Université de Montréal

Genetic Determinants of Rare Disorders and Complex Traits: Insights into the Genetics of Dilated Cardiomyopathy and Blood Cell Traits

par

Nathalie Chami

Département de Science Biomédicale

Faculté de médecine

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de doctorat

en Science Biomédicale

Avril 2017

DEDICATION

To my parents and brothers for all their love and everlasting sacrifices, for their patience, and continuous support and faith in me…And to my aunt, Fadia, who never saw the completion of this work, but who was an indispensable driver of my journey.

I love you.

RESUME

Les facteurs génétiques peuvent apporter des réponses à plusieurs questions que nous nous posons sur les traits humains, les maladies et la réaction aux médicaments, entre autres. Avec le temps, le développement continu d'outils d'analyse génétique nous a permis d'examiner ces facteurs et de trouver des explications pertinentes. Cette thèse explore plusieurs méthodes et outils génétiques, tels que le séquençage pan-exomique et le génotypage sur puce, dans un contexte d'analyse familial et populationnel pour étudier ces facteurs génétiques qui jouent un rôle dans une maladie rare, la cardiomyopathie dilatée (DCM), et dans deux traits complexes soient les globules rouges et les plaquettes.

DCM est une maladie rare qui est définie par un ventricule gauche dilaté et une dysfonction systolique. Environ 30% des cas de DCM sont héréditaires, et plus de 50 gènes ont été associés à un rôle dans la pathogénicité de DCM. Le dépistage génétique est un outil de référence dans la gestion clinique de DCM familiale. Par contre, pour la majorité des patients, les tests génétiques ne parviennent pas à identifier une mutation causale dans un gène candidat.

Les cellules sanguines remplissent une variété de fonctions biologiques, incluant le transport de l'oxygène, les fonctions immunologiques, ainsi que la guérison de plaies. Les niveaux de ces cellules et leurs paramètres auxiliaires sont mesurés par un test sanguin, et une différence avec les valeurs optimales peut signifier certains troubles. De plus, ces traits sont étudiés méticuleusement dans le contexte des maladies cardiovasculaires (CVD) où différents niveaux sont associés avec un risque variable de CVD ou sont des prédicteurs de complications de CVD.

J'ai examiné la DCM et les traits sanguins avec comme objectif de découvrir des nouvelles associations de mutations génétiques. Pour la DCM, j'ai évalué la pertinence d'un séquençage pan-exomique dans un environnement clinique. Je rapporte plusieurs nouvelles mutations dans des gènes candidats (*DSP, LMNA, MYH7, MYPN, RBM20, TNNT2*) et des mutations nonsenses dans deux gènes nouvellement associés (*TTN* et *BAG3*), et je démontre que les mutations nonsenses influencent la maladie d'une manière différente des autres mutations causales. Je rapporte aussi une mutation dans un nouveau gène, *FLNC*, qui cause une forme rare et distincte de cardiomyopathie. Pour l'étude des traits complexes, dans le grand consortium Blood Cell Consortium (BCX), j'ai utilisé l'exomechip pour disséquer le rôle des variantes rares et communes dans les globules rouges et les plaquettes. J'ai identifié 16 nouvelles régions génomiques associées avec les globules rouges et 15 avec les plaquettes, parmi lesquelles se retrouvent plusieurs variantes de basses fréquences (*MAP1A, HNF4A, ITGA2B, APOH*), et j'ai démontré un chevauchement significatif de régions associées avec d'autres traits, incluant les lipides.

Mes résultats sur la DCM ont mis en évidence le rôle de plusieurs gènes candidats, et suggèrent un traitement différent au niveau de la gestion clinique des patients qui portent des mutations dans *BAG3* et *FLNC*. En ce qui concerne les traits sanguins, mes résultats contribuent à enrichir le répertoire de régions associées avec ces traits, soulignant l'importance de l'utilisation de grands ensembles de données pour détecter les variantes rares ou de basses fréquences. La découverte de gènes dans les maladies rares et les traits complexes contribue à la compréhension des mécanismes sous-jacents qui ultimement favorisera de meilleurs diagnostics, gestions et traitements de maladies.

Mots-clés: cardiomyopathie dilatée, séquençage pan-exomique, analyse familial, cellules sanguines, exome chip, analyse populationnel, Blood Cell Consortium (BCX)

ABSTRACT

Genetic factors hold within them the answers to many questions we have on human traits, disease, and drug response among others. With time, the continuously advancing genetic tools have enabled us to examine those factors and provided and continue to provide astonishing answers. This thesis utilizes various methods of genetic tools such as exome sequencing and chip-based genotyping data in the context of both family and population-based analyses to interrogate the genetic factors that play a role in a rare disease, dilated cardiomyopathy (DCM), and in two complex traits, red blood cells and platelets.

DCM is a rare disease that is defined by a dilated left ventricle and systolic dysfunction. It is estimated that 30% of DCM cases are hereditary and more than 50 genes have been linked to play a role in the pathogenesis of DCM. Genetic screening of known genes is a gold standard tool in the clinical management of familial DCM. However, in the majority of probands, genetic testing fails to identify the causal mutation.

Blood cells play a variety of biological functions including oxygen transport, immunological functions, and wound healing. Levels of these cells and their associated indices are measured by a blood test, and deviation from optimal values may indicate certain disorders. Additionally, these traits are heavily studied in the context of cardiovascular disease (CVD) where different levels associate with a variable risk of CVD or are predictors of CVD complications or outcomes (for example, a higher level of white blood cells or lower level of hemoglobin).

I examined both DCM and blood cell traits and aimed to discover new mutations and variants that are associated with each. For DCM, I evaluated the value of whole exome

sequencing in a clinical setting, and I report a number of novel mutations in candidate genes (*DSP, LMNA, MYH7, MYPN, RBM20, TNNT2*) and truncating mutations in two newly established genes, *TTN* and *BAG3*, and I demonstrate that truncating mutations in the latter influence disease differently than other causal mutations. I also report a mutation in a novel gene, *FLNC* that causes a rare and distinct form of cardiomyopathy. In examining complex traits, I dissected the role of common and rare variants in red blood cells and platelets within a large consortium, the Blood Cell Consortium (BCX) using the ExomeChip, and identified 16 novel loci associated with red blood cell traits and 15 with platelet traits, some of which harbored low-frequency variants (*MAP1A, HNF4A, ITGA2B, APOH*), and demonstrated a substantial overlap with other phenotypes predominantly lipids.

My results on DCM establish the role of a number of candidate genes in this disorder and suggest a different course of clinical management for patients that carry mutations in *BAG3* and *FLNC*. As for blood cell traits, my results contributed to expanding the repertoire of loci associated with red blood cell and platelet traits and illustrate the importance of using large datasets to discover low-frequency or rare variants. Gene discovery in rare disease and complex traits gives insight into the underlying mechanisms which ultimately contributes to a better diagnosis, management, and treatment of disease.

Keywords: Dilated cardiomyopathy, whole-exome sequencing, family study, blood cell traits, exome chip, population study, Blood Cell Consortium (BCX)

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Dr. Guillaume Lettre. I would need a lot more than the lines I have here to express how thankful and grateful I am to the fate that has brought me to your lab. You have been an incredible mentor and wonderful human being. I have learned a lot from you be it in Science or life in general. Thank you for being extremely encouraging, understanding, and a true believer in what is right. I will always admire your unwavering honesty and work ethics that constitute a cornerstone in the way you deal with science and research and the people around you. I will proudly look back to the days that I spent in your lab… and proudly look up to you in my career going forward.

I would like to thank Dr. Mario Talajic for his guidance and support and for always finding the time to meet and discuss any issue even with his busy schedule. I am also grateful to the members of the thesis committee, Dr. Gregor Andelfinger and Dr. Marie-Pierre Dube for their insightful comments and suggestions. I would like to thank the members and participants of the Blood Cell Consortium. This has been a great learning experience and an enjoyable effort.

I would like to thank my lab mates for the beautiful time we have spent in the lab and for the fun environment we have all created. I would like to thank in particular Melissa Beaudoin, Ken Sin Lo, and Cecile Low Kam for always being there, for listening to me in the good and the bad times. You have been so wonderful! I will always cherish our friendship.

I am extremely grateful to two incredible people: Dr. Rafik Tadros and Laura Robb for all their support and help. I have learned a lot from you and I have enjoyed our encounters and random meetings which always had room for some laughter and fun despite the load of work.

I am infinitely grateful to my family. None of my dreams would have been possible without the continuous encouragement, unswerving sacrifices and tireless persistence of my family. Thank you mom and dad for your faith in me and for opening all the doors for me. As a mother, I now understand more than ever the meaning of your sacrifices; the meaning of sending me to pursue my goals without knowing when and if I will ever return back to you. I am also beyond grateful for my family in Canada for always being there for me ever since I arrived to Montreal, for attending to all my needs to create the best environment for me to succeed.

Last but not least, I am extremely grateful to my husband, Imad. Thank you for your faith in me, for always pushing me to do better, and for being so understanding and supportive when I needed you. You are my inspiration, my love and my soul! I am so lucky to have you by my side.

TABLE OF CONTENTS

Dedication i
Résumé ii
Abstract v
Acknowledgements vii
Table of Contents ix
List of Figures xiii
Supplementary Figures xv
List of Tables xvi
Supplementary Tables xviii
List of Abbreviations and Acronyms xix
CHAPTER 1: Introduction 1
1.1. Preface 1
1.2. Cardiomyopathy: a rare disorder 3
1.2.1. Relevant Clinical Background 3
1.2.2. DCM as a Genetic Disorder 14
1.3. Blood Cell number, size, or content: Classic human Complex Trait 26
1.3.1. Clinical Background 26
1.3.2. Genetic approaches to study complex traits 36
1.3.3. Genetic Findings of Blood Cell Traits 40
1.4. Research Objectives 59
1.5. Thesis Contributions to Knowledge 60
1.6. Contributions of Authors 62
CHAPTER 2: Nonsense mutations in BAG3 are associated with early-onset dilated cardiomyopathy in French Canadians. 65
2.1. Abstract 66
2.2. Introduction 68

2.3. Materials and Methods 69

2.4. RESULTS 72

2.5. DISCUSSION 81

2.6. Acknowledgements 84

2.7. Supplementary Information 85

CHAPTER 3: A splicing mutation in FLNC causes a rare form of cardiomyopathy. 89
3.1. Abstract 90

3.2. Introduction 91

3.3. Materials and Methods 93

3.4. Results 95

3.5. Discussion 102

CHAPTER 4: Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits 105
4.1. Abstract 106

4.2. Introduction 107

4.3. Subjects and Methods 109

4.4. Results 114

4.5. Discussion 130

4.6. Acknowledgements 133

CHAPTER 5: Platelet-related variants identified by exomechip meta-analysis in 157,293 individuals 136
5.1. Abstract 137

5.2. Introduction 138

5.3. Materials and Methods 141

5.4. Results 147

5.5. Discussion 159

5.6. Conclusions 165

5.7. Acknowledgements 166

CHAPTER 6: General Discussion 167

6.1. Lessons from Our Exome Sequencing Study 167

6.1.1. The BAG3 and FLNC mutations 167

6.1.2. The value of genetic findings in monogenic disease 168

6.1.3. Clinical impact of our study 169

6.1.4. The utility of whole exome sequencing 172

6.1.5. Exome sequencing caveats 173

6.1.6. Other challenges in monogenic disease studies 175

6.1.7. Current databases and variant prioritization 176

6.1.8. Proving pathogenicity 178

6.1.9. The hunt for modifier genes 180

6.2. Lessons from the exomechip study 182

6.2.1. Pleiotropy in blood cell traits 182

6.2.2. Rare variants and complex traits 183

6.2.3. Associated variants lying in Mendelian disease genes 185

6.2.4. The missing heritability problem 185

6.2.5. Functional experiments in blood cells 190

6.2.6. The value of identified variants in complex traits 192

6.3. The Goal of Personalized Medicine 194

6.4. Conclusions and final comments 197

Appendix 1: **Large-Scale Exome-wide Association Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases**. 200
Appendix 2: **Rare and Low-frequency Coding Variants in CXCR2 and Other Genes are Associated with Hematological Traits** 202
References 204

LIST OF FIGURES

Figure 1.1. Heart morphologies of the most common types of cardiomyopathy..... 11

Figure 1.2. Genetic heterogeneity of DCM..... 17

Figure 1.3. Functional groups of genes associated with dilated cardiomyopathy. 23

Figure 1.4. Hematopoietic stem cell differentiation..... 27

Figure 1.5. Erythropoiesis. 28

Figure 1.6. Platelets mediating hemostasis and thrombosis..... 35

Figure 1.7. The generally accepted model of disease susceptibility and variant allele frequency..... 39

Figure 1.8. Ideal study design to identify SNPs associated with human complex traits and diseases using genome-wide association studies (GWAS). 49

Figure 2.1. Pedigree of dilated cardiomyopathy families 1, 6 and 12. 77

Figure 2.2. Kaplan Meier curves of age of onset and severe adverse events of dilated cardiomyopathy (DCM) in carriers and non-carriers of *BAG3* nonsense mutations. 79

Figure 3.1. Pedigree of family 7 96

Figure 3.2. Fixed heart specimen from III.11 101

Figure 4.1. Quantile-Quantile (QQ) plots of single variant association results in the all ancestry meta-analyses for the seven red blood cell (RBC) traits analyzed 116

Figure 4.2. *CD36* expression in human erythroblasts. 123

Figure 4.3. Venn Diagram Summarizing Pleiotropic Effects for Genetic Variants Associated with Red Blood Cell Traits. 129

Figure 5.1. Study Design and Flow..... 149

Figure 5.2. Shared PLT and MPV Genetic Associations..... 155

SUPPLEMENTARY FIGURES

Supplementary Figure 2.1. Kaplan Meier curves of age of onset and clinical outcomes of dilated cardiomyopathy (DCM) in carriers and non-carriers of *TTN* nonsense mutations... 85

Supplementary Figure 4.1. Flow chart of the study design. 134

LIST OF TABLES

Table 1.1. Genes associated with dilated cardiomyopathy (DCM). 24

Table 1.2. Main blood cell traits routinely measured in standard complete blood count (CBC). 44

Table 1.3. Loci identified by GWAS that carry SNPs associated with at least two of the three main blood cell types. 51

Table 1.4. Orphan human syndromes mapped to a chromosomal band and characterized by a blood cell phenotype. 57

Table 2.1. Clinical characteristics for the 42 dilated cardiomyopathy (DCM) subjects that were whole-exome sequenced at the Montreal Heart Institute (MHI). 73

Table 2.2. Mutations identified in candidate dilated cardiomyopathy genes in this whole-exome DNA sequencing experiment. 76

Table 3.1. Clinical characteristics of the evaluated family members 97

Table 3.2. Variants retained prior to segregation analysis 100

Table 4.1. Association results of variants in novel loci associated with red blood cell (RBC) traits. 117

Table 4.2. Gene-based association results 126

Table 4.3. Overlap of red blood cell (RBC) markers with other blood cell traits and/or lipids. 128

Table 5.1. Novel associations (n=32) with PLT. 150

Table 5.2. Novel associations (n=18) with MPV. 151

Table 5.3. Variants associated with both PLT and MPV. 156

Table 5.4. Intersection of platelet associated variants with red blood cell (RBC) and white blood cell (WBC) traits ($p<0.0001$). 157

Table 5.5. Overlap of associations of platelet count (PLT) and mean platelet volume (MPV) variants with platelet reactivity ($p<0.001$). 158

SUPPLEMENTARY TABLES

Supplementary Table 2.1. Inclusion and exclusion criteria for probands with dilated cardiomyopathy (DCM). 86

Supplementary Table 2.2. Clinical characteristics of all family members for whom a potential pathogenic mutation was identified. 87

Supplementary Table 4.1. Expression quantitative trait loci (eQTL) results for variants associated with red blood cell phenotypes. 135

LIST OF ABBREVIATIONS AND ACRONYMS

AA	African-American
ARVC	Arrhythmogenic Right Ventricular Cardiomyopathy
AV	Atrio-ventricular
BCX	Blood Cell Consortium
CBC	complete blood count
CCS	Canadian Cardiovascular Society
CHD	Coronary heart disease
CVD	Cardiovascular disease
DCM	Dilated Cardiomyopathy
EA	European
EDMD	Emery-Dreifuss muscular dystrophy
EPO	Erythropoietin
eQTL	expression quantitative trait loci
ESP	NHLBI Exome Sequencing Project
ESR	erythrocyte sedimentation rate
ExAC	The Exome Aggregation Consortium
GCs	Glucocorticoids
gnomAD	genome Aggregation Database
GRS	genetic risk score
GTR	Genetic Testing Registry
GWAS	Genome-wide Association Studies
HbA	adult hemoglobin
HbF	fetal hemoglobin
HCM	Hypertrophic Cardiomyopathy
HCT	Hematocrit
HGB	Haemoglobin
HSCs	hematopoietic stem cells
IBD	identity-by-descent
ICD	implantable cardioverter defibrillator
IGF-1	growth factor 1
LD	linkage disequilibrium
LMM	Laboratory of Molecular Medicine
LOD	Logarithm of Odds Ratio
LV	left ventricular
LVEDD	left ventricular end-diastolic diameter
LVEF	Left ventricular ejection fraction
LVNC	Left ventricular non-compaction cardiomyopathy
MAF	minor allele frequency
MCH	Mean corpuscular hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MCV	Mean corpuscular volume

MHI	Montreal Heart Institute
MI	Mocardial infarction
MPN	myeloproliferative neoplasms
MPV	mean platelet volume
MRI	magnetic resonance imaging
NGS	Next generation sequencing
PCA	principle component analysis
PLT	PLT count
PPCM	peripartum cardiomyopathy
PVC	premature ventricular contraction
RBC	red blood cell
RCM	Restrictive Cardiomyopathy
RDW	Red blood cell distribution width
RV	right ventricle
SAECG	Signal Averaged ECG
SC	Sickle cell
SCD	sudden cardiac death
SCF	stem cell factor
SKAT	sequence kernel association test
SNP	single nucleotide polymorphism
TFs	transcription factors
TIA	transient ischemic attack
VAD	ventricular assist device
VCF	variant call file
VT	variable threshold
WBC	white blood cell
WES	Whole exome sequencing
WGS	whole genome sequencing

CHAPTER 1: INTRODUCTION

1.1.PREFACE

Genetic factors not only govern how we look or act but more importantly our susceptibility to disease. Ever since it was discovered that disease could be inherited according to the rules of Mendel in the early 1900s, genetics has become a compelling area of research and the key that holds the answers to the mysteries of human disease. A lot of what we know today regarding the causes of disease, prognosis, management, and treatment could be attributed to our understanding of genetics. Despite all these discoveries, many questions are left unanswered and serve as drivers to more research which has led to tremendous collaborations between scientists around the globe.

The influence of genetics on disease can be broadly divided into two main categories: i) monogenic and ii) complex. The first describes a group of diseases that are usually rare and inherited following the expected Mendelian ratios of offspring and caused by mutations that disrupt the function of a single locus or gene. Complex traits, on the other hand, arise due to the modest effect of many variants in multiple genes (polygenic), environmental factors, as well as their intricate interplay.

Hence, genetic factors influence each group differently, and it follows that the methods used to analyze each category as well as the interpretation and clinical relevance of the genetic findings are extremely distinct. This thesis addresses both categories. I will discuss the genetics of cardiomyopathy as an example of a rare disorder, and blood cell traits as an example of complex traits. I will talk about the clinical aspect of each group and the genetic approaches used within each to hunt for novel genes.

The thesis allows a broad discussion on the methodology, analysis, tools, and the challenges that are specific to each type of diseases/traits. It highlights the importance of the advances that we have achieved in genetics in both realms of human disease and how they are used to improve our knowledge and advance treatment options in the near and far future.

1.2.CARDIOMYOPATHY: A RARE DISORDER

1.2.1. Relevant Clinical Background

The American Heart Association defines cardiomyopathies as a heterogeneous group of heart muscle disorders associated with mechanical and/or electrical dysfunction that exhibit ventricular hypertrophy or dilation often leading to progressive heart failure [1]. Several classifications of cardiomyopathies exist. One broad classification divides cardiomyopathies into primary and secondary. Primary cardiomyopathy affects specifically the heart muscle, whereas in secondary cardiomyopathy the myocardium is affected as part of a systemic disorder, such as muscular dystrophy for instance, where multiple organs are involved [1].

Cardiomyopathies are further classified into several types based on clinical manifestation. The major ones being hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), arrhythmogenic right ventricular cardiomyopathy (ARVC), restrictive cardiomyopathy (RCM, and left ventricular non-compaction cardiomyopathy (LVNC). Often these disorders are familial and caused by a mutation in one gene.

1.2.1.1. Prevalence and Pathophysiology of DCM

DCM is the second most common type of cardiomyopathy after HCM with estimated prevalence of 1 in 2500 [2]. This estimate though comes from a study conducted between 1975-1984, where awareness and diagnostic tools were limited and thus it underestimates the actual disease prevalence. Hershberger *et al.* [3] relied on recent observations of DCM cases as well as estimates of other forms of cardiomyopathy to derive a new prevalence for DCM which was suggested to be 1:250 or 0.4%, i.e. a significant 10 fold increase from the original value.

DCM is characterized by left ventricular (LV) dilatation and systolic dysfunction in the absence of other causes of systolic impairment such as severe hypertension, coronary artery disease or severe valve disease. This enlargement impairs the systolic function of the heart which reduces the heart's ability to sufficiently pump blood to the body (**Figure 1.1**). DCM mainly involves the left ventricle, however, right ventricular dilatation and dysfunction may also manifest, but are not necessary for the diagnosis of the disease. DCM is the third cause of heart failure and leads to decline in LV contractile function, arrhythmias, conduction system abnormalities, and thromboembolism. Additionally, it is the major indication for cardiac transplantation.

Early compensation for systolic dysfunction and decreased cardiac output (cardiac output= stroke volume x heart rate) is achieved by increasing the stroke volume, heart rate, or both. This compensation is also accompanied by an increase in peripheral vascular tone which helps to maintain normal blood pressure. Compensation of low cardiac output is explained by the Frank-Starling mechanism which states that myocardial force at end-diastole increases as the cardiac muscle length increases. This leads to a greater amount of force as the muscle is stretched. However, overstretching leads to reduced myocardial contractility.

Decreased cardiac output leads to neurohormonal adaptations such as the activation of the renin-angiotensin-aldosterone system (RAAS) and sympathetic nervous systems in order to maintain organ perfusion. This is accomplished by maintaining systemic pressure by vasoconstriction and by restoration of cardiac output. Other factors include vasoconstrictor endothelin and the vasodilators atrial natriuretic peptide, brain natriuretic peptide, and nitric oxide. Natriuretic peptide levels are elevated in individuals with dilated cardiomyopathy.

In the short term, neurohumoral activation is beneficial in patients with HF since it contributes to cardiac output restoration and improves tissue perfusion by increasing cardiac contractility, vascular resistance and renal sodium retention. However, these compensatory mechanisms over the long term, lead to further myocardial dysfunction due to pulmonary and peripheral edema, increased afterload, and pathologic myocardial remodeling.

The changes in the normal performance of the heart and cardiomyocytes promotes ventricular remodeling, a process that leads to changes in the heart's size (increase in myocardial mass) and function in order to improve and maintain the LV performance. Although this results in immediate benefit, in the long run, LV remodeling becomes maladaptive leading to many changes including cardiomyocyte death and fibrosis or excessive deposition of collagen that eventually causes progressive contractile dysfunction.

The changes in the normal performance of the heart and cardiomyocytes promotes ventricular remodeling, a process that leads to changes in the heart's size (increase in myocardial mass) and function in order to improve and maintain the LV performance. Although this results in immediate benefit, in the long run, LV remodeling becomes maladaptive leading to many changes including cardiomyocyte death and fibrosis or excessive deposition of collagen that eventually causes progressive contractile dysfunction [4].

1.2.1.2. Clinical Presentation and Diagnosis

Affected individuals usually present with symptoms and signs of heart failure (e.g. syncope, shortness of breath upon exertion), arrhythmias, thromboembolic disease (e.g.

stroke), or even sudden cardiac death. Clinical diagnosis is made following cardiac imaging tests such as echocardiography and magnetic resonance imaging (MRI). Echocardiography allows cardiologists to evaluate the volume and functions of the four chambers of the heart and assess pulmonary pressure and other parameters. MRI provides both anatomical and functional information.

Age of presentation of DCM is extremely wide ranging from infancy to late adulthood, but most commonly manifests between ages 30-60 [3]. In fact, the clinical phenotype of DCM is extremely variable in terms of age of onset, characteristics and severity, across families and among members of the same family.

1.2.1.3. Treatment

The main goal of HF therapy is to reduce morbidity and mortality by reducing symptoms, improving quality of life and decreasing the rate of hospitalization. Modulation of neurohormonal mechanisms is an important therapeutic target. Therefore, pharmacologic management is targeted to counteract the deleterious effects of sympathetic nervous system activation, and the Renin-Angiotensin-Aldosterone System (RAAS) to ultimately minimize cardiac remodeling. Drugs that have been shown to reduce HF-related morbidity and mortality include beta adrenergic receptor blockers, angiotensin converting enzyme inhibitors, angiotensin receptor blockers and mineralocorticoid receptor antagonists. Diuretics alleviate HF symptoms by reducing filling pressures.

In addition to pharmacologic treatment, device therapy is considered in certain individuals. Biventricular pacing provides electromechanical coordination and improves ventricular synchrony in patients with severe systolic dysfunction and conduction delay (e.g. left bundle branch block). An implantable cardioverter-defibrillator (ICD) may be used for primary or secondary prevention of sudden cardiac death. Finally, a ventricular assist device may be used in patients with end stage HF, either as destination therapy or as a bridge to cardiac transplantation.

Individuals with severe systolic dysfunction and advanced stages of heart failure are considered for cardiac transplantation.

1.2.1.4. Familial DCM and Family Screening

It is estimated that 30-35% of all DCM cases are familial [1; 5]. The familial nature of DCM is established when two or more related family members meet the diagnostic definition of idiopathic DCM [6] or when there is a family history of sudden death or conduction system disease or skeletal myopathy.

In familial DCM, the disease is caused by a mutation in a gene that is most commonly transmitted in an autosomal dominant fashion. Autosomal recessive, X-linked, and mitochondrial forms also exist, albeit less frequent. A detailed family history is indispensable in order to establish the "familial" nature of idiopathic DCM. It is recommended that all first-degree relatives and family members of individuals with idiopathic DCM undergo clinical and image testing to determine the number of affected individuals particularly since they may remain asymptomatic for many years. Early diagnosis is essential in order to monitor affected individuals, and in some cases intervene clinically, for instance by administration of anti-

hypertensive drugs or implanting an ICD. The major challenge remains that even asymptomatic individuals that get tested may have normal imaging exams but still be carriers of the disease-causing mutation. These individuals may suffer from sudden cardiac death as a first symptom. Therefore, genetic screening of family members in an attempt to identify the causal mutation is as important as clinical testing in these families and provides additional vital information for familial disease management and clinical intervention.

1.2.1.5. DCM and Reduced Penetrance

Penetrance is defined as the proportion of carriers of a disease mutation that develop the disease. It is calculated using the family pedigree. Given that DCM is a late-onset disease, unaffected young individuals are often non-informative to such analyses since they may still develop the disease at a later time. For that reason, often penetrance is calculated by considering individuals at an older age. Familial DCM is described to have reduced penetrance, meaning that not all carriers of disease causing mutations show the disease phenotype. An evidence of reduced penetrance is shown once an unaffected individual has both an affected parent and offspring. The reason for reduced penetrance is not known, but it is speculated that it is likely explained by modifier genetics and or/ environmental factors.

1.2.1.6. Other types of cardiomyopathy

1.2.1.6.1. HCM

HCM is the most common form of cardiomyopathy with estimated prevalence of 1:500 [1] , although epidemiological studies may have underestimated its prevalence since some patients remain undiagnosed due to incomplete phenotypic expression. HCM is also the most common

cause of sudden cardiac death. It is a clinically heterogeneous disorder characterized by left ventricular wall thickening (**Figure 1.1**), without LV dilation or other cardiac disease [5]. The hypertrophy observed in HCM is idiopathic and is not due to other factors such as hypertension or aortic stenosis. Clinical manifestations include diastolic dysfunction, left ventricular outflow tract obstruction, ischemia, atrial fibrillation, abnormal vascular responses and, in 5% of patients, the disease progresses to systolic impairment leading in some cases to heart failure [7].

The cause of death in HCM patients is mainly attributable to sudden cardiac death, heart failure, and embolic stroke. However, thanks to advances in management strategies, disease-related mortality is comparable to that of the general population and is estimated to be 0.5%, a substantial improvement from previous estimates of 1.3-1.4% between 1981 and 2002 [8; 9].

Management of HCM includes, among others, conventional therapy for diastolic dysfunction, atrial fibrillation, ventricular arrhythmias, and heart failure[7]. HCM is a genetic disorder with an autosomal dominant inheritance. Mutations that cause HCM are mainly in sarcomeric genes, and mutations in *MYBPC3* and *MYH7* alone explain between 70-80% of HCM cases [3; 10]. The main challenges for HCM genetic discovery, lies in exploring the influence of modifier genetics on the disease heterogeneity.

1.2.1.6.2. ARVC

Arrhythmogenic right ventricular cardiomyopathy or ARVC is characterized by right ventricular degeneration, fatty or fibrofatty tissue replacement, and ventricular arrhythmias (**Figure 1.1**). In early stages of the disease, structural changes are confined to localized regions of the right ventricle (RV), which is referred to as the “triangle of dysplasia” which includes

the inflow tract, outflow tract, or apex of the RV [11]. The disease may progress to the left ventricle affecting the posterior lateral wall and evolve into a DCM phenotype leading to biventricular heart failure in advanced stages of the disease. A left-dominant type is also observed [12].

ARVC is less common than HCM (1:5000 estimated prevalence) but is an important cause of sudden cardiac death, particularly in young athletes [13; 14]. In fact, the risk of sudden cardiac death increases in ARVC patients upon exertion. The mechanism of SCD in ARVC is cardiac arrest due to sustained ventricular tachycardia or ventricular fibrillation.

Just like other cardiomyopathies, management of the disease aims to prevent disease progression, improve the quality of life, and reduction of mortality. Clinical management thus consist of lifestyle changes, pharmacological treatment (e.g. antiarrhythmic agents or beta-blockers), ablation therapy, ICD implantation, and heart transplantation[15]. ARVC is a genetic cardiomyopathy with mainly an autosomal dominant inheritance pattern. More than 10 genes have been implicated in ARVC and desmosome related genes explain a substantial proportion of ARVC cases [16].

Figure 1.1. Heart morphologies of the most common types of cardiomyopathy

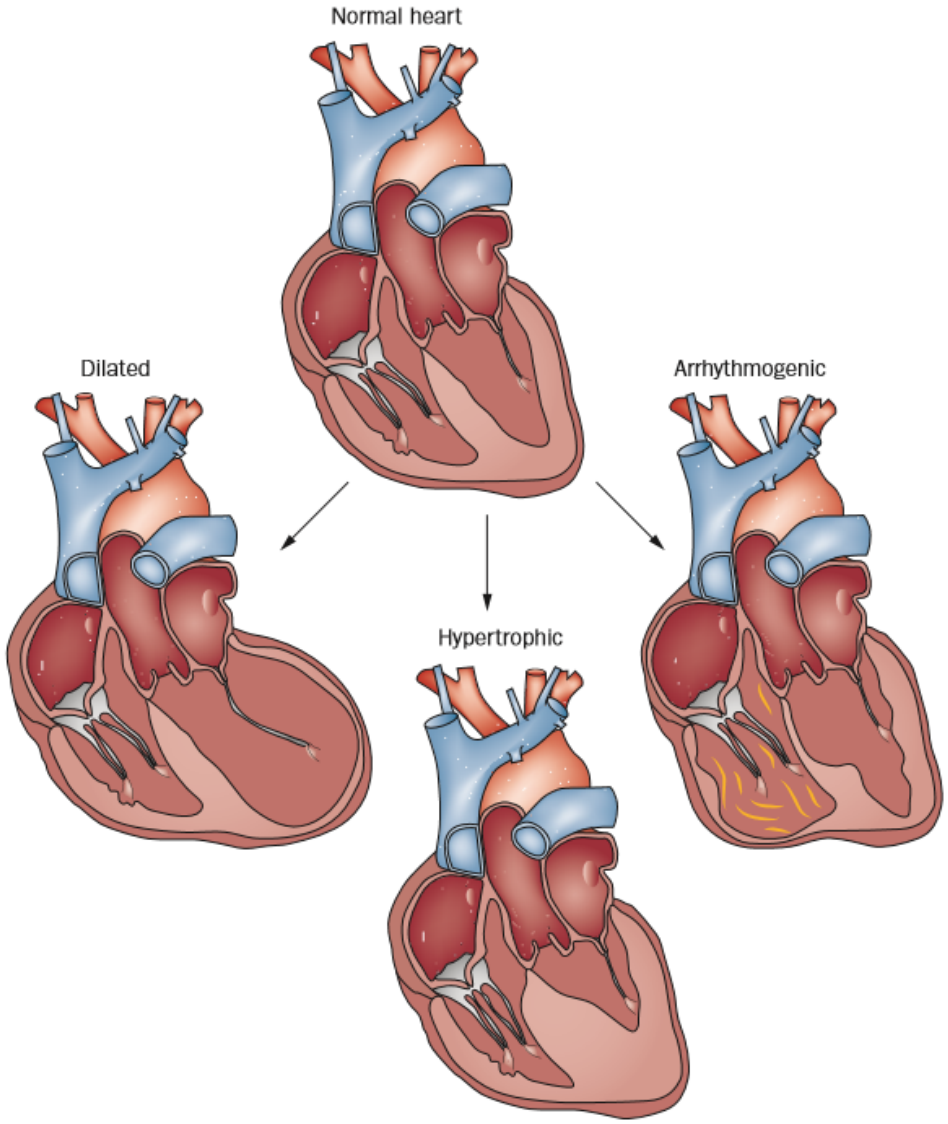


Figure 1.1. Compared to a normal heart (top), a dilated cardiomyopathy heart has a dilated left ventricle that leads to a systolic dysfunction which is measured clinically by the ejection fraction (EF) or the fraction of blood that is pumped to the rest of the body at each contraction. In hypertrophic cardiomyopathy, the ventricular wall is thickened leading to a decrease in the end-diastolic volume dimension. Arrhythmogenic right ventricular cardiomyopathy mainly affects the right ventricle and is characterized by fibro-fatty tissue replacement. Figure from [3].

1.2.1.6.3. Restrictive Cardiomyopathy (RCM)

RCM is thought to be the rarest type of cardiomyopathy. RCM causes increased stiffness in the myocardium that results in impaired ventricular filling in the presence of normal or reduced diastolic and/or systolic volumes and normal ventricular wall thickness [1; 5].

RCM does not appear to be a distinct type of cardiomyopathy as it results due to a functional rather than an anatomical defect and thus occurs in patients with other types of cardiomyopathy, mainly end-stage HCM or DCM. Nonetheless, it is thought that RCM is a genetic disease that often has an autosomal dominant inheritance. Genetic studies have implicated mainly sarcomeric genes with RCM [17; 18]; often patients in the studies had other types of cardiomyopathy with restrictive physiology [19; 20].

1.2.1.6.4. Left Ventricular Non-compaction Cardiomyopathy (LVNC)

It is not clear whether LVNC constitutes a distinct type of cardiomyopathy or is just a physiological manifestation shared by different types of cardiomyopathies. LVNC, first described as a "spongy myocardium" [21] is characterized by prominent trabecular meshwork and deep recesses in the ventricular wall that takes place in the early stages of embryogenesis [22]. Trabeculae are sheets of cardiomyocytes lined by endocardial cells that form in the ventricular cavity in early embryogenesis and increase the surface area for gas exchange. In LVNC, the myocardium is thickened and stiff.

LVNC is classified by the American Heart Association as a primary genetic cardiomyopathy [1], whereas the European Society of Cardiology considers it as an unclassified cardiomyopathy [5]. There are multiple etiologic bases for LVNC: 1) LVNC may occur in isolation ; 2) in association with other genetic diseases (cardiomyopathies) or congenital

disorders such as Ebstein's anomaly and other neuromuscular disorders [5] ; 3) may be acquired in other physiologic or pathologic conditions; 4) or it can be either permanent or transient. Hence, LVNC may originate during embryonic development or can be acquired "later in life". The population prevalence of isolated LVNC is 0.014% [23].

Clinical management of LVNC depends on the functional phenotype and related complications and includes ICD implantation, resynchronization therapy, and ablation procedure. LVNC can be familial or sporadic. Genes that have been implicated with this physiological cardiomyopathy are involved in several pathways including sarcomeric function[24], cytoskeletal organization[25], and notch signaling pathway[22].

1.2.1.7. Heart Failure

All different types of cardiomyopathy may lead to heart failure (HF) in advanced stages of the disease. HF is a major health burden worldwide. It is a relatively common condition with 50,000 new patients diagnosed each year in Canada [26]. The five year survival rate is estimated to be 50% [27] .

Despite the improvement in cardiovascular disease management and the therapeutic advances, heart failure incidents remain on the rise. The incidence of new cases of HF is increasing by 1% per year in individuals > 65 years, and it is thus estimated that the incidence of new cases may reach 1 million per year by the year 2050 in the US which is twice the current number [28]. There is a huge hope that genomics will further our understanding of the pathophysiology of heart failure which may aid in identifying individuals at higher risk based on the person's genetic/biomarker profile (disease prediction) and may lead to new drug discovery and individualized treatment.

1.2.2. DCM as a Genetic Disorder

Genetic causes of DCM account for 30-35% of cases. More than 50 genes have been shown or posited to play a role in DCM (**Table 1.1**), although less than half are thought to be definitively implicated in the disease and the rest are possible candidates. Ongoing and future studies will continuously clarify the status of DCM candidates as more well-characterized causal mutations in those genes are discovered. The genetic information gleaned so far has proven to be effective in disease diagnosis and management [29]. Genetic screening identifies a likely causal mutation in candidate genes in 30-35% of cases [30].

DCM genes encode for a wide variety of proteins that constitute the sarcomere, nuclear envelope, cytoskeleton, sarcolemma, ion channels, and intercellular junctions among others. DCM results when the function of the myocardium is weakened due to the disruption of muscle contraction, calcium homeostasis, functioning of the ion channels, or mechanic force of the myocardium.

Many of the genes causal of DCM overlap with other forms of cardiomyopathy and other disorders (e.g. channelopathies and neuromuscular disorders) [3] **Figure 1.2**. Indeed, the repertoire of DCM genes (**Table 1.1**) has expanded over the years with the evolving genetic "mapping" technologies that I will describe below.

1.2.2.1. Genetic Approaches in Monogenic Disease

1.2.2.1.1. Linkage Analysis

Most of the genes that were implicated in DCM were the result of linkage studies. Traditionally, linkage analysis served as a gene mapping tool that uses recombinant technology to map genetic markers or loci. If in a given pedigree, two genetic loci segregate

together more often than by random chance, they are said to be linked, in other words they lie close on the same chromosome. Linkage can be used to map a certain locus since the distance between two loci depends on the frequency of recombination between them. Hence, instead of distances expressed in base pairs, a genetic map gives distances in recombinational units (centimorgan, cM). This method can be exploited to locate an unknown disease locus in a pedigree by investigating the inheritance patterns of regions of the genome in affected and unaffected individuals to determine whether a genetic marker or locus is segregating with the disease. When linkage between a marker and disease is established, its extent is regarded as an approximation of the physical distance between the marker involved and the causative disease locus. Markers with known locations are used to locate the most likely position of the disease gene.

Linkage is tested by the Logarithm of Odds Ratio (LOD) which is the logarithm of the ratio of the probability of linkage by the probability of no linkage. Linkage analyses has been successful in identifying the genes linked to many Mendelian disorders such as Huntington's disease [31] , cystic fibrosis [32], and familial hypercholesterolemia [33].

For DCM, the first series of loci implicated in the disease were identified by using linkage analysis. The first wave of linkage studies identified regions of the human genome that harbored several candidate genes. The very first DCM gene identified by linkage analysis was dystrophin, DMD, in families with X-linked DCM [34; 35]. The first chromosomal region to be linked with autosomal DCM came a year later [36], although the gene was not identified. Following these studies, linkage analyses successfully identified more loci linked to DCM such as *ACTC1* [37], *LMNA* [38], *SCN5A* [39; 40] and others.

Despite the enormous contribution to gene discovery that resulted from linkage analyses, there were several caveats and challenges associated with this tool. Firstly, the method is successful in large pedigrees which are not always available. Second, if the disease is caused by a mutation in one gene only, then LOD scores from different families can be added to one another. However, the power of linkage is reduced in the presence of locus heterogeneity. For example, if the Mendelian trait can be caused by different genes in different families, then you would need more families since the analysis will rely solely upon families with high likelihoods of linkage to a particular disease locus. Third, the regions identified may span several genes, which is the reason why many earlier studies identified a region without being able to pinpoint the association to a specific gene. Hence, a fine-mapping step by sequencing the coding regions of candidate genes is required, which could be expensive and labor intensive, and thus a major limiting factor. For example, *TTN*, which was first identified in 2002 as a DCM gene [41], had been a candidate gene since 1999 by a linkage study [42]. The study had identified a linkage region that included several genes, and *TTN* was named as a possible candidate but further fine-mapping and sequencing of the region were hindered by its enormous size. However, thanks to the technological advances and the advent of next generation sequencing, the issues associated with linkage analyses could now be addressed.

Figure 1.2. Genetic heterogeneity of DCM

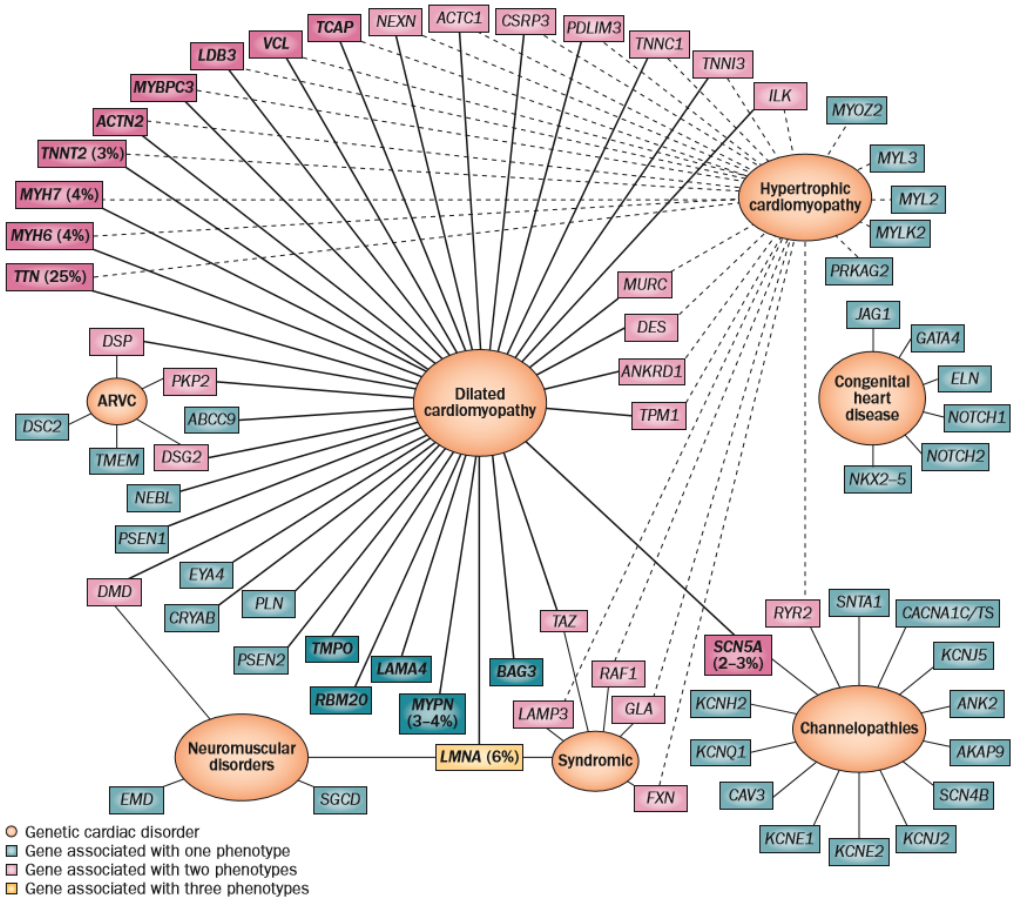


Figure 1.2. The major genes that play a role in the pathogenesis of DCM and their overlap with other cardiomyopathies as well as neuromuscular disorders and channelopathies highlighting the heterogeneity of DCM. The genes that account for the majority of DCM mutations are *TTN* and *LMNA*. ARVC: Arrhythmogenic right ventricular cardiomyopathy. Figure from [3].

1.2.2.1.2. Next generation sequencing (NGS)

Sanger sequencing was the first widely used method for sequencing. Sanger sequences are highly accurate, however, they are low throughput since they are restricted to a single DNA fragment at a time and a maximum length of 1000 bp. NGS, on the other hand, allows for massively parallel sequencing of the human genome in one experiment. Both targeted, that

is sequencing a list of candidate genes (or genomic regions), and whole-genome (or exome) sequencing can be achieved by this technology and it has been successful in identifying novel genes in both monogenic disease and complex traits.

1.2.2.1.2.1. Targeted NGS

For cardiomyopathy, the major gain of the advent of this technology was that it allowed the sequencing of *TTN* and *DMD1*, the two largest genes of the human genome which are both implicated in cardiomyopathy. Although it was already posited that they play a role in DCM, it was not feasible to sequence the whole gene to characterize and estimate the prevalence of mutations in those genes in cardiomyopathy patients. Not only did NGS allow the assertion of *TTN* as a cardiomyopathy gene [43; 44], but it also made it possible to add *TTN* to the available gene panels used in the clinics for screening patients.

1.2.2.1.2.2. Whole exome sequencing

Exome sequencing, in particular, targets the whole protein coding region (~ 20,000 genes) in a single experiment. It constitutes an attractive tool for studying monogenic diseases. Previous exome sequencing studies have served as a proof of concept that this strategy is successful for the identification of extremely rare or private causal mutations in known Mendelian disorders by sequencing the exomes of affected individuals and looking for variants segregating with disease and not present in controls or in public databases such as dbSNP [45]. Subsequent studies have been successful in identifying a number of novel genes for various Mendelian diseases and available databases of whole genome and exome sequences have become available and include sequencing data of several thousands of individuals which

improves data analysis and interpretation (see chapter 6 for a discussion). For DCM, 5 genes were discovered in one year only (*BAG3*, *ACSF3*, *AARS2*, *MRPL3*, and *GATAD1*) using exome sequencing [46; 47].

The price of sequencing one exome has dropped significantly in the past few years and currently reached ~ 600$ which makes it a very feasible technology. One of the major advantages of this technique is the untargeted approach where one examines the whole exome for novel genes without being limited to a hypothesis-driven analysis thereby increasing our chance for novel discoveries.

Unlike linkage, which yields linkage regions that need to be further fine-mapped to find the causal variation, exome sequencing enables analysts to narrow the findings to the causal mutation in an unbiased step, given a cautious and meticulous variant prioritization strategy (see chapter 6 for a discussion about the caveats of the strategy). It is imperative that the number of variants deemed to be likely pathogenic in a family decreases as the number of participants increases which facilitate the task of proving pathogenicity. However, it remains that a lower number of participants is required for exome sequencing compared to conventional linkage analyses which necessitate the availability of very large multiplex pedigrees. Another advantage of exome sequencing is that it can explore shared variants between affected members in a family as well as in sporadic cases in a population-based study, and thus is not limited to family studies.

1.2.2.2. Gene groups implicated in DCM

Genes that play a role in the pathogenesis of DCM display a wide heterogeneity and can be classified into several groups (**Figure 1.3** and **Table 1.1**). I will provide a general description of each group and its associated pathways and mechanisms that are thought to lead to DCM.

