described earlier. Before calculating the correlation matrix, the traits were adjusted for covariates and standardized, as PCA and FA are sensitive to scaling. For FA, principal components extraction with varimax rotation was utilized to extract factors. The principal components and factors were used to calculate scores (endophenotypes) for each individual in LLFS (related and controls) by multiplying the standardized hematologic trait values (or traits from the five health domains) by the eigenvectors. These endophenotypes were used in statistical genetic analysis, such as heritability, linkage and association. If the heritability of an endophenotype is significantly greater than zero, it implies that a similar set of genes underlies variation in the individual components of the endophenotype.

(c) Genetic correlations among hematologic traits: Based on our current understanding of the biology of the hematological traits, I expected that several of the traits would be genetically correlated. In other words, a gene or a set of genes influences variation in both traits, having pleiotropic effects. To quantify the underlying genetic relationship among traits, I performed bivariate analyses to estimate the genetic and environmental correlation between

different blood traits using the variance component framework[63]. The phenotypic correlation (ρ_P) between traits can be partitioned into additive genetic correlation (ρ_G) and (unmeasured) environmental correlation (ρ_E) as:

$$\rho_P = \rho_G \sqrt{h_2 r_1} \sqrt{h_2 r_2} + \rho_E \sqrt{1 - h_2 r_1} \sqrt{1 - h_2 r_2}$$

where h^2_{r1} and h^2_{r2} are residual heritabilities for traits 1 and 2 (estimation of heritability is discussed in section 1.6.3, below). The significance of the additive genetic correlation (ρ_G) among pairs of traits is tested by using the likelihood ratio test (LRT) and comparing the log likelihoods of a model in which ρ_G is constrained to 0 (null hypothesis of no genetic correlation between traits), to that of a model in which ρ_G is estimated for the traits. If the results of the test are significant, this is evidence for pleiotropy (i.e., a common set of genes influence both traits). The extent of covariation is assessed by a second test in which ρ_G is constrained to 1 (i.e., the covariation among traits is due to the same set of genes). The alternate hypothesis is that some genes affecting one trait do not influence the second trait and vice versa, if ρ_G is estimated to be significantly different from 0.

1.6.2 Relationship with Mortality (Cox Proportional Hazards Regression)

My colleagues and I also assessed the relationship between the five-domain endophenotypes and mortality using Cox proportional hazards regression. To assess the ability of the five-domain factors to predict mortality, we used the area under the receiver-operator curve method and calculated the concordance statistics (c-statistic). C-statistics from different models were compared using the method described by DeLong *et al.* (1988)[64]. We also assessed whether models including age alone, or endophenotype factors alone, or models including age, factors,

and sex were best at predicting mortality. These analyses were performed by Dr. Robert Boudreau and Tanushee Prasad.

1.6.3 Effects of Known Covariates and Heritability

To assess the effects of covariates (previously reported in the literature) and the heritability of the hematologic traits and the five-domain endophenotypes, I used the variance component framework as implemented in Sequential Oligogenic Linkage Analysis Routines, SOLAR[65]. This framework will also be used for the linkage and bivariate analyses. Briefly, this method partitions the total phenotypic variance (σ^2_P) into additive genetic (σ^2_G), environmental (σ^2_E) and unmeasured error (σ^2_e) components.

Heritability is defined as proportion of the total variance that is due to additive genetic factors ($h^2 = \sigma^2_G / \sigma^2_P$). Heritability of the hematologic traits and endophenotypes was estimated using variance decomposition methods which partition the phenotypic variation into three components: (1) measured covariates (σ^2_E) such as age and sex, (2) additive genetic factors (σ^2_G), estimated using the kinship between the pairs of relatives and (3) unmeasured error components (genetic and environmental). Mathematically, these components can be represented as:

$$y_i = \mu + \sum_{i=1}^{n} \beta_j X_{ij} + g_i + e_i$$

where μ is the overall mean, β_j is the regression coefficient for the X_{ij} (jth covariate for the ith individual), g_i is the additive genetic effect and e_i is the unmeasured error component. Pedigree-based maximum likelihood methods were used to estimate the model parameters. The significance of the parameters was tested using the likelihood ratio test (LRT) by comparing the likelihoods of models with and without the parameter in question. The LRT for effects of

covariates is approximately distributed as a chi-square distribution with 1 degree of freedom. For tests of heritability, the LRT follows a 50:50 mixture of a point mass at zero and a chi-square distribution with 1 degree of freedom. Residual heritability (h^2r) is defined as proportion of total trait variance due to additive genetic component after adjusting for measured environmental covariates.

The hematologic traits were assessed for effects of the following covariates (based on the literature); field center, age, sex, age-squared, BMI, smoking status, alcohol use, and menopausal status (see Background). Age, sex and field center were included in genetic models of the five-domain endophenotypes. The proportion of the phenotypic variation attributable to covariates was estimated by comparing the estimated variance in the model that includes all significant covariates to that excluding all significant covariates. For association analysis, the hematologic traits and five-domain endophenotypes were adjusted for principal components of genetic ancestry to account for population structure along with other covariates.

1.6.4 Association Analysis Studies

Long Life Family Study: A Genomewide Association study (GWAS) tests for the association of variant sites across the genome with the trait of interest without an *a priori* hypothesis, that is, they are hypothesis-generating studies. As described previously, the hematologic traits were adjusted for significant (p-value ≤ 0.1) measured covariates including sex, age, smoking status, BMI, menopause, alcohol use, field center and principal components of genetic ancestry. The five-domain endophenotypes were adjusted for sex, field center and age (if significant), and ancestry. A p-value ≤ 0.1 was chosen to ensure that we accounted for measured covariates that might influence the trait. Association between genotyped SNPs and covariate adjusted blood

traits and endophenotypes were tested (including spousal controls) using a linear mixed-effect model correcting for family structure. The kinship matrix was built with "lmekin" and "kinship" R functions[66]. Results were reported as negative logarithm of the *p*-value. SNPs were filtered from the analysis if they had a call rate < 98%, a minor allele frequency < 1% and a Hardy–Weinberg equilibrium *p*-value < 10^{-6}. As the GWA analysis involves millions of non-independent tests, a *p*-value $\leq 5 \times 10^{-8}$ was considered significant at the genomewide level and a *p*-value $\leq 5 \times 10^{-6}$ was considered suggestive for association. I also tested for genomic inflation using quantile–quantile (Q–Q) plots and calculated the genomic control inflation factor.

1.6.5 Linkage Analysis

Multipoint linkage analyses: Multipoint linkage analyses were done using an extension of the variance component method described previously that includes the effect of a presumed QTL (σ^2_{QTL}) as a component of genetic variance[65]. As implemented in SOLAR, the QTL effect was estimated based on the expected covariance of relatives due to their IBD at an arbitrary chromosomal location in tight linkage with the presumed QTL. Significance of the σ^2_{QTL} was assessed by the likelihood ratio test of a model that includes the QTL versus a model without the QTL, that is, the polygenic model. Results were reported as a LOD score (i.e. log10 of the likelihood ratio), that follows a 50:50 mixed distribution of a point mass at zero and 1 degree of freedom chi-square distribution. Loki[61] and SOLAR[65] were used for the MIBD estimation and linkage analyses because the LLFS families are relatively large and complex. Other programs, such as MERLIN[67], would require breaking the larger pedigrees into smaller pedigrees, thus, reducing the power to detect linkage. LOD scores ≥ 2.5 were considered to be suggestive evidence for linkage with a QTL, whereas LOD scores ≥ 3.3 were considered to be significant

evidence for a QTL. After detecting significant (or suggestive) evidence for linkage, I identified a region of interest under the linkage peak. I defined the region of interest as the chromosomal region contained within 1.5 LOD units on either side of the maximum LOD score[68].

Two-point linkage analyses: Linkage analyses using MIBDs derived from multiallelic loci are generally more powerful than analyses of IBDs from single SNPs. However, if the SNP is in high LD with a causal locus, it should provide strong evidence of co-segregation. To fine-map potential QTLs, I performed two-point (that is, single SNP) linkage analyses for each SNP in the area of interest for a specific linkage peak.

Conditional linkage analyses: Another method by which to fine-map a QTL region is to perform a conditional linkage analysis. I performed conditional analyses by including the most significant SNP (or SNPs) as covariates in my linkage analysis models (along with other covariates) and assessed whether the LOD score for linkage was reduced. If the SNP is in high LD with the QTL, the LOD score should decrease.

1.6.6 Replication in HABC Cohort

For replication of genomewide association or linkage signals, I performed association analyses of a subset of SNPs on hematologic and five-domain endophenotype traits, after including effects of significant covariates and ancestry in the model. Because the HABC participants were unrelated, I performed these analyses using ProbABEL (ProbABEL v. 0.4.1)[69].

Selection of SNPs for replication: My protocol for selecting SNPs to be replicated in the HABC cohort differed from methods used by large GWA consortia. In other words, I did not select the SNP with the most significant *p-value* in a region and test for replication in another cohort because I was concerned that I might not detect a "true" association for multiple reasons.

First, the sample sizes available in large consortia studies enable the detection of relatively small marginal effects of alleles in LD with causal QTL. That is, associations can be detected, even if there are differences in LD patterns, or allele frequencies, or genotype by environment interactions among the different cohorts. However, the HABC cohort is not large (n =1,600). Second, the LLFS families and HABC participants were ascertained using different criteria and this might affect LD patterns within a region of interest[70]. Third, recent reports indicate that human populations harbor many more unique rare variants than were expected[71]. In specific populations (or families), different (uncommon) causal variants might reside on different haplotypes within the same locus, thus a common GWA SNP may mark one variant but not the other. Fourth, I wanted to maximize the probability of detecting a true association and minimize the number of tests.

Briefly, at each possible QTL location, all SNPs with *p-values* < 10^{-5} were considered. Next, the SNP with the lowest *p-value* and also present in HABC was chosen (referred to as the "lead" SNP) and all SNPs that were in high LD with the lead SNP, that is, $r^2 > 0.8$, were excluded. Among the SNPs that remained (that is, not in high LD with the first lead SNP), a second "lead" SNP, with lowest *p-value* and also present in HABC, was chosen. Then all SNPs in high LD with the second lead SNP were excluded. This process continued until all SNPs were excluded (or chosen to be replicated).

2.0 PHENOTYPIC AND GENETIC CHARACTERIZATION OF HEMATOLOGIC TRAITS

2.1 INTRODUCTION

Blood traits are inherently correlated due to their development from common hematopoietic stem cells and their coordinated role in the immune response system. Figure 2.1 shows relationship among the hematologic traits.

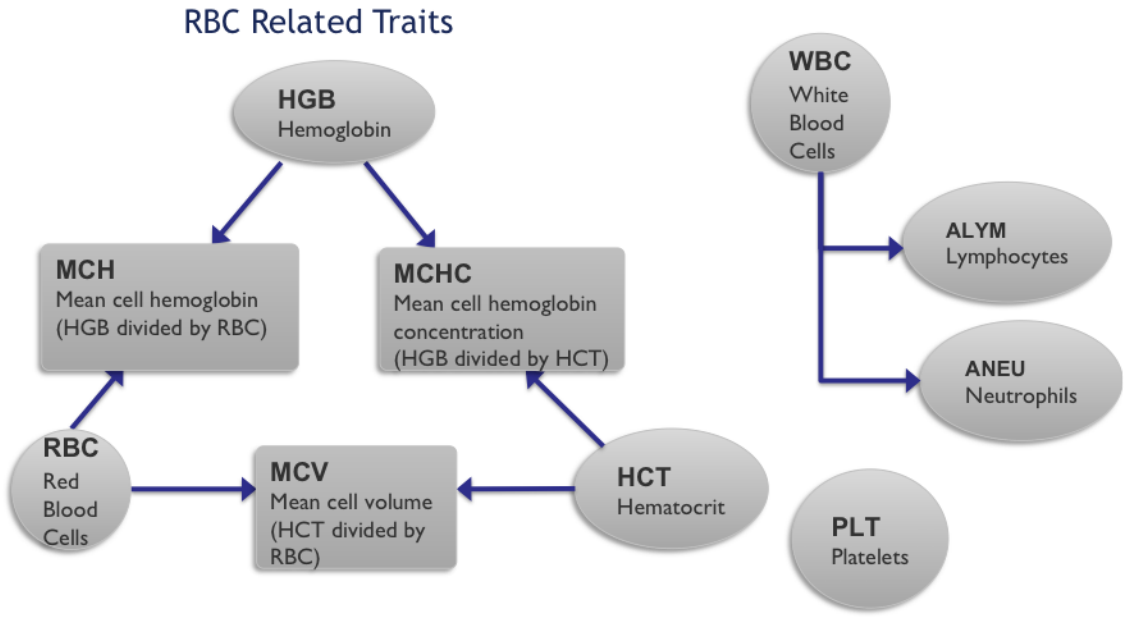

Figure 2.1: Hematologic Traits

As described in section 1.1, variations in the serum levels and counts of hematologic traits are hallmarks of age-related diseases, such as anemia. In addition, hematologic traits are known to be heritable (h^2 = 0.40 – 0.60) and numerous environmental factors have been correlated with serum levels and counts[28,29]. Many genetic variants have been associated with variation in the hematologic traits, but these variants account for only 1-6% of the phenotypic variation, indicating that many of the genes influencing hematologic traits have not yet been identified. Furthermore, many of these identified genetic variants are not known to be functional, nor have many genes with possible pleiotropic effects been identified. Finally, the relationship of hematologic traits to measures of health aging, such as the Healthy Aging Index (HAI)[22], is unknown.

This Chapter is a description of how I assessed the heritability and genetic correlations among the hematologic traits, and developed hematologic endophenotypes using data from participants in the Long Life Family Study (LLFS). As discussed in Chapter 1, analyses of endophenotypes may reveal genes with pleiotropic effects on the hematologic traits. In addition, I assessed the relationship of the hematologic traits with the Healthy Aging Index (Specific Aim 1). I also performed GWA and GWL analyses to detect QTLs that influence these traits (Specific Aim 2).

2.2 METHODS

2.2.1 Quality Control and Population Characteristics

LLFS is comprised of 4,535 individuals across 574 two-generation families: 1,515 in the older generation (mean age = 89.5 years), 2,255 in the offspring generation (mean age = 60.5 years) and 765 spousal controls (mean age = 60.8 years). These families were recruited from four sites, three in the US and one in Denmark, based on a measure of exceptional longevity[59]. A variety of demographic and phenotypic data were available. Family eligibility and ascertainment criteria have been described previously[60].

Prior to performing statistical and genetic analyses, I conducted a variety of quality control procedures; that is, I plotted the distributions of the traits and also compared the means and variances of the ten hematologic traits within and between sexes, generations, and study sites. In addition, because violation of the normality assumption can have an effect on type I error and power of the statistical methods that will be used in this study, the hematologic traits were assessed for normality and extreme outliers. Transformations were applied if required (and/or if commonly used in the literature) and outliers (values ± 4 standard deviation from the trait mean value) were removed.

2.2.2 Development of Hematologic Endophenotypes

To assess phenotypic correlation among the hematologic traits, I used hierarchical clustering and principal components analysis. As described in detail in section 1.6.1, because the participants in LLFS were not independent, I used an iterative, random sampling procedure to obtain unbiased

estimates of the correlation matrix. These correlations were used to develop dendrograms and heatmaps. In addition, these correlation matrices were used in the principal components analyses to develop hematologic endophenotypes.

2.2.3 Univariate and Bivariate Genetic Analyses

As mentioned in section 1.6.3, I estimated the heritability of each of the hematologic traits and the effects of covariates using the variance components framework as implemented in SOLAR[65]. Briefly, this method partitions the total phenotypic variance (σ^2_P) into additive genetic (σ^2_G), environmental (σ^2_E) and unmeasured error (σ^2_e) components. Covariates to be assessed in these analyses were selected based on the literature, especially GWA studies[37,39], for ease of comparison of my results to those from other studies. Effects of significant covariates were removed prior to the genomewide association and linkage analyses.

In addition to the univariate analyses, bivariate genetic analyses (described in section 1.6.1) were performed to quantify the underlying genetic and environmental relationships among the hematologic traits, as well as the relationship between the hematologic traits and the HAI. These bivariate analyses are an extension of the variance component framework described previously and is also implemented in SOLAR.

2.2.4 Genomewide Linkage Analyses

To detect QTLs influencing the hematologic traits or the endophenotypes, I performed multipoint linkage analyses as implemented in the program SOLAR[65]. This method is an extension of the variance component method described previously (section 1.6.5) that includes

the effect of a presumed QTL ($\sigma^2_{(QTL)}$) as a component of genetic variance. LOD scores ≥ 2.5 were considered to be suggestive evidence for linkage with a QTL, whereas LOD scores ≥ 3.3 were considered to be significant.

2.2.5 Genomewide Association Analyses

After adjusting for the effects of significant measured covariates, as well as the principal components for ancestry (see section 1.5.2.1), I performed GWA analyses on the hematologic traits and the hematologic endophenotypes. These analyses were performed using a linear mixed-effect model correcting for family structure. A detailed description of this method is in section 1.6.4. Results were reported as the negative logarithm of the *p*-value. SNPs were filtered from the analysis if they had a call rate < 98%, a minor allele frequency < 1% or a Hardy–Weinberg equilibrium *p*-value < 10^{-6}. For GWA analysis, a *p*-value ≤ 5×10^{-8} was considered significant at the genomewide level and a *p*-value ≤ 5×10^{-6} was considered suggestive for association.

2.3 RESULTS

2.3.1 Quality Control and Population Characteristics

Assessment of distributions of the hematologic traits by site, gender and generation revealed a few issues. For example, values for MCH (Mean Corpuscular Hemoglobin) differed between Denmark and US, where the values from Denmark were calculated to fewer digits (Figure 2.2). I was able to rescue this trait by recalculating MCH from serum HGB (Hemoglobin) concentration and RBC (Red Blood Cells) count. After assessment of all of the distributions, the following traits were transformed by natural logarithms to reduce non-normality: WBC (White Blood Cells), lymphocytes and neutrophil counts.

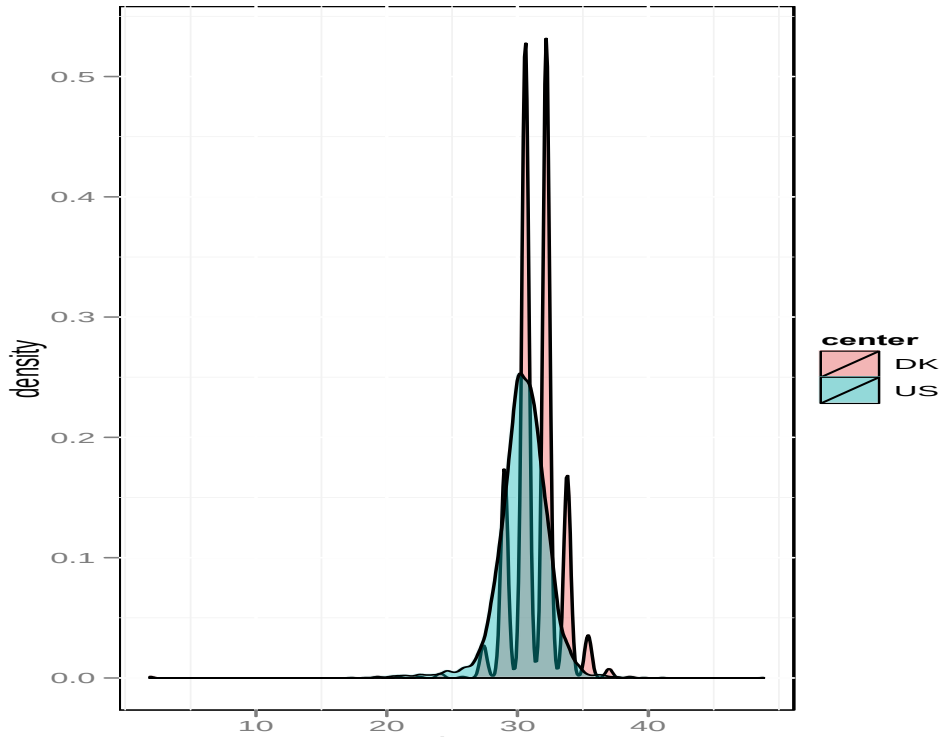

Figure 2.2: Problems with the MCH Data

Table 2.1 and Table 2.2 present the characteristics of the LLFS population by field centers and cohorts respectively, including the means (± SD), and sample sizes available for each of the hematologic traits, after data cleaning and removing outliers. The average age of the LLFS cohort was 70.2 years, 55% of the cohort were women, 7% were current smokers, and 36% consumed > 3 drinks/week. Almost 90% of the women were post-menopausal.

Table 2.1: Characteristics of the LLFS Cohort and Hematologic Traits by Field Center

Characteristics	**All**	**Pittsburgh**	**New York**	**Boston**	**Denmark**
	Mean ± SD or frequency (%)	Mean ± SD or frequency (%)	Mean ± SD or frequency (%)	Mean ± SD or frequency (%)	Mean ± SD or frequency (%)
N	4535	1202	935	1236	1162
Age (year)	70.25 ± 15.75	71.15 ± 15.91	74.43 ± 16.29	69.67 ± 15.94	66.56 ± 13.92
BMI (kg/m^2)	27.10 ± 4.96	27.72 ± 5.26	26.55 ± 4.66	27.59 ± 5.41	26.38 ± 4.21
Current Smoking	7%	5%	4%	3%	14%
Alcohol consumption (> 3 drink/week)	36%	22%	27%	33%	60%
Sex (%female)	55%	56%	53%	56%	54%
Menopause (% women)	89%	87%	92%	90%	88%
HCT (%)	41.88 ± 4.11	42.09 ± 3.97	41.58 ± 4.37	42.69 ± 4.28	41.04 ± 3.63
HGB (g/dL)	13.88 ± 1.42	13.86 ± 1.40	13.59 ± 1.52	13.90 ± 1.43	14.12 ± 1.28
RBC (10^{12}/L)	4.54 ± 0.48	4.56 ± 0.48	4.50 ± 0.51	4.57 ± 0.51	4.53 ± 0.43
MCH (pg/cell)	30.68 ± 1.99	30.48 ± 1.69	30.27 ± 2.05	30.53 ± 2.07	31.37 ± 1.97
MCHC (g/dL)	33.03 ± 1.42	32.92 ± 1.15	32.65 ± 1.27	32.56 ± 1.54	34.35 ± 0.84
MCV (fL/cell)	92.57 ± 5.57	92.66 ± 4.96	92.79 ± 5.79	93.92 ± 6.29	90.85 ± 4.68
WBC (10^9/L)	6.26 ± 2.25	6.49 ± 2.05	6.45 ± 1.98	6.22 ± 2.08	5.92 ± 2.73
ALYM (10^9/L)	1.92 ± 1.56	1.87 ± 1.15	1.95 ± 1.39	1.89 ± 1.43	1.97 ± 2.10
ANEU (10^9/L)	3.58 ± 1.46	3.86 ± 1.46	3.76 ± 1.42	3.61 ± 1.37	3.11 ± 1.47
PLT (10^9/L)	235.21 ± 62.38	234.85 ± 62.47	230.49 ± 61.70	238.31 ± 61.94	236.09 ± 63.14

Table 2.2: Characteristics of the LLFS Cohort and Hematologic Traits by Cohort

Characteristics	**Probands**	**Offspring**	**Controls**
	Mean ± SD or frequency (%)	Mean ± SD or frequency (%)	Mean ± SD or frequency (%)
N	1515	2255	765
Age (year)	89.52 ± 6.62	60.49 ± 8.29	60.85 ± 8.70
BMI (kg/m^2)	26.13 ± 4.32	27.64 ± 5.41	27.38 ± 4.47
Current Smoking	2%	9%	9%
Alcohol consumption (> 3 drinks/week)	23%	40%	49%
Sex (% female)	55%	58%	46%
Menopause (% women)	100%	86%	76%
HCT (%)	40.37 ± 4.32	42.66 ± 3.76	42.57 ± 3.82
HGB (g/dL)	13.20 ± 1.45	14.20 ± 1.28	14.29 ± 1.25
RBC (10^{12}/L)	4.32 ± 0.50	4.65 ± 0.43	4.66 ± 0.43
MCH (pg/cell)	30.70 ± 2.15	30.61 ± 1.90	30.83 ± 1.90
MCHC (g/dL)	32.70 ± 1.40	33.16 ± 1.38	33.41 ± 1.45
MCV (fL/cell)	93.92 ± 6.07	91.99 ± 5.17	91.60 ± 5.18
WBC (10^9/L)	6.80 ± 2.91	5.99 ± 1.79	5.98 ± 1.71
ALYM (10^9/L)	1.92 ± 2.38	1.93 ± 0.97	1.90 ± 0.64
ANEU (10^9/L)	4.04 ± 1.54	3.36 ± 1.35	3.31 ± 1.38
PLT (10^9/L)	226.01 ± 64.88	240.49 ± 60.63	237.80 ± 60.44

2.3.2 Development of Endophenotypes

To assess the correlation among hematologic traits and for the development of endophenotypes, the following methods were used.

2.3.2.1 Hierarchical Clustering

To assess phenotypic correlations among blood traits, hierarchical clustering was done after adjusting for significant covariates, and dendrograms and heatmaps were generated. Analyses revealed that phenotypic correlations among hematologic traits were similar between the related family member group and spousal control group (Table B1; Appendix). Subsequently, the average correlation matrix, using the iterative process described in section 1.6.1, was used to develop principal components. The heatmap and hierarchical clustering (Figure 2.3) illustrate these phenotypic correlations. There were three clusters of highly correlated traits: (1) RBC (Red Blood Cells), HCT (Hematocrit) and HGB (Hemoglobin); (2) MCV (Mean Corpuscular Volume) and MCH (Mean Corpuscular Hemoglobin), and (3) WBC (White Blood Cells) and ANEU (Neutrophils). Neutrophils are the most abundant type of white blood cells (WBC); hence they show high correlation with WBC. MCV and MCH traits were also moderately correlated with RBC. Platelets were moderately correlated only with WBC.

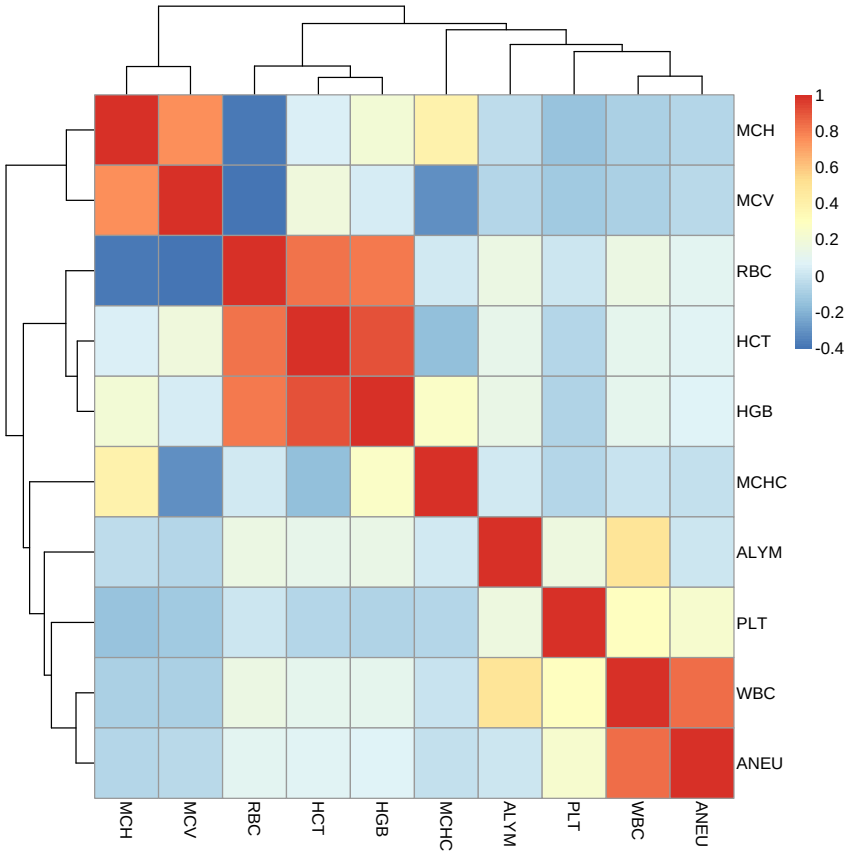

Figure 2.3: Phenotypic Correlations Among the Hematologic Traits in LLFS

2.3.2.2 Principal Component Analysis

Loadings for the first four components explain 81.6% of the phenotypic variation, and are shown in Table 2.3. The first composite phenotype (PC1) is strongly influenced by red blood cell related traits: HCT (Hematocrit), HGB (Hemoglobin) and RBC count, and accounts for 28.7% of the variability in the data. PC2 is comprised of a mix of red blood cell, white blood cell related traits and platelets, and accounts for 21.6% of variability. PC3 is comprised of MCH (Mean Corpuscular Hemoglobin), MCV (Mean Corpuscular Volume), WBC (White Blood Cells), and ANEU (Neutrophils) and accounts for 18% of the variation, whereas PC4 is defined by MCHC

(Mean Corpuscular Hemoglobin Concentration) and explains 13.3% of the variation. The principal components reflect the phenotypic correlations illustrated in the clustering analysis.

Table 2.3: Eigenvectors for the First Four Principal Components of Hematologic Endophenotypes

	PC1	**PC2**	**PC3**	**PC4**
Eigenvalue	4.25	2.91	2.07	1.85
% Variance explained	28.7	21.6	18.0	13.3
HCT	**-0.50**	**0.30**	0.03	0.24
HGB	**-0.50**	**0.33**	0.04	-0.12
RBC	**-0.55**	0.07	-0.27	0.02
MCH	0.12	**0.42**	**0.52**	-0.24
MCHC	-0.03	0.10	0.04	**-0.85**
MCV	0.14	**0.36**	**0.51**	**0.36**
WBC	-0.27	**-0.43**	**0.43**	-0.04
ALYM	-0.20	-0.18	0.20	-0.09
ANEU	-0.21	**-0.38**	**0.39**	0.01
PLT	-0.07	**-0.34**	0.14	0.03

Values in the bold indicate traits with the strongest contribution ≥ |0.3| to the PC.

2.3.3 Effects of Known Covariates and Heritability

Consistent with previous reports, women had higher levels of WBC count and PLT count and lower RBC count than males[72,73], and RBC count decreased with increasing age. Individuals with higher BMI had high RBC count (Table 2.4). As reported in the previous studies[74], smokers had higher WBC count than non-smokers. Drinkers (1-7 alcoholic drinks per week) had lower WBC count than non-drinkers and WBC count decreased further with increasing drinking. RBC count was also low in heavy drinkers (> 7 drinks per week). Estimates of σ^2_E ranged from 0.05 for ALYM (Lymphocytes) to 0.316 for HGB (Hemoglobin).

Table 2.4: Beta-Coefficients for Significant Covariates (p-value ≤ 0.10) for the Hematologic Traits

	ALYM	**ANEU**	**HCT**	**HGB**	**MCH**	**MCHC**	**MCV**	**PLT**	**RBC**	**WBC**
Sample Size (N)	4280	4287	4314	4314	4286	4449	4436	4301	4312	4308
Sex[a]		0.635	-2.575	-1.158	-0.324	-0.469	-0.197	31.177	-2.765	0.470
Age	-0.014	0.051	-0.099	-0.040	0.016	-0.016	0.093	-0.470	-0.155	0.037
Age × Age			-0.002	-0.001					-0.002	
Smoking[b]	1.106	1.411	1.309	0.430	0.709		2.400	11.951		1.827
BMI	0.051	0.077	0.096	0.031				-0.731	0.128	0.076
Menopause		-1.026		0.191		0.185				-0.668
Sex × Age			0.070	0.025	-0.008	0.005	-0.035		0.099	
Sex × BMI			-0.056	-0.022	-0.030			1.216	-0.049	
Sex × Smoke		0.603						-4.244	0.650	
BMI × Smoke				0.022						
Drinking1[c]	-0.316	-0.224	0.342	0.097	0.159		0.629	-4.295		-0.272
Drinking2[d]	-0.368	-0.472	0.375	0.150	0.876	0.116	2.431		-0.785	-0.418
NY	0.332		-0.802	-0.170	-0.159	0.119	-1.102			0.235
DK	0.462	-1.129	-2.171		0.510	1.726	-3.479		-0.602	-0.316
PT		0.266	-0.506			0.307	-1.100			0.182
σ2E[e]	0.049	0.137	0.268	0.316	0.112	0.277	0.112	0.090	0.264	0.125
Residual heritability ± SE	0.283 ± 0.035	0.259 ± 0.036	0.307 ± 0.033	0.268 ± 0.033	0.500 ± 0.035	0.645 ± 0.031	0.498 ± 0.033	0.421 ± 0.037	0.329 ± 0.034	0.317 ± 0.035

(a) effect of female sex with respect to male sex; (b) effect of smoking with respect to no smoking; (c) effect of 1-7 drinks per week with respect to no drinking; (d) effect of > 7 drinks per week with respect to no drinking; (e) proportion of variance due to covariates

After accounting for significant covariates, residual heritability estimates for the blood traits ranged from 0.259 to 0.645 and all were highly significant (Table 2.4). Heritability estimates for endophenotypes were 0.283 ± 0.033 (PC1), 0.381 ± 0.036 (PC2), 0.449 ± 0.036 (PC3), and 0.359 ± 0.031 (PC4). The heritabilities of the PCs were comparable to those of the blood traits. For example, the heritabilities of three variables, HCT, HGB, RBC count (the major components of PC1) have heritabilities equal to 0.307, 0.268 and 0.329 respectively; whereas PC1 has heritability equal to 0.283. Among all the blood traits and endophenotypes, PC4 has the highest heritability of 0.659, which is comparable to the heritability of MCHC, the main component for PC4 (see Table 2.3 and Table 2.4).

2.3.4 Genetic Correlations among Traits

I next estimated the genetic correlations between the hematologic traits, after adjusting for significant covariates, to assess whether the strong phenotypic correlations are due to pleiotropy (Table 2.5). Significant positive genetic correlations were observed between several traits: HCT and RBC count ($\rho_G = 0.775$) indicating that HCT and RBC share 60% ($\rho_G^2 = 0.775^2$) of the additive genetic variance and that the percentage of red blood cells in the serum by volume (HCT) and total red blood cell count are modulated by common genetic factors. Similarly significant positive correlations were also observed for HCT-RBC, HGB-RBC and MCH-MCV trait pairs. On the other hand, negative correlations were observed between red blood cell (RBC) numbers and size (MCV), and between RBC count and mean hemoglobin per RBC (MCH). For all the trait pairs, ρ_G was significantly different from one. Identification of QTLs influencing the genetically correlated traits may further reveal the genetic architecture of blood traits.

Table 2.5: Genetic Correlations Between Hematologic Traits

	HCT	HGB	RBC	MCH	MCHC	MCV	WBC	ALYM	ANEU	PLT
HCT	0.307	0.775	0.576		-0.363	0.312				
HGB		0.268	0.548	0.287	0.283					
RBC			0.329	-0.554		-0.526				
MCH				0.500	0.411	0.714				
MCHC					0.645	-0.313				
MCV						0.498				
WBC							0.317	0.650	0.893	0.216
ALYM								0.283	0.289	0.238
ANEU									0.259	
PLT										0.421

Only genetic correlations for which ρ_G is significantly different (p-value < 0.05) from zero are shown. Heritabilities for the hematologic traits are shown in the diagonal.

2.3.5 Genetic Correlation between Blood Traits and the Healthy Aging Index

To assess the relationship between hematologic traits and adverse health outcomes, I estimated the genetic correlation between hematologic traits and the Healthy Aging Index (HAI) (Table 2.6). HAI is a composite longevity phenotype, which includes measures of systolic blood pressure, pulmonary vital capacity, creatinine, fasting glucose and a modified mini-mental status examination score; and it has been shown to be a strong independent predictor of mortality in the Cardiovascular Health Study (CHS)[21,22].

Table 2.6: Genetic Correlations between HAI and Hematologic Traits

Trait	**h^2r**	**N**	**ρ_G**	**ρ_G SE**	***P*($\rho_G \neq 0$)**
HCT	0.307	3043	-0.190	0.106	0.075
HGB	0.268	3043	-0.275	0.116	**0.018**
MCH	0.500	3043	-0.241	0.091	**0.007**
MCHC	0.645	3043	-0.044	0.075	0.559
MCV	0.498	3043	-0.213	0.090	**0.017**
PLT	0.421	3043	-0.042	0.096	0.664
RBC	0.329	3043	-0.039	0.108	0.721
WBC	0.317	3043	0.219	0.102	**0.038**

Genetic correlations for which ρ_G is significantly different (p-value < 0.05) from zero are shown in bold.

Significant genetic correlations were observed between HAI and hematologic traits (HGB, MCH, MCV and WBC), indicating pleiotropy between these physiologic measures. Higher values of HAI are associated with increased mortality. Positive correlation of HAI with WBC count is consistent with the expectation that elevated WBC count is a hallmark of acute or

chronic systemic inflammation. Similarly, negative correlations of RBC indices (HGB, MCH, MCV) with HAI are consistent with the expectation that lower values of hemoglobin (anemia) have been shown to be associated with increased mortality. For all the trait pairs, ρ_G was found to be significantly different from 1.

2.3.6 Genomewide Linkage Results

A summary of suggestive and significant evidence for linked QTLs for ten hematologic traits and four composite endophenotypes are presented in Table 2.7; and I describe a few results below. These results are based on MIBD matrices derived from multi-locus haplotypes. Results from the other MIBD matrices derived using different SNP sets (PittA, PittB, and stLouis) were consistent with the results in Table 2.7 (Table B2; Appendix). I detected evidence of a significant QTL on chromosome 11p15.1 influencing RBC count, with LOD scores = 3.4. Another significant QTL influencing MCHC mapped to 10p12.3 with LOD scores of 3.7. I also identified several regions with suggestive evidence of linkage. Interestingly, a region on 2p13.3 was linked to PC4 with a LOD score of 3.2; this region was not detected by any single trait. Genomewide plots of all ten hematologic traits and four composite endophenotypes can be found in the Appendix (Figure B1-B14).

Table 2.7: Univariate LOD Scores

Trait	Region	cM (Mb)	LOD Score
PC4	2p13.3	92 (70.7)	3.2
HCT	3p25.3	27 (9.6)	2.7
ANEU	8p21.3	39 (21.0)	2.6
WBC	8q12.1	72 (58.1)	2.8
PLT	8p22	33 (17.8)	2.9
MCHC	10p12.3	45 (21.4)	3.7
PC4	10p12.1	53 (29.1)	2.5
RBC	11p15.1	38 (20.3)	3.4
RBC	11p15.2	26 (12.7)	2.5
PC1	11p15.2	27 (13.5)	2.5
RBC	11q24.1	134 (122.7)	3.0
PC1	17q12	61 (32.7)	2.5

2.3.7 Genomewide Association Analysis of Hematologic Traits

After performing genomewide association analyses on all of the hematologic traits and four composite endophenotypes, I calculated the genomic inflation factor (λ). Except for MCH (Mean Corpuscular Hemoglobin), the values of λ ranged from 1.00 – 1.07, indicating that the GWA results were not inflated by confounding factors, such as unrecognized population substructure. However, inflation was high for MCH ($\lambda = 1.1$), mainly due to deviation in the upper tail. Even after removing the SNPs with significant association, inflation remained somewhat high: $\lambda = 1.09$. Inspection of the GWA literature revealed that this result (higher inflation factor and a relatively high number of highly significantly associated SNPs) for MCH loci is typical, especially as compared with other blood traits[37,40]. The quantile – quantile (Q-Q) plots of the MCH and WBC are presented in Figures 2.4 and 2.5 (the other Q-Q plots are presented in the

Appendix; Figure B15 to Figure B28). Manhattan plots (-$\log_{10}$ transformed *p-values* against the physical positions) for hematologic traits and endophenotypes are presented in the Appendix (Figures B29 to B42).

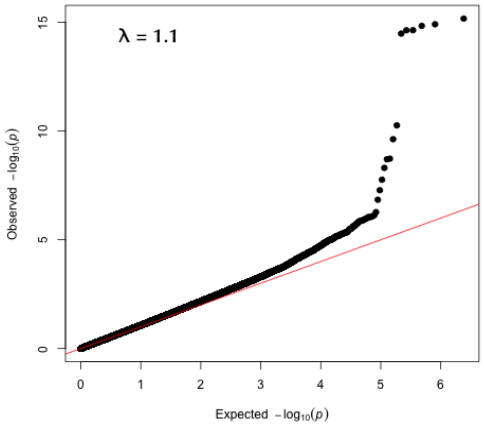

Figure 2.4: Q-Q Plot MCH

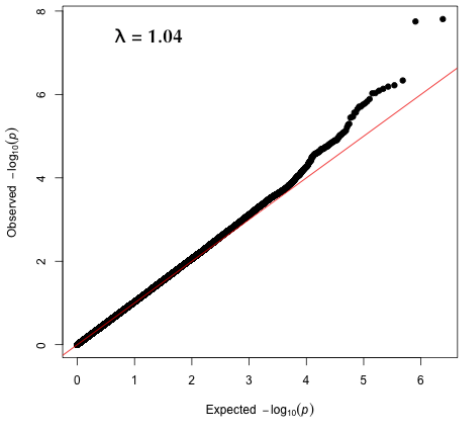

Figure 2.5: Q-Q Plot WBC

In total, I identified 32 SNPs belonging to five regions that were significantly associated (p-value < 5×10^{-8}) with four hematologic traits (MCH, MCV, RBC and WBC count) and two endophenotypes, PC2 and PC3. Table 2.8 lists the most significantly associated SNPs for the 9 significant trait-locus combinations. The chromosomal locations (and nearby genes) for these QTLs were 6p22.2 (*HFE*), 6p21.1, (*CCND3*), 6q23.3 (*HBS1L*), 17q21.1 (*PSMD3*), and 22q12.3 (*TMPRSS6*). These five loci are known to influence hematologic traits and have been reported by multiple studies[37,39,40,48]. As can be seen, results for GWA analyses for the composite endophenotypes were similar to the individual traits, with significant associations of PC2 and PC3 with the *TMPRSS6* and *HBS1L-MYB* region respectively.