1.2.2.2.1. Sarcomeric proteins

The sarcomere is the basic unit of the myofibril found in muscle cells (Figure **1.3**). Interaction between thin filaments and thick filaments stimulates muscle contraction. Contractile force generation and its propagation to neighboring cells is essential for heart function. Several sarcomeric genes have been associated with DCM such as *ACTC1* (actin), *MYBPC3, MYH6, MYH7, TNNC1, TNNI3, TNNT2, TPM1,* and *TTN.* Truncating mutations in *TTN* alone account for the majority of DCM cases and is estimated to be around 20% [43; 48]. TTN is a giant protein that includes > 34,000 amino acids and interacts with both thin and thick filaments to participate in sarcomere assembly and force generation. So far, it is believed that pathogenic mutations in *TTN* are nonsense, frameshift, and splice site mutations. Although missense variants in *TTN* are not considered to have an associated medical significance, some reports suggested that in certain cases missense variants may be pathogenic [49-51]

1.2.2.2.2. Structural proteins

Structural proteins provide structural integrity to the sarcolemma (the plasma membrane of muscle cells) and ensure that the contractile force is transmitted from the sarcomere to the sarcolemma and the extracellular matrix (Figure **1.3**). Genes in this group include *DES, DMD,*

CSRP3 , CAV3, and others. *DMD* was the first gene to be linked to DCM. *DMD* was first linked to Duchene muscular dystrophy which is an X-linked hereditary disease characterized by gradual muscle weakness and, often, cardiac disorders. Patients with cardiac disorders often develop heart failure at the last stage of the disease. However, mutations in *DMD* can cause X-linked DCM without clinical signs of skeletal muscle weakness [35]. *DMD*- associated DCM has a severe prognosis and leads to the death of the affected individuals at early age (10-20 years of age). Other structural proteins constitute the desmosome, an example of a filamentous system that is responsible for the propagation of contractile force from one cell to the other which maintains both the mechanical and electrical integrity of the heart and includes *DSP, DSG2, DSC2,* and *PKP2*. Desmosomal genes are mainly associated with ARVC, but mutations in those genes have also been described for DCM [52; 53].

1.2.2.2.3. Nuclear proteins

Nuclear proteins play an important role in chromatin organization and transcriptional regulation. Genes in this group that have been linked to DCM include *LMNA, EMD,* and *TMPO* (**Figure 1.3** and **Table 1.1**). *LMNA* mutations are the most frequent in this group and the second most frequent overall after *TTN*. They account for 5-8% of familial DCM cases and up to 11% of sporadic cases [54]. They are inherited mainly in an autosomal dominant pattern. *LMNA* encodes the lamin A/C protein which is located in the nuclear lamina. Mutations in *LMNA* are known to cause Emery-Dreifuss muscular dystrophy (EDMD) which is a genetic disorder characterized by progressive skeletal muscle weakness [55; 56]. Many patients with this disorder also suffer from DCM. Soon after, it became evident that mutations

in *LMNA* could cause DCM without the presence of skeletal muscle weakness [38]. *LMNA* associated DCM is usually accompanied with conduction system disease and arrhythmias [57]

1.2.2.2.4. Ion channel proteins

Calcium homeostasis plays a critical role in muscle contraction. Ca^{2+} enters the cardiomyocytes through voltage-gated channels present on the sarcolemma and triggers sarcomere contraction (Figure **1.3**). Dysfunction of these ion channels leads to muscular contraction deterioration causing DCM and other cardiac problems. Genes in this group include *SCN5A, KCNQ1, ABCC9,* and *PLN. SCN5A* encodes a sodium channel subunit which plays a critical role in the regulation of the heartbeat. Mutations in *SCN5A* are the most frequent and cause a modification in the electrical activity of sodium channels which destabilizes the contraction process of the cardiomyocytes. *SCN5A* mutations are associated with early-onset DCM accompanied with conduction abnormalities and atrial fibrillation [40].

1.2.2.2.5. Other proteins

Other proteins involved in DCM include transcription factors (TFs) such as *TBX5, TBX20, NKX2-5, GATA4, FOXD4,* and others. These TFs regulate the expression of cardiac structural and regulatory proteins necessary for the heart function. Many Z-disk proteins have also been implicated in DCM such as *ACTN2, LDB3*, and *MYPN*. The proteins encoded by these genes ensure the transmission of the contractile force between adjacent sarcomeres within the muscle fibers. Other genes are involved in signaling pathways (*BRAF*), heat shock proteins (*BAG3, CRYAB*), and the cytoskeleton (*PLEC, SGCD, VCL*).

Figure 1.3. Functional groups of genes associated with dilated cardiomyopathy.

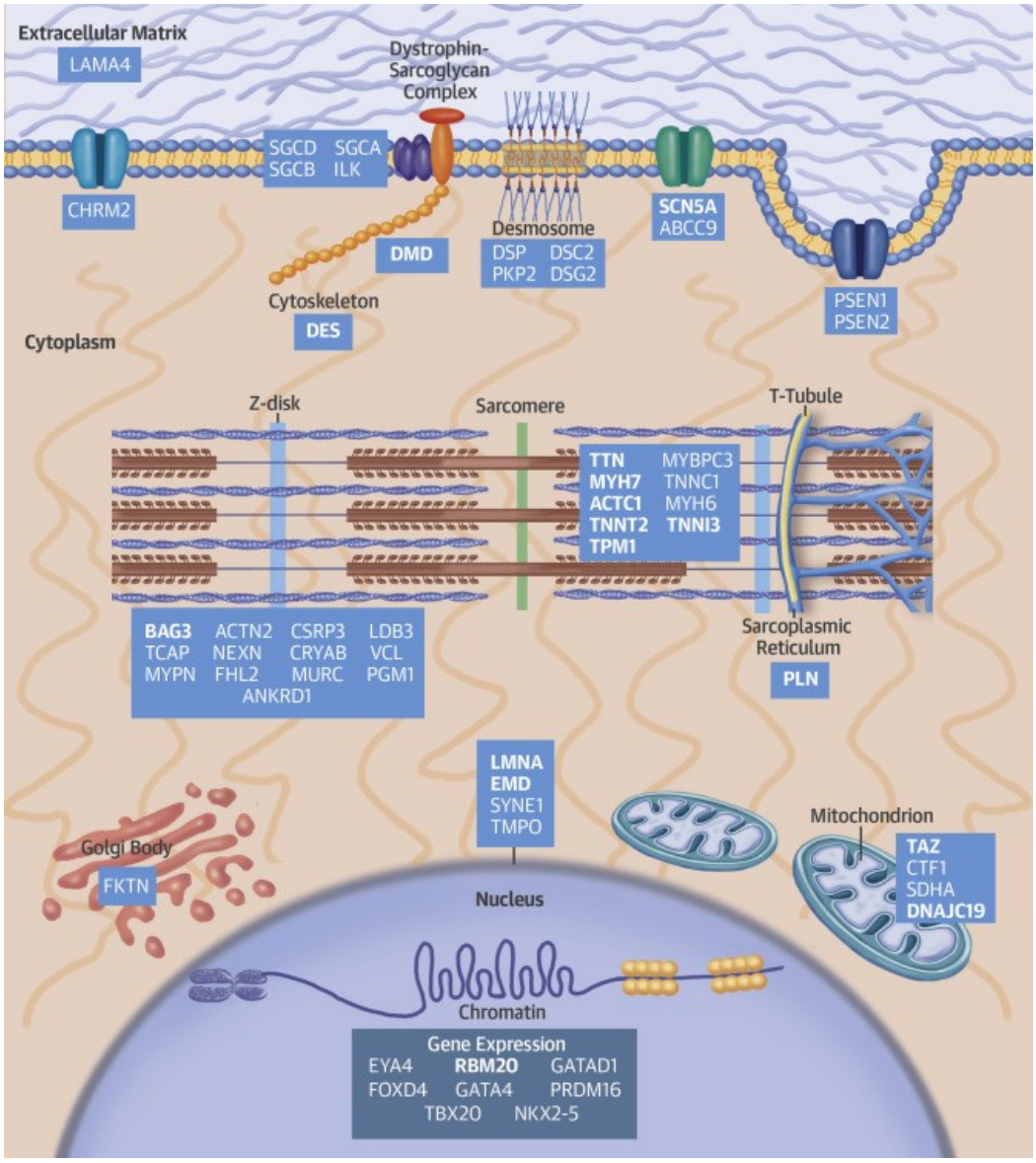


Figure 1.3. Examples of genes associated with DCM grouped based on where they are located in the cardiomyocyte. Proteins located on the nuclear envelope are involved in gene expression or in maintaining the structural integrity of the cytoskeleton. Proteins within the sarcomere are responsible for muscle contraction. Other proteins form filamentous systems such as desmosomes which connect adjacent cardiomyocytes. Ion channels maintain the electrical activity of the heart. Figure from [58]

Table 1.1. Genes associated with dilated cardiomyopathy (DCM).

Gene	Protein	Function
ABCC9	Sulfonylurea receptor 2	Regulates ion transport, essential for normal heart function
ACTC1	Actin, alpha cardiac muscle 1	Muscle contraction
ACTN2	Alpha-actinin-2	Anchor for myofibrillar actin filaments
ANKRD1	Ankyrin repeat domain-containing protein 1 (Cardiac ankyrin repeat protein) (Cytokine-inducible gene C-193 protein) (Cytokine-inducible nuclear protein)	Interacts with sarcomeric proteins: myopalladin and titin
BAG3	BAG family molecular chaperone regulator 3BAG family molecular chaperone regulator 3	Inhibits apoptosis
CHRM2	Muscarinic cholinergic receptor	Modulation of potassium channels
CRYAB	Alpha-crystallin B chain	Has chaperone-like activity
CSRP3	Cysteine and glycine-rich protein 3	Organization of cytosolic structures in cardiomyocytes
CTF1	Cardiotrophin-1	Cytokine activity
DES	Desmin	Muscle contraction
DMD	Dystrophin	Cytoskeleton integrity; Sarcolemma stability
DNAJC19	DNAJ (Hsp40) homolog	Heat shock protein
DOLK	Dolichol kinase	Plays a role in the Endoplasmic reticulum
DSC2	Desmocollin 2	Component of the desmosome
DSG2	Desmoglein 2	Component of the desmosome
DSP	Desmoplakin	Involved in the formation of desmosomal complexes
EMD	Emerin	Maintains function of skeletal and cardiac muscle
EYA4	Eyes absent homolog 4	Transcriptional activator
FBXO32	F-box protein 32/atrogin-1	FoxO family signaling
FHL2	Four and a half LIM domains protein 2	Extracellular membrane assembly
FKTN	Fukutin	Glycosylation of alpha-dystroglycan in skeletal muscle
FOXD4	Forkhead box protein D4	Transcription factor activity
GATAD1	GATA zinc finger domain-containing protein 1	Regulates gene expression
ILK	Integrin-linked kinase	Cellular signaling
JUP	Junction plakoglobin	Cytoskeleton integrity
LAMA4	Laminin subunit alpha-4	Component of the extracellular matrix
LDB3	LIM domain-binding protein 3	Cytoskeleton assembly
LMNA	Lamin A/C	Nuclear stability, chromatin structure and gene expression
MURC	Muscle-related coiled-coil protein	Involved inmyofibrillar organization
MYBPC3	Myosin binding protein C, cardiac	Component of the A bands in striated muscle; muscle contraction
MYH6	Myosin, heavy chain 6, cardiac muscle, alpha	Muscle contraction
MYH7	Myosin, heavy chain 7, cardiac muscle, beta	Muscle contraction
MYPN	Myopalladin	Component of the sarcomere
NEBL	Nebulette	Assembly of the Z-disk
NEXN	Nexilin (F actin binding protein)	Maintenance of Z line and sarcomere integrity
NKX2-5	NK2 homeobox 5	Transcription factor activity
PGM1	Phosphoglucomutase 1	Breakdown and synthesis of glucose

Gene	Protein	Function
PKP2	Plakophilin 2	Component of the desmosome
PLN	Phospholamban	Key regulator of cardiac diastolic function
PRDM16	PR domain containing 16	Transcriptional regulator
PSEN1	Presenilin 1	Intracellular signaling
PSEN2	Presenilin 2	intracellular signaling
RBM20	RNA-binding protein 20	Regulates splicing of TTN and other genes
SCN5A	Sodium channelprotein type 5 subunit alpha	Controls Na+ transport
SDHA	Succinate dehydrogenase enzyme	Plays a role in mitochondria
SGCA	α-sarcoglycan	Component of the sarcoglycan complex, a subcomplex of the dystrophin-glycoprotein complex
SGCB	β-sarcoglycan	Component of the sarcoglycan complex, a subcomplex of the dystrophin-glycoprotein complex
SGCD	δ-sarcoglycan	Component of the sarcoglycan complex
SYNE1	Nesprin-1	Cytoskeletal organization
TAZ	Tafazzin	expressed at high levels in cardiac and skeletal muscle;Some isoforms may be involved in cardiolipin (CL) metabolism
TBX20	T-box 20	Transcription factor activity
TCAP	Telethonin	Muscle assembly
TMPO	Thymopoietin	Structural organization of the nucleus
TNNC1	Troponin C, slow skeletal and cardiac muscles	Muscle contraction
TNNI3	Troponin I, cardiac muscle	Muscle Contraction
TNNT2	Troponin T, cardiac muscle	Muscle Contraction
TPM1	Tropomyosin alpha-1 chain	Muscle Contraction
TTN	Titin	Assembly and functioning of striated muscle
VCL	Vinculin	Cell-matrix adhesion and cell-cell adhesion

Table 1.1. The 59 genes that are implicated or posited to be implicated in DCM and a brief description of their function. Figure adapted from the supplementary information of [59].

1.3.BLOOD CELL NUMBER, SIZE, OR CONTENT: CLASSIC HUMAN COMPLEX TRAIT

1.3.1. Clinical Background

The major constituents of blood are the red blood cells, white blood cells, and platelets. Red blood cells are involved in oxygen transport to the organs of the body, white blood cells play a major role in immunological functions, and platelets are important for blood clotting. The focus of this thesis will be on red blood cells and platelets.

1.3.1.1. Red blood cells

Red blood cells (RBC)s, or erythrocytes are enucleated biconcave disks that are essential for gas exchange and for regulating vascular tone [60]. They are the most common cell type in blood. RBCs transport O_2 from pulmonary capillaries to tissue capillaries, in exchange for CO_2. The gases are mainly carried by hemoglobin, the major cytoplasmic protein of RBCs. The average diameter of an erythrocyte is 8 μm, although due to its biconcave shape and membrane flexibility, it can enter capillaries with a diameter < 8 μm [61]. As red cells age, their membranes lose their flexibility and become rigid.

1.3.1.1.1. RBCs life cycle

RBCs are produced from hematopoietic stem cells (HSCs) of the bone marrow (Figure **1.4**) and undergo a series of maturation steps before the production of a mature RBC, or erythrocyte (Figure **1.5**). Once HSCs start proliferating and differentiating, many blood cells can be produced from each individual stem cell. The different stages of RBC production, or erythropoiesis, are progenitor cells, proerythroblasts, erythroblasts, reticulocytes, and

eventually erythrocytes. The primary hormone which controls erythropoiesis is the kidney-derived cytokine, erythropoietin, or EPO which is activated in hypoxic conditions and leads to the stimulation of progenitor cells [62]. In addition to EPO, other hormones regulate erythropoiesis including insulin like stem cell factor (SCF), growth factor 1 (IGF-1), glucocorticoids (GCs), and IL-3, and IL-6, as well as a number of important transcription factors such as GATA-1, EKLF, SCL and LMO2 [63]. Through the differentiation steps, the cells undergo substantial changes, most importantly, decrease in cell size, enucleation and expulsion of other organelles [61]. The life cycle of human RBCs is 120 days after which they are removed from circulation by macrophages.

Figure 1.4. Hematopoietic stem cell differentiation.

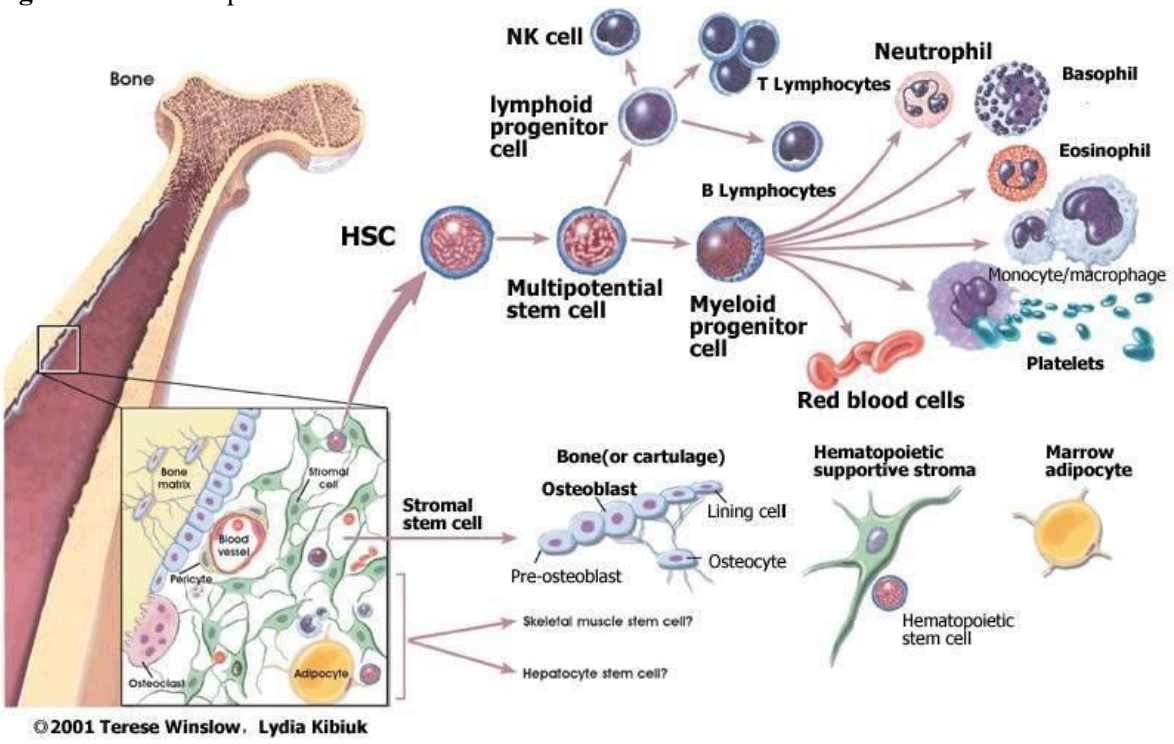


Figure 1.4. All blood cell types are derived from the hematopoietic stem cells (HSC) through hematopoiesis. HSCs are located in the bone marrow. An HSC differentiates into a multipotent stem cell that also differentiates into a lymphoid progenitor cell and a myeloid progenitor cell. Both red blood cells and platelets are derived from the myeloid progenitor cell.

Figure 1.5. Erythropoiesis.

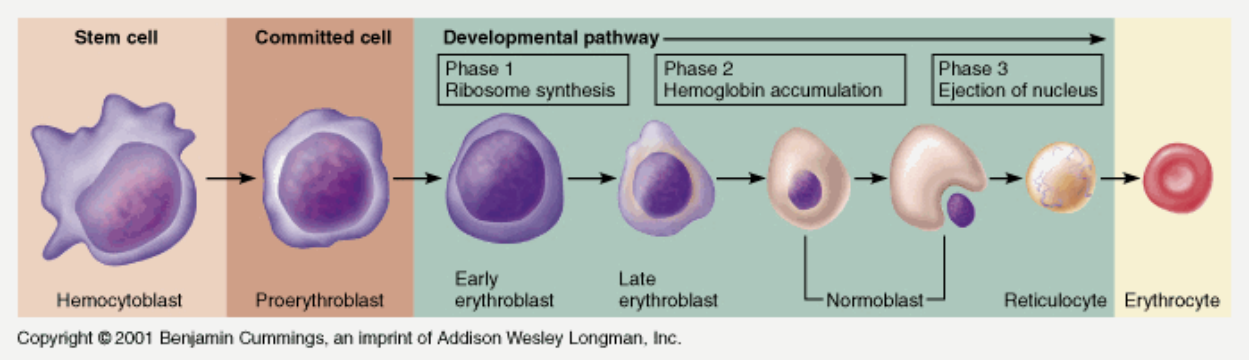


Figure 1.5. Erythropoiesis is the process that leads to the production of red blood cells or erythrocytes. In the bone marrow, a hemocytoblast differentiates into an erythroblast. In the differentiation stage from erythroblast to normoblast hemoglobin accumulation takes place. The normoblast expels its nucleus and organelles and becomes a reticulocyte. At this stage, reticulocytes are released into the circulation to eventually give rise to erythrocytes.

1.3.1.1.2. RBC levels and related disorders

Several disorders are associated with RBCs, the most common of which are anemias which lead to poor gas exchange. Indicators of anemia include reduced red cell numbers, hemoglobin content and hematocrit (percentage of red cells in blood). Other blood indices are also used to distinguish the different types of anemia, a more detailed discussion is presented below.

There are several causes of anemia. In addition to nutritional deficiency (iron, Vitamin B12, folates), anemia can result due to decreased EPO production (due to renal failure for example), when erythropoiesis is impaired or suppressed (eg. aplastic anemia, a rare disorder where the bone marrow does not make enough new blood cells), or due to defects in hemoglobin synthesis (e.g hypochromic anemia which is characterized by pale erythrocytes due to reduced hemoglobin) [61]. Iron deficiency anemias result when iron is not well absorbed

or transported. Hemolytic anemias are caused by severe destruction of RBCs (due to infections for instance, or mutations in genes such as *G6PD*).

On the other hand, overproduction of RBCs due to increased EPO production can cause polycythemias or erythrocytosis [64]. Immature RBCs can also become oncogenic albeit rarely, which results in nucleated RBCs in the circulation, a condition called erythroleukemia [63].

Hemoglobinopathies are disorders that result due to mutations in globin genes; such as thalassaemias which are caused by mutations in the adult α and β globin genes that change the cell morphology leading to red cell destruction. Another example is sickle cell (SC) disease, where an abnormal β globin protein results in the production of RBCs with a characteristic sickle shape [65]. These sickle RBCs are hard and inflexible, often forming clumps that stick to blood vessels increasing the risk of blocking the blood flow and resulting in various complications. Sickle blood cells have a lower life span compared to normal cells (10-12 days).

1.3.1.1.3. RBC indices

In addition to RBC count, there exist other vital RBC indices which are important indicators of an individual's health. The measurements of these indices are obtained by the CBC test. These calculations and values are generally determined by hematologic machines that analyze the different components of blood.

1.3.1.1.3.1. Hemoglobin and Hematocrit

Hemoglobin is the major protein in red blood cells and the carrier of O_2 and CO_2. Hemoglobin comprises four globin subunits that are connected together, 2α and 2β chains each surrounding the molecule heme. At the center of the heme molecule is iron which is essential for gaseous transport. Hemoglobin is responsible for the characteristic redness of blood and plays a major role in maintaining the shape of red blood cells. Abnormal hemoglobin is what gives the sickle shape of RBCs in SC disease. Hematocrit is defined as the ratio of the volume of RBCs to the volume of whole blood and is expressed as a percentage. Normal values range between 38-48% for hematocrit and between 12-18 gm/dl for hemoglobin depending on age and gender. Low levels of hemoglobin or hematocrit indicate anemia that results from decreased production, excessive loss, or destruction of red blood cells. High levels, on the other hand, can indicate polycythemia (excessive amount of RBCs).

1.3.1.1.3.2. Mean corpuscular volume (MCV) and Mean corpuscular hemoglobin (MCH)

MCV is the average volume (or size) of a red blood cell and is calculated using the hematocrit and red cell count values. Normal range may fall between 80 to 100 femtoliters. MCV is always interpreted along with MCH. Both measurements increase and decrease in the same conditions. MCH is the amount of hemoglobin in one red blood cell. This is a calculated value derived from the measurement of hemoglobin and the red cell count. Optimum values are between 28 and 32 micrograms. MCV and MCH are increased in conditions that include, folic acid or vitamin B12 deficiency anemia, and hypothyroidism. They are decreased on the other hand, in thalassemia and iron deficiency anemia.

1.3.1.1.3.3. Mean Corpuscular Hemoglobin Concentration (MCHC)

MCHC is very related to MCH and refers to the average concentration of hemoglobin in a given volume of red cells. The hemoglobin and hematocrit measurements are used to derive the value. Normal range is 32% to 36%. In most but not all cases, MCHC increase and decrease with MCH and MCV values.

1.3.1.1.3.4. Red blood cell distribution width (RDW)

RDW is a measurement of the variability of red cell size and is interpreted along with MCV, MCH, and MCHC mainly to distinguish different types of anemia. RDW is derived from the RBC distribution curve which is generated automatically by the hematologic analyzers. RDW measures the variability of the RBC width and not the actual width of individual cells. Normal range is 11% to 14% and indicates that the red cells are mostly the same size. Higher numbers indicate greater variation in size meaning that there are small and large red blood cells. Since newly made cells (reticulocytes), B12 and folic acid deficient cells are larger than iron deficient cells, RDW helps to clarify if an anemia has multiple components. Hence, RDW is increased in the presence of B12, folic acid, hemolytic anemia (premature destruction of erythrocytes), or sideroblastic anemia (failure to incorporate iron into the heme molecule), as well as other liver disease. It is decreased in iron deficiency or Vitamin B6 anemia, as well as rheumatoid arthritis.

1.3.1.2. Platelets

Platelets play a major role in wound healing and preventing excessive bleeding. They also participate in other biological functions such as immunity, inflammation, and tissue

regeneration. Platelets are discoid shaped cells and have a diameter of 0.5-3μm and are the smallest corpuscular components in the circulation. Platelets circulate in the blood for 10 days on average and are then removed by macrophages. Circulating platelets prevent blood loss at sites of vessel injury by adhering to the vessel and forming a platelet plug, or thrombus. Upon adhesion, platelets become activated and release a number of proteins and molecules which contribute to the stabilization of the platelet plug formed at the site of injury. The adhered platelets recruit more platelets which is referred to as "platelet aggregation" and form a thrombus at the site of injury.

1.3.1.2.1. Thrombopoeisis

Thrombopoiesis is the process in which platelets are formed. Platelets derive from megakaryocytes which derive from hematopoietic stem cells (Figure **1.4**). Thrombopoiesis is regulated primarily by thrombopoietin, which stimulates platelet production, and to a lesser extent by inflammatory stimuli such as IL-6.

1.3.1.2.2. Platelet indices

Platelet indices are useful to infer certain health conditions. The two major indices discussed here are platelet count and mean platelet volume.

1.3.1.2.2.1. Platelet count

Platelet count is simply the count of the number of platelets in the circulation. The normal range of platelet count in the blood is 150,000 - 450,000/μL blood. Deviation from the normal values indicates disorders of platelet number or function. Low platelet count

(thrombocytopenia) may indicate certain viral or bacterial infections, autoimmune diseases, liver disease, or bone marrow disorders. High platelet count (thrombocytosis) increases the risk of thrombotic events. Thrombocytosis can be caused by a bone marrow disease and referred to as essential or primary thrombocytosis, or it may be due to an underlying disorder (reactive) such as infections.

1.3.1.2.2.2. Mean Platelet Volume (MPV)

MPV is a parameter used to measure the size of platelets. Normal values range between 7 and 11 fl. Size of platelets vary within the same individual, and it has been shown reproducibly that larger platelets are metabolically more active and have a higher prothrombotic potential [66-68]. MPV is a simple way to evaluate platelet activity. In fact, higher MPV is associated with increased platelet aggregation and expression of adhesion molecules, indicators of platelet activity. Further, MPV is elevated in individuals with CVD risk factors such as diabetes mellitus, obesity, hypertension, and smoking, which makes it a heavily studied parameter for its potential to predict CVD [69-73].

1.3.1.2.3. Platelets and Cardiovascular Disease (CVD)

Several studies have shown that platelets are activated during atherothrombotic events. Platelets secrete a large number of components that control coagulation, inflammation, thrombosis, and atherosclerosis [74; 75]. In addition, antiplatelet drugs are proven to reduce the risk of cardiovascular events in patients with established coronary artery disease [76] suggesting that platelets play a role in atherothrombosis (**Figure 1.6**). Thrombus formation may become pathologic once platelet activation becomes exaggerated or unnecessary which leads to

formation of thrombi in intact tissues that could detach and cause either myocardial infarction or stroke [77; 78]

Platelet count has inconsistently been associated with CVD risk where some studies suggested that higher platelet count is associated with a higher risk of CVD [79-81] and others showing that patients with atherothrombotic events have a lower platelet count[82; 83]. One explanation could be that when an atherothrombotic event occurs, platelets would rush to the thrombotic site and thus there is less platelets in circulation [84]. On the other hand, MPV studies have been more consistent suggesting that MPV is positively associated with thrombotic events [80; 82; 85-87]. Further studies have shown that MPV could in fact predict the occurrence of MI independent of other CVD factors [86; 88]. These reports highlight the clinical importance of platelet parameters in the context of CVD, although these studies still suffered from the influence of other comorbidities and drug therapies as the patient subjects, for the most part, were individuals with either some form of cardiovascular disease or at a high risk for it and did not represent the general population.

Figure 1.6. Platelets mediating hemostasis and thrombosis.

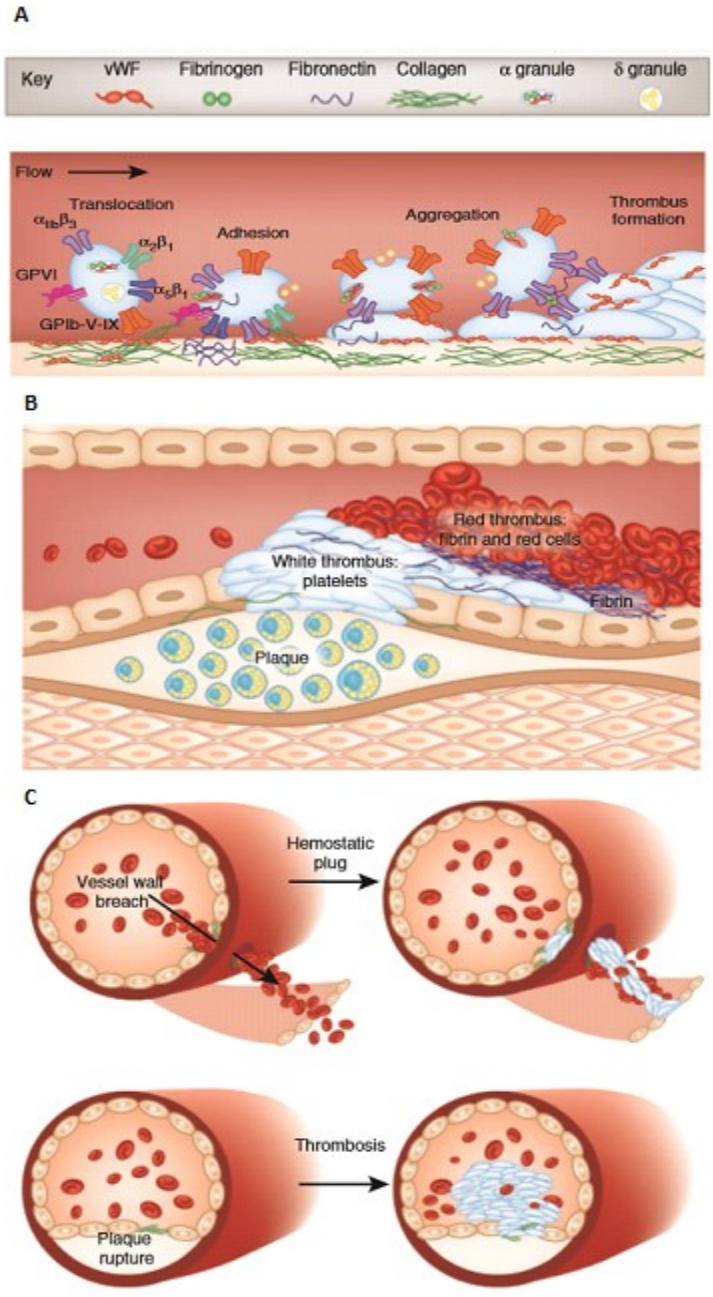


Figure 1.6. A) Platelets are captured at the site of injury (adhesion) from flowing blood. They are activated and recruit more platelets that aggregate at the site (aggregation) as well as other cellular components to form a thrombus. B) Arterial thrombi at a site of atherosclerosis plaque rupture. A platelet-rich white thrombus is formed as well as a red thrombus composed of red blood cells and fibrin. C) A hemostatic plug is formed when there is an injury in the vessel wall. The plug forms at the wall of the vessel and does not extend to the lumen. A thrombotic plug builds up on an atherosclerotic plaque and extends to the lumen, restricting blood flow. A hemostatic plug forms within minutes of injury to stop bleeding, whereas an arterial thrombi can form over a long period of time by building up over existing plaque. Figure adapted from [89]

1.3.2. Genetic approaches to study complex traits

1.3.2.1. Linkage and Association studies

Complex traits are thought to be the result of an interplay between genetic and environmental factors. Twin and family studies have provided evidence of genetic heritability for those common diseases. For example the heritability of MI is ~ 40-60% [90] and blood cell traits is ~ 37-57% [91], whereas the heritability of schizophrenia is estimated to be up to 80% [92]. Thus uncovering the genetic factors that contribute to complex traits would give a lot of insight about the underlying pathophysiology and mechanisms of disease.

The different approaches that are used to study the genetics of complex traits were enabled by the advances in genomic and computational technologies. Linkage and association studies are the two primary methods used to locate disease genes. Linkage studies are best suited to assess the co-segregation of disease loci within families and were very successful in identifying genes for monogenic diseases as mentioned above. However, linkage is not a powerful tool to capture variation of low effect size and hence has had limited success in gene discovery for complex traits [93] which are influenced by multiple small effect variants.

Population-based analyses provide more power to detect small effect variants and are more suited to study the genetics of complex traits. Prior to 2007, the majority of association studies came as candidate gene studies which attempted to test the association between genetic markers and disease susceptibility by sequencing genes that are suspected to be implicated in the etiology of a the studied trait. However, owing to the technological advances, it became possible to carry out association analyses genome-wide and in an unbiased approach that is not limited to candidate genes.

1.3.2.2. Genome-wide Association Studies (GWAS)s

In the mid-1990s, a systemic genome wide approach to association studies was proposed [94-96] and suggested the creation of a catalogue of human genetic variants to be tested for association with disease risk. In 2000, the first draft of the human genome project was completed which facilitated the creation of a detailed map of genetic variation throughout the human genome. This led to the discovery of millions of SNPs. Many of those are inherited together as haplotype blocks through a phenomenon called linkage disequilibrium (LD), or the nonrandom association between alleles of different SNPs [97]. The HapMap project [98] created a comprehensive map of haplotype blocks and LD estimates between SNPs which allows researchers to genotype a subset of SNPs and impute or predict the variant genotypes of the others. Imputation thus makes it possible to analyze a vast genomic region without the cost of genotyping millions of variants.

These milestones in the genomic era paved the way to efficiently detect the role of genetic variation in modulating various phenotypes and disease susceptibilities using genome-wide association studies (GWAS)s. GWASs relied mainly on the "common disease common variant" hypothesis, which postulates that common diseases are influenced by common variants. The rationale of this hypothesis is driven by the idea that since those diseases are common in the population, then their underlying genetic factors must be common as well. While mutations underlying monogenic diseases are extremely rare since they usually cause deleterious phenotypes and are thus under purifying selection, on the other end of the spectrum, variants that have relatively less impact on reproductive fitness due to their small phenotypic effects have reached high frequencies in the populations and have been postulated to be responsible for much of the genetic etiology of common diseases [94-96] (Figure **1.7**). More

than 32,000 variants have been associated with various traits and diseases and can be found in the GWAS catalogue (https://www.ebi.ac.uk/gwas/), the majority of which are common and have a small individual effect size on disease. The exact identity of the genes driving the association of a particular SNP with a trait though, cannot be determined on the basis of GWA studies data alone. The most proximal genes are only "best guesses" for the gene containing the causal variant until the functional mechanism is revealed.

Figure 1.7. The generally accepted model of disease susceptibility and variant allele frequency

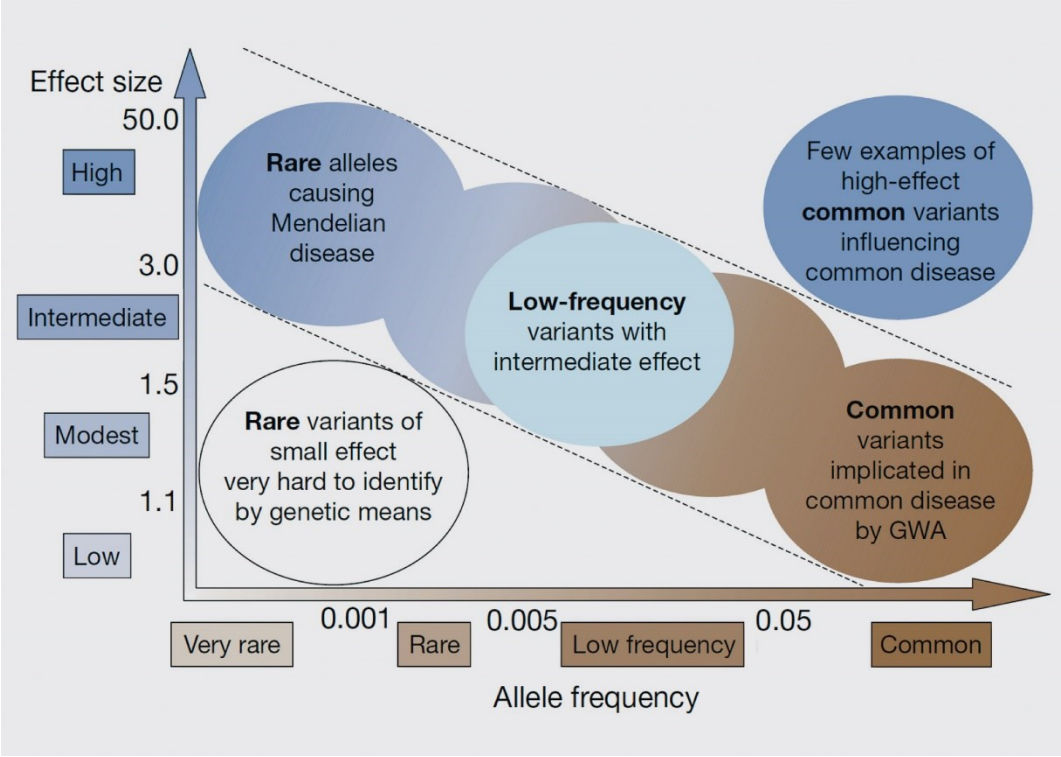


Figure 1.7. In this model, allele frequency is inversely related to the effect size. Common variants are expected to have a low effect size on disease and rare variants which are subject to strong purifying selection would have a stronger effect size on the disease. Allele frequencies that fall in the middle would have an intermediate effect size. It is worth mentioning that this is the general expected model based on the CDCV hypothesis, but some variants will certainly deviate from the model, for instance, not all rare and low-frequency variants would have a high or intermediate effect size. Figure from [99]

Following from the conclusion that common variants have a low effect size, it is necessary to look for other factors of genetic variation. The term "missing heritability" has been coined to refer to the genetic factors that may explain the remaining genetic component of the phenotypic variance [100]. It is generally accepted that the CDCV hypothesis does not give the complete picture and that complex traits are generally influenced by a multitude of common (MAF > 5%), low-frequency (1% < MAF < 5%), and rare variants (MAF < 1%) (the definition of cutoffs varies), with small to strong effect sizes in addition to other intricate

biological processes such as epigenetic modifications, gene-gene and gene-environment interactions (see chapter 6 for a discussion).

1.3.2.3. The birth of the exomechip

GWAS studies have been successful in identifying a large number of common variants associated with blood cells and other complex traits. However, the variants identified for the most part have a low effect size and explain a small proportion of any given trait's phenotypic variance. In an attempt to increase the ability of capturing rare variation with higher effect size, the exomechip was created. The exomechip array is enriched for rare coding variants chosen from existing sequenced genome and exome datasets. It also includes more than 5,000 GWAS tag SNPs which allows to assess to a certain extent the independent association of additional low-frequency coding variants. Using the exomechip has been successful in identifying several novel low frequency variants for various traits such as lipid traits [101; 102], blood pressure [103; 104], diabetes [105], and height [106].

1.3.3. Genetic Findings of Blood Cell Traits

I will provide a summary of genetic findings of blood cell traits through the following review.

1.3.3.1. Lessons and Implications from GWAS Findings of Blood Cell Phenotypes

Authors:

Nathalie Chami and Guillaume Lettre

Reference:

Chami N, Lettre, G. Lessons and Implications from GWAS Findings of Blood Cell Phenotypes. Genes (Basel). 2014 Jan 27;5(1):51-64.

Authors' Contributions:

NC and GL wrote the manuscript.

1.3.3.1.1. Abstract

Genome-wide association studies (GWAS) have identified reproducible genetic associations with hundreds of human diseases and traits. The vast majority of these associated single nucleotide polymorphisms (SNPs) are non-coding, highlighting the challenge in moving from genetic findings to mechanistic and functional insights. Nevertheless, large-scale (epi)genomic studies and bioinformatic analyses strongly suggest that GWAS hits are not randomly distributed in the genome but rather pinpoint specific biological pathways important for disease development or phenotypic variation. In this review, we focus on GWAS discoveries for the three main blood cell types: red blood cells, white blood cells and platelets. We summarize the knowledge gained from GWAS of these phenotypes and discuss their possible clinical implications for common (e.g. anemia) and rare (e.g. myeloproliferative neoplasms) human blood-related diseases. Finally, we argue that blood phenotypes are ideal to study the genetics of complex human traits because they are fully amenable to experimental testing.

Keywords: GWAS; hemoglobin; hematocrit; red blood cell; erythrocyte; white blood cell; leukocyte; platelet; human genetics

1.3.3.1.2. Genetics of Red Blood Cells, White Blood Cells and Platelets

Blood is mostly composed of plasma and blood cells and plays a major role in a variety of functions involved in general human homeostasis: it transports oxygen, nutrients and hormones to tissues, removes waste, performs immunological functions and contributes tissue damage repair through coagulation. The main three blood cell types carry out most of these activities: red blood cells (RBC, or erythrocytes) transport oxygen, white blood cells (WBC, or leukocytes) coordinate some of the immune responses, and platelets are the bricks that form blood clots to prevent excessive bleeding. All of these cell types originate through proliferation and differentiation from common precursors (hematopoietic stem cells) [107].

An aberrant number, size or feature of the three main blood cell types characterizes multiple human diseases (**Table 1.2**). In many cases, the triggering factor is of environmental origin, often poor nutrition or infections (*e.g.* malaria, HIV). Germline and somatic mutations can also cause severe blood disorders, such as mutations in glucose-6 phosphate dehydrogenase (*G6PD*) which is responsible for chronic hemolytic anemia or mutations in oncogenes or tumor suppressor genes that result in leukemia. It is also known that blood cell phenotypes vary between healthy individuals, and that some of this inter-individual variation is controlled by genetics. In a large study of healthy Sardinians (N=6,148), the heritability estimates for RBC, WBC and platelet counts were, respectively, 0.67, 0.38 and 0.53 [108]. Similar heritability estimates were obtained when analyzing phenotype concordance in healthy monozygotic and dizygotic twins from the United Kingdom [91]. These results indicate that a

large fraction of the phenotypic variation in these blood traits is controlled by DNA sequence variants segregating in healthy individuals.

Table 1.2. Main blood cell traits routinely measured in standard complete blood count (CBC).

Trait	**Description**	**Unit**
Red blood cell (RBC) count	Count of RBC per microliter	Million cells per microliter ($x10^6$/uL)
Hemoglobin (HGB)	Hemoglobin concentration	Gram per deciliter (g/dL)
Hematocrit (HCT)	Fraction of blood that contains hemoglobin	Percentage (%)
Mean corpuscular hemoglobin (MCH)	Amount of hemoglobin per RBC	Picogram (pg)
Mean corpuscular volume (MCV)	Average volume of RBC	Femtoliter (fL)
MCH concentration (MCHC)	Hemoglobin divided by hematocrit	Gram per deciliter (g/dL)
RBC distribution width (RDW)	Distribution of RBC volume	Percentage (%)
White blood cell (WBC) count	Number of WBC per liter (include all main subtypes)	Billion cells per liter ($x10^9$/L)
Platelet (PLT) count	Number of PLT per liter	Billion cells per liter ($x10^9$/L)
Mean platelet volume (MPV)	Average platelet volume	Femtoliter (fL)

The clinical importance of this heritable variation in blood cell phenotypes is unclear. However, it is interesting that epidemiological studies have detected links between WBC or platelet counts and the risk to suffer from cardio- and cerebrovascular diseases [109-111]. As for most epidemiological observations, however, it is difficult to determine if changes in hematological parameters are pathological or reflect consequences of disease manifestation. Using Mendelian randomization methodologies, in which inherited genetic variants associated with hematological traits are used as instruments to test the causal effect of the traits on diseases, may provide an answer to this question [112]. Such an approach was successfully used to determine that LDL-cholesterol and triglyceride levels, but unlikely HDL-cholesterol levels, are causes of coronary artery diseases [113; 114]. Understanding how DNA polymorphisms modulate blood cell phenotypes in health (and diseases) could provide new opportunities to study hematopoiesis, improve their use in medicine as biomarkers and maybe even help in the development of new drugs. To this list, we would also add that hematological traits are ideal phenotypes to further our understanding of the genetics of human complex diseases and traits because experimental systems exist to functionally validate genetic findings.

1.3.3.1.3. Genome-Wide Association Studies (GWAS) for Blood Cell Phenotypes

Before GWAS, little was known about the role of SNPs and other common DNA sequence variants on normal variation in blood cell phenotypes. Candidate gene DNA sequencing experiments have identified mutations in the globin loci, but also in the erythropoietin receptor (*EPOR*) and hemochromatosis (*HFE*) genes [115; 116]. Genome-wide linkage studies also found a few reproducible signals, most notably a linkage peak on chromosome 6q23 that encompasses the MYB transcription factor [117; 118]. These findings could not, however, explain the heritability of these blood cell phenotypes in normal individuals.

As for many other complex human traits and diseases, the capacity to test associations with genotypes across the genome by GWAS opened a new world. Prior to the GWAS era, genetic association studies often had sample sizes that were too small and were limited to testing only known genes [119]. With GWAS, it became possible to genotype all genes independently of previous knowledge. Blood cell traits are particularly amenable to the GWAS approach because they are routinely and accurately measured in large cohorts, and initial findings can be tested for replication in other cohorts because it is easy to harmonize these phenotypes (Figure 1.8) [120]. In general, one of the main challenges for GWAS has been to pinpoint functional genes and variants associated with a given trait. Although this remains a challenge, blood cell traits are particularly well-suited for genetic and functional follow-up. As mentioned earlier, fine-mapping by dense genotyping and DNA re-sequencing is possible because the traits are usually available in most cohorts or biobanks, including participants of different ethnicities (see below). There is also the possibility to test the functions of new genes

in cell culture systems or model organisms because the phenotypes are often cell autonomous and the assays already well-developed. Using this approach, investigators showed that SNPs at 6p21.1 modulate erythrocyte traits through a regulatory effect on the cyclin D3 (*CCND3*) gene [121]. Large-scale gene silencing and other functional experiments in fruit flies, zebrafish and mice were also used to validate several new genes involved in platelet and RBC development within loci identified by GWAS [122; 123].

All the steps described in (Figure 1.8) now take advantage of powerful bioinformatic tools and other resources freely available on the web. For instance, comparative genomics has identified DNA bases that are conserved through evolution and therefore more likely to be functionally important [124]. There are also software that can predict based on conservation and physicochemical properties whether a DNA polymorphism that changes an amino acid is likely detrimental or not [125; 126]. We can also quickly query large gene expression datasets to determine if the genes near an associated SNP are expressed in the relevant tissue(s) for the phenotypes of interest (as an example, see reference [127]). And when genotypes are available, it is possible to test *in silico* if the GWAS SNPs (or SNPs in linkage disequilibrium) control gene expression through regulatory mechanisms, that is if the variants are expression quantitative trait loci (eQTL) [128]. The ENCODE and Roadmap Epigenomics Projects have used next-generation DNA sequencing applications, including DNAse I hypersensitive sites mapping and chromatin immunoprecipitation with antibodies against several histone tail modifications (ChIP-seq), to define regulatory sequences in human cell lines and tissues [129-131]. Using a complementary approach (FAIRE-seq), Paul et *al*. identified regions of open chromatin in primary human blood cells and showed that SNPs associated with RBC and

platelet phenotypes are enriched in these regions [132]. All this vast genomic information is useful in prioritizing causal genes and variants at GWAS loci, and investigators are developing algorithms to facilitate its integration [133; 134].

Figure 1.8. Ideal study design to identify SNPs associated with human complex traits and diseases using genome-wide association studies (GWAS).

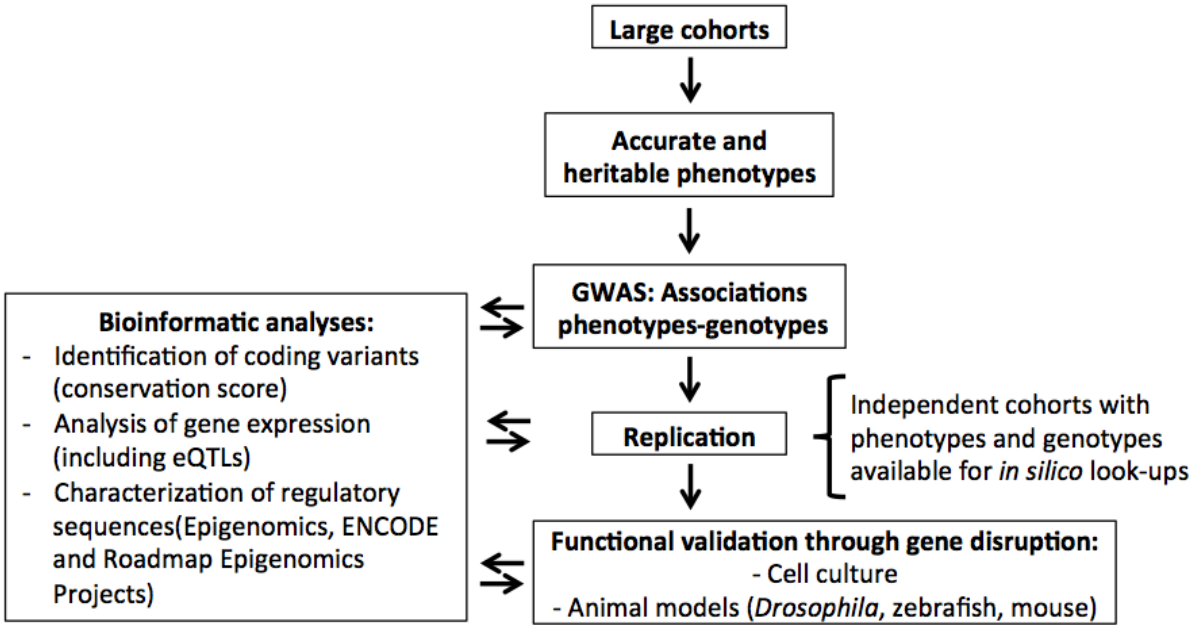


Figure 1.8. For blood cell phenotypes, GWAS were particularly successful because sample sizes are large, phenotypes are easy to measure and are accurate, and well-characterized experimental models already exist.

Several GWAS for hematological traits have already been published [122; 123; 135-151]. The largest studies, carried out in Europeans or individuals of European ancestry, have so far identified at genome-wide significance (*P*-value $<5x10^{-8}$) 75, 10 and 68 SNPs associated with RBC, WBC and platelet traits respectively [122; 123; 150]. The lower number of SNPs associated with WBC count could be explained by a lower heritability (see above), but also because the sample size for the WBC GWAS was smaller (N=11,823) in comparison with the GWAS for RBC (N=135,367) and platelet (N=66,867) traits. Despite their large number, these variants only explain a small fraction of the heritable variation in these phenotypes (<10%). They are, however, not random but clustered near genes involved in relevant biological pathways and enriched for regulatory functions by expression quantitative trait loci (eQTL) and epigenomic analyses. Most loci are associated with a single blood cell type but by comparing the different studies, we found seven loci that are associated with at least two different cell types (**Table 1.3**). These include *SH2B3*, a gene that encodes the adapter protein LNK that interacts with JAK2 and modulates JAK-STAT signaling in hematopoietic cells, and *MYB*, that encodes a transcription factor essential for definitive hematopoiesis. Both *SH2B3* and *MYB* SNPs are associated with the three main blood cell types. The other loci presented in **Table 1.3** include genes associated with a combination of two phenotypes, maybe suggesting different functions in different hematopoietic lineages.

Table 1.3. Loci identified by GWAS that carry SNPs associated with at least two of the three main blood cell types.

Locus	Location	RBC	WBC	Platelet	References
TMCC2	1q32.1	Caucasian		Caucasian	122; 123
ARHGEF3	3p14.3	African American		Caucasian	122; 135; 141; 143
LRRC16A	6p22.2	African American		African American	136; 142
HBS1L-MYB	6q22-q23.3	African American/ Caucasian/Japanese	Caucasian	African American/ Caucasian	122; 123; 136; 137; 139; 140; 142
IL-6	7p21		Japanese	Japanese	152
RCL1	9p24.1-p23	Caucasian		Caucasian / Japanese	122; 123; 137; 139
SH2B3	12q24	Caucasian	Caucasian	Caucasian / Japanese	122; 137-140; 143

Table 1.3. For each association, we report the ethnic group in which the genetic associations were found. We also listed only one gene per locus, although for many loci, the causal gene is unknown. RBC: red blood cell; WBC: white blood cell.

1.3.3.1.4. Some Loci Associated with Blood Cell Traits Are Population-Specific

It is difficult to compare association results for hematological traits across different populations because the sample size of the respective GWAS, and thus the statistical power to discover associations, is very different. For instance for RBC phenotypes, the largest studies in Caucasians and African Americans included, respectively, 135,367 and 16,496 participants [123; 136]. Despite this caveat, many of the loci found in African Americans or Asians were also present in Caucasians; this general transferability of results across ethnic groups has been observed for other complex human traits [153; 154]. For blood cell traits, however, there are notable exceptions. A SNP upstream of the Duffy antigen/receptor for chemokines (*DARC*) gene explains a large fraction of the variation in WBC and neutrophil counts, and is responsible for benign neutropenia [155]. This variant, which is monomorphic in Caucasians, is under positive selection in persons of African ancestry because it provides protection against *Plasmodium vivax* malaria infections. Similarly, genetic variation near the *α-globin*, the *β-globin* and the *G6PD* genes are associated with RBC indices in Africa-derived populations and are relatively common in frequency because they provide a selective advantage against malaria infections. These observations suggest that as we continue to query the human genome for associations with blood cell phenotypes, integrating evidence of natural selection would be a powerful approach.

1.3.3.1.5. Genetic Modifiers of Disease Severity

Several human diseases, which afflict a large fraction of the human population, are characterized by abnormally low or high counts of the three main blood cell types, or some unusual values for their features or contents. Anemia is a decrease of RBC count and hemoglobin levels (<11g/dL in women or <13g/dL in men) and is characterized by a wide spectrum of symptoms from simple fatigue to heart failure [156]. The World Health Organization estimates that anemia affects 1.62 billion people in the World [42]. The main causes of anemia are poor nutrition and iron deficiency, infections (*e.g.* malaria) and RBC diseases such as the hemoglobinopathies. Although the effect size of an individual SNP associated with RBC count or hemoglobin levels is not sufficient to cause anemia, a combination of hemoglobin-reducing alleles at many SNPs could have an impact on the risk to develop this disorder. Maybe more importantly, without causing anemia itself, this genetic score could influence clinical severity in at-risk populations (*e.g.* children with a small number of hemoglobin-increasing alleles that live in a region where malaria is endemic). Since anemia is mostly frequent in Africa and South-East Asia, it is critical to continue to search for genetic associations with hemoglobin levels in these populations [42].

There are many other human diseases that are diagnosed, like anemia, through abnormal counts of the main blood cell types (*e.g.* cancers). One example are myeloproliferative neoplasms (MPNs), diseases of the bone marrow characterized by excess cell production [157]. By far, the main cause of MPNs is a somatic gain-of-function mutation in the kinase gene *JAK2* (Val617Phe), which activates cell proliferation in the myeloid lineage [158; 159], and

changes platelet formation and reactivity [160]. It has never been tested whether SNPs associated with blood cell counts could modify complication risk in MPN patients with a *JAK2* (Val617Phe) mutation. For instance, MPN patients are at high risk of stroke, but it is unknown if such patients that also carry a large number of platelet-increasing alleles are at even higher stroke risk. Such analyses, on MPNs but also all other diseases characterized by a blood phenotype, are simple and could test the role that SNPs associated with normal variation in hematological traits may have on our risk to develop more severe disorders and related complications [123].

1.3.3.1.6. BCL11A Modifies Clinical Severity in Hemoglobinopathies

In adults, hemoglobin (HbA) is composed of two α- and two β-globin subunits that form a tetramer with the heme moiety to transport oxygen from the lungs to the different organs. Prior to birth, the *β-globin* gene is silent and the β-globin subunits are encoded by the *γ-globin* genes to form fetal hemoglobin (HbF). The switch from HbF to HbA production is a transcriptionally and epigenetically tightly regulated process [161]. For most healthy individuals, the switch itself has no clinical impact. However, for β-thalassemia and sickle cell disease patients with mutations in the *β-globin* gene, understanding and modulating the globin switch is currently the most promising therapeutic strategy. Conceptually, this is easy to appreciate: if the disease-causing mutations are in the *β-globin* gene, then re-activating *γ-globin* gene expression to form "normal" β-globin subunits would bypass the problem. This approach is supported by an extensive literature on the natural history of hemoglobinopathies and epidemiological studies [162]. For instance, it has been shown that sickle cell disease patients

that normally produce more HbF have better survival prognostic and less severe disease complications than patients with low HbF levels [163-165].

Although as adults we mostly produce HbA, we continue to make residual levels of HbF. Inter-individual variation in HbF levels is highly heritable (h^2~0.6-0.9) [108; 166]. Genetic investigations, including GWAS, have identified common genetic variation at three loci (*BCL11A*, *HBS1L-MYB* and *β-globin*) that have strong phenotypic effects and that together explain almost half of the heritable variation in HbF levels [167-170]. These HbF-associated SNPs are also associated with clinical severity in β-hemoglobinopathy patients: transfusion-dependency in β-thalassemia and painful crises in sickle cell disease [169; 171; 172]. This again emphasizes the importance of HbF as a strong modifier of severity for these diseases.

BCL11A encodes a transcription factor that had no known function in the globin switch before its discovery in two GWAS for HbF levels [167; 169]. Since then, we have learned that BCL11A is a potent transcriptional repressor of *γ-globin* gene expression and that its inactivation in the erythroid lineage can treat a sickle cell disease mouse model through re-activation of HbF production [173; 174]. More recently, both genetic and molecular fine-mapping work has determined that HbF-associated SNPs located in a *BCL11A* intron disrupt en erythroid enhancer that controls *BCL11A* expression [175]. This model was confirmed by targeted deletion of the enhancer through genome engineering that blocked *BCL11A* expression and re-activated *γ-globin* gene expression and HbF production [16]. As genome editing methods are rapidly improving, this proof-of-concept experiment suggests a new therapeutic strategy in which the *BCL11A* enhancer would be deleted *ex vivo* in a

hemoglobinopathy patient's cells to re-activate HbF production, and the cells would then be transplanted back to the patient [176]. The characterization of *BCL11A* and its role in HbF production serves as a powerful example to illustrate the success of GWAS from new biology to potentially innovative therapy.

1.3.3.1.7. Orphan Blood Cell Diseases

Although we did not assess the statistical significance of the enrichment, we observed that many of the SNPs associated with blood cell traits are located near genes that are mutated in severe hematological disorders and inherited in a Mendelian fashion. These include SNPs near *HK1* (hemolytic anemia), *TMPRSS6*, *HFE* and *TFR2* (iron deficiency) or *TUBB1* (thrombocytopenia). This observation is similar to the situation of many other complex human phenotypes (e.g. lipids, height, diabetes) where GWAS have identified hypomorphic alleles near human syndrome genes for related phenotypes. As such, the long list of loci found by GWAS provides a framework to investigate human syndromes characterized by aberrant blood features, mapped to a chromosome arm by linkage studies, but where the gene culprit has not been identified yet.

To investigate this hypothesis, we queried the Online Mendelian Inheritance in Man (OMIM) database [177]. In a non-exhaustive search, we identified four such orphan diseases where the genomic locations overlap with SNPs identified by GWAS (**Table 1.4**). For three of the diseases, GWAS findings suggest a strong candidate gene (*IL5*, *LIPC*, *NUDT19*) for re-sequencing in affected individuals. As we continue to map these rare blood disorders, cross-referencing with GWAS hits may provide a strong filter to prioritize genes for genetic testing.

Table 1.4. Orphan human syndromes mapped to a chromosomal band and characterized by a blood cell phenotype.

Mendelian genetics : orphan syndromes				Genome-wide association studies				
Locus	Disease	OMIM#	Description	SNP	Position	Phenotype	Candidate-gene(s)	Ref.
5q31	Familial eosinophilia	131400	Characterized by peripheral hypereosinophilia with or without other organ involvement	rs4143832	chr5:131,862,977	Eosinophil count	IL5	[138]
6p21	Macroblobulinemia, susceptibility to Waldenstrom	153600	Malignant B-cell neoplasm characterized by lymphoplasmacytic infiltration of the bone marrow and hypersecretion of monoclonal immunoglobulin M (IgM) protein	rs2517524	chr6:31,025,713	White blood cell	*HLA* region	[150]
15q21	Dyserythropoietic anemia, congenital type III	105600	Characterized by nonprogressive mild to moderate hemolytic anemia, macrocytosis in the peripheral blood, and giant multinucleated erythroblasts in the bone marrow	rs1532085	chr15:58,683,366	Hemoglobin	LIPC	[123]
19q13	Transient erythroblastopenia of childhood	227050	Red blood cell aplasia	rs3892630	chr19:33,181484	Mean corpuscular volume	NUDT19	[123]

Table 1.4. Orphan human syndromes mapped to a chromosomal band and characterized by a blood cell phenotype. Only such syndromes that overlap with a locus identified by GWAS for the corresponding blood cell trait are included in this table. We generated this list by querying the Online Mendelian Inheritance in Man (OMIM) database with the following keywords: anemia, blood, hemoglobin, leukopenia, neutropenia, platelet, thrombocytopenia.

1.3.3.1.8. Conclusions

GWAS have identified hundreds of loci that carry common genetic variants associated with RBC, WBC and platelet phenotypes. Many of these genetic associations still need to be linked to causal genes and genetic variants, yet because tractable cellular and animal models are available, this might be simpler for blood cell traits than it is for most complex human phenotypes. By design, GWAS interrogate common DNA variants, leaving untested low-frequency and rare sequence variation. The development of next-generation DNA sequencing platforms and exome genotyping arrays now provides the tools to test the role of this rarer genetic variation on blood cell phenotypes. Much criticism has been raised against GWAS because identified SNPs have poor predictive value; this is also true for SNPs associated with blood cell traits. However, this observation needs to be counter-balanced by the potential gain in improving our understanding of human biology in health and disease. GWAS blood cell trait loci provide new opportunities to study hematopoiesis, natural selection and the various ways common segregating DNA sequence variants can modify disease severity, paving the way for the development of more specific therapies.

1.3.3.1.9. Acknowledgements

This work was funded by grants from the Doris Duke Charitable Foundation (2012126), the Canadian Institute of Health Research (123382) and the Canada Research Chair Program.

1.3.3.1.10. Conflicts of Interest

The authors declare no conflict of interest.

1.4.RESEARCH OBJECTIVES

Although we have seen significant strides in gene discovery for both DCM and blood cell traits, many more genes remain to be identified. For DCM, more than 50 genes have been implicated with the disease, however mutations in those genes explain less than half of DCM cases. As I discussed above, identifying the genetic cause of DCM facilitates the clinical management of affected family members, and also improves our understanding of DCM and heart failure. On the other hand, although we know a lot more now about the genetic basis of blood-cell traits, its genetic architecture remains largely unknown (so is the case for other complex traits). GWAS studies did not provide as much explanation as was originally anticipated which warranted more research and different methods to contribute to identifying novel factors, particularly low-frequency variants that would partially account for the unexplained phenotypic variance. Moreover, the relationship between blood cell traits and CVD remains elusive. Hence, the objectives that I undertook in this thesis can be divided into two main components:

1) Discover novel genes and genetic mutations that cause dilated cardiomyopathy
2) Discover novel genetic factors that play a role in blood cell traits.

To achieve these objectives I used a variety of methods and tools that differed significantly between the analyses of the Mendelian vs complex traits. I ran both family-based and population-based analyses, and used both sequencing and chip-based genetic data.

1.5.THESIS CONTRIBUTIONS TO KNOWLEDGE

In chapter 2, I describe an exome sequencing study on French Canadian families followed at the MHI genetic clinic. Only 30% of probands with familial DCM get a positive result upon genetic screening at the MHI. By using exome sequencing, I show that we can improve sensitivity of genetic testing by more than two fold. I also identified a novel mutation in the *BAG3* gene that was causal of DCM in three families of the same region in Quebec. This work was the first to demonstrate that truncating mutations in *BAG3* are associated with early age of onset, where DCM patients carrying the *BAG3* truncating mutations developed the disease, on average, ten years earlier than their counterparts who carried causal mutations in other genes. This result had direct impact in clinical evaluation and follow-up of patients with mutations in this gene at the MHI and in other hospitals. We also identified five novel nonsense mutations in the giant gene *TTN* that segregated with disease and accounted for the highest number of families with mutations in the same gene.

In chapter 3, I describe a family with an atypical form of cardiomyopathy, one that does not conform to the known types of this disorder. I show that a mutation in *FLNC*, a gene that has an established role in myofibrillar myopathy causes a distinct cardiomyopathy characterized by fibrosis and sudden cardiac death suggesting that asymptomatic carriers of truncating mutations in *FLNC* may require clinical intervention to prevent SCD-associated arrhythmia.

In chapter 4, I carried out a large and multi-ethnic study as part of the blood cell consortium (BCX). I analyzed seven red blood cell traits in (130k) individuals from five ancestries: Europeans, African Americans, South Asians, East Asians, and Hispanics and

identified 16 novel genes not previously reported to be associated with red blood cells, contributing by that to the repertoire of genes involved in red blood cell biology. The use of an African American population in the study allowed the discovery of an association between *CD36* and red blood cell traits. This work also demonstrated that the identified mutation may have a functional role, by showing that there is reduced expression of *CD36* in erythroblasts of individuals that carry the identified nonsense variant.

In chapter 5, I describe a large scale analysis study with platelets within BCX. We explored genetic associations with platelet count (PLT) and mean platelet volume (MPV) and identified 15 novel loci. Among the identified genes, there were 8 genes that play a role in platelet reactivity.

1.6.CONTRIBUTIONS OF AUTHORS

Chapter 2:

Chami N*, Tadros R*, Lemarbre F, Lo KS, Beaudoin M, Robb L, Labuda D, Tardif JC, Racine N, Talajic M, Lettre G. Nonsense mutations in BAG3 are associated with early-onset dilated cardiomyopathy in French Canadians. Can J Cardiol. 2014 Dec;30(12):1655-61.

*contributed equally

I contributed to the design of the study. I performed all genetic and statistical analyses in all stages of the project. I interpreted the data and wrote the first draft of the manuscript. Rafik Tadros contributed to the design of the study, recruited patients, analyzed all clinical data and contributed to the writing of the manuscript. Laura Robb and Francois Lemarbre recruited patients. Ken Sin Lo prepared the sequencing pipeline and provided bio-informatic support. Melissa Beaudoin performed genotyping experiments. Damian Labuda contributed samples. Guillaume Lettre and Mario Talajic directed the study and contributed to the study design, the interpretation of results, and to the writing and review of the manuscript. All authors read and edited the manuscript.

Chapter 3:

Chami N, Tadros R, Lo KS, Beaudoin M, Robb L, Evelyne Naas, Talajic M, Lettre G. A splicing mutation in FLNC causes a rare form of cardiomyopathy. *In preparation*

I contributed to the design of the study, performed all genetic and statistical analyses, interpreted the results and wrote the first draft of the manuscript. Rafik Tadros contributed to the design of the study, recruited patients, analyzed all clinical data and contributed to the writing of the paper. Laura Robb and Evelyne Naas recruited patients. Ken Sin Lo prepared the sequencing pipeline and provided bio-informatic support. Melissa Beaudoin performed genotyping experiments. Guillaume Lettre and Mario Talajic contributed to the study design and the interpretation of results.

Chapter 4:

Nathalie Chami*, Ming-Huei Chen*, Andrew J. Slater*, John D. Eicher, Evangelos Evangelou, Salman M. Tajuddin *et al*, on behalf of the Blood Cell Consortium. Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. Am J Hum Genet. 2016 Jul 7; 99 (1):8-21

*contributed equally

I led the analyst team of the red blood cell (RBC) group of the Blood Cell Consortium. I performed genetic and statistical analysis, contributed to the interpretations of results and wrote the first draft of the manuscript. Ming-Huei Chen and Andew Slater performed genetics and statistical analyses and contributed to interpretations of results. Guillaume Lettre, Alex P. Reiner, Andrew D. Johnson, and Paul L. Auer supervised and designed the experiments. Guillaume Lettre oversaw the RBC team, contributed to the writing and final review of the paper. All authors read and edited the paper.

Chapter 5 :

John D. Eicher*, Nathalie Chami*, Tim Kacprowski*, Akihiro Nomura*, Ming-Huei Chen, Lisa R. *et al*, on behalf of the Blood Cell Consortium. Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 individuals. Am J Hum Genet. 2016 Jul 7; 99 (1):40-55

*contributed equally

John D. Eicher lead the platelet team of the Blood Cell Consortium, performed genetic and statistical analysis, contributed to the interpretations of results and wrote the first draft of the manuscript. I performed genetic and statistical analyses and contributed to interpretation of results and writing of the manuscript. Tim Kacprowski and Akihiro Nomura performed genetic and statistical analysis and contributed to interpretation of results. Andrew D. Johnson, Guillaume Lettre, Alex P. Reiner, and Paul L. Auer supervised and designed the experiments. Andrew D. Johnson oversaw the platelet team, contributed to the writing and final review of the paper. All authors read and edited the paper.

CHAPTER 2: NONSENSE MUTATIONS IN BAG3 ARE ASSOCIATED WITH EARLY-ONSET DILATED CARDIOMYOPATHY IN FRENCH CANADIANS.

Authors: Nathalie Chami*, Rafik Tadros*, Francois Lemarbre, Ken Sin Lo, Melissa Beaudoin, Laura Robb, Damian Labuda, Jean-Claude Tardif, Normand Racine, Mario Talajic, Guillaume Lettre.

***** These authors contributed equally to this work.

Reference: Chami N*, Tadros R*, Lemarbre F, Lo KS, Beaudoin M, Robb L, Labuda D, Tardif JC, Racine N, Talajic M, Lettre G. Nonsense mutations in BAG3 are associated with early-onset dilated cardiomyopathy in French Canadians. Can J Cardiol. 2014 Dec;30(12):1655-61.

2.1.ABSTRACT

Background: Dilated cardiomyopathy (DCM) is a major cause of heart failure that may require heart transplantation. Approximately one third of DCM cases are familial. Next-generation DNA sequencing of large panels of candidate genes (ie, targeted sequencing) or of the whole exome can rapidly and economically identify pathogenic mutations in familial DCM.

Methods: We recruited 64 individuals from 26 DCM families followed at the Montreal Heart Institute Cardiovascular Genetic Center and sequenced the whole exome of 44 patients and 2 controls. Both affected and unaffected family members underwent genotyping for segregation analysis.

Results: We found 2 truncating mutations in BAG3 in 4 DCM families (15%) and confirmed segregation with disease status by linkage (log of the odds [LOD] score = 3.8). BAG3 nonsense mutations conferred a worse prognosis as evidenced by a younger age of clinical onset (37 vs 48 years for carriers and noncarriers respectively; P = 0.037). We also found truncating mutations in TTN in 5 families (19%). Finally, we identified potential pathogenic mutations for 9 DCM families in 6 candidate genes (DSP, LMNA, MYH7, MYPN, RBM20, and TNNT2). We still need to confirm several of these mutations by segregation analysis.

Conclusions: Screening an extended panel of 41 candidate genes allowed us to identify probable pathogenic mutations in 69% of families with DCM in our cohort of mostly French-Canadian patients. We confirmed the prevalence of TTN nonsense mutations in DCM.

Furthermore, to our knowledge, we are the first to present an association between nonsense mutations in BAG3 and early-onset DCM.

2.2.INTRODUCTION

Dilated cardiomyopathy (DCM) is a major cause of heart failure representing the main reason for cardiac transplantation [178]. One third of DCM cases are familial [179-181] and the recognition of the familial nature of the disease is important for screening family members. The genetics of DCM is complex with more than 40 genes involved [3; 182]. As recommended in clinical practice guidelines, genetic testing is now routinely performed in familial DCM for the purpose of screening family members [183; 184]. Unfortunately, due to its large genetic heterogeneity, the yield of genetic testing targeting a small number of genes is modest (15-30%) as compared to that of hypertrophic or arrhythmogenic cardiomyopathies (30-70%) [183-185].

Advances in sequencing technologies made it possible to perform whole-exome sequencing (WES), where the protein coding regions of the whole-genome are targeted, at a reasonable cost. WES is an unbiased and efficient method to uncover potential pathogenic mutations in disease without previous assumptions about candidate genes or pathways and has proven to be successful at identifying causal mutations in several genetic disorders[45; 186; 187]. We applied WES on clinically well-characterized individuals with familial DCM in order to identify pathogenic mutations in those families. Here, we present our initial focused search for pathogenic mutations (missense, nonsense, frameshift and splice site variants) in the 41 known DCM candidate genes.

2.3.MATERIALS AND METHODS

Participants

The project was approved by the ethics committee at the Montreal Heart Institute (MHI) and conforms to the principles outlined in the Declaration of Helsinki. Individuals were recruited from the MHI Cardiovascular Genetic Center and signed informed consents. Inclusion criteria for probands were: left ventricular ejection fraction (LVEF) <45%, left ventricular dilatation end-diastolic diameter >117% of predicted value[188], and a first degree relative with DCM or a familial history of premature sudden cardiac death. Inclusion and exclusion criteria are detailed in **Supplementary Table 2.1.** Family members were recruited for segregation analysis; each was classified as *affected*, *unaffected* or *borderline* based on published criteria[188]. With the exception of two individuals who were tested five years prior to enrollment, all subjects had an echocardiogram performed within the previous three years. Prior to current study, clinical targeted genetic testing was performed in 20 of the 26 probands in the MHI molecular laboratory using Sanger sequencing. The DCM panel includes *SCN5A*, *LMNA*, *TNNT2*, *TNNI3*, *MYBPC3* and *MYH7,* in agreement with a published Canadian Cardiovascular Society (CCS) position statement on genetic testing[184]. Staff in the research laboratory was blinded to clinical testing results. To compare the sensitivity of WES and clinical Sanger sequencing, both families with and without identified mutations in the clinical laboratory were included in this study.

Whole-exome DNA Sequencing (WES)

We sequenced the exome of 44 participants using the Illumina HiSeq2000 instrument and a paired-ends 2x101 base pairs protocol. We used Illumina's TruSeqExome Enrichment Kit that targets 62 megabases, including exons from 20,794 genes. Details of the WES protocol are described the **Supplementary Note**.

Sanger sequencing

We confirmed mutations identified by WES using Sanger capillary sequencing. For *TTN*, we only validated novel nonsense mutations. The primer sequences are in **Supplementary Table S2**. We also genotyped by capillary sequencing *BAG3* p.Arg309stop in the DNA of 192 unrelated French Canadians from Gaspesia (**Supplementary Note**)[189].

Linkage analysis

Seventeen additional affected and unaffected members from two families (#1 and 6) with the *BAG3* p.Arg309stop nonsense mutations were recruited and genotyped by Sanger sequencing to test if the mutation segregates with disease. Within families 1, 6 and 12, we carried out linkage analysis in Merlin[190] using an autosomal dominant model, a recombination fraction $\theta=0$ and a disease prevalence of 0.0004 [2]). Since the *BAG3* p.Arg309stop mutation is not present in public databases, we chose a disease allele frequency of 0.01%; lower allele frequencies had an insignificant effect on the calculated LOD score.

Genome-wide DNA genotyping

In order to look for relatedness among individuals from families 1, 6 and 12 that carry *BAG3* p.Arg309stop, we performed genome-wide DNA genotyping using the Illumina OmniExpress BeadChip array and calculated pairwise identity-by-descent (IBD) metrics. Details are presented in the **Supplementary Note**.

Statistical analysis

We examined whether *BAG3* or *TTN* mutations are associated with an earlier age of onset or adverse outcomes defined as cardiovascular mortality, cardiac transplantation or ventricular assist device (VAD) implantation. To avoid recruitment biases, we excluded patients identified during routine screening for this analysis. Kaplan-Meier curves were generated using the *survfit* function in R[191]. To test the association between age of onset and *BAG3* or *TTN* carrier status, we used the QFAM-total procedure implemented in PLINK that uses permutations to take into account family structure[192].

2.4.RESULTS

Study population

We recruited and sequenced the whole-exome of 44 individuals from 26 DCM families: 42 DCM patients and 2 unaffected family members that we used as controls. The clinical characteristics of the 42 patients are described in **Table 2.1**. All probands had normal coronary angiography, except one (Family 7) who did not undergo angiography due to absence of risk factors and autopsy proven DCM in a deceased family member.

Whole-exome sequencing and variant prioritization

The summary of the sequencing results is presented in Supplementary Table S3. We achieved a mean coverage of 62X, corresponding to 83% of the targeted bases sequenced at ≥20X. We identified 192,464 DNA sequence variants, including 38,248 not catalogued in public databases (dbSNP build 139 and 1000 Genomes Project release 14)[193; 194]. To identify potential pathogenic DCM mutations, we only considered non-synonymous coding (missense, nonsense and frameshift) or splice site variants, with a minor allele frequency (MAF) ≤0.001 in the NHLBI Exome Sequencing Project (ESP) data[195]. Of these, we initially prioritized 58 variants that lie in any of the 41 previously reported DCM candidate genes (Supplementary Tables S4 and S5)[3; 182]. In families with more than one recruited affected subject, only mutations that segregated in at least another affected individual were further considered.

Table 2.1. Clinical characteristics for the 42 dilated cardiomyopathy (DCM) subjects that were whole-exome sequenced at the Montreal Heart Institute (MHI).

Characteristics	Values
Male sex N (%)	21 (50)
Current age (years)[1]	52 ± 14
Age of onset (years)[1]	44 ± 12
French-Canadian descent N (%)	34 (81)
LVEF (%)[1]	22 ± 12
LVEF < 35% N (%)	34 (81)
LVEDD (millimeters)[1]	65 ± 10
History of NYHA class III-IV Heart Failure N (%)	28 (67)
History of VAD implantation N (%)	5 (12)
History of cardiac transplantation N (%)	15 (36)
History of ICD implantation N (%)	28 (67)
History of ventricular arrhythmia N (%)	7 (17)
Coronary angiography (number performed/% abnormal)	32/0
SAECG (number performed/% abnormal)	11/91
Cardiac MRI (number performed/%abnormal)	18/100

Table 2.1. LVEF: left ventricular ejection fraction; LVEDD: left ventricular end-diastolic diameter; NYHA: New York Heart Association; VAD: ventricular assist device; ICD: implantable cardioverter-defibrillator; SAECG: Signal Averaged ECG; MRI: magnetic resonance imaging.[1]Mean ± standard deviation.

Truncating variants in BAG3

A truncating mutation in *BAG3* was identified in three apparently unrelated DCM families (**Table 2.2**). Family 1 is a large French-Canadian family with many affected individuals, including cases of sudden cardiac death and four transplant recipients (Figure **2.1** and Supplementary Table **2.2**). The proband was diagnosed with postpartum cardiomyopathy at the age of 30 and underwent urgent cardiac transplantation. The subsequent diagnosis of clinical DCM in first-degree relatives prompted the diagnosis of familial DCM. Prior to our study, the proband underwent negative genetic testing at MHI and also at the Laboratory of Molecular Medicine (Harvard) where she was tested for ten genes in 2009.

We sequenced the exome of eight affected individuals in Family 1. We identified a nonsense mutation in *BAG3* (p.Arg309stop) (**Table 2.2**) that segregated in all sequenced individuals. We also identified the same mutation in Families 6 and 12 (Figure **2.1**, Supplementary Note and Supplementary Table **2.2**). In the proband of Family 9, we also found another novel nonsense *BAG3* mutation (p.Ser249stop).

To confirm the pathogenicity of the *BAG3* mutations, we enrolled additional affected and unaffected family members from Families 1 and 6. We could not recruit additional members from Families 9 and 12. These individuals underwent cardiac imaging if not performed within the previous three years and we confirmed by capillary sequencing their *BAG3* p.Arg309stop carrier status. Genotype and phenotype information appear in Figure **2.1** and Supplementary Table **2.2**. In summary, all genotype negative individuals are unaffected. In Family 1, three individuals carry the mutation but are not clearly affected: 1.10 (24 years

old) is unaffected but still young, 1.9 (45 years old) had a normal echocardiogram 5 years prior to enrolment but was not available for clinical re-evaluation, and 1.16 (67 years old) has a mildly depressed LVEF (50%) but no left ventricular dilation and thus does not meet criteria for DCM. Interestingly, this last individual has been taking an angiotensin receptor blocker for many years for hypertension, which could have halted disease progression [196]. In Family 6, *BAG3* p.Arg309stop was fully penetrant (Figure **2.1**). Individual 6.11 has borderline DCM with an LV end-diastolic diameter >112% predicted but normal LVEF at the age of 37 years old. In summary, p.Arg309stop mutation segregated with disease status with high penetrance (95% if we consider individuals ≥40 years old). We carried out linkage analysis with all individuals from Families 1, 6 and 12 (N=30) who are ≥40 years old and for whom an echocardiogram was performed in the last three years. We calculated a LOD score of 3.8 for *BAG3* p.Arg309stop, indicating that the probability that this mutation segregates with disease status by chance in these families is approximately 1 in 4,000. Note that including individual 1.10 (young age) and 1.9 (not clinically tested in the last four years) in the analysis yields a LOD score of 3.2.

Table 2.2. Mutations identified in candidate dilated cardiomyopathy genes in this whole-exome DNA sequencing experiment.

Gene	Position (chr:pos)	Variant ID	Mutation	Annotation	Amino acid change	Polyphen prediction	Minor allele frequency ESP/1000G	MHI Families	Segregation confirmed?	Identified by clinical testing
LMNA	1:156,084,983	ss836897369	c.274C>T	Missense	p.Leu92Phe	probably damaging	-/-	26	No	Yes
LMNA	1:156,100,500	ss836897320	c.449C>T	Missense	p.Thr150Ile	possibly damaging	-/-	14	No	Yes
LMNA	1:156,106,048	rs61094188	c.1201C>T	Missense	p.Arg401Cys	probably damaging	-/-	3	Yes	No
LMNA	1:156,106,981	ss836897354	c.1566CG>C	frameshift	p.Asn524Thr fs*23	-	-/-	21	Yes	Yes
LMNA	1:156,108,510	rs142000963	c.1840C>T	Missense	p.Arg614Cys	probably damaging	0.001/-	19	No	Yes
TNNT2	1:201,333,470	ss836897393	c.415C>T	Missense	p.Arg139Cys	probably damaging	-/-	4	No	Yes
TTN	2:179,428,202	ss836897424	c.55462G>T	Nonsense	p.Gly18488stop	-	-/-	17	No	No*
TTN	2:179,429,822	-	c.53842C>T	Nonsense	p.Arg17948stop	-	-/-	8	Yes	No*
TTN	2:179,440,999	ss836897463	c.42665G>A	Nonsense	p.Trp14222stop	-	-/-	11	Yes	No*
TTN	2:179,470,369	ss836897473	c.26458G>T	Nonsense	p.Glu8820stop	-	-/-	5	No	No*
TTN	2:179,505,980	ss836897490	c.13428T>A	Nonsense	p.Lys4476stop	-	-/-	2	Yes	No*
DSP	6:7,571,710	ss836897516	c.1796T>G	Missense	p.Met599Arg	probably damaging	-/-	10	No	No*
MYPN	10:69,908,204	ss836897657	c.343C>T	Missense	p.Arg115Cys	possibly damaging	-/-	20	No	No*
RBM20	10:112,572,062	ss836897756	c.1907G>A	Missense	p.Arg636His	probably damaging	-/-	19	No	No*
BAG3	10:121,432,095	ss836897768	c.746C>A	Nonsense	p.Ser249stop	-	-/-	9	No	No*
BAG3	10:121,435,991	ss836897771	c.925C>T	Nonsense	p.Arg309stop	-	-/-	1	Yes	No*
								6	Yes	No*
								12	Yes	No*
MYH7	14:23,893,327	ss836897779	c.2711G>A	Missense	p.Arg904His	probably damaging	-/-	16	Yes	Yes

Table 2.2. Genomic coordinates are on build UCSC hg19. Segregation was confirmed in nine families. *:gene is not on the DCM panel for clinical testing at the Montreal Heart Institute

Figure 2.1. Pedigree of dilated cardiomyopathy families 1, 6 and 12.

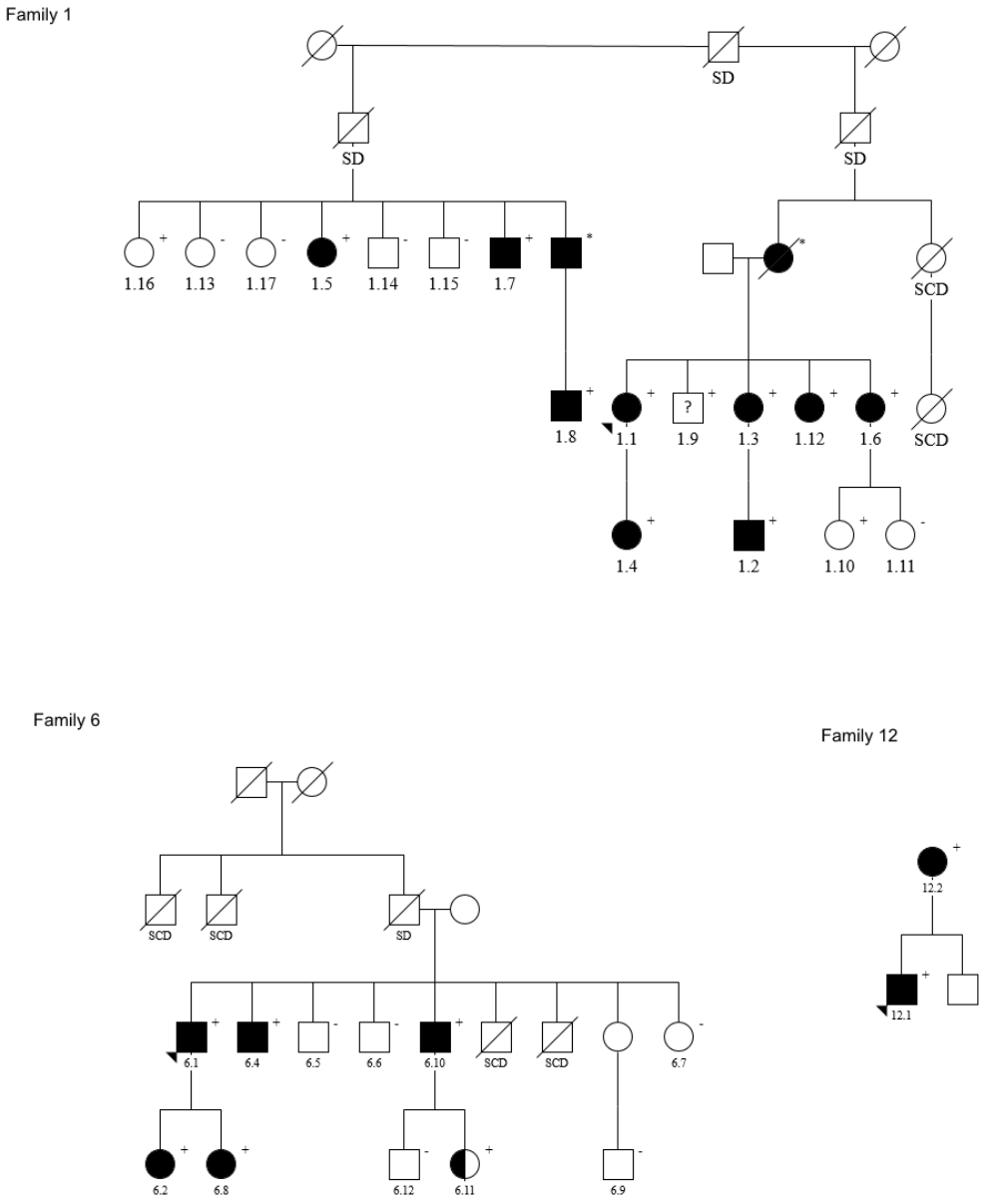


Figure 2.1. Probands are indicated by an arrow. Shaded and half-shaded symbols signify affected and borderline individuals respectively. “+”: a carrier of the *BAG3* mutation, p.Arg309stop; “-”: subject is negative for the mutation; "?": unknown affection status; SCD: sudden cardiac death; SD: sudden death. An asterisk (*) refers to an obligate carrier.

To evaluate the clinical importance of the *BAG3* p.Arg309stop mutation in our DCM patient population, we compared carriers and non-carriers in terms of age of clinical onset and severe adverse events: heart transplantation, implantation of a ventricular assist device (VAD), or cardiovascular death. Our analyses did not include patients recruited for screening to avoid bias and we used permutations to accounts for family structure. Interestingly, carrying the *BAG3* p.Arg309stop mutation was significantly associated with younger age of clinical onset (37 vs. 48, P=0.037; Figure **2.2**). The *BAG3* p.Arg309stop mutation does not modify the risk of severe adverse events in DCM patients (P=0.74) (Figure **2.2**).

This *BAG3* p.Arg309stop mutation is absent from public databases[193-195], but was previously reported in a European DCM pedigree[197]. Interestingly, Families 1, 6 and 12 are all originally from the Gaspésie region in Quebec. A genetic analysis based on identity-by-descent (IBD) revealed a third degree relatedness between individuals of these three families. We also carried out haplotype analyses to determine if *BAG3* p.Arg309stop is a founder mutation in Quebec. Details are described in the Supplementary Note.

Figure 2.2. Kaplan Meier curves of age of onset and severe adverse events of dilated cardiomyopathy (DCM) in carriers and non-carriers of *BAG3* nonsense mutations.

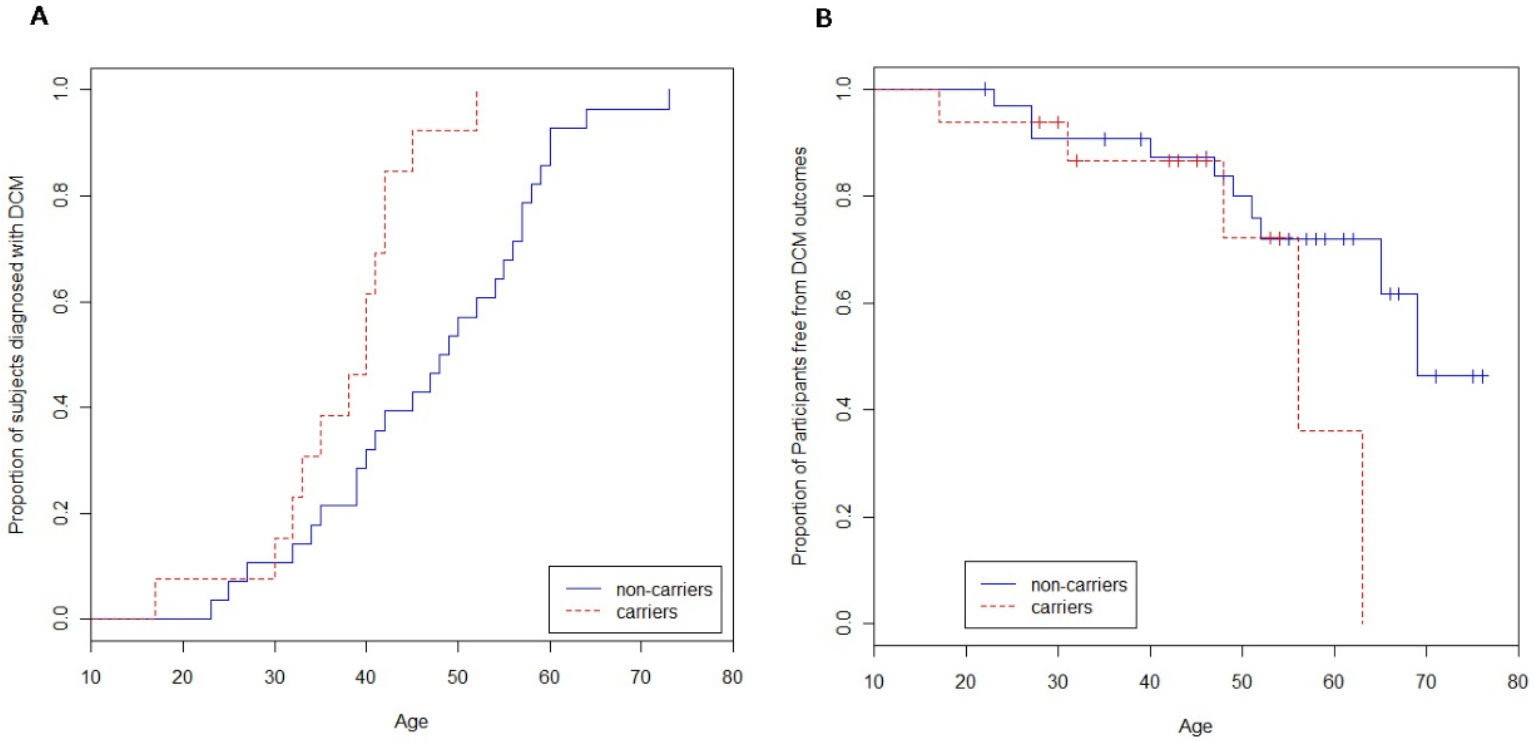


Figure 2.2. (**A**) Age of onset in DCM cases (N=41) including both probands and family members that presented to the hospital with DCM symptoms. Individuals diagnosed during screening are excluded. (**B**) Freedom of severe adverse events designated by heart transplantation, implantation of a ventricular assist device (VAD), or cardiovascular death in (N=47). Censored subjects are denoted with a hatch mark.

Mutations in other DCM candidate genes

We found five novel nonsense mutations in *TTN* in five families (Table **2.1**, Supplementary Table **2.2** and Supplementary Note). For Families 2, 8 and 11, (Supplementary Figure 4.1. Flow chart of the study design.), we confirmed segregation of the *TTN* nonsense mutation in one additional affected or borderline individual, but for Families 5 and 17 we could not recruit other members. We did not find DCM probands with frameshift *TTN* mutations. *TTN* nonsense mutations were not associated with earlier age of onset or adverse clinical outcomes (Supplementary Figure **2.1**).

Prior to our study, 20 probands underwent clinical testing for mutations in *SCN5A*, *LMNA*, *TNNT2*, *TNNI3*, *MYBPC3* and *MYH7* at the MHI. Likely pathogenic mutation for six probands were identified: in *LMNA* (Families 14, 19, 21, 26), *MYH7* (Family 16) and *TNNT2* (Family 4) (**Table 2.2** and Supplementary Table **2.2**). These mutations are described in the Supplementary Note. Our WES approach captured all six variants identified by clinical testing. We also found a missense mutation in *LMNA* (p.Arg401Cys) in the proband of Family 3.

Besides *BAG3*, *TTN* and the six genes routinely tested at the MHI, we examined 33 additional DCM genes. We identified missense mutations in *MYPN*, *DSP* and *RBM20* (**Table 2.2,** Supplementary Table **2.2** and Supplementary Note). These mutations are absent from ESP and 1000 Genomes Project databases[194; 195], although the same *RBM20* mutation (p.Arg636His) was previously identified in DCM patients[198; 199]. The proband from Family 19

carries both *LMNA* and *RBM20* missense mutations. We still need to confirm by segregation the pathogenicity of these mutations.

2.5.DISCUSSION

Our whole-exome DNA sequencing experiment in 26 DCM families identified rare and potentially pathogenic mutations in the following DCM candidate genes: *DSP*, *LMNA*, *MYH7*, *MYPN*, *RBM20* and *TNNT2* in nine families. The remaining nine families for whom we identified a potential pathogenic mutation carry truncating alleles in *TTN* or *BAG3*. Our study reinforces the role of *BAG3* in DCM. Our multiplex pedigrees allowed us to demonstrate that *BAG3* carries highly penetrant DCM mutations that are associated with a worse prognosis characterized by earlier age of onset. Our study adds to the clinical knowledge gleaned so far about *BAG3*.

The role of BAG3 in DCM

BAG3 encodes the Bcl-2-associated athanogene 3 protein which is a co-chaperone of heat shock proteins that localizes to the Z disk[200] and was previously linked to DCM[197; 201; 202]. Knocking-down *bag3* translation in a zebrafish model induced a heart failure phenotype[201]. Villard et *al*. reported the same nonsense mutation in *BAG3* (p.Arg309stop) in two related DCM patients of European origin[197]. To our knowledge, no one else has reported this mutation, and it remains absent from public databases[193-195]. Given the fact that Families 1, 6 and 12 are originally from the Gaspesie region in Quebec, we tested for the widespread presence of the *BAG3* p.Arg309stop allele in this region. We did not identify any carriers

among 192 healthy Gaspesians. Additional genetic analyses in 3,953 French Canadians did not identify potential *BAG3* p.Arg309stop carriers. The best model to explain this result is that the *BAG3* p.Arg309stop mutation arose on a *BAG3* haplotype that is common in the French-Canadian population (haplotype frequency is 9%). Although we cannot formerly rule out that the European[197] and French-Canadian DCM patients that carry the *BAG3* p.Arg309stop mutation share a common ancestor, the simplest explanation is that *BAG3* p.Arg309stop is a rare familial DCM mutation that has occurred twice independently. Recently, Campbell et *al*. combined WES and haplotype analysis to determine that a novel missense variant in *TNNT2* observed in two DCM families was likely due to independent mutational events[203]. In our case, we note that the *BAG3* p.Arg309stop mutation occurs due to a C > T nucleotide change within a CpG site. It has been suggested that DNA methylation at CpG sites can create mutation hotspots[204].