Table 2.8: Most Significant Hematologic Traits by SNP Combinations Obtained from GWA Analyses

SNP	Region	Trait	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs79220007	6p22.2	MCH	26098474	C/T	0.049	*HFE*	2511	0.50	0.08	3.08×10^{-9}
rs3218086	6p21.1	MCV	41910064	A/G	0.166	*CCND3*	intron-variant	0.84	0.15	2.01×10^{-8}
rs9376090	6q23.3	RBC	135411228	C/T	0.244	*HBS1L*	33201	-0.61	0.11	9.51×10^{-9}
rs9376090	6q23.3	PC3	135411228	C/T	0.244	*HBS1L*	33201	0.23	0.04	3.28×10^{-10}
rs6920211	6q23.3	MCV	135431318	C/T	0.230	*HBS1L*	53291	0.88	0.13	3.45×10^{-11}
rs9494145	6q23.3	MCH	135432552	C/T	0.212	*HBS1L*	54525	0.37	0.05	6.73×10^{-16}
rs4065321	17q21.1	WBC	38143548	C/T	0.445	*PSMD3*	intron-variant	0.28	0.05	1.56×10^{-8}
rs855791	22q12.3	MCH	37462936	T/C	0.446	*TMPRSS6*	missense	-0.24	0.04	2.37×10^{-10}
rs855791	22q12.3	PC2	37462936	T/C	0.446	*TMPRSS6*	missense	0.18	0.03	4.05×10^{-8}

I also obtained evidence of suggestive associations (*p*-value < 5×10^{-6}) for 300 SNPs from 91 regions (Table B3; Appendix) with one or more hematologic traits or endophenotypes. Of these 91 QTLs, 35 have previously been reported to be associated with one or more hematologic traits. In addition, 12 of the 91 QTLs were associated with both composite endophenotypes and one or more hematologic traits, showing overlap between endophenotypes and hematologic traits. Additionally, 24 of these 91 QTLs were only associated with endophenotypes (and none of the hematologic traits). Of these 24 loci, 4 have previously been reported to be associated with hematologic traits in other populations.

Table B4 (Appendix) presents the top associated SNPs for the 121 trait-locus combinations that reached the suggestive threshold of *p*-value < 5×10^{-6}. A window of length ± 60 kb surrounding each of these 121 SNP-trait pairs identified 158 genes, of which 50 genes have previously been reported to be associated with hematologic traits (that is, within ± 60 kb of known GWA hits). Among the 108 novel genes with suggestive associations, few genes *(*e.g., *STAT3, DACH1* etc.) are known from functional studies to play a role in hematopoiesis, however

to my knowelge, variants in or near these genes have not yet been reported to be associated with hematologic traits[75,76].

2.4 DISCUSSION

Residual heritability of the hematologic traits ranged from 0.26 to 0.65 and these results are similar to those reported in other studies. The effects of measured covariates accounted for 5 – 32% of the phenotypic variation within each trait (Table 2.4).

The hierarchical clustering and bivariate genetic analyses revealed three clusters of highly phenotypically and genetically correlated traits: (1) RBC count, HCT, and HGB concentration, (2) MCV and MCH, and (3) ANEU and WBC count (Figure 2.3 and Table 2.5). The first derived endophenotype (PC1) reflects the phenotypic (and genetic) relationships among the blood traits, that is, PC1 is primarily comprised of RBC count, HCT (Hematocrit) and HGB (Hemoglobin) concentrations. However, the remaining PCs comprise multiple RBC and WBC-related traits, as well as platelets. Thus, genetic analyses of the hematologic traits and the endophenotypes may reveal differing sets of genes that influence these traits.

In general, linkage analyses may reveal uncommon variants segregating within families. Linkage analyses of the hematologic traits and endophenotypes revealed suggestive evidence for 12 QTLs. The highest LOD score (3.7) was for a QTL on chromosome 10p12 that influenced MCHC. The next highest LOD score (3.4) was for RBC count on chromosome 11p15.1. These results indicate that novel QTLs for hematologic traits may be segregating in the LLFS and I have begun to follow up on some of these signals; the linkage results for RBC count are discussed in more detail in Chapter 3.

Genomewide association studies are another method by which to identify potential QTLs that influence hematologic traits and endophenotypes. I performed GWA analyses and obtained significant evidence for five QTLs and these regions have been associated with hematologic traits in previous studies[37,38]. I replicated the genomewide significant association of WBC with the *PSMD3-CSF3* region and of RBC-related traits with *HBS1L-MYB, HFE, TMPRSS6* and *CCND3* (Table 2.8). Genetic variants in iron homeostasis genes *HFE* and *TMPRSS6* have been reported to be associated with red blood cell related traits (HGB, MCH, MCV, HCT)[37,39,43]. TMPRSS6 is a type II plasma membrane serine protease and plays an important role in iron hemostasis[44]. The nonsynonymous mutations (C282Y and H63D) in the *HFE* (High Iron Fe) gene are used routinely to confirm the clinical diagnosis of hereditary hemochromatosis[46]. The *HBS1L-MYB* intergenic region on chromosome 6q23 has been identified by multiple studies as a key regulator of blood traits. SNPs within this region have been shown to be associated with MCV, RBC, MCH, MCHC, HCT[37,38,39], WBC[40] and PLT[41], although the functional variants at most QTLs have not yet been identified. Finally, the *PSMD3-CSF3* region on chromosome 17q21.1 has been significantly associated with WBC count and neutrophil count in the European and Japanese populations, respectively[37,52]. A priori, CSF3 (Colony Stimulating Factor 3) was the most likely candidate gene in this region because it encodes a cytokine that regulates granulocyte production. However, Okada *et al.*[52] reported that the most significant SNP in the region is also associated with expression levels of proteasome 26S subunits non-ATPase 3 (*PSMD3*) and not with *CSF3*. Further studies are required to identify the functional genetic variants within this region that regulate the counts of WBC and its subtypes. Lastly, common variants in a member of the cyclin-D family, *CCND3* (6p21.1), that regulate CDKs (cyclin-dependent kinases) have been reported to be associated with MCH, MCV and RBC[37,39,40].

In addition to the five significant QTLs, I identified 91 QTLs that achieved the suggestive threshold for significance of association with hematologic traits or endophenotypes. These 91 QTLs encompass 158 genes within ± 60kb of the most significant SNP in the QTL region of interest. Of these 158 genes, 50 have been previously reported[38,39,41,47,48,54,77], but 108 have not. At first glance, a few of these 108 novel genes may also influence hematologic traits. For example, as a result of functional studies, *STAT3* and *DACH1* have been shown to play a role in hematopoiesis. However, to my knowledge, there are no reports that variants in these genes are associated with hematologic traits[75,76]. Of these 91 identified QTLs showing suggestive association with hematologic traits and endophenotypes in LLFS, 35 have previously been reported to be associated with one or more hematologic traits, demonstrating that the LLFS population has sufficient power to detect QTLs that influence hematologic traits. Furthermore, analyses of hematologic endophenotypes in addition to the individual traits increased my ability to detect QTLs. Analyses of haplotypes and/or sequencing regions of interest within the LLFS may reveal functional variants. Of potentially greater interest, however, are the 108 novel genes that may influence hematologic traits; and as stated above, a few of them are known to be involved in hematopoiesis. One of my next steps will be to try to replicate the association with these novel genes using other populations, such as the HABC cohort. Results of these replication studies may reveal additional genes and perhaps novel biological pathways that influence hematologic traits.

3.0 GENOMEWIDE LINKAGE STUDY OF RED BLOOD CELLS IN LLFS

3.1 INTRODUCTION

Hemoglobin disorders, such as sickle cell anemia and β-thalassemia, are among the most commonly inherited monogenic disorders in the world, especially in tropical regions of the world[2]. According to one estimate, a minimum of 332,000 children are born each year with a hemoglobin disorder[4]. Anemia is also common in the elderly population. Anemia and hemoglobin concentrations have been shown to be associated with adverse outcomes such as disability, hospitalization, morbidity and mortality in older adults[5,9,10,11].

Results from linkage studies have identified significant evidence for quantitative trait loci (QTL) influencing RBC (Red Blood Cell) count on chromosomes 19p13, 12p13, 11p15.2, and 18p11.32[31,34]. In addition, GWA studies, including large consortia, such as CHARGE (Cohorts for Heart and Aging Research in Genetic Epidemiology) and HaemGen have identified many loci associated with RBC and RBC related traits such as MCV (Mean Corpuscular Volume), MCH (Mean Corpuscular Hemoglobin), and MCHC (Mean Corpuscular Hemoglobin Concentration). HaemGen[37], the first large consortium on hematological parameters, analyzed data on 13,943 individuals and have identified a total of 6 loci (22q12.3, 6p21.1, 6p21.3, 22q12, 6q23 and 7q22) that influence variation in red blood cell traits (RBC, MCV and MCH). The HaemGen consortium is comprised of six European population based studies with average age in

these studies ranging from 41.4 to 61.2 years. The next large consortium, CHARGE[39], analyzed data on 24,167 individuals of European ancestry for 6 red blood cell traits (RBC, HCT, MCH, MCHC, MCV and HGB) and identified 23 loci associated with at least one of the red blood cell traits, 17 of which were novel. The largest of the meta-analysis for red blood cell related traits was reported by *van der Harst et al.,* 2012[48], which included 135,367 individuals of European and South Asian ancestry. In total, they reported 75 loci with evidence of association, 43 of which were novel. However, similar to results of meta-analyses of many other phenotypes and diseases[49], the identified variants explain little of the observed inter-individual variation in RBC count. For example, the CHARGE consortium reported that the most significantly associated SNPs with RBC count, at the two loci, i.e., *HBS1L-MYB* and *EPO*, explained 0.85% of the variation in RBC count[39]. Furthermore, the majority of the SNPs associated with the hematologic traits are not known to be functional variants.

One major drawback for all of these studies is that they are limited to individuals of European ancestry. Also as is the case with meta-analysis, heterogeneity among different cohorts can give rise to false positives/negatives.

To enhance our understanding of the possible genetic mechanisms involved in the regulation of red blood cells, data from the Long Life Family Study (LLFS), a large, two-generation family-based cohort study designed to elucidate the genes and environmental factors that influence exceptional aging, were analyzed. Because these families were selected based on exceptional aging, this study provides the unique opportunity to potentially identify novel loci involved in the regulation of blood traits that may influence exceptional survival.

In chapter 2, I described the association analyses of hematologic traits. In this chapter, I will present the results of linkage analysis for RBC. Linkage analyses for other blood traits were also preformed but are not presented in this thesis due to time constraints.

3.2 SUBJECTS AND METHODS

3.2.1 Study Subjects

The Long Life Family Study comprises a total of 4,535 individuals in 574 two-generation families. These families were recruited by four sites, three in the US and one in Denmark, based on a measure of exceptional longevity[59]. A variety of demographic and phenotypic data were available. After applying quality control measures, over 2.2 million assayed genetic markers and 18.1 million imputed SNPs (Based on 1000 Genomes) were available for analysis. Details of this study have been described in section 1.4.1. For the current study, phenotypic and genotypic data were available on 4,529 individuals in all 574 families.

Genotypic and phenotypic data from Health, Aging and Body Composition Study (HABC) was used to replicate our findings from the LLFS population. Details of HABC are given in section 1.4.2. For the current study, genotypic and phenotypic data were available on 1,297 unrelated European Americans between the ages of 71 and 82. A total of 2.5 million genotyped and imputed SNPs were available for analysis, as well as two principle components for admixture. The individuals were drawn equally from two sites, 45% from Pittsburgh, Pennsylvania, and 55% from Memphis, Tennessee.

3.2.2 Phenotypes

Characteristics of the LLFS and HABC cohorts are presented in Table 3.1. In general, the proband generation in LLFS is older than HABC. The proband generation had the lowest smoking rates (2%) and the highest alcohol use rates (66%).

Table 3.1: LLFS and HABC Characteristics

	LLFS			**HABC**
Characteristics	Proband (N = 1511)	Offspring (N = 2253)	Control (N = 765)	HABC (N =1297)
	mean or N (SD/%)	mean or N (SD/%)	mean or N (SD/%)	mean or N (SD/%)
Age	89.51 (6.62)	60.48 (8.29)	60.86 (8.70)	75.71 (2.81)
Sex	M - 685 (45%) F - 825 (55%)	M - 952 (42%) F - 1301 (58%)	M - 411 (54%) F - 354 (46%)	M - 682 (53%) F - 615 (47%)
Current smoking	29 (2%)	208 (9%)	68 (9%)	68 (5%)
Alcohol (> 1drink/week)	997 (66%)	928 (41%)	231 (30%)	486 (38%)
Menopause	826 (100%)	1114 (86%)	269 (76%)	615 (100%)
Field Center	BU - 401 (26%) DK - 211 (14%) NY - 452 (30%) PT - 447 (30%)	BU - 621 (28%) DK - 580 (26%) NY - 406 (18%) PT - 646 (29%)	BU - 212 (28%) DK - 367 (48%) NY - 79 (10%) PT - 107 (14%)	Mem - 718 (55%) PT - 579 (45%)
BMI (kg/m^2)	26.13 (4.33)	27.65 (5.41)	27.37 (4.46)	26.58 (4.19)
RBC ($10^{12}/L$)	4.32 (0.49)	4.65 (0.42)	4.66 (0.43)	4.50 (0.43)

BU = Boston; NY = New York; PT = Pittsburgh; Mem = Memphis; DK = Denmark

3.2.3 Statistical analysis of LLFS data

Before performing the genetic analyses, the RBC count was assessed for non-normality and outliers; a total of six values ± 4 standard deviations from the trait mean value were removed. Based on the literature, the following covariates were included in our model of RBC count: field

center, age, sex, age-squared, BMI, smoking status, alcohol use, and menopausal status[26,27], as well as principle components for admixture.

Linkage analysis: After adjusting for the above covariates, genomewide multipoint linkage analysis (GWL) was performed using an extension of the variance component method described previously (section 1.6.5) that includes the effect of a presumed QTL ($\sigma^2_{(QTL)}$) as a component of genetic variance, as implemented in SOLAR[65]. A logarithm of the odds (LOD) score ≥ 2.5 (or ≥ 3.3) was considered to be genomewide statistically suggestive (or significant) evidence for the presence of a QTL.

Fine-mapping: To identify candidate genes under the linkage peaks, two approaches were used: i) two-point variance components linkage analysis with each SNP in the region of interest and ii) family-based association analysis using imputed and genotyped SNPs in the region of interest. Although two-point (or single SNP) linkage analyses are generally not as powerful as multipoint analyses, two-point linkage analysis can be informative if the tested SNP is in strong LD with the causal variant. Likewise, if a SNP is in high LD with a causal variant, association analyses using assayed or imputed SNPs should be informative. The region of interest is usually defined as the region contained within one-LOD units of the highest LOD-score. These results were visually assessed in combination with the location of recombination hot-spots, known genes, and regulation sites[78].

Replication in HABC: The same covariates, as well as two principal components (PCs) for population substructure, were used to adjust RBC count in HABC. SNPs selected from LLFS were tested in HABC. Association analyses were done using ProbABEL (ProbABEL v. 0.4.1)[69] software.

3.3 RESULTS

RBC counts in the LLFS and HABC cohorts were not transformed prior to analyses. The covariates accounted for 26.4% and 12.7% of the variation in LLFS and HABC, respectively. Consistent with previous reports, women had lower RBC counts than males and RBC count also decreased with age (Table 3.2). RBC count was also low in heavy drinkers (> 7 drinks per week). The residual heritability of RBC count was 33% in LLFS, which was lower than what was reported by the Framingham Heart Study (h^2r: 56%)[31].

Table 3.2: Relationship between RBC and Covariates

		LLFS	**HABC**
		***β*-coefficient (*p*-value)**	***β*-coefficient (*p*-value)**
Sex[a]		-2.77 (1.19×10^{-20})	-1.45 (0.03)
Age		-0.16 (8.73×10^{-126})	
Age^2		-0.002 (7.97×10^{-8})	
Smoking[b]			0.47 (0.08)
BMI		0.13 (3.85×10^{-8})	0.02 (2.95×10^{-6})
Menopause			
Sex × Age		0.10 (1.38×10^{-29})	0.02 (0.03)
Sex × BMI		-0.05 (0.06)	
Sex × Smoke		0.65 (0.06)	
BMI × Smoke			-0.02 (0.10)
Drinking	D1[c]		
	D2[d]	-0.79 (1.16×10^{-4})	
Field Center	NY		-0.15 (1.45×10^{-8})
	DK	-0.60 (0.001)	
	PT		
Principal Component (admixture) for HABC		NA	0.98 (0.07)
Proportion of variance explained by covariates (%)		26.4	12.7

Only covariates with significant effect at $\alpha = 0.1$ are included in the genetic model.

(a) effect of female sex with respect to male sex; (b) effect of smoking with respect to no smoking; (c) effect of 1-7 drinks per week with respect to no drinking; (d) effect of > 7 drinks per week with respect to no drinking

3.3.1 Linkage Analysis

Genomewide linkage analyses were performed using MIBD matrices calculated using haplotypes as described in section 1.4.3. As can be seen in Figure 3.1, linkage analyses detected significant evidence for a QTL influencing RBC count on chromosome 11p15.1 (LOD = 3.4) as well as two suggestive signals on 11p15.2 (LOD = 2.5) and 11q24 (LOD = 3.0).

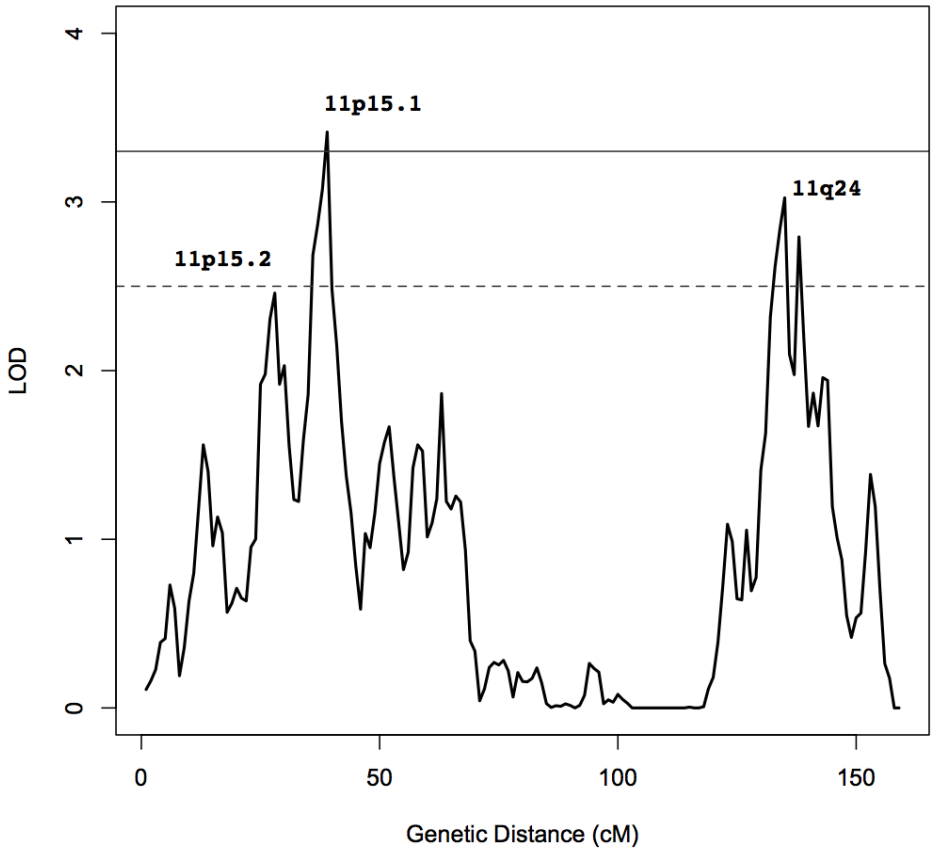

Figure 3.1: Chromosome 11 Univariate Linkage Results for RBC Count

Unlike signals from GWA analyses, a linkage peak usually spans large regions of genome and multiple genes can be located beneath the peak. I identified the one-LOD score

confidence interval under each peak as the region of interest within which to perform fine-mapping studies.

3.3.2 Fine Mapping of QTL at 11p15.2

The one-LOD unit confidence interval for the peak on 11p15.2 ranged from 23 to 30 centimorgans (cM). I next performed family-based association analysis and two-point linkage analysis using data on each SNP in this region.

Association analyses: A total of 16,348 SNPs (both genotyped and imputed) with MAF > 0.01 were assessed using family-based association analysis. I identified two SNPs with *p*-values < 10^{-4} (Figure 3.2): rs12421307 (*p*-value = 1.48×10^{-5}) and rs12419484 (*p*-value = 3.89×10^{-5}).

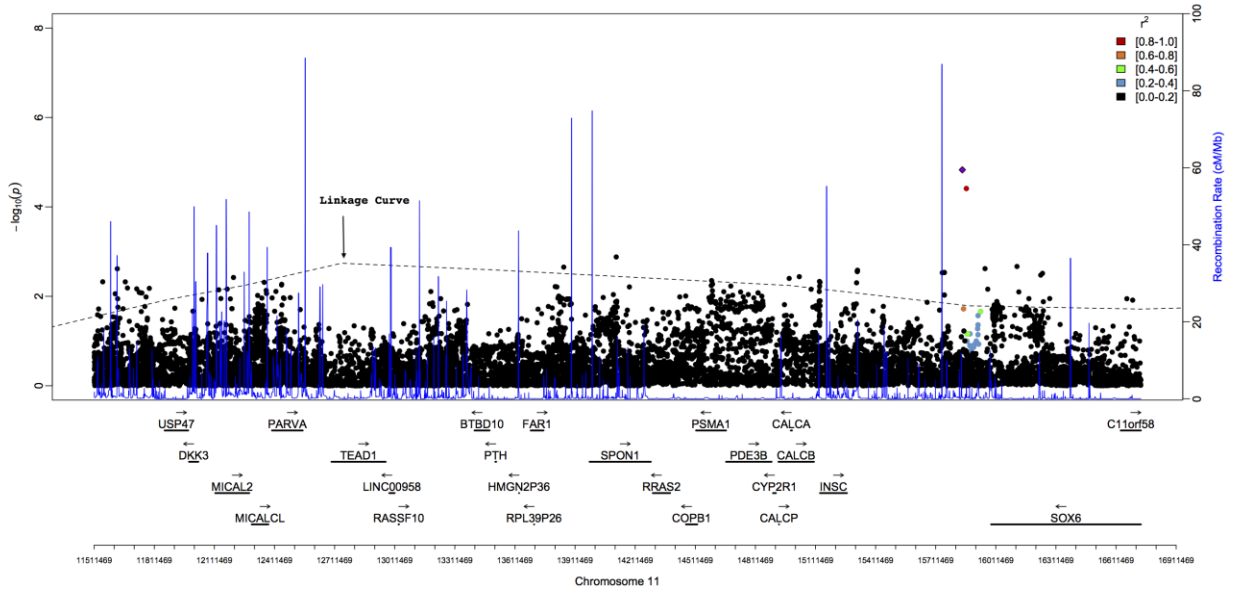

Figure 3.2: Association Analysis for Peak at 11p15.2 for RBC Count

Both SNPs are located to the right of a strong recombination peak (at 15.76 Mb), in a region that contains the gene *SOX6*. rs12421307 is 143 kb and rs12419484 is 123 kb downstream of *SOX6.* None of the SNPs were significant after adjusting for multiple testing (that is, Bonferroni *p*-value < 3.06×10^{-6}), however, this correction is conservative because it assumes independent tests and the SNPs that I tested are highly correlated.

Two-point linkage analyses: I also performed two-point linkage analysis under the linkage peak. A maximum LOD score of 3.27 was obtained for rs1484419. This SNP is 242 kb downstream of *SOX6.* Table 3.3 presents the results of two-point linkage analysis for SNPs with LOD > 2.5. rs1484419 is not in strong LD ($r^2 > 0.8$) with any of these SNPs except rs1385165 (Figure B43; Appendix). All of these SNPs are contained in a region that is 64 kb to 242 kb downstream of *SOX6.*

Table 3.3: Results of Association Analysis (*p*-value < 10^{-4}) and Two-Point Linkage Analysis (LOD > 2.5) for Peak at 11p15.2 for RBC Count

SNP	Position	minor/major allele	MAF	Two-point LOD Score	Association *p*-value
rs16931878	15744950	C/T	0.103	2.98	0.11
rs1484419	15745640	C/T	0.168	3.27	0.13
rs1011824	15746997	T/C	0.095	2.64	0.09
rs11023692	15748289	A/G	0.096	2.65	0.11
rs11023696	15758914	G/A	0.103	2.73	0.25
rs11023698	15760125	T/C	0.096	2.52	0.12
rs1385165	15763070	G/A	0.165	2.61	0.19
rs1385164	15779599	A/G	0.197	2.56	0.21
rs11023714	15785074	T/C	0.198	2.60	0.24
rs12421307	15844310	T/C	0.012	-	1.48×10^{-5}
rs12419484	15864406	T/C	0.012	-	3.89×10^{-5}
rs12576777	15877891	C/T	0.362	2.56	0.27
rs882148	15880959	T/C	0.339	3.23	0.45
rs1866821	15886799	A/C	0.310	3.02	0.33
rs722317	15923562	T/C	0.460	2.67	0.80

In LLFS, the results of the single SNP association and linkage analysis across the region of interest indicate that a QTL for RBC count may be located downstream of *SOX6*.

Replication of SOX6 downstream region in HABC: Based on the linkage and association analyses, I identified a 246 kb (15741469-15987995) region downstream of *SOX6*. The lower limit of the region was defined by the high recombination peak at 15.74 Mb on chromosome 11. Fifteen SNPs in LLFS that had *-logp* value for association greater than 2 or two-point LOD scores equal to or greater than 2.5, were selected for replication using HABC data. Of these 15 SNPs, 4 SNPs gave clear evidence of replication with *p*-value less than the stringent Bonferroni corrected *p*-value of 0.0033 (Table 3.4).

Table 3.4: Replication of *SOX6* Downstream SNPs in HABC for RBC Count

SNP	Position	minor/ major allele	MAF	LLFS			HABC	
				Beta	LOD	*p*-value	Beta	*p*-value
rs12365492	15744252	C/G	0.152	-0.392	-	2.97×10^{-3}	-0.079	1.20×10^{-3}
rs16931878	15744950	C/T	0.103	-	2.98	1.05×10^{-4}	0.073	7.72×10^{-3}
rs1484419	15745640	C/T	0.168	-	3.28	5.15×10^{-5}	0.065	5.18×10^{-3}
rs1011824	15746997	T/C	0.095	-	2.64	2.44×10^{-4}	0.082	2.67×10^{-3}
rs11023692	15748289	A/G	0.096	-	2.65	2.41×10^{-4}	0.082	2.70×10^{-3}
rs12361668	15754676	A/G	0.151	-0.340	-	9.36×10^{-3}	-0.073	7.49×10^{-4}
rs11023696	15758914	G/A	0.103	-	2.73	1.94×10^{-4}	0.071	6.87×10^{-3}
rs11023698	15760125	T/C	0.096	-	2.52	3.33×10^{-4}	0.078	5.07×10^{-3}
rs1385165	15763070	G/A	0.165	-	2.61	2.63×10^{-4}	0.058	9.97×10^{-3}
rs1385164	15779599	A/G	0.197	-	2.56	2.96×10^{-4}	0.035	0.08
rs11023714	15785074	T/C	0.198	-	2.60	2.73×10^{-4}	0.034	0.09
rs12576777	15877891	C/T	0.362	-	2.56	2.95×10^{-4}	0.014	0.41
rs882148	15880959	T/C	0.339	-	3.23	5.69×10^{-5}	0.009	0.59
rs1866821	15886799	A/C	0.310	-	3.02	9.56×10^{-5}	0.001	0.94
rs722317	15923562	T/C	0.460	-	2.67	2.28×10^{-4}	-0.013	0.43

3.3.3 Fine Mapping of QTL at 11p15.1

The one-LOD unit confidence interval for the peak on 11p15.1 ranged from 35 to 40 centimorgans (cM). This region contains two very large genes (*NAV2*; 770 kb, *NELL1*; 906 kb) and four other genes (*DBX1, HTATIP2, PRMT3* and *SLC6A5*). None of these genes have any a priori evidence to be involved in the regulation of erythropoiesis. To fine-map this region, I performed family-based association analysis, two-point linkage analysis and SNP conditional analysis using data on each SNP in the region.

Association Analyses: A total of 6,046 SNPs (both genotyped and imputed with MAF > 0.01) were assessed using family-based association analysis. Figure B44 (Appendix) presents the results of association analysis. Table 3.5 presents the results of association analysis for SNPs with *p*-value < 3.2×10^{-3} (-*logp* > 2.5). The strongest association was observed for an indel c11_21127449 (*p*-value = 2.76×10^{-4}), an intronic variant in *NELL1*, although the majority of the variants listed in Table 3.5 are intronic or downstream of *NAV2*.

Table 3.5: Results of Association Analysis (*p*-value < 3.2×10^{-3}) for Peak at 11p15.1 for RBC Count

SNP	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs139479126	19610526	T/C	0.011	*NAV2*	intron-variant	1.97	0.66	3.02×10^{-3}
c11_19850210 INDEL	19850210	D/R	0.217	*NAV2*	intron-variant	-0.36	0.12	2.22×10^{-3}
rs4447157	19988487	G/T	0.116	*NAV2*	intron-variant	-0.44	0.14	2.32×10^{-3}
rs7933978	19988666	C/G	0.410	*NAV2*	intron-variant	-0.29	0.09	2.11×10^{-3}
rs4757873	19992506	T/A	0.286	*NAV2*	intron-variant	-0.30	0.10	2.84×10^{-3}
rs4757894	20144722	G/A	0.178	*NAV2*	1087	-0.40	0.12	1.16×10^{-3}
rs12288745	20148723	C/G	0.367	*NAV2*	5088	0.30	0.10	2.27×10^{-3}
rs75897432	20290238	C/G	0.032	*HTATIP2*	-95085	0.84	0.28	2.89×10^{-3}
rs185091958	21014487	C/T	0.023	*NELL1*	intron-variant	1.49	0.49	2.13×10^{-3}
c11_21127449 INDEL	21127449	R/D	0.010	*NELL1*	intron-variant	2.88	0.79	2.76×10^{-4}

Two-point linkage analyses: Table 3.6 presents the results of two-point linkage analysis for SNPs with LOD > 2.5. All six SNPs with LOD > 2.5 fall within an 83 kb intronic region of *NELL1*. The highest two-point LOD score of 3.47 (*p*-value = 3.23×10^{-5}) was obtained for rs1401790. This SNP is in strong LD ($r^2 > 0.8$) with rs1611930, rs12295675 and rs9630161 (Figure B45; Appendix).

Table 3.6: Two-Point Linkage Analysis (LOD > 2.5) for Peak at 11p15.1 for RBC Count

SNP	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Two-Point LOD Score	*p*-value
rs1611930	20737138	A/G	0.207	*NELL1*	intron-variant	3.28	5.14×10^{-5}
rs1401790	20737599	C/T	0.200	*NELL1*	intron-variant	3.47	3.23×10^{-5}
rs12295675	20747171	T/C	0.163	*NELL1*	intron-variant	2.88	1.36×10^{-4}
rs9630161	20749657	A/G	0.162	*NELL1*	intron-variant	2.91	1.25×10^{-4}
rs12284819	20756016	A/G	0.102	*NELL1*	intron-variant	2.67	2.27×10^{-4}
rs79559057	20820962	C/T	0.091	*NELL1*	intron-variant	2.56	2.96×10^{-4}

Conditional linkage analyses: For conditional linkage analysis, 92 SNPs from association analysis (-*logp* > 2) and 14 SNPs from two-point linkage analysis (LOD score > 2) were tested. Results of the SNP conditional analysis for SNPs that decreased the LOD score at 11p15.1 by greater than 0.30 are presented in Table 3.7. Most of the SNPs are intronic variants in *NAV2* and two SNPs are in the intergenic region between *DBX1* and *HTATIP2*.

Fine-mapping under the linkage peak at 11p15.1 provided evidence for *NAV2, NELL1, HTATIP2* and *DBX1*. Multiple QTLs in the regions may be responsible for the linkage peak.

Table 3.7: SNP Conditional Analysis for Peak at 11p15.1 for RBC Count

SNP	Position	minor/major allele	MAF	Nearby Gene	Position Near Gene	Decrease in LOD Score
rs7933978	19988666	C/G	0.409	*NAV2*	intron-variant	0.31
c11_19989667_INDEL	19989667	R/D	0.293	*NAV2*	Intron-variant	0.33
rs10741806	19989730	A/T	0.292	*NAV2*	intron-variant	0.33
rs4757872	19990087	G/A	0.294	*NAV2*	intron-variant	0.33
rs10833208	19990885	G/A	0.293	*NAV2*	intron-variant	0.34
rs10833209	19990890	G/A	0.293	*NAV2*	intron-variant	0.35
rs4757873	19992506	T/A	0.286	*NAV2*	intron-variant	0.37
rs4757913	20264424	G/A	0.256	*DBX1*	82683	0.31
rs7119037	20266391	G/C	0.242	*DBX1*	84650	0.35

Replication in HABC: For replication in HABC, 82 SNPs that had -*logp* values greater than 2 for association or two-point LOD scores greater than 2, were selected. In the replication, none of the SNPs reached the Bonferroni corrected *p*-value of 6×10^{-4}. Only one SNP, rs4757922, showed nominal evidence of replication (*p*-value < 0.05). This SNP is 68 kb upstream of *HTATIP2 (DBX1 - HTATIP2* intergenic*).*

3.3.4 Fine Mapping of QTL at 11q24

The one-LOD unit confidence interval for the peak on 11q24 ranged from 131 to 138 centimorgans (cM). I next performed family-based association analysis and two-point linkage analysis using data on each SNP in this region.

Association analyses: To fine-map the chromosome 11q24 peak at 134 cM, a total of 11,898 SNPs (both genotyped and imputed) with MAF > 0.01 were assessed using family-based association analysis. Table 3.8 presents the results of association analysis for SNPs with *p*-values $< 3.2 \times 10^{-3}$.

Table 3.8: Results of Association Analysis (p-value < 3.2×10^{-3}) for Peak at 11q24 for RBC Count

SNP	Position	minor major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	p-value
rs61905533	121847033	G/C	0.016	*LOC100507165*	-52114	-1.43	0.46	2.07×10^{-3}
rs873590	122670390	T/G	0.216	*UBASH3B*	intron-variant	-0.35	0.11	1.75×10^{-3}
c11_122717117 INDEL	122717117	I/R	0.490	*CRTAM*	0	-0.30	0.10	2.11×10^{-3}
rs7130937	123053321	A/G	0.025	*CLMP*	intron-variant	0.89	0.29	2.49×10^{-3}
rs949064	123061707	C/G	0.025	*CLMP*	intron-variant	0.87	0.29	2.69×10^{-3}
rs112875189	123413666	A/G	0.134	*GRAMD1B*	intron-variant	-0.51	0.15	8.33×10^{-4}
rs17455332	123437978	C/T	0.136	*GRAMD1B*	intron-variant	-0.45	0.14	1.53×10^{-3}
rs117186816	123936128	G/A	0.053	*OR10G7*	26433	-0.70	0.23	2.22×10^{-3}
c11_123969180 INDEL	123969180	D/R	0.024	*OR10G7*	59485	0.91	0.31	2.98×10^{-3}
rs11606663	123973977	T/G	0.069	*VWA5A*	-12170	-0.61	0.19	1.58×10^{-3}
rs117036580	124366328	T/A	0.013	*OR8B12*	-46345	-1.23	0.39	1.71×10^{-3}

The most strongly associated SNP in the region was rs112875189 (p-value = 8.33×10^{-4}), an intronic variant in *GRAMD1B.*

Two-point linkage analysis: Table 3.9 presents the results of two-point linkage analysis for SNPs with LOD > 2.5. The highest two-point LOD score of 4.89 (p-value = 1.04×10^{-6}) was obtained for rs6590081 (5 kb upstream of *OR8B12*) followed by rs61904445 (LOD = 4.02; p-value = 8.35×10^{-6}), which is located 8 kb upstream of *CLMP*.

Table 3.9: Two-Point Linkage Analysis (LOD > 2.5) for Peak at 11q24 for RBC Count

SNP	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	LOD	*p*-value
rs682011	121544285	A/G	0.464	*SORL1*	39351	2.55	3.04×10^{-4}
rs10502260	121567897	T/C	0.302	*SORL1*	62963	3.43	3.54×10^{-5}
rs7935547	121740317	A/C	0.200	*LOC100507165*	-158830	2.99	1.04×10^{-4}
rs7938656	121740615	C/A	0.200	*LOC100507165*	-158532	2.86	1.41×10^{-4}
rs61904445	123074371	T/A	0.149	*CLMP*	8456	4.02	8.35×10^{-6}
rs4245047	123094797	T/C	0.199	*CLMP*	28882	2.95	1.14×10^{-4}
rs559254	123096933	A/G	0.208	*CLMP*	31018	3.54	2.72×10^{-5}
rs3018105	123097158	A/G	0.208	*CLMP*	31243	3.02	9.72×10^{-5}
rs525854	123098651	A/G	0.198	*CLMP*	32736	2.96	1.11×10^{-4}
rs2846054	123305966	G/T	0.488	*LOC100128242*	intron-variant	2.70	2.12×10^{-4}
rs1275085	123513161	T/C	0.095	*SCN3B*	synonymous	3.25	5.49×10^{-5}
rs1720343	123514384	C/T	0.102	*SCN3B*	intron-variant	2.73	1.94×10^{-4}
rs1720340	123518917	G/A	0.092	*SCN3B*	intron-variant	2.98	1.06×10^{-4}
rs1720339	123519255	G/A	0.093	*SCN3B*	intron-variant	2.99	1.04×10^{-4}
rs1453631	123925974	A/G	0.323	*OR10G7*	16279	3.00	1.01×10^{-4}
rs6590081	124419307	C/A	0.411	*OR8B12*	5780	4.89	1.04×10^{-6}
rs10893265	124433714	A/C	0.096	*OR8A1*	-6272	2.71	2.04×10^{-4}
rs10893269	124442947	T/C	0.096	*OR8A1*	2017	2.70	2.10×10^{-4}

Conditional linkage analyses: For conditional linkage analysis, 62 SNPs from the association analysis (-*logp* > 2) and 43 SNPs from the two-point linkage analysis (LOD score > 2) were tested. Results of the SNP conditional analysis for SNPs that decreased the LOD score at 11q24 by greater than 0.25 are presented in Table 3.10.

Table 3.10: SNP Conditional Analysis for Peak at 11q24 for RBC Count

SNP	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Decrease in LOD score
rs140632358	122706042	A/G	0.018	*CRTAM*	-3239	0.27
rs112875189	123413666	A/G	0.134	*GRAMD1B*	intron-variant	0.27
rs73025504	124252963	T/A	0.013	*OR8B2*	missense	0.35
rs11604549	124255717	C/G	0.042	*OR8B2*	2509	0.29

The largest decrease was observed for rs73025504. This SNP is in the coding region of *OR8B2* and results in a missense mutation (serine to threonine). Another SNP, rs11604549, which is 2 kb upstream of *OR8B2*, also decreased the LOD score by 0.29. None of these 4 SNPs presented in Table 3.10 are in LD with each other ($r^2 < 0.2$). When these 4 SNPs were simultaneously included in the linkage model, the multipoint LOD score decreased by 1.13. Replication of SNPs from this region in HABC was not attempted in part due to low MAF.

3.4 DISCUSSION

Hematologic traits are routinely used in clinical practice because abnormal readings are markers for a variety of diseases such as anemia, polycythemia, etc., as well as adverse outcomes in older adults. Identification of genes that influence hematologic traits will increase our understanding of the risk factors that affect diseases such as anemia, and may lead to better interventions and treatments. I performed genomewide linkage analyses using data from the LLFS and obtained suggestive and significant evidence for QTLs influencing RBC counts in the region of chromosome 11p15.

I obtained the most promising result by following up the suggestive evidence of linkage of a QTL located at chromosome 11p15.2 with RBC count. This result was particularly interesting because investigators from the Framingham Heart Study previously reported linkage of a QTL for RBC counts at 11p15.2[34]. Results from fine-mapping this region under the linkage peak (using family-based association analysis and two-point linkage analysis) in the LLFS was most consistent with a QTL located in the region downstream of the *SOX6* gene. From this region, I selected 15 SNPs that displayed the strongest relationship with RBC count, and then replicated this relationship using data from the HABC cohort; 4 out of 15 SNPs were significantly associated with RBC count (Table 3.4).

SOX6 is a member of the Sox transcription factor gene family. This gene family is defined by high mobility group (HMG) domain. *SOX6* acts as a regulator of gene expression and previous studies have shown *SOX6* to be involved in development of the central nervous system[79], muscle[80] and cartilage[81]. *SOX6* has also been shown to be essential for efficient erythropoiesis in both fetus and adult erythropoietic tissue in mice, and it plays an important role in the development and maturation of erythroid cells[82,83]. It mediates this effect by directly binding to the promoter of the *Bcl2l1* gene, which codes for Bcl-xL, an anti-apoptotic factor, and enhancing its expression[83]. *Bcl2l1* has previously been shown to be involved in erythroid cell survival[84]. *Sox6*$^{-/-}$ mice fetuses show anemia associated with defective maturation and decreased survival of RBC[82]. Apart from acting as an enhancer of erythropoiesis, *SOX6* is also involved in the developmental transition from fetal to adult hemoglobin. Xu *et al.* have shown that during erythroid maturation, SOX6 interacts physically and cooperates with BCL11A, a known repressor of gamma-globin gene expression, to regulate the expression of globin genes[85].

For the linkage peak at 11p15.1, six genes (*NAV2, DBX1, HTATIP2, PRMT3, SLC6A5 and NELL1)* lay within the region of interest under the linkage peak. Association analysis and two-point linkage analysis of SNPs under this peak provided weak evidence that variants in or near *NAV2, DBX1, HTATIP2* and *NELL1* may influence RBC counts. However, selected SNPs in these regions were not replicated at a Bonferroni level of significance. *NAV2* (neutron navigator) is a member of the neuron navigator gene family and may play a role in cellular growth and migration. *NELL1* (NEL-Like 1 (chicken)) encodes a protein that contains epidermal growth factor (EGF) like repeats and may play a role in cell growth and differentiation.

For the linkage peak at 11q24, association analysis, two-point linkage analysis and SNP conditional analysis indicated that the QTL influencing this peak may be present in a region marked by olfactory receptor genes. A missense mutation in *OR8B2* (Olfactory Receptor, Family 8, Subfamily B, Member 2) decreased the LOD score by 0.35. Interestingly, Framingham Heart Study investigators also previously reported the association of olfactory receptors with MCH (Mean Corpuscular Hemoglobin) at 11q12.1[34]. Solovieff *et al.* have reported the association of a region, upstream of the β-globin gene cluster, containing olfactory receptor genes at 11p15.4 with fetal hemoglobin in sickle cell anemia[86]. Recently, Shim *et al.*, have shown that the smell perception in Drosophila is involved in maintenance of the hematopoietic system[87].