Truncating mutations in TTN

TTN encodes a 33,000 amino acids protein that is important for sarcomere assembly and contractile forces in striated muscle. Several studies implicated *TTN* in DCM[41; 42; 44; 205] and a recent report[43] suggested that truncating mutations in *TTN* are an important genetic cause of DCM, a result corroborated by our study. We demonstrate that 19% (5/26) of familial DCM cases carry a truncating *TTN* mutation. In agreement with Herman *et al.*[43], *TTN* nonsense mutations were not associated with earlier age of onset or more severe outcomes in our study (**Supplementary Figure 2.1**). It will be important to validate segregation of the identified nonsense *TTN* mutations in Families 5 and 17 as recent data suggest that not all truncating *TTN* alleles are fully penetrant or even pathogenic[44; 206; 207].

Families or probands-only?

WES generates an almost exhaustive catalogue of coding mutations found in a given patient. It is therefore a very attractive approach to identify the cause of rare Mendelian diseases, and works particularly well with diseases in which one or few genes are mutated. In its simplest form, you sequence a series of unrelated probands and find the one gene in which they all carry a private mutation. In the case of DCM, however, the probands-only strategy is difficult because >40 genes are implicated (and the list continues to grow). This approach worked well for *TTN*, but the prevalence of truncating mutations in this gene in DCM patients is very high (20-25%)[43]. For all other known DCM genes, the prevalence of mutations is small (<5%) and it is difficult to build a convincing statistical argument by simply testing unrelated patients. Even more challenging would be the validation of a pathogenic mutation found in a single affected individual without family members. Functional studies in cellular or animal models could provide hints, yet the extrapolation of phenotypes observed in cells or mice to humans is not straightforward. For these reasons, we advocate that careful segregation analyses should remain the gold-standard criterion to evaluate the candidacy of new DCM genes. The recruitment of family members is also essential to achieve the main goals of our DCM screening program: (1) preventing fatal cardiovascular events and (2) genetic counseling.

2.6.ACKNOWLEDGEMENTS

We thank all participants and acknowledge the technical support of the Beaulieu-Saucier MHI Pharmacogenomic Center.

Funding Sources

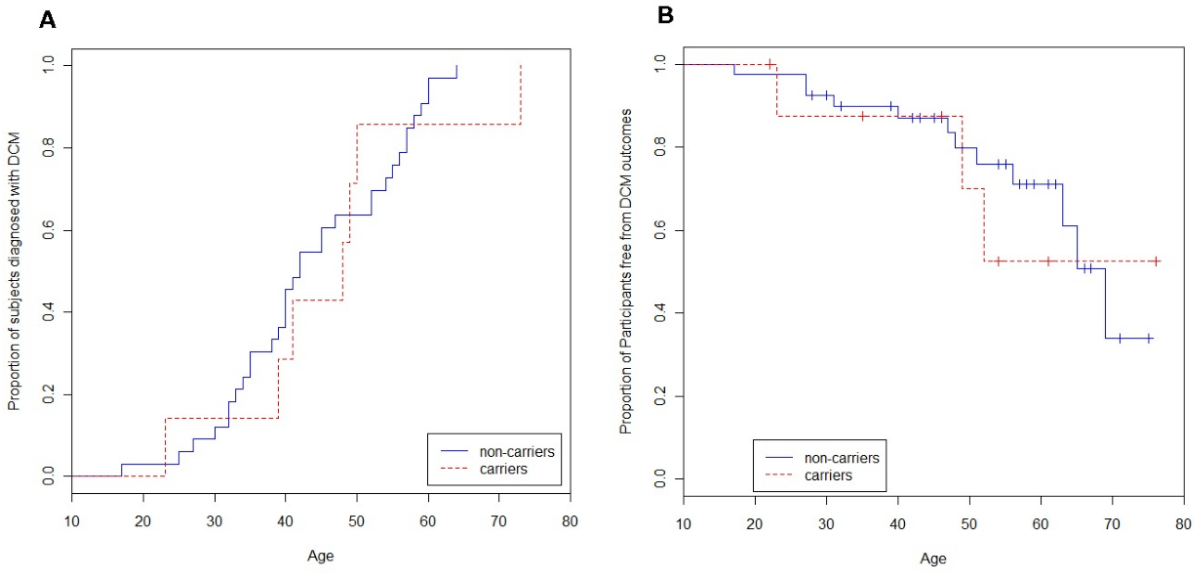
This work was supported by the Centre of Excellence in Personalized Medicine (CEPMed); the “Fonds de Recherche Santé Québec” (FRQS); the “Réseau de Médecine Génétique Appliquée” (RMGA) of FRQS; the Canada Research Chair program; the Marvin and Philippa Chair in Cardiology; and the MHI Foundation. There are no relationships with industry. The funding sources had no involvement in the study design or the interpretation of the results.

Disclosures

None.

2.7.SUPPLEMENTARY INFORMATION

Supplementary Figure 2.1. Kaplan Meier curves of age of onset and clinical outcomes of dilated cardiomyopathy (DCM) in carriers and non-carriers of *TTN* nonsense mutations.



Supplementary Figure 2.1. (A) Age of onset in DCM cases (N=40) including both probands and family members that presented to the hospital with DCM symptoms. (B) Freedom of clinical outcomes designated by heart transplantation, implantation of a ventricular assist device (VAD), or death in 45 individuals. Censored subjects are denoted with a hatch mark. In both cases, analyses comparing DCM patients with or without nonsense mutations were non-significant (P=0.57 and P=0.27 respectively), consistent with a previous report[43].

Supplementary Table 2.1. Inclusion and exclusion criteria for probands with dilated cardiomyopathy (DCM).

Inclusion	Exclusion
Ejection fraction of the left ventricle <0.45.	Systemic diseases, pericardial diseases, congenital heart disease.
Left ventricular end-diastolic diameter >117% of the predicted value corrected for age and body surface area which corresponds to 2 standard deviations of the predicted normal limit +5%.	Coronary heart disease.
At least one affected family member and/or 1st degree family history of sudden cardiac death below the age of 35.	Systemic arterial hypertension (>160/100 mmHg).
	History of excess alcohol consumption.
	Clinical, sustained and rapid supraventricular arrhythmias.
	Evidence of DCM due to infections.

Supplementary Table 2.2. Clinical characteristics of all family members for whom a potential pathogenic mutation was identified.

FID	ID	Ethnicity	Affection Status	AO	Context of diagnosis	LVEF (%)	LVEDD (mm)	Transp/VAD	VA	Death	Mutation
1	1.1	FC	Affected	30	Cardiogenic shock	5	75	Transp.	No	No	*BAG3*; p.Arg309X
1	1.2	FC	Affected	26	Screening	37	59	NC	No	No	*BAG3*; p.Arg309X
1	1.3	FC	Affected	42	Heart failure	40	58	NC	No	No	*BAG3*; p.Arg309X
1	1.4	FC	Affected	17	Cardiogenic shock	15	71	Transp.	No	No	*BAG3*; p.Arg309X
1	1.5	FC	Affected	38	Heart failure	20	66	NC	No	No	*BAG3*; p.Arg309X
1	1.6	FC	Affected	35	Heart failure	30	64	NC	No	No	*BAG3*; p.Arg309X
1	1.7	FC	Affected	45	Heart failure	16	89	Transp.	No	No	*BAG3*; p.Arg309X
1	1.8	FC	Affected	40	Heart failure	10	78	NC	No	No	*BAG3*; p.Arg309X
1	1.9*	FC	Unknown	NA	NA	60	47	NA	No	No	*BAG3*; p.Arg309X
1	1.10	FC	Not affected	NA	NA	65	47	NA	No	No	*BAG3*; p.Arg309X
1	1.11	FC	Not affected	NA	NA	63	46	NA	No	No	negative
1	1.12	FC	Affected	32	Palpitations	36	55	NC	No	No	*BAG3*; p.Arg309X
1	1.13	FC	Not affected	NA	NA	62	48	NA	No	No	negative
1	1.14	FC	Not affected	NA	NA	62	52	NA	No	No	negative
1	1.15	FC	Not affected	NA	NA	NA	NA	NA	No	No	negative
1	1.16	FC	Not affected	NA	Screening	50	52	NA	No	No	*BAG3*; p.Arg309X
1	1.17	FC	Not affected	NA	NA	68	47	NA	No	No	negative
2	2.1	FC	Affected	48	Heart failure	41	61	NC	No	No	*TTN*; p.Lys4476X
2	2.2	FC	Affected	41	Heart failure	15	80	Transp.	Yes	No	*TTN*; p.Lys4476X
2	2.3	FC	Not affected	NA	NA	60	47	NA	No	No	*TTN*; p.Lys4476X
2	2.4	FC	Affected	32	Screening	39	56	NC	No	No	*TTN*; p.Lys4476X
2	2.5	FC	Not affected	NA	NA	64	56	NA	No	No	negative
3	3.1	FC	Affected	54	Heart failure	22	64	NC	No	No	*LMNA*;p.Arg401Cys
3	3.2	FC	Affected	56	Heart failure	35	71	NC	No	No	*LMNA*;p.Arg401Cys
4	4.1	FC	Affected	63	Palpitations	24	61	NC	No	No	*TNNT2*; p.Arg139Cys
5	5.1	FC	Affected	39	Heart failure	20	73	NC	No	No	*TTN*; p.Glu8820X
6	6.1	FC	Affected	33	Heart failure	12	72	Transp.	No	No	*BAG3*, p.Arg309X
6	6.2	FC	Affected	23	Screening	47	58	NC	No	No	*BAG3*, p.Arg309X
6	6.4	FC	Affected	42	Atrial fibrillation	20	66	NC	No	No	*BAG3*, p.Arg309X
6	6.5	FC	Not affected	NA	NA	65	45	NA	No	No	negative
6	6.6	FC	Not affected	NA	NA	65	53	NA	No	No	negative
6	6.7	FC	Not affected	NA	Chest pain	60	54	NA	No	No	negative
6	6.8	FC	Affected	17	Screening	42	59	NC	Yes	No	*BAG3*, p.Arg309X
6	6.9	FC	Not affected	NA	NA	65	45	NA	No	No	negative
6	6.10	FC	Affected	58	Stroke	missing	missing	Transp.	Yes	No	*BAG3*, p.Arg309X
6	6.11	FC	Borderline	NA	NA	65	52	NA	No	No	*BAG3*, p.Arg309X
6	6.12	FC	Not affected	NA	NA	65	49	NA	No	No	negative
8	8.1	FC	Affected	23	Heart failure	5	70	Transp.	No	No	*TTN*; p.Arg17948X
8	8.2	FC	Affected	49	TIA	15	78	Transp.	No	No	*TTN*; p.Arg17948X
9	9.1	European	Affected	41	Heart failure	10	80	NC	No	No	*BAG3*; p.Ser249X

10	10.1	FC	Affected	25	Stroke	15	68	Transp.	Yes	Yes (Immuno-suppressants non-compliance)	DSP; p.Met599Arg
11	11.1	FC	Affected	50	Heart failure	10	59	NC	No	No	*TTN*; p.Trp14222X
11	11.2	FC	Borderline	20	Screening	55	41	NC	No	No	*TTN*; p.Trp14222X
12	12.1	FC	Affected	40	Heart failure	10	76	NC	No	No	*BAG3*; p.Arg309X
12	12.2	FC	Affected	52	Heart failure	12	70	Transp.	No	No	*BAG3*; p.Arg309X
14	14.1	FC	Affected	47	Heart failure	25	67	Transp.	Yes	No	*LMNA*; p.Thr150Ile
16	16.1	FC	Affected	27	Cardiogenic shock	5	91	Transp.	No	No	*MYH7*; p.Arg904His
16	16.2	FC	Affected	52	Syncope	20	59	NC	No	No	*MYH7*; p.Arg904His
16	16.3	FC	Not Affected	NA	Screening	51	50	NA	No	No	negative
16	16.4	FC	Not Affected	NA	Screening	65	46	NA	No	No	negative
19	19.1	FC	Affected	56	Heart failure	25	61	NC	No	No	LMNA; p.Arg614Cys RBM20; p.Arg636His
20	20.1	FC	Affected	55	Heart failure	25	58	NC	No	No	*MYPN;* p.Arg115Cys
21	21.1	Other	Affected	34	Stroke & AV block	17	61	Transp.	Yes	Yes (Immediate post transplant)	*LMNA*, p.Asn524Thr fs*23
21	21.2	Other	Affected	42	Stroke & AV block	20	49	NC	Yes	No	*LMNA*, p.N524T fs*23
26	26.1	FC	Affected	45	Heart failure	20	62	VAD	Yes	Yes (Brain hemorrage)	*LMNA;* p.Leu92Phe

Supplementary Table 2.2. FID: Family ID; AO: age of onset; LVEF: left ventricular ejection fraction; LVEDD: left ventricular end-diastolic diameter; Transp.: Cardiac Transplantation; VAD: ventricular assist device; VA: ventricular arrhythmia; FC: French Canadian; NC: not candidate; NA: not applicable; TIA: transient ischemic attack; AV: Atrio-ventricular ;*: recent data not available.

CHAPTER 3: A SPLICING MUTATION IN FLNC CAUSES A RARE FORM OF CARDIOMYOPATHY.

Authors: Nathalie Chami, Rafik Tadros, Ken Sin Lo, Melissa Beaudoin, Laura Robb, Evelyn Naas, Mario Talajic, Guillaume Lettre.

Reference: Nathalie Chami, Rafik Tadros, Ken Sin Lo, Melissa Beaudoin, Laura Robb, Evelyn Naas, Mario Talajic, Guillaume Lettre. A splicing mutation in FLNC causes a rare form of cardiomyopathy. *In preparation*

3.1. ABSTRACT

Cardiomyopathies are a group of heart muscle disorders. Inherited forms of cardiomyopathies are usually caused by one mutation in a variety of genes that play a role in multiple mechanisms. Genetic screening of affected members identifies a causal mutation in only ~ 30% of cases. We sought to identify the cause of cardiomyopathy in a family that presented with a distinct form of cardiomyopathy characterized by a history of sudden cardiac death (SCD), fibrosis, arrhythmias, and a variable degree of dilated cardiomyopathy and for which genetic screening did not identify a causal mutation.

We performed whole exome sequencing on 11 family members including 4 affected individuals. Additionally, we sequenced the exome of a formalin fixed and paraffin embedded (FFPE) heart tissue sample of a deceased individual that suffered a SCD. We identified a splicing mutation in exon 44 of the *FLNC* gene that was carried by all affected individuals. Pathological analyses of the heart sample of the deceased individual revealed characteristics of left-dominant arrhythmogenic cardiomyopathy and extensive fibrosis. Our results suggest that FLNC plays a role in cardiomyopathy and particularly in an atypical form with severe prognosis.

3.2.INTRODUCTION

Cardiomyopathies are a group of disorders that weaken the heart muscle function and may lead to heart failure and cardiac transplantation[208]. Non ischemic dilated cardiomyopathy (DCM) is a type of cardiomyopathy with prevalence =1 in 2700 adults [2], although recent estimates suggest that it may be much higher [3]. DCM is characterized by left ventricular enlargement and systolic dysfunction and is associated with high morbidity and mortality rates.

More than 50 genes [58] have been linked to DCM and a number of those have arisen recently with the advances in genomic technologies. The advent of next generation sequencing (NGS) enabled faster and cheaper DNA sequencing compared to the traditional Sanger method. As a result, it is now possible to expand the list of genes that are screened during genetic testing for DCM. The cardiomyopathy panel at the Montreal Heart Institute (MHI) genetic clinic, for instance, now includes 35 genes compared to the previous panel that included only 6 genes. NGS also made it possible to carry out whole exome sequencing (WES) which queries all genes of the human genome in an untargeted approach and thus provides a great potential for novel gene discoveries for cardiomyopathies. Despite all these promising advances, in the majority of cases, screening probands for cardiomyopathy genes does not identify the causal mutation.

As part of an ongoing study at the MHI that aims at identifying novel genetic mutations associated with dilated cardiomyopathy[59], we recruited a family with an atypical

cardiomyopathy characterized by left ventricular fibrosis, arrhythmia and sudden cardiac death (SCD). We sequenced the exomes of 5 affected individuals including that of a deceased family member who suffered a sudden cardiac death and whose heart was preserved at the MHI, in addition to 6 unaffected or asymptomatic family members.

3.3. MATERIALS AND METHODS

Participants

Members of family 7 were recruited at the MHI Cardiovascular Genetics Center. All participants signed an informed consent. The project was approved by the ethics committee at the MHI and conforms to the principles outlined in the Declaration of Helsinki. All family members underwent phenotyping including a cardiac magnetic resonance (CMR) imaging with gadolinium enhancement and/or echocardiography if CMR is not available or contra-indicated, a standard and signal-averaged electrocardiogram, holter monitoring and exercise testing. Electrocardiographic examinations were systematically assessed by a cardiac electrophysiologist while all CMR were reviewed by an experienced CMR expert. The family was recruited on the basis of atypical cardiomyopathy with left ventricular dilatation and systolic dysfunction, malignant arrhythmia and left ventricular fibrosis. Family members were classified as *affected* if they have a clinical DCM as defined by [188] or fibrosis detected on CMR (late gadolinium enhancement, LGE) or at post-mortem cardiac examination. Minor LGE with an estimated scar mass <5g was considered non-diagnostic because of the non-specificity of this finding. The proband (III.9) previously underwent diagnostic genetic testing in 2008 at the MHI molecular laboratory which consisted of Sanger sequencing of 11 genes associated with cardiomyopathy (*MYH7, MYBPC3, TNNT2, TNNI3, SCN5A, LMNA, DSP, PKP2, DSG2, DSC2, TMEM43*). No disease-causing mutation was found.

Whole exome sequencing

We sequenced the exomes of 11 living family members at the pharmacogenomics center of MHI using the Illumina HiSeq2000 instrument and a paired-ends 2x101 base pairs protocol.

We used Illumina's TruSeqExome Enrichment Kit that targets 62 megabases, including exons from 20,794 genes. More information about the sequencing protocol and variant calling is published elsewhere [59].

DNA extraction and sequencing of III.11

A formalin fixed and paraffin embedded (FFPE) heart tissue was obtained for individual III.11. We obtained consent from the subject's family to extract and sequence DNA from the available tissue. DNA was extracted using GeneRead gDNA kit for FFPE samples from Qiagen (cat no 180134) according to the protocol version 03/2014). Briefly, a 10uM slice was used. Following step 5, 100uL of deparafinization solution was added followed by another 3 min of incubation at 56°C. The elution was done in 30uL. The low molecular weight DNA was subsequently sent to the Genome center in Montreal to carry out the library preparation using NEB Ultra II kit (SeqCap Exome from Roche Nimblegen) using 100ng. The exome sequencing was performed on Illumina HiSeq2000 with PE100bp using Illumina TruSeq DNA v3. We carried out the variant calling with the other family members sequenced at the pharmacogenomics center of the MHI. Total number of variants and TS/TV ratios were comparable between ID III.11 and the rest of the family members.

3.4. RESULTS

Phenotypic characterization

The family (**Figure 3.1**) was referred for cardiovascular genetics evaluation when individual III.11 died suddenly while cycling at the age of 32. Past medical history was uneventful. Three years prior to the death of III.11, his father (II.6) died at the age of 57 of cardiogenic shock. He was diagnosed with DCM at the age of 46 after presenting with a syncopal event. At diagnosis, the left ventricle (LV) was dilated (diastolic diameter = 65mm) and showed systolic dysfunction (LV ejection fraction, LVEF = 30%). His follow-up was characterized by recurrent ventricular tachycardia and progressive heart failure requiring an implantable cardioverter defibrillator (ICD) with resynchronization therapy and amiodarone. In the few months prior to his death, the patient was recurrently admitted for appropriate ICD shocks with deterioration of LV function.

Familial cardiac evaluation was performed in 11 additional family members (Table **3.1** and **Figure 3.1**). Four additional individuals were found to be affected and seven were unaffected or undiagnostic. III.6 had minor ventricular LGE that was thought to be non-specific (estimated mass <5g) and is thus said to have non-diagnostic findings. Clinical genetic testing of III.9 that was performed at the MHI did not identify any disease-causing variant.

Figure 3.1. Pedigree of family 7

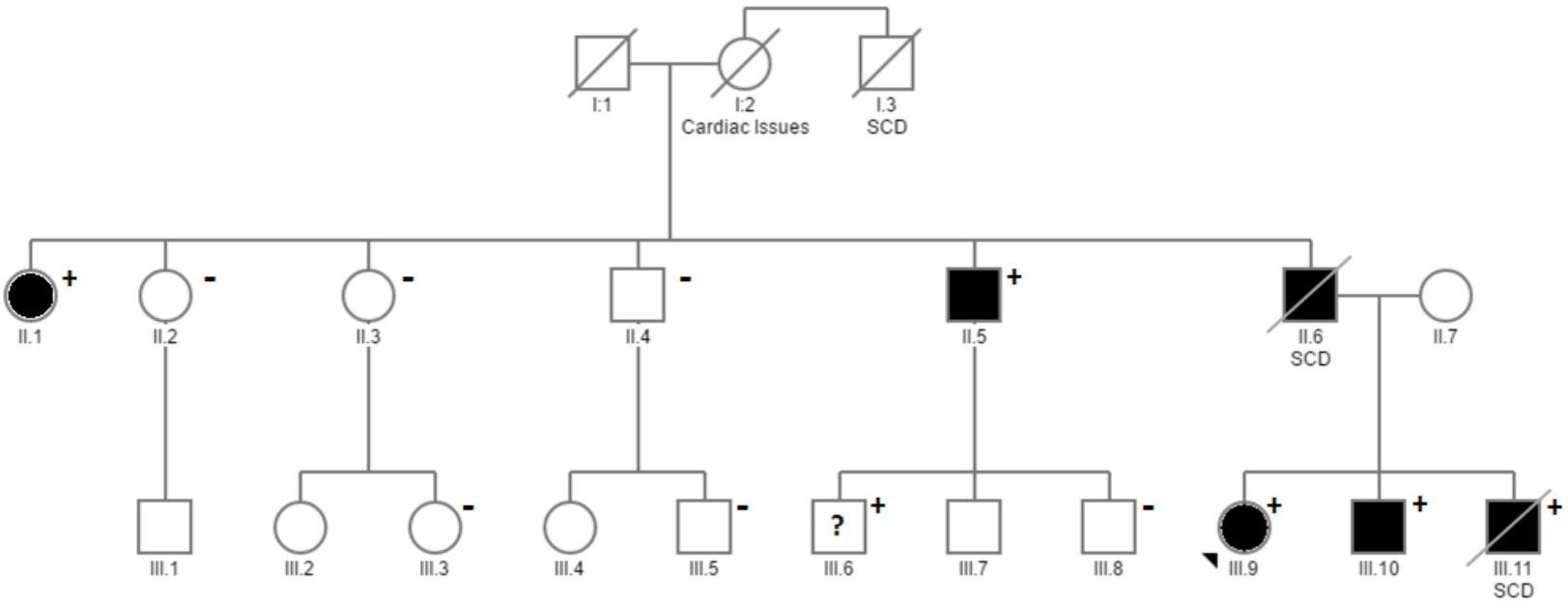


Figure 3.1. The proband is indicated by an arrow. Shaded squares and circles indicate affected individuals. A "+" sign refers to carriers of the *FLNC* mutation. A "-" sign means noncarriers. A "?" denotes unknown affection status. Those without a sign have not been genetically tested. SCD: sudden cardiac death. II.6 was considered as an obligate carrier in the segregation analysis.

Table 3.1. Clinical characteristics of the evaluated family members

ID	Phenotype	Age	PR	QRS	QTc	SAECG	Exercise testing	Holter PVCs/24h	ICD	Clinical Arrhythmia	LVEDV (mL/m2)		Fibrosis	FLNC
II.1	Affected	63	154	86	404	Abnormal(1/3)	None	342	No	None	66.3	66	5g	+
II.2	Unaffected	53	162	82	417	Normal	1 isolated PVC	NA	No	None	71.5	63	n	-
II.3	Unaffected	60	138	82	432	Abnormal(1/3)	NA	21	No	None	50.6	72	2.1g	-
II.4	Unaffected	58	164	100	414	Abnormal(1/3)	NA	NA	No	None	6II.5	62	no	-
II.5	Affected	53	178	116	414	Abnormal(3/3)	Isolated PVCs	148	Primary	Recurrent NSVT	92	68	yes	+
II.6	Affected	57	156	116	467	NA	NA	NA	Secondary	Recurrent VT Atrial fibrillation	65mm (TTE)	30 (TTE)	NA	(+)
III.3	Unaffected	26	156	90	402	Abnormal(1/3)	None	1	No	None	82.4	67	1.2g	-
III.5	Unaffected	23	182	90	391	Abnormal(3/3)	NA	NA	No	None	80.6	67	2.4g	-
III.6	Non-diagnostic	22	140	96	399	Abnormal(1/3)	None	0	No	None	96.6	67	4.5g	+
III.8	Unaffected	28	178	102	409	Abnormal(1/3)	None	0	No	None	95.4	62	n	-
III.9	Affected	27	128	98	410	Abnormal(1/3)	Isolated PVCs	379	Primary	None	93.3	59	yes	+
III.10	Affected	30	144	92	378	NA	NA	NA	Primary	Appropriate ATP (VT at 230bpm)	NA	67	yes	+
III.11	Affected	32	NA	NA	NA	NA	NA	NA	No	SCD	Post-mortem: Moderate biventricular dilatation Extensive epicardial fibrosis	+		

Table 3.1. SAECG: signal-averaged ECG; PVC: premature ventricular contraction; ICD: implantable cardioverter defibrillator; LVEDV: left ventricular end-diastolic volume; LVEF: left ventricular ejection fraction; NSVT: non-sustained ventricular tachycardia; ATP: anti-tachycardia pacing; VT: ventricular tachycardia; SCD: sudden cardiac death; TTE: trans-thoracic echocardiogram NA: non-available

Variant prioritization and segregation analysis

To look for candidate pathogenic mutations, we followed a series of steps. First, we kept variants shared between the 5 individuals considered to be clearly affected (II.1,II.5, III.9, III.10, III.11) which yielded 27,206 variants. We then kept nonsynonymous coding (missense, nonsense and frameshift) or splice site variants, with a minor allele frequency (MAF) ≤ 0.0002 in the ExAC dataset [209] and 51 remained. To further narrow down the list, we filtered against our own cardiomyopathy dataset that includes 78 individuals. Instead of choosing only private mutations, we were less conservative since we and others have demonstrated that distinct DCM families may share the same mutation[59]. So we assigned a MAF cutoff of 7% (or 11 carriers). Nine variants were retained and are listed in Table **3.2**. Given that young unaffected individuals are less informative for segregation analyses in a late onset disease, we only considered older unaffected family members (age > 45) for the segregation analyses. We removed variants that were present in ≥ 2 unaffected older individuals (age > 45) and only two variants were subsequently considered: a splice variant in exon 45 of *FLNC* and a missense variant (Thr372Ala) in *MUC21*. The latter is deemed benign by two mutation prediction tools, polyphen and SIFT. Further, the gene is not expressed in the heart tissue according to the human protein atlas (proteinatlas.org) which also includes the GTEx dataset. Hence, this variant does not constitute a candidate mutation and only the *FLNC* splice mutation was finally considered. The *FLNC* variant was carried by 6 family members, 5 of which are affected, and the sixth carrier is 26 years old with undiagnostic status (see methods). The variant is also absent from any of the queried databases (1000 genomes, ESP, and ExAC).

Pathology analysis of III.11

Macroscopic examination of a section of the myocardium of individual III.11 demonstrated cardiomegaly (440g) with moderate predominantly left-sided ventricular dilatation and mild cardiomyocyte hypertrophy (Figure **3.2**). Extensive interstitial fibrosis with fatty infiltration predominantly affecting the left ventricular epicardium was noted Figure **3.2**. Cardiac histology confirmed the nature of the fibrous and fibro-fatty infiltration of the myocardium in the outer third and subepicardium of the left ventricle. Besides secondary hypertrophy of the cardiomyocytes, there was no obvious or specific cellular abnormality demonstrated on histology, including a reduction in myofibrils or the presence of vacuoles.

Table 3.2. Variants retained prior to segregation analysis

Position	ExAC MAF	II.1	II.2	II.3	II.4	II.5	III.3	III.5	III.6	III.8	III.9	III.10	III.11	amino acid change	gene
7:128496798	–	+	–	–	–	+	–	–	+	–	+	+	+	exon 44	FLNC
1:207850854	0.00002		+	+	–	+	–	–	–	+	+	+	+	Pro73Arg	CR1L
1:209800293	0.00010	+	+	+	–	+	–	–	–	+	+	+	+	Phe506Ile	LAMB3
1:40533278	–	+	+	+	+	+	–	–	+	–	+	+	+	Pro233Ala	CAP1
10:124166131	0.00009	+	–	–	+	+	–	+	+	+	+	+	+	Asn95Ser	PLEKHA1
11:709588	0.00010	–	+	+	+	+	+	–	–	+	+	+	+	Met27Thr	EPS8L2
6:30955066	0.00015	–	–	–	+	+	–	+	–	+	+/+	+	+	Thr372Ala	MUC21
7:156742787	–	+	+	+	+	+	+	+	–	+	+	+	+	Ala119Val	NOM1
7:158555855	0.00002	+	+	+	+	+	+	+	–	+	+	+	+	Ala416Va	ESYT2

Table 3.2. The 9 nonsynonymous coding variants with ExAC Minor allele frequency (MAF) < 0.0002 that are shared between all affected members of family 7 (II.1, II.5, III.9, III.10, and III.11) and that were retained prior to segregation analysis. Individuals III.6 has an unknown affection status and the remaining individuals are unaffected. All variants are missense except the FLNC variants, is a splice site mutation. AC and AF are allele count and allele frequency respectively from our DCM sample of 78 individuals.

Figure 3.2. Fixed heart specimen from III.11

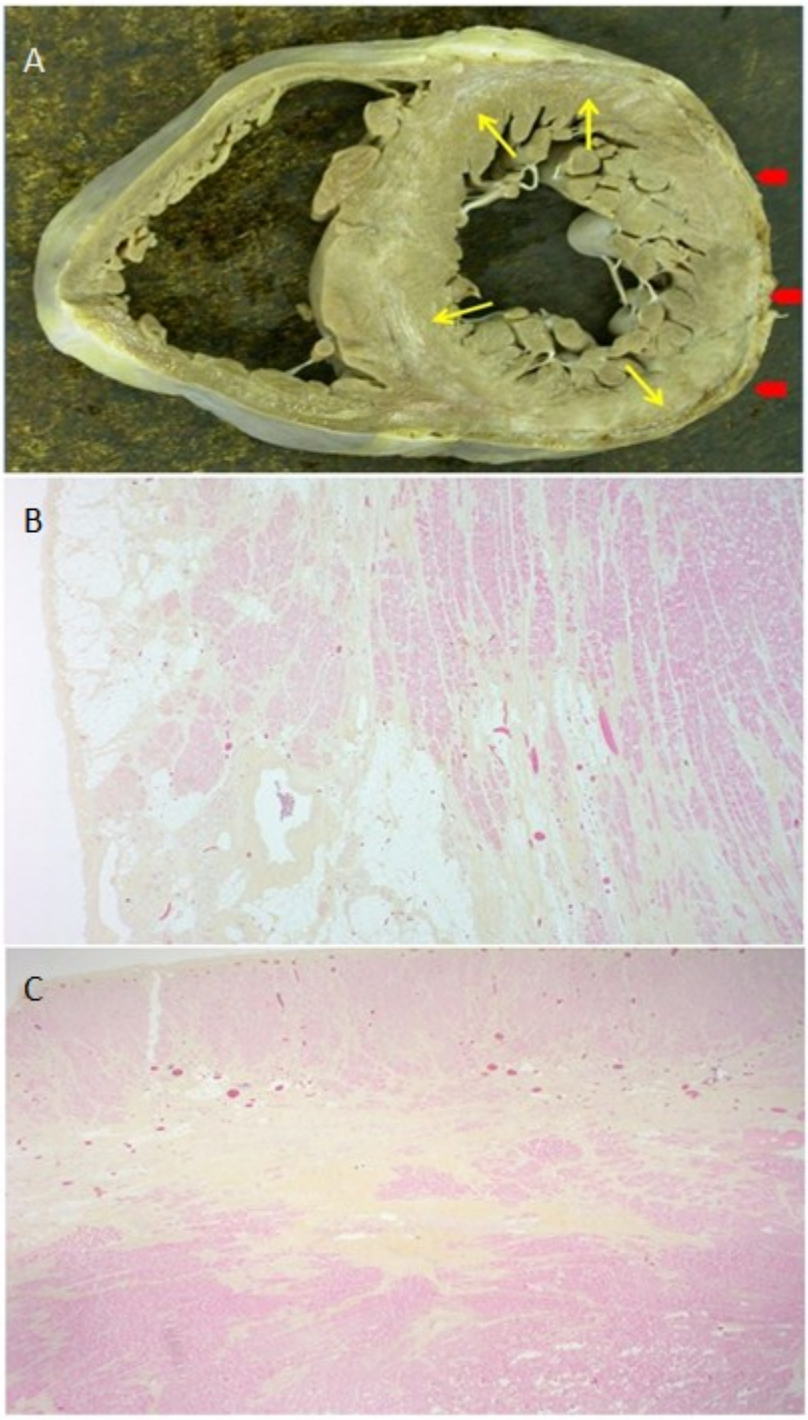


Figure 3.2. **A.** short axis of the heart showing extensive fibrosis. Yellow arrows indicate the circumferential fibrous band in the outer third of the left ventricle. The red arrows indicate the subepicardial fibro-fatty infiltration. **B.** Subepicardium of the left ventricle free wall showing fibro-fatty infiltration. **C.** Fibrous band running in the outer third of the left ventricle free wall.

3.5. DISCUSSION

We present a family with a splicing mutation in *FLNC* and an atypical form of cardiomyopathy characterized by left ventricular dilatation with or without systolic dysfunction, arrhythmia, the presence of fibrosis, and a history of SCD with additional overlap with left ventricular arrhythmogenic cardiomyopathy (III.11). Hence, the cardiac phenotype observed does not conform to a distinct type of cardiomyopathy with a stark phenotypic heterogeneity seen among family members.

FLNC: A role in myofibrillar myopathies

Filamins are a family of muscle proteins that have an actin-binding domain and involved in various processes from organization of the cytoskeleton, membrane stabilization, to signal transduction and transcription [210; 211]. Filamin C is largely produced in skeletal and cardiac muscles and interacts with a large number of muscle proteins at the sarcolemma [212; 213]. *FLNC* has long been known to be involved in myofibrillar myopathies[214-218] which are often accompanied by a manifestation of cardiomyopathies. However, *FLNC* soon emerged as a player in cardiomyopathies in individuals in whom muscular weakness and ailments were not necessarily present. Mutations in *FLNC* were linked to HCM in 2014 [219] and more recently with restrictive [220] and dilated cardiomyopathy [221]. A recent study [222] evaluated the association between truncating variants in *FLNC* and cardiomyopathies in 28 unrelated families. Interestingly, they report that truncating variants in *FLNC* cause an overlapping phenotype of dilated and left dominant arrhythmogenic cardiomyopathies with the presence of fibrosis, similar to the phenotype we observe in the family presented here. However, the majority of the patients described in that study also suffered from systolic dysfunction, a typical characteristic

of DCM, which is not the case for family 7. The report by Begay *et al* [221], also described atypical manifestation of DCM in three families that carried splicing mutations in *FLNC*. The same study also demonstrated using a zebrafish model that the splicing variants led to a reduction in cardiac *FLNC* protein with Z-disc and sarcomere disorganization. These reports further confirm the role of *FLNC* in cardiomyopathy and provide strong support for the addition of *FLNC* on cardiomyopathy screening panels.

The fact that the same mutation causes different characteristics of cardiomyopathy within the same family is intriguing. Other mutations have been shown to manifest as two distinct cardiomyopathy phenotypes. For example, the same mutation in *TNNT2* can manifest as HCM in one individual and DCM in another adding yet a second level to the complexity to the heterogeneity observed in these disorders [3].

Affected individuals of family 7 who carried the *FLNC* mutations also carried two other missense variants in genes that interact with actin and play a role in cytoskeletal organization (*CAP1* and *EPS8L2*). We postulate a framework that implicates the segregating *FLNC* mutation as the cause of DCM, due to the nature of the mutation, i.e a splicing unreported mutation, in a gene predominantly expressed in cardiac muscle with an established role in maintenance of muscle integrity, and in which the other variants may be contributing to the distinct phenotypes seen in the affected individuals. Indeed, it would be interesting to explore whether other families with *FLNC* truncating mutations also carry variants in genes involved in actin binding and muscle function. Multiplex families would be particularly informative as they would allow to evaluate if such variants have any modifier effect on phenotype outcomes within and across families.

Conclusions

According to the genetic testing registry, only two labs in the US test for FLNC as part of a cardiomyopathy panel (probably due to its known association with hypertrophic cardiomyopathy in 2014) (Fulgent Genetics and Invitae) and none in Canada. It is also not present on commercial panels (e.g the TruSight cardiomyopathy panel of Illumina). The results presented here corroborate other work [221; 222] and support the role of *FLNC* truncating mutations in cardiomyopathy suggesting that it should be a novel addition to existing gene panels. A common theme in these reports is the atypical manifestation of cardiomyopathy that includes arrhythmia, fibrosis and sudden cardiac death. Studies have shown that fibrosis is associated with a poor prognosis in patients with heart failure [223; 224], ventricular arrhythmias [225; 226], and with a higher likelihood of ICD therapy [227]. Therefore, individuals with the atypical phenotype described here and with *FLNC* truncating mutations should be considered for ICD implantation for primary prevention of sudden cardiac death.

CHAPTER 4: EXOME GENOTYPING IDENTIFIES PLEIOTROPIC VARIANTS ASSOCIATED WITH RED BLOOD CELL TRAITS

Chami N*, Chen MH*, Slater AJ*, Eicher JD, Evangelou E, Tajuddin SM, Love-Gregory L, Kacprowski T, Schick UM, Nomura A, Giri A, Lessard S, Brody JA, Schurmann C, Pankratz N, Yanek LR, Manichaikul A, Pazoki R, Mihailov E, Hill WD, Raffield LM, Burt A, Bartz TM, Becker DM, Becker LC, Boerwinkle E, Bork-Jensen J, Bottinger EP, O'Donoghue ML, Crosslin DR, de Denus S, Dubé MP, Elliott P, Engström G, Evans MK, Floyd JS, Fornage M, Gao H, Greinacher A, Gudnason V, Hansen T, Harris TB, Hayward C, Hernesniemi J, Highland HM, Hirschhorn JN, Hofman A, Irvin MR, Kähönen M, Lange E, Launer LJ, Lehtimäki T, Li J, Liewald DC, Linneberg A, Liu Y, Lu Y, Lyytikäinen LP, Mägi R, Mathias RA, Melander O, Metspalu A, Mononen N, Nalls MA, Nickerson DA, Nikus K, O'Donnell CJ, Orho-Melander M, Pedersen O, Petersmann A, Polfus L, Psaty BM, Raitakari OT, Raitoharju E, Richard M, Rice KM, Rivadeneira F, Rotter JI, Schmidt F, Smith AV, Starr JM, Taylor KD, Teumer A, Thuesen BH, Torstenson ES, Tracy RP, Tzoulaki I, Zakai NA, Vacchi-Suzzi C, van Duijn CM, van Rooij FJ, Cushman M, Deary IJ, Velez Edwards DR, Vergnaud AC, Wallentin L, Waterworth DM, White HD, Wilson JG, Zonderman AB, Kathiresan S, Grarup N, Esko T, Loos RJ, Lange LA, Faraday N, Abumrad NA, Edwards TL, Ganesh SK*, Auer PL*, Johnson AD*, Reiner AP*, Lettre G*

*These authors contributed equally to this study

Reference: Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. Nathalie Chami*, Ming-Huei Chen*, Andrew J. Slater*, John D. Eicher, Evangelos Evangelou, Salman M. Tajuddin *et al*, Am J Hum Genet. 2016 Jul 7; 99 (1):8-21

4.1.ABSTRACT

Red blood cell (RBC) traits are important heritable clinical biomarkers and modifiers of disease severity. To identify novel coding genetic variants associated with these traits, we conducted meta-analyses of seven RBC phenotypes in 130,273 multi-ethnic individuals from studies genotyped on an exome array. Following conditional analyses and replication in 27,480 independent individuals, we identified 14 new RBC loci. We found low-frequency missense variants in *MAP1A* (rs55707100, minor allele frequency (MAF=3.3%, $P=2x10^{-10}$ for hemoglobin (HGB)) and *HNF4A* (rs1800961, MAF=2.4%, $P<3x10^{-8}$ for hematocrit (HCT) and HGB). In African Americans, we identified a nonsense variant in *CD36* associated with higher RBC distribution width (rs3211938, MAF=8.7%, $P=7x10^{-11}$), and showed that it is associated with lower *CD36* expression and strong allelic imbalance in *ex vivo* differentiated human erythroblasts. We also identified a rare missense variant in *ALAS2* (rs201062903, MAF=0.2%) associated with lower mean corpuscular volume and mean corpuscular hemoglobin ($P<8x10^{-9}$). Mendelian mutations in *ALAS2* are a cause of sideroblastic anemia and erythropoietic protoporphyria. Gene-based testing highlighted three rare missense variants in *PKLR*, a known gene of Mendelian non-spherocytic hemolytic anemia, associated with HGB and HCT (SKAT $P<8x10^{-7}$). The novel rare, low-frequency, and common RBC variants showed pleiotropy, being also associated with platelet, white blood cell, and lipid traits. Our association results and functional annotation suggest the involvement of new genes in human erythropoiesis. We also confirm that rare and low-frequency variants play an important role in the architecture of complex human traits, although their phenotypic effect is generally smaller than originally anticipated.

4.2.INTRODUCTION

One in four cells in the human body is a mature enucleated red blood cell (RBC), also called an erythrocyte. RBC mean lifespan in adults is 100-120 days, requiring constant renewal. To that end, we produce on average 2.4 million RBCs per second in the bone marrow. This massive, yet well-orchestrated cell proliferation process is necessary to accommodate RBCs' main function, to transport oxygen from the lungs to the peripheral organs, and carbon dioxide from the organs to the lungs. Hemoglobin (HGB), the metalloprotein that constitutes by far the most abundant biomolecule found in mature RBC, is responsible for oxygen transport. In addition to their critical role in the circulatory system, RBCs also have secondary, often less appreciated, functions. Within blood vessels, they respond to shear stress and produce the vasodilator nitric oxide to regulate vascular tonus[228]. RBCs participate in antimicrobial strategies to fight hemolytic pathogens[229] or in the inflammatory response, acting as a reservoir for multiple chemokines[230]. Furthermore, the direct involvement of RBC in adhering to the vascular endothelium or supporting thrombin generation may help to promote blood coagulation or thrombosis[231; 232].

Given the paramount importance of RBCs in physiology, it is not surprising that monitoring their features is common practice in medicine to assess the overall health of patients. An excessive number of circulating RBCs (erythrocytosis[MIM: 133100]) can suggest a primary bone marrow disease, a myeloproliferative neoplasm such as polycythermia vera (MIM: 263300), or chronic hypoxemia due to congenital heart defects. Low HGB concentration and hematocrit (HCT) levels (anemia) may indicate inherited HGB or RBC structural gene mutations, malnutrition, or kidney diseases. By considering the volume (mean corpuscular volume (MCV)), hemoglobin content (mean corpuscular hemoglobin (MCH) and

mean corpuscular hemoglobin concentration (MCHC)), or the distribution width (RDW) of RBCs, a physician can distinguish between the different causes of anemia (*e.g.* microcytic/ hypochromic due to iron deficiency[233]). In addition, epidemiological studies have correlated high RDW values with a worse prognosis in heart failure patients[234]. RDW is also an independent predictor of overall mortality in healthy individuals, as well as a predictor of mortality in patients with various conditions such as cardiovascular diseases, obesity, malignancies, and chronic kidney disease[235-239].

RBC count and indices vary among individuals, and 40-90% of this phenotypic variation is heritable[108; 117; 240; 241]. Identifying the genes and biological pathways that contribute to this inter-individual variation in RBC traits could highlight modifiers of severity and/or therapeutic options for several hematological diseases. Already, large-scale genome-wide association studies (GWAS) have found dozens of single nucleotide polymorphisms (SNPs) associated with one or several of these RBC traits[123; 242]. However, owing to their design, GWAS are largely insensitive to rare (minor allele frequency [MAF] <1%) and low-frequency (1%≤MAF<5%) genetic variants. Using an exome array, we previously performed an association study for HGB and HCT in 31,340 European-ancestry individuals and identified rare coding or splice site variants in the erythropoietin and β-globin genes[243]. Within the framework of the Blood-Cell Consortium (BCX)[244; 245], we now report a larger genotyping-based exome survey of seven RBC traits conducted in up to 130,273 individuals, including 23,896 participants of non-European ancestry. With this experiment, our initial goals were to expand the list of rare and low-frequency coding or splice site variants associated with RBC traits and to explore whether the exome array can complement the GWAS approach to fine-map RBC causal genes.

4.3.SUBJECTS AND METHODS

Study participants

The Blood-Cell Consortium (BCX) aims at identifying novel common and rare variants associated with blood-cell traits using an exome array. BCX is comprised of more than 134,021 participants from 24 discovery cohorts of five ancestries: European, African American, Hispanic, East Asian, and South Asian. Detailed description of the participating cohorts is provided in **Table S1**. BCX is interested in the genetics of all main hematological measures and is divided into three main working groups: RBC, white blood cell (WBC)[244], and platelet (PLT)[245]. For the RBC working group, we analyzed seven traits available in up to 130,273 individuals: RBC count ($x10^{12}$/L), HGB (g/dL), HCT (%), MCV (fL), MCH (pg), MCHC (g/dL), and RDW (%)(**Table S2**). The BCX procedures were in accordance with the institutional and national ethical standards of the responsible committees and proper informed consent was obtained.

Genotyping and quality-control steps

Participants from the different studies were genotyped on one of the following exome chip genotyping arrays: Illumina ExomeChip v1.0, Illumina ExomeChip v1.1_A, Illumina ExomeChip-12 v1.1, Affymetrix Axiom Biobank Plus GSKBB1, Illumina HumanOmniExpressExome Chip. Genotypes were then called either 1) with the Illumina GenomeStudio GENCALL and subsequently recalled using zCALL; or 2) by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Exome Chip effort[246] (**Table S1**). The same quality-control steps were followed by each participating study. We excluded variants with low genotyping success rate (<95%, except for WHI that

used a cutoff <90%) (**Table S3)**. Samples with call rate <95% (except for SOLID-TIMI 52 and STABILITY that used 94.5% and 93.5% cutoffs, respectively) after joint or zCALL calling and with outlying heterozygosity rate were also excluded. Other exclusions were deviation from Hardy-Weinberg equilibrium ($P<1x10^{-6}$) and gender mismatch. We performed principle component analysis (PCA) or multidimensional scaling (MDS) and excluded sample outliers from the resulting plots through visual inspection, using populations from the 1000 Genomes Project to anchor these analyses. Keeping only autosomal and X chromosome variants for the analysis, we aligned all variants on the forward strand and created a uniform list of reference alleles using the --force alleles command in PLINK[247]. Finally, an indexed variant call fomat file (VCF) was created by each study and checked for allele alignment and any allele or strand flips using the checkVCF package (https://github.com/zhanxw/checkVCF). Prior to performing meta-analyses of the association results provided by each participating study, we ran the EasyQC protocol[248] to check for population allele frequency deviations and proper trait transformation in each cohort.

Phenotype modeling and association analyses

When possible, we excluded individuals with blood cancer, leukemia, lymphoma, bone marrow transplant, congenital or hereditary anemia, HIV, end-stage kidney disease, dialysis, splenectomy, and cirrhosis, and those with extreme measurements of RBC traits (**Table S1**). We also excluded individuals that are on erythropoietin treatment or chemotherapy. Additionally, we excluded pregnant women and individuals with acute medical illness at the time the complete blood count (CBC) was done. For the seven RBC traits, within each study, we adjusted for age, age-squared, gender, the first 10 principle components, and, where applicable, other study specific covariates such as study center using a linear regression model.

Within each study, we then applied inverse normal transformation on the residuals and tested the variants for association with the ExomeChip variants using either RVtests (version 20140416) (http://genome.sph.umich.edu/wiki/RvTests) or RAREMETALWORKER.0.4.9 (http://genome.sph.umich.edu/wiki/RAREMETALWORKER).

Discovery meta-analyses

Score files generated by RVtests or RAREMETALWORKER from each participating study were used to carry out meta-analyses of the single variant association results using RareMETALS v.5.9[249]. All analyses were performed separately in each of European (EA) and African-American (AA) ancestries. In the multi-ancestry meta-analyses, we combined individuals of European, African-American, Hispanic, East-Asian, and South-Asian ancestries (All). We included variants with allele frequency difference between the highest and lowest MAF <0.3 for European and African-American ancestries, and <0.6 for the combined ancestry meta-analyses. For the gene-based analyses, we used score files and variance-covariance matrices from the study-specific association results, and applied the sequence kernel association test (SKAT)[250] and variable threshold (VT) algorithms[251] in RareMETALS considering only missense, nonsense and splice site variants with a MAF <1%. Gene-based analyses were also stratified by ancestry. Significance thresholds were determined using Bonferroni correction assuming ~250,000 independent variants ($P<2x10^{-7}$ for the single variant analyses) and ~17,000 genes tested on the ExomeChip ($P<3x10^{-6}$ for the gene-based tests).

Conditional analysis and replication

In order to identify independent signals, we performed conditional analyses. In each round of conditional analysis, we conditioned on the most significant single variant in a 1 Mb

window. These conditional analyses were performed at the meta-analysis level using RareMETALS. We repeated this step until there were no new signals identified in each region, defined as a $P<2x10^{-7}$. We then checked for linkage disequilibrium (LD) within the list of variants that was retained from the conditional analyses. For variants that were in moderate-to-strong LD ($r^2 \geq 0.3$), we kept the most significant. We attempted replication of the final list of independent variants in eight additional studies that contributed a total of 27,480 individuals (N=21,473 for EA and N=6,007 for AA) (**Table S4**). The division of discovery and replication samples was dictated by timing because we collected all groups we were aware of for initial discovery and then found others who could participate only much later and hence were used for replication. These studies followed similar analytical procedures and steps as those followed by the discovery analysis (see above). A joint meta-analysis of the discovery and the replication results was carried out using a fixed-effects model and inverse-variance weighting as implemented in METAL[252]. We considered as replicated markers those with a nominal $P_{replication}<0.05$ and an effect on phenotype in the same direction as in the discovery results.

Allelic imbalance and expression of CD36

We checked for allelic imbalance of the rs3211938 variant in *CD36* as well as the expression of the gene in 12 samples of fetal liver erythroblasts obtained from anonymous donors. Details on the protocol including RNA extraction and sequencing can be found elsewhere[253]. We calculated the difference in the ratio of reads of the reference allele (T) and the alternate allele (G) of rs3211938. Briefly, reads overlapping rs3211938 were counted using samtools (v 1.1) mpileup software using genome build hg19. We kept uniquely mapping reads using -q 50 argument (mapping quality > 50) and sites with base quality >10. Statistical significance of the difference in the ratio of reads between the reference allele and the alternate

allele was assessed using a binomial test. For each sample, we summed all reads overlapping all heterozygous SNPs and calculated the expected ratio within each SNP allele combination. Reads that fall in the top 25th coverage percentile were down-sampled so that the highest covered sites do not bias the expected ratio[254]. For rs3211938, the expected T:G ratio was 0.507.

Expression quantitative trait loci (eQTL) analysis

We cross-referenced our list of novel variants with over 100 separate expression quantitative trait loci (eQTL) published datasets. Datasets were collected through publications, publically available sources, or private collaborations. A general overview of a subset of >50 eQTL studies has been published[255], with specific citations for >100 datasets included in the current query followed here. A complete list of tissues and studies used can be found in the **Supplemental Data**. We considered SNPs that are themselves expression SNPs (eSNP) when they meet a $P<0.0001$ threshold or when they are in LD ($r^2>0.3$) with the best eSNP ($P<0.0001$).

4.4.RESULTS

Single-variant meta-analyses

We meta-analyzed ExomeChip results for seven RBC-related phenotypes (RBC count, HCT, HGB, MCH, MCHC, MCV, and RDW) available in up to 130,273 individuals from 24 studies and five ancestries (**Tables S1-S3** and **Supplementary Figure 4.1**). Across these different phenotypes, a total of 226 variants reached exome-wide significance ($P<2x10^{-7}$) in the combined ancestry analyses (**Figure 4.1**and **Figure S2**). Given that some of these RBC traits are correlated (**Figure S3**), these associated variants highlight 71 different loci (defined using a 1 Mb interval). Overall, we observed only modest inflation of the test statistics (λ_{GC}=1.03-1.05), suggesting little confounding due to technical artifacts, population stratification, or cryptic relatedness.

In order to identify independent variants, we performed conditional analyses at the meta-analysis level adjusting for the effect of the most significant variant in a 1 Mb region in a stepwise manner (**Subjects and Methods**). After this analysis, we obtained a list of 126 independent variants associated with at least one RBC trait at $P<2x10^{-7}$ (**Table S5**). Selecting only variants that lie more than 1 Mb away from a known GWAS locus resulted in 23 independent variants located within 20 novel RBC loci (**Table 4.1**). We attempted to replicate these 126 variants in 8 independent cohorts totaling 27,480 participants (**Table S5**). Overall, we observed a strong replication, with 94 of the 126 variants showing consistent direction of effect between the discovery and replication analyses (binomial $P=3x10^{-8}$, **Table S5**). Of the 23 novel variants, we replicated 16 at nominal $P<0.05$ for at least one RBC trait (binomial

P=3x10^{-16}, **Table 4.1**). Out of these 16 novel and replicated RBC variants, there are five missense variants, including two variants with MAF<5% in *MAP1A* (MIM: 600178) and *HNF4A* (MIM: 600281) and one nonsense variant in *CD36.* (**Table 4.1**). Among the remaining nine novel and replicated RBC variants, there are five intronic, one synonymous, one 5′ UTR, and one intergenic marker (**Table 4.1**).

Figure 4.1. Quantile-Quantile (QQ) plots of single variant association results in the all ancestry meta-analyses for the seven red blood cell (RBC) traits analyzed

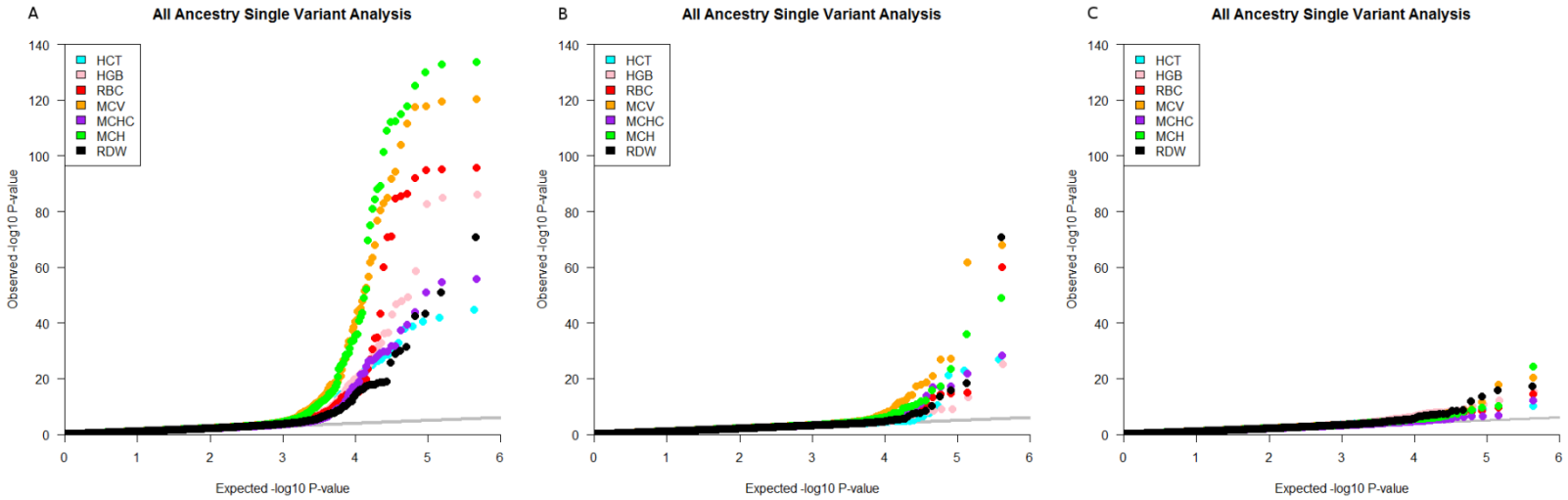


Figure 4.1. **(A)** Distribution of the single variant results for all variants tested on the exome array. (**B**) Only markers with a minor allele frequency <5% are shown here. (**C**) Variants outside of known RBC GWAS regions. Variants that are within 1 Mb from a previously published RBC GWAS locus were excluded for this QQ plot. Abbreviations are as follows: HCT, hematocrit; HGB, hemoglobin; RBC, red blood cell count; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin; RDW, red blood cell distribution width.

Table 4.1. Association results of variants in novel loci associated with red blood cell (RBC) traits.

Marker Info							Discovery				Replication		Combined	
Trait	A1/A2	SNP	Annotation	Gene	N	AF (A2)	Beta (SE)	P-value	N	AF (A2)	Beta (SE)	P-value	Beta (SE)	P-value
RDW-EA	A/G	rs10903129*	intron	TMEM57-RHD	45573	0.544	0.037(0.007)	1.19E-07	18475	0.560	0.023(0.011)	0.0373	0.033(0.006)	2.41E-08
RDW-All	A/G	rs10903129*	intron	TMEM57-RHD	56194	0.568	0.034(0.006)	9.58E-08	24474	0.600	0.021(0.01)	0.0252	0.03(0.005)	1.32E-08
HCT-All	C/T	rs4072037*	synonymous	MUC1	109875	0.554	0.025(0.005)	5.82E-08	25006	0.563	0.038(0.009)	5.96E-05	0.027(0.004)	3.47E-11
HGB-All	T/C	rs780094	intron	GCKR	130,273	0.626	0.024 (0.004)	7.14E-08	3162	0.626	-0.012(0.026)	0.6410	0.023(0.044)	1.68E-07
RBC-All	C/A	rs2230115*	missense	ZNF142	74488	0.509	0.033(0.006)	9.74E-09	27442	0.477	0.024(0.01)	0.0167	0.031(0.005)	7.11E-10
HCT-All	A/C	rs3772219*	missense	ARHGEF3	109875	0.338	-0.028(0.005)	2.38E-09	25006	0.366	-0.021(0.01)	0.0292	-0.027(0.004)	2.56E-10
HGB-All	A/C	rs3772219*	missense	ARHGEF3	130273	0.336	-0.026(0.004)	3.76E-09	27749	0.367	-0.02(0.009)	0.0331	-0.025(0.004)	4.33E-10
HCT-EA	G/A	rs236985*	intron	AFF1	87444	0.394	0.032(0.005)	3.89E-10	19968	0.405	0.02(0.011)	0.0626	0.03(0.005)	1.14E-10
RBC-EA	G/A	rs236985	intron	AFF1	60231	0.393	0.034(0.006)	3.50E-08	21435	0.405	0.023(0.011)	0.0273	0.031(0.005)	4.22E-09
HGB-EA	G/T	rs442177*	intron	AFF1	106377	0.595	-0.034(0.005)	3.97E-13	21743	0.586	-0.029(0.01)	0.0052	-0.033(0.004)	8.23E-15
RDW-EA	A/G	rs10063647*	intron	LINC01184-SLC12A2	45573	0.463	-0.05(0.007)	1.72E-13	18475	0.480	-0.033(0.011)	0.0018	-0.045(0.006)	2.88E-15
RDW-All	A/G	rs10063647*	intron	LINC01184-SLC12A2	56194	0.506	-0.044(0.006)	2.11E-12	24474	0.545	-0.03(0.01)	0.0014	-0.04(0.005)	2.37E-14
RDW-EA	C/T	rs10089*	utr_5p	LINC01184-SLC12A2	45573	0.21	0.051(0.008)	8.45E-10	16692	0.215	0.058(0.014)	2.71E-05	0.053(0.007)	1.15E-13
RDW-All	C/T	rs10089*	utr_5p	LINC01184-SLC12A2	56194	0.207	0.044(0.008)	4.08E-09	22691	0.208	0.045(0.012)	0.0001	0.044(0.006)	2.73E-12
HGB-All	C/A	rs35742417*	missense	RREB1	130273	0.174	0.030 (0.005)	1.17E-08	4074	0.207	0.065 (0.028)	0.0190	0.032 (0.005)	1.50E-09
RDW-AA	T/G	rs3211938*	nonsense	CD36	6666	0.087	0.174(0.031)	2.36E-08	5999	0.086	0.139(0.032)	1.83E-05	0.161(0.025)	7.09E-11
RDW-All	T/G	rs3211938*	nonsense	CD36	55510	0.012	0.171(0.029)	5.29E-09	22691	0.023	0.139(0.032)	1.61E-05	0.157(0.022)	5.12E-13
RDW-EA	A/T	rs2954029*	intergenic	TRIB1	45573	0.46	0.036(0.007)	1.53E-07	16692	0.466	0.026(0.011)	0.0210	0.034(0.006)	1.29E-08
RDW-All	A/T	rs2954029*	intergenic	TRIB1	56194	0.439	0.032(0.006)	1.83E-07	22691	0.432	0.021(0.01)	0.0298	0.029(0.005)	2.54E-08
MCH-All	T/C	rs2487999	missense	OBFC1	66318	0.869	0.047(0.009)	4.12E-08	26749	0.861	0.025(0.013)	0.0601	0.041(0.007)	1.75E-08
MCH-AA	G/A	rs1447352	intergenic	MTNR1B	8273	0.557	0.089(0.016)	1.85E-08	5038	0.562	-0.022(0.02)	0.2713	0.07(0.014)	1.08E-06
HGB-EA	C/T	rs55707100*	missense	MAP1A	106377	0.033	-0.071(0.013)	1.65E-08	21743	0.0223	-0.099(0.033)	0.0028	-0.075(0.012)	2.31E-10
MCV-AA	A/G	rs2667662*	intron	TELO2	10849	0.725	-0.099(0.015)	1.79E-10	5034	0.724	-0.093(0.022)	3.02E-05	-0.098(0.014)	7.32E-12
MCV-AA	C/A	rs2240140*	missense	SRRM2	8525	0.118	0.134(0.025)	7.08E-08	6002	0.124	0.106(0.027)	0.0001	0.128(0.022)	5.24E-09
HCT-EA	T/C	rs8080784	intron	BCAS3-TBX2	79344	0.158	-0.039(0.007)	2.62E-08	19968	0.148	0.011(0.014)	0.4349	-0.029(0.006)	3.39E-06
HGB-EA	C/T	rs8068318	intron	BCAS3-TBX2	106377	0.722	-0.026(0.005)	1.53E-07	21743	0.730	-0.021(0.011)	0.0565	-0.025(0.005)	2.55E-08

Marker Info							Discovery				Replication		Combined	
MCV-EA	C/T	rs4911241*	intron	NOL4L	61462	0.241	-0.04(0.007)	1.25E-08	21714	0.252	-0.025(0.012)	0.0302	-0.036(0.006)	2.01E-09
RDW-EA	C/T	rs4911241*	intron	NOL4L	45573	0.242	0.043(0.008)	5.79E-08	18475	0.240	0.049(0.012)	7.44E-05	0.045(0.007)	2.01E-11
RDW-All	C/T	rs4911241*	intron	NOL4L	56194	0.235	0.038(0.007)	1.56E-07	24474	0.222	0.044(0.011)	6.10E-05	0.04(0.006)	4.60E-11
HCT-EA	C/T	rs1800961*	missense	HNF4A	79344	0.024	0.083(0.015)	1.44E-08	19968	0.033	0.082(0.028)	0.0037	0.083(0.013)	1.91E-10
HGB-EA	C/T	rs1800961*	missense	HNF4A	98277	0.032	0.073(0.013)	2.53E-08	21743	0.032	0.062(0.027)	0.0232	0.071(0.012)	1.93E-09
HCT-All	C/T	rs1800961*	missense	HNF4A	100313	0.022	0.077(0.014)	2.31E-08	25006	0.027	0.091(0.028)	0.0010	0.08(0.012)	9.88E-11
HGB-All	C/G	rs738409	missense	PNPLA3	130273	0.223	0.028(0.005)	2.24E-08	4074	0.218	0.053(0.027)	0.0504	0.029(0.005)	4.81E-09
MCH-EA	G/A	rs201062903	missense	ALAS2	52758	0.002	-0.324(0.053)	7.32E-10	5855	0.001	-0.291(0.235)	0.215	-0.323(0.052)	5.81E-10
MCH-All	G/A	rs201062903	missense	ALAS2	65067	0.002	-0.322(0.051)	3.36E-10	10893	0.001	-0.276 (0.224)	0.218	-0.321(0.051)	2.68E-10
MCV-EA	G/A	rs201062903	missense	ALAS2	60211	0.002	-0.285 (0.049)	7.11E-09	5044	0.001	-0.178 (0.248)	0.472	-0.282 (0.049)	6.11E-09

Table 4.1. Variants in novel loci with $P<2x10^{-7}$ and that were retained after conditional analyses are presented here. All variants are >1Mb apart from a known GWAS signal for RBC traits. Allele frequency and effect size are given for the alternate (A2) allele. Replication was carried in six cohorts for EA and two cohorts for AA and was performed in RareMetals; meta-analyses of the discovery and replication cohorts is presented under "Combined" and was carried in METAL. Asterisks (*) indicate variants that replicated with a nominal $p < 0.05$. Abbreviations are as follows: EA, European American; AA, African American; All, combined ancestry (EA + AA + Asians + Hispanics); A1, reference allele; A2, alternate allele; N, sample size; AF, allele frequency; SE, standard error; HCT, hematocrit; HGB, hemoglobin; RBC, red blood cell count; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin; RDW, red blood cell distribution width.

Prioritization of candidate genes and genetic variants

Our single-variant analyses in EA samples identified one rare missense variant in *ALAS2* (MIM: 301300) associated with MCV and MCH (rs201062903, p.Pro507Leu [c.1559C>T], MAF = 0.2%) (**Table 4.1**). The association with this variant did not replicate, potentially because of limited statistical power (the replication sample size for this rare marker was 5,044; see also Discussion). *ALAS2* encodes 5-aminolevulinate synthase 2, the rate-controlling enzyme of erythroid heme synthesis. Additionally, rare mutations in *ALAS2* cause X-linked sideroblastic anemia (MIM: 300751) and erythropoietic protoporphyria (MIM: 300752). Thus, despite the lack of replication, *ALAS2* remains an excellent candidate gene to modulate RBC traits. The *ALAS2* p.Pro507Leu variant, which is not reported in the ClinVar database, maps between two amino acids (Tyr506 and Thr508) that are important for catalytic activity and known to be mutated in cases of sideroblastic anemia[256].

Two low-frequency missense variants identified in our analyses implicate *MAP1A* and *HNF4A* for the first time in RBC biology (**Table 4.1**). *MAP1A* encodes microtubule-associated protein 1A, a gene highly expressed in the nervous system and mostly studied in the context of neuronal diseases, although it is expressed in many additional tissues, including hematopoietic cells[257]. Deletion of *MAP1A* in the mouse causes defects in synaptic plasticity[258]. This observation is interesting given that inactivation of *ANK1*, another gene that encodes a cytoskeleton protein and is expressed in neurons and RBCs, is associated with neurological dysfunction in the mouse and spherocytosis and hemolytic anemia in humans (MIM: 182900). Our meta-analyses confirmed two known independent *ANK1* variants associated with MCHC: an intronic SNP (rs4737009, MAF=19.8%, P=1.3x10^{-8}) and a low-frequency missense variant

(rs34664882, p.Ala1462Val, MAF=2.9%, P=1.7x10^{-16}) (**Table S5;** N.P., U.M.S., J.B.-J., and M.-H.C., unpublished data). [123].

In the accompanying BCX PLT article[245], we report that the same *MAP1A* rs55707100 allele (p.Pro2349Leu) associated here with decreased HGB concentration is also associated with increased PLT count. Furthermore, recent studies have identified associations between rs55707100 and HDL-cholesterol and triglycerides levels[259]. Adding to the complexity, the GTex dataset indicates that rs55707100 is an expression quantitative trait locus (eQTL) for the *ADAL* gene (P_{eQTL}=9x10^{-11}) but not for *MAP1A*[260]. *ADAL* is a poorly characterized adenosine deaminase-like protein that is highly expressed in human erythroblasts. However, the eQTL association between rs55707100 and *ADAL* could simply reflect "LD shadowing" from nearby markers that are much stronger eQTL variants for *ADAL*. Indeed, rs3742971 (a common variant located in *ADAL*'s 5′ UTR) is in partial LD with rs55707100 (r^2 = 0.18 in European populations from the 1000 Genomes Project) and strongly associated with *ADAL* expression levels ($p_{eQTL} = 6 \times 10^{-49}$).

The second low-frequency missense variant associated with HGB and HCT maps within the coding sequence of the transcription factor HNF4A (**Table 4.1**). This marker, rs1800961 (p.Thr117Ile [c.350C>T]), has previously been associated with HDL- and total cholesterol, C-reactive protein, fibrinogen, and coagulation factor VII levels[261-264]. Mutations in HNF4A cause maturity-onset diabetes of the young (MODY [MIM: 125851]) and a common intronic SNP in HNF4A (rs4812829) has been associated with type 2 diabetes (MIM: 125853) risk[265]. The missense rs1800961 associated with HGB and HCT is only in weak LD with rs4812829 (r2 = 0.021 in EA populations from the 1000 Genomes Project). Querying recently released

ExomeChip data from Type 2 Diabetes Genetics (Web Resources), we found that rs1800961 is also associated with T2D risk in ~82,000 participants ($p = 9.5 \times 10-7$, odds ratio = 1.16). *HNF4A* is expressed in the kidney and could influence HGB and HCT through the regulation of erythropoietin production[266]. It is also abundantly expressed in the liver, where it could indirectly affect HGB and HCT levels through an effect on blood lipid levels (see Discussion). *HNF4A* is detectable at low levels in erythroblasts, and the BLUEPRINT Project has found that some *HNF4A* isoforms may be more highly expressed in this cell type (**Figure S4**)[267].

In AA, we identified a nonsense variant (rs3211938, p.Tyr325Ter [c.975T>G], MAF = 8.7%, $p = 7.1 \times 10^{-11}$) in *CD36* associated with RDW. This variant displays a wide variation in allele frequency between AA and EA ($MAF_{EA} = 0.01\%$). The association is slightly improved in the ancestry-combined meta-analysis ($p = 5.1 \times 10^{-13}$) because there is also evidence of association in Hispanics (MAF = 1.9%, p = 0.022) (**Table 4.1**). We examined a dataset of *ex vivo* differentiated human erythroblasts to check if this nonsense *CD36* variant (rs3211938) shows allelic imbalance (AI)[253]. All samples were homozygous at rs3211938 for the reference allele with the exception of one heterozygous sample (FL11). FL11 had the lowest level of *CD36* expression among the 12 samples tested and demonstrated strong AI where we observe 705 sequence reads for the reference allele (T) versus 126 for the alternate allele (G) ($p = 4.9 \times 10^{-95}$; **Figure 4.2**). To confirm this finding in independent samples, we queried the GTEx dataset, which has compiled RNA-sequencing and genotype information from multiple human tissues[260]. GTEx does not include data for human erythroblasts. However, it detected a strong eQTL effect of rs3211938 on *CD36* expression in whole blood ($p_{eQTL} = 1.1 \times 10-15$), with carriers of the G-allele expressing less CD36 (Figure S5). Furthermore, GTEx reported

evidence for moderate AI in multiple tissues for CD36-rs3211938, with the G-allele being under-represented among sequence reads (Figure S5). These results strongly support our observations in human erythroblasts.

eQTL analysis

To prioritize additional causal genes at RBC loci that contain non-coding variants, we cross-referenced our list of novel variants with over 100 published eQTL datasets (**Subjects and Methods**). Overall, 12 variants were significant eQTLs in a wide variety of tissues (**Table S6**). The most interesting eQTL finding is the association between rs10903129, a common marker associated with RDW in our analyses and located within an intron of *TMEM57* (MIM: 610301), and the expression of *RHD* (MIM: 111680) in whole blood. *RHD* is located 112 kb downstream of *TMEM57* and encodes the D antigen of the clinically significant Rhesus (Rh) blood group. rs10903129 has also been associated with total cholesterol levels and erythrocyte sedimentation rate (ESR) [268; 269]. The association with ESR is particularly intriguing given that it is considered a non-specific indicator of inflammation. As described above, RDW is also abnormal in chronic diseases, such as atherosclerosis and diabetes, which have an important inflammation component.

Figure 4.2. *CD36* expression in human erythroblasts.

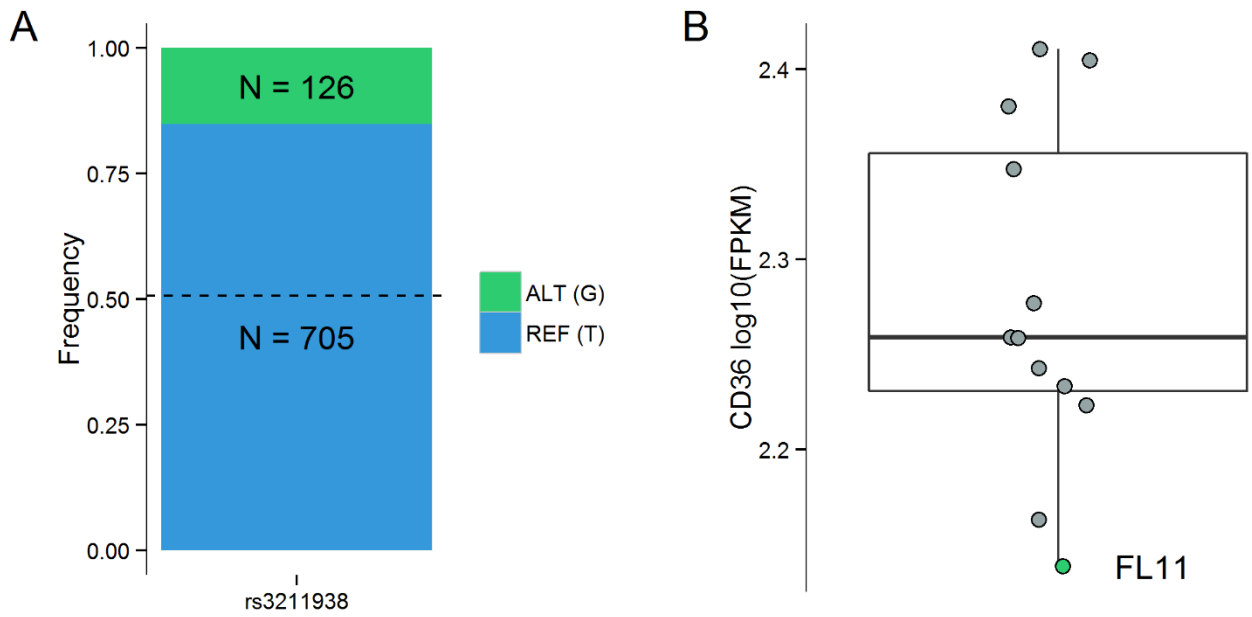


Figure 4.2. (**A**) In a dataset of 12 human fetal liver erythroblasts, all samples were homozygous at rs3211938 for the reference T-allele with the exception of one heterozygous sample (FL11). FL11 demonstrated strong allelic imbalance: we observe 705 reads for the reference allele (T) and 126 reads for the alternate allele (G)(binomial P=4.9x10^{-95}). (**B**) FL11 (in green) shows the lowest *CD36* expression level when compared to the other 11 samples. Abbreviation is as follows: FPKM, fragments per kilobase of transcript per million mapped reads.

Gene-based association testing

Despite our large sample size, statistical power remains limited for rare variants of weak-to-moderate phenotypic effect. To try to capture these genetic factors, we performed gene-based testing by aggregating coding and splice site variants with MAF < 1% within each gene (Subjects and Methods). The SKAT analyses identified two genes: *ALAS2* associated with MCH and *PKLR* (MIM: 609712) associated with HGB and HCT (**Table 4.2**). The *ALAS2* signal was driven by a single rare missense variant (rs201062903) and was described above. *PKLR* encodes the erythrocyte pyruvate kinase (PK) that catalyzes the last step of glycolysis. PK deficiency, usually caused by recessive mutations, is one of the main causes of non-spherocytic hemolytic anemia (MIM: 266200). In fact, one of the variants identified in our meta-analysis (rs116100695, p.Arg486Trp [c.1456T>G], MAF = 0.3%, $beta_{HGB}$ = −0.242 g/dl, $p_{HGB} = 1.2 \times 10^{-5}$) is a frequent cause of PK deficiency in Italian and Spanish subjects[270; 271]. This variant was confirmed in the replication cohorts ($p_{replication}$ = 0.039; Table S7). Two additional *PKLR* rare missense variants contribute to the gene-based association statistic with HGB and HCT: rs61755431 (p.Arg569Gln [c.1706G>A], MAF = 0.2%, betaHGB = −0.179 g/dl, pHGB = 0.006) and rs8177988 (p.Val506Ile [c.1516G>A], MAF = 0.6%, betaHGB = +0.116 g/dl, pHGB = 0.003). It is noteworthy that the p.Val506Ile substitution is associated with increased HGB concentration given that this amino acid maps to a PKLR structural domain necessary for protein interaction. [272]. This heterogeneity of effect among the *PKLR* missense variants might explain why SKAT's result is more significant than VT's for this gene (**Table 4.2**). A third gene, *ALPK3*, was identified only in the VT analysis for association with MCHC (**Table 4.2**). *ALPK3* encodes a kinase previously implicated in cardiomyocyte

differentiation [273]. We could not test for replication because of the rarity of ALPK3's coding alleles (Table S7).

Table 4.2. Gene-based association results

				VT	SKAT			
Trait	Gene	N	variants analyzed	P-value	P-value	Top variant	Top-variant MAF	Top-variant P-value
HGB-EA	PKLR	106,377	15	1.92E-05	7.02E-07	rs116100695	0.003	1.17E-05
HGB-All	PKLR	130,273	15	0.00016	6.57E-07	rs116100695	0.003	1.94E-05
HCT-All	PKLR	109,875	15	3.96E-05	7.95E-07	rs116100695	0.003	2.49E-05
MCH-EA	ALAS2	54,009	11	4.78E-06	5.79E-07	rs201062903	0.002	7.32E-10
MCHC-All	ALPK3	84,841	28	1.95E-06	0.793	rs202037221	$3.0x10^{-5}$	0.0005

Table 4.2. Gene-based results of the VT and SKAT algorithms for genes associated with RBC traits at $p < 3 \times 10^{-6}$. We analyzed non-synonymous coding (nonsense, missense) and splice site variants with a minor allele frequency (MAF) < 1%. Abbreviations are as follows: EA, European American; All, combined ancestry (EA + AA + Asians + Hispanics); n, sample size; HCT, hematocrit; HGB, hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin.

RBC variants and pleiotropic effects

Besides the overlap within the RBC traits themselves, we identified seven novel RBC variants associated with other blood-cell type traits or with lipid levels (**Figure 4.3** and **Table 4.3**). To assess whether the genetic associations with RBC traits are independent of lipid levels, we performed additional analyses in a subset of BCX participants from three of our studies (FHS, MHIBB, and WHI) ranging from ∼10,000 to 23,000 individuals. We repeated the association analyses for five RBC loci (*TMEM57-RHD* rs10903129, *AFF1* rs442177, *TRIB1* rs2954029, *MAP1A* rs55707100, and *HNF4A* rs1800961) additionally adjusting for the respective lipid trait and combined the results across the three studies using fixed-effect meta-analysis (Table S8). There was little or no change in the effect size or p values associated with the five RBC trait loci upon adjustment for the corresponding lipid trait, suggesting that the RBC and lipid associations are independent of one another and thus represent true "pleiotropic" genetic effects.

A correlated response to or role in inflammation might explain why some of the RBC variants are also associated with WBC, PLT, or lipid traits. Another plausible explanation for

the concomitant association of several markers with RBC, WBC, and PLT phenotypes could be a more general effect of these genes on the proliferation or differentiation of hematopoietic progenitor cells. This is most likely the case for *JAK2* (MIM: 147796) and *SH2B3* (MIM: 605093), two key regulators of hematopoietic cells (**Figure 4.3**). In this category, we also observed two novel findings, *AFF1* (MIM: 159557) and *NOL4L*, which are associated with RBC and WBC phenotypes and have been previously implicated in leukemia [274; 275]. Finally, we identified a novel missense variant in *ARHGEF3* (MIM: 612115) associated with HGB and HCT. In addition to its association with PLT traits, *ARHGEF3* plays a role in the regulation of iron uptake and erythroid cell maturation[276].

Table 4.3. Overlap of red blood cell (RBC) markers with other blood cell traits and/or lipids.

SNP	Position	A1/A2	EAF	Annotation	Gene	Trait	Beta	*P*-value
rs10903129	1:25768937	A/G	0.568	intron	TMEM57-RHD	RDW	0.037	1.19E-07
						TC[268]	0.061	5.40E-10
						PLT	-0.021	7.06E-06
rs3772219	3:56771251	A/C	0.338	missense	ARHGEF3	HCT*	-0.028	2.38E-09
						HGB*	-0.026	3.76E-09
						PLT	0.031	5.93E-10
rs442177	4:88030261	G/T	0.595	intron	AFF1	HGB	-0.034	3.97E-13
						TG[262]	-0.031	1.00E-18
						BASO	-0.030	1.99E-05
rs2954029	8:126490972	A/T	0.439	intergenic	TRIB1	RDW	0.036	1.53E-07
						TG[262]	-0.076	1.00E-107
rs55707100	15:43820717	C/T	0.033	missense	MAP1A	HGB	-0.071	1.65E-08
						PLT	0.095	7.03E-14
						TG[101]	0.090	1.40E-17
rs4911241	20:31140165	C/T	0.241	intron	NOL4L	MCV	-0.040	1.25E-08
						RDW	0.043	5.79E-08
						BASO	-0.051	1.35E-10
						MONO	-0.033	3.57E-05
rs1800961	20:43042364	C/T	0.032	missense	HNF4A	HCT	0.083	1.44E-08
						HGB	0.073	2.53E-08
						HDL[262]	-0.127	2.00E-34

Table 4.3. Shown here are significant novel variants from the RBC traits association analyses that overlap with other blood-cell traits or with lipids. Results for the white blood cell and platelet traits are from the Blood-Cell Consortium, and results for lipids are from the published literature. Results are presented for European-ancestry individuals, except in the presence of an asterisk (*), which stands for result from "All" ancestry. The allele frequency and direction of the effect (beta) is given for the A2 allele. Abbreviations are as follows: A1, reference allele; A2, alternate allele; AF, allele frequency; HCT, hematocrit; HGB, hemoglobin; MCV, mean corpuscular volume; RDW, red blood cell distribution width; TC, total cholesterol; PLT, platelet; TG, triglycerides; WBC, white blood cells; BASO, basophils; MONO, monocytes; HDL, HDL cholesterol.

Figure 4.3. Venn Diagram Summarizing Pleiotropic Effects for Genetic Variants Associated with Red Blood Cell Traits.

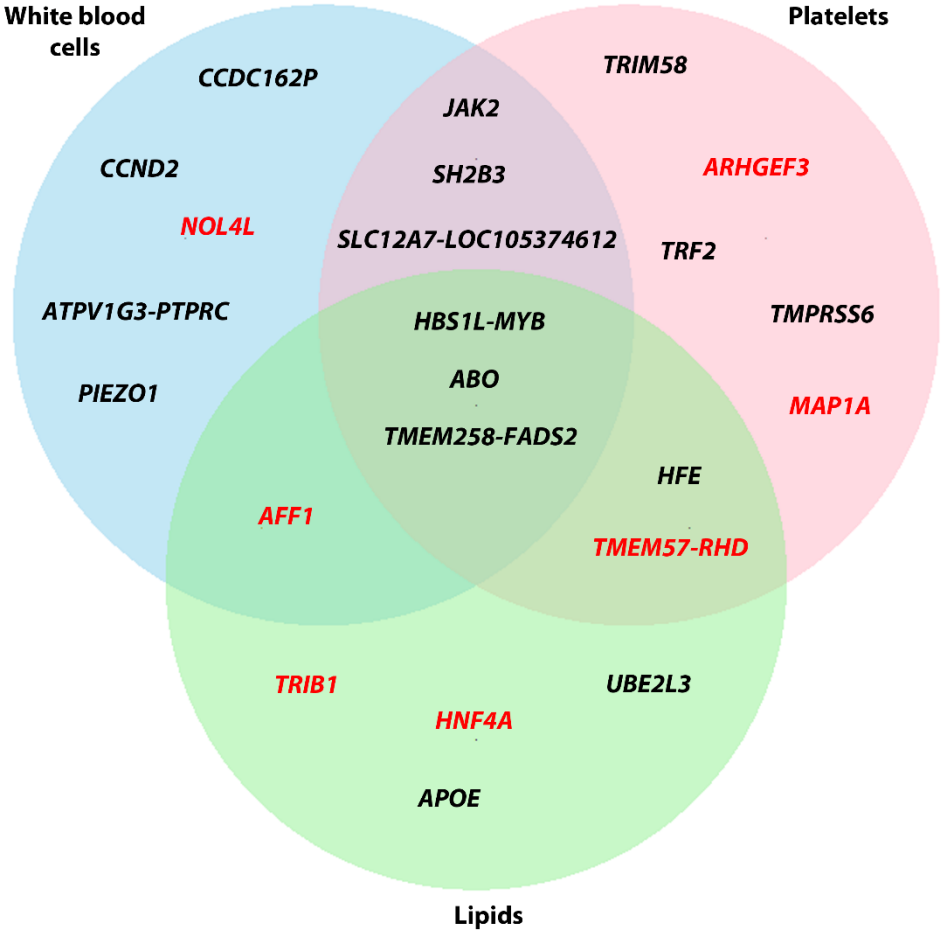


Figure 4.3. We considered variants only if their association p values with white blood cell (WBC) traits, platelet (PLT) traits, or with lipid levels was $p < 1 \times 10^{-4}$. Results for WBC and PLT are from the accompanying Blood-Cell Consortium articles[244; 245]. Results for lipids have previously been published (**Table 4.3**). Genes highlighted in red are novel RBC trait findings.

4.5.DISCUSSION

We present multi-ethnic meta-analyses of seven RBC traits using ExomeChip results of 130,273 individuals. Our statistical thresholds to declare significance at the discovery stage ($p < 2 \times 10-7$ in the single-variant analyses) was adjusted for the approximate number of variants genotyped on the ExomeChip (Bonferroni correction for 250,000 variants), but we decided not to adjust it for the seven RBC phenotypes tested because of the high correlation between some of these traits (Figure S3). Instead, we relied on independent replication to distinguish true from probably false positive associations. Despite the limited size of our replication set (27,480 individuals), it was encouraging to detect a strong replication of direction of effect for known and novel RBC variants, suggesting a low false discovery rate. In total, we identified 23 novel variants associated with RBC traits in the single-variant analyses and a collection of three rare missense variants in PKLR associated with HGB and HCT in the gene-based analyses. Out of the 23 novel RBC variants, 16 were replicated at $p < 0.05$ in the independent samples (**Table 4.1**). To inform our replication criteria, we conducted a power analysis using a sample size of 20,000 and considering multiple combinations of allele frequencies and effect sizes. Based on allele frequency and effect size, one of our most difficult to replicate variants was rs1800961 (MAF = 0.022, Beta = 0.028). However, we still had approximately 56% power to detect this association in the replication stage.

We identified a nonsense variant in *CD36* associated with RDW in African Americans. CD36 is a type B scavenger receptor located on the surface of many cell types, including endothelial cells, platelets, monocytes, and erythrocytes. CD36 is a marker of erythroid

progenitor differentiation[277] and might also be involved in macrophage-mediated clearance of red cells[278]. Furthermore, CD36 plays a role in many biological pathways such as lipid metabolism/transport and atherosclerosis, hemostasis, and inflammation[279]. The nonsense *CD36* variant identified in our RDW meta-analysis (rs3211938, **Table 4.1**) has previously been associated with platelet count, HDL-cholesterol and C-reactive protein levels in African Americans[135; 280], and malaria resistance in Africans[281; 282]. The *CD36* locus shows a signature of natural selection in African-ancestry populations[283] and the MAF of rs3211938 varies widely between continents: in the 1000 Genomes Project, the minor allele is absent from European populations but reaches frequency of 24-29% in some African populations[284]. To characterize the molecular mechanism by which rs3211938 may impact RDW, we identified one heterozygous sample among a collection of *ex vivo* differentiated human erythroblasts[253]. In erythroblasts from this donor, we noted a strong allelic imbalance (**Figure 4.2**). Importantly, this result was confirmed in independent samples from the GTex dataset (Figure S5). At the molecular level, this CD36 expression phenotype could be explained by nonsense-mediated mRNA decay or the regulatory effect of non-coding genetic variants in LD with rs3211938.

We observed that many new RBC variants are pleiotropic, being often associated with more than one RBC index as well as with WBC, PLT, and lipid traits (**Figure 4.3**). These shared effects could imply that the underlying causal genes at these RBC loci generally controlled blood cell proliferation or modulate inflammatory responses. An additional explanation for the link between RBC traits and lipid variants might be the cholesterol content of RBC membranes. As mentioned earlier, RBC corresponds to a large fraction (~25%) of the

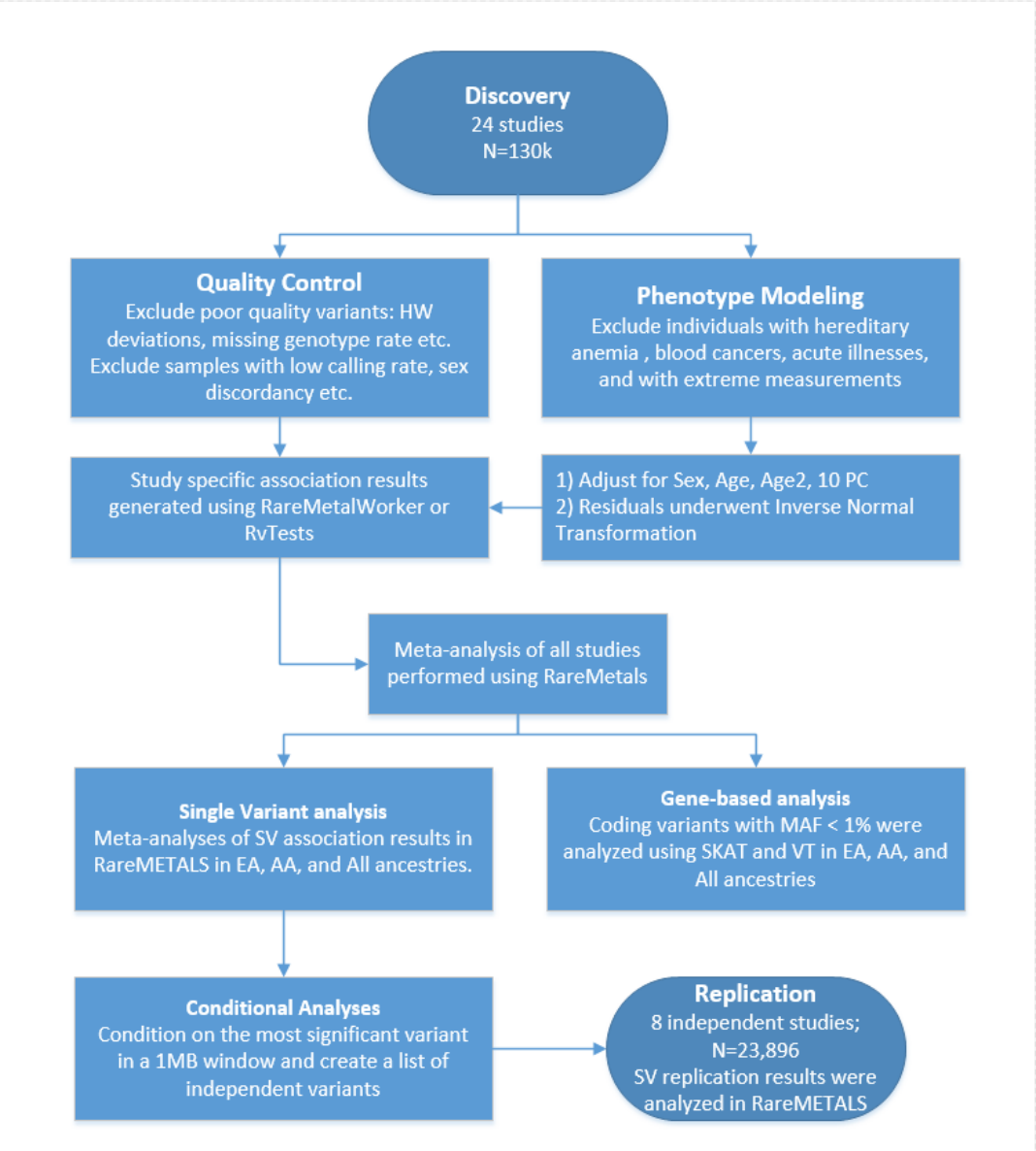
cells found in the human body. Genetic variation that modulates RBC count or volume could impact circulating lipid levels. In support of this hypothesis, it has been observed that a thalassemia allele is strongly associated with cholesterol levels in the Sardinian population[285]. In total, we found ten loci associated with lipid levels and RBC indices, including four novel RBC variants (*AFF1*, *TMEM57-RHD*, *TRIB1*, *HNF4A*) (**Figure 4.3**).

In summary, our multi-ethnic meta-analyses have expanded the genetic knowledge of erythrocyte biology and identified a number of common, low-frequency, and rare RBC variants. Many of the new RBC variants are pleiotropic, affecting other complex traits such as WBC, PLT, and blood lipid levels. Although our report demonstrates the utility of the ExomeChip for genetic discovery, it also highlights the challenge to attribute gene causality based only on association results. This is particularly evident for loci with common variants, for which coding and non-coding markers are often statistically equivalent. For instance, we found no examples of RBC coding variants that entirely explain RBC GWAS signals among the seven loci that had both a sentinel GWAS variant and ExomeChip coding markers. Although increasing sample sizes will continue to yield additional RBC loci, it has become incredibly clear that only a combination of well-powered genetic studies, transcriptomic and epigenomic surveys, and functional experiments (e.g., using genome editing) will ultimately pinpoint causal variants and genes that control RBC phenotypes.

4.6. ACKNOWLEDGEMENTS

We thank all participants, staff, and study coordinating centers. We also thank Raymond Doty and Jan Abkowitz for discussion of the ALAS2 finding. We would like to thank Liling Warren for contributions to the genetic analysis of the SOLID-TIMI-52 and STABILITY datasets. Young Finns Study (YFS) acknowledges the expert technical assistance in the statistical analyses by Ville Aalto and Irina Lisinen. Estonian Genome Center, University of Tartu (EGCUT) thanks co-workers at the Estonian Biobank, especially Mr. V. Soo, Mr. S. Smith, and Dr. L. Milani. Airwave thanks Louisa Cavaliero who assisted in data collection and management as well as Peter McFarlane and the Glasgow CARE, Patricia Munroe at Queen Mary University of London, and Joanna Sarnecka and Ania Zawodniak at Northwick Park for their contributions to the study. This work was supported by the Fonds de Recherche du Queébec-Santeé (FRQS, scholarship to N.C.), the Canadian Institute of Health Research (Banting-CIHR, scholarship to S.L. and operating grant MOP#123382 to G.L.), and the Canada Research Chair program (to G.L.). P.L.A. was supported by NHLBI R21 HL121422-02. N.A.A. is funded by NIH DK060022. A.N. was supported by the Yoshida Scholarship Foundation. S.K. was supported by a Research Scholar award from the Massachusetts General Hospital (MGH), the Howard Goodman Fellowship from MGH, the Donovan Family Foundation, R01HL107816, and a grant from Fondation Leducq. Additional acknowledgments and funding information is provided in the Supplemental Data.

Supplementary Figure 4.1. Flow chart of the study design.



Supplementary Figure 4.1. Data was contributed from 24 studies for the discovery phase. We applied QC steps to remove low quality variants and samples. We also excluded individuals with extreme phenotypes. After the proper adjustments, we applied inverse normal transformation on the residuals. We then performed study-specific association analyses using RV test or RareMetalWorker followed by QC of the individual association results. We meta-analyzed the association results using RareMetals and performed single variant (SV) and gene-based analyses. Additionally, we performed conditional analyses on the SV results, and attempted replication of the significant independent markers in the replication phase which comprised 8 independent studies. Abbreviations are as follows: HW: Hardy Weinberg; PC: principle components; SKAT: Sequence Kernel Association Test; VT: Variable threshold test; EA; European ancestry; AA: African American ancestry; All: combined ancestry (EA + AA + Asians + Hispanics).