Conclusions: My genetic studies in a novel population, LLFS, may have identified additional genes that influence RBC count, a marker of health, especially among older adults. However, these studies are just the beginning; I have not identified any causal genetic variants that are associated with prediction of healthy aging. In fact, the "nearby genes" that I have identified may not be involved. Nevertheless, additional studies might include association studies performed in other elderly populations, investigations of gene expression in hematopoietic

precursor cells, and perhaps other bioinformatic studies. Furthermore, the most interesting nearby gene, *SOX6*, has also been implicated in the development of the central nervous system[79], muscle[80] and cartilage[81]. Unfortunately, LLFS does not have good measurements of the latter traits, thus, studies of these characteristics in additional populations should also be fruitful.

4.0 RELATIONSHIP OF LLFS ENDOPHENOTYPES TO MORTALITY AND REPLICATION IN THE HEALTH AGING AND BODY COMPOSITION COHORT

4.1 INTRODUCTION

One of the hypotheses in the field of aging is that aging is a fundamental biologic process that eventually leads to age-related diseases and disability, rather than a conglomeration of multiple diseases. Thus, identification of environmental factors or sets of genes that influence aging could lead to insights or interventions to better enable a long and healthy life for all individuals. Longevity and healthy aging, however, are complex traits and longevity is not readily amenable to many types of epidemiologic and genetic studies because long wait times are required. Furthermore, "functional longevity" or "disease-free" survival is the desired outcome, not "longevity," per se. Although disease-free survival is the trait of interest and could be measured at different ages, precisely defining and measuring it is more challenging. One common measure uses individual self-reports and/or administrative records of disease. However, Newman and colleagues (2008)[21] have shown that continuous measures of subclinical disease are more reflective of exceptional survival. They developed a composite longevity phenotype, Healthy Aging Index (HAI) and estimated HAI for participants in the Cardiovascular Health Study (CHS). HAI includes measures from systolic blood pressure, pulmonary vital capacity, creatinine, fasting glucose and modified mini-mental status examination score. They showed that

only 1.7% of individuals in their 70's could be considered disease-free and that this latter group of individuals had very low mortality rates = 7/1000 person years[21]. A variation of this index has subsequently been shown to be heritable (Sanders *et al.*, in press) and studies to detect and identify QTLs that influence this measure are ongoing (R. Minster *et al.*, manuscript in preparation). I have also previously shown that HAI is significantly, genetically correlated with several hematologic traits (section 2.3.5). However, the HAI is only one measure of healthy aging and many others can be derived.

My colleagues in the LLFS previously developed five heritable endophenotypes comprised of linear combinations (using factor analysis) of 28 traits across 5 health domains: cognition, cardiovascular, metabolic, physical activity, and pulmonary[23]. These endophenotypes may better characterize exceptional survival than any single trait and may facilitate identification of specific loci that influence exceptional longevity. Although these endophenotypes were heritable, their relationship to mortality was not known. As described in this chapter, I assessed whether any of these endophenotypes were correlated with measures of mortality in LLFS. Furthermore, I validated these endophenotypes and their relationships with mortality, using data from the Health Aging and Body Composition cohort (HABC).

4.2 MATERIALS AND METHODS

4.2.1 Development of Endophenotypes in LLFS

Using an earlier 'freeze' of the LLFS data, Matteini and colleagues (2010)[23] derived heritable endophenotypes, based on 28 trait values across 5 domains. These analyses included 480

families consisting of 3,224 participants, which were available at that time. I first extended this work by including all of the 4,472 LLFS participants belonging to 574 families in the analysis.

For endophenotype development, selection of 28 traits and factor analysis was done as described by Matteini and colleagues[23]. Outliers (± 4 SD away from the mean) were removed and the same transformations (natural logarithm transformation for triglycerides, pulse pressure, creatinine, systolic blood pressure, HDL, glucose, glycosylated hemoglobin, waist circumference and square-root transformation for average and maximum grip) were applied as described by Matteini and colleagues[23]. Endophenotype scores were calculated for each individual by multiplying the eigenvectors for each factor to their corresponding generation-adjusted standardized, transformed trait values. Heritability was estimated using the variance component framework (described in section 1.6.3), as implemented in SOLAR[65]. Covariates in the heritability analysis included age, gender and recruitment site.

4.2.2 Replication in HABC

For replication, data on 1,794 individuals for all the traits except those belonging to the cognition domain were available in HABC. In developing endophenotypes, outliers (± 4 SD away from the mean) were removed (between 1 and 26 values were removed). The same transformations were applied as those in LLFS. Factor analysis was done using the 'principal' function (psych package) in R with varimax rotation, initially allowing for five factors. For subsequent analyses, endophenotype scores for each factor were calculated by multiplying the eigenvectors for factors to their respective standardized trait values. For calculating the endophenotype score for the second factor, loadings from factor 2 (F2) and factor 3 (F3) were linearly combined, as the

predominant variables in F2 and F3 from HABC (based on the loadings) were similar to those in F2 in LLFS (see Results).

Because HABC lacked data on the cognition domain, the construction of endophenotypes as described above was not ideal. Therefore, I performed additional studies. First, I removed the cognitive domain variables from LLFS, re-derived eigenvectors and restricted the analyses to four factors only. Then I estimated heritability and performed mortality analyses on LLFS re-derived endophenotypes F1R and F2R, to determine if I obtained similar results to those described above. I also re-derived eigenvalues in HABC, restricting the outcomes to four endophenotypes (F1R, F2R, F3R, F4R). Thus, the "R" factors result from factor analyses performed without cognitive variables and constrained to four eigenvectors only. Results from the "R" factors for LLFS and HABC are presented in the Appendix.

4.2.3 Association of Endophenotypes with Mortality

As a preliminary assessment of a possible relationship between the endophenotypes and mortality, I estimated the correlation between the endophenotypes and age of death. I also categorized the endophenotypes into tertiles and then performed survival analyses using the Kaplan-Meier method; significance was determined by log rank test.

Because the results of my initial crude analyses were tantalizing, a more sophisticated analysis of the possible association of endophenotypes with mortality was tested using Cox proportional hazard analysis. My LLFS colleagues, Dr. Robert Boudreau and Tanushree Prasad performed these latter analyses.

4.3 RESULTS

4.3.1 Population Characteristics

Population characteristics for LLFS and HABC are summarized in Table 4.1. Average age in the LLFS cohort was 68 and in HABC it was 74 with a narrow range (69 - 80 years).

Table 4.1: Population Characteristics of Individuals with Endophenotype Data in LLFS and HABC

	LLFS (N = 4472)			**HABC (N = 1794)**		
	Mean or N	SD or Freq (%)	Outliers Removed	Mean or N	SD or Freq (%)	Outliers Removed
Age	68.67	14.90	NA	73.77	2.86	NA
Sex (Females)	2462	55%	NA	855	48%	NA
Field Centers	BU: 1146 DK: 1179 NY: 958 PT: 1189	BU: 26% DK: 26% NY: 21% PT: 27%	NA	MEM: 935 PT: 859	MEM: 52% PT: 48%	NA
Cognition						
Animal recall	20.18	6.39	0	Few and different measures of cognition		
Vegetable recall	13.92	4.66	0			
Digit forward	8.33	2.20	0			
Digit back	6.46	2.29	0			
Immediate memory	12.02	4.49	0			
Delayed memory	10.41	4.87	0			
Cardiovascular						
Presence of hypertension	2243	50%	NA	776	43%	NA
Systolic BP (mm Hg)	131.38	21.64	8	133.51	19.68	1
Diastolic BP (mm Hg)	77.51	11.23	3	69.93	11.01	0
Pulse pressure	53.69	17.34	18	63.41	16.62	5
Total cholesterol (mg/dL)	200.39	41.75	3	201.15	37.02	2
HDL cholesterol (mg/dL)	59.17	17.09	8	51.64	15.73	6
LDL cholesterol (mg/dL)	118.83	35.44	3	119.51	32.85	2
Triglyceride (mg/dL)	108.80	57.69	36	148.98	76.28	12
Metabolic						
Presence of diabetes	304	7%	NA	191	11%	NA
BMI (kg/m^2)	27.14	4.65	17	26.50	4.06	4
Creatinine (mg/dL)	1.03	0.24	29	1.01	0.23	5
Glucose (mg/dL)	94.02	15.65	44	98.94	21.38	26
Glycosylated hemoglobin (%)	5.59	0.46	43	6.09	0.74	14
Waist circumference (cm)	94.64	13.69	7	98.92	11.70	4
Physical Activity						
Average grip strength (kg)	28.99	11.66	6	29.32	9.80	0
Maximum grip strength (kg)	29.84	11.88	8	32.06	10.37	1
Gait speed (m/sec)	1.06	0.29	0	1.25	0.22	2
Total physical activity	10.27	2.64	0	10.35	1.37	3
Pulmonary						
FEV1/FEV6 (%)	76.99	6.84	15	76.55	7.32	3

Table 4.1 Continued

FEV1	2474.15	860.98	0	2284.44	654.11	0
FEV6	3201.03	1045.02	0	2984.02	800.15	0
Presence of lung disease	572	13%	NA	269	15%	NA

BU = Boston; NY = New York; PT = Pittsburgh; Mem = Memphis; DK = Denmark

4.3.2 Estimation of Endophenotypes and Heritability

Eigenvalues and eigenvectors for the first five factors for LLFS are shown in Table 4.2. The first factor (F1) is mainly dominated by the physical and pulmonary domain and explains 13.9 % of the variation. The second factor (F2) is dominated by measures from metabolic and cardiovascular domains and explains 10.7% of the variation. The third factor (F3) is comprised of measures solely from the cognition domain and explains 9.3% of the variation. The fourth factor (F4) is characterized mainly by blood pressure related traits (hypertension, systolic BP, diastolic BP and pulse pressure) and explains 9.1% of the variation. The fifth factor (F5) includes cardiovascular measures and explains 7.7% of the variation. Loadings of the variables and the variation explained by them were strikingly similar to those reported by Matteini and colleagues (2010)[23] using a smaller number of participants (Table B5; Appendix).

Genetic factors contribute a large proportion of variation in endophenotypes (Table 4.3). Estimates of heritability for the F1 and F2 ($h^2_{F1} = 0.51$; $h^2_{F2} = 0.39$) were higher than what was previously reported ($h^2_{F1} = 0.39$; $h^2_{F2} = 0.27$)[23]. Heritability estimates for the rest of the endophenotypes ($h^2_{F3} = 0.38$; $h^2_{F4} = 0.21$; $h^2_{F5} = 0.23$) were comparable to those previously reported. Age, sex, and recruitment center (coded as 3 dummy variables) explained a large proportion of variation in F1, which is dominated by the physical and pulmonary domains.

Table 4.2: Results of Factor Analyses for LLFS

	F1	**F2**	**F3**	**F4**	**F5**
Eigenvalue	4.01	3.48	2.54	2.31	1.86
% Variance explained	13.9	10.7	9.3	9.1	7.7
Cognition					
Animal recall	0.18	-0.12	0.53	-0.07	0.06
Vegetable recall	-0.13	-0.16	0.58	-0.05	0.08
Digit forward	0.03	0.05	0.42	-0.09	-0.18
Digit backward	0.04	0.06	0.51	-0.07	-0.10
Immediate memory	0.01	0.04	0.80	0.02	0.04
Delayed memory	0.01	0.01	0.80	0.01	0.06
Cardiovascular					
Presence of hypertension	-0.06	0.22	-0.09	0.69	-0.10
Systolic BP	-0.02	0.03	-0.08	0.96	0.09
Diastolic BP	0.22	-0.01	-0.04	0.65	0.16
Pulse pressure	-0.18	0.04	-0.07	0.79	0.00
Total cholesterol	-0.08	-0.13	-0.04	0.10	0.94
HDL cholesterol	-0.26	-0.61	0.09	0.06	0.12
LDL cholesterol	0.03	-0.03	-0.08	0.06	0.93
Triglycerides	0.02	0.56	-0.05	0.09	0.44
Metabolic					
Presence of diabetes	-0.14	0.45	0.03	-0.00	-0.17
Estimated BMI	0.02	0.74	0.04	0.12	0.12
Creatinine	0.31	0.27	-0.22	-0.06	-0.02
Glucose	-0.03	0.53	-0.04	0.09	-0.01
Glycosylated hemoglobin	-0.19	0.56	0.02	0.02	-0.01
Waist Circumference	0.19	0.77	-0.06	0.06	0.03
Physical Activity					
Average grip strength	0.88	0.16	-0.05	0.01	-0.08
Maximum grip strength	0.88	0.17	-0.06	0.01	-0.08
Gait speed	0.45	-0.25	0.28	0.02	0.08
Total physical activity	0.42	-0.21	0.27	0.07	0.14
Pulmonary					
Presence of lung disease	-0.14	0.05	-0.03	-0.03	-0.06
FEV1	0.86	0.03	0.04	-0.10	0.00
FEV6	0.88	0.00	0.02	-0.10	-0.03
FEV1/FEV6 ratio	0.07	0.10	0.09	-0.05	0.13

Table 4.3: Estimation of Covariate Effects and Heritability of the Five-Domain Endophenotypes in LLFS

	F1	F2	F3	F4	F5
Heritability					
N	4302	4302	4302	4302	4302
h^2r (S.E.)	0.51 (0.04)	0.39 (0.04)	0.38 (0.04)	0.21 (0.04)	0.23 (0.04)
p-value	5.19×10^{-51}	4.08×10^{-33}	6.48×10^{-37}	3.04×10^{-12}	2.38×10^{-14}
Covariates					
Age	-0.05	0.01	-0.03	0.03	-
Sex	-5.14	-2.16	0.98	0.27	0.95
NY	-	-0.43	-	-	-0.18
DK	0.67	-0.28	-1.01	1.26	1.17
PT	-	0.41	-	0.16	-0.17
Variance Explained by covariates	0.53	0.16	0.09	0.07	0.11

Beta-coefficients for covariates with $p < 0.1$ are shown
NY = New York; DK = Denmark; PT = Pittsburgh

4.3.3 Replication of Five-Domain Endophenotypes in HABC Cohort

Eigenvalues and eigenvectors for the first five endophenotypes for HABC are presented in Table 4.4. The first factor is made up of physical and pulmonary function measures, similar to LLFS results. Interestingly, as we didn't have a cognition domain in HABC, which solely defines the third factor in LLFS, the second factor in HABC is divided into two parts (F2 and F3). The second factor in LLFS is made up of the cardiovascular and metabolic domains. In HABC, F2 is made up of the metabolic domain and F3 is made up of measures mainly from the metabolic domain and also from the cardiovascular and physical domains. Results from F4 and F5 are very similar to corresponding factors in LLFS and are dominated by blood pressure and lipid-related measures, respectively.

I also performed factor analyses (and constrained to four factors only) using LLFS data without the cognitive variables, and on HABC data. The eigenvectors and eigenvalues are

presented in Tables B6 and B7 in the Appendix. Again, the eigenvectors for the four "non-cognition" endophenotypes are similar to those reported in Tables 4.2 and 4.4.

Table 4.4: Factor Analysis Results for the First Five Endophenotypes in HABC (no measures of the Cognition Domain are available for HABC)

	F1	**F2**	**F3**	**F4**	**F5**
Eigenvalue	4.25	2.91	2.07	1.85	1.49
% Variance explained	18.3	10.1	10.0	9.7	9.1
Cardiovascular					
Presence of hypertension	-0.02	0.14	0.22	0.49	-0.14
Systolic BP	-0.03	0.00	-0.01	0.97	0.07
Diastolic BP	0.19	-0.15	0.11	0.50	0.20
Pulse pressure	-0.16	0.10	-0.09	0.79	-0.04
Total cholesterol	-0.22	0.00	-0.02	0.04	0.94
HDL cholesterol	-0.46	-0.25	-0.38	-0.02	0.15
LDL cholesterol	-0.01	0.03	-0.01	-0.00	0.92
Triglycerides	-0.06	0.19	0.43	0.16	0.22
Metabolic					
Presence of diabetes	0.01	0.79	0.02	0.03	-0.08
Estimated BMI	0.11	0.11	0.83	0.05	0.02
Creatinine	0.50	0.09	0.19	0.07	-0.08
Glucose	0.14	0.82	0.24	0.08	-0.00
Glycosylated hemoglobin	0.01	0.82	0.12	-0.00	0.05
Waist Circumference	0.21	0.10	0.83	0.03	-0.04
Physical Activity					
Average grip strength	0.87	0.07	0.09	0.01	-0.14
Maximum grip strength	0.87	0.07	0.10	0.00	-0.15
Gait speed	0.44	-0.07	-0.35	0.00	0.05
Total physical activity	0.40	0.03	-0.39	0.00	0.04
Pulmonary					
Presence of lung disease	-0.18	0.13	0.04	-0.00	-0.14
FEV1	0.84	-0.08	0.03	-0.08	0.11
FEV6	0.87	-0.03	-0.02	-0.09	0.05
FEV1/FEV6 ratio	0.07	-0.15	0.16	-0.00	0.21

4.3.4 Relationship of the First Two Five-Domain Endophenotypes in LLFS with Mortality

I next assessed whether the two most dominant endophenotypes (F1 and F2) were associated with mortality. These two factors had the highest residual heritabilities and were predominantly comprised of traits across multiple health domains. Furthermore, using data from HABC, we obtained similar eigenvectors for the first two factors.

Preliminary analyses. I obtained a significant correlation between F1 and age at death (r^2 = 0.15, *p*-value = 0.0002), and a borderline significant correlation between F2 and age at death (r^2 = -0.08, *p*-value = 0.05). Results of the Kaplan-Meier survival analysis were consistent with the crude correlation results, that is, mortality differed significantly among the F1 tertiles (*p*-value = 0.001; Figure 4.1). No significant difference in mortality was detected for F2 tertiles (*p*-value = 0.8) (results not shown). Based on these results, my colleagues performed Cox proportional hazard analyses.

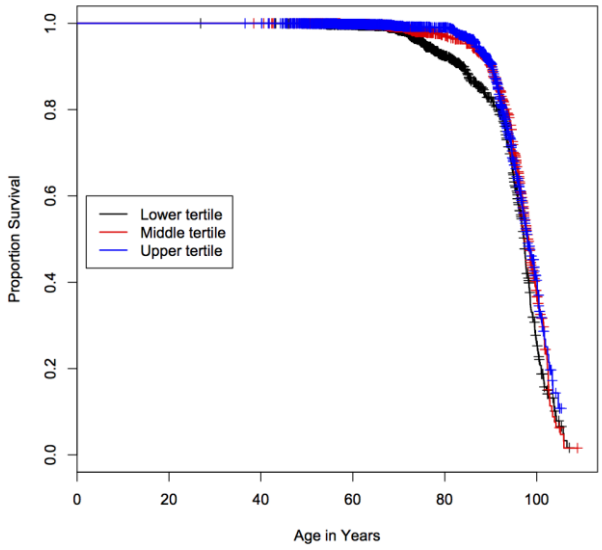

Figure 4.1: Survival Function Plot by F1 Tertiles

Hazard analyses in LLFS. In LLFS, 636 (14.38%) of the total 4,424 individuals were deceased after a mean follow-up time of 4.4 ± 1.2 years. In the offspring generation, 3.16% (99/3135) were deceased and 41.66% (537/752) individuals in the proband generation were deceased. We performed Cox regression analyses to assess whether F1, F2, or both improved prediction of mortality compared to baseline age and sex. We expected that baseline age and sex would be a strong predictor of mortality. The results of these analyses are presented in Table 4.5. As can be seen, baseline age plus sex were significant predictors of mortality; the 95% confidence intervals (CI) of the hazard ratios (HR) for both variables did not encompass 1.0. In addition, the area under the (prediction) curve (AUC) = 0.898, indicating that these two variables were able to discriminate between those who died versus those who did not die after 4.4 years of follow-up. The addition of the first endophenotype (F1) increased discrimination slightly (from 0.898 to 0.903). More interestingly, the effect of F1 was significant (the 95% CI of the hazard ratio did not encompass 1.0), and it attenuates (that is, explains) 3.36% of the effect of age. In contrast, F2 did not have a significant effect and did not contribute to the discriminatory ability of the model. These results indicate that F1 is a significant predictor of mortality.

Table 4.5: Results from Cox Regression Models Including Baseline Age and Gender

Model	HR (95% CI)	Attenuation (%) of Age HR	AIC	AUC
Baseline age + gender	Age: 1.119 (1.111, 1.128) F: 0.693 (0.592, 0.810)	Reference	8851.68	0.898
Baseline age + F1 + gender (F = female)	Age: 1.115 (1.106, 1.123) F1: 0.898 (0.872, 0.925) F: 0.424 (0.345, 0.521)	3.36 %	8803.73	0.903
Baseline age + F2 + gender (F = female)	Age: 1.119 (1.111, 1.128) F2: 1.030 (0.998, 1.063) F: 0.723 (0.614, 0.851)	0 %	8850.24	0.898
Baseline age + F1 + F2 + gender (F = female)	Age: 1.115 (1.106, 1.123) F1: 0.900 (0.873, 0.926) F2: 1.021 (0.990, 1.053) F: 0.442 (0.357, 0.547)	3.36 %	8803.96	0.903

Further inspection, however, revealed that there was a strong effect of cohort (offspring versus proband) on F1 (see Figure B46; Appendix) even after incorporating effects of sex. This difference would impact the Cox regression models. Therefore, we incorporated "offspring generation" in our models. As can be seen in Table 4.6, F1 is still significant (the 95% CI of the hazard ratio does not encompass 1.0) and the discrimination of the model including F1 is still slightly greater than that for baseline age plus gender alone. However, now F1 explains a larger proportion of the age effect (26%).

Table 4.6: Results from Cox Regression Models Including Baseline Age, Gender, and Generation

Model	HR (95% CI)	Attenuation (%) of Age HR	AIC	AUC
Baseline age + gender (F = female)	Age: 1.119 (1.111, 1.128) F: 0.693 (0.592, 0.810)	Reference	8851.68	0.898
Baseline age + generation + gender (F = female)	Age: 1.120 (1.105, 1.135) Gen: 1.021 (0.702, 1.485) F: 0.693 (0.592, 0.810)	- 0.84 %	8853.67	0.897
Baseline age + F1 + generation + gender (F = female)	Age: 1.087 (1.070, 1.103) F1: 0.870 (0.841, 0.900) Gen: 0.445 (0.292, 0.679) F: 0.368 (0.296, 0.458)	26.9 %	8791.38	0.902
Baseline age + F2 + generation + gender (F = female)	Age: 1.119 (1.104, 1.134) F2: 1.030 (0.998, 1.063) Gen: 0.988 (0.679, 1.438) F: 0.723 (0.614, 0.852)	0 %	8852.24	0.899
Baseline age + F1 + F2 + generation + gender (F = female)	Age: 1.086 (1.070, 1.103) F1: 0.871 (0.842, 0.900) F2: 1.026 (0.995, 1.058) Gen: 0.435 (0.285, 0.663) F: 0.385 (0.308, 0.481)	27.7 %	8790.79	0.902

Proportional hazard analyses in the HABC cohort: We next performed Cox regression analyses in the HABC cohort. In the HABC cohort, 936 (52.17%) of the total 1,794 individuals were deceased after a mean follow-up time of 10.9 ± 3.6 years. Results of the analyses including baseline age, sex, F1, and F2 are presented in Table 4.7. The discriminatory ability of this model is lower, most likely because of the longer follow-up time. However, the results of the analyses

in the HABC cohort are similar to the results obtained in LLFS: F1 is a significant predictor, independent of age, and the model including F1 increases discrimination versus the model without F1, AUC increases from 0.651 to 0.677, respectively. Furthermore, F1 attenuates the effect of age by 19.6%. In HABC, F2 is also a significant predictor of mortality.

Again, as described above, these analyses are not ideal, however, my colleagues and I did not have time to complete analyses of additional factors within the time frame of this dissertation. Some of the limitations of these analyses are presented in the discussion.

Table 4.7: Results from Cox Regression Models in HABC

Model	HR (95% CI)	Attenuation (%) of Age HR	AIC	AUC
Baseline age + gender (F = female)	Age: 1.112 (1.087, 1.137) F: 0.686 (0.602, 0.781)	Reference	12709.23	0.651
Baseline age + F1 + gender (F = female)	Age: 1.090 (1.065, 1.116) F1: 0.911 (0.885, 0.937) F: 0.384 (0.309, 0.478)	19.6 %	12670.29	0.677
Baseline age + F2 + gender (F = female)	Age: 1.113 (1.088, 1.138) F2: 1.037 (1.019, 1.056) F: 0.739 (0.645, 0.846)	- 0.9 % (↑)	12695.59	0.660
Baseline age + F1 + F2 + gender (F = female)	Age: 1.089 (1.064, 1.115) F1: 0.905 (0.880, 0.932) F2: 1.042 (1.024, 1.061) F: 0.405 (0.326, 0.504)	20.5 %	12651.88	0.687

4.4 DISCUSSION

The incidence and prevalence of all common complex diseases increase with increasing age. Therefore, identifying the underlying environmental and genetic causes of such an increase may enable society to develop interventions to increase functional longevity or healthy aging. In the current chapter, I have used factor analyses to extend the results of Matteini and colleagues (2010) to the complete LLFS cohort. The composition and eigenvectors of the first five

endophenotypes are strikingly similar to those reported previously. Furthermore, I was able to recover similar endophenotypes in another population comprised of (initially) healthy individuals, the HABC cohort. The composition and eigenvectors of the HABC endophenotypes were very similar to those obtained in LLFS. Because data on cognition variables were not available in HABC, I also re-derived four factors for both of the LLFS (after excluding the cognitive domain variables) and HABC cohorts. As presented in the Appendix (Tables B6 and B7), the eigenvectors for the remaining four domains (and traits) were still similar to the initial results in LLFS.

Although the specific traits from the five health domains were hypothesized to influence healthy aging, we did not know whether these endophenotypes were associated with mortality. We performed Cox regression analyses to assess whether the endophenotype factors were predictive of mortality. Our analyses revealed that F1 was a significant predictor of mortality, independent of age. Higher F1 values were associated with lower mortality (HR < 1.0). This result is consistent with our expectations because F1 is comprised of improved physical and pulmonary function; the loadings for the predominant components (e.g., grip strength, FEV1, etc.) were positive. After incorporating a covariate for generation, F1 explained 26.9% of the effect of age, in addition to its independent effect on mortality. Remarkably, analyses of F1 in the HABC cohort revealed similar effects: higher F1 was associated with lower mortality and F1 attenuated 19.6% of the effect of age. In HABC, F2 was also significantly associated with mortality. Higher values of F2 were associated with increased mortality. This result is also consistent with the expectations as the predominant components were increased frequency of diabetes, and increased glucose and glycosylated hemoglobin concentrations. These mortality

analyses need to be redone using the eigenvectors obtained from the four-factor analyses of traits after exclusion of the cognitive domain traits.

Finally, in addition to the significant, but independent effects of F1 (and F2 in HABC) on mortality, these endophenotypes were also moderately heritable; residual heritability ranged from 0.51 to 0.21. Thus, linkage and association analyses of these endophenotypes may reveal sets of genes that influence healthy aging. Furthermore, results of these analyses (both those reported in this chapter and in the appendix) indicate that the HABC cohort should be a reasonable cohort for replication of possible effects of quantitative trait loci identified in the LLFS.

5.0 ASSOCIATION ANALYSES OF ENDOPHENOTYPES OF LONG AND HEALTHY LIFE: THE LONG LIFE FAMILY STUDY

5.1 INTRODUCTION

Studies in animal models, such as the nematode *C. elegans*, have revealed several genes that have dramatic effects on longevity[88]. Additional genes with strong effects on longevity have been reported in yeast, flies, and mice[89,90,91]. Although longevity is known to be heritable in humans[92], identification of specific genes that influence longevity in humans, especially healthy aging and longevity, has been challenging[58]. Numerous linkage, candidate gene, and GWA studies have been performed, but except for *apolipoprotein E*, *FOXO3*, and a QTL on chromosome 3, the results of these studies have been inconsistent (reviewed by Brooks-Wilson; 2013)[58].

As described in Chapter 4, the LLFS five-domain endophenotypes were derived from phenotypes that were individually associated with morbidity and mortality. In addition, these endophenotypes were replicated in the HABC cohort and factor 1 (F1) was significantly associated with mortality, independent of age in both LLFS and HABC cohorts. Finally, all of the five-domain endophenotypes (F1 – F5) were heritable. In this chapter, I report the results of genomewide association analyses performed on F1 and F2. I did not analyze data on F3 because I did not have a replication population available. I also did not analyze F4 and F5 at this time.

5.2 MATERIALS AND METHODS

5.2.1 Endophenotypes in LLFS

Endophenotypes (and genotypes) from five health domains (F1 to F5) were available on 4,302 individuals. As described in section 4.3.2, the dominant factor (F1) in LLFS and HABC was predominantly comprised of traits from the pulmonary and physical function domains, whereas F2 in LLFS was comprised of measures from the cardiovascular and metabolic health domains. Factor 3 (F3) in LLFS consisted of traits from the cognitive domain. Factors 4 and 5 (F4 and F5) represented different components of the cardiovascular health domain, that is, blood pressure related traits and cholesterol-related traits, respectively.

5.2.2 Genotype Data in LLFS

As described in section 1.5.2, the LLFS data were available on 4,693 participants for 2.2 million assayed genotypes and 18.3 million imputed genotypes. Assayed markers with < 98% call rate and a high Mendelian error rate were excluded, as well as data on individuals with < 97% genotyping call rate. For imputed SNP data, SNP genotypes ranged from 0 - 2, representing the probability of the numbers of minor alleles. SNPs with imputation quality < 0.3 were not included in the analysis. Assayed and imputed SNPs with MAF < 0.01 were not included in the GWA analyses.

5.2.3 Genomewide Association in LLFS

Briefly, for each of the five-domain endophenotypes, family based genomewide association analyses were performed using a linear mixed-effect model correcting for family structure. I also included sex, age, recruitment site and ancestry PCs as covariates in the models. GWA analyses were performed using genotypic information available on 2.2 million assayed genotypes. The genomic inflation factor λ was calculated using the GenABEL package in R. Thresholds for suggestive and significant levels of association were p-value $< 5 \times 10^{-6}$ and p-value $< 5 \times 10^{-8}$, respectively.

To better characterize the chromosomal regions of interest obtained from the GWA analyses, all imputed and assayed SNPs were tested for association within a 2 Mb window surrounding the lead SNP (that is, the SNP with lowest p-value). Using LocusZoom (version 1.1)[93], the $-log_{10}$ transformed p-values for each of the analyzed SNPs were plotted against their physical location on the chromosome. In addition, recombination rates derived from HapMap (build GRCh37) were also plotted.

5.2.4 Replication in HABC Cohort for GWA and GWL Results

Data on 879 male and 784 females of European American ancestry was available from the HABC cohort. Endophenotypes were developed as described in Chapters 1 and 4. The following covariates were included in the models for association: age, sex, recruitment center and ancestry PCs. Association analyses were performed using ProbABEL (ProbABEL v. 0.4.1)[69].

As described in Chapter 4, variables in the cognition domain were not measured in the HABC cohort. In LLFS, F3 is solely defined by the cognition domain. In consequence, the

second LLFS factor (F2) appears to be represented by two factors, F2 and F3 in the HABC cohort. Therefore, to replicate signals obtained from association analyses of F2 in LLFS, I tested for association with F2 and F3 from the HABC cohort.

To minimize the number of association tests and maintain a reasonable probability of detecting a "true" association, I used several criteria in an iterative process to select a set of non-redundant SNPs for replication in the HABC cohort. See section 1.6.6 for rationale and details. Briefly, at each possible QTL location, all SNPs with *p*-values $< 10^{-5}$ were considered. Next, the SNP with lowest *p*-value and also present in HABC was chosen (the "lead" SNP) and all SNPs that were in high LD with the lead SNP, that is, $r^2 > 0.8$, were excluded. Among the SNPs that remained (that is, not in high LD with the first lead SNP), a second "lead" SNP was chosen based on the lowest *p*-value. Then all SNPs in high LD with the second lead SNP were excluded. This process continued until all SNPs were excluded (or chosen to be replicated).

5.3 RESULTS

I first assessed the distribution of the *p*-values using a Q-Q plot (Figure 5.1 and 5.2). As can be seen, the distributions of *p*-values for both factors were consistent with the distribution that is expected under the null hypothesis. Genomic inflation factor values were 1.05 and 1.02 for F1 and F2 respectively, which are within acceptable limits for GWA studies. Deviations from the null distribution (higher values of λ) may indicate unrecognized population substructure or issues that could result in inflated *p*-values that, in turn, would lead to incorrect conclusions.

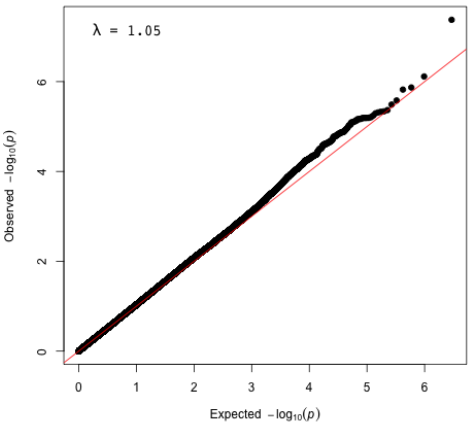

Figure 5.1: Q-Q Plot F1

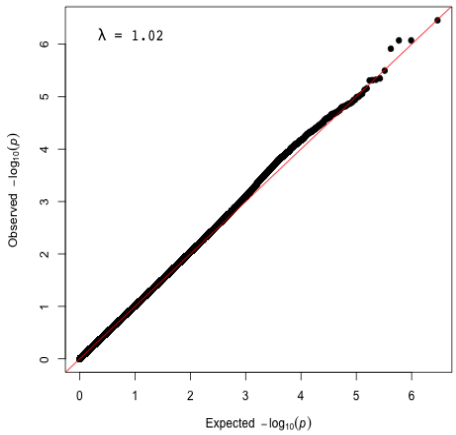

Figure 5.2: Q-Q Plot F2

5.3.1 Genomewide Association Results for Factor 1

Results of the genomewide association analysis for F1 revealed evidence for one QTL with genomewide significance (Figure 5.3). This possible QTL was located on 10p15 at rs7896849 (p-value = 4.21×10^{-8}) that is 19 kb downstream of *KLF6* gene. As can be seen, fine-mapping of assayed and imputed SNPs within a 2 Mb region around rs7896849 revealed additional SNPs associated with F1 at suggestive levels of significance (p-value < 5×10^{-6}; Figure 5.4). These SNPs are located between two recombination hotspots (the blue peaks) that flank the *KLF6* locus.

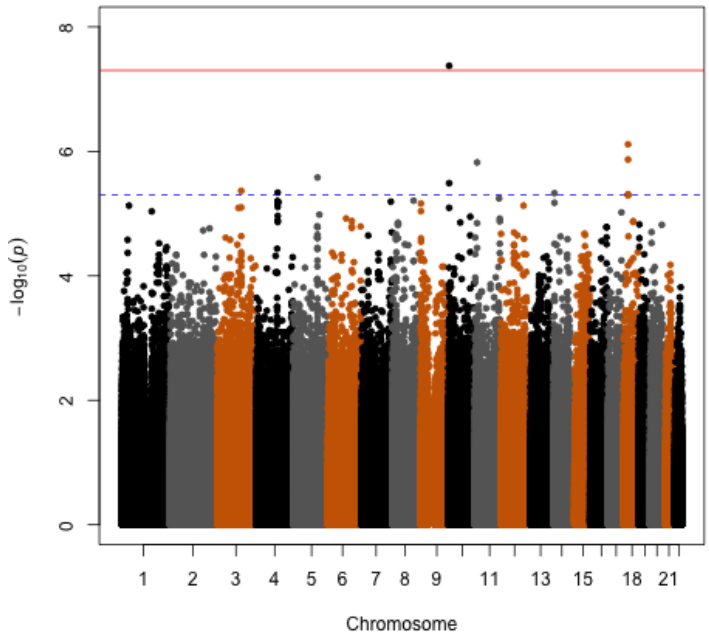

Figure 5.3: Manhattan Plot for F1

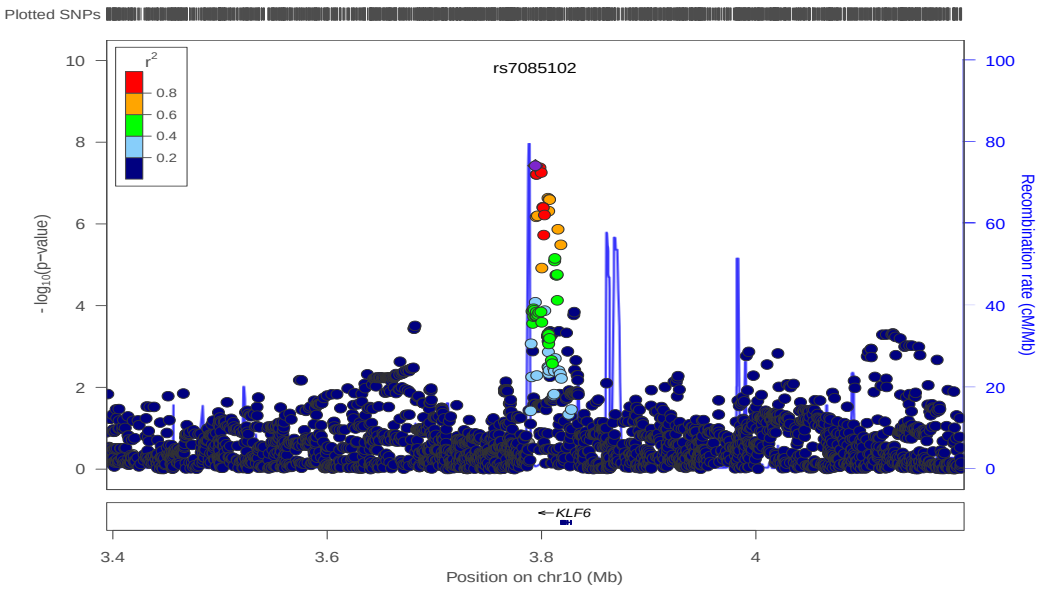

Figure 5.4: Regional Association Plot for the Locus at 10p15 for F1

In addition to the genomewide significant results, I also obtained suggestive evidence of association (p-values < 5×10^{-6}) for eight SNPs in six chromosomal regions (Table 5.1). Among the suggestive loci, the lowest p-value was obtained on chromosome 18q11.2 at rs10853653 (p-value = 7.69×10^{-7}). This SNP is 57 kb upstream of the *ZNF521* gene (Figure 5.5). Again, multiple SNPs were in high to moderate LD ($r^2 > 0.6$) and located between two recombination hotspots upstream of the *ZNF521* locus.

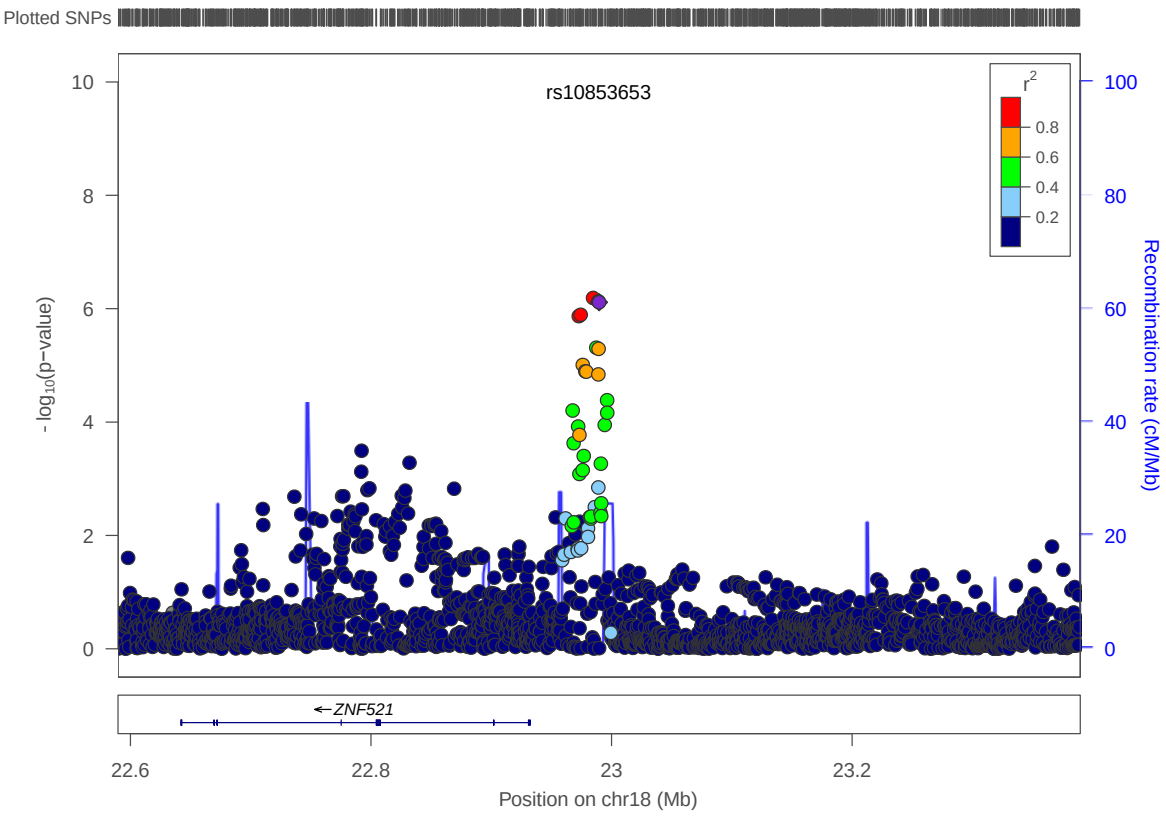

Figure 5.5: Regional Association Plot for the Locus at 18q11.2 for F1

Table 5.1: Results of GWA Analyses for F1 (p-value < 5×10^{-6})

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	p-value
rs17266628	3q21.1	122127926	G/A	0.216	*FAM162A*	intron-variant	-0.325	0.071	4.31×10^{-6}
rs10516563	4q25	111677722	G/T	0.126	*MIR297*	-104024	-0.404	0.088	4.61×10^{-6}
kgp5641805	5q23.2	125274062	T/C	0.403	*GRAMD3*	-421771	0.277	0.059	2.63×10^{-6}
rs7896849	10p15	3798495	A/C	0.364	*KLF6*	-19194	0.335	0.061	4.21×10^{-8}
rs2279414	10p15	3818058	G/A	0.306	*KLF6*	downstream-variant-500B	0.293	0.063	3.24×10^{-6}
rs988667	11p15.3	12010581	C/T	0.096	*DKK3*	intron-variant	0.474	0.098	1.51×10^{-6}
rs1958682	14q12	25598386	A/G	0.128	*STXBP6*	79467	0.404	0.088	4.70×10^{-6}
rs4548961	18q11.2	22972726	A/C	0.118	*ZNF521*	40746	0.446	0.092	1.35×10^{-6}
rs11083124	18q11.2	22987297	C/A	0.252	*ZNF521*	55317	0.306	0.067	4.87×10^{-6}
rs10853653	18q11.2	22989604	T/C	0.147	*ZNF521*	57624	0.413	0.083	7.69×10^{-7}

I also performed fine-mapping of the remaining 5 QTLs in Table 5.1 and identified an additional 39 variants with p-values < 5×10^{-6}. The complete list of significant/suggestive variants is presented in Table B8 (Appendix).

5.3.2 Tests for Replication of GWA Results for Factor 1

To test for replication in the HABC cohort, I identified 14 SNPs marking the 7 loci using the iterative method previously described (section 1.6.6). Results of the tests for replication are presented in Table 5.2. A Bonferroni corrected p-value $\leq 3.6 \times 10^{-3}$ was considered to be significant evidence for replication. There was no replication of the potential QTL for F1 on chromosome 10p15, however, there was a significant association (p-value = 8.19×10^{-4}) between the F1 and rs7240975 on chromosome 18q11.2. This SNP is 57 kb upstream of *ZNF521*.

Table 5.2: Results for Replication of QTLs for F1 in the HABC Cohort

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	LLFS		Replication HABC	
							Beta	*p*-value	Beta	*p*-value
rs16832958	3q21.1	122018493	T/C	0.244	*CASR*	12765	-0.306	7.94×10^{-6}	-0.087	0.340
rs17266628	3q21.1	122127926	G/A	0.216	*FAM162A*	intron-variant	-0.325	4.31×10^{-6}	-0.034	0.722
rs144691425	4q25	111683003	C/T	0.122	*MIR297*	-98743	-0.426	2.68×10^{-6}	0.071	0.554
rs465236	5q23.2	125273245	G/C	0.405	*GRAMD3*	-422588	0.279	2.25×10^{-6}	-0.016	0.839
rs7085102	10p15	3794279	T/G	0.366	*KLF6*	-23410	0.338	3.76×10^{-8}	0.010	0.906
rs10795073	10p15	3806127	C/T	0.371	*KLF6*	-11562	0.312	2.34×10^{-7}	0.056	0.503
rs1906143	10p15	3815373	T/C	0.305	*KLF6*	-2316	0.305	1.35×10^{-6}	0.049	0.584
rs988667	11p15.3	12010581	C/T	0.096	*DKK3*	intron-variant	0.474	1.51×10^{-6}	0.100	0.469
rs1958682	14q12	25598386	A/G	0.128	*STXBP6*	79467	0.404	4.70×10^{-6}	0.025	0.837
rs1241492	14q12	25605964	T/C	0.159	*STXBP6*	87045	0.366	4.91×10^{-6}	0.018	0.870
rs10445494	18q11.2	22974335	A/G	0.117	*ZNF521*	42355	0.447	1.28×10^{-6}	0.242	0.048
rs7237853	18q11.2	22984635	T/C	0.147	*ZNF521*	52655	0.416	6.45×10^{-7}	0.237	0.038
rs11083124	18q11.2	22987297	C/A	0.252	*ZNF521*	55317	0.306	4.87×10^{-6}	0.257	0.005
rs7240975	18q11.2	22989234	A/G	0.176	*ZNF521*	57254	0.355	5.10×10^{-6}	0.352	8.19×10^{-4}

5.3.3 Genomewide Association Results for Factor 2

Results of GWA analysis for the second factor (F2) are presented as a Manhattan plot in Figure 5.6; a total of 1,470,015 SNPs were tested for association. None of the loci reached the genomewide significant threshold of 5×10^{-8}, however, nine SNPs from seven regions reached the suggestive threshold of significance (*p*-value $< 5 \times 10^{-6}$; Table 5.3).

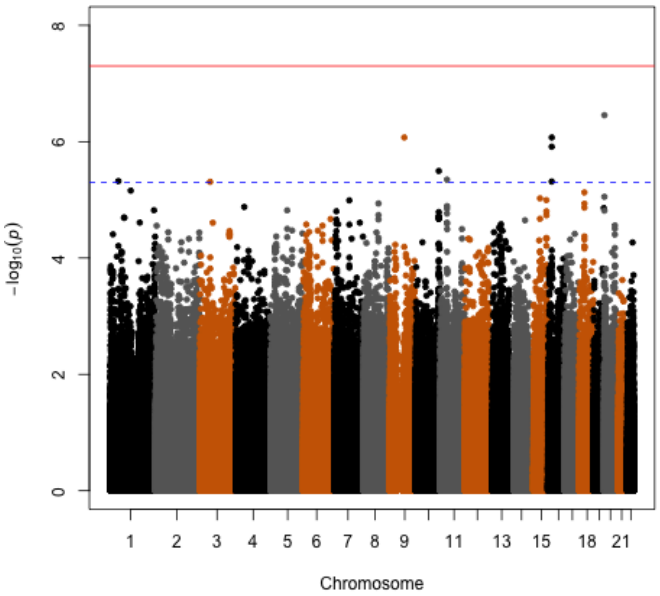

Figure 5.6: Manhattan Plot for F2

Table 5.3: Results of GWA Analyses for F2 (*p*-value < 5 x 10^{-6})

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs12088087	1p34.1	45170012	G/A	0.390	*C1orf228*	intron-variant	0.276	0.060	4.76×10^{-6}
rs9830791	3p14.3	57091575	A/G	0.479	*ARHGEF3*	intron-variant	-0.267	0.058	4.91×10^{-6}
rs765468	9q21.13	78906740	G/T	0.014	*PCSK5*	intron-variant	1.176	0.239	8.47×10^{-7}
rs1710313	10q26.2	127734513	C/A	0.305	*ADAM12*	intron-variant	0.302	0.065	3.19×10^{-6}
rs80354775	11p12	37439843	A/G	0.026	*C11orf74*	759043	0.868	0.189	4.49×10^{-6}
rs12102869	16p12.3	19918987	C/T	0.155	*GPRC5B*	23043	-0.400	0.081	8.47×10^{-7}
rs9926784	16p12.3	19941968	C/T	0.170	*GPRC5B*	46024	-0.358	0.078	4.84×10^{-6}
rs11648621	16p12.3	19973008	G/A	0.197	*GPR139*	-70043	-0.360	0.074	1.22×10^{-6}
rs203544	20p13	1195784	G/A	0.401	*C20orf202*	7023	-0.314	0.061	3.50×10^{-7}

The strongest evidence for association of a QTL with F2 was observed on chromosome 20p13 (with rs203544, *p*-value = 3.50×10^{-7}) and on chromosome 16p12.3 (with rs12102869, *p*-value = 8.47×10^{-7}).

Fine-mapping of the seven suggestive regions was done using assayed and imputed data, in a 2 Mb window with lead SNP, that is, the SNP with the lowest *p*-value at the mid-point. Fine-mapping identified additional 40 suggestive variants. Fine-mapping of the region of interest on chromosome 16p23.3 is presented in Figure 5.7.

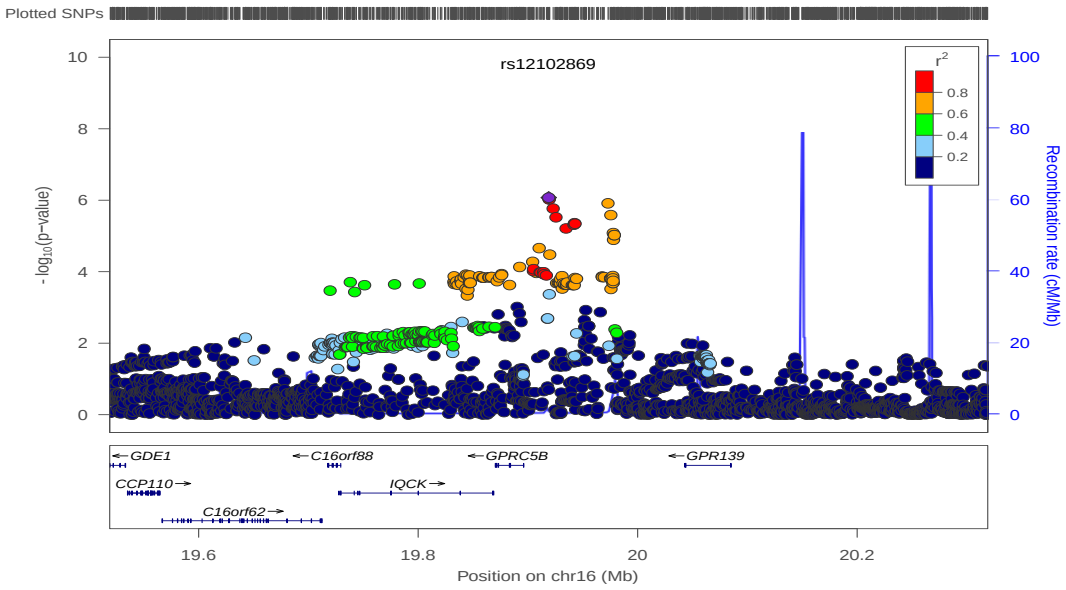

Figure 5.7: Regional Association Plot for the Locus at 16p12.3 for F2

A complete list of suggestive variants (*p*-values < 5×10^{-6}) is presented in Table B9 (Appendix).

5.3.4 Tests for Replication of GWA Results for Factor 2

Tests for replication of QTLs for F2 were also performed using the European American cohort in HABC. As described in Chapter 4, the largest eigenvalues for components of F2 in LLFS are represented in both F2 and F3 in HABC. Therefore, I performed association analyses for HABC endophenotypes F2 and F3. As described in the methods section, I selected a non-redundant (r^2 <

0.8) set of 11 SNPs marking the seven QTLs. Two of these selected SNPs failed quality control in the HABC cohort (MAF < 0.05). Results for the remaining nine variants are presented in Table 5.4. I considered a Bonferroni *p*-value ≤ 0.0028 to be significant.

Table 5.4: Results for replication of QTLs for Factor 2 in the HABC cohort

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	LLFS		HABC Replication			
							Beta	*p*-value	Beta (F2)	*p*-value (F2)	Beta (F3)	*p*-value (F3)
rs56164117	1p34.1	45173351	C/T	0.393	*C1orf228*	intron-variant	0.287	3.10×10^{-6}	0.067	0.936	0.083	0.072
rs9830791	3p14.3	57091575	A/G	0.479	*ARHGEF3*	intron-variant	-0.267	4.91×10^{-6}	0.065	0.538	0.076	0.081
rs1531331	10q26.2	127734012	T/G	0.304	*ADAM12*	intron-variant	0.302	3.01×10^{-6}	0.072	0.140	0.086	0.622
rs1278322	10q26.2	127820570	G/T	0.425	*ADAM12*	intron-variant	0.273	6.81×10^{-6}	0.066	0.929	0.078	0.684
rs12102869	16p12.3	19918987	C/T	0.155	*GPRC5B*	23043	-0.400	8.47×10^{-7}	0.087	0.438	0.099	0.008
rs11648621	16p12.3	19973008	G/A	0.197	*GPR139*	-70043	-0.360	1.22×10^{-6}	0.077	0.756	0.090	0.068
rs203547	20p13	1195245	T/A	0.482	*C20orf202*	6484	0.294	1.04×10^{-6}	0.065	0.386	0.075	0.870
rs203545	20p13	1195688	C/G	0.401	*C20orf202*	6927	-0.317	2.56×10^{-7}	0.064	0.748	0.075	0.247
rs203541	20p13	1196943	C/G	0.336	*C20orf202*	8182	-0.310	1.00×10^{-6}	0.066	0.977	0.080	0.275

None of the replication SNPs achieved the Bonferroni level of significance; the strongest association was obtained for F3 with rs12102869 (*p*-value = 0.008).

5.4 DISCUSSION

Over the past decade, several investigators have performed association studies (both candidate gene and GWA) on long-lived individuals to identify loci that may contribute to 'desirable phenotypes,' such as longevity and healthy aging. Genetic association studies of longevity are

often challenging because of the necessity of identifying appropriate controls for long-lived cases. The most appropriate controls would be individuals from the same cohort as the long-lived cases who were already deceased. The limitation of this best-case study design is that few such longitudinal studies have been done, and thus the numbers of cases and controls is fairly small for GWA analyses[57]. Nonetheless, multiple candidate gene and GWA studies have been performed using data from long-lived individuals, but most of the results of these studies have been inconsistent, perhaps indicating the genetic and environmental complexity underlying longevity. The best replicated findings are for variation at the *ApoE* locus and *FOXO3A*[57]. The effects of variation at *ApoE* on longevity are well-replicated and well-known[94,95]. The *FOXO3A* gene lies within the insulin/insulin-like growth factor 1 signaling pathway, a pathway that is known to extend lifespan in several animal models. In particular, *FOXO3A* is an orthologue of the Daf-16 locus that influences lifespan in *C. elegans*[96].

Fewer association studies of healthy aging phenotypes have been performed, partly because healthy aging has been defined in various ways, including the absence of various disease or morbidities at a pre-defined "older" age (such as event-free survival) or the presence of desirable traits, such as mobility, at specific "older" age. None of them have reported genomewide significant results[97].

In the current study, I performed GWA analyses using data on five endophenotypes derived from five health domains (cognition, pulmonary function, cardiovascular and metabolic health, and physical activity) that were hypothesized to influence healthy aging[23]. Genotypic and phenotypic data were available on 4,302 individuals. In Chapter 4, I reported that these five-domain endophenotypes were heritable (as previously reported by Matteini *et al.*, 2010)[23], similar to those obtained in another population (i.e., the HABC cohort), and that the most

dominant endophenotype (F1) was significantly correlated with mortality in both the LLFS and HABC populations. As discussed in Chapter 4, the derivation of the endophenotypes in the HABC cohort was not ideal, because HABC does not have data on similar cognitive measures. However, preliminary results of GWA analyses of endophenotypes derived without the cognitive domain traits in LLFS (see Chapter 4) are similar to the results reported here (see Tables B10 and B11 and Figures B47 and B48; Appendix).

I obtained suggestive or significant evidence for QTLs associated with endophenotype F1 at seven chromosomal regions (Table 5.1). The most significant result (p-value $< 5 \times 10^{-8}$), was obtained for a SNP that was ~ 19 kb downstream of the *KLF6* gene. *KLF6* (Kruppel-Like factor 6) belongs to a family of zinc-finger containing transcription factors involved in several biological processes such as differentiation, proliferation and development[98,99]. *KLF6* is a tumor suppressor gene and mutations in this gene have previously been associated with increased risk of prostate cancer[100]. Although this result is potentially interesting, I was not able to replicate this association in the HABC cohort; as the p-values were > 0.05 for SNPs marking this QLT region of interest. However, I also obtained suggestive evidence for a QTL influencing the F1 endophenotype on chromosome 18q11.2; SNPs in the QTL region of interest were significantly associated with F1 in the HABC cohort (p-value $= 8 \times 10^{-4}$). As can be seen in Table 5.2, the four SNPs that marked the 18q11.2 region and were assayed in both the LLFS and HABC cohorts had similar effects on the F1 endophenotype. These SNPs are located upstream of *ZNF521*. ZNF521 protein is a transcription factor, containing 30 kruppel-like zinc fingers and has been shown to play a role in erythroid cell differentiation[101,102].

I also obtained suggestive evidence that QTLs in seven chromosomal regions (Table 5.3) might influence variation in the F2 endophenotype. As stated previously, the F2 endophenotype

is predominantly comprised of traits from the cardiovascular and metabolic health domains in LLFS. In the HABC cohort, this endophenotype is represented by two factors, F2 and F3. Unfortunately, none of the QTL regions were significantly associated with F2 or F3 in HABC, although rs12102869 marking the QTL on chromosome 16p12.3 was nominally associated with F3 in HABC. The intergenic region between *GPRC5B* - *GPR139* on 16p12.3 has been shown to be associated with BMI[103]. Furthermore, endophenotype F3 in HABC is predominantly characterized by obesity related traits (waist circumference, BMI), so this result may reflect a true relationship, although it is not statistically significant.

In conclusion, my analyses indicate that a QTL influencing a healthy aging endophenotype predominantly comprised of pulmonary and physical function domains may be located on chromosome 18q11.2 near the *ZNF521* locus. Although I discuss the effects of coding genes, such as *ZNF521*, located near QTLs for endophenotypes F1 and F2, I recognize that these QTLs may be marking yet unknown regulatory regions, or regions under epigenetic control. Additional analyses of these regions may identify a gene or suite of genes that influences underlying pathways that contribute to both health domains.

6.0 LINKAGE ANALYSES OF FIVE-DOMAIN ENDOPHENOTYPES IN THE LONG LIFE FAMILY STUDY

6.1 INTRODUCTION

Numerous linkage studies of longevity and healthy aging have been performed using data on long-lived sibships and families. Linkage QTLs for measures of longevity tend not to overlap with linkage QTLs for measures of healthy aging[58]. Also, with few exceptions, the linkage-derived QTLs for healthy aging phenotypes do not overlap across studies[104,105]. Furthermore, with the exception of a linkage signal on 19q13 that was attributable to variation at the apolipoprotein E locus (*ApoE*)[105], none of the genes underlying the QTL linkage signals have been identified. Non-replication of the linkage QTLs for healthy aging may represent differences in the definition of the traits, the complexity of healthy aging, and/or the reflection of the underlying genetic heterogeneity. Thus, there is a need for additional large studies of healthy aging phenotypes.

In this chapter, I report my results of linkage analyses of the first two factors of the five-domain endophenotype derived in the LLFS.

6.2 METHODS

6.2.1 Endophenotypes in LLFS

Endophenotypes (and genotypes) from five health domains were available on 4,302 individuals. As described in section 4.3 (see Table 4.2), the dominant endophenotype (F1) in LLFS and HABC was predominantly comprised of traits from the pulmonary and physical function domains, whereas the second dominant endophenotype, F2, in LLFS was comprised of measures from the cardiovascular and metabolic health domains.

6.2.2 Genotype Data

As described in detail in section 1.5.3, multiallelic haplotypes were derived (using ZAPLO)[62], cleaned for Mendelian and recombination inconsistencies, and then multipoint IBD estimates were calculated using the LOKI program[61]. In addition, for fine-mapping under the QTL linkage peaks, I used a subset of the 2.2 million assayed genotypes and 18.3 million imputed genotypes (details are provided in section 1.5.1). Assayed markers with < 98% call rate and a high Mendelian error rate were excluded, as well as data on individuals with < 97% genotyping call rate. For imputed SNP data, SNPs with imputation quality < 0.3 were not included in the analysis. Assayed and imputed SNPs with MAF < 0.01 were not included in the fine-mapping association analyses.

6.2.3 Genomewide Linkage Analyses

Briefly, for each of the five-domain endophenotypes, family-based genomewide linkage analyses were done using an extension of the variance component framework described previously (see section 1.6.5), which includes the effect of a presumed QTL (σ^2QTL) as a component of genetic variance[65] . I also included sex, age, and recruitment site as covariates in the models. After detecting significant (or suggestive) evidence for linkage, I identified a region of interest under the linkage peak. I defined the region of interest as the chromosomal region contained within 1.5 LOD units on either side of the maximum LOD score[68].

6.2.4 Fine-Mapping

Two-point linkage analysis: Although linkage analyses performed using MIBDs derived from multiallelic loci are generally more powerful than analyses of IBDs from single SNPs, if the SNP is in high LD with a causal locus, it should provide strong evidence of co-segregation. To fine-map potential QTLs, I performed two-point (that is, single SNP) linkage analyses for each SNP in the area of interest for a specific linkage peak. All analyses were done using SOLAR[65] . To facilitate comparisons of association and linkage results, I obtained *p*-values for two-point LOD scores. LOD scores were converted to corresponding chi-square statistics (λ^2), calculated as: $\lambda^2 = \text{LOD} \times 2\log_e(10)$, where λ^2 has one degree of freedom.

Association analyses: To assess association of SNPs under the linkage peak, I used a linear mixed-effect model correcting for family structure. Details of the methodology are described in section 1.6.4. All assayed and imputed SNPs under the linkage peak were filtered from the analysis if they had a call rate < 98%, a minor allele frequency < 1% and a Hardy–

Weinberg equilibrium p-value < 10^{-6}. Results were reported as negative logarithm of the p-value.

Conditional linkage analyses: SNP conditional analysis was performed to assess whether SNPs identified through association and two-point linkage analyses can account for the observed linkage signal. To perform SNP conditional analysis, each SNP was coded (0, 1 or 2) as count/dosage of the minor allele. Linkage analysis was performed by including the SNP as covariate, and decrease in LOD score was observed.

6.3 RESULTS

6.3.1 Genomewide Linkage Analysis for Endophenotypes

Genomewide linkage analyses (GWL) for endophenotypes were performed using MIBD matrices calculated from multiple-SNP haplotypes, as described in section 1.5.3. Linkage analysis was performed for the first five endophenotypes (F1 - F5) in LLFS. Except for F2, none of the GWL analyses revealed significant evidence for linkage with a QTL (LOD score > 3.3) for any of the endophenotypes; the maximum LOD scores were 2.41 on chromosome 3, 3.98 on chromosome 1, 2.18 on chromosome 10, 1.77 on chromosome 10, and 2.55 on chromosome 6 for F1, F2, F3, F4, and F5, respectively (See Figures B49-B53; Appendix). The maximum LOD score of a QTL for F2 was obtained on chromosome 1q43 (maximum LOD = 3.98 at 257 cM followed by LOD = 3.06 at 266 cM; Figure 6.1). Additionally, I also obtained suggestive evidence (LOD > 2.5) for linkage of QTLs for F2 on chromosome 10p12.33 (LOD = 2.53) and 17q23.2 (LOD = 2.74).

In the subsequent sections, I describe my fine-mapping analyses of the two linkage peaks on chromosome 1q43 at 257 cM and 266 cM.

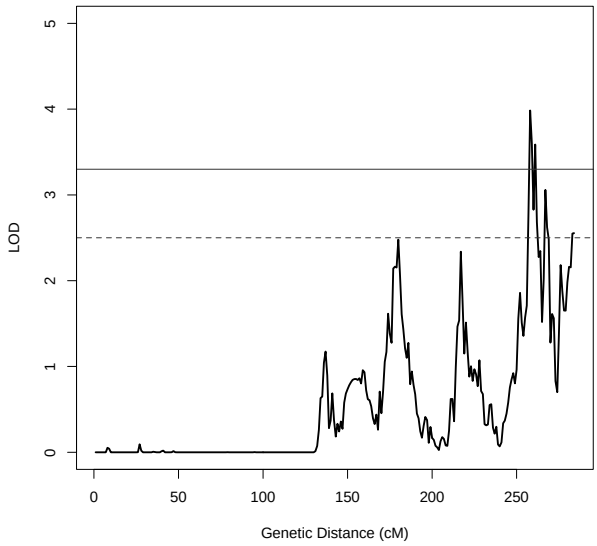

Figure 6.1: Multipoint Linkage Results for F2 on Chromosome 1

6.3.2 Fine-Mapping of QTL for F2 on 1q43: Peak at 257 cM

The region under the highest linkage peak on 1q43 ranged from 256 – 261 cM (representing the chromosomal region contained within 1.5 LOD units from the peak – see Figure 6.1). To fine-map the region of interest, I performed family-based association analysis, two-point linkage analysis and SNP conditional analysis using data on each SNP in the region.

Association analyses: A total of 8,866 SNPs (both genotyped and imputed) with MAF > 0.01 were assessed using family-based association analysis. Figure B54 (Appendix) presents the results of association analysis. None of the SNPs were significant after adjusting for multiple testing (that is, Bonferroni p-value $< 5.64 \times 10^{-6}$), however, this correction is conservative because it assumes independent tests and the SNPs that I tested were highly correlated. Table 6.1

presents the results of association analysis for SNPs with *p*-value < 3.2×10^{-3} (*-logp* > 2.5). The strongest association was observed for rs78101891 (*p*-value = 1.44×10^{-4}). This SNP was not in strong LD with any of the other SNPs listed in Table 6.1 (Figure B55; Appendix) and is located 132 kb downstream of *LOC339535*, a RNA gene. Out of 14 SNPs with *p*-value < 3.2×10^{-3}, 10 SNPs are within 200 kb upstream or downstream of *LOC339535.*

Table 6.1: Results of Association Analyses Between F2 and SNPs Under the Chromosome 1q43 257 cM Peak: SNPs with *p*-values < 3.2×10^{-3} are Listed

SNP	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs11585046	238269084	A/G	0.017	*LOC100130331*	177518	0.952	0.316	2.60×10^{-3}
rs116212833	238355441	C/A	0.016	*LOC100130331*	263875	0.994	0.318	1.80×10^{-3}
rs9661072	238472344	A/C	0.142	*LOC339535*	-171392	-0.252	0.085	3.13×10^{-3}
rs148797824	238472814	D/R	0.142	*LOC339535*	-170922	-0.259	0.085	2.41×10^{-3}
rs72759209	238472831	A/G	0.142	*LOC339535*	-170905	-0.259	0.085	2.41×10^{-3}
rs72759217	238474899	A/G	0.141	*LOC339535*	-168837	-0.265	0.085	1.91×10^{-3}
rs12142622	238476102	T/C	0.141	*LOC339535*	-167634	-0.265	0.085	1.93×10^{-3}
rs116503884	238489929	T/C	0.137	*LOC339535*	-153807	-0.269	0.089	2.37×10^{-3}
rs78101891	238511675	C/T	0.019	*LOC339535*	-132061	-1.175	0.309	1.44×10^{-4}
rs2392812	238524006	G/A	0.448	*LOC339535*	-119730	0.174	0.059	3.11×10^{-3}
rs191256743	238571563	T/C	0.014	*LOC339535*	-72173	-1.438	0.424	7.08×10^{-4}
rs116233900	238729433	C/T	0.030	*LOC339535*	80125	0.607	0.171	3.97×10^{-4}
rs141388658	239066354	A/C	0.010	*LOC339535*	417046	-1.105	0.326	7.05×10^{-4}
rs79930989	239505485	T/C	0.042	*LOC100505872*	-44340	-0.525	0.171	2.14×10^{-3}

Two-point linkage analyses: I also performed two-point linkage analysis of all assayed SNPs under the linkage peak (Table 6.2). The highest two-point LOD score (LOD = 4.11) was obtained for rs1361664 and the next highest two-point LOD score (LOD = 3.42) was obtained for rs1342078. These two SNPs were in modest LD with each other and are located 38 kb and 45 kb downstream of *LOC339535*, respectively. Similar to the results of association analysis, most of the highest two-point LOD SNPs are in the region around *LOC339535* (Table 6.2). LOD

scores were plotted along with recombination information to better characterize the region of interest (Figure B56; Appendix). LD among top two-point LOD SNPs is presented in Figure B57 (Appendix).

Table 6.2: Results of Two-Point Linkage Analyses for F2 Under the Chromosome 1q43 257cM Peak: SNPs with Two-Point LOD > 2.5 are Listed

SNP	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Two-Point LOD Score	*p*-value
rs1342078	238598399	A/G	0.228	*LOC339535*	-45337	3.42	3.61×10^{-5}
rs1361664	238605709	G/A	0.231	*LOC339535*	-38027	4.11	6.76×10^{-6}
rs4578212	238643498	T/A	0.249	*LOC339535*	-238	2.66	2.35×10^{-4}
rs574819	238777528	T/C	0.226	*LOC339535*	128220	2.72	1.98×10^{-4}
rs2841340	238934301	T/C	0.414	*LOC339535*	284993	2.72	2.02×10^{-4}
rs12145112	239170783	C/T	0.169	*LOC100505872*	-379042	2.78	1.72×10^{-4}

Conditional linkage analyses: Next, to assess whether the top SNPs (identified through association analysis and two-point linkage analysis) can account for the observed QTL, conditional linkage analyses were performed. Eighty-one SNPs from association analysis (-*logp* > 2) and 22 SNPs from two-point linkage analysis (LOD > 2) were included as covariates, individually, in the multipoint linkage analysis. In general, these SNPs did not reduce the maximum LOD score at 257 cM very much. I identified one SNP (rs78101891) that decreased the LOD score by 0.4 followed by rs2171907, which decreased the LOD score by 0.2 (Table 6.3). rs78101891 was also the most strongly associated SNP identified by family-based association analyses. These two SNPs were not in LD with one another and conditional linkage analyses including both of these SNPs decreased the LOD score by 0.56. Despite inclusion of

these two SNPs, significant evidence for linkage remains, indicating that these two SNPs are not in strong LD with the QTL on chromosome 1q43.

Table 6.3: Results of Conditional Linkage Analyses for Two SNPs with the Largest Effects on the 257 cM Peak

SNP	Position	minor/major allele	MAF	Nearby Gene	Position Near Gene	Decrease in LOD Score
rs78101891	238511675	C/T	0.019	*LOC339535*	-132061	0.40
rs2171907	238551688	T/C	0.414	*LOC339535*	-92048	0.26

6.3.3 Fine-Mapping of QTL for F2 on 1q43: Peak at 266 cM

Association analyses: To fine-map the chromosome 1q43 peak at 266 cM, a total of 6,270 SNPs (both genotyped and imputed) with MAF > 0.01 were assessed using family-based association analysis. None of the SNPs were significant after adjusting for multiple testing (that is, Bonferroni p-value < 7.97×10^{-6}), however this threshold is conservative. The most strongly associated SNP in the region was rs149740839 (p-value = 3.36×10^{-4}), an intronic variant in *RGS7*, followed by rs78344277 (p-value = 3.61×10^{-4}), which is an intronic variant in *PLD5* (Figure B58; Appendix).

Table 6.4 presents the results of association analysis for SNPs with p-values < 3.2×10^{-3}. Out of 10 SNPs with p-values < 3.2×10^{-3}, 4 SNPs are intronic variants in *RGS7*, two SNPs are intronic variants in *OPN3* and 2 SNPs are intronic variants in *MAP1LC3C* and *PLD5*. The most strongly associated SNP, rs149740839, was in moderate LD with rs181122729 ($r^2 = 0.51$). Two SNPs (rs3765811, rs3753216) located in the intronic region of *OPN3* were in high LD with each other ($r^2 = 0.97$; Figure B59).

Table 6.4: Results of Association Analyses Between F2 and SNPs Under the Chromosome 1q43 266 cM Peak: SNPs with *p*-values < 3.2×10^{-3} are Listed

SNP	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs181122729	240948350	T/C	0.014	*RGS7*	intron-variant	-1.293	0.365	4.04×10^{-4}
rs149740839	240952346	T/C	0.023	*RGS7*	intron-variant	-0.985	0.274	3.36×10^{-4}
rs200483920	241122000	D/R	0.019	*RGS7*	intron-variant	-0.912	0.287	1.53×10^{-3}
rs138462826	241128697	A/G	0.012	*RGS7*	intron-variant	-0.993	0.319	1.88×10^{-3}
rs3765811	241763789	G/A	0.327	*OPN3*	intron-variant	0.206	0.063	1.14×10^{-3}
rs3753216	241766551	G/A	0.327	*OPN3*	intron-variant	0.190	0.063	2.50×10^{-3}
rs201823920	242120345	R/D	0.132	*MAP1LC3C*	-38447	-0.286	0.094	2.41×10^{-3}
rs114367814	242159710	T/G	0.033	*MAP1LC3C*	intron-variant	0.537	0.166	1.23×10^{-3}
rs115256942	242219908	C/T	0.024	*PLD5*	-31788	0.645	0.218	3.15×10^{-3}
rs78344277	242286976	A/G	0.154	*PLD5*	intron-variant	0.291	0.081	3.61×10^{-4}

Two-point linkage analyses: Table 6.5 presents the results of two-point linkage analysis for the region of interest under the linkage peak on chromosome 1q43 at 266 cM. The highest two-point LOD score of 5.26 was obtained for rs28449276. This SNP is located in the intergenic region between *MAP1LC3C* and *PLD5* (Figure B60; Appendix). Two-point LOD scores > 2.5 were obtained for four other SNPs from this intergenic region; these SNPs were in moderate LD with each other, but not with rs28449276 (Figure B61; Appendix). I also obtained a two-point LOD score = 3.78 for rs261861 (an intronic variant in *RGS7*) and a LOD score = 2.67 for rs3765814 (an intronic variant in *OPN3*). The SNPs near *PLD5* were in moderate to high LD with each other, but none of the other SNPs were in LD with each other (Figure B61; Appendix).

Table 6.5: Results of Two-Point Linkage Analyses Between F2 and SNPs Under the Chromosome 1q43 266 cM Peak: SNPs with Two-Point LOD > 2.5 are Listed

SNP	Position	Minor/ major allele	MAF	Nearby Gene	Position Near Gene	Two-Point LOD Score	*p-value*
rs261861	241095576	A/G	0.459	*RGS7*	intron-variant	3.78	1.52×10^{-5}
rs3765814	241773027	T/C	0.168	*OPN3*	intron-variant	2.66	2.30×10^{-4}
rs28449276	242177677	C/A	0.325	*MAP1LC3C*	15326	5.26	4.27×10^{-7}
rs10158939	242223343	A/G	0.314	*PLD5*	-28353	2.80	1.65×10^{-4}
rs9428912	242233299	A/T	0.484	*PLD5*	-18397	2.56	2.95×10^{-4}
rs9428536	242233314	T/C	0.464	*PLD5*	-18382	3.04	9.09 × 10-5
rs28718783	242233583	G/A	0.479	*PLD5*	-18113	2.75	1.87×10^{-4}

Conditional linkage analyses: Results of SNP conditional analyses are presented in Table 6.6. The top 39 SNPs from association analysis *(-logp* > 2) and the 15 top SNPs from two-point linkage analysis (two-point LOD score > 2) were tested for SNP conditional analysis. The largest decrease of 0.42 was observed for rs115256942. None of the SNPs presented in Table 6.6 were in high LD ($r^2 > 0.8$) with each other (Figure B62; Appendix).

Table 6.6: Results of Conditional Linkage Analyses for Five SNPs with the Largest Effects on the 266 cM Peak LOD Score

SNP	Position	Minor/ major allele	MAF	Nearby Gene	Position Near Gene	Decrease in LOD score
rs181122729	240948350	T/C	0.014	*RGS7*	intron-variant	0.33
rs149740839	240952346	T/C	0.023	*RGS7*	intron-variant	0.31
rs2090689	242146145	T/A	0.473	*MAP1LC3C*	-12728	0.32
rs115256942	242219908	C/T	0.024	*PLD5*	-31788	0.42
rs199789153	242241731	I/R	0.025	*PLD5*	-9958	0.33

I next assessed whether a combination of these SNPs would reduce the linkage signal. I first used stepwise regression in R to select a subset of these SNPs that influenced F2. Based on these analyses, rs181122729, rs2090689 and rs115256942 were all included as covariates in the

multipoint linkage model. In combination, these three SNPs decreased the LOD score by 0.93 for the chromosome 1 peak at 266 cM.

Conditional linkage analyses across both chromosome 1q43 peaks: I next analyzed the combined effect of the SNPs across both of the two linkage peaks at 1q43. Using results from stepwise regression analyses, I selected 2 SNPs (rs78101891, rs2171907) from the peak at 257 (Table 6.3) and 3 SNPs (rs181122729, rs2090689 and rs115256942) from the peak at 266, and included all of them as covariates in the multipoint linkage model. presents the results of the combined SNP conditional analysis. In combination, these five SNPs decreased the LOD score by 1.11 at 257 cM and by 1.13 for the peak at 266 cM. Thus, these SNPs, in combination, account for a substantial proportion of the QTL linked with F2 on chromosome 1q43.

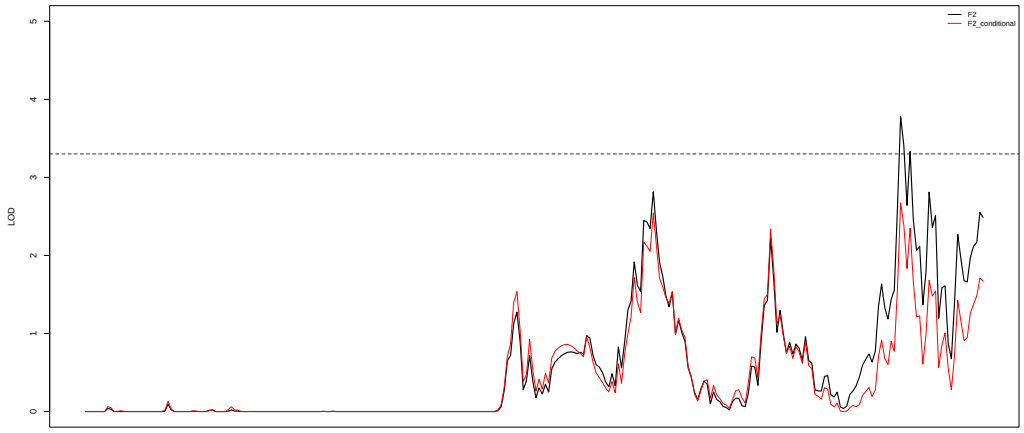

Figure 6.2: Results of Original and Conditional Multipoint Linkage Analyses of F2 on Chromosome 1

6.4 DISCUSSION

Over the past decade, several investigators have performed linkage analyses on data from long-lived sibships, as well as families of long-lived individuals to identify loci that may contribute to 'desirable phenotypes,' such as longevity and healthy aging. Healthy aging has been defined in various ways, including the absence of various diseases or morbidities at a pre-defined "older" age or the presence of desirable traits, such as mobility, at a specific "older" age.

Multiple studies of longevity have indicated the presence of a QTL on chromosome 3[106,107]. Additional linkage signals have been reported for several other chromosomes, but these have not been replicated[58].

Fewer linkage studies of healthy aging phenotypes have been performed. Edwards and colleagues (2012) defined successful aging in the Amish as individuals who were cognitively intact, high functioning and without depression. Their studies of 263 individuals in 12 sub-pedigrees revealed a QTL on chromosome 6q25[108]. A large linkage meta-analysis of 2,118 full sib-pairs greater than 90 years old (The European Genetics of Healthy Aging, GEHA) revealed linkage at 14q11, 17q12, 19p13, and 19q13[105]. Subsequent fine-mapping performed in a subset of 1,228 unrelated 90-year olds versus 1,907 controls indicated that the 19q13 QTL was likely due to variation at *ApoE*. Subsequent inclusion of the *ApoE2* and *ApoE4* alleles in the linkage model accounted for the signal at 19q13 in the GEHA study.

In the current study, I analyzed five endophenotypes derived from five health domains (cognition, pulmonary function, cardiovascular and metabolic health, and physical activity) that were hypothesized to influence healthy aging[23]. Genotypic and phenotypic data were available on 4,302 individuals in 574 pedigrees, thus this is one of the largest linkage analysis studies of a

healthy aging phenotype. In Chapter 4, I reported that these five-domain endophenotypes were heritable (as previously reported by Matteini *et al.*, 2010[23]), repeatable in another population, and that the most dominant endophenotype is correlated with mortality. Thus, I performed linkage analyses to assess whether any QTLs co-segregated with these endophenotypes.

I obtained significant evidence (LOD = 3.98) that a QTL located on chromosome 1q43 influenced the second endophenotype (F2). F2 is predominantly comprised of traits from the cardiovascular and metabolic health domains (see Table 4.2). Two peaks were present in this region – one at 257 cM and the other at 266 cM and the regions of interest under each of these peaks encompassed 2.2 and 1.5 Mb, respectively, and > 13 known loci. I next performed single-SNP linkage and association analyses under the two linkage peaks in an effort to narrow the region of interest and, perhaps, identify the QTL. Results of these analyses indicated that variation near several loci might influence F2, including: *LOC339535, RGS7, MAP1LC3C*, and *PLD5* (Tables 6.3 and 6.6). I next performed conditional linkage analyses that included SNPs near the above genes, separately, in the analytical model; however, these SNPs did not remove the evidence for linkage. Finally, I included the most significant SNPs near the above genes (*LOC339535, RGS7, MAP1LC3C*, and *PLD5*), simultaneously in the conditional linkage model and removed a substantial proportion of the evidence for linkage, but not all.

These results may indicate that variation at several closely linked loci influence variation in the F2 endophenotype. As stated previously, the F2 endophenotype is predominantly comprised of traits from the cardiovascular and metabolic health domains. And a case can be made that each of the four implicated genes (*LOC339535, RGS7, MAP1LC3C*, and *PLD5*) may influence fundamental variation in F2. Previously this region has been shown to be associated (*p*-value < 10^{-5}) with childhood obesity[109]. However, except for *PLD5*, none of the associated SNPs

are located within the "nearby" genes, thus, they may not be 'marking' variation that influences any or all of these four genes. Instead, these SNPs may be marking yet unknown regulatory regions, or regions under epigenetic control.

In conclusion, my analyses indicate that a QTL influencing a healthy aging endophenotype predominantly comprised of cardiovascular and metabolic health domains may be located on chromosome 1q43. Additional analyses of this region may identify a gene or suite of genes that influences underlying pathways that contribute to both health domains.

Future analyses: I have not yet attempted to replicate my results in another cohort for multiple reasons. First, replicating the linkage analyses results are difficult because few family studies of older individuals (mean age ~ 70 years) contain measures of the same traits from the five health domains. Second, the F2 endophenotype is represented by two factors in HABC, therefore determining which SNPs (or haplotypes or loci) to assess in each factor is not straightforward. However, as described in Chapters 4 and 5, initial results indicate that four endophenotypes derived in LLFS without the cognitive domain traits, are similar to endophenotypes 1, 2, 4, and 5 in LLFS, and also similar to factors 1, 2, 3, and 4 in HABC. Third, I have not yet performed heterogeneity analyses, nor identified possible at-risk families co-segregating with F2. Such analyses may reveal specific haplotypes that could be assessed in other studies.

7.0 CONCLUSION

Some investigators hypothesize that aging is a fundamental biological process that eventually leads to age-related diseases and disability, rather than the result of a conglomeration of multiple diseases. Identification of genes that influence multiple age-related disorders or health-related physiological conditions, that is, genes with pleiotropic effects, could provide support for this hypothesis. However, even if this hypothesis is incorrect, identification of genes that influence healthy aging phenotypes and/or age-related diseases, such as anemia, might reveal novel biological pathways that could be used to develop methods of prophylaxis or early interventions.

In the current study, I applied a variety of statistical genetic methods to a unique set of long-lived individuals and their families, the Long Life Family Study. My goal was to identify loci that influence hematologic traits, as well as endophenotypes derived from five domains of health; these endophenotypes may better characterize exceptional survival than any single trait. Hematological phenotypes (e.g., counts of white blood cells, red blood cells and platelets) are heritable, play important roles in immune response, oxygen carrying and blood clotting, and are associated with age-related diseases, such as anemia. Although genetic studies have identified multiple variants that are associated with hematologic traits, these known variants account for little of the heritable variation. Furthermore, the relationship between these variants and susceptibility to age-related health outcomes is unclear. Similarly, my LLFS colleagues previously constructed several endophenotypes comprised of traits that are presumed to correlate

with healthy aging[23]. These endophenotypes were heritable, but analyses had not been done to try to identify loci that might affect these endophenotypes nor was the relationship with mortality known. In the next section, I present the major results of my studies.