Supplementary Table 4.1. Expression quantitative trait loci (eQTL) results for variants associated with red blood cell phenotypes.

POS	eSNP	Gene	Trait	Tissue	eSNP.p	Transcript	r2	BestESNP	BestESNP
1:25768937	rs10903129	TMEM57	RDW-EA/All	Whole blood	2.67E-128	RHD	0.669	rs909832	9.81E-198
1:155162067	rs4072037	MUC1	HCT-All	CD16+ neutrophils	2.30E-05	THBS3	1	rs2066981	2.30E-05
2:219509618	rs2230115	ZNF142	RBC-All	CD16+ neutrophils	7.26E-17	CYP27A1	1	rs10187066	7.26E-17
3:56771251	rs3772219	ARHGEF3	HCT/HGB-All	Whole blood	3.10E-21	ARHGEF3	0.682	rs2046823	1.16E-27
3:56771251	rs3772219	ARHGEF3	HCT/HGB-All	Peripheral blood mononuclear cells	4.55E-15	ARHGEF3	SameSNP	rs3772219	4.55E-15
4:88008782	rs236985	AFF1	HCT/RBC-EA	Peripheral blood mononuclear cells	4.42E-18	AFF1	0.932	rs442177	1.05E-18
4:88030261	rs442177	AFF1	HGB-EA	Peripheral blood mononuclear cells	1.05E-18	AFF1	same SNP	rs442177	1.05E-18
5:127371588	rs10063647	LINC01184	RDW-EA/All	Peripheral blood mononuclear cells	2.78E-16	FLJ33630	0.327	rs2250127	2.03E-40
5:127371588	rs10063647	LINC01184	RDW-EA/All	CD14+ monocytes	1.48E-12	FLJ33630	0.327	rs3749748	3.24E-38
5:127522543	rs10089	SLC12A2	RDW-EA/All	Whole blood	2.78E-09	FBN2	0.002	rs764369	9.81E-198
7:80300449	rs3211938	CD36	RDW-AA/All	Whole blood	6.40E-14	CD36	SameSNP	rs3211938	6.40E-14
10:105659826	rs2487999	OBFC1	MCV-All	Liver	2.05E-14	OBFC1	SameSNP	rs2487999	2.05E-14
17:59483766	rs8068318	TBX2	HGB-EA	Fibroblasts	4.09E-06	C17ORF82	1	rs2240736	4.09E-06
17:59483766	rs8068318	TBX2	HGB-EA	Monocytes (CD14+)	9.97E-07	CCDC47	0.527	rs9905140	2.73E-07
20:31140165	rs4911241	NOL4L	MCV/RDW-EA; RDW-All	Peripheral blood mononuclear cells	7.65E-11	ASXL1	0.293	rs6141282	1.85E-22
20:31140165	rs4911241	NOL4L	MCV/RDW-EA; RDW-All	Whole blood	4.37E-07	ASXL1	0.293	rs3746612	9.13E-18

Supplementary Table 4.1. eSNP, SNP associated with the gene expression phenotype; eSNP.p, eQTL association P-value; *r*2, linkage disequilibrium in European populations between the eSNP (from the RBC analyses) and the best eSNP for a given gene; Best_eSNP, best reported eSNP for the gene tested; Best_eSNP.p, eQTL Pvalue for the best eSNP for the gene tested.

CHAPTER 5:PLATELET-RELATED VARIANTS IDENTIFIED BY EXOMECHIP META-ANALYSIS IN 157,293 INDIVIDUALS

Authors: Eicher JD*, Chami N*, Kacprowski T*, Nomura A*, Chen MH, Yanek LR, Tajuddin SM, Schick UM, Slater AJ, Pankratz N, Polfus L, Schurmann C, Giri A, Brody JA, Lange LA, Manichaikul A, Hill WD, Pazoki R, Elliot P, Evangelou E, Tzoulaki I, Gao H, Vergnaud AC, Mathias RA, Becker DM, Becker LC, Burt A, Crosslin DR, Lyytikäinen LP, Nikus K, Hernesniemi J, Kähönen M, Raitoharju E, Mononen N, Raitakari OT, Lehtimäki T, Cushman M, Zakai NA, Nickerson DA, Raffield LM, Quarells R, Willer CJ, Peloso GM, Abecasis GR, Liu DJ; Global Lipids Genetics Consortium, Deloukas P, Samani NJ, Schunkert H, Erdmann J; CARDIoGRAM Exome Consortium; Myocardial Infarction Genetics Consortium, Fornage M, Richard M, Tardif JC, Rioux JD, Dube MP, de Denus S, Lu Y, Bottinger EP, Loos RJ, Smith AV, Harris TB, Launer LJ, Gudnason V, Velez Edwards DR, Torstenson ES, Liu Y, Tracy RP, Rotter JI, Rich SS, Highland HM, Boerwinkle E, Li J, Lange E, Wilson JG, Mihailov E, Mägi R, Hirschhorn J, Metspalu A, Esko T, Vacchi-Suzzi C, Nalls MA, Zonderman AB, Evans MK, Engström G, Orho-Melander M, Melander O, O'Donoghue ML, Waterworth DM, Wallentin L, White HD, Floyd JS, Bartz TM, Rice KM, Psaty BM, Starr JM, Liewald DC, Hayward C, Deary IJ, Greinacher A, Völker U, Thiele T, Völzke H, van Rooij FJ, Uitterlinden AG, Franco OH, Dehghan A, Edwards TL, Ganesh SK, Kathiresan S, Faraday N*, Auer PL*, Reiner AP*, Lettre G*, Johnson AD*

*These authors contributed equally to this study

Reference: Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 individuals. John D. Eicher*, Nathalie Chami*, Tim Kacprowski*, Akihiro Nomura*, Ming-Huei Chen, Lisa R. *et al*, Am J Hum Genet. 2016 Jul 7; 99 (1):40-55

5.1.ABSTRACT

Platelet production, maintenance, and clearance are tightly controlled processes indicative of platelets' important roles in hemostasis and thrombosis. Platelets are common targets for primary and secondary prevention of several conditions and monitored clinically by complete blood counts, specifically with measurements of platelet count (PLT) and mean platelet volume (MPV). Identifying genetic contributors of PLT and MPV can reveal mechanistic insights into platelet biology and their role in disease. Therefore, we formed the Blood Cell Consortium (BCX) to perform a large-scale meta-analysis of exome chip association results for PLT and MPV in up to 157,293 and 57,617 multi-ethnic individuals, respectively. With increased sample size and use of the low-frequency/rare coding variant enriched exome chip platform, we sought to identify genetic variants associated with PLT and MPV. In addition to confirming 47 known PLT and 20 known MPV associations, we identified 32 novel PLT and 18 novel MPV associations across the allele frequency spectrum, including rare large effect (*FCER1A*), low-frequency (*IQGAP2*, *MAP1A*, *LY75*), and common (*ZMIZ2*, *SMG6*, *PEAR1*, *ARFGAP3*/*PACSIN2*) variants. Several variants associated with PLT/MPV (*PEAR1*, *MRVI1*, *PTGES3*) were also associated with platelet reactivity. In concurrent BCX analyses, there was overlap of platelet associated variants with red (*MAP1A*, *TMPRSS6*, *ZMIZ2*) and white blood cell (*PEAR1*, *ZMIZ2*, *LY75*) traits, suggesting common regulatory pathways with shared genetic architecture among these hematopoietic lineages. Our large-scale exome chip effort successfully identified numerous novel genes and variants associated with platelet traits and further indicate that several complex quantitative hematological, lipids, and cardiovascular traits share genetic factors.

5.2.INTRODUCTION

The number and size of circulating blood cells are determined by multiple genetic and environmental factors, and deviations from normal are common manifestations of human disease. The three major cell types—red blood cells (RBCs), white blood cells (WBCs), and platelets—have distinct biological roles, with platelets serving as important mediators of hemostasis and wound healing. Platelet count (PLT) and mean platelet volume (MPV), a measure of platelet size, are clinical blood tests that are used to screen for and diagnose disease. A number of well-described rare genetic disorders, including Bernard-Soulier Syndrome [286], Glanzmann's Thrombasthenia (MIM: 273800), and Wiskott-Aldrich Syndrome (MIM: 301000), as well as common conditions such as acute infection are characterized by abnormalities in the number, size, and/or reactivity of circulating blood platelets. MPV has also been reported to be an independent risk factor for myocardial infarction in population-based studies[88]. Accordingly, anti-platelet medications including aspirin, ADP/PAR receptor blockers, and GIIb/IIIa inhibitors that reduce platelet reactivity are common forms of primary and secondary prevention for several cardiovascular conditions including stroke and myocardial infarction [287; 288]. Investigating the biological mechanisms that govern platelet number (PLT) and size (MPV) can provide insights into platelet development and clearance, and has the potential to enhance our understanding of human diseases.

Genome-wide association studies (GWAS) have successfully identified numerous loci associated with PLT and MPV[122; 142-144; 149; 289-292]. To date, the largest GWAS of PLT (n=66,867) and MPV (n=30,194) identified 68 distinct associated loci[122]. Subsequent functional experiments of several identified genes, including *ARHGEF3* (MIM: 612115),

DNM3 (MIM: 611445), *JMJD1C* (MIM: 604503), and *TPM1* (MIM: 191010), demonstrated their roles in hematopoiesis and megakaryopoesis[293], as well as the potential of human genetic association methods to identify genetic factors that functionally contribute to platelet biology and dysfunction in disease.

Despite these successes, much of the heritability of these traits remains unexplained [294]. GWAS studies of PLT and MPV have largely focused on more common (minor allele frequency [MAF] > 0.05) genetic variation with many of the associated markers located in intronic or intergenic regions. The examination of rare (MAF < 0.01) and low-frequency (MAF: 0.01-0.05) genetic variation, particularly that in protein coding regions, has the potential to identify causal variants. Indeed, recent studies reaching sample sizes of 31,340 individuals have identified rare to low-frequency coding variants, associated with PLT in *MPL* (MIM: 159530), *CD36* (MIM: 173510), and *JAK2* (MIM: 147796), among others [135; 243]. Studies with larger sample size are needed to further characterize the contribution of rare and low-frequency genetic variation to PLT and MPV.

To conduct such a study of platelet related traits, we formed the Blood Cell Consortium (BCX) to perform a large scale meta-analysis of exome chip association results of blood cell traits. In this report, we describe results from a meta-analysis of exome chip association data in up to 157,293 and 57,617 multi-ethnic participants for PLT and MPV, respectively. The exome chip is a customized genotyping platform enriched for rare to low-frequency coding as well as common variants previously identified in GWAS of complex disorders and traits. With

increased sample size and use of the exome chip array, our goal was to identify novel rare, low-frequency, and common variant associations with PLT and MPV.

5.3.MATERIALS AND METHODS

Study participants

The Blood Cell Consortium (BCX) was formed to identify novel genetic variants associated with blood cell traits using the exome chip platform. As the BCX is interested in the genetics of common hematological measures, our collaborative group is divided into three main working groups: RBC, WBC, and platelet[244; 295]. For the platelet working group, our sample is comprised of up to 157,293 participants from 26 discovery and replication cohorts of five ancestries: European (EA), African-American (AA), Hispanic, East Asian, and South Asian. Detailed descriptions of the participating cohorts are provided in the Tables S1-S4. All participants provided informed consent, and all protocols were approved by the participating studies' respective institutional review boards. In the platelet working group, we analyzed two traits: PLT ($x10^9$/L of whole blood) and MPV (fL) (Table S3).

Genotyping and Quality Control

Each participating study used one of the following exome chip genotyping arrays: Illumina ExomeChip v1.0, Illumina ExomeChip v1.1_A, Illumina ExomeChip-12 v1.1, Illumina ExomeChip-12 v1.2, Affymetrix Axiom Biobank Plus GSKBB1, or Illumina HumanOmniExpressExome Chip (Table S2). Genotypes were called either 1) using a combination of the Illumina GenomeStudio and zCall software or 2) the exome chip joint calling plan developed by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Table S2) [246]. Standard quality control criteria were applied by each study. Exclusion criteria included: 1) sample call rates, 2) excess heterozygosity rate, 3) Hardy-Weinberg equilibrium p-values < $1x10^{-6}$, and 4) sex mismatch.

Additionally, ancestry was confirmed through principal components or multi-dimensional scaling analyses using linkage disequilibrium (LD) pruned markers ($r^2 < 0.2$) with MAF > 1%. Scatterplots anchored using the 1000 Genomes Project populations (http://www.1000genomes.org/) were visually inspected, and ancestry outliers excluded. We only included autosomal and X chromosome variants. All remaining variants (including monomorphic variants) were aligned to the forward strand and alleles checked to ensure that the correct reference allele was specified. We performed study specific level quality control on each trait association results using EasyQC [248]. We plotted variant allele frequencies from each study against ethnicity specific reference population allele frequencies to identify allele frequency deviations and presence of flipped alleles. Following all quality control procedures, each study generated an indexed variant call file (VCF) for subsequent analyses that was checked for allele alignment using the checkVCF package (https://github.com/zhanxw/checkVCF).

Association analysis

To assess the association between the blood cell traits and exome chip variants in the BCX, we considered blood cell traits measured in standard peripheral complete blood counts. When possible, we excluded individuals with blood cancer, leukemia, lymphoma, bone marrow transplant, congenital or hereditary anemia, HIV, end-stage kidney disease, dialysis, splenectomy, and cirrhosis, and those with extreme measurements of platelet traits. We also excluded individuals on erythropoietin treatment as well as those on chemotherapy. Additionally, we excluded women that were pregnant and individuals with acute medical illness at the time of complete blood count.

For platelet traits, we used raw values of PLT ($x10^9$/L) and MPV (fL). In each participating study, residuals for PLT and MPV were calculated from linear regression models adjusted for age, age^2, sex, study center (where applicable), and principal components. We then transformed residuals using the rank-based inverse normal transformation. To confirm proper trait transformation in each cohort, a scatter plot of the median standard error versus study specific sample size was visually inspected for deviations using EasyQC[248]. Autosomal and X chromosome variants were then tested for association with each blood cell trait using either RvTests (http://genome.sph.umich.edu/wiki/RvTests) or RAREMETALWORKER (http://genome.sph.umich.edu/wiki/RAREMETALWORKER). Within individual cohorts, we performed analyses in ancestry-stratified groups: EA, AA, Hispanic, East Asian, and South Asian. Both statistical packages generate single variant association score summary statistics, variance-covariance matrices containing LD relationships between variants within a 1MB window, and variant-specific parameters including minor allele frequency, chromosome position, strand, genotype call rate, and Hardy-Weinberg equilibrium p-values.

Discovery association meta-analysis

We performed ancestry-stratified (EA and AA) and combined all ancestry (All) meta-analyses of single variant association results using the Cochran-Mantel-Haenszel approach implemented in RareMETALS (http://genome.sph.umich.edu/wiki/RareMETALS) [249]In the multi-ancestry meta-analyses (All), we combined individuals of EA, AA, Hispanic, South Asian, and East Asian ancestries. We included variants in the meta-analysis if the genotype call rate was ≥ 95%, Hardy-Weinberg equilibrium p-values > $1x10^{-7}$, and allele frequency difference was <0.30 or <0.60 for ancestry-specific (EA and AA) or combined all ancestry (All) analyses, respectively [248]. Heterogeneity metrics (I^2 and heterogeneity p value) were

calculated using METAL[252]. Using single variant score statistics and variance-covariance matrices of LD estimates, we performed two types of gene-based tests: (1) variable threshold (VT) burden test with greatest power when all rare variants in a gene are associated with a trait [251] and (2) and sequence kernel association test (SKAT) [250] with greatest power when rare variants in a gene have opposing direction of effects. For all gene-based tests, we only considered missense, nonsense, and splice site SNVs with MAF ≤1%. Similar to the single variant meta-analyses, gene-based results were generated for each major ancestry group (EA and AA) and for the combined multi-ancestry (All) samples.

Conditional analysis

To identify independent signals, we performed step-wise conditional analyses conditioning on the most significant single variant in a 1MB window in RareMETALS. This procedure was repeated until there was no new signal identified in each region, defined as a p-value corrected for the number of markers tested in each ancestry group. For discovery and conditional single variant analyses, the corrected threshold was: AA $p<3.03x10^{-7}$, EA $p<2.59x10^{-7}$, and All $p<2.20x10^{-7}$. For gene-based tests, the significance threshold corrected for the number of genes tested: AA $p<2.91x10^{-6}$, EA $p<2.90x10^{-6}$, and All $p<2.94x10^{-6}$. In regions like chromosome 12q24 with known extended LD structure spanning more than 1MB, we performed a step-wise conditional analysis in GCTA to disentangle 7 independent PLT-associated SNVs (Table S9)[296], conditioning on the most significant variant in the region.

Replication meta-analysis

We attempted to replicate PLT and MPV associations with independent SNVs that reached significance levels in 6 independent cohorts (Figure 1, Table S4). Single variant association results of the 6 independent cohorts were combined in RareMETALS.

Contributing replication cohorts adhered to identical quality control and association analysis procedures described previously for the discovery phase. The results of discovery and replication phases were further combined using fixed effects inverse variance weighted meta-analysis in METAL[252].

Platelet Function Exome Chip Lookup

Two BCX cohorts, GeneSTAR and the Framingham Heart Study (FHS), measured platelet aggregation in a subset of genotyped participants. Platelet aggregation measures are described in detail elsewhere and briefly below (Table S18)[297]. Both studies isolated platelet-rich plasma from fasting blood samples and measured platelet aggregation after addition of agonists using a four-channel light transmission aggregometer (Bio/Data Corporation). FHS (Offspring Exam 5) tested aggregation for periods of 4 minutes after administration of ADP (0.05, 0.1, 0.5, 1.0, 3.0, 5.0, 10.0 and 15.0 μM) and 5 minutes after administration of epinephrine (0.01, 0.03, 0.05, 0.1, 0.5, 1.0, 3.0, 5.0 and 10.0 μM), as well as lag time(s) to aggregation with 190 μg/ml calf skin–derived type I collagen (Bio/Data Corporation). Threshold concentrations (EC_{50}) were determined as the minimal concentration of agonist required to produce a >50% aggregation. The maximal aggregation response (% aggregation) was also determined for each participant at each concentration tested. GeneSTAR recorded maximal aggregation (% aggregation) for periods of 5 minutes after ADP (2.0 and 10.0 μM) and 5 minutes after epinephrine administration (2.0 and 10.0 μM), as well as lag time(s) to aggregation with equine tendon–derived type I collagen (1, 2, 5 and 10 μg/ml). Exome chip genotyping, quality control, and association analyses adhered to methods described previously for PLT and MPV analysis. We queried independent SNVs associated with PLT (n=79) and/or

MPV (n=38) in these platelet aggregation association results and report platelet aggregation associations with $p<0.001$.

Further Variant Annotation

In addition to primary analyses completed in this investigation, we took advantage of several existing resources to annotate our associated SNVs. Associated variants were cross-referenced with CADD scores for exome chip [298]. The Global Lipids Genetics Consortium (GLGC) and the Myocardial Infarction and Coronary Heart Disease (MICHD) exome-chip studies have each performed independent exome chip analysis of lipids traits and coronary heart disease (CHD)[262; 299]. The CHD phenotype reflected a composite endpoint that included MI, CHD, coronary artery bypass graft, and hospitalized angina, among others[299]. Similar to the platelet aggregation lookups, we queried our list of PLT and/or MPV associated SNVs against their exome chip association results for lipids and CHD. We report lipid and CHD associations with $p<0.0001$. From a curated collection of over 100 separate expression quantitative trait loci (QTL) datasets, we conducted a more focused query of whether platelet loci were also associated with transcript expression in blood, arterial and adipose related tissues. A general overview of a subset of >50 eQTL studies has been published (Supplemental Data)[255]. Separately, we queried transcripts in loci corresponding to novel associated variants and/or marginally associated variants showing replication to assess their platelet expression levels using the largest platelet RNA-seq dataset to date (n=32 patients with MI)[300].

5.4.RESULTS

Discovery Meta-Analysis

In our discovery phase, we performed a meta-analysis of the associations of 246,925 single nucleotide variants (SNVs) with PLT and MPV in up to 131,857 and 41,529 individuals, respectively (**Figure 5.1**, Figures S1-S2, Tables S1-S4). After the initial meta-analyses, we ran conditional analyses to identify independent loci and found 79 independent PLT and 38 independent MPV SNVs (**Table 5.1**,**Table 5.2**, Tables S5-S8). One association, rs12692566 in *LY75-CD302*, with PLT in EA did not surpass the initial discovery statistical significance threshold but surpassed the threshold when conditioned on nearby rs78446341 ($p = 2.48 \times 10^{-7}$). There were no associations unique to the AA ancestry group, which had a limited sample size (Tables S10 and S11). Single variant meta-analysis results for each ancestry grouping that met our significance thresholds are available in the Supplement (Tables S10 and S11). Additionally, full discovery meta-analysis results are available online (Web Resources).

Of these independently associated single variants, 32 PLT and 18 MPV variants were in loci not previously reported (**Table 5.1** and **Table 5.2**). Of these 32 PLT loci, 4 had previously been identified as MPV loci (**Table 5.1**), and 10 of the 18 MPV loci had previously been identified with PLT (**Table 5.2**) [122; 143; 243].

Of the independent loci in our study, 23 SNVs showed association with both PLT and MPV (**Table 5.3**, **Figure 5.2**). All but one (rs6136489 intergenic to *SIRPA* (MIM: 602461) and *LOC727993*) had opposite directions of effects for PLT and MPV, indicative of the strong inverse statistical correlation between these traits.

Associated variants ranged in allele frequency and included rare, low-frequency, and common SNVs. Most of the previously unreported associations were with common variants (PLT n = 25, MPV n = 15), although associations with low-frequency (PLT n = 6, MPV n = 2) and rare (PLT n = 1, MPV n = 1) variants were observed. Rare (PLT n = 6, MPV n = 1) SNVs associated with PLT and MPV had larger effects compared to common and low-frequency SNVs (Tables 1, 2, and S5–S8). A large majority of associated SNVs did not exhibit heterogeneous effects; however, one previously unreported association with *MRVI1* and a few known associated loci (e.g., *MYL2*/*SH2B3*/*ATXN2*, *ARHGEF3*, *WDR66*/*HPD*, and *JAK2*) did show moderate to substantial heterogeneity across discovery studies (Table S23). Gene-based tests of missense, nonsense, and splice-site rare variants that found significant results largely reflected rare and low-frequency single variant results, with variants in *TUBB1* (MIM: 612901), *JAK2*, *LY75* (MIM: 604524), *IQGAP2* (MIM: 605401), and *FCER1A* (MIM: 147140) showing associations (Tables S12 and S13).

Figure 5.1. Study Design and Flow

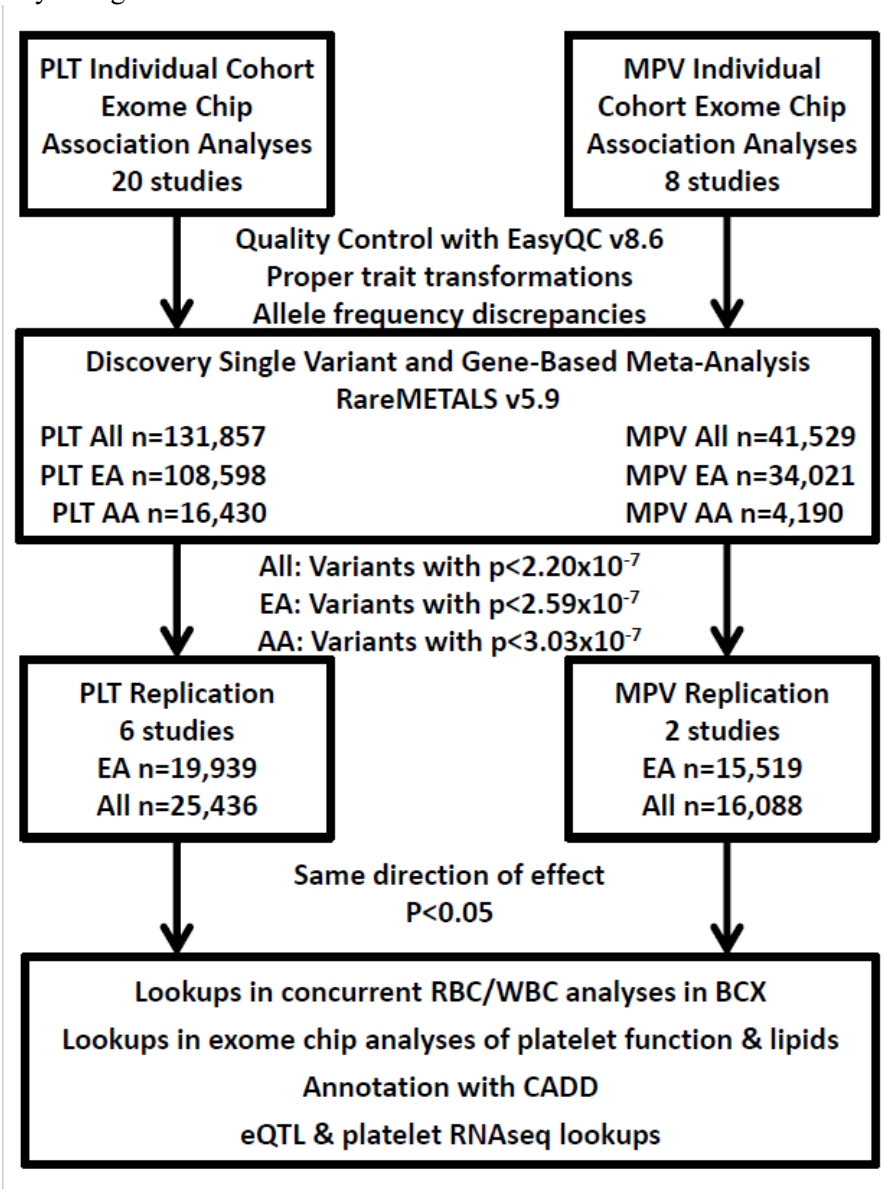


Figure 5.1. Study Design and Flow Individual study level association analyses were performed using RareMetalWorker or RVTests. To perform quality control of individual study association results, we used EasyQC v8.6 to ensure proper trait transformations, to assess allele frequency discrepancies, and to evaluate other metrics. We then combined results in meta-analysis with RareMETALS v5.9 in three groups: African ancestry (AA), European ancestry (EA), and combined all five (AA, EA, Hispanic-Latino, East Asian, South Asian) ancestries (All). Independent variants identified by conditional analysis in RareMETALS with a p-value less than the threshold corrected for multiple testing (All: $p<2.20\times10^{-7}$, EA: $p<2.59\times10^{-7}$, AA: $p<3.03\times10^{-7}$) were carried forward for replication. Markers showed replication if they had $p<0.05$ in the same direction of effect in the replication analyses. Associated markers were further annotated using various resources: (1) concurrent BCX exome chip analyses of RBC and WBC traits, (2) on-going exome chip analyses of platelet aggregation, quantitative lipids, and coronary heart disease (CHD) traits, (3) severity prediction by CADD, (4) an internal database of reported eQTL results, and (5) platelet RNA-seq data.

Table 5.1. Novel associations (n=32) with PLT.

							Discovery				Replication			Combined	Discovery				Replication			Combined
Marker	rsID	Ref/Alt	Function	AACh.	Gene	CADD[$]	n	EAF	Beta	P-value	n	Beta	P-value	P-value	n	EAF	Beta	P-value	n	Beta	P-value	P-value
1:25674785	rs3091242	C/T	intron		TMEM50A	12.7	100605	0.54	-0.026	9.68E-08	19939	-0.017	0.124	3.85E-08	122438	0.50	-0.02	1.03E-05	25436	-0.0084	0.39	1.24E-05
1:156869047	rs12566888	G/T	intron		PEAR1	1.5	108598	0.094	0.040	1.42E-07	19939	0.061	0.00126	1.17E-09	131857	0.16	0.034	2.09E-08	25436	0.047	0.000431	5.71E-11
1:159275786	rs200731779	C/G	missense	L114V	FCER1A	23.5	101368	1.5E-05	-2.96	2.48E-07	19939	NA	NA	2.48E-07	124627	1.2E-05	-2.96	2.48E-07	25436	NA	NA	2.48E-07
2:113841030	rs6734238	A/G	intergenic		IL1F10/IL1RN	1.7	86947	0.41	0.022	9.55E-06	19939	0.0075	0.487	1.64E-05	106744	0.41	0.026	7.19E-09	25436	0.015	0.117	3.77E-09
2:160676427	rs12692566	C/A	missense	K1321N	LY75-CD302	15.8	108598	0.82	-0.029	9.19E-07	19939	-0.042	0.0025	1.23E-08	131857	0.83	-0.026	2.27E-06	25436	-0.05	7.84E-05	3.65E-09
2:160690656	rs78446341	G/A	missense	P1247L	LY75-CD302	24.1	108598	0.02	0.092	4.16E-09	19939	0.14	5.01E-05	1.98E-12	131857	0.018	0.094	3.06E-10	25436	0.13	9.23E-05	1.97E-13
3:124377326	rs56106611*	T/G	missense	S2030A	KALRN	19.6	100605	0.012	0.11	3.51E-08	19939	0.11	0.00714	8.51E-10	123864	0.01	0.11	8.59E-08	25436	0.11	0.00737	2.14E-09
3:185529080	rs1470579	A/C	intron		IGF2BP2	6.3	108598	0.32	-0.028	1.08E-07	19939	-0.0073	0.562	2.82E-07	131857	0.38	-0.023	6.07E-07	25436	-0.012	0.272	5.15E-07
4:100045616	rs1126673	C/T	ncRNA	V393I	LOC100507053	6.1	108598	0.69	0.026	6.38E-08	19939	0.019	0.0963	1.81E-08	131857	0.71	0.025	1.87E-08	25436	0.014	0.168	1.12E-08
5:158603571	rs1473247*	T/C	intron		RNF145	4.5	108598	0.27	-0.029	3.01E-08	19939	-0.022	0.0832	7.28E-09	131857	0.32	-0.026	1.32E-08	25436	-0.025	0.0185	7.66E-10
6:31380529	rs2256183	A/G	intron		MICA	5.6	108598	0.56	0.03	6.78E-07	19939	-0.022	0.104	2.60E-06	131857	0.59	0.028	2.13E-07	20552	0.011	0.389	3.20E-07
7:44808091	rs1050331	T/G	3'UTR		ZMIZ2		100605	0.47	0.037	1.32E-15	19939	0.036	0.00058	3.28E-18	122438	0.48	0.035	3.09E-17	25436	0.031	0.00088	1.26E-19
9:100696203	rs755109	T/C	intron		HEMGN	3.3	108598	0.37	0.028	2.87E-09	19939	0.039	0.000684	1.17E-11	131857	0.34	0.028	9.03E-11	25436	0.044	2.18E-05	2.59E-14
10:94839642	rs2068888	G/A	nearGene-3		EXOC6	5.7	108598	0.45	-0.023	2.81E-07	19939	-0.012	0.266	2.47E-07	131857	0.44	-0.022	1.19E-07	25436	-0.012	0.212	8.61E-08
11:8751889	rs3794153	C/G	missense	K316N	ST5	23.7	107555	0.45	-0.027	7.28E-09	19939	-0.026	0.0153	3.57E-10	125167	0.40	-0.027	2.19E-09	25436	-0.023	0.0247	1.74E-10
11:61609750	rs174583	C/T	intron		FADS2	13.8	100605	0.34	0.031	8.79E-09	19939	0.048	0.000122	1.03E-11	121747	0.34	0.028	4.72E-09	25436	0.042	0.00011	4.42E-12
11:119060963	rs45535039	T/C	3'UTR		CCDC153		64720	0.28	0.04	4.02E-10	1933	0.071	0.0531	8.48E-11	81768	0.28	0.04	2.5E-12	2546	0.056	0.0856	6.25E-13
12:6502742	rs11616188	G/A	nearGene3		LTBR	3.7	108598	0.42	-0.025	1.26E-08	19939	-0.031	0.00359	1.81E-10	131857	0.37	-0.025	7.57E-09	25436	-0.033	0.00107	4.20E-11
12:54687232	rs10506328*	A/C	intron		NFE2	9.4	86947	0.64	0.033	5.63E-11	19939	0.06	5.88E-08	2.01E-16	110206	0.69	0.038	3.79E-15	25436	0.059	2.33E-08	2.73E-21
12:89745477	rs2279574	C/A	missense	V114L	DUSP6	23.5	108598	0.54	-0.023	2.47E-07	19939	-0.0082	0.442	4.28E-07	131857	0.50	-0.021	1.57E-07	25436	-0.006	0.531	4.04E-07
12:111785515	rs61745424	G/A	missense	E1221K	CUX2	2.3	108598	0.025	-0.064	2.36E-06	18923	-0.085	0.00679	6.49E-08	131857	0.023	-0.068	1.37E-07	25436	-0.073	0.0143	6.30E-09
14:53657823	rs2784521	A/G	nearGene-5		DDHD1		108598	0.83	0.025	1.62E-05	19939	0.0096	0.486	2.24E-05	131857	0.76	0.028	2.92E-08	25436	0.01	0.363	5.56E-08
15:43820717	rs55707100	C/T	missense	P2349L	MAP1A	23.4	108598	0.032	0.095	7.03E-14	19939	0.073	0.0387	9.53E-15	131857	0.028	0.092	6.85E-14	25436	0.082	0.0162	3.77E-15
17:2143460	rs10852932	G/T	intron		SMG6	0.8	108598	0.36	-0.024	1.82E-06	19939	-0.042	0.000893	1.42E-08	131857	0.39	-0.025	4.79E-08	25436	-0.036	0.000699	2.15E-10
17:42463054	rs76066357	G/C	missense	L147V	ITGA2B	6.6	78524	0.014	-0.17	6.92E-16	19939	-0.19	2.88E-05	1.05E-19	96684	0.013	-0.16	1.92E-15	25436	-0.18	6.00E-05	5.78E-19
17:64210580	rs1801689	A/C	missense	C325R	APOH	23.4	108598	0.036	0.083	6.34E-12	19939	0.13	2.44E-05	1.82E-15	131857	0.032	0.090	8.64E-15	25436	0.12	2.03E-05	1.57E-18
19:38912764	rs892055	A/G	missense	I18T	RASGRP4	7.7	108598	0.34	0.029	5.3E-10	19939	0.018	0.0987	2.01E-10	131857	0.38	0.025	3.49E-09	25436	0.017	0.0813	9.96E-10
19:51727962	rs3865444	C/A	5'UTR		CD33	3.8	86947	0.32	-0.026	1.11E-06	19939	-0.034	0.00252	1.27E-08	106744	0.29	-0.026	2.1E-07	25436	-0.032	0.00303	2.59E-09
20:1923734	rs6136489*	T/G	intergenic		SIRPA LOC727993	3.0	108598	0.34	-0.033	8.69E-13	19939	-0.028	0.0124	4.00E-14	131857	0.39	-0.030	1.8E-12	25436	-0.024	0.013	8.78E-14
22:37462936	rs855791	A/G	missense	V605D	TMPRSS6	23.6	108598	0.56	-0.031	3.96E-11	19939	-0.017	0.13	2.34E-11	131857	0.60	-0.029	2.34E-11	25436	-0.022	0.0352	2.97E-12
22:43206950	rs1018448	A/C	missense	S355R	ARFGAP3	22.4	108598	0.54	-0.028	4.02E-10	19939	-0.0053	0.618	2.62E-09	131857	0.59	-0.025	1.55E-09	25436	-0.0065	0.515	6.13E-09
22:44324727	rs738409	C/G	missense	I148M	PNPLA3	3.4	108598	0.23	-0.042	1.49E-14	19939	-0.042	0.00175	1.03E-16	131857	0.22	-0.044	1.33E-18	25436	-0.038	0.00161	9.73E-21

Table 5.1. We show variants in previously unreported loci (n = 32) and retained after conditional analyses in European ancestry (EA) ($p < 2.59 \times 10^{-7}$) and all ancestry (All) ($p < 2.20 \times 10^{-7}$) analyses. Associations in African ancestry (AA) had previously been reported in the literature (Table S10). Asterisks (*) indicate variants (20/32) showing evidence of replication ($p < 0.05$, same direction of effect). If multiple genes/transcripts were annotated to a variant, the transcript most expressed in Eicher et al.[300] (Table S22) was selected. Full results and annotations are available in Table S5. Abbreviations are as follows: PLT, platelet count; MPV, mean platelet volume; REF, reference allele; ALT, alternate allele; EAF, effect allele frequency.
*Previous association with MPV, [$]Scaled CADD score. **Abbreviations:** PLT, platelet count; MPV, mean platelet volume; REF, reference allele; ALT, alternate allele; AACh, amino acid change; EAF, effect allele frequency

Table 5.2. Novel associations (n=18) with MPV.

							European Ancestry (EA)								Combined All Ancestry (All)							
							Discovery				Replication			Combined	Discovery				Replication			Combined
Marker	rsID	Ref/Alt	Function	AAChange	Gene	CADD[$]	n	EAF	Beta	P-value	n	Beta	P-value	P-value	n	EAF	Beta	P-value	n	Beta	P-value	P-value
1:25889632	**rs6687605**	T/C	missense	Ser202Thr	LDLRAP1	0.009	34021	0.53	0.046	8.27E-12	15519	0.025	0.0374	1.80E-09	41529	0.51	0.046	9.92E-11	16088	0.024	0.0358	3.80E-11
1:247719769	**rs56043070***	G/A	Splice donor		GCSAML	18	34021	0.069	0.094	1.3E-09	15519	0.19	4.48E-16	1.12E-21	41529	0.064	0.092	2.25E-10	16088	0.19	3.66E-16	2.42E-22
1:248039294	rs1339847*	G/A	missense	Val322Ile	TRIM58	22.7	30569	0.10	-0.10	1.47E-13	15519	-0.037	0.0544	9.31E-13	37415	0.10	-0.11	2.18E-17	16088	-0.032	0.0977	1.06E-15
5:75960968	rs34968964*	G/C	missense	Glu436Gln	IQGAP2	22	34021	0.0049	0.32	7.65E-09	15519	0.12	0.0918	1.99E-08	41529	0.004	0.32	2.11E-09	16088	0.11	0.106	8.18E-09
5:75964507	**rs34950321***	C/T	missense	Thr447Ile	IQGAP2	32	34021	0.018	0.18	7.8E-10	15519	0.14	0.00149	6.03E-12	41529	0.016	0.17	2.61E-09	16088	0.14	0.00159	1.86E-11
5:75996909	**rs34592828***	G/A	missense	Arg1012Gln	IQGAP2	26.5	34021	0.037	0.22	1.72E-27	15519	0.16	2.73E-09	1.61E-34	41529	0.032	0.23	1.68E-31	16088	0.16	2.95E-09	2.98E-38
6:25605091	rs1012899*	G/A	missense	Gly1182Ser	LRRC16A	9.9	34021	0.77	0.051	1.4E-07	15519	0.012	0.417	1.24E-06	41529	0.77	0.042	1.32E-06	16088	0.016	0.273	2.50E-06
6:36393816	**rs664370**	A/G	missense	Val32Ala	PXT1	10.2	34021	0.30	-0.034	8.03E-05	15519	-0.025	0.0561	1.39E-05	41529	0.35	-0.042	5.77E-08	16088	-0.028	0.0278	7.23E-09
8:106593207	rs2343596*	C/A	intron		ZFPM2	2.5	34021	0.31	0.062	2.02E-13	15519	0.012	0.357	3.32E-11	41529	0.38	0.052	1.59E-11	16088	0.012	0.339	4.35E-10
8:145001031	rs55895668*	T/C	missense	His1327Arg	PLEC	21.9	34021	0.43	-0.042	5.94E-07	15519	-0.013	0.350	2.19E-06	41529	0.47	-0.041	1.23E-07	16088	-0.011	0.409	5.97E-07
11:10673739	**rs4909945**	T/C	missense	Ile11Val	MRVI1	19.1	34021	0.68	-0.048	1.25E-08	15519	-0.035	0.00841	5.19E-10	41529	0.71	-0.041	3.96E-07	16088	-0.035	0.00742	1.06E-08
14:55611839	rs11125	A/T	missense	Gln201His	LGALS3	13.9	30569	0.078	-0.091	1.55E-08	15519	-0.037	0.117	2.76E-08	38077	0.07	-0.09	4.22E-09	16088	-0.037	0.117	7.21E-09
15:65157482	**rs2010875***	C/T	missense	Pro290Ser	PLEKHO2	0.3	21732	0.14	-0.076	1.33E-07	14581	-0.042	0.0162	2.10E-08	28290	0.15	-0.063	3.01E-07	14581	-0.042	0.0162	2.43E-08
17:33884804	**rs10512472***	T/C	missense	Gln93Arg	SLFN14	22.2	34021	0.18	-0.059	1.37E-08	15519	-0.059	0.000196	1.12E-11	41529	0.18	-0.058	3.15E-10	16088	-0.059	0.00012	1.67E-13
19:45162189	**rs35385129**	C/A	missense	Arg391Ser	PVR	9.3	34021	0.16	-0.058	6.24E-08	15519	-0.044	0.00736	2.01E-09	41529	0.15	-0.055	3.00E-08	16088	-0.043	0.00713	8.79E-10
20:1546911	rs2243603	C/G	missense	Ala252Pro	SIRPB1	0.4	34021	0.77	0.044	5.89E-06	938	0.077	0.167	2.62E-06	41529	0.79	0.049	4.58E-08	1507	0.088	0.0778	1.25E-08
22:43206950	**rs1018448**	A/C	missense	Ser355Arg	ARFGAP3	22.4	34021	0.55	0.056	1.13E-12	15519	0.051	1.78E-05	1.04E-16	41529	0.60	0.055	1.52E-13	16088	0.05	2.16E-05	1.68E-17
X:57622607	**rs1997715**	G/A	3'UTR		ZXDB		34021	0.26	0.048	1.93E-09	938	0.084	0.0583	4.26E-10	41529	0.35	0.04	4.58E-08	1507	0.08	0.0399	8.88E-09

Table 5.2. We show variants in novel MPV loci and retained after conditional analyses in European Ancestry (EA) ($p<2.59x10-7$) and All Ancestry (All) ($p<2.20x10-7$) analyses. There were no novel associations in African Ancestry (AA). Chromosome positions are human genome build hg19. **Bolded** variants (11/18) showed evidence of replication ($p<0.05$, same direction of effect). If multiple genes/transcripts were annotated to a variant, the transcript more expressed in Eicher et al. 2015 (Table S20) was selected. *Previous association with PLT, [$]Scaled CADD score. **Abbreviations:** MPV, mean platelet volume; PLT, platelet count; REF, reference allele; ALT, alternate allele; AAChange, amino acid change; EAF, effect allele frequency.

Replication and Marginally Associated Variants

We attempted to replicate our associations in six independent cohorts (PLT n = 25,436, MPV n = 16,088) (**Figure 5.1**, Table S4). Of the loci not previously associated, 20/32 PLT and 11/18 MPV variants showed evidence of replication with $p < 0.05$ and the same direction of effect (Table **5.1** and Table **5.2**). In addition to the significant SNVs in our discovery analysis, we carried forward 13 PLT and 10 MPV sub-threshold variants that approached discovery significance thresholds with p values ranging from 2.47×10^{-7} to 1.99×10^{-6} (Tables S14 and S15). Of these, 7/13 PLT and 4/10 MPV showed associations in same direction of effect with $p < 0.05$ and surpassed significance thresholds when discovery and replication results were combined (Tables S14 and S15).

Intersection with Other Cardiovascular and Blood Traits

The BCX also completed analyses of RBC and WBC traits, so we cross-referenced our list of PLT- and MPV-associated SNVs with the results of the other blood cell traits[244; 295]. Of our replicated platelet loci previously unreported in the literature, six SNVs in *TMPRSS6* (MIM: 609862), *MAP1A* (MIM: 600178), *PNPLA3* (MIM: 609567), *FADS2* (MIM: 606149), *TMEM50A* (MIM: 605348), and *ZMIZ2* (MIM: 611196) showed association with RBC-related traits ($p < 0.0001$) (Table **5.4**). Similarly, five replicated platelet SNVs previously unreported in the literature in *PEAR1* (MIM: 610278), *CD33* (MIM: 159590), *SIRPA*, *ZMIZ2*, and *LY75* showed association with WBC-related traits ($p < 0.0001$) (Table **5.4**). To explore possible shared genetic associations of platelet size/number with platelet reactivity, we examined the association of PLT/MPV-associated SNVs with platelet reactivity to collagen, epinephrine, and ADP in GeneSTAR and FHS. Eight SNVs associated with PLT and/or MPV were also associated with platelet reactivity ($p < 0.001$) (Table **5.5**, Tables S16-S17). The most strongly

associated SNVs were located in genes implicated with platelet reactivity in prior GWASs, including *PEAR1*, *MRVI1* (MIM: 604673), *JMJD1C*, and *PIK3CG* (MIM: 601232)[297]. However, we did observe new suggestive relationships between platelet reactivity and SNVs in *PTGES* (MIM: 607061), *LINC00523*, and *RASGRP4* (MIM: 607320) (Table **5.5**).

In addition to examining possibly shared genetic associations with blood cell-specific traits, we queried our list of associated platelet SNVs against independent Exomechip genotyping efforts in lipids and CHD by the GLGC, CARDIoGRAM Exome Consortium, and Myocardial Infarction Genetics Consortium Exomechip studies[301,299]. Numerous platelet-associated SNVs (n = 37), including those in *GCKR* (MIM: 600842), *FADS1* (MIM: 606148), *FADS2*, *MAP1A*, *APOH* (MIM: 138700), and *JMJD1C*, showed association with one or more lipids traits ($p < 0.0001$) (Table S20). Far fewer (n = 4; *MYL2* [MIM: 160781], *SH2B3* [MIM: 605093], *BRAP* [MIM: 604986], *APOH*) showed association with CHD ($p < 0.0001$) (Table S20).

Annotation of Associated Variants

We used various resources to annotate our platelet-associated variants. First, we used CADD to predict the putative functional severity of associated variants[298]. As expected, rare and low-frequency coding SNVs were predicted to be more severe than common, non-coding variation (Table **5.1**, Table **5.2**, Tables S5-S6). To assess potential impact on gene expression, we queried our list of platelet associated SNVs against a collection of results from existing eQTL datasets[255]. Many (n=67) platelet-associated SNVs were also associated with gene expression in blood, arterial, or adipose tissues (Table S21). These included the reported *trans*-eQTL rs12485738 in *ARHGEF3* with several platelet-related transcript targets (e.g., GP1BA, GP6, ITGA2B, MPL, TUBB1, and VWF)[302], as well as eQTLs in newly identified PLT/MPV

loci (e.g., rs1018448 with ARFGAP3/PACSIN2, rs1050331 with ZMIZ2, and rs174546 with FADS1/FADS2/TMEM258 expression). Using platelet RNA-seq data from 32 subjects with MI, we found that almost all of the genes closest to our previously unreported associated SNVs or marginal SNVs with evidence of replication were expressed in platelets indicating the feasibility of potential functional roles in the relevant target cell type (Table S22).

Figure 5.2. Shared PLT and MPV Genetic Associations.

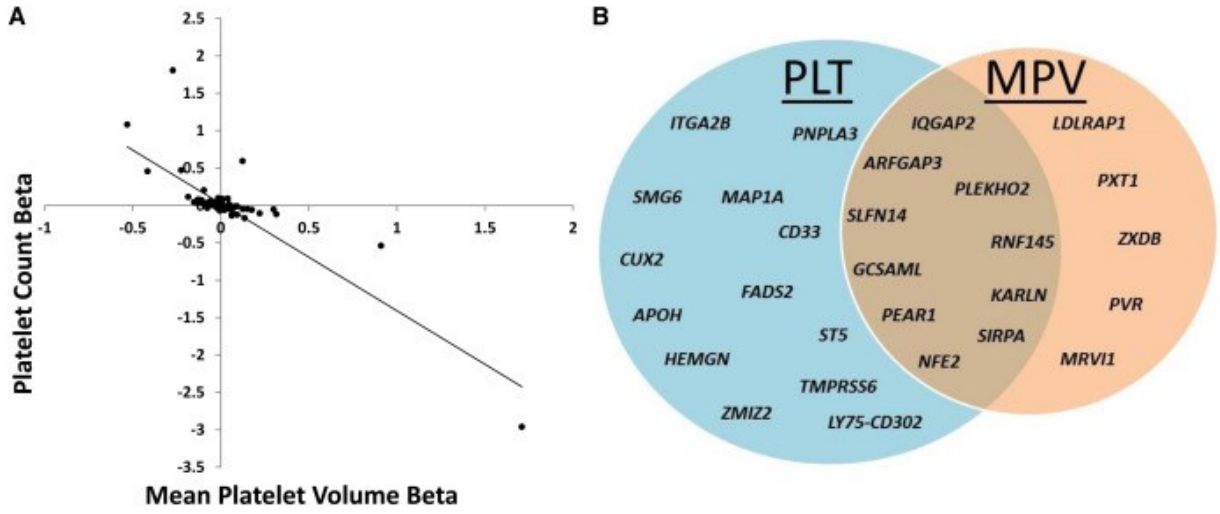


Figure 5.2.
(A) Comparing PLT and MPV effects sizes (r = −0.84) in European ancestry (EA) analyses of all identified SNVs identified (n = 124). Examined SNPs include all those from **Table 5.1**, **Table 5.2**, S5–S9, S14, and S15.
(B) 56 independent SNVs showed association to PLT only, and 15 independent SNVs were associated with MPV only. 23 independent SNVs were associated with both PLT and MPV. Named genes indicate that the association was not previously reported in the literature.