7.1 SUMMARY OF MAJOR RESULTS

General question #1: What is the genetic architecture of hematologic traits in the LLFS and are these traits related to measures of healthy aging?

As expected, all of the hematologic traits were moderately to highly heritable in the LLFS, similar to the reports in other studies. The magnitude and direction of effects of covariates, such as smoking, were also similar to previous reports. Hierarchical cluster analyses as well as estimates of genetic correlations among the hematologic traits revealed three general clusters: (1) measures of RBC counts, volume, and overall HGB; (2) measures of average size and hemoglobin content of RBCs; and (3) WBC related traits. Low RBC counts are indicative of anemias (cluster 1), whereas the RBC indices (cluster 2) are used to distinguish between anemia types. Although these traits are genetically correlated, the correlation is not perfect; therefore, I expected to find genes that influenced multiple traits within a cluster, as well as genes that only were associated with one trait. Indeed, this result has been observed in large GWAS studies[39]. Some of the hematologic traits were also genetically correlated with the Healthy Aging Index (Table 2.5). Because the Healthy Aging Index is comprised of non-hematologic traits, the finding of genetic correlation between HAI and hematologic traits is consistent with the hypothesis that part of the aging phenomenon is attributable to a fundamental biological process that is influenced by genetic variation.

General question #2: Do previously identified genetic variants (that is QTLs), as well as novel QTLs, influence the hematologic traits and endophenotypes in LLFS?

I identified 91 QTLs at the suggestive or significant threshold of evidence for association with hematologic traits, 35 of which have previously been reported by other studies. Importantly, using linkage and association analyses, I detected approximately 100 additional genes that had not been previously associated with hematologic traits. The most promising of these additional loci resulted from a linkage signal on chromosome 11p15.2 for RBC count. This QTL had previously been reported by investigators from the Framingham Heart Study[34]. Additional analyses in the LLFS implicated a region downstream of the *SOX6* gene, and SNPs in this region were also significantly associated with RBC count in the HABC cohort. *SOX6* is a transcription factor and further investigation of its function may lead to new treatments or prevention methods for anemia.

Aim 3: Are any of the healthy aging endophenotypes associated with mortality in LLFS and is this relationship replicable in another cohort?

Using factor analyses on data on traits from five health domains (cognition, physical activity, cardiovascular, metabolic, and pulmonary), I developed the same five endophenotypes that had been previously reported by Matteini and colleagues (2010)[23]. The most dominant factor, F1, is mainly comprised of the physical activity and pulmonary domains. The second factor, F2, is dominated by measures from metabolic and cardiovascular domains, whereas F3 comprises measures solely from the cognition domain. The fourth factor, F4, is characterized mainly by blood pressure related traits (hypertension, systolic BP, diastolic BP and pulse pressure) and F5 predominantly includes cardiovascular measures (Table 4.2). In addition, I obtained similar endophenotypes in the HABC cohort, especially when I compared

endophenotypes derived in LLFS after excluding the cognitive domain traits. In fact, not only were the compositions of the individual endophenotypes similar, the eigenvectors were strikingly similar between the two cohorts (Tables 4.2 and 4.4 and Appendix tables B6 and B7). Furthermore, F1 was significantly associated with reduced mortality in both LLFS and HABC, and this effect was independent of age and sex. These results indicate that a few of these endophenotypes may reflect fundamental biological processes that are associated with aging. Finally, these endophenotypes were moderately heritable (residual heritability ranged from 0.21 – 0.51), indicating that genetic variation is likely to contribute to these underlying biological processes.

Aim 4: Do novel QTLs influence variation in any of the healthy aging endophenotypes?

As discussed previously, very few statistical genetic studies have been performed to identify genes that influence healthy aging. Using GWA analyses, I identified a QTL on chromosome 18q11.2 (upstream from the *ZNF521* locus) that was associated with the F1 endophenotype. This endophenotype is comprised predominantly of the pulmonary and physical activity domains. I subsequently was able to replicate this association between F1 and SNPs near *ZNF521* in the HABC cohort. To my knowledge, this would be one of the first loci reported to influence variation in a healthy aging endophenotype. Additional studies need to be done to identify functional variants and determine the mechanism of action of this region on healthy aging.

In addition to QTLs identified by GWA analyses, I have also identified a QTL for the F2 endophenotype on chromosome 1q43. There are several genes of interest in this region, including *MAP1LC3C* and *PLD5*; however, I have not yet tried to replicate the relationship between SNPs in this region and F2 in another population.

7.2 FUTURE DIRECTIONS

Although the results of my studies are exciting, they represent a beginning. I have detected SNPs associated with hematologic and health-related endophenotypes, but these SNPs are unlikely to be causal. None of the associated SNPs are located in exons; almost all are located in 5' and 3' untranslated regions and thus they may be associated with regulatory variants. Furthermore, the associated SNPs may not regulate the "nearest" gene.

There are many analyses and experiments that can be done to try to identify the possible causal genes and causal variants, and the following is a description of a few approaches. First, I would try to replicate my results in another ancestry group. HABC consists of European Americans and African American participants. My initial replication analyses were performed in the European American cohort. Analyses of the African American cohort would strengthen my conclusions. Also, because LD differs between European and African American ancestry groups, these analyses may facilitate additional fine-mapping and eventual identification of causal variants.

Second, for the QTLs identified via linkage analyses, I could identify larger families with strong evidence for linkage. I would then assess co-segregation of haplotypes and the trait of interest within these families. Such analyses might facilitate identification of the potential causal genes or variants; however, these analyses are difficult to do with quantitative traits.

Third, I would perform some bioinformatic investigations to determine which genes within a QTL region of interest were expressed in specific tissues. This method could be useful for the hematologic traits, but it is not clear what tissues would be appropriate for the health-domain endophenotypes. I have investigated whether the various associated SNPs are in regions

of phylogenetic conservation, or whether they are likely to be regulatory elements, etc., based on ENCODE data. My initial analyses did not reveal anything especially notable.

Fourth, we could perform targeted sequencing in the region surrounding the QTL to identify rare variants or sets of rare variants that may be causal. For sequencing to be useful to identify rare variants influencing the continuous traits, I would need to identify a narrower region for sequencing or identify a set of individuals within families who appeared to be segregating "high" or "low" levels of the hematologic traits or endophenotypes.

7.3 PUBLIC HEALTH IMPACT

One of the goals of medicine and public health is to increase functional longevity. Anemia and other age-related blood cell trait abnormalities have been shown to be associated with adverse outcomes such as disability, hospitalization, morbidity and mortality in older adults[5,9,10,11,18]. Results from NHANES III (conducted between 1988-1994) indicated that 11.0% of men and 10.2% of women ≥ 65 years of age were anemic. Because the number of older adults is increasing both in the US and globally, the frequency of age-related hematologic disorders is likely to increase. In addition to hematologic disorders, the prevalence of all age-related disorders will increase as the US and global populations age. Therefore, knowledge of the genetic and environmental factors that influence composite traits of healthy aging, such as the Healthy Aging Index[22] or endophenotypes derived from multiple domains of health[23] would also be fruitful. Specifically, the identification of genes or novel biological pathways that regulate hematologic traits and/or healthy aging phenotypes could lead to insights and possible future

interventions to delay the onset of hematologic diseases, increase functional longevity, and concomitantly decrease the burden of age-related diseases on public health.

APPENDIX A

ABBREVIATIONS

Table A1: Abbreviations

1000HG	1000 Human Genome
AIC	Akaike Information Criteria
ALYM	Absolute Lymphocyte Count
ANEU	Absolute Neutrophil Count
ApoE	Apolipoprotein E
AUC	Area Under Curve
BMI	Body Mass Index
CAD	Coronary Artery Disease
CCND3	Cyclin D3
CDK	Cyclin Dependent Kinase
CDK6	Cyclin Dependent Kinase 6
CEPH	Centre d'Etude du Polymorphisme Humain
CHARGE	Cohorts for Heart and Aging Research in Genetic Epidemiology
CHD	Coronary Heart Disease
CHS	Cardiovascular Health Study
CI	Confidence Intervals
CIDR	Center for Inherited Disease Research
cM	Centimorgan
CSF3	Colony Stimulating Factor 3
DACH1	Dachshund Homolog 1 (Drosophila)
DARC	Duffy antigen receptor for chemokines
df	Degrees of Freedom
EDTA	Ethylene Diamine Tetraacetate
ENCODE	Encyclopedia of DNA Elements

Table A1 continued

EPO	Erythropoietin
FEV	Forced expiratory volume
FLoSS	Family Longevity Selection Score
FOXO3A	Forkhead Box Protein O3A
GEHA	Genetics of Healthy Aging
GWA	Genomewide Association
GWAS	Genomewide Association Study
GWL	Genomewide Linkage
HAI	Healthy Aging Index
HABC	Health Aging and Body Composition
HbF	Fetal Hemoglobin
HBS1L	Hsp70 Subfamily B Suppressor 1-Like
HCT	Hematocrit
HDL	High Density Lipoprotein
HFE	High Iron Fe
HGB	Hemoglobin
HLA	Human Leukocyte Antigen
HMIR	*HBS1L-MYB* intergenic region
HMG	High Mobility Group
HR	Hazard Ratios
HWE	Hardy-Weinberg Equilibrium
kb	Kilobase
KLF6	Kruppel-Like Factor 6
IBD	Identity By Descent
LD	Linkage Disequilibrium
LDL	Low Density Lipoprotein
LLFS	Long Life Family Study
LOD	Logarithm of Odds
LRT	Likelihood Ratio Test
MAF	Minor Allele Frequency
Mb	Megabase
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MCV	Mean Corpuscular Volume
MI	Myocardial infarction
MIBD	Multipoint Identity by Descent
MPV	Average platelets volume
NAV2	Neuron Navigator 2
NELL1	Neural Epidermal Growth Factor-Like 1
NHANES III	Third National Health and Nutrition Examination Survey

Table A1 continued

OR8B2	Olfactory Receptor, Family 8, Subfamily B, Member 2
PC	Principal component
PCA	Principal Components Analysis
PLT	Platelets
PSMD3	Proteasome 26S subunits non-ATPase 3
Q–Q	Quantile–Quantile
QTL	Quantitative Trait Loci
RBC	Red Blood Cells
SNP	Single Nucleotide Polymorphism
SOLAR	Sequential Oligogenic Linkage Analysis Routines
SOX6	SRY (Sex determining region Y)-Box 6
STAT3	Signal Transducer And Activator Of Transcription 3
TFR2	Transferrin receptor 2
TMPRSS6	Transmembrane Protease Serine 6
WBC	White Blood Cells
WHO	World Health Organization
ZNF521	Zinc Finger Protein 521

APPENDIX B

TABLES AND FIGURES

Table B1: Phenotypic Correlation among Blood Traits for Related Family Members and Spousal Controls

	HCT	HGB	RBC	MCH	MCHC	MCV	WBC	ALYM	ANEU	PLT
HCT	1	0.905	0.822	0.051	-0.164	0.178	0.112	0.123	0.075	-0.057
HGB	0.891	1	0.817	0.209	0.266	0.034	0.11	0.132	0.066	-0.076
RBC	0.788	0.784	1	-0.384	0.023	-0.404	0.153	0.144	0.099	0.01
MCH	-0.006	0.151	-0.487	1	0.389	0.749	-0.088	-0.035	-0.067	-0.141
MCHC	-0.25	0.211	-0.03	0.347	1	-0.308	-0.009	0.026	-0.023	-0.056
MCV	0.174	0.024	-0.462	0.774	-0.314	1	-0.086	-0.056	-0.053	-0.112
WBC	0.128	0.127	0.096	0.021	-0.013	0.03	1	0.485	0.841	0.294
ALYM	0.056	0.045	0.075	-0.059	-0.023	-0.043	0.433	1	0.011	0.169
ANEU	0.121	0.129	0.07	0.071	0.01	0.065	0.858	-0.006	1	0.241
PLT	-0.062	-0.113	-0.03	-0.104	-0.108	-0.038	0.294	0.212	0.203	1

Upper half of the matrix shows phenotypic correlations for related family members and lower half shows phenotypic correlation for spousal controls.

Table B2: Univariate LOD Scores for Hematologic Traits and Endophenotypes Using Different SNP Sets

Trait	Region	cM (Mb)	LOD Score			
			PittA	PittB	StLouis	Haplotypes
PC4	2p13.3	92 (70.7)	2.5	2.7	3.0	3.2
HCT	3p25.3	27 (9.6)	3.2	2.7	3.0	2.7
MCV	3q25.1	163 (151.7)	2.5	2.3	2.8	2.1
ANEU	8p21.3	39 (21.0)	2.1	2.2	2.5	2.6
WBC	8q12.1	72 (58.1)	3.3	3.0	2.6	2.8

Table B2 continued

Trait	Region	cM (Mb)	LOD Score			
			PittA	PittB	StLouis	Haplotypes
PLT	8p22	33 (17.8)	2.4	2.1	2.4	2.9
MCHC	10p12.3	45 (21.4)	3.8	4.2	3.8	3.7
PC4	10p12.1	53 (29.1)	2.6	2.5	3.0	2.5
RBC	11p15.1	38 (20.3)	1.9	2.5	2.8	3.4
RBC	11q24.1	134 (122.7)	2.3	2.1	2.4	3.0
RBC	11p15.2	26 (12.7)	2.1	2.1	2.7	2.5
PC1	11p15.2	27 (13.5)	1.5	2.1	2.6	2.5
PC1	17q12	61 (32.7)	2.9	2.7	2.8	2.5

Table B3: QTLs Showing Suggestive Association (p-value < 5×10^{-6}) with Hematologic Traits and Endophenotypes

Chromosome	LLFS		Previous Studies		
	Trait	Genes within 60 kb of top SNP	Trait	Genes within 60 kb of reported SNPs	Author
1p36.31	PC1	*ESPN; TNFRSF25; PLEKHG5; NOL9; TAS1R1*			
1p33	PC3	*FAAH; DMBX1*	MCV; MCH	*PDZK1IP1; TAL1; STIL*	vanderHarst P[48]
1p32.3	PC3	*OSBPL9*			
1p31.1	RBC	*LRRC7*			
1p22.3	WBC	*COL24A1*			
1q32.1	RBC	*NR5A2*	RBC; MCV; MCHC; MCH; MPV; PLT	*ATP2B4; SNORA77; NFASC; CNTN2; TMEM81; RBBP5; DSTYK; TMCC2; NUAK2; KLHDC8A*	vanderHarst P[48]; Soranzo N[47]; Gieger C[41]
1q42.3	PC4				
1q44	PC1; PC4	*OR6F1; OR14A2; OR14K1; OR1C1; OR14A16; HSD17B7P1*	PLT; MCV; RBC	*OR2W5; C1orf150; OR2C3; HSD17B7P1; OR11L1; TRIM58; OR2W3; OR2T8; OR2AJ1; OR2L13*	Gieger C[41]; vanderHarst P[48]; Kamatani Y[40]
2p21	HGB	*PRKCE*	PLT; RBC; HGB; HCT	*THADA; PRKCE*	Gieger C[41]; vanderHarst P[48]; Ganesh SK[39]; Kamatani Y[40]
2q13	MCH	*ACOXL*	MCV; MCH	*ACOXL; BCL2L11*	vanderHarst P[48]
2q24.3	PC3	*XIRP2*			
2q32.2	ALYM				
3q25.1	PC2	*WWTR1; COMMD2; C3orf16; RNF13*			
4p15.31	MCHC				
4q28.3	ALYM				
4q34.1	PC4	*HPGD*			
4q35.1	ANEU	*ENPP6*			
4q35.2	MCH				
5p15.33	PC1; RBC		MCHC; RBC	*SLC12A7; SLC6A18; TERT; CLPTM1L*	Kamatani Y[40]
5q13.3	ANEU	*SV2C*	MPV; PLT	*IQGAP2; F2R*	Gieger C[41]
5q34	PC2				
5q35.1	HGB	*DOCK2; FAM196B*			
5q35.2	PLT				
6p22.2	HGB; MCH; PC2	*HIST1H2BB; HIST1H3C; HIST1H1C; HFE; HIST1H4C; HIST1H1T; HIST1H2BC; HIST1H2AC; HIST1H1E; HIST1H2BD*	MCV; MCH; HGB; MCHC; PLT; HCT	*LRRC16A; SCGN; HIST1H2AA; HIST1H2BA; HIST1H2BPS1; SLC17A4; SLC17A1; SLC17A3; SLC17A2; TRIM38; HIST1H1A; HIST1H3A; HIST1H4A; HIST1H4B; HIST1H3B; HIST1H2AB; HIST1H2BB; HIST1H3C; HIST1H1C; HFE; HIST1H4C; HIST1H1T; HIST1H2BC; HIST1H2AC; HIST1H1E; HIST1H2BD; HIST1H2BE; HIST1H4D; HIST1H3D; HIST1H2AD; HIST1H2BF; HIST1H4E; HIST1H2BG; HIST1H2AE; HIST1H3E; HIST1H1D; HIST1H4F;*	vanderHarst P[48]; Ganesh SK[39]; Gieger C[41]; Kullo IJ[38]; Soranzo N[47]

Table B3 continued

Chromosome	LLFS		Previous Studies		
				HIST1H4G; HIST1H3F; HIST1H2BH; HIST1H3G; HIST1H2BI; HIST1H4H; BTN3A2; BTN2A2; BTN3A1; BTN3A3; BTN2A1; BTN1A1; HMGN4; ABT1;	
6p22.1	MCH; MCV	*ZSCAN23; COX11P1; GPX6; GPX5*	MCV; HGB; MCH; RBC; HCT; MCHC	*PRSS16; HIST1H4K; HIST1H2BN; HIST1H2AK; HIST1H2AL; HIST1H1B; HIST1H3I; HIST1H4L; HIST1H3J; HIST1H2AM; HIST1H2BO; RNU7-26P; OR2B2; OR2W6P; OR4D1; OR2B6; ZSCAN12; ZSCAN23; COX11P1; GPX6; GPX5; SCAND3; OR12D3; UBD; SNORD32B; RNF39; RPP21*	Ganesh SK[39]; vanderHarst P[48]
6p21.33	PC1; RBC; MCH	*HLA-C; SNORA38; APOM; SNORD52; SNORD48*	HGB; MCH; WBC; RBC; LYMPH; MCV; PLT	*RANP1; MIR877; CDSN; CCHCR1; HLA-C; SNORD84; SNORD117*	vanderHarst P[48]; Nalls MA[110]; Kamatani Y[40]; Gieger C[41]
6p21.32	PC1; RBC; HCT	*RNF5; GPSM3; HLA-DPB2*	RBC; HGB; HCT; PLT		vanderHarst P[48]; Gieger C[41]
6p21.2	MCV	*MDGA1*			
6p21.1	PC3; MCV	*USP49; MED20; BYSL; CCND3*	MCV; MCH; RBC; HGB; HCT	*MDFI; TFEB; PGC; FRS3; PRICKLE4; TOMM6; USP49; MED20; BYSL; CCND3; TAF8; C6orf132; VEGFA;*	vanderHarst P[48]; Ganesh SK[39]; Kamatani Y[40]; Soranzo N[47]
6q21	MCV	*SOBP; SCML4*	MONO; EOS; MCH; MCV; RBC; MCHC	*HLA-C; SNORD84; SNORD117; ARMC2; SESN1; C6orf182; CCDC162; CD164; PPIL6; SMPD2; MICAL1; AKD1; FIG4*	Okada Y[50]; vanderHarst P[48]; Ganesh SK[39]; Kamatani Y[40]
6q23.3	PC3; RBC; PLT; MCV; MCH	*HBS1L*	MCH; MCV; RBC; fHGB; HCT; MCHC; HGB; PLT; WBC	*ALDH8A1; HBS1L; MYB; MIR548A2*	vanderHarst P[48]; Ganesh SK[39]; Lettre G[111]; Kullo IJ[38]; Gieger C[41]; Kamatani Y[40]; Soranzo N[47]
6q25.1	WBC		RBC	*TAB2*	Yang Q[34]
7p22.3	ALYM ; PLT	*CHST12; LFNG; TTYH3; AMZ1; GNA12*			
7p21.3	HCT				
7p15.3	MCH				
7p14.3	PC1; RBC	*CPVL; CHN2*			
7p14.1	MCH	*ELMO1*			
7q21.13	MCH				
7q22.3	RBC	*PRKAR2B*	PLT; MPV		Gieger C[41]; Soranzo N[47]
7q31.2	ANEU	*TES; CAV2*			
7q33	WBC		MCH	*NUP205; SLC13A4*	vanderHarst P[48]
8p11.1	PLT				
8q12.1	MCH	*FAM110B*			
8q13.3	MCV	*KCNB2*			
8q24.3	WBC	*TRAPPC9*	PLT	*PLEC; MIR661; PARP10; GRINA*	Gieger C[41]
9p24.2	PC1	*GLIS3*	RBC; HCT	*GLIS3*	vanderHarst P[48]
9p24.1	PLT	*AK3; RCL1*	PLT; MCV; MCH; RBC	*CDC37L1; AK3; RCL1; MIR101-2; PTPRD*	Gieger C[41]; Soranzo N[47]; Ganesh SK[39]; vanderHarst P[48]; Yang Q[34]
9p21.1	ALYM	*LINGO2*			
9p13.2	PLT	*GRHPR; ZBTB5; POLR1E; FBXO10*			
9q31.3	HCT	*SUSD1*			
9q32	MCH	*ZNF618*			
9q34.3	PC2	*LCN1; OBP2A; PAEP; GLT6D1*			
10p15.1	PC4	*IL15RA; IL2RA; RBM17*			
10p14	HCT				
10p11.21	MCH	*CCNY; GJD4; FZD8*			
10q22.1	ALYM	*UNC5B; SLC29A3*	HCT; HGB; MCV; RBC; MCH	*HK1; C10orf27; SGPL1; PCBD1*	Ganesh SK[39]; vanderHarst P[48]; Yang Q[34]
10q22.2	MCH	*C10orf11*			
10q23.31	WBC	*CH25H; LIPA; IFIT2*			
10q26.13	PC3	*DMBT1; C10orf120*			
10q26.3	HCT				
11p14.1	PC3				

Table B3 continued

Chromosome	LLFS		Previous Studies		
11q23.3	HGB	*FXYD6; TMPRSS13*	PLT	*NLRX1; PDZD3; CCDC153; CBL*	Gieger C[41]
11q25	PC3				
12p11.22	HGB		MPV	*FAR2; ERGIC2*	Gieger C[41]
12p11.21	ANEU				
12p11.1	ANEU				
12q22	PC3				
12q23.3	MCH	*CHST11*			
12q24.12	ALYM	*SH2B3; ATXN2*	HCT; HGB; PLT; RBC; MCHC	*CUX2; FAM109A; SH2B3; ATXN2; BRAP; ACAD10; ALDH2; MAPKAPK5*	vanderHarst P[48]; Gieger C[41]; Ganesh SK[39]; Soranzo N[47]; Kamatani Y[40]
12q24.13	ALYM	*PTPN11*	HGB; HCT; RBC; PLT	*TMEM116; ERP29; NAA25; TRAFD1; C12orf51; PTPN11; MIR1302-1*	vanderHarst P[48]; Ganesh SK[39]; Soranzo N[47]; Gieger C[41]
12q24.21	PC2		MCHC		vanderHarst P[48]
12q24.31	HCT	*KNTC1; GPR81*	MCV; RBC; MCH; MPV; PLT	*CABP1; MLEC; UNC119B; ACADS; PSMD9; WDR66; NCOR2*	vanderHarst P; Gieger C[41]; Meisinger C[112]; Soranzo N[47]
13q21.33	HCT	*DACH1*			
14q11.2	WBC		MCHC	*OR4N2; OR4K2; OR4K5; OR4K1*	vanderHarst P[48]
14q23.1	PC2				
14q24.1	PLT	*ACTN1*	PLT; MCV	*RAD51L1; GALNTL1; ERH; SLC39A9*	Gieger C[41]; vanderHarst P[48]
14q32.13	PC1	*SNHG10; SCARNA13; GLRX5*			
14q32.31	MCV	*PPP2R5C*	MCH; PLT	*RAGE; ZNF839; CINP; TECPR2; RCOR1*	vanderHarst P[48]; Gieger C[41]
15q21.2	ALYM	*ATP8B4*			
15q21.3	PC4; MCHC	*ZNF280D*	RBC; HGB	*PRTG; NEDD4; LIPC*	vanderHarst P[48]
16q12.2	PC3	*FTO*			
16q22.1	ANEU	*WWP2; MIR140; CLEC18A*	RBC; MCH; MCV	*CTCF; RLTPR; ACD; PARD6A; C16orf48; C16orf86; GFOD2; RANBP10; TSNAXIP1; CENPT; THAP11; NUTF2; EDC4; NRN1L; PSKH1; CTRL; PSMB10; LCAT; SLC12A4; DPEP3; DPEP2; DDX28; DUS2L; NFATC3; ESRP2; PLA2G15; SLC7A6; SLC7A6OS; PRMT7; CDH3*	vanderHarst P[48]
16q24.1	PC2	*ATP2C2; KIAA1609*			
17q21.1	WBC; PC3; ANEU	*ORMDL3; GSDMA; PSMD3; CSF3; MED24; SNORD124*	WBC; NEUT	*GSDMB; ORMDL3; GSDMA; PSMD3; CSF3; MED24; SNORD124; THRA*	Soranzo N[47]; Nalls MA[110]; Okada Y[52]
17q21.2	WBC	*STAT5A; STAT3; PTRF*			
17q22	ALYM				
17q23.2	PLT	*TBX2; C17orf82; TBX4*			
17q24.2	PC1	*PITPNC1; NOL11; SNORA38B*			
19p13.2	HGB; PC3; MCH; MCV	*DOCK6; TSPAN16; RAB3D; TMEM205; CCDC159; RTBDN; MAST1; DNASE2; KLF1; GCDH; SYCE2; FARSA; CALR; RAD23A*	MCV; MCH; RBC	*ZNF490; ZNF791; MAN2B1; WDR83; C19orf56; DHPS; FBXW9; TNPO2; SNORD41; HOOK2; JUNB; PRDX2; RNASEH2A; RTBDN; MAST1; DNASE2; KLF1; GCDH; SYCE2; FARSA; CALR; RAD23A; GADD45GIP1; DAND5; NFIX*	Ganesh SK[39]; vanderHarst P[48]
21q21.1	PC2				
22q12.3	MCH; PC2; MCV	*C22orf33; TST; MPST; KCTD17; TMPRSS6; IL2RB*	MCH; MCV; HGB; HCT; MCHC	*C22orf28; BPIL2; FBXO7; SYN3; CSF2RB; C22orf33; TST; MPST; KCTD17; TMPRSS6; IL2RB*	vanderHarst P[48]; Soranzo N[47]; Ganesh SK[39]; Benyamin B[43]; Chambers JC[45]; Kullo IJ[38]

Table B4: Trait-Locus Combinations with Suggestive Association (*p-value* < 5×10^{-6}) for Hematologic Traits

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value	Trait
rs2986747	1p36.31	6562784	G/A	0.181	*PLEKHG5*	intron-variant	-0.231	0.049	2.17E-06	PC1
rs76108244	1p33	46931854	G/A	0.012	*LOC729041*	20564	-0.791	0.149	1.05E-07	PC3
rs78250115	1p32.3	52136855	C/T	0.146	*OSBPL9*	intron-variant	-0.218	0.045	1.36E-06	PC3
kgp9335112	1p32.3	52142548	A/G	0.143			-0.212	0.045	2.97E-06	PC3
rs12034677	1p31.1	70454240	C/T	0.413	*LRRC7*	intron-variant	-0.426	0.093	4.76E-06	RBC
rs6687339	1p22.3	86638265	G/A	0.135	*COL24A1*	16159	0.338	0.070	1.48E-06	WBC
rs2816998	1q32.1	200067248	C/T	0.155	*NR5A2*	intron-variant	-0.575	0.125	4.38E-06	RBC
rs16844140	1q42.3	235019266	T/C	0.059	*PP2672*	123620	0.256	0.054	2.51E-06	PC4
rs78367025	1q44	244431317	T/C	0.065	*C1orf100*	-84773	-0.381	0.076	5.35E-07	PC1
rs4925570	1q44	247935307	G/A	0.165	*OR9H1P)*	-2698	0.159	0.034	3.93E-06	PC4
rs12613391	2p21	46301750	A/G	0.108	*PRKCE*	intron-variant	-0.192	0.041	3.04E-06	HGB
rs4849120	2q13	111599601	G/A	0.299	*ACOXL*	intron-variant	0.199	0.041	1.16E-06	MCH
rs114711819	2q24.3	167688587	A/G	0.014	*XIRP2*	-56442	0.651	0.136	1.65E-06	PC3
rs13394281	2q32.2	191650662	G/T	0.458	*NAB1*	93181	0.227	0.049	3.71E-06	ALYM
rs73870405	3q25.1	149492807	G/A	0.026	*ANKUB1*	intron-variant	0.469	0.102	4.88E-06	PC2
rs938840	4p15.31	18851987	T/G	0.288	*LCORL*	826570	0.133	0.028	2.51E-06	MCHC
rs115156266	4q28.3	133208971	C/T	0.055	*PCDH10*	-861660	-0.499	0.108	4.06E-06	ALYM
rs2612659	4q34.1	175433338	C/A	0.300	*HPGD*	intron-variant	0.131	0.028	2.94E-06	PC4
rs34493244	4q35.1	185122787	G/A	0.304	*ENPP6*	intron-variant	0.347	0.069	5.75E-07	ANEU
rs6816228	4q35.2	188160927	C/T	0.144	*LOC339975*	-64335	-0.258	0.054	1.80E-06	MCH
rs16872928	5p15.33	4211418	G/T	0.029	*IRX1*	610010	0.565	0.114	7.25E-07	PC1
rs16872928	5p15.33	4211418	G/T	0.029	*IRX1*	610010	-1.281	0.278	3.99E-06	RBC
kgp4026960	5p15.33	4212406	T/C	0.029			0.556	0.113	8.16E-07	PC1
rs6859341	5q13.3	75496098	G/A	0.326	*SV2C*	*intron-variant*	-0.333	0.068	1.09E-06	ANEU
rs4530779	5q34	164194594	G/A	0.216	*LOC100507193*	*intron-variant*	0.183	0.039	3.36E-06	PC2
rs17071870	5q35.1	169341337	C/T	0.058	*DOCK2*	*intron-variant*	-0.257	0.055	3.64E-06	HGB
rs606095	5q35.2	175022459	G/A	0.471	*HRH2*	*-62654*	-5.876	1.273	4.04E-06	PLT
rs79220007	6p22.2	26098474	C/T	0.049	*HFE*	*2511*	0.287	0.059	9.70E-07	HGB
rs79220007	6p22.2	26098474	C/T	0.049	*HFE*	*2511*	0.501	0.084	3.08E-09	MCH
rs79220007	6p22.2	26098474	C/T	0.049	*HFE*	*2511*	-0.398	0.074	9.02E-08	PC2
rs35889911	6p22.1	28467606	A/G	0.055	*GPX6*	*-3555*	0.393	0.080	9.07E-07	MCH
rs35889911	6p22.1	28467606	A/G	0.055	*GPX6*	*-3555*	1.118	0.241	3.61E-06	MCV
kgp11969343	6p21.33	31240312	C/T	0.302			0.200	0.041	8.53E-07	PC1
rs113215453	6p21.33	31315501	A/G	0.270			-0.497	0.101	9.90E-07	RBC
rs2736157	6p21.33	31600820	C/T	0.166	*PRRC2A*	intron-variant	0.237	0.050	1.96E-06	PC1
rs486416	6p21.33	31856070	C/T	0.287	*EHMT2*	intron-variant	0.202	0.041	7.52E-07	MCH
rs3130303	6p21.32	32205867	G/A	0.128	*NOTCH4*	12149	0.286	0.055	2.39E-07	PC1
rs3129716	6p21.32	32657436	C/T	0.102	*HLA-DQB1*	20993	-0.728	0.149	1.13E-06	RBC
rs2071354	6p21.32	33044388	C/T	0.147	*HLA-DPA1*	intron-variant	-0.556	0.109	3.90E-07	HCT
rs2051072	6p21.2	37561865	G/T	0.225	*MDGA1*	-38420	-0.639	0.135	2.40E-06	MCV
rs3218086	6p21.1	41910064	A/G	0.166	*CCND3*	intron-variant	0.201	0.042	1.56E-06	PC3
rs3218086	6p21.1	41910064	A/G	0.166	*CCND3*	intron-variant	0.843	0.150	2.01E-08	MCV
kgp22761599	6q21	108000797	G/A	0.165			0.742	0.150	7.59E-07	MCV
rs9376090	6q23.3	135411228	C/T	0.244	*HBS1L*	33201	0.227	0.036	3.28E-10	PC3
rs9376090	6q23.3	135411228	C/T	0.244	*HBS1L*	33201	-0.605	0.105	9.51E-09	RBC
rs6920211	6q23.3	135431318	C/T	0.230	*HBS1L*	53291	7.964	1.521	1.72E-07	PLT
rs6920211	6q23.3	135431318	C/T	0.230	*HBS1L*	53291	0.881	0.133	3.45E-11	MCV
rs9494145	6q23.3	135432552	C/T	0.212	*HBS1L*	54525	0.366	0.045	6.73E-16	MCH
rs12205882	6q25.1	150820346	A/G	0.258	*IYD*	94153	-0.275	0.056	9.26E-07	WBC
rs886626	7p22.3	2498052	A/C	0.325	*LOC100288594*	11439	0.243	0.053	4.68E-06	ALYM
rs10950842	7p22.3	2753929	C/A	0.462	*AMZ1*	reference	-6.560	1.272	2.60E-07	PLT
rs10249915	7p21.3	9018291	C/T	0.369	*NXPH1*	225747	0.372	0.080	3.25E-06	HCT
rs59376676	7p15.3	23940721	C/T	0.370	*STK31*	68677	0.179	0.039	4.06E-06	MCH
rs73087002	7p14.3	29146595	A/G	0.175	*CPVL*	intron-variant	-0.242	0.049	9.88E-07	PC1
rs73087002	7p14.3	29146595	A/G	0.175	*CPVL*	intron-variant	0.568	0.121	2.81E-06	RBC
rs2080410	7p14.1	37286864	T/C	0.366	*ELMO1*	intron-variant	-0.184	0.039	2.32E-06	MCH
rs1860586	7q21.13	89581246	A/C	0.368	*DPY19L2P4*	-167642	0.184	0.038	1.88E-06	MCH
rs117533401	7q22.3	106667546	T/C	0.022	*PRKAR2B*	-17661	-1.421	0.308	4.13E-06	RBC
kgp8981518	7q31.2	115899857	T/C	0.129			-0.465	0.097	1.73E-06	ANEU
rs34870036	7q33	137933825	A/G	0.117	*AKR1D1*	130291	0.361	0.077	2.72E-06	WBC
rs6474463	8p11.1	43383815	T/C	0.201	*POTEA*	165690	7.665	1.608	1.93E-06	PLT
rs7010991	8q12.1	59037323	A/G	0.336	*FAM110B*	intron-variant	0.190	0.039	1.45E-06	MCH
rs72653580	8q13.3	73846853	C/T	0.031	*KCNB2*	intron-variant	-1.476	0.322	4.81E-06	MCV
rs13282061	8q24.3	141300377	T/C	0.059	*TRAPPC9*	intron-variant	0.537	0.106	4.63E-07	WBC
rs17273930	9p24.2	4090724	G/A	0.287	*GLIS3*	intron-variant	0.194	0.042	3.40E-06	PC1

Table B4 continued

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value	Trait
rs13284412	9p24.1	4778777	A/G	0.192	*RCL1*	-14069	7.836	1.634	1.67E-06	PLT
rs10757733	9p21.1	28366839	A/G	0.296	*LINGO2*	intron-variant	0.263	0.054	1.20E-06	ALYM
rs730283	9p13.2	37483926	G/T	0.304	*POLR1E*	-2116	6.943	1.418	1.02E-06	PLT
rs1367057	9q31.3	114831697	T/G	0.057	*SUSD1*	intron-variant	-0.815	0.163	5.77E-07	HCT
rs1999203	9q32	116652328	C/T	0.048	*ZNF618*	intron-variant	-0.407	0.088	4.19E-06	MCH
rs74792450	9q34.3	138473253	A/G	0.068	*LOC100130954*	intron-variant	-0.297	0.064	3.51E-06	PC2
rs12722522	10p15.1	6078553	A/G	0.109	*IL2RA*	intron-variant	0.199	0.042	1.97E-06	PC4
rs17364530	10p14	9152705	C/T	0.024	*LOC100507163*	-164904	1.188	0.247	1.62E-06	HCT
rs77013689	10p11.21	35882638	G/A	0.042	*GJD4*	-11784	0.449	0.095	2.56E-06	MCH
rs10823721	10q22.1	73058558	G/A	0.169	*UNC5B*	intron-variant	-0.309	0.066	2.53E-06	ALYM
rs11001499	10q22.2	77555270	C/T	0.469	*C10orf11*	intron-variant	0.175	0.037	2.30E-06	MCH
rs1412444	10q23.31	91002927	A/G	0.336	*LIPA*	intron-variant	0.243	0.052	3.47E-06	WBC
rs7913531	10q26.13	124456033	A/C	0.376	*C10orf120*	-1205	0.154	0.033	2.73E-06	PC3
rs117478505	10q26.3	130623086	A/C	0.027	*MGMT*	-642459	1.127	0.244	3.84E-06	HCT
rs9666212	11p14.1	29531097	C/T	0.220	*KCNA4*	-500224	0.189	0.038	6.73E-07	PC3
rs56703391	11q23.3	117788879	T/A	0.196	*TMPRSS13*	intron-variant	-0.152	0.033	4.32E-06	HGB
rs2116390	11q25	133019923	G/T	0.372	*OPCML*	intron-variant	-0.157	0.032	1.01E-06	PC3
rs79443175	12p11.22	30117100	G/A	0.010	*TMTC1*	179492	-0.567	0.123	4.25E-06	HGB
rs76158898	12p11.21	33174972	C/T	0.025	*PKP2*	123303	1.002	0.202	6.81E-07	ANEU
rs75057219	12p11.1	34026631	T/C	0.026	*ALG10*	-146613	0.890	0.194	4.64E-06	ANEU
rs74375663	12q22	93620637	A/G	0.098	*LOC643339*	intron-variant	0.238	0.051	3.86E-06	PC3
rs2248220	12q23.3	104822346	G/A	0.320	*CHST11*	-26350	-0.201	0.040	5.35E-07	MCH
rs10774625	12q24.12	111910219	G/A	0.486	*ATXN2*	intron-variant	-0.229	0.050	4.11E-06	ALYM
rs11066320	12q24.13	112906415	A/G	0.457	*PTPN11*	intron-variant	0.236	0.050	2.16E-06	ALYM
rs2484594	12q24.21	115288601	T/C	0.032	*TBX3*	164694	0.448	0.091	8.46E-07	PC2
rs78120748	12q24.21	115289334	G/A	0.032	*TBX3*	165427	0.448	0.091	8.46E-07	PC2
rs34773022	12q24.31	123123414	A/G	0.246	*KNTC1*	12504	0.414	0.090	4.03E-06	HCT
rs6562681	13q21.33	72331984	C/T	0.339	*DACH1*	intron-variant	0.410	0.081	4.33E-07	HCT
rs10146835	14q11.2	22323046	T/C	0.433	*TRA@*	73	0.252	0.051	6.49E-07	WBC
rs35169499	14q23.1	59462088	A/G	0.312	*DAAM1*	-193323	0.159	0.035	4.73E-06	PC2
rs10136833	14q24.1	69401094	C/T	0.060	*ACTN1*	intron-variant	-13.240	2.655	6.34E-07	PLT
rs75665537	14q32.13	96024374	G/A	0.084	*GLRX5*	12831	0.317	0.067	2.35E-06	PC1
rs79282233	14q32.31	102241839	A/G	0.032	*PPP2R5C*	intron-variant	-1.620	0.327	7.51E-07	MCV
rs2414009	15q21.2	50354310	G/A	0.171	*ATP8B4*	intron-variant	0.299	0.064	3.68E-06	ALYM
rs77677780	15q21.3	56951904	T/C	0.109	*ZNF280D*	intron-variant	0.216	0.041	1.17E-07	PC4
rs77677780	15q21.3	56951904	T/C	0.109	*ZNF280D*	intron-variant	-0.201	0.041	8.10E-07	MCHC
rs16952730	16q12.2	54018921	A/G	0.293	*FTO*	intron-variant	-0.168	0.035	1.39E-06	PC3
rs3748387	16q22.1	69974546	C/T	0.332	*WWP2*	reference	0.317	0.069	4.95E-06	ANEU
rs4782985	16q24.1	84537527	A/C	0.044	*KIAA1609*	intron-variant	0.370	0.079	3.31E-06	PC2
rs4065321	17q21.1	38143548	C/T	0.445	*PSMD3*	intron-variant	0.278	0.049	1.56E-08	WBC
kgp6557113	17q21.1	38146264	G/A	0.367			0.150	0.032	2.35E-06	PC3
rs8066582	17q21.1	38146929	T/C	0.445	*PSMD3*	*intron-variant*	0.147	0.031	2.18E-06	PC3
rs8066582	17q21.1	38146929	T/C	0.445	*PSMD3*	*intron-variant*	0.313	0.064	9.78E-07	ANEU
rs72823022	17q21.2	40513732	C/T	0.080	*STAT3*	*intron-variant*	0.437	0.092	2.25E-06	WBC
rs77014629	17q22	50865595	C/T	0.035	*LOC100506650*	*-197385*	0.617	0.135	4.72E-06	ALYM
rs758596	17q23.2	59544863	T/C	0.268	*TBX4*	*intron-variant*	-6.644	1.450	4.75E-06	PLT
rs3760220	17q24.2	65713350	G/T	0.121	*LOC100507049*	*intron-variant*	-0.266	0.058	4.27E-06	PC1
rs412934	19p13.2	11405518	T/C	0.331	*TSPAN16*	*-1360*	0.126	0.027	4.44E-06	HGB
rs11085824	19p13.2	13001547	G/A	0.364	*GCDH*	*upstream-variant-2KB*	0.152	0.032	2.09E-06	PC3
rs11085824	19p13.2	13001547	G/A	0.364	*GCDH*	*upstream-variant-2KB*	0.200	0.038	1.44E-07	MCH
rs11085824	19p13.2	13001547	G/A	0.364	*GCDH*	*upstream-variant-2KB*	0.606	0.115	1.43E-07	MCV
rs2823126	21q21.1	16561704	A/G	0.029	*NRIP1*	*124597*	0.451	0.099	5.00E-06	PC2
rs855791	22q12.3	37462936	T/C	0.446	*TMPRSS6*	missense	-0.238	0.037	2.37E-10	MCH
rs855791	22q12.3	37462936	T/C	0.446	*TMPRSS6*	missense	0.181	0.033	4.05E-08	PC2
rs855791	22q12.3	37462936	T/C	0.446	*TMPRSS6*	missense	-0.560	0.114	8.25E-07	MCV