Table 5.3. Variants associated with both PLT and MPV.

rsID	Gene	PLT	MPV
rs12566888	PEAR1	↑	↓
rs1668873	TMCC2	↑	↓
rs56043070	GCSAML	↓	↑
rs12485738	ARHGEF3	↑	↓
rs56106611	KALRN	↑	↓
rs34592828	IQGAP2	↓	↑
rs1012899	LRRC16A	↓	↑
rs342293	PIK3CG	↓	↑
rs2343596	ZFPM2	↓	↑
rs10761731	JMJD1C	↑	↓
rs11602954	BET1L	↑	↓
rs10506328	NFE2	↑	↓
rs2958154	PTGES3	↓	↑
rs7961894	WDR66	↓	↑
rs1465788	ZFP36L1	↑	↓
rs2297067	EXOC3L4	↑	↓
rs2138852	TAOK1	↓	↑
rs10512472	SLFN14	↑	↓
rs11082304	CABLES1	↓	↑
rs6136489*	SIRPA/LOC727993	↓	↓
rs41303899	TUBB1	↓	↑
rs6070697	TUBB1	↑	↓
rs1018448	ARFGAP3	↓	↑

Table 5.3. All variants listed here showed association with both PLT and MPV in the opposite direction of effect as indicated by the arrows, except for rs6136489 (denoted by asterisk) which showed association with decreased PLT and decreased MPV. **Abbreviations:** PLT, platelet count; MPV, mean platelet volume.

Table 5.4. Intersection of platelet associated variants with red blood cell (RBC) and white blood cell (WBC) traits (p<0.0001).

SNP	MarkerName	Gene	PLT	Trait	Other Blood Cell
rs855791	22:37462936	TMPRSS6	↓	MCH, MCV, HGB MCHC, HCT	↑
rs855791	22:37462936	TMPRSS6	↓	RDW	↓
rs55707100	15:43820717	MAP1A	↑	HGB, MCH, HCT, MCHC	↓
rs174583	11:61609750	FADS2	↑	RDW	↓
rs174583	11:61609750	FADS2	↑	HGB, RBC, HCT, MCHC	↑
rs738409	22:44324727	PNPLA3	↓	HCT, HGB	↑
rs3091242	1:25674785	TMEM50A	↓	RDW	↑
rs1050331	7:44808091	ZMIZ2	↑	MCH, MCV	↓
rs1050331	7:44808091	ZMIZ2	↑	WBC	↑
rs6734238a	2:113841030	IL1F10/IL1RN	↑	MCH	↓
rs6734238a	2:113841030	IL1F10/IL1RN	↑	WBC, NEU	↑
rs12566888	1:156869047	PEAR1	↑	WBC, NEU, MON	↓
rs3865444	19:51727962	CD33	↓	WBC	↓
rs6136489	20:1923734	SIRPA/LOC727993	↓	WBC, LYM	↓
rs2256183a	6:31380529	MICA	↑	BAS	↑
rs12692566	2:160676427	LY75-CD302	↓	WBC	↓

Table 5.4. We cross-referenced novel variants associated with platelet count (PLT) and/or mean platelet volume (MPV) in RBC and WBC association analyses in the Blood Cell Consortium (BCX). Here, we show RBC/WBC associated platelet variants with p<0.0001. Full details of RBC/WBC associations are shown in Table S16 and Table S17. Arrows denote direction of effect for the platelet and other blood cell trait(s). **Abbreviations:** BCX, Blood Cell Consortium; RBC, red blood cell; WBC, white blood cell; PLT, platelet count; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; HGB, hemoglobin; MCHC, mean corpuscular hemoglobin concentration; HCT, hematocrit; RDW, red blood cell distribution width; PLT, platelet count; NEU, neutrophil; MON, monocyte; LYM, lymphocyte; BAS, basophil. a: Marker not replicated in platelet analyses

Table 5.5. Overlap of associations of platelet count (PLT) and mean platelet volume (MPV) variants with platelet reactivity ($p<0.001$).

rsID	Gene	PLT	MPV	Agonist(s)*	Direction of Effects**
rs12566886	PEAR1	↑	↓	Epi, ADP, Collagen	↓↓↓
rs10761731	JMJD1C	↑	↓	Epi, ADP	↑↑
rs12355784	JMJD1C	↑	Ns	Epi	↑
rs342293	PIK3CG	↓	↑	Epi	↓
rs4909945	MRVI1	ns	↓	Epi, ADP	↓↓
rs2958154	PTGES3	↓	↑	Collagen	↑
rs12883126	LINC00523	↑	Ns	Epi	↑
rs892055	RASGRP4	↑	Ns	Epi	↓

Table 5.5. Variants were examined using platelet reactivity phenotypes (Table S16) in GeneSTAR and the Framingham Heart Study (FHS). Arrows denote direction of effect for PLT, MPV, and platelet reactivity. Multiple arrows refer to direction for respective agonist for platelet reactivity. Detailed association results for platelet reactivity are given in Table S17. *Platelet reactivity associations with $p<0.001$. **Collagen measurements reflect lag time to aggregation, so direction of effect has been flipped to denote a negative direction of effect as less reactive and positive direction of effect as more reactive. Abbreviations are as follows: PLT, platelet count; MPV, mean platelet volume; ns, not significant ($p>0.05$), Epi, epinephrine.

5.5.DISCUSSION

Here, we present a large-scale meta-analysis of Exomechip association data with two clinical platelet measurements, PLT and MPV. By combining Exomechip association results in 157,293 and 57,617 participants, respectively, we detected numerous associations with rare, low-frequency, and common variants. There was substantial overlap of our platelet associations with concurrent Exomechip association findings for RBC and WBC traits, indicating shared genetic influence on regulatory and functional mechanisms among the three different blood cell lineages[244; 295]. More surprisingly, we observed shared associations of platelet and lipids loci. The identification of shared blood cell and lipids associations as well as identifying genes with entirely new associations reveals candidates for further examination in order to elucidate the mechanisms underlying platelet development and function.

Using Exomechip to Identify Previously Unreported Genetic Associations

Using the Exomechip that has an emphasis on rare and infrequent coding variation, we found associations with variants that ranged from common to rare in allele frequency. We attempted to replicate independent associations, although our replication cohorts were underpowered to associations of rare variants. To inform our replication criteria, we conducted a power analysis by using a sample size of 20,000 and considering multiple combinations of allele frequencies and effect sizes. Based on allele frequency and effect size, our most difficult to replicate variant was rs56106611 (MAF = 0.012, Beta = 0.11). However, we still had approximately 80% power to detect this association in the replication stage. Despite this, replication of extremely rare variants remains a challenge. For example, there were

associations with rare coding variants with large effect sizes in *FCER1A*, *MPL*, *JAK2*, *SH2B3*, *TUBB1*, and *IQGAP2*[135; 243]. The overall effect size of these rare variants must be validated in independent studies. The PLT-associated and predicted deleterious variant rs200731779 in *FCER1A* (p.Leu114Val) had a large effect ($\beta = -2.96$) in discovery analyses, but could not be replicated in available samples due to its extremely rare allele frequency (MAF = 1.48×10^{-5} in EA). The affected amino acid is extracellularly positioned near the interface of two Ig-like domains, an area of the protein critical for FC-IgE interaction as shown through its crystal structure, biochemical data, and mutagenesis studies[303-306]. Other variants in *FCER1A*, a subunit of the allergy response IgE receptor and basophil differentiation factor, have previously been associated with IgE levels and monocyte counts[146; 307]. Increased platelet activation has been postulated to contribute to or be a consequence of allergic and inflammatory responses[308]. Our association of rare deleterious variation in *FCER1A* to reduced PLT provides a further link between platelet biology and allergy response.

Although SNVs in *IQGAP2* have previously been associated with PLT, we detected independent *IQGAP2* low-frequency and rare missense variants associated with increased MPV (Table **5.2**, Figures S3 and S4)[122; 243] Located proximal to thrombin receptor *F2R* (MIM: 187930), IQGAP2 functions in the cytoskeletal dynamics in response to thrombin-induced platelet aggregation[286]. We did not observe *IQGAP2* associations with platelet aggregation, which may be due to the rare/low-frequency nature of the SNVs and the absence of thrombin-induced aggregation data in the available cohorts. Nonetheless, the associations of rare and low-frequency variants in *IQGAP2* further strengthen its contribution to platelet biology. In addition to *IQGAP2*, we observed other low-frequency associations, including

nonsynonymous coding variants in *ITGA2B* (MIM: 607759), *LY75*, *MAP1A*, and *APOH*. The SNV rs76066357 in *ITGA2B*, a gene implicated in Glanzmann's thrombasthenia (MIM: 273800), was associated with decreased PLT (Table **5.1**). Moreover, *ITGA2B* codes for the platelet glycoprotein alpha-IIb, which is part of the target receptor of GIIb/IIIa inhibitors (e.g., eptifibatide and abciximab) used in the acute management of acute coronary syndromes. Although ClinVar lists rs76066357 as pathogenic (ID: 216944) with limited evidence, rs76066357 is a non-rare, predicted benign variant that contributes to population variability in PLT in our study as opposed to a severe Mendelian disorder of platelet reactivity.[309] Previous studies do suggest a potential role for variants in *ITGA2B* and *ITGB3* (MIM: 173470) leading to thrombocytopenia as well as abnormalities in platelet reactivity.[310]

In addition to rare and low-frequency variant associations, we detected previously unreported associations for PLT and MPV at 25 and 15 common loci, respectively. For example, a common missense SNV rs1018489 in *ARFGAP3* (MIM: 612439) showed association with decreased PLT and increased MPV. This variant is an eQTL for both *ARFGAP3* and neighboring gene *PACSIN2* (MIM: 604960) in blood tissues (Table S21, Figures S5 and S6). Although the possible role of the androgen receptor (AR) gene target and cellular secretory factor *ARFGAP3* is unknown in platelets,[311-313] *PACSIN2* functions in the formation of the megakaryocyte demarcation membrane system during platelet production through interactions with FlnA[314]. Genetic variation that influences *PACSIN2* expression may hinder the formation of the megakaryocyte demarcation membrane system and lead to the production of fewer but larger and potentially more reactive platelets. We also observed several other novel associations with common variants, including those in *SMG6* (MIM:

610963), a mediator of embryonic stem cell differentiation through nonsense-mediated decay, and *LY75*, an endocytotic immunity-related receptor highly expressed on dendritic cells where it is involved in recognition of apoptotic and necrotic cells. [315-317]

Overlap with Other Platelet and Blood Cell Traits

There was substantial overlap of variants associated with both PLT and MPV (n = 23) as well as a strong negative correlation in effect sizes, consistent with the documented negative correlation between the two traits in population studies (**Figure 5.2**).[318]Only rs6136489, a reported eQTL for SIRPA, showed the same direction of effect for both PLT and MPV. SIRPA directly interacts with CD47, and SIRPA/CD47 signaling plays an important role in platelet clearance and the etiology of immune thrombocytopenia purpura[318-320]. Knockout *Sirpa* mice exhibit thrombocytopenia phenotypes, although they have similar MPV to control animals[320]. How genetic variation in *SIRPA* influences MPV in addition to its demonstrated contribution to PLT remains to be characterized. In addition to shared associations of PLT and MPV, there was overlap in the parallel Exomechip analyses of platelet reactivity. Largely mirroring results from previous GWASs, markers within *PEAR1*, *JMJD1C*, *PIK3CG*, and *MRVI1* showed the strongest associations with PLT/MPV and platelet reactivity[297; 321-323]. Other PLT/MPV-associated markers in *PTGES3*, *LINC00523*, and *RASGRP4* showed marginal associations. Notably, PTGES3 is linked to prostaglandin synthesis and the RasGRP family has been shown to have functional roles in blood cells including in platelet adhesion[324]. The association of platelet reactivity genes, particularly *PEAR1* and *MRVI1*, with PLT/MPV further supports a biological relationship between processes that control platelet function, megakaryopoiesis, and clearance[325; 326]. However, these large-scale association analyses are

unable to demonstrate whether these shared associations indicate shared biological mechanisms or simply reflect the epidemiological correlations among these traits.

In addition to platelet traits, there was substantial overlap of genetic associations with RBC and WBC traits examined by the BCX[244; 295]. The shared genetic associations with the two other primary blood cell lineages further supports other studies proposing that mechanisms that govern platelet size and number also influence RBC and WBC traits[327]. In BCX analyses, rs1050331 in the 3′ UTR of *ZMIZ2* was associated with increased PLT, mean corpuscular hemoglobin (MCH), and mean corpuscular volume (MCV), as well as with decreased WBC count[244; 295]. rs1050331 is also an eQTL for ZMIZ2 expression in whole blood (Table S21)[328]. There are known sex differences in cell counts, with females consistently having higher PLT and mixed results on MPV[329; 330]. Similar to well-established PLT- and MPV-associated transcriptional regulator JMJD1C, ZMIZ2 directly interacts with AR to modulate AR-mediated transcription and influences mesodermal development, and thus genetic variation in *ZMIZ2* could potentially contribute to hormonally mediate differences in PLT across genders[331-333]. Also associated with increased PLT and decreased RBC indices was rs55707100 in *MAP1A*. Though typically examined in a neurological context, *MAP1A* is involved in microtubule assembly, a process important in blood cell development and function[334]. Our observed association of *MAP1A* and its expression in platelets and RBCs suggests that the known role of *MAP1A* in developmental and cytoskeletal processes in neural tissues may extend to blood cells (Table S22). How these shared genetic factors specifically influence the development, maintenance, or clearance of multiple blood cell types remains to be determined.

Overlap with Non-Blood Cell Traits

Although the overlap with other blood cell traits may be intuitive, we also observed overlap with quantitative lipids traits. In cross-trait lookups, several known PLT/MPV loci confirmed in this study (e.g., *JMJD1C*, *GCKR*, and *SH2B3*) showed associations with lipids traits, and several known lipids loci showed association to PLT/MPV (e.g., *FADS1*, *FADS2*, *APOH*, and *TMEM50A*). Moreover, *SH2B3*, which is also expressed in human vascular endothelial cells where it modulates inflammation, has been associated with blood pressure and the risk of MI[138; 335; 336]. Our study further suggests that a regulation of platelets could also contribute to potential implication of *SH2B3* in the development of cardiovascular diseases. The associated SNVs in the *FADS1*/*FADS2* locus (rs174546 and rs174583) are eQTLs for multiple lipid-related transcripts in blood-related tissues, including TMEM258, FADS1, FADS2, and LDLR (Table S21)[328]. Intriguingly, expression of TMEM258 has also been shown to be a transcriptional regulatory target of cardiovascular disease implicated *CDKN2B-AS1* (MIM: 613149), a region marginally associated with PLT (discovery EA $p = 1.00 \times 10^{-6}$, replication EA $p = 0.0577$, combined EA $p = 1.56 \times 10^{-7}$) (Table S14)[301; 337; 338]. Our genetic association results link the underlying genetic architecture of platelet and lipids traits as suggested by previous epidemiological, genetic, and animal studies[330; 338-341]. However, these observed shared genetic associations do not demonstrate whether these reflect direct genetic pleiotropy or indirect relationships. Several variants previously implicated in lipids (e.g., *FADS1*, *FADS2*, *SH2B3*, *TMEM50A*, and *GCKR*) have stronger associations with lipids traits relative to our platelet associations, suggesting that their primary effects are on lipids

pathways (Table S20). Determining the directionality and causality among genetic variants, lipids, and platelets remains an important future step in dissecting which genetic variants may reveal new insights into platelet biology.

5.6.CONCLUSIONS

By performing a large meta-analysis of Exomechip association results, we identified rare, low-frequency, and common variants that influence PLT and MPV. Despite our ability to detect numerous associations with SNVs across a wide range of allele frequencies, the Exomechip interrogated a limited fraction of genomic variation. Sequencing-based studies across the genome in large sample sizes will be necessary to fully assess the contribution of variants across the allele frequency spectrum, particularly of rare variants in intergenic regions. Nonetheless, our results identify several intriguing genes and genetic mechanisms of platelet biology. Many of these associations overlapped with related blood cell and lipids traits, pointing to common mechanisms underlying their development and maintenance. Because blood cells share developmental lineages and several of our platelet-associated genes have known developmental or transcriptional regulatory functions, we hypothesize that the origins of these shared genetic associations are mainly in blood cell development in the bone marrow. How these genes function and interact in RBC, WBC, and platelet development will need to be tested in future experiments in both functional and human-based studies. Advances in these domains could provide key insights into genes that influence human blood disorders and reveal new mechanisms for the development of novel therapeutic applications

5.7. ACKNOWLEDGEMENTS

We thank all participants and study coordinating centers. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute, the NIH, or the U.S. Department of Health and Human Services. The Framingham Heart Study (FHS) authors acknowledge that the computational work reported on in this paper was performed on the Shared Computing Cluster, which is administered by Boston University's Research Computing Services. The MHI Biobank acknowledges the technical support of the Beaulieu-Saucier MHI Pharmacogenomic Center. We would like to thank Liling Warren for contributions to the genetic analysis of the SOLID-TIMI-52 and STABILITY datasets. The University Medicine Greifswald is a member of the Caché Campus program of the InterSystems GmbH. The SHIP and SHIP-TREND samples were genotyped at the Helmholtz Zentrum München. Estonian Genome Center, University of Tartu (EGCUT) would like to acknowledge Mr. V. Soo, Mr. S. Smith, and Dr. L. Milani. The Airwave Health Monitoring Study thanks Louisa Cavaliero who assisted in data collection and management as well as Peter McFarlane and the Glasgow CARE, Patricia Munroe at Queen Mary University of London, and Joanna Sarnecka and Ania Zawodniak at Northwick Park. FINCAVAS thanks the staff of the Department of Clinical Physiology for collecting the exercise test data. Young Finns Study (YFS) acknowledges the expert technical assistance in statistical analyses by Irina Lisinen.

CHAPTER 6:GENERAL DISCUSSION

6.1.LESSONS FROM OUR EXOME SEQUENCING STUDY

6.1.1. The BAG3 and FLNC mutations

In chapters 2 and 3, we identified mutations in BAG3 and FLNC. Both genes have important roles in skeletal and cardiac muscle and have been implicated in myofibrillar myopathy.

BAG3 is a co-chaperone of the heat shock proteins (HSP)s and has various roles such as apoptosis, autophagy, and cell adhesion. BAG3 has an apoptotic activity which is mediated by its role as a co-chaperone. It is suggested that a loss of its anti-apoptotic activity leads to degeneration of muscle fibers. Aimura et al, demonstrated in a BAG3 knockout mouse model that they displayed apoptotic nuclei in the striated muscles resulting in a severe form of skeletal myopathy and cardiomyopathy[342] . Another role associated with BAG3 is autophagy and the degradation of misfolded proteins. Hindering autophagy initiation may force the cells to choose death instead because they would have no mechanism to degrade the oxidized proteins. This cell death could then result in deteriorating muscle tissue, like that seen in Bag3 knockout mice. Moreover, BAG3 has a role in cell-adhesion and cytoskeleton integrity. The Arg309X mutation that we identified in chapter 2 is in the PXXP domain which modulates cell adhesion. Disruption of cell adhesion may lead to cell detachment from matrix and thus to muscle degeneration.

FLNC belongs to the Filamin family of actin-binding proteins. It is expressed predominantly in skeletal and cardiac muscle. In skeletal muscle, the majority of FLNC localizes to the Z-disk. In cardiac muscle, FLNC is found in intercalated discs [343]. FLNC plays

a major role in development and remodelling of the Actin cytoskeleton. It has a highly conserved structure with 2 actin-binding sites and 24 immunoglobulin (Ig) repeats. The Ig repeats act as an interface of interaction with binding proteins which has been thought to be important for the dimerization of FLNC enabling them to bundle or cross-link filaments. It is believed that dimerization of FLNC is an important step for its targeting to the cytoskeleton, and hence to ensure its function [344].

The mechanism by which mutations in FLNC result in muscle disease is not understood. However, haploinsufficiency has been suggested as one possible mechanism in distal myopathy[345]and in dilated cardiomyopathy[221] leading to a decrease in FLNC levels. The splice site that we describe in chapter 3 is an acceptor-splice site at exon 45. We did not functionally validate the mutation, however, aberrant splicing would lead to truncation of the last Ig domains. This may result in the disruption of the FLNC dimerization which may subsequently affect the integrity of the cytoskeleton. Truncation of the dimerization domain results in the loss of secondary structure of the mutant protein which makes it less stable and more susceptible to degradation by proteolytic enzymes [214; 346]

6.1.2. The value of genetic findings in monogenic disease

Genetic studies have provided exciting findings for monogenic disease. As a result, a vast array of clinical genetic testing panels is now available for patients with family history of heritable cardiovascular conditions including cardiomyopathies and arrhythmias. The discovery of causative mutations will reveal more information regarding the mechanisms of

these conditions which will have important implications for diagnosis, prognosis, and intervention.

With the recent advances in genetic technologies more genes are being added to panels. Further, the overlap of genes that cause different forms of cardiomyopathy led to the creation of pan-cardiomyopathy gene panels that include DCM, HCM, and ARVC genes. Other genes have also been shown to cause several forms of cardiovascular diseases such as *SCN5A* which is associated with DCM, Brugada syndrome and long QT syndrome. Hence, pan-arrhythmia and pan-cardio genetic testing panels are now available and include as many as 100 genes. With the costs of DNA sequencing rapidly decreasing, and with the number of novel genes identified on the rise, it is likely that the gene panels will soon be replaced with whole exome or genome sequencing which will make it possible to consult those results when new discoveries emerge.

For families, genetic testing provides many advantages. First, members negative for the causal mutation do not need to undergo regular screening. Second, the information will be useful to clinically intervene in mutation carriers for example through therapeutics or ICD implantation before they become symptomatic which may prevent or improve CVD mortality[347]. Lastly, this information will be useful for prenatal counseling and decision making.

6.1.3. Clinical impact of our study

Our exome sequencing study was initiated in parallel with many other studies that used the same technique to hunt for DCM genes. Despite the fact that we were not the first to report the implication of *BAG3* with DCM (chapter 2), our study provided unequivocal evidence of

that association because we had access to two multiplex pedigrees (the third family was small) which provided the power needed to confirm such an association. This is important because often, variants are attributed to be causal without sufficient evidence. For that reason, there is a debate on whether all DCM associated genes are in fact true candidates (see below), and hence, well powered studies, like ours, serve to establish the implication of the reported gene with DCM. For the same reason (power), we were able to study the impact of *BAG3* truncating variants in those families and report for the first time that *BAG3* is associated with early-onset DCM. The immediate clinical impact of our work was the addition of *BAG3* on the MHI's DCM existing panel that comprised at the time 6 genes in compliance with a position statement from the Canadian Cardiovascular Society [27]. The MHI was not the only hospital that originally did not screen for *BAG3*, other clinics (such as the Laboratory of Molecular Medicine (LMM) at Harvard) that utilize more comprehensive gene panels also did not have *BAG3* on their list at the time, and it is now added due to the rising evidence of its role in DCM.

Exome sequencing of a family with atypical cardiomyopathy allowed us to find a mutation in *FLNC* (chapter 3). Up until now, *FLNC* is not considered a DCM gene in the sense that it is not present on any screening panel (LMM, MHI) nor on commercial panels (illumina cardiomyopathy panel). In parallel to our finding, another study implicated *FLNC* in atypical cardiomyopathy[222]. In the family we analyzed, the disease presents with overlapping features of cardiomyopathy. It has a characteristic of DCM, left ventricular dilatation, but without systolic dysfunction (which is the major other requirement for the diagnosis of DCM), and pathologically a left dominant arrhythmogenic cardiomyopathy. Three patients of the four that we had analyzed also have fibrosis, the most prominent of which was present in the 32 year

old deceased individual. Our report provided additional evidence to the notion that *FLNC*, originally thought to be linked to myofibrillar myopathy, plays a role in cardiomyopathy independent of any skeletal muscle problems. The clinical impact of our study and the others[221; 222] that implicate *FLNC* is highlighting that the truncating mutations in *FLNC* may subject carriers to a higher risk of SCD and may require an ICD implantation as a preventive measure. Indeed, one of the family members in our study that had an ICD as a prevention had two appropriate shocks, confirming the critical role that such as intervention may have on *FLNC* truncating mutation carriers.

The vast majority of DCM mutations are missense. In a study conducted by Norton *et al*, the authors examined a list of all the variants that had been identified until then. Out of the 198 variants, the vast majority were missense, (83%) and 6% were truncating. With the subsequent analysis of *TTN*, more truncating variants in this gene have been described in many families. Truncating mutations are considered to be more pathogenic than a missense variant that alters one amino acid, however, this is not always true[206]. Truncating mutations are predominantly reported in *TTN*[43; 48]. In our two reports, the mutations in *BAG3* and *FLNC* were also truncating and we suggest that the impact on disease may be stronger with the first causing an earlier onset and the second causing an overlapping phenotype and fibrosis which is associated with degeneration of contractile function. Moreover, in both studies, the *BAG3* and *FLNC* families had a strong history of SCD.

A few points to mention in the context of the modifier effect of *BAG*3 and the possible association of *FLNC* and SCD. 1) Although SCD is a possible outcome of DCM and other cardiomyopathies, not all DCM families have a history of SCD. 2) ICD implantation is not always considered as this procedure is not free of possible complications. Clinicians follow a

thorough assessment of the patient and his/her family history, health condition, and stage of disease to weigh the risk and benefit of such a procedure. An ICD is usually considered when there is a history of SCD. However, *whether* and *when* to implement this procedure also vary and the decision to adopt this course of action is very subjective and in many instances depends on the physician following the case. 3) We do not expect that all mutations would impact the risk of SCD and thus establishing that correlation with specific mutations will help to categorize patients and improve clinical management.

Taking the three points into consideration, if a correlation between specific mutations and SCD or age of onset can be reliably established, this will play a significant role in orienting clinicians towards the right interventions or treatments and when to implement them. If the carrier has a mutation that is correlated with early onset, like we have shown for *BAG3,* then an intervention in a 17 year old asymptomatic carrier may not be too early as would have been speculated had this correlation not been established. And if an asymptomatic carrier has a truncating mutation in *FLNC*, then we suggest that an ICD may also be required, although we did not have a statistically robust evidence as we had had for *BAG3*, but we depend on observations from our study and the others on *FLNC* [221; 222] Hence, more studies are warranted to reliably establish the correlation between *FLNC* and SCD. In summary, genetic findings would provide a more solid support to any interventional procedure and may harmonize the strategies adopted by clinicians.

6.1.4. The utility of whole exome sequencing

Targeted NGS identifies pathogenic mutations in known genes. These panels are enriched for mutations in the candidate genes, and hence one can only identify a very rare variant that segregates with disease in a candidate gene, but may not itself be the causal one. In one of our families in chapter 2, the proband had a very rare likely causal mutation in *LMNA*. However, whole exome sequencing revealed another mutation in another gene called *RBM20*. The same mutation was published in a recent paper where they proved its segregation with disease in a huge multiplex family [199]. *RBM20* is on the list of genes targeted in genetic panels in only 21 labs compared to LMNA (65 labs) according to the Genetic Testing Registry (GTR) (https://www.ncbi.nlm.nih.gov/gtr/). Chances are that the majority of the labs would have ranked the *LMNA* as the pathogenic mutation in this family. The value of exome sequencing is that the data would still be available once new findings emerge and thus can be consulted. The cardiomyopathy panels will likely not keep up with the pace of research discoveries. Finally, with the current cost of NGS, one can carry a whole exome sequencing experiment for a similar price as targeted NGS [46]. It is likely that with the decrease of the cost of whole genome sequencing (WGS), it will also be an option in the future and will allow for discovery of non-coding variation and CNVs, although data interpretation will still be a major hurdle.

An example where WES was particularly successful is highlighted in the Finding of Rare Disease Genes (FORGE) Canada Consortium effort which aims to identify novel genes implicated in rare pediatric genetic disorders such as muscular diseases, birth defects and intellectual disability[348]. The consortium has successfully identified more than 67 novel genes for a wide range of rare disorders[220; 348].

6.1.5. Exome sequencing caveats

6.1.5.1. Coverage

Exome sequencing targets the exons of the genome and is thus enriched for coding exonic variants and does not provide adequate coverage of other types of variations. As such, findings for Mendelian disease have mainly been coding variants, but we did not sufficiently explore noncoding variation that may affect disease (e.g through epigenetics and gene regulation as discussed earlier). These aspects can not be addressed by exome sequencing. In addition to SNPs, genetic variation may be due to small insertions or deletions, or CNVs. Exome sequencing does not detect insertions and deletions reliably where calling methods give variable results. And thus these types of variants remain not fully explored with exome sequencing.

6.1.5.2. Too much data poses interpretation issues

WES is a very attractive approach for discovery of genes implicated in rare disorders. However, despite the exciting prospects of WES, this technology has several limitations. Sequencing technologies yield a huge amount of variants. Although some of these will be pathogenic or risk alleles, the majority will be benign. Strategies used in order to assign pathogenicity to variants include evidence of segregation of the variant within a family, its conservation across species and the use of computational prediction tools, as well as filtering variants using public sequencing databases such as the 1000 genomes project, ESP, and ExAC datasets (see below for discussion). For some families, the filtering process leads to the identification of a "likely" pathogenic variant in a known gene. Subsequent segregation analysis in other unaffected (preferably older) members of the family is very helpful to increase the likelihood that the mutation identified is in fact pathogenic. However, when one

does not find pathogenic variants in known genes, it becomes extremely challenging to identify the causal mutation (and thus novel genes) in these families. Usually, several variants thought to be "likely pathogenic" are retained by the end of the filtering steps and the interpretation of the results becomes very difficult. What makes it even more challenging is the fact that often mutations are private to a family, which means catalogues of published pathogenic variants such as Clinvar or HGMD that one can consult to prioritize pathogenic mutations will not always be useful. Of course, biological knowledge about the gene in which the mutation is present in terms of its function, the pathway it is associated with or its interaction with other (possibly heart failure) genes is instrumental in ranking the likelihood of variants. Identifying other families that carry other pathogenic mutations in the same gene will also support its involvement in disease. The best way to confirm pathogenicity would be to test those variants functionally, however, this is not practical in a clinical setting where every novel variant segregates in one family or few families. The next few years will likely witness more tweaking of the current technologies and tools to improve the sensitivity of identifying novel mutations and genes. Further, expanding current sequencing databases will greatly aid in prioritizing variants.

6.1.6. Other challenges in monogenic disease studies

Although the genetic approaches have improved our ability to conduct analyses for gene discovery, it remains that there are certain challenges that hinder that discovery and that are in many instances outside the control of the clinician or analyst. And unlike research that involves big numbers of individuals, studies that rely on family pedigrees are affected by individual-related issues, which would be negligible in population-based studies.

The main challenge is that several people are needed in a pedigree in order to reliably implicate a disease gene. This is particularly important when a mutation in a known gene has not been found, and we are considering a bigger set of possible novel candidates. The first issue is family compliance; this could take several forms such as lack of interest of family members that are needed to carry out a segregation analysis, or familial disputes that limit our freedom to contact other members to participate, or in many cases, simply the lack of availability of members either in small families or families with history of SCD and only one affected individual.

Another important challenge is phenotyping of family members. When a history of DCM is known, then imaging and clinical examination of family members will be evaluated differently. In other words, any abnormality would be considered as a sign of the disease because of the family history, and that same abnormality may have been ignored in another individual. This makes it complicated to assign the phenotypic status of such participating family members and in many cases, these are left unanalyzed in the genetic study due to an "undetermined" affection status which further reduces the number of participating individuals.

6.1.7. Current databases and variant prioritization

The compelling results that were achieved using the sequencing technologies have prompted various research groups to join forces and aggregate their datasets into public databases that serve as a resource for researchers and genetic clinicians and that help to overcome the interpretational challenges that accompany the ever so growing genetic data. As a result multiethnic databases such as the 1000 genomes, ESP, and ExAC have been created. The 1000 genomes aims at providing a catalogue for most of the genetic variants that have a

frequency of at least 1% in the populations studied. The dataset now contains genomes of more than 2,000 samples. ESP includes exome sequencing data from more than 6,000 individuals from various projects that include well-phenotyped and diverse populations. Recently, the 1000 genomes, ESP and many other contributing projects comprising disease-specific and population genetic studies were combined to create yet a more comprehensive resource, ExAC which includes the exome sequences of more than 60,000 samples. Last October, the ExAC dataset has expanded even more and the genome Aggregation Database (gnomAD) now includes exome sequences from more than 123,000 samples and whole genome sequencing from more than 15,000 samples.

These datasets made it possible to attain calculations of allele frequencies for the variants that were found in the participating individuals and thus became indispensable resources. As these databases grow in size, they make it easier to interpret sequencing data. One can look up a variant of interest to know its frequency in the general population and also to compare frequencies between the different ethnicities (although this is not yet perfect as the sample sizes contributed by the various ethnicities largely differ). From the assumption that the frequency of Mendelian disease-causing alleles are extremely rare or private mutations, then we would expect the disease-causing allele to either be very rare or even absent from these datasets (since the dataset is a subset of the general population). Studies have used MAF cutoffs to filter out variants in their analyses of sequencing data in order to facilitate the task of prioritizing variants in search of the pathogenic ones. There is no official way to decide on an MAF cutoff, but usually it is a function of the prevalence of the disease in question, the rarer the disease the less frequent the causal mutations is going to be and the lower the MAF threshold. It is important that the cutoff is not too stringent so as not to filter out potentially

pathogenic variants nor too loose to avoid the inclusion of false positives. Whiffin *et al* suggested a statistical framework that in addition to disease prevalence uses knowledge of previous disease causing variants to determine an MAF cutoff [349] to be applied in exome sequencing studies.

6.1.8. Proving pathogenicity

Traditionally, to determine pathogenicity of variants causal of DCM, linkage studies provided a strong statistical evidence of linkage followed by functional validation. Currently, many WES studies on cardiomyopathies rely on bioinformatics methods to assess pathogenecity. We take into consideration population frequency, the conservation of the site of the variant, the predicted effect on function- by using software such as polyphen and SIFT, segregation of the variant (for monogenic diseases), whether the gene is expressed in the heart, and its interaction with other genes that play a role in the disease of interest (discussed above).

The American College of Medical Genetics and Genomics (ACMGG) has recommended certain guidelines to establish the pathogenicity of variants using typical types of variant evidence (e.g., population data, computational data, functional data, and segregation data).[350]. They classify variants into "pathogenic", "likely pathogenic", "benign", "likely benign", or "uncertain significance". The classification depends on the level of evidence that in turn ranges from very strong evidence of pathogenicity to very strong evidence that the variant is benign. The criteria to assign the level of pathogenicity are very intricate, where there exist many possible scenarios for each variant class and can be consulted here[350]. For example, a variant is considered pathogenic if it satisfies one condition of the "very strong" evidence conditions plus one strong evidence condition, if it has two strong evidence conditions, or one

strong evidence and 3 moderate evidence conditions and so on. Therefore, a truncating variant with evidence of disease segregation is considered pathogenic. A truncating variant is also considered pathogenic in the absence of family data if it has been functionally proven to cause the disease. A missense variant is considered pathogenic if it segregates in the family AND if it has been functionally proven to cause the disease.

In a large pedigree, that includes more than two affected people, such as the case in our *FLNC* report, proving pathogenicity will be more reliable that in smaller pedigrees. Multiplex pedigrees are ideal because they would allow a LOD score calculation which would be a robust evidence of linkage, as is the case with our *BAG3* study. However, there is no doubt that the most convincing proof of pathogenicity would be to test those variants experimentally. Appropriate experimental methods can be selected depending on the class of variant and the feasibility and cost of the experiment. Overall, to prove pathogenicity of those variants, the main outcomes that one seeks are that the gene is disrupted or that the variant led to the disease phenotype in cells derived from the patient or in a well-validated *in vivo* model such as mouse or zebrafish. For both *in vitro* and *in vivo* models one needs to show that the introduction of the variant, or an engineered vector carrying the variant, into a cell line or animal model resulted in the disease phenotype and ideally that the phenotype can be rescued by addition of wild-type gene product or specific knockdown on the risk allele of the variant.

For DCM, the majority of variants thought to be causal of the disease are coding variants. However, reports that scrutinized the list of reported DCM variants argue that many of those reported as pathogenic are not in fact pathogenic [349; 351]. Walsh *et al* [351]considered all reported DCM variants in HGMD and cross checked the allele frequencies of those variants in the ExAC catalogue, and found that 19.6% of individuals in ExAC carry reported DCM

disease-causing variants which exceed the disease prevalence by far. Hence, not all reported variants are causal and giving higher priority to published variants in clinical genetic testing is not recommended. As the database continues to expand and as clinical genetic testing moves to larger gene panels and whole-exome and genome sequencing, variant interpretation will increasingly improve and a better interpretation of disease variants will likely be achieved.

6.1.9. The hunt for modifier genes

Even when the causal mutation is identified, the phenotypic heterogeneity observed in patients within the same family will be a challenge. Who has a worse prognosis? Who is at more risk of sudden cardiac death? For instance, it is extremely crucial in the management of cardiomyopathy to know who is at a higher risk of arrhythmias. Mutations in *SCN5A* have been associated with arrhythmias in family-based and population-based studies [40; 352; 353]. These studies facilitate the stratification of individuals and in choosing appropriate therapies such as implanting an ICD in *SCN5A* mutations carriers. Future studies may implicate other genes that may modify severity or cause a higher risk of sudden cardiac death. The current genetic technologies have the potential of identifying variants that may act as modifiers of disease, but such studies will require a large number of unrelated individuals and will need to be replicated before they can be clinically relevant.

DCM displays phenotypic heterogeneity, meaning that the manifestation of disease and its prognosis differs considerably between patients. DCM also has both 1) genetic heterogeneity, meaning that there is not only one gene that causes the disease as is the case for other mendelian traits like Huntingon or cystic fibrosis, rather there is a considerable number of genes that have been liked to DCM; 2) and allelic heterogeneity since the same mutation

may have a different impact on its carriers within the same family. Carriers of the same mutation may exhibit very distinct phenotypic characteristics and in certain cases a different type of cardiomyopathy[3]. Hence, in addition to identifying the causal mutation, it is essential that we explore whether there are genetic factors that modify the disease characteristics or severity. Understanding the sources of this heterogeneity would impact the clinical management of the disease.

The concept of "modifier genetics" in the context of DCM remains largely unexplored. Unraveling the genetic factors that would render a DCM patient at a higher risk to have arrhythmias or a more severe prognosis would be very valuable for clinicians to make the right clinical decisions in managing the affected families. Within the same context, identifying the factors that may protect certain carriers from developing the disease (reduced penetrance, see introduction) would also be extremely important for the clinical management of DCM and would lead to a better understanding of the underlying mechanisms and the pathophysiology of this disease. Such discoveries might also guide the development of tailored therapies for DCM. In our paper for example, a carrier of the *BAG3* mutation had not developed the disease at age 67. We do know that she had been taking beta-blocker medication for a while. It is hard to believe that this may have halted the disease, but it could be a combination of the medication, lifestyle or environmental factors as well as other genetic variants that she may carry and her affected relatives do not.

Tackling the issue of modifier genetics requires extremely large sample sizes of DCM patients to have enough power to detect the associated variants. One would gather a large sample of DCM probands and conduct an association study. Information about all the relevant clinical parameters and complications, as well as sequencing data for both affected and

unaffected individuals would be required. Since we are looking at variants that modify the disease, and not cause it, then the rationale of using exome date is insufficient here since modifiers may be noncoding variants. One would then need to look for variants that are present at a significantly higher frequency in individuals with a disease-related characteristic or complication, for example arrhythmia, or heart failure, than with individuals without (association testing). These variants are likely to be common and if they are true "modifiers", then their effect on the disease would be apparent once the causal mutation is present. It is also possible to conduct such analyses using family data, but that would require many multiplex pedigrees in order to have sufficient patients with or without the complication or phenotypic parameter that we are studying, and that is more difficult to obtain. In addition, individuals that are related will also share most of their genetic data, thus unrelated individuals would pose less challenges.

6.2.LESSONS FROM THE EXOMECHIP STUDY

6.2.1. Pleiotropy in blood cell traits

Pleiotropy is a phenomenon where a single genetic locus influences multiple traits. As shown in our results in chapters 4 and 5, the identified variants are associated with more than one blood cell phenotype across all three major cell types, red blood cells, white blood cells and platelets, as well as other traits such as lipids, T2D, obesity, etc. Since blood cells are involved in a variety of biological processes, it would be plausible that some of the genetic loci would have pleiotropic effects on a number of hematological and other related traits. Pleiotropy poses challenges in analyzing and interpreting association studies. This overlap makes it harder to pinpoint the direct effect of the SNP. In order to disentangle the actual

effect of the genetic variation, multivariate association analyses that account for the correlation between the traits is required. Such analyses provide additional statistical power to detect novel genes contributing to pleiotropic events and may give new insights into the biology of the overlapping traits.

6.2.2. Rare variants and complex traits

Gene-based tests such as SKAT, or Variable-Threshold (VT) tests allow to combine information across variants and evaluate the aggregation of the effects of multiple variants in a gene or region. Using these methods and aggregating the effects or low-frequency variants, we identified novel genes implicated in blood cell traits such as *PKLR* and *ITGA2B* (chapters 4 and 5A). Methods such as whole exome sequencing, low-depth WGS, and the exome chip have been suggested as genetic tools to capture rare variants. However, they are also associated with some limitations. For example, low-depth sequencing has limited accuracy for identifying rare variants, whole exome sequencing is limited for the exome and the exome chip is limited to targeted regions. In general, all methods will have their advantages and disadvantages, and there will probably not be a perfect one, rather a combination of all would yield the most results.

In our exome chip study we were able to identify 12 novel rare and low-frequency independent variants for red blood cells (chapter 4) and platelet traits (chapter 5). The exome chip has also been successful in enabling the identification of rare variants with other traits such as blood pressure [103; 104] and lipid traits [102]. Since the exome chip also includes GWAS tag SNPs, then it is possible to run conditional analyses by correcting for the GWAS signal to capture the rare variant that may be driving the association in known genes. In chapter 4, we

ran conditional analyses correcting for known red blood cells signals and we found a rare variant that was independent of the GWAS signal, a rare variant in the *ANK1* gene. The same result was presented in another study [354]. A rare splice variant in *HBB* (MAF = 0.008%) which we published in another work (appendix 2) [243] was also independent of known GWAS signals in this locus. But we could not find rare variants that fully explained a GWAS signal. In other words, the GWAS signal was not lost when we adjusted for a given rare variant, and both SNPs are independently contributing to the phenotype where the variation is not exclusively driven by one of the variants. In appendix 2, using the exome chip, we demonstrate that a rare missense variant in *EPO* (MAF=0.5%) associated with red blood cell traits was also independent of the strongest GWAS signal. In the same study, using the exome chip allowed the identification of a novel gene for WBC count, *CXCR2*, that has not been discovered in GWAS studies and that harbored several rare and low-frequency variants contributing to the association. Perhaps the best demonstration of the utility of the exome chip was lately proven in the GIANT consortium [106] where 83 rare and low-frequency coding variants (MAF < 5%) were reported to be associated with height and having higher effect sizes compared to common variants. The study comprised more than 700k individuals proving that the combination of the exome chip with very large sample sizes are invaluable to uncover novel rare variation.

In addition to the exomechip, sequencing studies are thought to capture rare variants. A WES study for blood cell traits successfully identified a rare variant in *GFI1B* associated with lower platelet count and using genome editing and knockdown experiments showed that the variant plays a role in suppressing platelet production [355]. Following our exome chip studies, a GWAS in the UK Biobank and INTERVAL cohorts that included more than 173k individuals

identified hundreds of novel rare and low-frequency variants associated with blood cell indices [356]. The success of a GWAS to capture rare variation is heavily due to the availability of whole genome sequencing reference panels such as the 1000 genomes [357] , UK10K [358] ,and the Haplotype Reference Consortium[359] projects that include more rare variants than what was originally included in previous GWAS studies, and thus highly improve the quality of genotyping and imputation to better capture rare and low-frequency variants in association analyses.

6.2.3. Associated variants lying in Mendelian disease genes

In chapters 4 and 5, we identified rare and low-frequency variants in several genes that cause Mendelian forms of blood disorders such as *ALAS2*, which is mutated in sideroblastic anemia and *PKLR* implicated in non-spherocytic hemolytic anemia. For platelet traits, we found a low-frequency variant associated with decreased platelet count in *ITGA2B,* a gene mutated in the rare disorder, Glanzmann's thrombasthenia. Previous resequencing studies of genes implicated in Mendelian disorders have revealed that rare variants in those genes can contribute to various complex traits at a population level. *LDLR* mutations for example, can cause both, the rare disorder familial hypercholersterolemia (FH) which manifests at a young age and a complex form of hypercholerterolemia that manifests in the fourth or fifth decades of life. These results suggest that there exist common pathways between rare disorders and complex traits that when perturbed ultimately lead to disease.

6.2.4. The missing heritability problem

The term “missing heritability” has been coined to refer to the genetic factors that may explain the remaining genetic component of the phenotypic variance [100]. Solving the missing heritability has a great impact on human health. The heritability estimates the contribution of genetic information to disease. It is believed that for complex traits, the effect or influence comes from several hundreds of variants and the accumulation of their effects. Therefore, if we do not know the majority of these contributing factors, then we can not take this information to the next level, which is the translation of these findings into clinical use. The hope is that our knowledge of our genetic information will help us 1) predict disease and 2) achieve tailored treatment (refer to the section on “The Goal of Personalized Medicine” for a broader discussion). Knowing our genetic risk can help us to predict disorders and thus prevent them when possible. For example, people who have a genetic susceptibility to MI will be advised to have a specific diet and lifestyle. Solving the missing heritability is key for individualized treatment. Without it, our ability to translate even the information that we already know is limited. For example, using the current knowledge of breast cancer genetics, genetic tests have bene developed that aim to categorize patients into groups based on how well they will respond to chemotherapy treatment or how likely will they have cancer recurrence. These are created based on our knowledge of a certain number of genes that have been validated. The objectives behind these tests is what we want the genetic information to achieve however, we can only interpret the results of these tests with great caution because a lot is not known yet. In other words, the absence of certain alleles in one individual does not mean that he does not carry other risk alleles that we have not identified yet.

It is thought that the missing heritability lies in low-frequency (1% < MAF < 5%), and rare variants (MAF < 1%) (the definition of cutoffs varies), with small to strong effect sizes in

addition to other intricate biological processes such as epigenetic modifications, gene-gene and gene-environment interactions. However, it is possible to improve our ability to detect the "missing heritability" by giving importance to certain aspects of data analysis.

6.2.4.1. Well defined phenotypes

Having a phenotype that is not well defined can dilute signals and contribute to the missing heritability. A well-defined phenotype makes the distinction between cases and controls more pronounced and more easily detectable. Difference in the analyzed phenotypes may be due to 1) disease phenotypic heterogeneity or 2) poorly measured phenotypes. For complex traits with a wide range of phenotype signatures it is essential to analyze these sub-groups separately in order to capture the genetic contribution for each. Once these phenotypes are combined, this will lead to a decrease in power to identify genetic variation.

Some phenotypes or traits are not measured in a standard fashion (such as blood pressure) in different studies. Thus, one challenge is to implement a uniform procedure when data comes from different studies and countries, which is the case of the consortia or meta-analyses which leads to introduction of noise. It is generally believed that larger sample sizes will account for issues in defining and measuring a trait (and the differences between the different parties contributing to the study). However, I think it is equally important to seek having a clean phenotype as a way to increase power to detect novel findings.