Table B5: Results of Factor Analyses for the First Three Factors for Mattieni *et al.*, 2010, the Current Study and HABC

	LLFS [Matteini 2010] (N = 3600)			**LLFS (N = 4472)**			**HABC (N = 1794)**		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
Cognition									
Animal recall	0.17	-0.07	0.56	0.18	-0.12	0.53	Few and different measures of cognition		
Vegetable recall	-0.14	-0.12	0.60	-0.13	-0.16	0.58			
Digit forward	0.06	0.04	0.46	0.03	0.05	0.42			
Digit back	0.04	0.03	0.56	0.04	0.06	0.51			
Immediate memory	0.00	0.01	0.78	0.01	0.04	0.80			
Delayed memory	0.01	-0.01	0.78	0.01	0.01	0.80			
Cardiovascular									
Presence of hypertension	-0.09	0.11	-0.07	-0.06	0.26	-0.09	-0.02	0.14	-0.14
Systolic BP (mm Hg)	-0.06	0.05	-0.06	-0.02	0.03	-0.08	-0.03	0.00	0.07
Diastolic BP (mm Hg)	0.14	-0.01	0.00	0.22	-0.01	-0.04	0.19	-0.15	0.20
Pulse pressure	-0.17	0.08	0.03	-0.18	0.04	-0.07	-0.16	0.10	-0.04
Total cholesterol (mg/dL)	-0.09	-0.14	-0.04	-0.08	-0.13	-0.04	-0.22	0.00	0.94
HDL cholesterol (mg/dL)	-0.29	-0.56	0.10	-0.26	-0.61	0.09	-0.46	-0.25	0.15
LDL cholesterol (mg/dL)	0.02	-0.07	-0.07	0.03	-0.03	-0.08	-0.01	0.03	0.92
Triglyceride (mg/dL)	0.05	0.52	-0.08	0.02	0.56	-0.05	-0.06	0.19	0.22
Metabolic									
Presence of diabetes	-0.17	0.59	0.02	-0.14	0.45	0.03	0.01	0.79	-0.08
BMI (kg/m^2)	0.20	0.66	0.00	0.02	0.74	0.04	0.11	0.11	0.02
Creatinine (mg/dL)	0.35	0.21	-0.16	0.31	0.27	-0.22	0.50	0.09	-0.08
Glucose (mg/dL)	-0.07	0.67	-0.01	-0.03	0.53	-0.04	0.14	0.82	0.00
Glycosylated hemoglobin (%)	-0.19	0.68	0.03	-0.19	0.56	0.02	0.01	0.82	0.05
Waist circumference (cm)	0.17	0.68	-0.08	0.19	0.77	-0.06	0.21	0.10	-0.04
Physical Activity									
Average grip strength (kg)	0.88	0.14	-0.02	0.88	0.16	-0.05	0.87	0.07	-0.14
Maximum grip strength (kg)	0.88	0.14	-0.02	0.88	0.17	-0.06	0.87	0.07	-0.15
Gait speed (m/sec)	0.42	-0.20	0.31	0.45	-0.25	0.28	0.44	-0.07	0.05
Total physical activity	0.42	-0.15	0.01	0.42	-0.21	0.27	0.40	0.03	0.04
Pulmonary									
FEV1/FEV6 (%)	0.10	0.07	-0.02	-0.14	0.05	-0.03	-0.18	-0.13	-0.14
FEV1	0.85	0.00	0.08	0.86	0.03	0.04	0.84	-0.08	0.11
FEV6	0.86	0.00	0.07	0.88	0.00	0.02	0.87	-0.03	0.05
Presence of lung disease	-0.15	0.10	0.05	0.07	0.10	0.09	0.07	-0.15	0.21

Table B6: Results of Factor Analysis (Four Factor Solution) for LLFS without the Cognition Domain

	RF1	**RF2**	**RF3**	**RF4**
Eigenvalue	3.96	3.21	2.50	1.86
% Variance explained	17.3	13.7	11.7	9.7
Cognition				
Animal recall	Not Included			
Vegetable recall				
Digit forward				
Digit backward				
Immediate memory				
Delayed memory				
Cardiovascular				
Presence of hypertension	-0.07	0.22	**0.70**	-0.10
Systolic BP	-0.02	0.02	**0.96**	0.09
Diastolic BP	0.22	-0.00	**0.66**	0.16
Pulse pressure	-0.18	0.03	**0.79**	-0.01
Total cholesterol	-0.08	-0.17	0.12	**0.94**
HDL cholesterol	-0.23	**-0.64**	0.05	0.09
LDL cholesterol	0.02	-0.06	0.08	**0.93**
Triglycerides	-0.00	**0.54**	0.10	**0.47**
Metabolic				
Presence of diabetes	-0.16	**0.45**	-0.01	-0.17
Estimated BMI	-0.01	**0.72**	0.12	0.15
Creatinine	0.28	0.31	-0.02	0.01
Glucose	-0.05	**0.53**	0.09	-0.03
Glycosylated hemoglobin	-0.21	**0.54**	0.02	-0.01
Waist Circumference	0.15	**0.78**	0.07	0.07
Physical Activity				
Average grip strength	**0.87**	0.22	0.01	-0.08
Maximum grip strength	**0.87**	0.22	0.01	-0.08
Gait speed	**0.49**	-0.27	-0.01	0.06
Total physical activity	**0.46**	-0.23	0.04	0.12
Pulmonary				
Presence of lung disease	-0.14	0.05	-0.02	-0.05
FEV1	**0.86**	0.07	-0.11	0.01
FEV6	**0.88**	0.05	-0.10	-0.03
FEV1/FEV6 ratio	0.08	0.09	-0.06	0.15

Table B7: Results of Factor Analysis for HABC (Four Factor Solution)

	RF1	RF2	RF3	RF4
Eigenvalue	4.25	2.91	2.07	1.85
% Variance explained	18.0	13.6	9.6	9.1
Cardiovascular				
Presence of hypertension	-0.04	0.26	**0.50**	-0.13
Systolic BP	-0.03	0.00	**0.97**	0.07
Diastolic BP	0.20	-0.01	**0.51**	0.23
Pulse pressure	-0.17	0.00	**0.79**	-0.07
Total cholesterol	-0.26	-0.03	0.04	**0.92**
HDL cholesterol	**-0.43**	**-0.48**	-0.02	0.13
LDL cholesterol	-0.05	0.02	-0.01	**0.91**
Triglycerides	-0.09	**0.44**	0.16	0.24
Metabolic				
Presence of diabetes	-0.08	**0.58**	0.00	-0.16
Estimated BMI	0.09	**0.66**	0.07	0.10
Creatinine	**0.49**	0.24	0.07	-0.05
Glucose	0.04	**0.76**	0.06	-0.06
Glycosylated hemoglobin	-0.09	**0.68**	-0.03	-0.03
Waist Circumference	0.20	**0.66**	0.05	0.04
Physical Activity				
Average grip strength	**0.86**	0.19	0.01	-0.12
Maximum grip strength	**0.86**	0.19	0.00	-0.12
Gait speed	**0.45**	-0.25	-0.01	0.03
Total physical activity	**0.39**	-0.22	-0.01	0.01
Pulmonary				
Presence of lung disease	-0.19	0.11	-0.01	-0.15
FEV1	**0.84**	0.03	-0.08	0.15
FEV6	**0.87**	0.03	-0.09	0.08
FEV1/FEV6 ratio	0.08	0.00	0.01	0.24

Table B8: Complete List of Variants (p-value < 5×10^{-6}) for F1 for LLFS

SNP	Region	Position	minor/major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs17266628	3q21.1	122127926	G/A	0.216	*FAM162A*	intron-variant	-0.325	0.071	4.31E-06
c3_122199493_INDEL	3q21.1	122199493	D/R	0.044			-0.834	0.179	3.43E-06
rs4626276	4q25	111649989	C/A	0.132	*PITX2*	86507	-0.394	0.086	4.85E-06
rs1906610	4q25	111665917	A/G	0.118	*PITX2*	102435	-0.430	0.092	3.14E-06
rs60409120	4q25	111677395	T/C	0.128	*MIR297*	-104351	-0.402	0.088	4.91E-06
rs10516563	4q25	111677722	G/T	0.126	*MIR297*	-104024	-0.404	0.088	4.61E-06
rs142641595	4q25	111681539	A/G	0.111	*MIR297*	-100207	-0.448	0.097	4.23E-06
rs79687642	4q25	111682614	G/T	0.124	*MIR297*	-99132	-0.410	0.089	4.27E-06
rs144691425	4q25	111683003	C/T	0.122	*MIR297*	-98743	-0.426	0.091	2.68E-06
rs4833443	4q25	111684643	T/C	0.124	*MIR297*	-97103	-0.411	0.089	3.79E-06
rs643154	5q23.2	125243687	A/G	0.405	*GRAMD3*	-452146	0.274	0.059	3.59E-06
c5_125252828_INDEL	5q23.2	125252828	R/D	0.404			0.275	0.059	3.14E-06
rs192645244	5q23.2	125256696	C/T	0.404	*GRAMD3*	-439137	0.281	0.060	2.72E-06
rs451573	5q23.2	125263605	C/T	0.404	*GRAMD3*	-432228	0.276	0.059	2.98E-06
rs465236	5q23.2	125273245	G/C	0.405	*GRAMD3*	-422588	0.279	0.059	2.25E-06
c5_125274052_INDEL	5q23.2	125274052	R/I	0.412			0.290	0.059	1.03E-06
kgp5641805	5q23.2	125274062	T/C	0.403			0.277	0.059	2.63E-06
rs445513	5q23.2	125274830	G/A	0.405	*GRAMD3*	-421003	0.277	0.059	2.71E-06
rs432924	5q23.2	125276015	C/T	0.405	*GRAMD3*	-419818	0.276	0.059	3.04E-06
rs412655	5q23.2	125286928	T/G	0.405	*GRAMD3*	-408905	0.275	0.059	3.31E-06
c10_3790907_INDEL	10p15.2	3790907	D/R	0.327			0.375	0.066	1.72E-08

Table B8 continued

SNP	Region	Position	minor/major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs7085102	10p15.2	3794279	T/G	0.366	*KLF6*	-23410	0.338	0.061	3.76E-08
rs11252070	10p15.2	3795085	T/G	0.314	*KLF6*	-22604	0.317	0.064	6.68E-07
rs7089867	10p15.2	3795187	A/G	0.364	*KLF6*	-22502	0.332	0.061	6.14E-08
rs7090260	10p15.2	3795522	A/G	0.364	*KLF6*	-22167	0.332	0.061	6.15E-08
rs7918940	10p15.2	3796222	G/C	0.317	*KLF6*	-21467	0.318	0.064	6.32E-07
rs7896849	10p15.2	3798495	A/C	0.364	*KLF6*	-19194	0.335	0.061	4.21E-08
rs2171301	10p15.2	3799730	T/G	0.372	*KLF6*	-17959	0.331	0.061	5.53E-08
rs3829199	10p15.1	3801384	T/C	0.382	*KLF6*	-16305	0.310	0.061	3.88E-07
rs3750859	10p15.1	3801549	C/T	0.382	*KLF6*	-16140	0.309	0.061	4.00E-07
rs2279417	10p15.1	3802112	G/C	0.381	*KLF6*	-15577	0.293	0.061	1.87E-06
rs2279419	10p15.1	3802761	C/T	0.386	*KLF6*	-14928	0.304	0.061	6.05E-07
rs10795073	10p15.1	3806127	C/T	0.371	*KLF6*	-11562	0.312	0.060	2.34E-07
rs11252075	10p15.1	3806700	C/T	0.384	*KLF6*	-10989	0.307	0.061	4.86E-07
rs11252076	10p15.1	3807145	C/T	0.373	*KLF6*	-10544	0.310	0.060	2.48E-07
rs4242761	10p15.1	3807517	G/A	0.372	*KLF6*	-10172	0.310	0.060	2.54E-07
c10_3813216_INDEL	10p15.1	3813216	D/R	0.301			0.305	0.064	1.70E-06
c10_3813218_INDEL	10p15.1	3813218	D/R	0.308			0.300	0.063	2.15E-06
rs1906143	10p15.1	3815373	T/C	0.305	*KLF6*	-2316	0.305	0.063	1.35E-06
rs2279414	10p15.1	3818058	G/A	0.306	*KLF6*	downstream-variant-500B	0.293	0.063	3.24E-06
rs988667	11p15.3	12010581	C/T	0.096	*DKK3*	intron-variant	0.474	0.098	1.51E-06
rs1958682	14q12	25598386	A/G	0.128	*STXBP6*	79467	0.404	0.088	4.70E-06
rs1241492	14q12	25605964	T/C	0.159	*STXBP6*	87045	0.366	0.080	4.91E-06
rs4548961	18q11.2	22972726	A/C	0.118	*ZNF521*	40746	0.446	0.092	1.35E-06
rs10445494	18q11.2	22974335	A/G	0.117	*ZNF521*	42355	0.447	0.092	1.28E-06
rs7237853	18q11.2	22984635	T/C	0.147	*ZNF521*	52655	0.416	0.084	6.45E-07
rs11083124	18q11.2	22987297	C/A	0.252	*ZNF521*	55317	0.306	0.067	4.87E-06
rs7244729	18q11.2	22989014	T/G	0.146	*ZNF521*	57034	0.414	0.084	7.40E-07
rs10853653	18q11.2	22989604	T/C	0.147	*ZNF521*	57624	0.413	0.083	7.69E-07

Table B9: Complete List of Variants (*p*-value < 5×10^{-6}) for F2 for LLFS

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs12088087	1p34.1	45170012	G/A	0.390	*C1orf228*	intron-variant	0.276	0.060	4.76E-06
rs56164117	1p34.1	45173351	C/T	0.393	*C1orf228*	intron-variant	0.287	0.061	3.10E-06
rs9830791	3p14.3	57091575	A/G	0.479	*ARHGEF3*	intron-variant	-0.267	0.058	4.91E-06
rs765468	9q21.13	78906740	G/T	0.014	*PCSK5*	intron-variant	1.176	0.239	8.47E-07
rs1674898	10q26.2	127733726	C/A	0.302	*ADAM12*	intron-variant	0.301	0.065	3.57E-06
rs1531331	10q26.2	127734012	T/G	0.304	*ADAM12*	intron-variant	0.302	0.065	3.01E-06
rs1710313	10q26.2	127734513	C/A	0.305	*ADAM12*	intron-variant	0.302	0.065	3.19E-06
rs1710315	10q26.2	127735048	T/C	0.304	*ADAM12*	intron-variant	0.302	0.065	3.06E-06
rs148460957	11p12	37402851	A/G	0.023	*C11orf74*	722051	0.938	0.197	2.05E-06
rs11034117	11p12	37414156	T/C	0.026	*C11orf74*	733356	0.880	0.189	3.38E-06
c11_37419207_INDEL	11NA	37419207	I/R	0.026			0.874	0.189	3.86E-06
rs11034125	11p12	37422961	T/C	0.025	*C11orf74*	742161	0.876	0.190	3.93E-06
rs11034126	11p12	37423106	A/G	0.027	*C11orf74*	742306	0.923	0.186	7.09E-07
rs1916074	11p12	37434719	A/C	0.027	*C11orf74*	753919	0.916	0.186	8.43E-07
rs80354775	11p12	37439843	A/G	0.026	*C11orf74*	759043	0.868	0.189	4.49E-06
rs10836742	11p12	37492760	T/A	0.028	*C11orf74*	811960	0.928	0.189	8.91E-07
rs12102869	16p12.3	19918987	C/T	0.155	*GPRC5B*	23043	-0.400	0.081	8.47E-07
rs9921401	16p12.3	19919338	G/C	0.154	*GPRC5B*	23394	-0.400	0.081	9.19E-07
rs9921480	16p12.3	19919440	G/C	0.154	*GPRC5B*	23496	-0.399	0.081	9.50E-07
rs3885610	16p12.3	19923044	C/T	0.154	*GPRC5B*	27100	-0.391	0.081	1.71E-06
rs28482811	16p12.3	19925612	C/T	0.154	*GPRC5B*	29668	-0.381	0.081	3.00E-06
c16_19925837_INDEL	16p12.3	19925837	D/R	0.155			-0.378	0.082	3.65E-06
rs9926784	16p12.3	19941968	C/T	0.170	*GPRC5B*	46024	-0.358	0.078	4.84E-06

Table B9 continued

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs4613074	16p12.3	19942527	C/T	0.171	*GPRC5B*	46583	-0.360	0.078	4.39E-06
rs7205054	16p12.3	19943026	G/C	0.170	*GPRC5B*	47082	-0.359	0.078	4.55E-06
rs11648621	16p12.3	19973008	G/A	0.197	*GPR139*	-70043	-0.360	0.074	1.22E-06
rs9646255	16p12.3	19975601	A/T	0.192	*GPR139*	-67450	-0.353	0.075	2.60E-06
rs203550	20p13	1194416	A/C	0.402	*C20orf202*	5655	-0.317	0.062	2.79E-07
rs203549	20p13	1194648	C/A	0.402	*C20orf202*	5887	-0.317	0.062	2.75E-07
c20_1194678_ INDEL	20p13	1194678	I/R	0.395			-0.324	0.062	2.09E-07
rs203548	20p13	1195014	G/A	0.401	*C20orf202*	6253	-0.317	0.062	2.69E-07
rs203547	20p13	1195245	T/A	0.482	*C20orf202*	6484	0.294	0.060	1.04E-06
rs203546	20p13	1195501	A/G	0.401	*C20orf202*	6740	-0.317	0.061	2.58E-07
rs203545	20p13	1195688	C/G	0.401	*C20orf202*	6927	-0.317	0.061	2.56E-07
c20_1195705_ INDEL	20p13	1195705	D/R	0.401			-0.317	0.062	2.57E-07
rs203544	20p13	1195784	G/A	0.401	*C20orf202*	7023	-0.314	0.061	3.50E-07
rs203543	20p13	1195808	G/C	0.401	*C20orf202*	7047	-0.317	0.061	2.59E-07
rs203541	20p13	1196943	C/G	0.336	*C20orf202*	8182	-0.310	0.063	1.00E-06
rs203538	20p13	1197487	T/C	0.498	*C20orf202*	8726	0.284	0.061	3.08E-06
rs203537	20p13	1198599	T/C	0.484	*RAD21L1*	-8261	0.291	0.060	1.44E-06
rs182193	20p13	1199205	G/A	0.467	*RAD21L1*	-7655	-0.297	0.062	1.89E-06
rs1090517	20p13	1199421	T/A	0.401	*RAD21L1*	-7439	-0.312	0.062	4.43E-07
rs1090516	20p13	1199487	A/G	0.483	*RAD21L1*	-7373	0.290	0.060	1.50E-06
rs1090515	20p13	1199566	A/G	0.441	*RAD21L1*	-7294	-0.314	0.066	1.78E-06
rs1090513	20p13	1200581	G/T	0.491	*RAD21L1*	-6279	0.282	0.061	3.62E-06
rs1090512	20p13	1200909	A/G	0.494	*RAD21L1*	-5951	0.281	0.060	3.31E-06
rs1090511	20p13	1201771	G/A	0.484	*RAD21L1*	-5089	0.285	0.060	2.21E-06
rs1090510	20p13	1202366	T/C	0.483	*RAD21L1*	-4494	0.278	0.060	3.83E-06
rs430731	20p13	1205093	T/C	0.401	*RAD21L1*	-1767	-0.300	0.062	1.23E-06

Table B10: Results of GWA Analyses for RF1 (p-value < 5 x 10^{-6}) for LLFS (Without Cognition; Four Factor Solution)

SNP	Region	Position	minor/ major allele	MAF	Nearby gene	Position nearby gene	Beta	SE	*p*-value
rs2016469	3q13.13	108023965	T/C	0.361	*HHLA2*	intronic	0.278	0.060	3.66E-06
kgp5641805	5q23.2	125274062	T/C	0.403		NA	0.273	0.059	3.38E-06
rs4740660	9p24.3	227554	G/C	0.144	*DOCK8*	intronic	-0.373	0.082	4.93E-06
rs7896849	10p15.2	3798495	A/C	0.364	*KLF6*	-19194	0.330	0.061	5.45E-08
rs2279414	10p15.1	3818058	G/A	0.306	*KLF6*	DV-500B	0.288	0.062	4.08E-06
rs988667	11p15.3	12010581	C/T	0.096	*DKK3*	intronic	0.460	0.098	2.57E-06
rs1958682	14q12	25598386	A/G	0.128	*STXBP6*	79467	0.401	0.088	4.88E-06
rs4548961	18q11.2	22972726	A/C	0.118	*ZNF521*	40746	0.445	0.092	1.23E-06
rs11083124	18q11.2	22987297	C/A	0.252	*ZNF521*	55317	0.307	0.067	4.02E-06
rs7240975	18q11.2	22989234	A/G	0.176	*ZNF521*	57254	0.355	0.077	4.37E-06
rs10853653	18q11.2	22989604	T/C	0.147	*ZNF521*	57624	0.414	0.083	6.18E-07

Table B11: Results of GWA Analyses for RF2 (p < 5 x 10^{-6}) for LLFS (Without Cognition; Four Factor Solution)

SNP	Region	Position	minor/ major allele	MAF	Nearby Gene	Position Near Gene	Beta	SE	*p*-value
rs12088087	1p34.1	45170012	G/A	0.390	*C1orf228*	intron-variant	0.274	0.060	4.25E-06
rs765468	9q21.13	78906740	G/T	0.014	*PCSK5*	intron-variant	1.157	0.236	9.58E-07
rs1710313	10q26.2	127734513	C/A	0.305	*ADAM12*	intron-variant	0.299	0.064	3.00E-06
rs80354775	11p12	37439843	A/G	0.026	*C11orf74*	759043	0.881	0.187	2.42E-06
rs12102869	16p12.3	19918987	C/T	0.155	*GPRC5B*	23043	-0.404	0.080	4.66E-07
rs9926784	16p12.3	19941968	C/T	0.170	*GPRC5B*	46024	-0.362	0.077	2.81E-06
rs11648621	16p12.3	19973008	G/A	0.197	*GPR139*	-70043	-0.362	0.073	7.81E-07
rs203544	20p13	1195784	G/A	0.401	*C20orf202*	7023	-0.315	0.061	2.15E-07

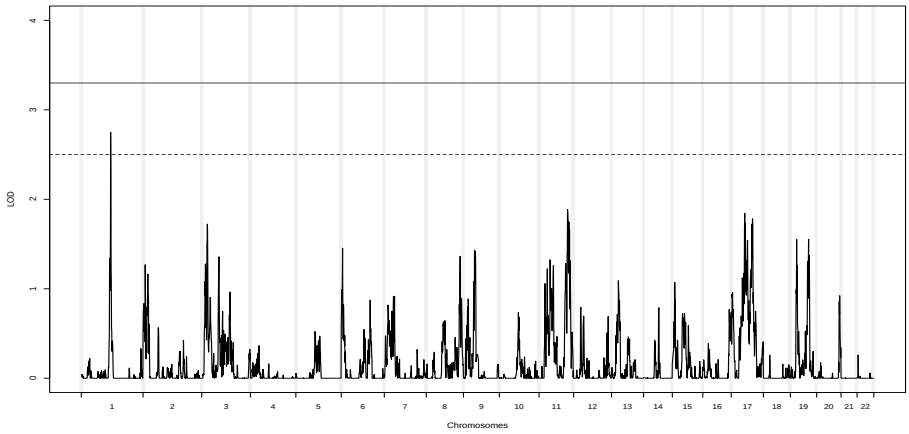

Figure B1: HGB Linkage Plot

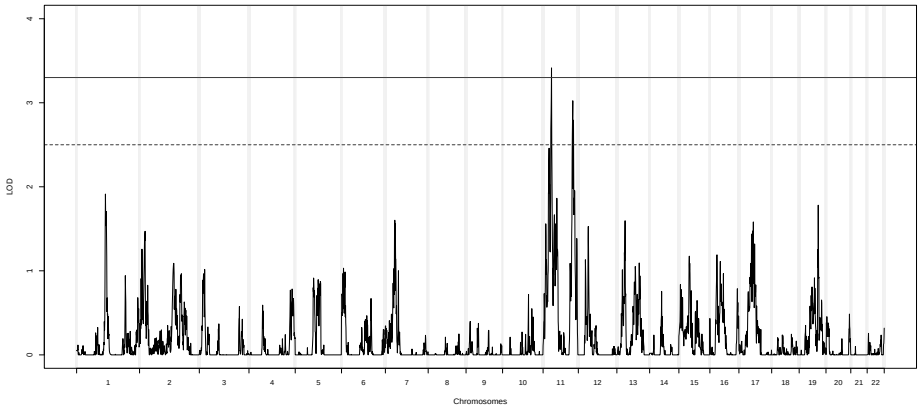

Figure B2: RBC Linkage Plot

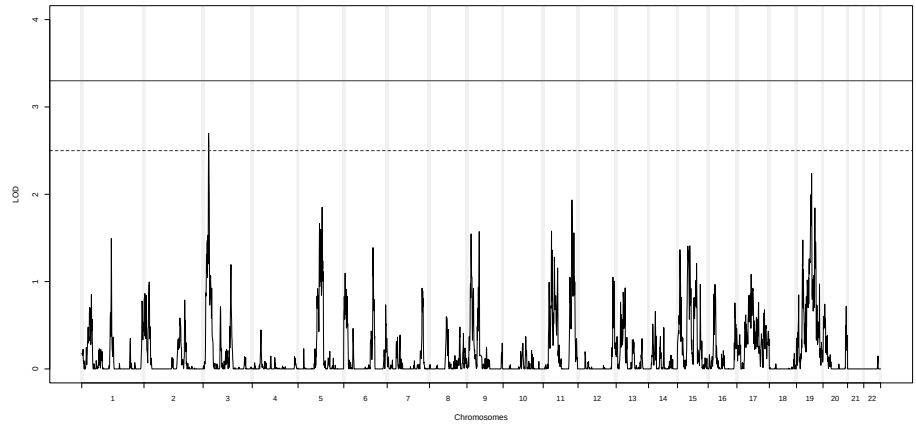

Figure B3: HCT Linkage Plot

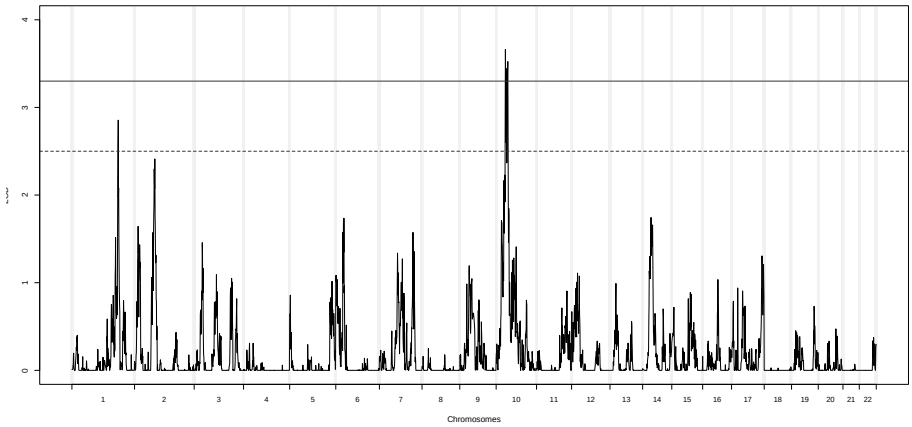

Figure B4: MCHC Linkage Plot

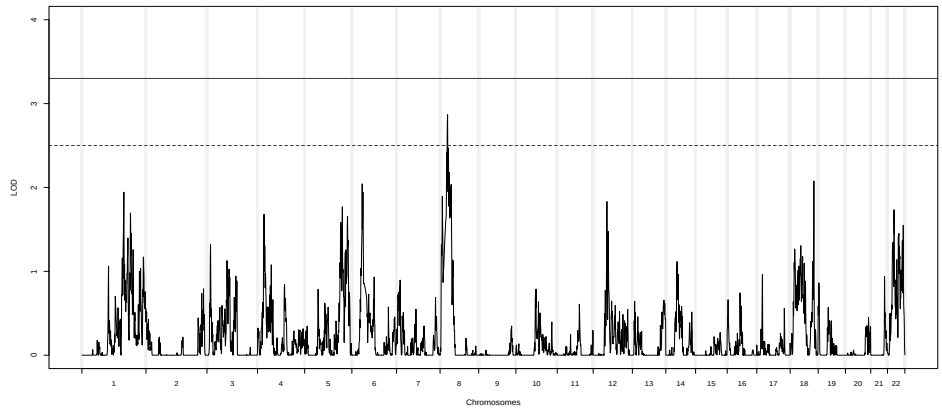

Figure B5: PLT Linkage Plot

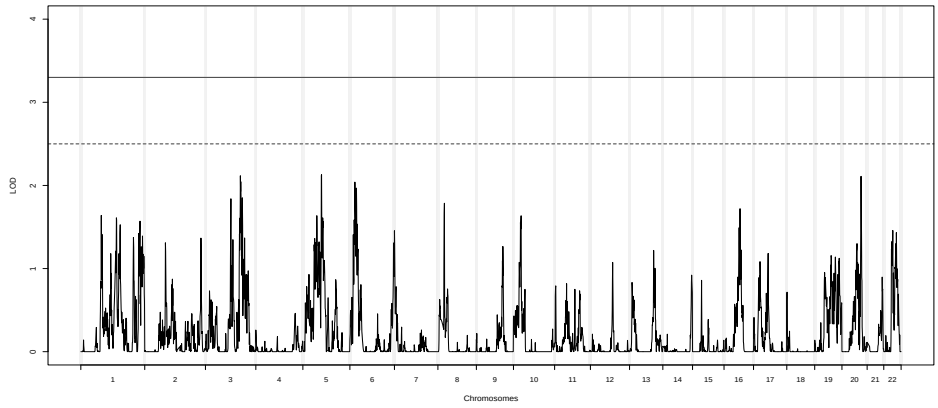

Figure B6: MCV Linkage Plot

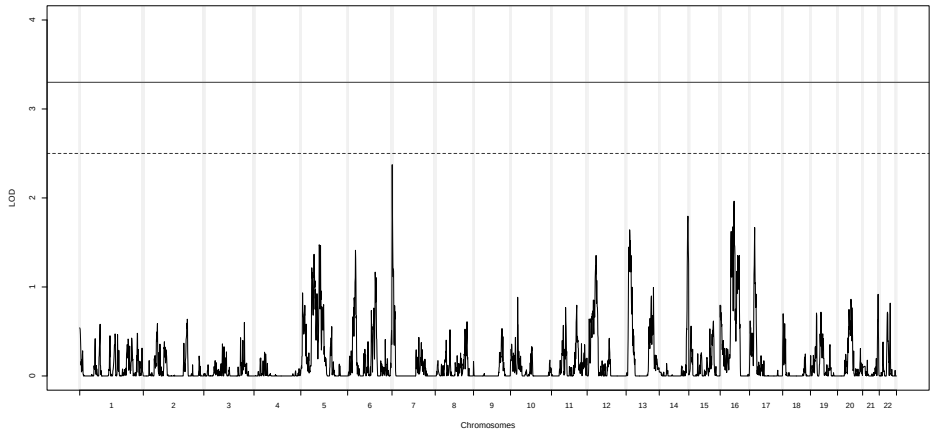

Figure B7: MCH Linkage Plot

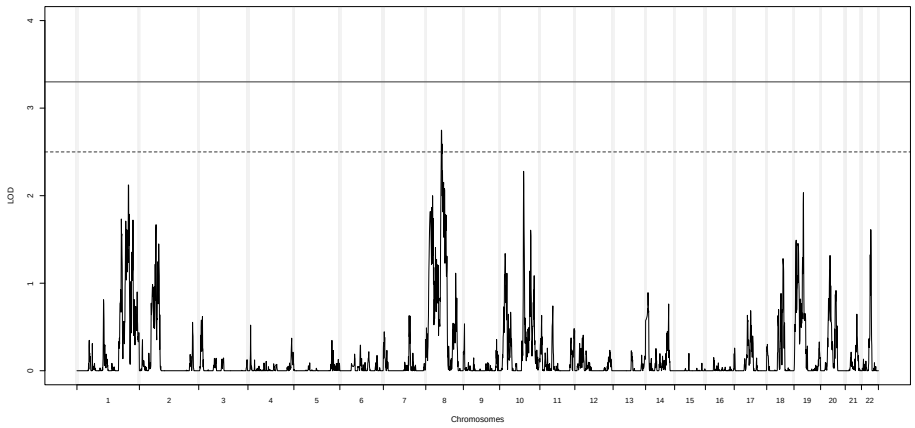

Figure B8: WBC Linkage Plot

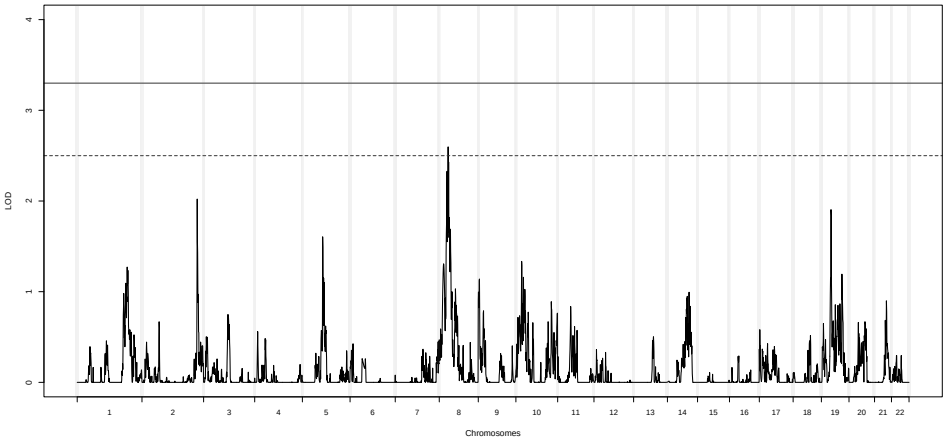

Figure B9: ANEU Linkage Plot

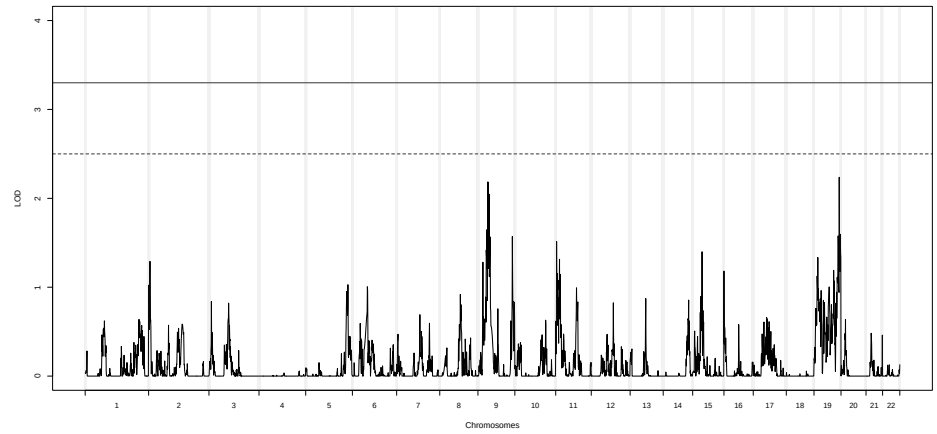

Figure B10: ALYM Linkage Plot

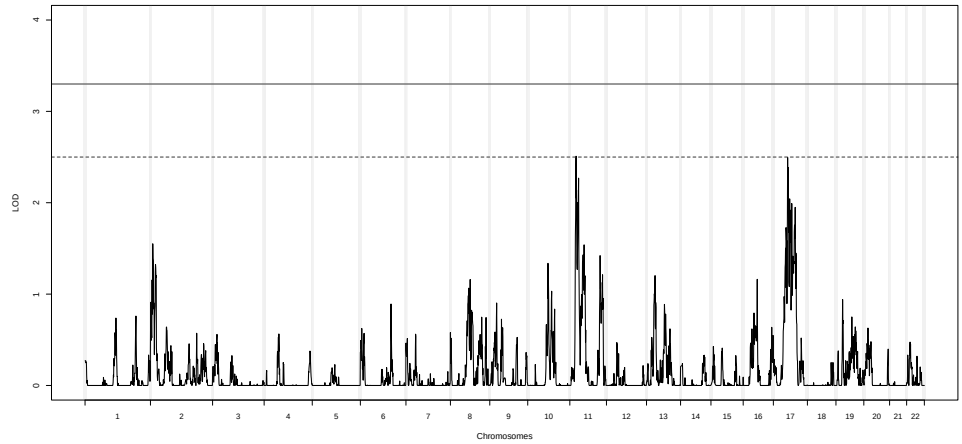

Figure B11: PC1 Linkage Plot

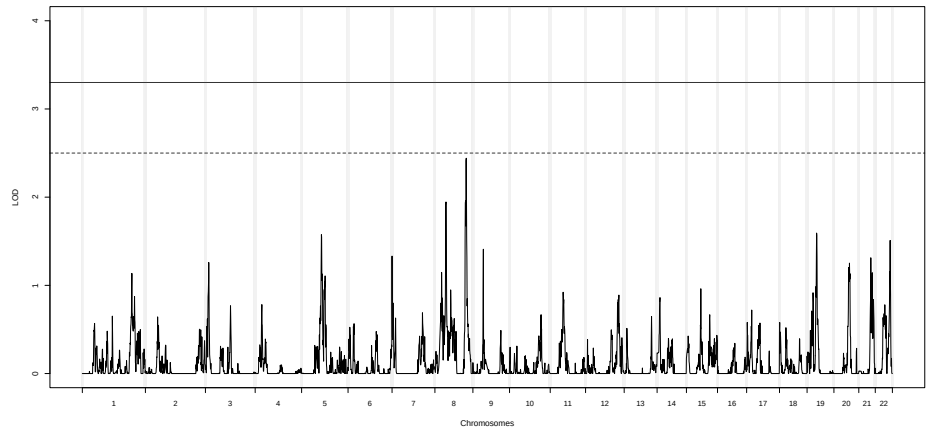

Figure B12: PC2 Linkage Plot

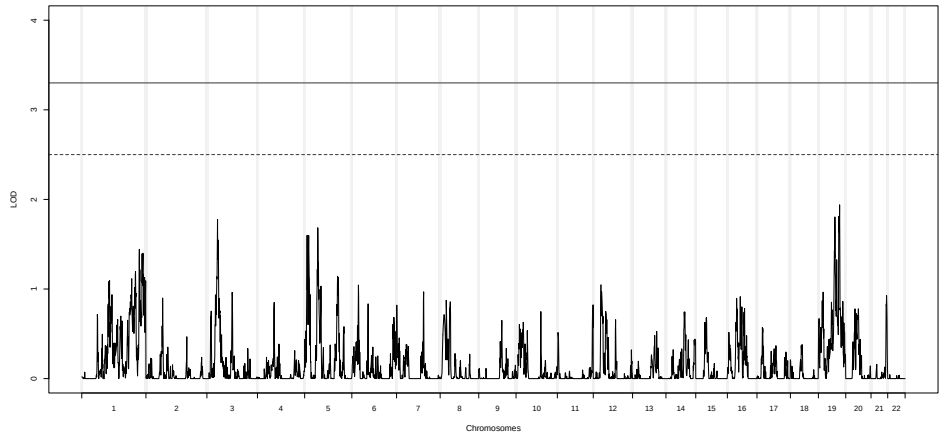

Figure B13: PC3 Linkage Plot

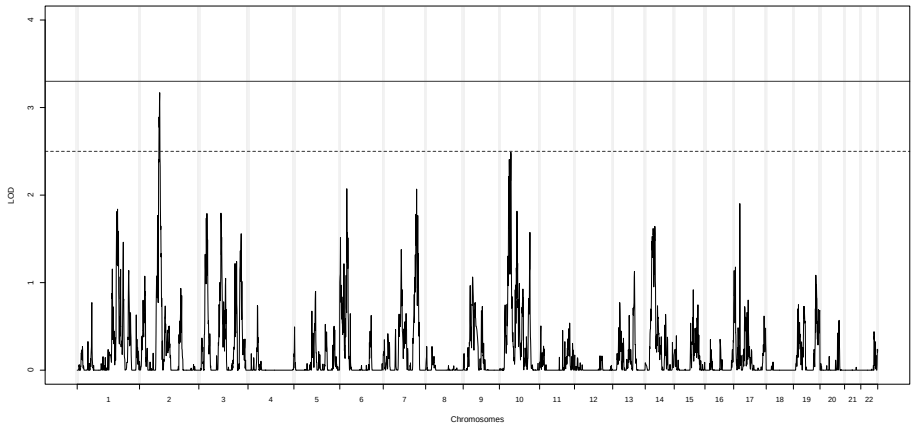

Figure B14: PC4 Linkage Plot

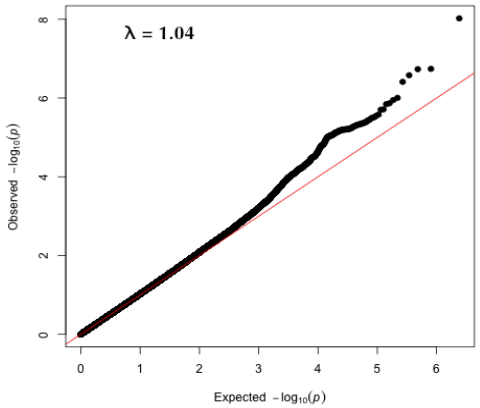

Figure B15: Q-Q Plot RBC

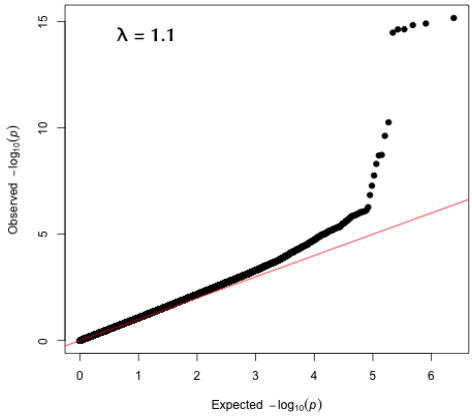

Figure B16: Q-Q Plot MCH

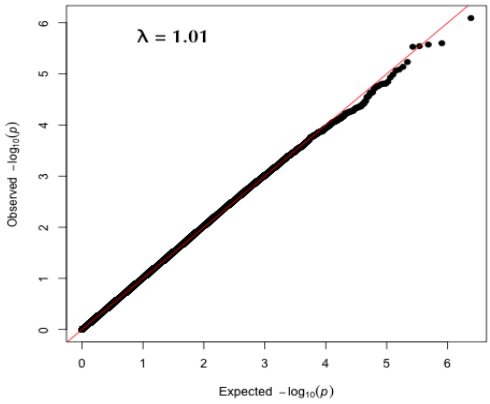

Figure B17: Q-Q Plot MCHC

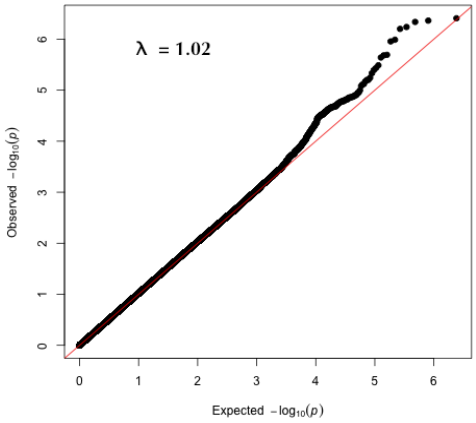

Figure B18: Q-Q Plot HCT

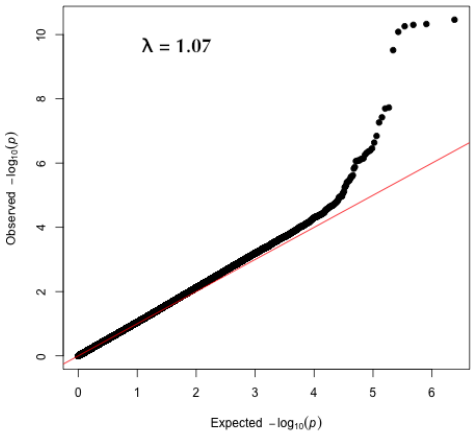

Figure B19: Q-Q Plot MCV

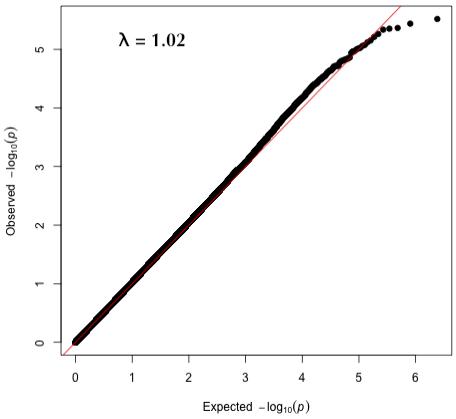

Figure B20: Q-Q Plot HGB

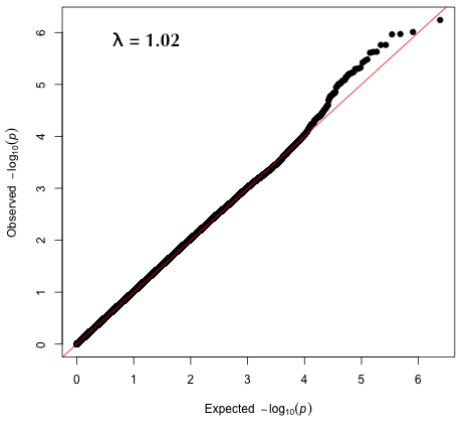

Figure B21: Q-Q Plot NEUT

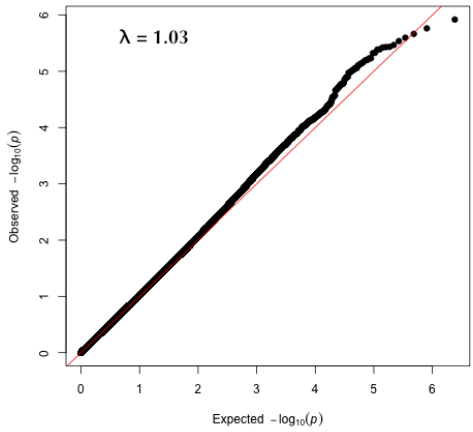

Figure B22: Q-Q Plot ALYM

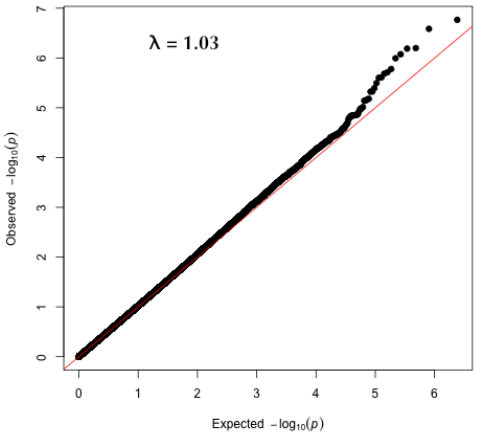

Figure B23: Q-Q Plot PLT

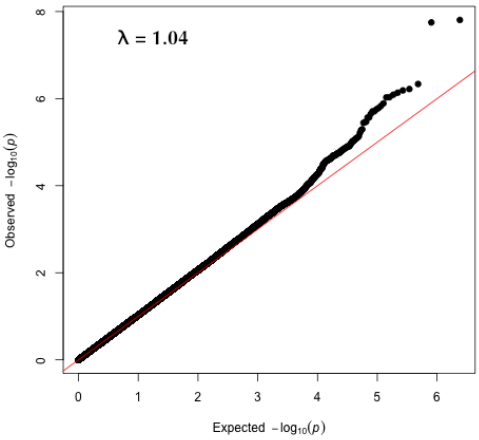

Figure B24: Q-Q Plot WBC

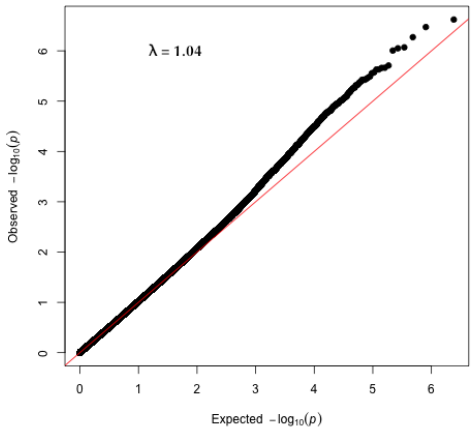

Figure B25: Q-Q Plot PC1

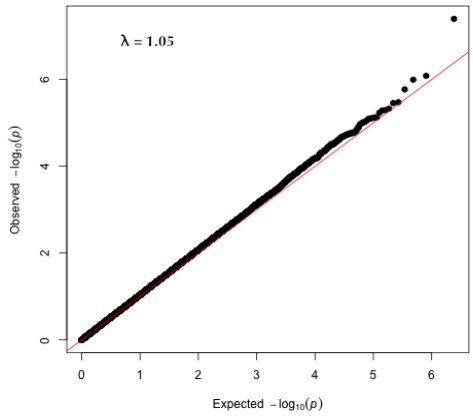

Figure B26: Q-Q Plot PC2

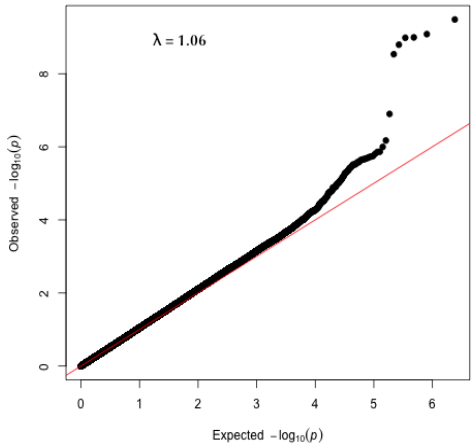

Figure B27: Q-Q Plot PC3

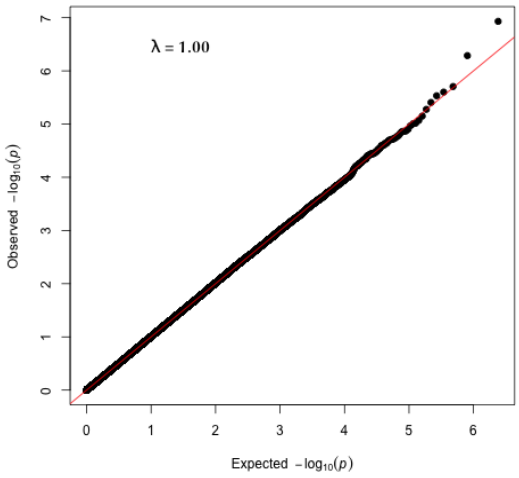

Figure B28: Q-Q Plot PC4

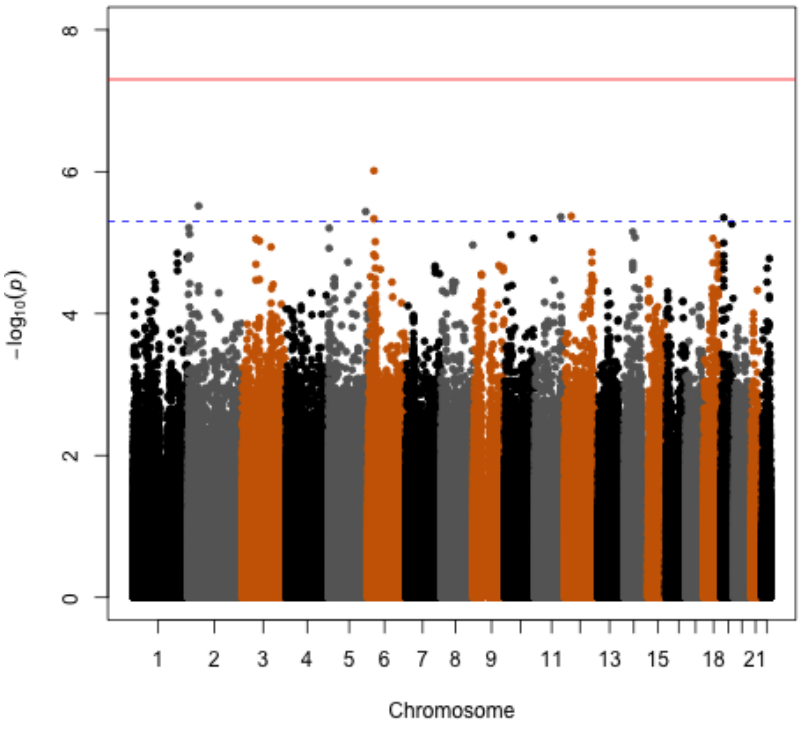

Figure B29: Manhattan Plot HGB

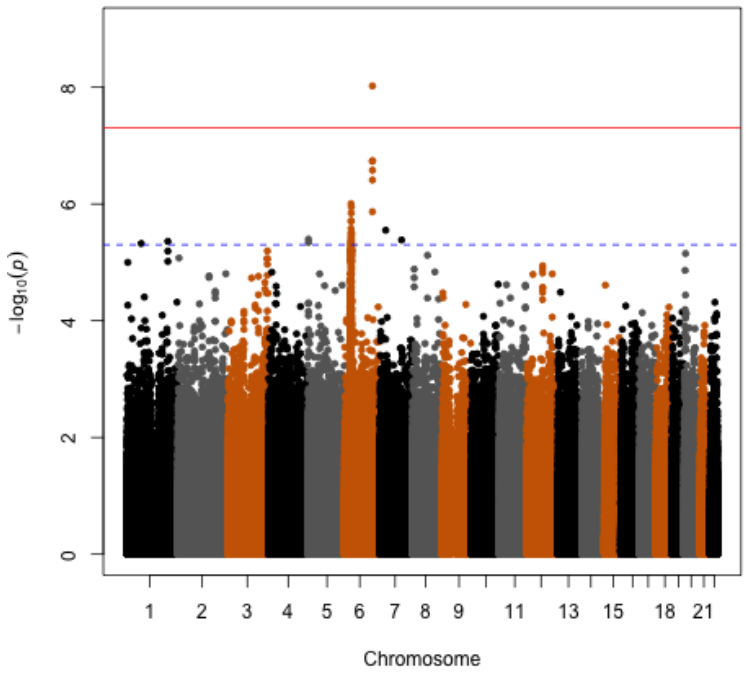

Figure B30: Manhattan Plot RBC

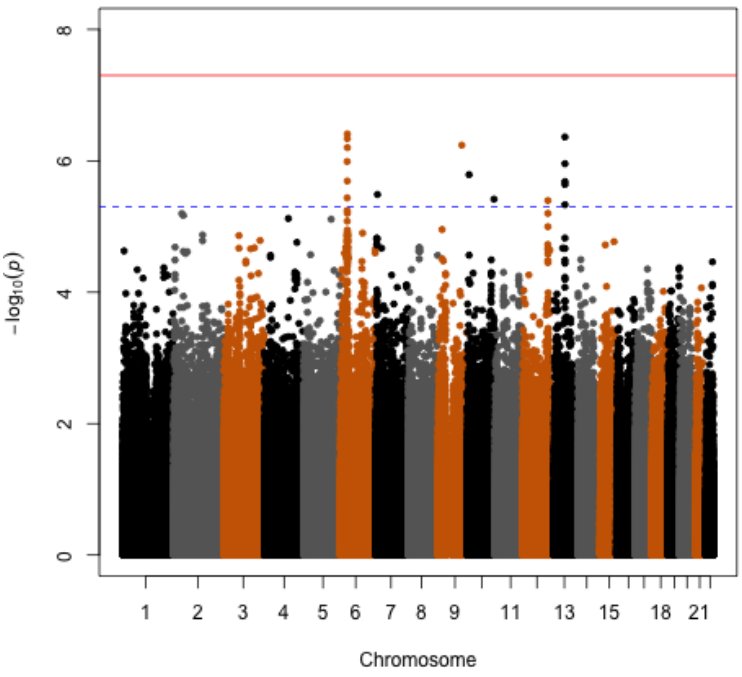

Figure B31: Manhattan Plot HCT

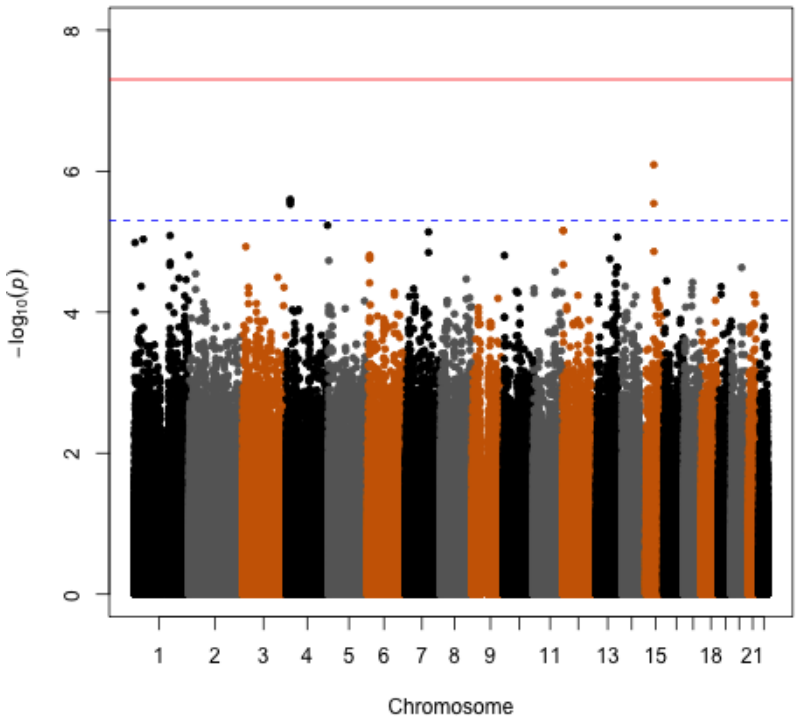

Figure B32: Manhattan Plot MCHC

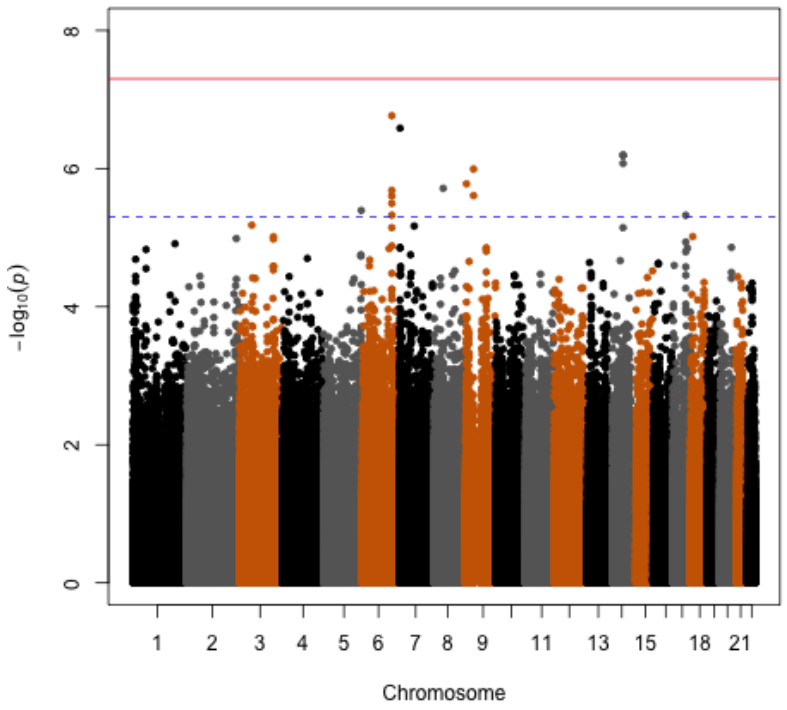

Figure B33: Manhattan Plot PLT

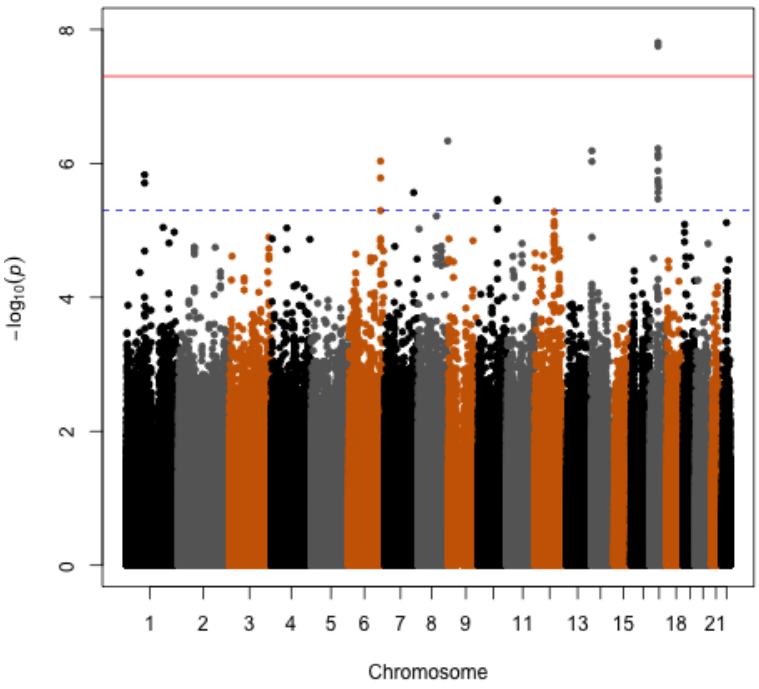

Figure B34: Manhattan Plot WBC

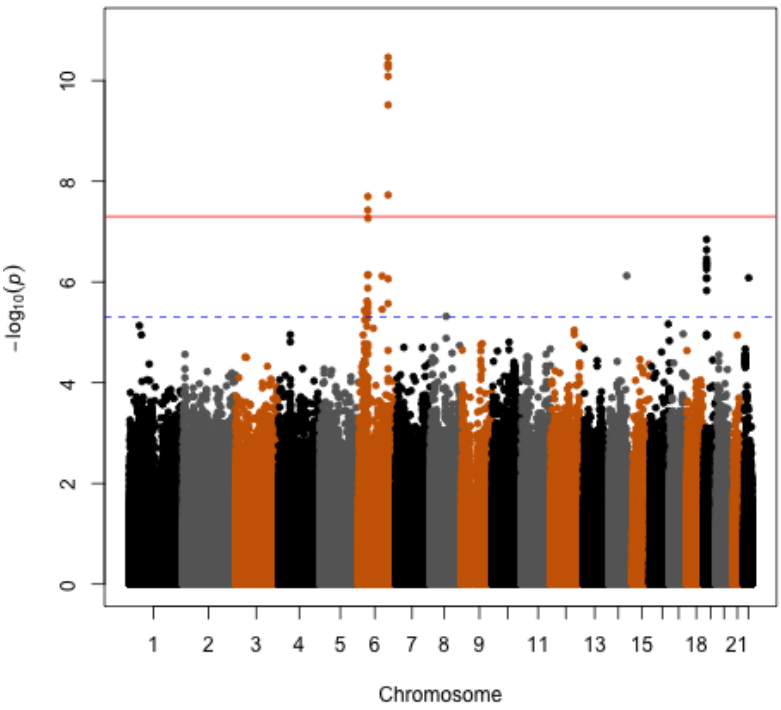

Figure B35: Manhattan Plot MCV

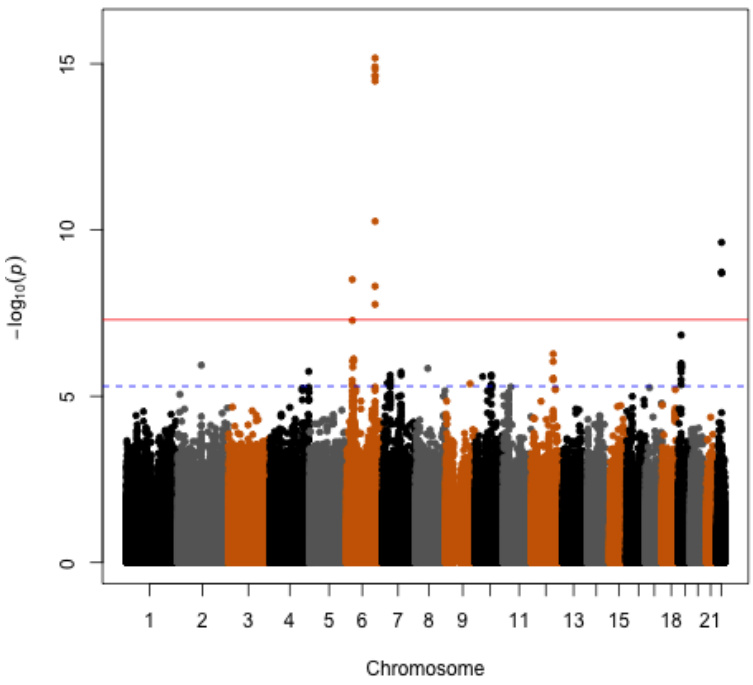

Figure B36: Manhattan Plot MCH

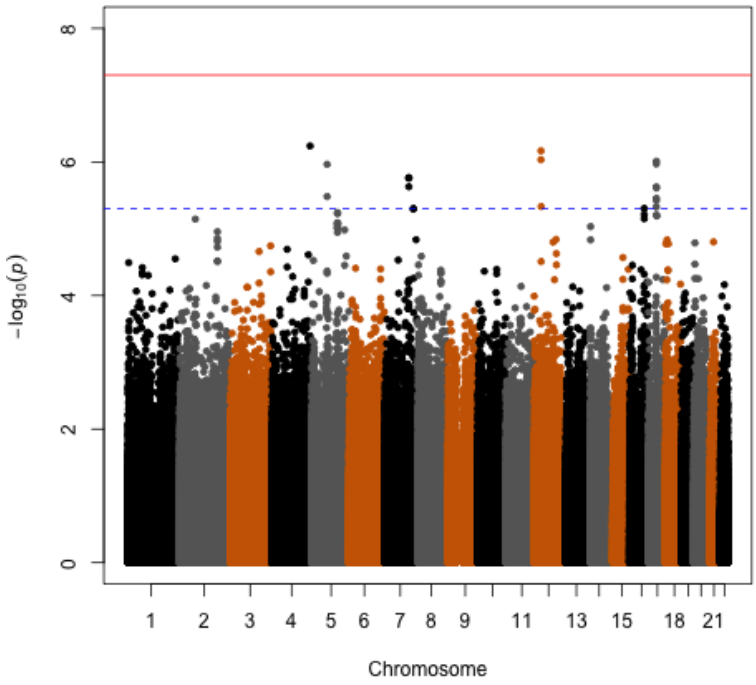

Figure B37: Manhattan Plot ANEU

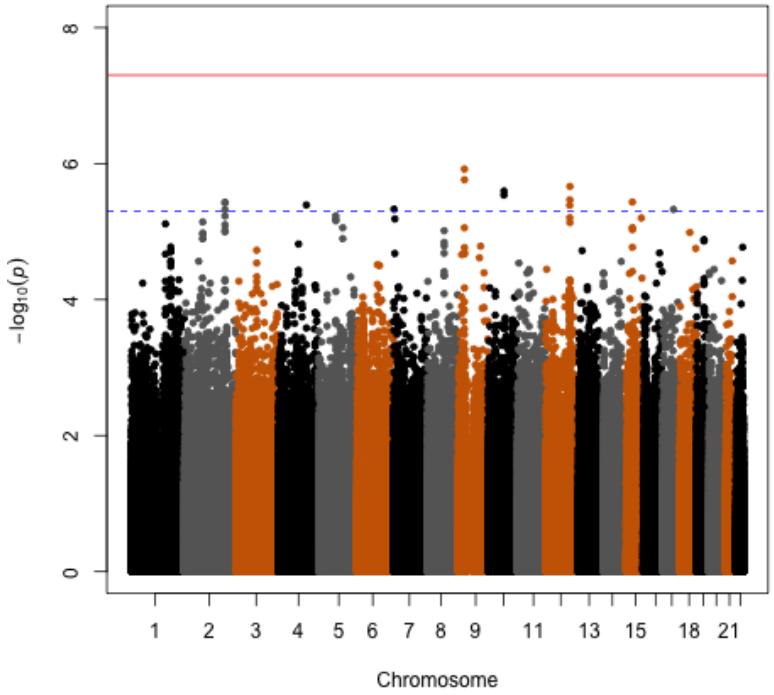

Figure B38: Manhattan Plot ALYM

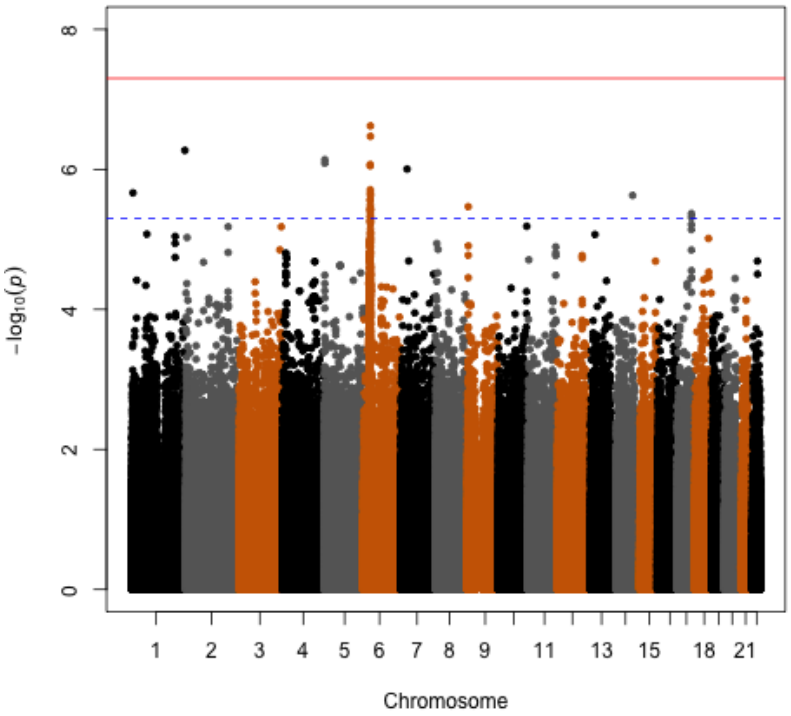

Figure B39: Manhattan Plot PC1

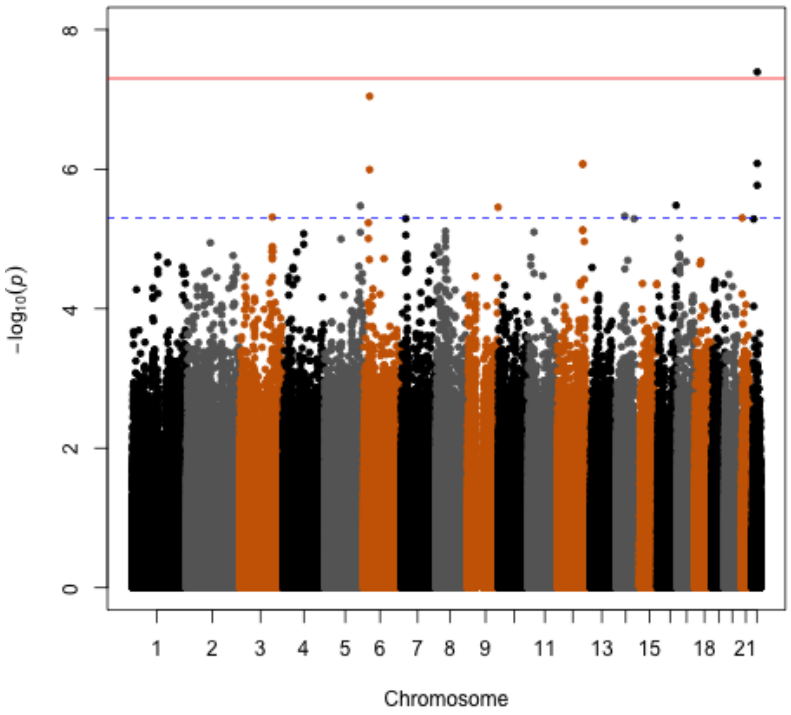

Figure B40: Manhattan Plot PC2

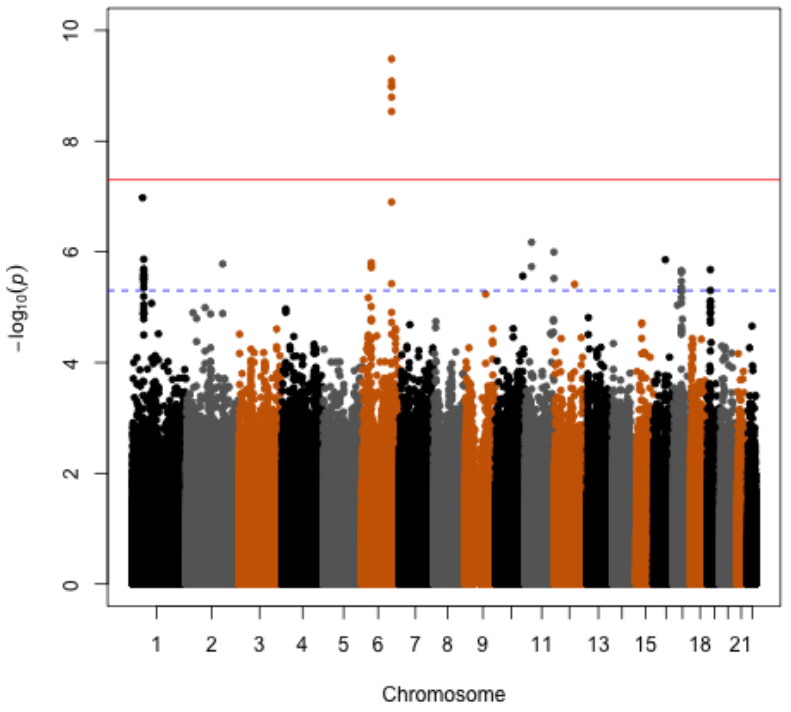

Figure B41: Manhattan Plot PC3

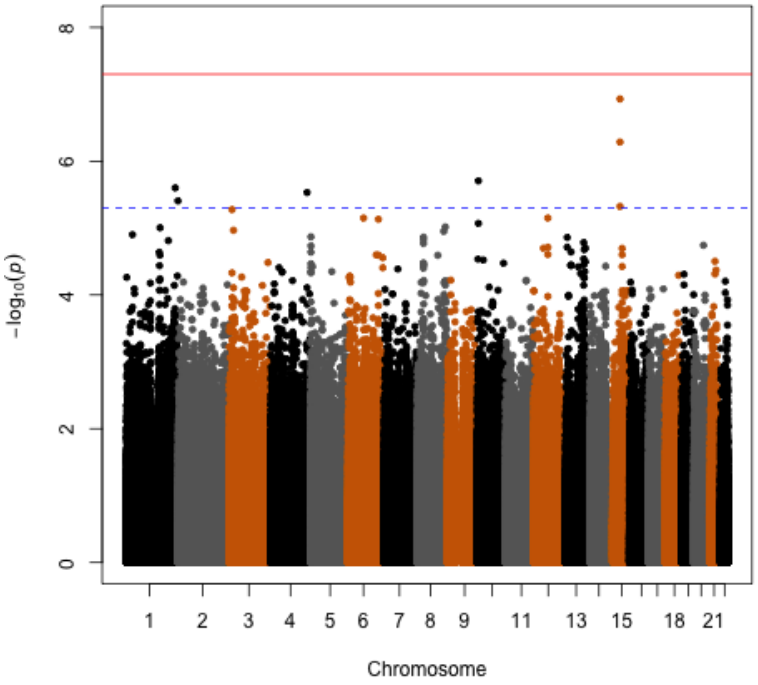

Figure B42: Manhattan Plot PC4

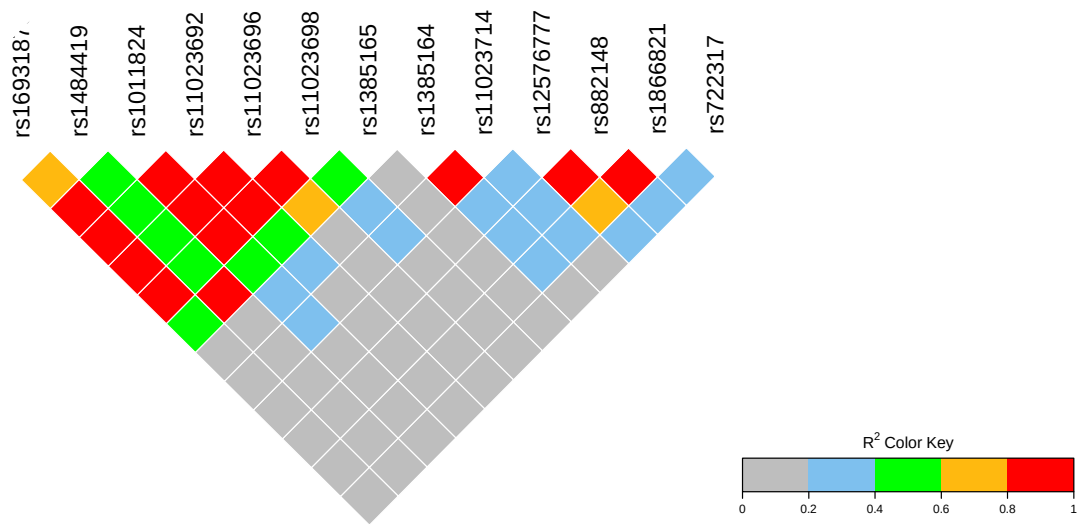

Figure B43: LD Plot for Two-point Linkage SNPs for Peak at 11p15.2 (LOD > 2.5) for RBC Count

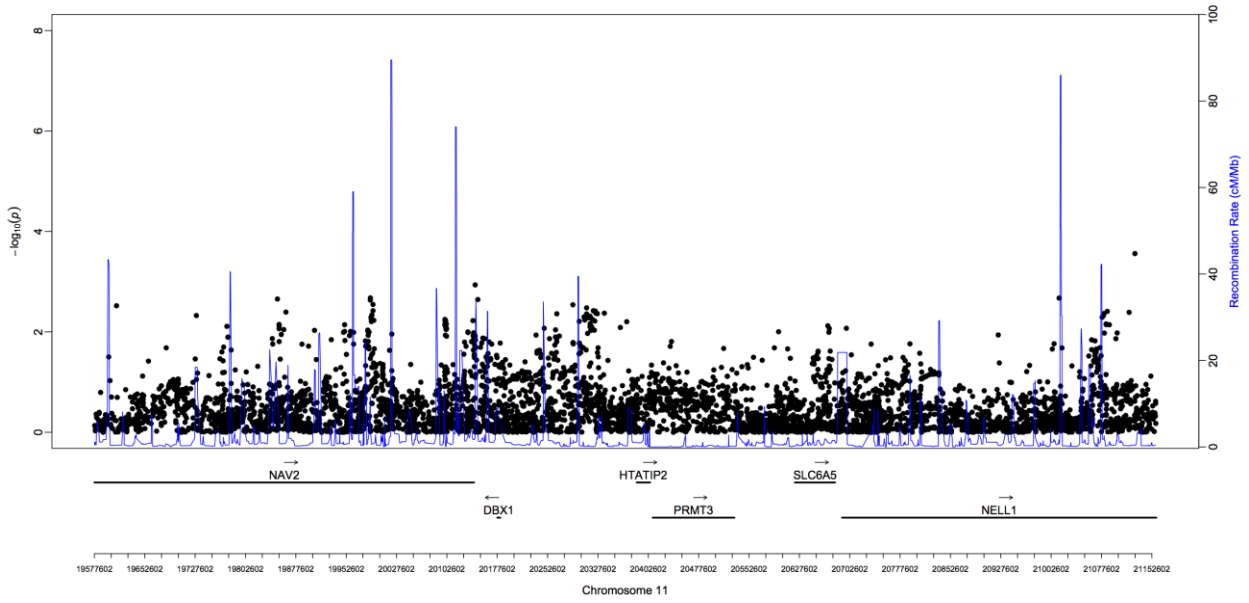

Figure B44: Association Analysis for Peak at 11p15.1 for RBC Count

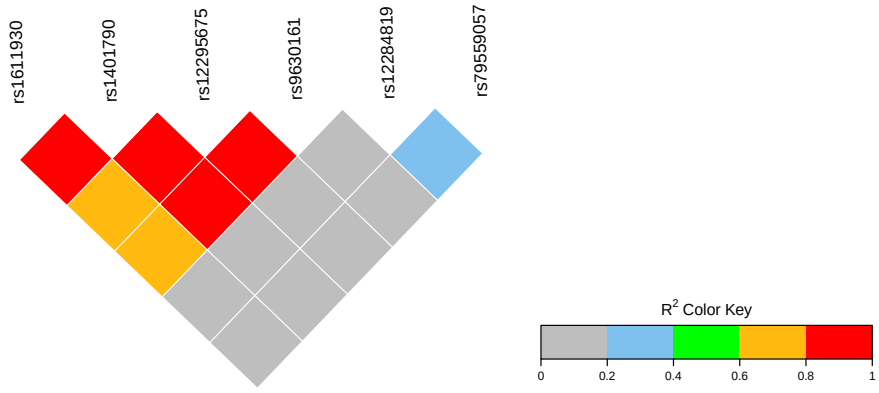

Figure B45: LD Plot for Two-point Linkage SNPs for Peak at 11p15.1 (LOD > 2.5) for RBC Count

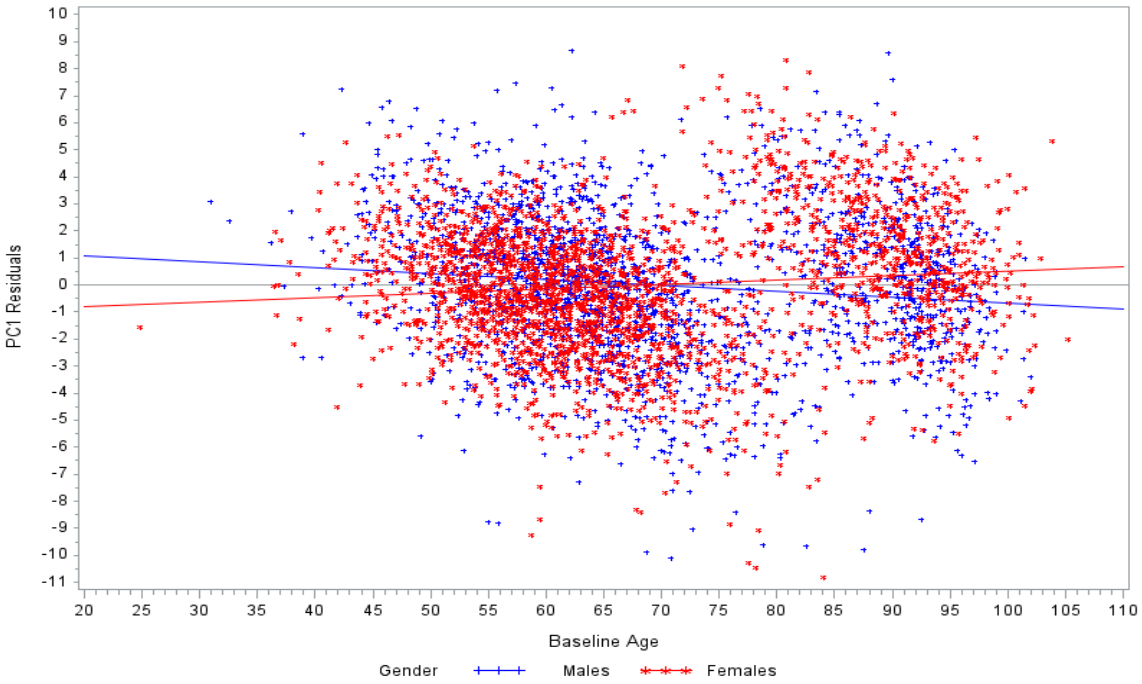

Figure B46: Scatter Plot of F1 Residuals (After Adjusting for Gender)

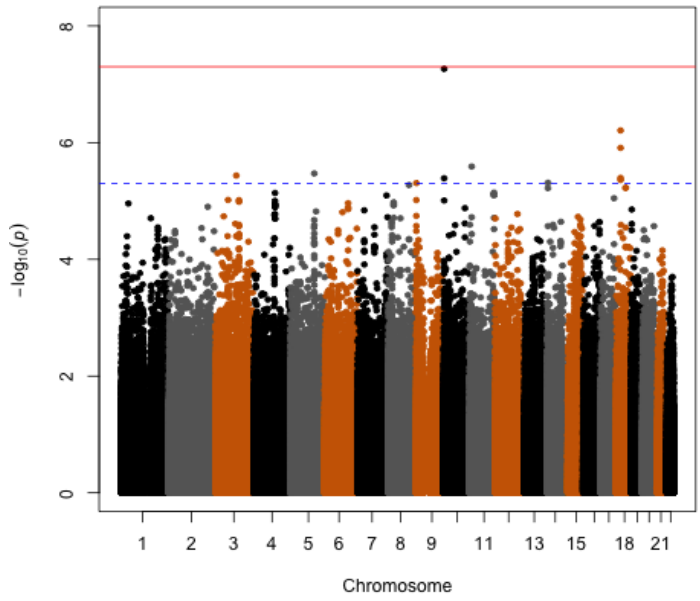

Figure B47: Manhattan Plot for RF1 for LLFS (Without Cognition; Four Factor Solution)

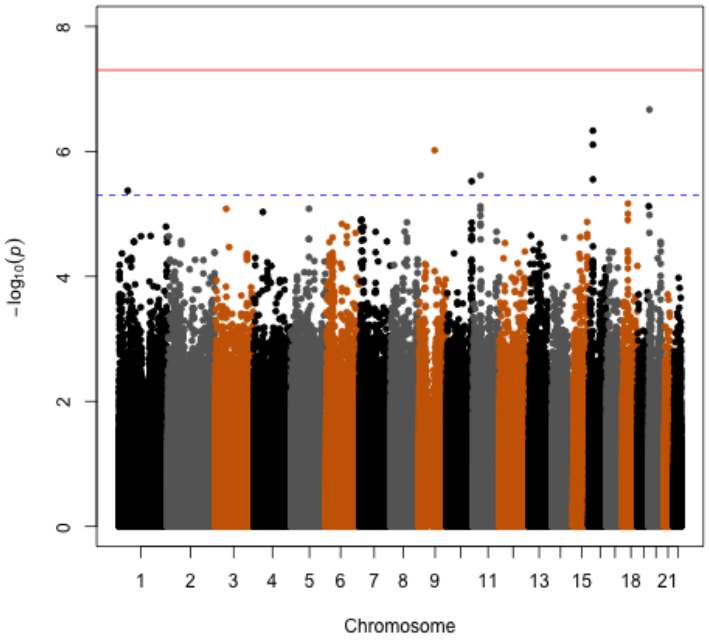

Figure B48: Manhattan Plot for RF2 for LLFS (Without Cognition; Four Factor Solution)

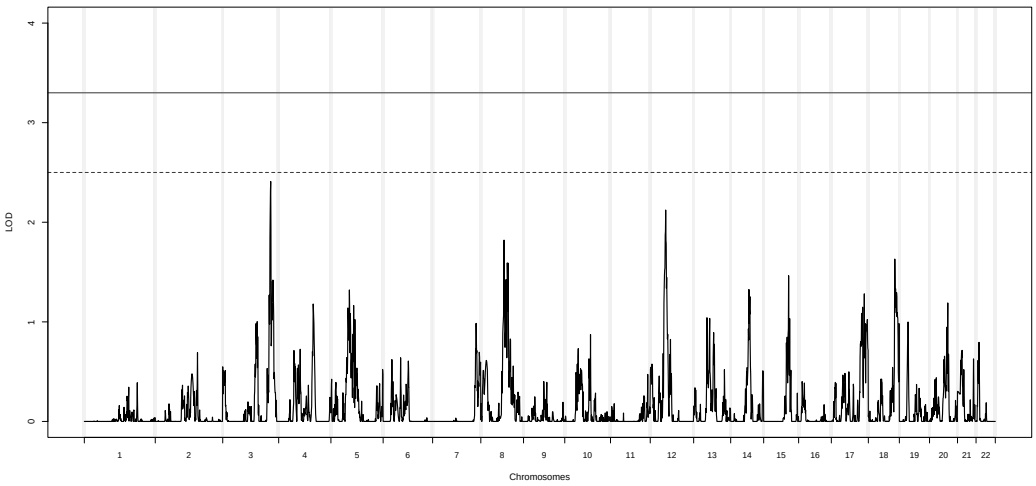

Figure B49: F1 Linkage Plot

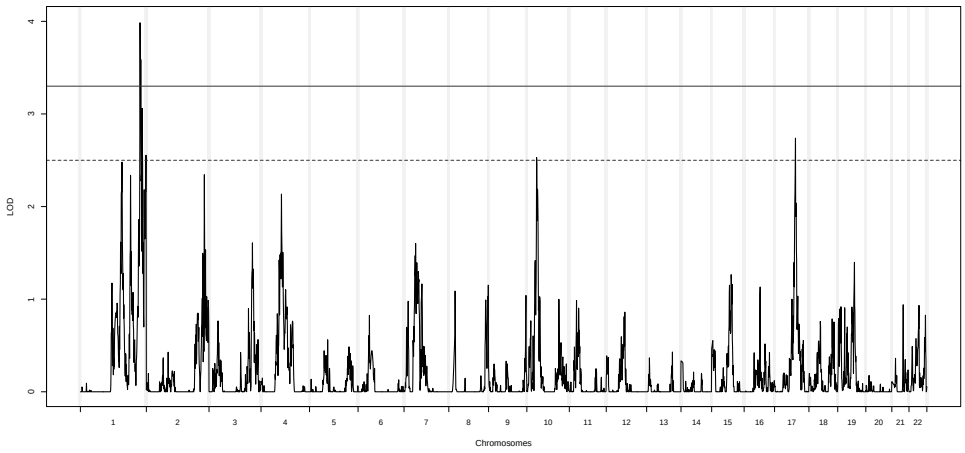

Figure B50: F2 Linkage Plot

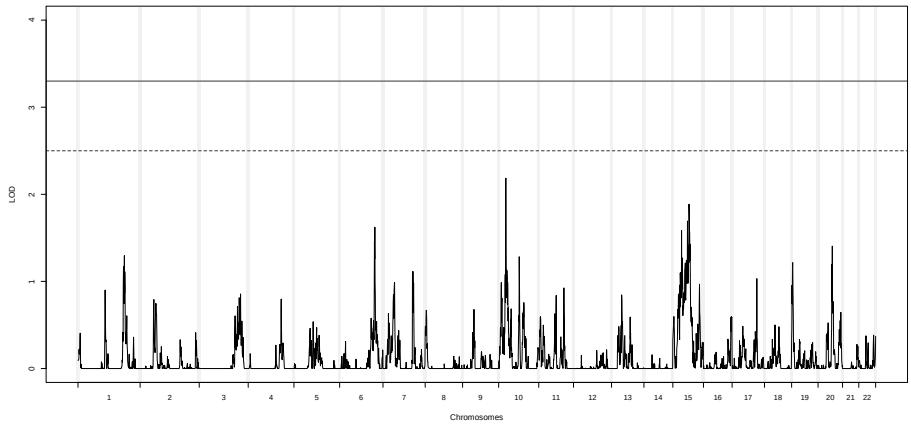

Figure B51: F3 Linkage Plot

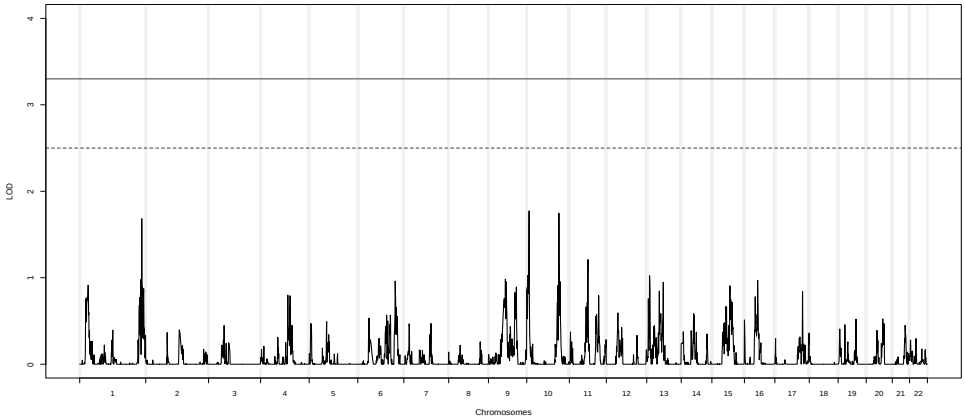

Figure B52: F4 Linkage Plot

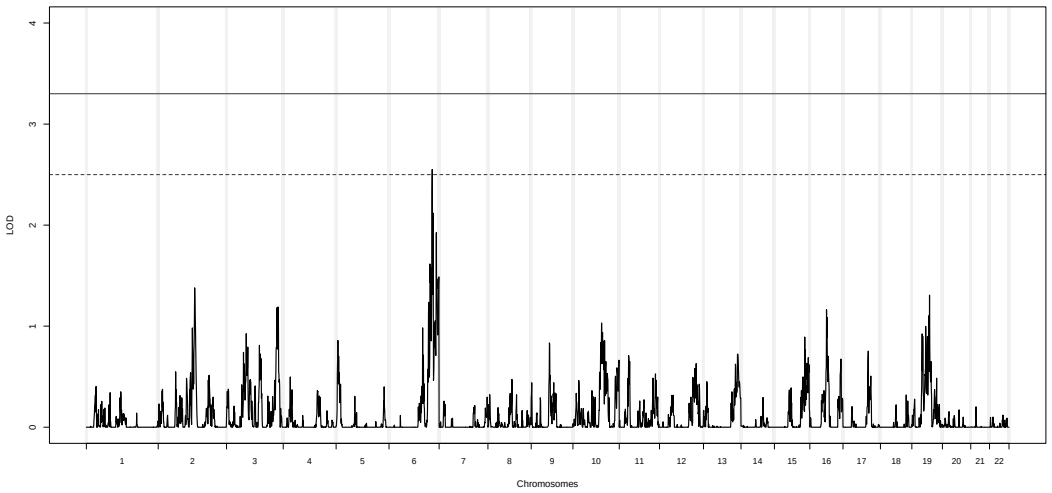

Figure B53: F5 Linkage Plot

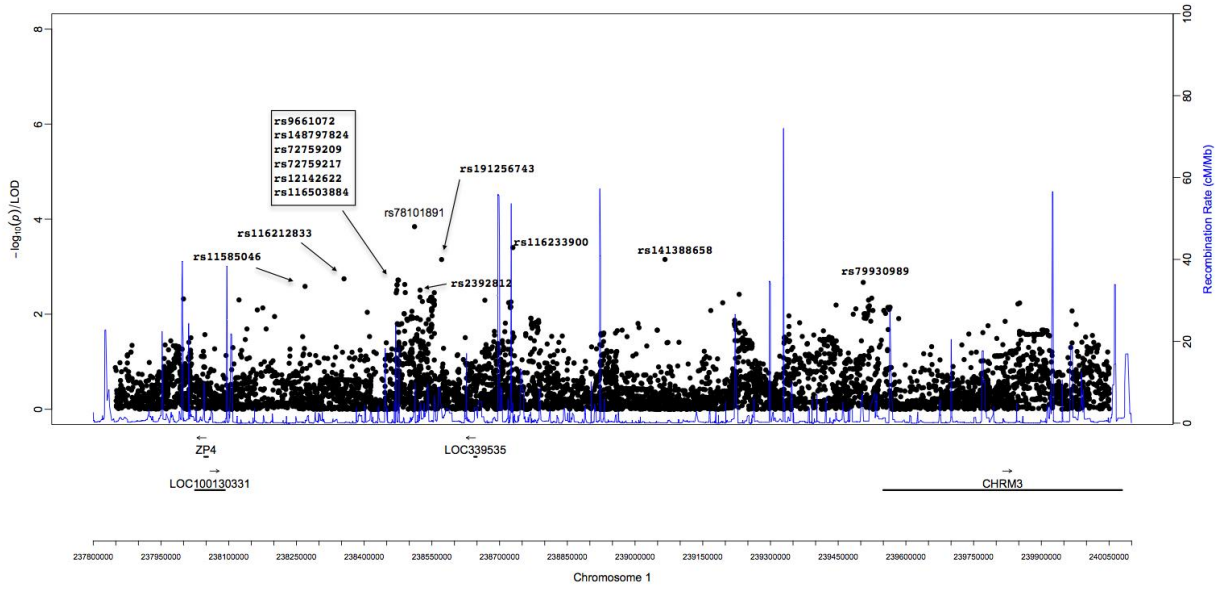

Figure B54: Association Analyses between F2 and SNPs under the chromosome 1q43 257cM peak

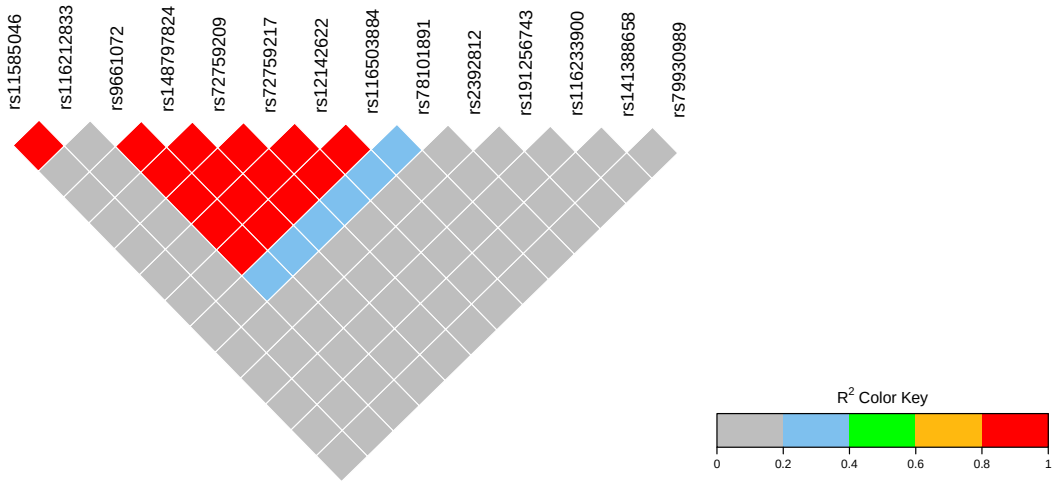

Figure B55: LD Plot for SNPs (*p*-values < 3.2×10^{-3}) usnder the Chromosome 1q43 257 cM Peak for F2

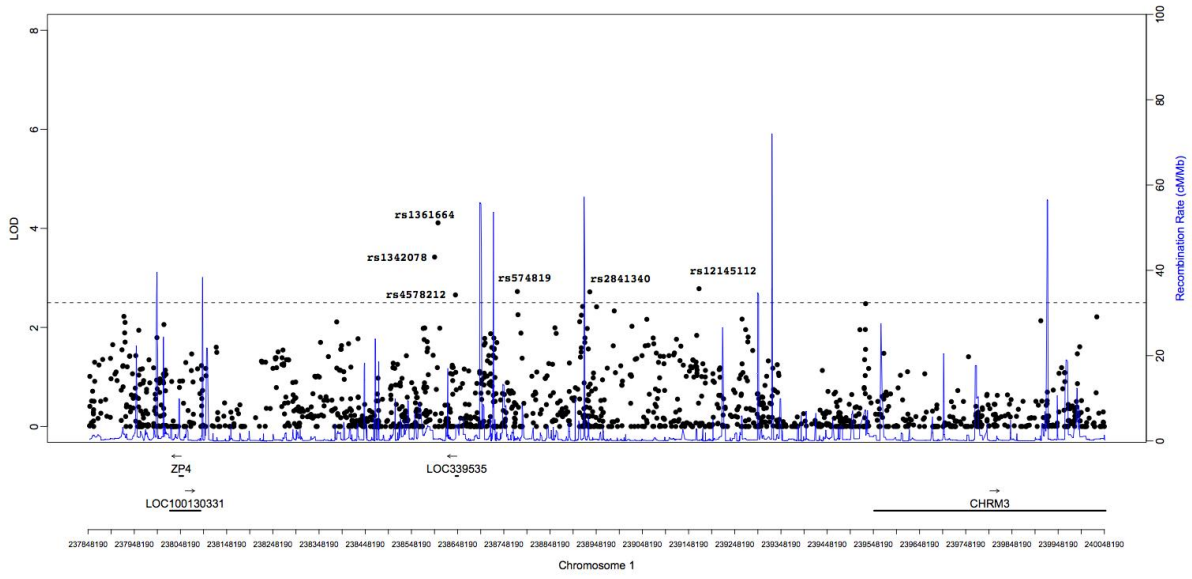

Figure B56: Two-point Linkage Analyses for F2 under the Chromosome 1q43 257cM Peak

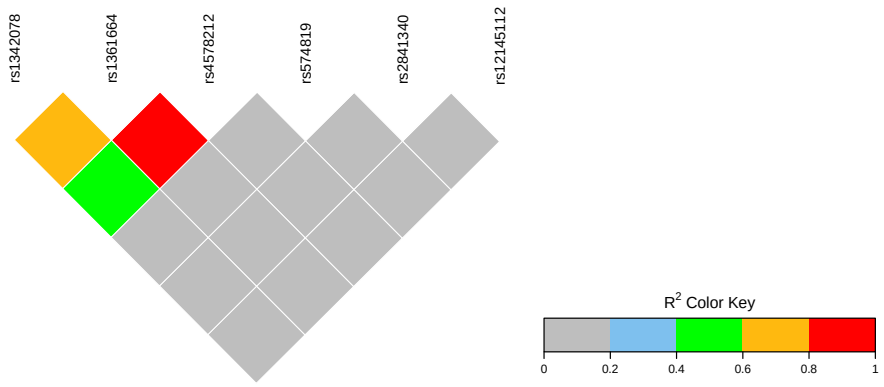

Figure B57: LD Plot for SNPs (Two-Point LOD > 2.5) under the Chromosome 1q43 257cM Peak for F2

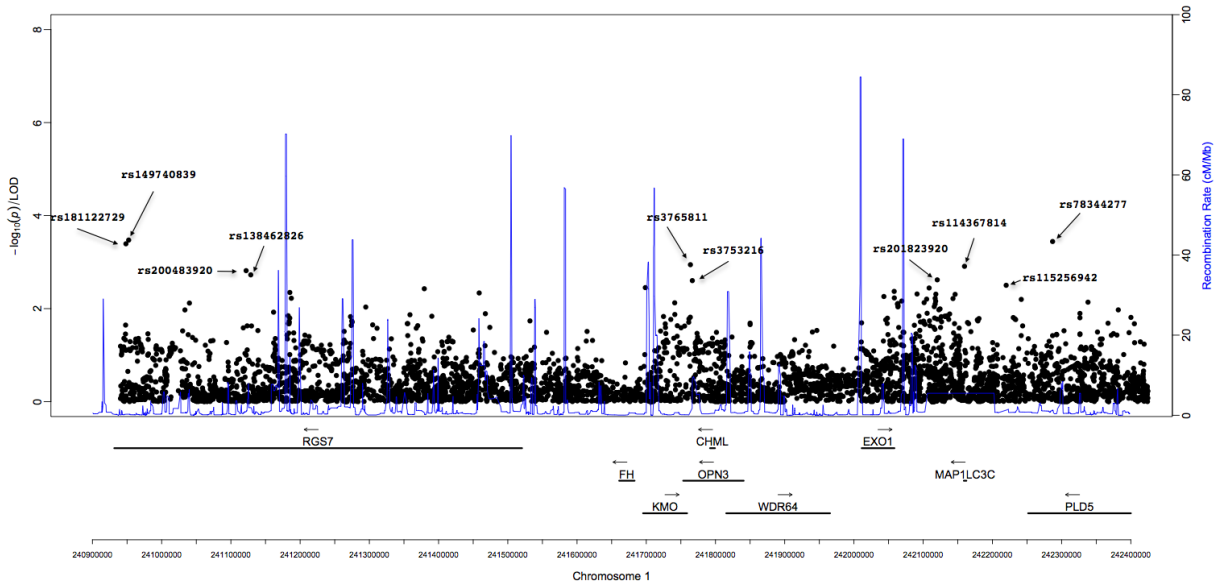

Figure B58: Association Analyses between F2 and SNPs under the Chromosome 1q43 266cM Peak

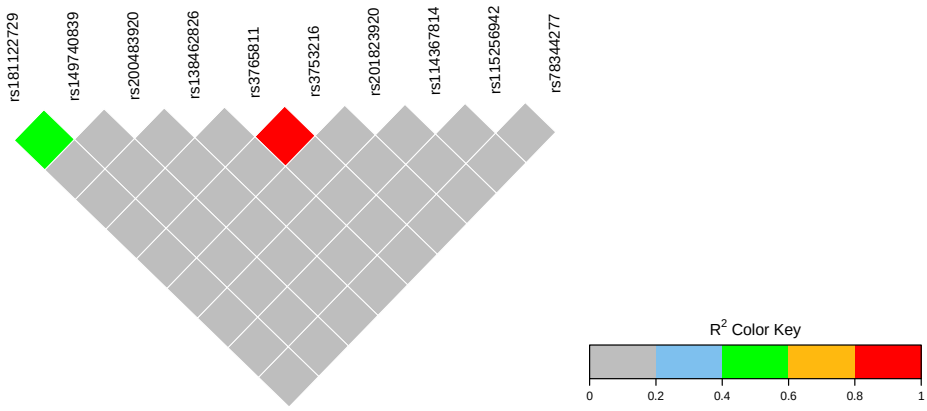

Figure B59: LD Plot for SNPs (*p*-values < 3.2×10^{-3}) under the Chromosome 1q43 266 cM Peak for F2

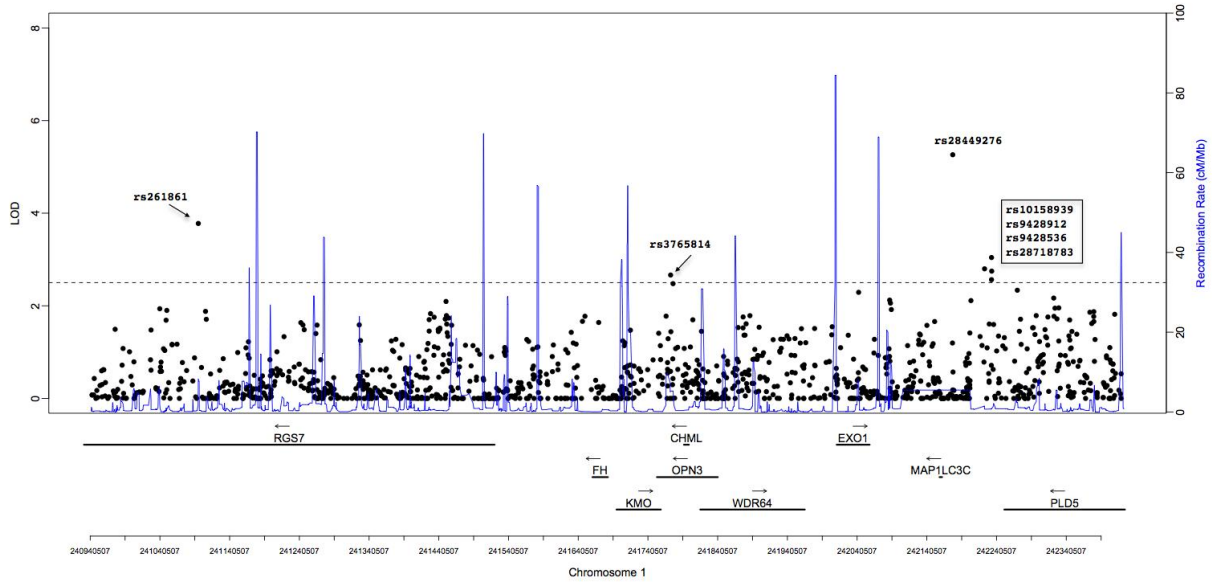

Figure B60: Two-point linkage analyses for PC2 under the chromosome 1q43 266cM peak

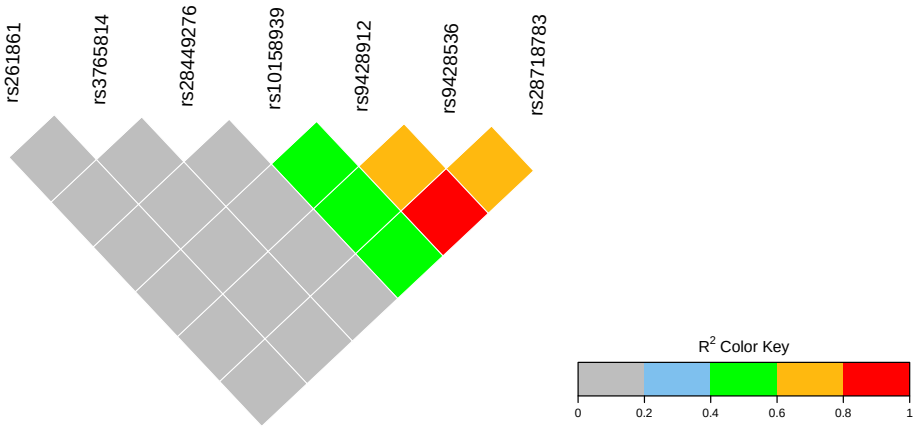

Figure B61: LD Plot for SNPs (Two-Point LOD > 2.5) under the Chromosome 1q43 266 cM Peak for F2

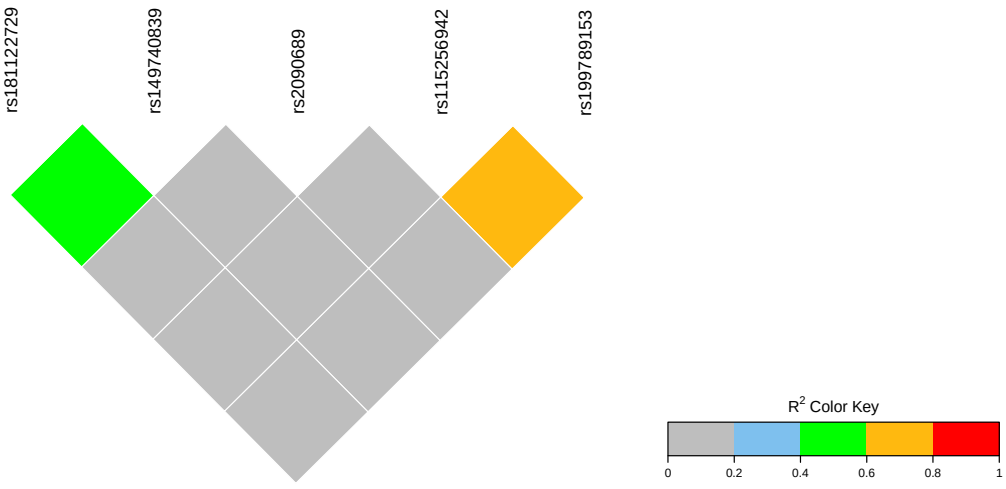

Figure B62: LD Plot for SNPs with the Largest Effect on the 266 cM Peak LOD Score for F2

BIBLIOGRAPHY

1. Hoffbrand, V. & Moss, P. *Essential Haematology*. (Wiley-Blackwell, 2011).
2. Weatherall, D. J. & Clegg, J. B. Inherited haemoglobin disorders: an increasing global health problem. *Bull. World Health Organ.* **79,** 704–712 (2001).
3. Ferreira, A. *et al.* Sickle hemoglobin confers tolerance to Plasmodium infection. *Cell* **145,** 398–409 (2011).
4. Modell, B. & Darlison, M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull. World Health Organ.* **86,** 480–487 (2008).
5. Guralnik, J. M., Eisenstaedt, R. S., Ferrucci, L., Klein, H. G. & Woodman, R. C. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. *Blood* **104,** 2263–2268 (2004).
6. Skjelbakken, T., Langbakk, B., Dahl, I. M. S., Løchen, M.-L.Tromsø Study. Haemoglobin and anaemia in a gender perspective: the Tromsø Study. *Eur. J. Haematol.* **74,** 381–388 (2005).
7. Garn, S. M., Smith, N. J. & Clark, D. C. Lifelong differences in hemoglobin levels between Blacks and Whites. *J Natl Med Assoc* **67,** 91–96 (1975).
8. Perry, G. S., Byers, T., Yip, R. & Margen, S. Iron nutrition does not account for the hemoglobin differences between blacks and whites. *The Journal of nutrition* (1992).
9. Penninx, B. W. J. H., Pahor, M., Woodman, R. C. & Guralnik, J. M. Anemia in old age is associated with increased mortality and hospitalization. (2006).
10. Culleton, B. F. *et al.* Impact of anemia on hospitalization and mortality in older adults. *Blood* **107,** 3841–3846 (2006).
11. Patel, K. V. Epidemiology of Anemia in Older Adults. *Seminars in Hematology* **45,** 210–217 (2008).
12. Izaks, G. J., Westendorp, R. G. & Knook, D. L. The definition of anemia in older persons. *JAMA* **281,** 1714–1717 (1999).
13. Patel, K. V. *et al.* Racial variation in the relationship of anemia with mortality and mobility disability among older adults. *Blood* **109,** 4663–4670 (2007).
14. Harris, T. B. *et al.* Associations of elevated Interleukin-6 and C-Reactive protein levels with mortality in the elderly. *The American Journal of Medicine* **106,** 506–512 (1999).
15. Ensrud, K. & Grimm, R. H., Jr. The white blood cell count and risk for coronary heart disease. *American Heart Journal* **124,** 207–213 (1992).
16. Hoffman, M., Blum, A., Baruch, R., Kaplan, E. & Benjamin, M. Leukocytes and coronary heart disease. *Atherosclerosis* (2004).
17. Friedman, G. D., Klatsky, A. L. & Siegelaub, A. B. The leukocyte count as a predictor of myocardial infarction. *The New England journal …* (1974).

18. Shankar, A. *et al.* Association between circulating white blood cell count and cancer mortality: a population-based cohort study. *Arch. Intern. Med.* **166,** 188–194 (2006).
19. Thaulow, E., Erikssen, J., Sandvik, L., Stormorken, H. & Cohn, P. F. Blood platelet count and function are related to total and cardiovascular death in apparently healthy men. *Circulation* **84,** 613–617 (1991).
20. Taniguchi, A. *et al.* Platelet count is independently associated with insulin resistance in non-obese Japanese type 2 diabetic patients. *Metabolism* **52,** 1246–1249 (2003).
21. Newman, A. B., Boudreau, R. M., Naydeck, B. L., Fried, L. F. & Harris, T. B. A physiologic index of comorbidity: relationship to mortality and disability. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **63,** 603–609 (2008).
22. Sanders, J. L. *et al.* Heritability of and Mortality Prediction With a Longevity Phenotype: The Healthy Aging Index. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* (2013). doi:10.1093/gerona/glt117
23. Matteini, A. M. *et al.* Heritability Estimates of Endophenotypes of Long and Health Life: The Long Life Family Study. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **65A,** 1375–1379 (2010).
24. Perls, T. T. *et al.* Life-long sustained mortality advantage of siblings of centenarians. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 8442–8447 (2002).
25. Christensen, K. *et al.* Genetic and environmental influences on functional abilities in Danish twins aged 75 years and older. *J Gerontol A Biol Sci Med Sci* **55,** M446–52 (2000).
26. Eschwege, E. *et al.* Blood cells and alcohol consumption with special reference to smoking habits. *J. Clin. Pathol.* **31,** 654–658 (1978).
27. Chalmers, D. M., Levi, A. J., Chanarin, I., North, W. R. & Meade, T. W. Mean cell volume in a working population: the effects of age, smoking, alcohol and oral contraception. *Br. J. Haematol.* **43,** 631–636 (1979).
28. Whitfield, J. B. & Martin, N. G. Genetic and environmental influences on the size and number of cells in the blood. *Genet. Epidemiol.* **2,** 133–144 (1985).
29. Evans, D. M., Frazer, I. H. & Martin, N. G. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res* **2,** 250–257 (1999).
30. Fisch, I. R. & Freedman, S. H. Smoking, oral contraceptives, and obesity. Effects on white blood cell count. *JAMA* **234,** 500–506 (1975).
31. Lin, J.-P. *et al.* Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study. *Am. J. Hematol.* **82,** 605–610 (2007).
32. Lin, J.-P., O'Donnell, C. J., Levy, D. & Cupples, L. A. Evidence for a gene influencing haematocrit on chromosome 6q23-24: genomewide scan in the Framingham Heart Study. *Journal of Medical Genetics* **42,** 75–79 (2005).
33. Hsieh, M. M., Everhart, J. E. & Byrd-Holt, D. D. Prevalence of neutropenia in the US population: age, sex, smoking status, and ethnic differences. *Ann Intern …* (2007).
34. Yang, Q., Kathiresan, S., Lin, J.-P., Tofler, G. H. & O'Donnell, C. J. Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. *BMC Med Genet* **8,** S12 (2007).
35. Craig, J. E. *et al.* Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nature Genetics* **12,** 58–64 (1996).

36. 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. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 11346–11351 (2007).
37. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics* **41,** 1182–1190 (2009).
38. Kullo, I. J., Ding, K., Jouni, H., Smith, C. Y. & Chute, C. G. A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. *PLoS ONE* **5,** e13011 (2010).
39. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nature Publishing Group* **41,** 1191–1198 (2009).
40. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature Publishing Group* **42,** 210–215 (2010).
41. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480,** 201–208 (2011).
42. Farrell, J. J. *et al.* A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood* **117,** 4935–4945 (2011).
43. Benyamin, B. *et al.* Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. *Nature Genetics* **41,** 1173–1175 (2009).
44. Ramsay, A. J., Hooper, J. D., Folgueras, A. R., Velasco, G. & López-Otín, C. Matriptase-2 (TMPRSS6): a proteolytic regulator of iron homeostasis. *Haematologica* **94,** 840–849 (2009).
45. Chambers, J. C. *et al.* Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nature Genetics* **41,** 1170–1172 (2009).
46. Hanson, E. H., Imperatore, G. & Burke, W. HFE gene and hereditary hemochromatosis: a HuGE review. Human Genome Epidemiology. *American Journal of Epidemiology* **154,** 193–206 (2001).
47. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics* **41,** 1182–1190 (2009).
48. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492,** 369–375 (2012).
49. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
50. Okada, Y. *et al.* Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. *PLoS Genet* **7,** e1002067 (2011).
51. Lo, K. S. *et al.* Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Hum. Genet.* **129,** 307–317 (2011).
52. Okada, Y. *et al.* Common variations in PSMD3-CSF3 and PLCB4 are associated with neutrophil count. *Hum. Mol. Genet.* **19,** 2079–2085 (2010).
53. McKeigue, P. M. Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am. J. Hum. Genet.* **60,** 188–196 (1997).

54. 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. *The American Journal of Human Genetics* **82,** 81–87 (2008).
55. Michon, P. *et al.* Duffy-null promoter heterozygosity reduces DARC expression and abrogates adhesion of the P. vivax ligand required for blood-stage infection. *FEBS Lett.* **495,** 111–114 (2001).
56. Cheng, C. L., Gao, T. Q., Wang, Z. & Li, D. D. Role of insulin/insulin-like growth factor 1 signaling pathway in longevity. *World J Gastroenterol* (2005).
57. Newman, A. B. & Murabito, J. M. The Epidemiology of Longevity and Exceptional Survival. *Epidemiol Rev* – (2013). doi:10.1093/epirev/mxs013
58. Brooks-Wilson, A. R. Genetics of healthy aging and longevity. *Hum. Genet.* **132,** 1323–1338 (2013).
59. Sebastiani, P. P. *et al.* A family longevity selection score: ranking sibships by their longevity, size, and availability for study. *CORD Conference Proceedings* **170,** 1555–1562 (2009).
60. Newman, A. B. *et al.* Health and function of participants in the Long Life Family Study: A comparison with other cohorts. *Aging (Albany NY)* **3,** 63–76 (2011).
61. Heath, S. C. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.* **61,** 748–760 (1997).
62. O'Connell, J. R. Zero-recombinant haplotyping: applications to fine mapping using SNPs. *Genet. Epidemiol.* **19 Suppl 1,** S64–S70 (2000).
63. Almasy, L. & Blangero, J. Variance component methods for analysis of complex phenotypes. *Cold Spring Harb Protoc* **2010,** pdb–top77 (2010).
64. DeLong, E. R. E., DeLong, D. M. D. & Clarke-Pearson, D. L. D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44,** 837–845 (1988).
65. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62,** 1198–1211 (1998).
66. Pinheiro, J. C. & Bates, D. M. Linear mixed-effects models: basic concepts and examples. (2000).
67. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30,** 97–101 (2002).
68. Conneally, P. M. *et al.* Report of the Committee on Methods of Linkage Analysis and Reporting. *Cytogenet. Cell Genet.* **40,** 356–359 (1985).
69. Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11,** 134 (2010).
70. Pollin, T. I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322,** 1702–1705 (2008).
71. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337,** 100–104 (2012).
72. Bain, B. J. & England, J. M. Normal haematological values: sex difference in neutrophil count. *Br Med J* **1,** 306–309 (1975).
73. Bain, B. J. Platelet count and platelet size in males and females. *Scand J Haematol* **35,** 77–79 (1985).
74. Ishizaka, N. *et al.* Relationship between smoking, white blood cell count and metabolic syndrome in Japanese women. *Diabetes Res Clin Pract* **78,** 5–5 (2007).

75. Jenkins, B. J., Roberts, A. W., Najdovska, M., Grail, D. & Ernst, M. The threshold of gp130-dependent STAT3 signaling is critical for normal regulation of hematopoiesis. *Blood* **105,** 3512–3520 (2005).
76. Lee, J.-W. *et al.* DACH1 regulates cell cycle progression of myeloid cells through the control of cyclin D, Cdk 4/6 and p21Cip1. *Biochemical and Biophysical Research Communications* **420,** 91–95 (2012).
77. Ding, K. *et al.* Genetic Loci implicated in erythroid differentiation and cell cycle regulation are associated with red blood cell traits. *Mayo Clin. Proc.* **87,** 461–474 (2012).
78. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
79. Hamada-Kanazawa, M. *et al.* Suppression of Sox6 in P19 cells leads to failure of neuronal differentiation by retinoic acid and induces retinoic acid-dependent apoptosis. *FEBS Lett.* **577,** 60–66 (2004).
80. Cohen-Barak, O. *et al.* Sox6 regulation of cardiac myocyte development. *Nucleic Acids Res.* **31,** 5941–5948 (2003).
81. Lefebvre, V., Behringer, R. R. & de Crombrugghe, B. L-Sox5, Sox6 and Sox9 control essential steps of the chondrocyte differentiation pathway. *Osteoarthr. Cartil.* **9 Suppl A,** S69–75 (2001).
82. Dumitriu, B. Sox6 cell-autonomously stimulates erythroid cell survival, proliferation, and terminal maturation and is thereby an important enhancer of definitive erythropoiesis during mouse development. *Blood* **108,** 1198–1207 (2006).
83. Dumitriu, B. *et al.* Sox6 Is Necessary for Efficient Erythropoiesis in Adult Mice under Physiological and Anemia-Induced Stress Conditions. *PLoS ONE* **5,** e12088 (2010).
84. Wagner, K. U. *et al.* Conditional deletion of the Bcl-x gene from erythroid cells results in hemolytic anemia and profound splenomegaly. *Development* **127,** 4949–4958 (2000).
85. Xu, J. *et al.* Transcriptional silencing of -globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev* **24,** 783–798 (2010).
86. Solovieff, N. *et al.* Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood* **115,** 1815–1822 (2010).
87. Shim, J. *et al.* Olfactory control of blood progenitor maintenance. *Cell* **155,** 1141–1153 (2013).
88. Ayyadevara, S., Alla, R., Thaden, J. J. & Shmookler Reis, R. J. Remarkable longevity and stress resistance of nematode PI3K-null mutants. *Aging Cell* **7,** 13–22 (2008).
89. Tissenbaum, H. A. & Guarente, L. Model organisms as a guide to mammalian aging. *Dev. Cell* **2,** 9–19 (2002).
90. Paaby, A. B. & Schmidt, P. S. Dissecting the genetics of longevity in Drosophila melanogaster. *Fly (Austin)* **3,** 29–38 (2009).
91. Yuan, R., Peters, L. L. & Paigen, B. Mice as a mammalian model for research on the genetics of aging. *ILAR J* **52,** 4–15 (2011).
92. Murabito, J. M., Yuan, R. & Lunetta, K. L. The search for longevity and healthy aging genes: insights from epidemiological studies and samples of long-lived individuals. *J Gerontol A Biol Sci Med Sci* **67,** 470–479 (2012).
93. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–2337 (2010).

94. Bathum, L. *et al.* Apolipoprotein e genotypes: relationship to cognitive functioning, cognitive decline, and survival in nonagenarians. *J Am Geriatr Soc* **54,** 654–658 (2006).
95. Jacobsen, R. *et al.* Increased effect of the ApoE gene on survival at advanced age in healthy and long-lived Danes: two nationwide cohort studies. *Aging Cell* **9,** 1004–1009 (2010).
96. Hsin, H. & Kenyon, C. Signals from the reproductive system regulate the lifespan of C. elegans. *Nature* **399,** 362–366 (1999).
97. Walter, S. *et al.* A genome-wide association study of aging. *Neurobiology of Aging* **32,** 2109.e15–2109.e28 (2011).
98. Pearson, R. R., Fleetwood, J. J., Eaton, S. S., Crossley, M. M. & Bao, S. S. Kruppel-like transcription factors: A functional family. *Int J Biochem Cell Biol* **40,** 6–6 (2008).
99. Bieker, J. J. Krüppel-like factors: three fingers in many pies. *J Biol Chem* (2001).
100. Narla, G. *et al.* A germline DNA polymorphism enhances alternative splicing of the KLF6 tumor suppressor gene and is associated with increased prostate cancer risk. *Cancer Res.* **65,** 1213–1222 (2005).
101. Matsubara, E. *et al.* The role of zinc finger protein 521/early hematopoietic zinc finger protein in erythroid cell differentiation. *J Biol Chem* **284,** 3480–3487 (2009).
102. Bond, H. M. *et al.* Early hematopoietic zinc finger protein-zinc finger protein 521: a candidate regulator of diverse immature cells. *Int J Biochem Cell Biol* **40,** 848–854 (2008).
103. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42,** 937–948 (2010).
104. Edwards, D. R. V. *et al.* Successful aging shows linkage to chromosomes 6, 7, and 14 in the Amish. *Annals of Human Genetics* **75,** 516–528 (2011).
105. Beekman, M. *et al.* Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Aging Cell* **12,** 184–193 (2013).
106. Boyden, S. E. & Kunkel, L. M. High-density genomewide linkage analysis of exceptional human longevity identifies multiple novel loci. *PLoS ONE* **5,** e12432 (2010).
107. Kerber, R. A., O'Brien, E., Boucher, K. M., Smith, K. R. & Cawthon, R. M. A genome-wide study replicates linkage of 3p22-24 to extreme longevity in humans and identifies possible additional loci. *PLoS ONE* **7,** e34746 (2012).
108. Edwards, D. R. V. *et al.* Linkage and association of successful aging to the 6q25 region in large Amish kindreds. *Age (Dordr)* **35,** 1467–1477 (2013).
109. Comuzzie, A. G. *et al.* Novel genetic Loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLoS ONE* **7,** e51954–e51954 (2012).
110. Nalls, M. A. *et al.* Multiple loci are associated with white blood cell phenotypes. *PLoS Genet* **7,** e1002113 (2011).
111. 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. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 11869–11874 (2008).
112. Meisinger, C. *et al.* A Genome-wide Association Study Identifies Three Loci Associated with Mean Platelet Volume. *The American Journal of Human Genetics* **84,** 66–71 (2009).