6.2.4.2. The value of non-European Ethnicities

Although we know today a lot more about certain ethnicities like East Asians and South Asians and intriguing results have been discovered in non-European populations, the

overwhelming majority of studies have been performed on Caucasians. In chapter 4, I report a novel association between a nonsense variant in *CD36* and red blood cell traits in the African American population. This result was not significant in the European cohort which highlights the gain in knowledge that we could achieve in exploring non-European populations. Some variants will have higher allele frequencies in non-European populations compared to Europeans which will increase power. The variant may also be monomorphic in European populations, but not in other ethnicities in which case any association would only be detected in non-European ethnic groups. For example, the *G6PD* locus association with red blood cells is African American-specific. Conversely, rare splice variant, rs33971440, in *HBB* associated with hemoglobin and hematocrit levels is only detected in Europeans. A GWAS in Latin Americans identified novel associations with white blood cell traits [360] not previously seen in Europeans. Another advantage of exploring different ethnicities is that the LD patterns differ between populations, and hence a genotyped SNP may be in greater LD with a causal variant in one population compared to another. To make use of the difference in LD, it is important that genotyping chips include genetic variation of diverse ethnic groups not only Europeans. Finally, the environmental factors will differ between one population and another. Hence certain associations may have a greater effect in one population compared to another due to gene-environment or even gene-gene interactions.

Given the variable allele frequencies and environmental backgrounds, the information gleaned from the plethora of genetic findings in Europeans may not apply to other populations. There exists a wealth of genetic information that lies in each ethnic population and thus studies in diverse populations is extremely crucial for gene discovery and to contribute to explaining the missing heritability.

6.2.4.3. Gene-Environment interactions

Other factors that may explain the missing heritability lie in the effects of gene-gene and gene-environment interactions. Exploring the effects of gene-environment interactions are necessary to better understand the underlying biology and pathophysiology of disease, Numerous reports have addressed these types of interactions, however, for the most part, these results have not been replicated. Challenges in conducting such analyses are often due to the variability in measuring environmental exposures between studies. Further, lifestyle measures such as diet and exercise are usually self-reported. Additionally, many of the environmental factors of interest are correlated; hence an interaction between a variant and one environmental variable may be driven by its correlation with another variable. Further, more power is required to detect interactions compared to genotype phenotype associations. One way to overcome the power limitation is to meta-analyze data from several studies. However, meta-analyses often dichotomize continuous variables to account for between-study heterogeneity which leads to a loss of power. In addition, meta-analyses usually consider only one interaction at a time which may overestimate the interaction with a particular variable since the latter may be correlated with other variables. A possible solution could be to fit more than one environmental variable jointly. This however, requires very large sample sizes since there is a loss of power once many variables are included in the model. A recent gene-environment interaction study made use of the large UK biobank dataset and analyzed more than 100k individuals [361]. The study found an interaction between the *FTO* locus and several environmental variables including alcohol consumption, diet, physical activity and others by using a joint model. Mores studies in this context are required to elucidate the contribution of

gene-environmental interactions to disease heritability. It is possible that their overall perceived contribution is overestimated.

6.2.4.4. Rare variants and significance thresholds

Analyses of rare variants require very large samples to detect them. However, it is possible that we are being stringent in our P-value thresholds and we are losing some true positives. For example, in chapter 5, 7/13 PLT variants and 4/10 MPV variants that did not pass the significance threshold in the discovery analysis were nonetheless replicated in the replication cohorts and surpassed significance thresholds when the discovery and replication results were combined. These variants would have been ignored it we strictly relied on the significance thresholds. Within this context, it is also difficult to replicate very rare variants. For example, in chapter 4, we identified a rare variant in *ALAS2* associated with MCH. This gene is implicated in sideroblastic anemia and thus we have likely identified a true signal, however, the variant did not replicate, most probably because the replication samples in which this variant was present was very small. Hence, it is possible that some of the missing heritability lies in rare variants that we are overlooking in our analyses.

6.2.5. Functional experiments in blood cells

The majority of SNPs identified to be associated with blood cell traits are non-coding and may be involved in gene regulation and hence expression. Gene expression can be regulated by several factors such as transcriptional regulatory networks, enhancers, methylation etc. Polymorphisms in regulatory elements may modify the levels of gene transcripts. Consequently, transcript abundance can be measured and considered as a

quantitative trait. The use of both whole genome association studies and the measurement of global gene expression permits the discovery of expression quantitative trait locus (eQTL)s. Once a genetic variant is identified to be associated with a trait, a genome-wide eQTL mapping data can be examined to check if the variant is associated with quantitative transcript levels. In chapters 4 and 5, we used eQTL databases in order to check whether any of the associated variants for red blood cell traits and platelets are expressed in relevant tissues. eQTL studies can be used as a general method to help identify a set of target genes. The emergence of large-scale genomics projects such as ENCODE, and other efforts, is aiding in attaining a better understanding of the non-coding regions of the human genome. These studies have benefitted from next generation sequencing technologies to generate genome-wide maps of functional elements such as regulatory elements. Such studies help in prioritizing variants by studying the overlap with molecular features or interactions. For example, a candidate causal variant may overlap with a sequence motif within a known binding site for a particular transcription factor, giving clues that the variant may be functional[362].

As discussed in chapter 1, hematological traits are amenable to functional experiments which is extremely crucial to validate findings from association studies. First, blood and its cell types are easily accessible. Hematological measurements are normally available in most cohorts or biobanks. Second, blood cell types can be differentiated and new genes can be tested in cell culture systems or model organisms. In chapter 4, we were able to functionally test the nonsense mutation in *CD36* using differentiated erythroblasts and show that there was a reduction in the expression of *CD36* in heterozygotes of the identified variant.

To validate whether a gene causes a given phenotype, techniques such as CRISPR/Cas9 or gene knockdown approaches in cellular models or model organisms may be applied. Unlike monogenic diseases, complex traits are thought to be caused by many low-effect size variants, hence testing the function of those variants requires a framework that integrates all variants and test them simultaneously. A GWAS study has demonstrated platelet phenotype of 11 novel genes by silencing them in model organisms [122]. Antisense morpholino silencing of *ARHGEF3* in zebrafish lead to ablation of both primitive erythropoiesis and thrombocyte formation, and a novel role has been ascribed to *ARHGEF3* in the regulation of iron uptake and erythroid cell maturation [122;https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916502/ -voxs12217-bib-0014][276]. A recent report attempted to study the function of genes associated with platelet size and number. To achieve this they selected 24 loci among the 68 loci identified by GWAS studies and knocked down their expression in a zebrafish model using morpholino oligonucleotides. As a result, they were able to describe the role of 9 genes whose role in thrombopoiesis was not previously known [363].

Advancing technologies to expand and differentiate pluripotent stem cells into blood cells (e.g. megakaryocytes/platelets [364],erythroid progenitors/RBC[365], macrophages[366]) for clinical and commercial applications enable their use as a model system of hematopoiesis. There is potential to produce and bank all blood subtypes which will allow to study the effect of variation from the start of differentiation with hematopoietic stem cells towards production of mature blood cells.

6.2.6. The value of identified variants in complex traits

GWAS and exomechip studies identified thousands of variants associated with hundreds of diseases. Many of these are not themselves causal of disease but they likely tag a causal variant. But what is the prognostic value of those variants? Unlike monogenic diseases, the variants associated with complex traits will have a low impact on disease susceptibility. However, findings have contributed greatly to our understanding of complex disease. For instance, they provide new understandings of the underlying biology, or can be useful in certain approaches to establish causation pathways between an exposure and an outcome that is often unclear in epidemiological studies alone (due to issues such as reverse causation) through Mendelian randomization studies. We also know that some variants interact with drug response (e.g clopidogrel) or contributed to the discovery of drug targets such as *PCSK9* [367]. Hence, although the effect size on disease is low *per se*, findings for complex traits have proven to be extremely useful. Therefore, out of the numerous findings, it is important to establish methods to prioritize those that matter and that have biological relevance. There is a dire need to design a framework which combines the functional evidence with the statistical support and helps in prioritizing variants based on an experimentally proven function.

Computational methods will also be crucial. Recently, a study combined next generation sequencing, bioinformatics, and clinical data to derive a diagnostic workflow. The tool allows to prioritize variants based on both pathogenicity and similarity of the patients' phenotype to described diseases. They correctly ranked genes based on the number of pathogenic variants 86% of the time and predicted disease in a prospective study [368].

Current knowledge of associated variants made it possible to design genetic risk scores and test their ability to predict disease. Genetic risk scores have a main advantage which is that they remain stable throughout life and can be helpful in predicting disease at any age. In the

context of CVD, risk scores that integrated variants associated with CVD or with CVD risk factors have moderately improved disease prediction over traditional risk factors[369; 370]. To be clinically useful, genetic risk scores need to provide significant improvements in disease prediction. Recently, statistical and computational methods that allow for prioritizing variants or better integration of genetic risk scores into the assessment of clinical risk have been suggested [371]. It is also possible that with both the discovery of more well-characterized variants and the prioritization of known signals, genetic risk scores will be more robust. Second, once we know more about the genetic basis of diseases (solving the missing heritability) and the biological relevance of those genetic factors elucidated then it will be possible to choose variants for risk integration. Machine learning tools will also be useful to prioritize candidate genes.

Associated rare variants would have a better prognostic value and will be easier to link to genes since they are predominantly in coding regions. As discussed earlier, they also tend to have a larger phenotypic effect. If variants are coding, then their associated mechanism such as the disruption of the protein function, would be easier to interpret. Hence, rare variants may be considered "actionable" variants with a better potential for scientific and/or clinical value such as therapeutic targets [372].

Given the complex nature of common diseases, statistical and analytical methods need to be designed that will ultimately lead to additional variant discovery, classification and interpretation.

6.3.THE GOAL OF PERSONALIZED MEDICINE

The advances in genomics have revealed a trove of genetic information that helped predicting certain diseases (monogenic) and achieve targeted treatments. It is because of the successes of genetic studies and the feasibility of genetic tests – facilitated by advances in technologies – that the notion of "personalised medicine" became a goal worth pursuing.

Examples of personalised medicine have already had significant impact on health. As presented in this thesis, the clinical management of monogenic diseases has already benefitted greatly from genetic findings. Patients with causal mutations for cardiomyopathy are considered for preventative interventions (eg. ICD) and asymptomatic family members are genetically screened for mutations. It is likely that when the picture of modifier genetics becomes more lucid, disease classification and hence personalized medicine and treatment will also be possible. Screening for *BRCA1* and *BRCA2* mutations in breast cancer is another example of personalized medicine and has a large benefit in terms of mortality reduction. Although the prevalence of these mutations is very small, they confer more than 70% lifetime risk for breast and ovarian cancer. According to the US Preventive Services Task Force, recommendations for genetic testing for *BRCA1* and *BRCA2* mutations in high risk women in the US have been issued since 2005.

However, for the majority of conditions, we have yet to prove that the additional knowledge gained from genetic information, could not otherwise be provided by the assessment of risk factors and family history. As for polygenic traits, as mentioned above, the prognostic value of the genetic findings is still very limited. In the context of pharmacogenomics, it has been shown that some variants are associated with a different drug response. The main strategic challenge is applying those tests in the clinic. Not only do we need to prove their clinical utility, but also their cost-effectiveness. Will it be more cost-

effective to go with the “trial and error” method to decide what a suitable dose of warfarin is for patients, or to genotype them? Which tests should be applied in routine practice?

The Centers for Disease Control and Prevention (CDC) office of public health genomics (https://phgkb.cdc.gov/GAPPKB/topicStartPage.do), uses guidelines based on the FDA recommendations, the research evidence, and the clinical evidence, to classify genetic testing into three categories: tier 1 (supports implementation in practice), tier 2 (clinically valid, but more evidence required), tier 3 (not yet recommended)[373]. This is a very effective way in providing guidelines that can be consulted in order to decide on implementing genetic tests. These tests are either related to prediction of disease, drug dosage, or drug choice. Currently, more than 100 drugs have a pharmacogenomics information label on them reflecting the great accomplishments of genetic findings and their integration within genetic testing. For instance, a drug in tier 1, cetuximab is ineffective in 40% of colorectal cancer patients [374]. Those that have a mutation in the *KRAS* gene will not benefit from the drug. The genetic test will help stratify patients based on their genotype and treat them accordingly. Other drugs like clopidogrel, for which there is considerable evidence for inter-individual variability in response due to a variants in the *CYP2C19* gene[375], has not been introduced in clinical practice yet illustrating an example where clinical evidence exists but the clinical utility has yet to be proven.

In addition to the scientific impediments, there are other educational, social, and strategic challenges. On the level of education, there is a clear disparity on genomic medicine education among clinicians. There is an overload of data from genomic studies that clinicians and health care professionals need to keep up with. Further, surveys demonstrated that even when physicians are familiar with genomic medicine, some of them would not use the results of

genomic testing to guide their clinical decisions and disease management[376]. Therefore, programs to educate primary care professionals and physicians will be essential to learn more about the genomic evidence in order to appreciate the value of genomic medicine.

We need strategies to manage the social impact of personalized medicine and have a plan to answer questions like: Should the government or insurance companies have access to an individual's genetic data? How do we prevent genetic discrimination in the workplace? Scientists and lawyers have worked for so long to address these issues, however, they remain issues that trouble the minds of individuals. The next few years will have provided a more complete picture about the genetic basis of disease and with more findings, the goal of personalized medicine will be more attainable.

6.4.CONCLUSIONS AND FINAL COMMENTS

In conclusion, the work presented in this thesis contributed new genetic findings for DCM and blood cell traits. For DCM, I identified truncating mutations in *BAG3* that predispose carriers to early onset DCM and contributed in adding yet another evidence for the importance of the inclusion of *BAG3* in gene panels. I also identified a truncating mutation in a novel gene, *FLNC* that causes a distinct type of cardiomyopathy with fibrosis, arrhythmias, and history of SCD. For blood cell traits, we identified 16 novel genes for red blood cells and 15 novel for platelet traits in a very large, multiethnic and well powered study. These findings have important implications in understanding the biology of blood cell traits and contribute to expanding the list of genes involved in these traits.

Future directions of genetic studies in both realms of human disease will focus on addressing the challenges of each. For monogenic disease, the current technologies will enable studies to continue to identify more genes implicated in disease, but also to consider types of variants not well captured by exome sequencing studies such as insertions, deletions, CNVs etc. More rigorous steps of variant validation are also required in order to limit false positives. Collaborations among different research groups are invaluable and will increase power to conduct such types of analyses owing to the small number of people with rare diseases. Additionally I believe that clinical genetic testing and research studies still need to go hand in hand in the search of novel genes. While gene panels are more feasible in a clinical context, I believe that families with negative results should always be considered for exome sequencing as a part of research study in order to expand the list of causal genes. Finally, finding modifier variants will have a substantial impact on disease classification and individual-level treatment.

For complex traits, current and future studies will focus on explaining the remaining phenotypic variance of disease. The use of the exomechip and of exome sequencing was successful in identifying rare variants for some traits but not for others. Strategies to capture rare variants in addition to the current ones will also require studies in isolated populations and investigating the extremes of the population distribution. Efforts in including diverse ethnic groups have already been fruitful and warrant more samples and the inclusion of more ethnic groups. Exploring the effect of the environmental factors and their interactions with genetic factors will also likely contribute to explaining a portion of the missing heritability. Another challenge is to find the value of the thousands of variants that had been identified. The current state is that we are overloaded with findings, the majority of which is yet to be proven meaningful. The rapidly decreasing cost of genomic technologies will make it easier to

address many of our questions. More importantly, the coming years will greatly rely on tailored analytical and statistical methods, more computational tools, and functional studies in order to complete the puzzle and make sense of the information that we have found to increase the prognostic value of genetic findings and hence the evidence of their clinical utility.

APPENDIX 1: LARGE-SCALE EXOME-WIDE ASSOCIATION ANALYSIS IDENTIFIES LOCI FOR WHITE BLOOD CELL TRAITS AND PLEIOTROPY WITH IMMUNE-MEDIATED DISEASES.

Authors: Tajuddin SM, Schick UM, Eicher JD, Chami N, Giri A, Brody JA, Hill WD, Kacprowski T, Li J, Lyytikäinen LP, Manichaikul A, Mihailov E, O'Donoghue ML, Pankratz N, Pazoki R, Polfus LM, Smith AV, Schurmann C, Vacchi-Suzzi C, Waterworth DM, Evangelou E, Yanek LR, Burt A, Chen MH, van Rooij FJ, Floyd JS, Greinacher A, Harris TB, Highland HM, Lange LA, Liu Y, Mägi R, Nalls MA, Mathias RA, Nickerson DA, Nikus K, Starr JM, Tardif JC, Tzoulaki I, Velez Edwards DR, Wallentin L, Bartz TM, Becker LC, Denny JC, Raffield LM, Rioux JD, Friedrich N, Fornage M, Gao H, Hirschhorn JN, Liewald DC, Rich SS, Uitterlinden A, Bastarache L, Becker DM, Boerwinkle E, de Denus S, Bottinger EP, Hayward C, Hofman A, Homuth G, Lange E, Launer LJ, Lehtimäki T, Lu Y, Metspalu A, O'Donnell CJ, Quarells RC, Richard M, Torstenson ES, Taylor KD, Vergnaud AC, Zonderman AB, Crosslin DR, Deary IJ, Dörr M, Elliott P, Evans MK, Gudnason V, Kähönen M, Psaty BM, Rotter JI, Slater AJ, Dehghan A, White HD, Ganesh SK, Loos RJ, Esko T, Faraday N, Wilson JG, Cushman M, Johnson AD, Edwards TL, Zakai NA, Lettre G, Reiner AP, Auer PL.

Reference: Large-Scale Exome-wide Association Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases. Tajuddin SM, Schick UM, Eicher JD, Chami N, Giri A, Brody JA *et al*. Am J Hum Genet. 2016 Jul 7;99(1):22-39

ARTICLE

Large-Scale Exome-wide Association Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases

Salman M. Tajuddin,[1,80] Ursula M. Schick,[2,3,80] John D. Eicher,[4] Nathalie Chami,[5,6] Ayush Giri,[7] Jennifer A. Brody,[8] W. David Hill,[9,10] Tim Kacprowski,[11,12] Jin Li,[13] Leo-Pekka Lyytikäinen,[14,15] Ani Manichaikul,[16] Evelin Mihailov,[17] Michelle L. O'Donoghue,[18] Nathan Pankratz,[19] Raha Pazoki,[20] Linda M. Polfus,[21] Albert Vernon Smith,[22,23] Claudia Schurmann,[2,3] Caterina Vacchi-Suzzi,[24] Dawn M. Waterworth,[25] Evangelos Evangelou,[26,27] Lisa R. Yanek,[28] Amber Burt,[29] Ming-Huei Chen,[4] Frank J.A. van Rooij,[20] James S. Floyd,[8] Andreas Greinacher,[30] Tamara B. Harris,[31] Heather M. Highland,[21,32] Leslie A. Lange,[33] Yongmei Liu,[34] Reedik Mägi,[17] Mike A. Nalls,[35] Rasika A. Mathias,[36] Deborah A. Nickerson,[37] Kjell Nikus,[38,39] John M. Starr,[9,40] Jean-Claude Tardif,[5,6] Ioanna Tzoulaki,[26,27] Digna R. Velez Edwards,[41] Lars Wallentin,[42] Traci M. Bartz,[43] Lewis C. Becker,[44] Joshua C. Denny,[45] Laura M. Raffield,[33] John D. Rioux,[5,6] Nele Friedrich,[12,46] Myriam Fornage,[47] He Gao,[26] Joel N. Hirschhorn,[48,49] David C.M. Liewald,[9,10] Stephen S. Rich,[16] Andre Uitterlinden,[20,50,51] Lisa Bastarache,[45] Diane M. Becker,[28] Eric Boerwinkle,[21,52] Simon de Denus,[6,53] Erwin P. Bottinger,[2]

(Author list continued on next page)

White blood cells play diverse roles in innate and adaptive immunity. Genetic association analyses of phenotypic variation in circulating white blood cell (WBC) counts from large samples of otherwise healthy individuals can provide insights into genes and biologic pathways involved in production, differentiation, or clearance of particular WBC lineages (myeloid, lymphoid) and also potentially inform the genetic basis of autoimmune, allergic, and blood diseases. We performed an exome array-based meta-analysis of total WBC and subtype counts (neutrophils, monocytes, lymphocytes, basophils, and eosinophils) in a multi-ancestry discovery and replication sample of ~157,622 individuals from 25 studies. We identified 16 common variants (8 of which were coding variants) associated with one or more WBC traits, the majority of which are pleiotropically associated with autoimmune diseases. Based on functional annotation, these loci included genes encoding surface markers of myeloid, lymphoid, or hematopoietic stem cell differentiation (*CD69, CD33, CD87*), transcription factors regulating lineage specification during hematopoiesis (*ASXL1, IRF8, IKZF1, JMJD1C, ETS2-PSMG1*), and molecules involved in neutrophil clearance/apoptosis (*C10orf54, LTA*), adhesion (*TNXB*), or centrosome and microtubule structure/function (*KIF9, TUBD1*). Together with recent reports of somatic *ASXL1* mutations among individuals with idiopathic cytopenias or clonal hematopoiesis of undetermined significance, the identification of a common regulatory 3′ UTR variant of *ASXL1* suggests that both germline and somatic *ASXL1* mutations contribute to lower blood counts in otherwise asymptomatic individuals. These association results shed light on genetic mechanisms that regulate circulating WBC counts and suggest a prominent shared genetic architecture with inflammatory and autoimmune diseases.

Introduction

White blood cells (WBCs) are major constituents of the blood and lymphatic system. They are classified into two lineages: myeloid (neutrophils, basophils, eosinophils, and monocytes) and lymphoid (lymphocytes). Lineage commitment of hematopoietic stem cells involves precise transcriptional and epigenetic regulation, creating the

[1]Laboratory of Epidemiology and Population Sciences, National Institute on Aging, NIH, Baltimore, MD 21224, USA; [2]The Charles Bronfman Institute for Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [3]The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [4]Population Sciences Branch, National Heart Lung and Blood Institute, The Framingham Heart Study, Framingham, MA 01702, USA; [5]Department of Medicine, Université de Montréal, Montréal, QC H3T 1J4, Canada; [6]Montreal Heart Institute, Montréal, QC H1T 1C8, Canada; [7]Division of Epidemiology, Institute for Medicine and Public Health, Vanderbilt University, Nashville, TN 37235, USA; [8]Department of Medicine, University of Washington, Seattle, WA 98101, USA; [9]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK; [10]Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK; [11]Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald and Ernst-Moritz-Arndt University Greifswald, Greifswald 17475, Germany; [12]DZHK (German Centre for Cardiovascular Research), partner site Greifswald, Greifswald, Germany; [13]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Palo Alto, CA 94305, USA; [14]Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland; [15]Department of Clinical Chemistry, University of Tampere School of Medicine, Tampere 33014, Finland; [16]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA; [17]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia; [18]TIMI Study Group, Cardiovascular Division, Brigham and Women's Hospital, Boston, MA 02115, USA; [19]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55454, USA; [20]Department of Epidemiology, Erasmus University Medical Center, Rotterdam 3000, the Netherlands; [21]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [22]Icelandic Heart Association, 201 Kopavogur, Iceland; [23]Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland; [24]Department of Family,

(Affiliations continued on next page)

APPENDIX 2: RARE AND LOW-FREQUENCY CODING VARIANTS IN CXCR2 AND OTHER GENES ARE ASSOCIATED WITH HEMATOLOGICAL TRAITS

Authors: Auer PL, Teumer A, Schick U, O'Shaughnessy A, Lo KS, Chami N, Carlson C, de Denus S, Dubé MP, Haessler J, Jackson RD, Kooperberg C, Perreault LP, Nauck M, Peters U, Rioux JD, Schmidt F, Turcot V, Völker U, Völzke H, Greinacher A, Hsu L, Tardif JC, Diaz GA, Reiner AP, Lettre G.

Reference: Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Auer PL, Teumer A, Schick U, O'Shaughnessy A, Lo KS, Chami N *et al*. Nat Genet. 2014 Jun;46(6):629-34

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2014 December 01.

Published in final edited form as:

Nat Genet. 2014 June ; 46(6): 629–634. doi:10.1038/ng.2962.

Rare and low-frequency coding variants in *CXCR2* and other genes are associated with hematological traits

Paul L. Auer[1,2], **Alexander Teumer**[3], **Ursula Schick**[2], **Andrew O'Shaughnessy**[4], **Ken Sin Lo**[5], **Nathalie Chami**[5], **Chris Carlson**[2], **Simon de Denus**[5,6], **Marie-Pierre Dubé**[5,6], **Jeff Haessler**[2], **Rebecca D. Jackson**[7], **Charles Kooperberg**[2], **Louis-Philippe Lemieux Perreault**[5], **Matthias Nauck**[8], **Ulrike Peters**[2,9], **John D. Rioux**[5,6], **Frank Schmidt**[3], **Valérie Turcot**[5], **Uwe Völker**[3], **Henry Völzke**[10], **Andreas Greinacher**[11], **Li Hsu**[2], **Jean-Claude Tardif**[5,6], **George A. Diaz**[4,12,13], **Alexander P. Reiner**[2,9,13], and **Guillaume Lettre**[5,6,13]

[1]School of Public Health, University of Wisconsin-Milwaukee, 1240 N. 10th Street, Milwaukee WI, 53201, USA.

[2]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle WA, 98109, USA.

[3]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Germany.

[4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

[5]Montreal Heart Institute, 5000 Bélanger Street, Montréal, Quebec, H1T 1C8, Canada.

[6]Université de Montréal, 2900 Boul. Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada.

[7]Division of Endocrinology, Diabetes, and Metabolism, Ohio State University, 376 W 10th Avenue, Columbus OH, 43210, USA.

[8]Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Germany.

[9]Department of Epidemiology, University of Washington School of Public Health, 1959 NE Pacific Street, Seattle WA, 98195, USA.

[10]Institute for Community Medicine, University Medicine Greifswald, Germany.

[11]Institute for Immunology and Transfusion Medicine, University Medicine Greifswald, Germany.

Correspondence: George A. Diaz, george.diaz@mssm.edu, Tel.: 212-659-6790, Fax: 212-849-2508. Alexander P. Reiner, apreiner@u.washington.edu, Phone: 206-667-2710, Fax: 206-667-4142. Guillaume Lettre, guillaume.lettre@umontreal.ca, Phone: 514-376-3330, Fax: 514-593-2539.

[13]These authors co-directed the study.

Author contributions

PLA, GAD, APR and GL conceived and designed the experiments. PLA, AT, US, AO, KSL, GAD, APR and GL performed the experiments. PLA, AT, US, AO, KSL, GAD, APR and GL analyzed the data. All authors contributed reagents and materials. PLA, GAD, APR and GL Wrote the paper with contributions from all authors.

Competing financial interests

The authors declare no competing financial interests.

REFERENCES

1. Maron, B.J., Towbin, J.A., Thiene, G., Antzelevitch, C., Corrado, D., Arnett, D., Moss, A.J., Seidman, C.E., Young, J.B., American Heart, A., et al. (2006). Contemporary definitions and classification of the cardiomyopathies: an American Heart Association Scientific Statement from the Council on Clinical Cardiology, Heart Failure and Transplantation Committee; Quality of Care and Outcomes Research and Functional Genomics and Translational Biology Interdisciplinary Working Groups; and Council on Epidemiology and Prevention. Circulation 113, 1807-1816.
2. Codd, M.B., Sugrue, D.D., Gersh, B.J., and Melton, L.J., 3rd. (1989). Epidemiology of idiopathic dilated and hypertrophic cardiomyopathy. A population-based study in Olmsted County, Minnesota, 1975-1984. Circulation 80, 564-572.
3. Hershberger, R.E., Hedges, D.J., and Morales, A. (2013). Dilated cardiomyopathy: the complexity of a diverse genetic architecture. Nature reviews Cardiology 10, 531-547.
4. Mann, D.L., and Bristow, M.R. (2005). Mechanisms and models in heart failure: the biomechanical model and beyond. Circulation 111, 2837-2849.
5. Elliott, P., Andersson, B., Arbustini, E., Bilinska, Z., Cecchi, F., Charron, P., Dubourg, O., Kuhl, U., Maisch, B., McKenna, W.J., et al. (2008). Classification of the cardiomyopathies: a position statement from the European Society Of Cardiology Working Group on Myocardial and Pericardial Diseases. European heart journal 29, 270-276.
6. Burkett, E.L., and Hershberger, R.E. (2005). Clinical and genetic issues in familial dilated cardiomyopathy. Journal of the American College of Cardiology 45, 969-981.
7. Sen-Chowdhry, S., Jacoby, D., Moon, J.C., and McKenna, W.J. (2016). Update on hypertrophic cardiomyopathy and a guide to the guidelines. Nature reviews Cardiology 13, 651-675.
8. Maron, B.J., Casey, S.A., Poliac, L.C., Gohman, T.E., Almquist, A.K., and Aeppli, D.M. (1999). Clinical course of hypertrophic cardiomyopathy in a regional United States cohort. Jama 281, 650-655.
9. Elliott, P.M., Gimeno, J.R., Thaman, R., Shah, J., Ward, D., Dickie, S., Tome Esteban, M.T., and McKenna, W.J. (2006). Historical trends in reported survival rates in patients with hypertrophic cardiomyopathy. Heart 92, 785-791.
10. Raghow, R. (2016). An 'Omics' Perspective on Cardiomyopathies and Heart Failure. Trends in molecular medicine 22, 813-827.
11. Corrado, D., Basso, C., Thiene, G., McKenna, W.J., Davies, M.J., Fontaliran, F., Nava, A., Silvestri, F., Blomstrom-Lundqvist, C., Wlodarska, E.K., et al. (1997). Spectrum of clinicopathologic manifestations of arrhythmogenic right ventricular cardiomyopathy/dysplasia: a multicenter study. Journal of the American College of Cardiology 30, 1512-1520.
12. Sen-Chowdhry, S., Syrris, P., Ward, D., Asimaki, A., Sevdalis, E., and McKenna, W.J. (2007). Clinical and genetic characterization of families with arrhythmogenic right ventricular dysplasia/cardiomyopathy provides novel insights into patterns of disease expression. Circulation 115, 1710-1720.

13. Thiene, G., Nava, A., Corrado, D., Rossi, L., and Pennelli, N. (1988). Right ventricular cardiomyopathy and sudden death in young people. The New England journal of medicine 318, 129-133.
14. Saberniak, J., Hasselberg, N.E., Borgquist, R., Platonov, P.G., Sarvari, S.I., Smith, H.J., Ribe, M., Holst, A.G., Edvardsen, T., and Haugaa, K.H. (2014). Vigorous physical activity impairs myocardial function in patients with arrhythmogenic right ventricular cardiomyopathy and in mutation positive family members. European journal of heart failure 16, 1337-1344.
15. Corrado, D., Wichter, T., Link, M.S., Hauer, R., Marchlinski, F., Anastasakis, A., Bauce, B., Basso, C., Brunckhorst, C., Tsatsopoulou, A., et al. (2015). Treatment of arrhythmogenic right ventricular cardiomyopathy/dysplasia: an international task force consensus statement. European heart journal 36, 3227-3237.
16. Christensen, A.H., Benn, M., Bundgaard, H., Tybjaerg-Hansen, A., Haunso, S., and Svendsen, J.H. (2010). Wide spectrum of desmosomal mutations in Danish patients with arrhythmogenic right ventricular cardiomyopathy. Journal of medical genetics 47, 736-744.
17. Kaski, J.P., Syrris, P., Burch, M., Tome-Esteban, M.T., Fenton, M., Christiansen, M., Andersen, P.S., Sebire, N., Ashworth, M., Deanfield, J.E., et al. (2008). Idiopathic restrictive cardiomyopathy in children is caused by mutations in cardiac sarcomere protein genes. Heart 94, 1478-1484.
18. Wu, W., Lu, C.X., Wang, Y.N., Liu, F., Chen, W., Liu, Y.T., Han, Y.C., Cao, J., Zhang, S.Y., and Zhang, X. (2015). Novel Phenotype-Genotype Correlations of Restrictive Cardiomyopathy With Myosin-Binding Protein C (MYBPC3) Gene Mutations Tested by Next-Generation Sequencing. Journal of the American Heart Association 4.
19. Menon, S.C., Michels, V.V., Pellikka, P.A., Ballew, J.D., Karst, M.L., Herron, K.J., Nelson, S.M., Rodeheffer, R.J., and Olson, T.M. (2008). Cardiac troponin T mutation in familial cardiomyopathy with variable remodeling and restrictive physiology. Clinical genetics 74, 445-454.
20. Purevjav, E., Arimura, T., Augustin, S., Huby, A.C., Takagi, K., Nunoda, S., Kearney, D.L., Taylor, M.D., Terasaki, F., Bos, J.M., et al. (2012). Molecular basis for clinical heterogeneity in inherited cardiomyopathies due to myopalladin mutations. Human molecular genetics 21, 2039-2053.
21. Bellet, S. (1932). Congenital heart disease with multiple cardiac anomalies: report of a case showing aortic atresia, fibrous scar in myocardium and embryonal sinusoidal remains. Am J Med Sci, 458-465.
22. Luxan, G., Casanova, J.C., Martinez-Poveda, B., Prados, B., D'Amato, G., MacGrogan, D., Gonzalez-Rajal, A., Dobarro, D., Torroja, C., Martinez, F., et al. (2013). Mutations in the NOTCH pathway regulator MIB1 cause left ventricular noncompaction cardiomyopathy. Nature medicine 19, 193-201.
23. Arbustini, E., Weidemann, F., and Hall, J.L. (2014). Left ventricular noncompaction: a distinct cardiomyopathy or a trait shared by different cardiac diseases? Journal of the American College of Cardiology 64, 1840-1850.
24. Klaassen, S., Probst, S., Oechslin, E., Gerull, B., Krings, G., Schuler, P., Greutmann, M., Hurlimann, D., Yegitbasi, M., Pons, L., et al. (2008). Mutations in sarcomere protein genes in left ventricular noncompaction. Circulation 117, 2893-2901.
25. Ichida, F., Tsubata, S., Bowles, K.R., Haneda, N., Uese, K., Miyawaki, T., Dreyer, W.J., Messina, J., Li, H., Bowles, N.E., et al. (2001). Novel gene mutations in patients with left ventricular noncompaction or Barth syndrome. Circulation 103, 1256-1263.

26. Ross, H., Howlett, J., Arnold, J.M., Liu, P., O'Neill, B.J., Brophy, J.M., Simpson, C.S., Sholdice, M.M., Knudtson, M., Ross, D.B., et al. (2006). Treating the right patient at the right time: access to heart failure care. The Canadian journal of cardiology 22, 749-754.
27. Arnold, J.M., Liu, P., Demers, C., Dorian, P., Giannetti, N., Haddad, H., Heckman, G.A., Howlett, J.G., Ignaszewski, A., Johnstone, D.E., et al. (2006). Canadian Cardiovascular Society consensus conference recommendations on heart failure 2006: diagnosis and management. The Canadian journal of cardiology 22, 23-45.
28. Braunwald, E. (2013). Heart failure. JACC Heart failure 1, 1-20.
29. Morales, A., and Hershberger, R.E. (2015). The Rationale and Timing of Molecular Genetic Testing for Dilated Cardiomyopathy. The Canadian journal of cardiology 31, 1309-1312.
30. Japp, A.G., Gulati, A., Cook, S.A., Cowie, M.R., and Prasad, S.K. (2016). The Diagnosis and Evaluation of Dilated Cardiomyopathy. Journal of the American College of Cardiology 67, 2996-3010.
31. Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306, 234-238.
32. Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L., et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245, 1066-1073.
33. Abifadel, M., Varret, M., Rabes, J.P., Allard, D., Ouguerram, K., Devillers, M., Cruaud, C., Benjannet, S., Wickham, L., Erlich, D., et al. (2003). Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nature genetics 34, 154-156.
34. Towbin, J.A., Hejtmancik, J.F., Brink, P., Gelb, B., Zhu, X.M., Chamberlain, J.S., McCabe, E.R., and Swift, M. (1993). X-linked dilated cardiomyopathy. Molecular genetic evidence of linkage to the Duchenne muscular dystrophy (dystrophin) gene at the Xp21 locus. Circulation 87, 1854-1865.
35. Muntoni, F., Cau, M., Ganau, A., Congiu, R., Arvedi, G., Mateddu, A., Marrosu, M.G., Cianchetti, C., Realdi, G., Cao, A., et al. (1993). Brief report: deletion of the dystrophin muscle-promoter region associated with X-linked dilated cardiomyopathy. The New England journal of medicine 329, 921-925.
36. Kass, S., MacRae, C., Graber, H.L., Sparks, E.A., McNamara, D., Boudoulas, H., Basson, C.T., Baker, P.B., 3rd, Cody, R.J., Fishman, M.C., et al. (1994). A gene defect that causes conduction system disease and dilated cardiomyopathy maps to chromosome 1p1-1q1. Nature genetics 7, 546-551.
37. Olson, T.M., Michels, V.V., Thibodeau, S.N., Tai, Y.S., and Keating, M.T. (1998). Actin mutations in dilated cardiomyopathy, a heritable form of heart failure. Science 280, 750-752.
38. Fatkin, D., MacRae, C., Sasaki, T., Wolff, M.R., Porcu, M., Frenneaux, M., Atherton, J., Vidaillet, H.J., Jr., Spudich, S., De Girolami, U., et al. (1999). Missense mutations in the rod domain of the lamin A/C gene as causes of dilated cardiomyopathy and conduction-system disease. The New England journal of medicine 341, 1715-1724.
39. Olson, T.M., and Keating, M.T. (1996). Mapping a cardiomyopathy locus to chromosome 3p22-p25. The Journal of clinical investigation 97, 528-532.
40. McNair, W.P., Ku, L., Taylor, M.R., Fain, P.R., Dao, D., Wolfel, E., Mestroni, L., and Familial Cardiomyopathy Registry Research, G. (2004). SCN5A mutation associated with dilated cardiomyopathy, conduction disorder, and arrhythmia. Circulation 110, 2163-2167.

41. Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitas, K., Sasse-Klaassen, S., Seidman, J.G., Seidman, C., Granzier, H., Labeit, S., et al. (2002). Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. Nature genetics 30, 201-204.
42. Siu, B.L., Niimura, H., Osborne, J.A., Fatkin, D., MacRae, C., Solomon, S., Benson, D.W., Seidman, J.G., and Seidman, C.E. (1999). Familial dilated cardiomyopathy locus maps to chromosome 2q31. Circulation 99, 1022-1026.
43. Herman, D.S., Lam, L., Taylor, M.R., Wang, L., Teekakirikul, P., Christodoulou, D., Conner, L., DePalma, S.R., McDonough, B., Sparks, E., et al. (2012). Truncations of titin causing dilated cardiomyopathy. The New England journal of medicine 366, 619-628.
44. Norton, N., Li, D., Rampersaud, E., Morales, A., Martin, E.R., Zuchner, S., Guo, S., Gonzalez, M., Hedges, D.J., Robertson, P.D., et al. (2013). Exome sequencing and genome-wide linkage analysis in 17 families illustrate the complex contribution of TTN truncating variants to dilated cardiomyopathy. Circulation Cardiovascular genetics 6, 144-153.
45. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272-276.
46. Norton, N., Li, D., and Hershberger, R.E. (2012). Next-generation sequencing to identify genetic causes of cardiomyopathies. Current opinion in cardiology 27, 214-220.
47. Galmiche, L., Serre, V., Beinat, M., Assouline, Z., Lebre, A.S., Chretien, D., Nietschke, P., Benes, V., Boddaert, N., Sidi, D., et al. (2011). Exome sequencing identifies MRPL3 mutation in mitochondrial cardiomyopathy. Human mutation 32, 1225-1231.
48. Hershberger, R.E., and Morales, A. (2015). Dilated Cardiomyopathy Overview. In GeneReviews(R), R.A. Pagon, M.P. Adam, H.H. Ardinger, S.E. Wallace, A. Amemiya, L.J.H. Bean, T.D. Bird, N. Ledbetter, H.C. Mefford, R.J.H. Smith, et al., eds. (Seattle (WA).
49. Roncarati, R., Viviani Anselmi, C., Krawitz, P., Lattanzi, G., von Kodolitsch, Y., Perrot, A., di Pasquale, E., Papa, L., Portararo, P., Columbaro, M., et al. (2013). Doubly heterozygous LMNA and TTN mutations revealed by exome sequencing in a severe form of dilated cardiomyopathy. European journal of human genetics : EJHG 21, 1105-1111.
50. Hinson, J.T., Chopra, A., Nafissi, N., Polacheck, W.J., Benson, C.C., Swist, S., Gorham, J., Yang, L., Schafer, S., Sheng, C.C., et al. (2015). HEART DISEASE. Titin mutations in iPS cells define sarcomere insufficiency as a cause of dilated cardiomyopathy. Science 349, 982-986.
51. Hastings, R., de Villiers, C.P., Hooper, C., Ormondroyd, L., Pagnamenta, A., Lise, S., Salatino, S., Knight, S.J., Taylor, J.C., Thomson, K.L., et al. (2016). Combination of Whole Genome Sequencing, Linkage, and Functional Studies Implicates a Missense Mutation in Titin as a Cause of Autosomal Dominant Cardiomyopathy With Features of Left Ventricular Noncompaction. Circulation Cardiovascular genetics 9, 426-435.
52. Posch, M.G., Posch, M.J., Geier, C., Erdmann, B., Mueller, W., Richter, A., Ruppert, V., Pankuweit, S., Maisch, B., Perrot, A., et al. (2008). A missense variant in desmoglein-2 predisposes to dilated cardiomyopathy. Molecular genetics and metabolism 95, 74-80.
53. Garcia-Pavia, P., Syrris, P., Salas, C., Evans, A., Mirelis, J.G., Cobo-Marcos, M., Vilches, C., Bornstein, B., Segovia, J., Alonso-Pulpon, L., et al. (2011). Desmosomal protein gene mutations in patients with idiopathic dilated cardiomyopathy undergoing cardiac transplantation: a clinicopathological study. Heart 97, 1744-1752.
54. Taylor, M.R., Fain, P.R., Sinagra, G., Robinson, M.L., Robertson, A.D., Carniel, E., Di Lenarda, A., Bohlmeyer, T.J., Ferguson, D.A., Brodsky, G.L., et al. (2003). Natural history of dilated

cardiomyopathy due to lamin A/C gene mutations. Journal of the American College of Cardiology 41, 771-780.
55. Bonne, G., Di Barletta, M.R., Varnous, S., Becane, H.M., Hammouda, E.H., Merlini, L., Muntoni, F., Greenberg, C.R., Gary, F., Urtizberea, J.A., et al. (1999). Mutations in the gene encoding lamin A/C cause autosomal dominant Emery-Dreifuss muscular dystrophy. Nature genetics 21, 285-288.
56. Raffaele Di Barletta, M., Ricci, E., Galluzzi, G., Tonali, P., Mora, M., Morandi, L., Romorini, A., Voit, T., Orstavik, K.H., Merlini, L., et al. (2000). Different mutations in the LMNA gene cause autosomal dominant and autosomal recessive Emery-Dreifuss muscular dystrophy. American journal of human genetics 66, 1407-1412.
57. Brodt, C., Siegfried, J.D., Hofmeyer, M., Martel, J., Rampersaud, E., Li, D., Morales, A., and Hershberger, R.E. (2013). Temporal relationship of conduction system disease and ventricular dysfunction in LMNA cardiomyopathy. Journal of cardiac failure 19, 233-239.
58. Burke, M.A., Cook, S.A., Seidman, J.G., and Seidman, C.E. (2016). Clinical and Mechanistic Insights Into the Genetics of Cardiomyopathy. Journal of the American College of Cardiology 68, 2871-2886.
59. Chami, N., Tadros, R., Lemarbre, F., Lo, K.S., Beaudoin, M., Robb, L., Labuda, D., Tardif, J.C., Racine, N., Talajic, M., et al. (2014). Nonsense mutations in BAG3 are associated with early-onset dilated cardiomyopathy in French Canadians. The Canadian journal of cardiology 30, 1655-1661.
60. Ellsworth, M.L., Ellis, C.G., Goldman, D., Stephenson, A.H., Dietrich, H.H., and Sprague, R.S. (2009). Erythrocytes: oxygen sensors and modulators of vascular tone. Physiology 24, 107-116.
61. Klinken, P.S. (2002). Red Blood Cells. Int J of Biochem Cell Biol.
62. Koury, M.J., Bondurant, M.C., and Atkinson, J.B. (1987). Erythropoietin control of terminal erythroid differentiation: maintenance of cell viability, production of hemoglobin, and development of the erythrocyte membrane. Blood cells 13, 217-226.
63. Shivdasani, R.A., and Orkin, S.H. (1996). The transcriptional control of hematopoiesis. Blood 87, 4025-4039.
64. Prchal, J.T., and Sokol, L. (1996). "Benign erythrocytosis" and other familial and congenital polycythemias. European journal of haematology 57, 263-268.
65. Herrick, J. (1910). Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. Archives of internal medicine 6, 517-521.
66. Karpatkin, S. (1969). Heterogeneity of human platelets. I. Metabolic and kinetic evidence suggestive of young and old platelets. The Journal of clinical investigation 48, 1073-1082.
67. Karpatkin, S. (1969). Heterogeneity of human platelets. II. Functional evidence suggestive of young and old platelets. The Journal of clinical investigation 48, 1083-1087.
68. Kamath, S., Blann, A.D., and Lip, G.Y. (2001). Platelet activation: assessment and quantification. European heart journal 22, 1561-1571.
69. Kario, K., Matsuo, T., and Nakao, K. (1992). Cigarette smoking increases the mean platelet volume in elderly patients with risk factors for atherosclerosis. Clinical and laboratory haematology 14, 281-287.
70. Coban, E., Ozdogan, M., Yazicioglu, G., and Akcit, F. (2005). The mean platelet volume in patients with obesity. International journal of clinical practice 59, 981-982.
71. Papanas, N., Symeonidis, G., Maltezos, E., Mavridis, G., Karavageli, E., Vosnakidis, T., and Lakasas, G. (2004). Mean platelet volume in patients with type 2 diabetes mellitus. Platelets 15, 475-478.

72. Nadar, S., Blann, A.D., and Lip, G.Y. (2004). Platelet morphology and plasma indices of platelet activation in essential hypertension: effects of amlodipine-based antihypertensive therapy. Annals of medicine 36, 552-557.
73. Pathansali, R., Smith, N., and Bath, P. (2001). Altered megakaryocyte-platelet haemostatic axis in hypercholesterolaemia. Platelets 12, 292-297.
74. Coppinger, J.A., Cagney, G., Toomey, S., Kislinger, T., Belton, O., McRedmond, J.P., Cahill, D.J., Emili, A., Fitzgerald, D.J., and Maguire, P.B. (2004). Characterization of the proteins released from activated platelets leads to localization of novel platelet proteins in human atherosclerotic lesions. Blood 103, 2096-2104.
75. Gawaz, M., Langer, H., and May, A.E. (2005). Platelets in inflammation and atherogenesis. The Journal of clinical investigation 115, 3378-3384.
76. Baigent, C., Blackwell, L., Collins, R., Emberson, J., Godwin, J., Peto, R., Buring, J., Hennekens, C., Kearney, P., Meade, T., et al. (2009). Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. Lancet 373, 1849-1860.
77. Ruggeri, Z.M. (2002). Platelets in atherothrombosis. Nature medicine 8, 1227-1234.
78. Mackman, N. (2008). Triggers, targets and treatments for thrombosis. Nature 451, 914-918.
79. Thaulow, E., Erikssen, J., Sandvik, L., Stormorken, H., and Cohn, P.F. (1991). Blood platelet count and function are related to total and cardiovascular death in apparently healthy men. Circulation 84, 613-617.
80. Pizzulli, L., Yang, A., Martin, J.F., and Luderitz, B. (1998). Changes in platelet size and count in unstable angina compared to stable angina or non-cardiac chest pain. European heart journal 19, 80-84.
81. Ly, H.Q., Kirtane, A.J., Murphy, S.A., Buros, J., Cannon, C.P., Braunwald, E., Gibson, C.M., and Group, T.S. (2006). Association of platelet counts on presentation and clinical outcomes in ST-elevation myocardial infarction (from the TIMI Trials). The American journal of cardiology 98, 1-5.
82. Ranjith, M.P., Divya, R., Mehta, V.K., Krishnan, M.G., KamalRaj, R., and Kavishwar, A. (2009). Significance of platelet volume indices and platelet count in ischaemic heart disease. Journal of clinical pathology 62, 830-833.
83. Gonzalez-Porras, J.R., Martin-Herrero, F., Gonzalez-Lopez, T.J., Olazabal, J., Diez-Campelo, M., Pabon, P., Alberca, I., and San Miguel, J.F. (2010). The role of immature platelet fraction in acute coronary syndrome. Thrombosis and haemostasis 103, 247-249.
84. Lordkipanidze, M. (2012). Platelet turnover in atherothrombotic disease. Current pharmaceutical design 18, 5328-5343.
85. Martin, J.F., Bath, P.M., and Burr, M.L. (1991). Influence of platelet size on outcome after myocardial infarction. Lancet 338, 1409-1411.
86. Klovaite, J., Benn, M., Yazdanyar, S., and Nordestgaard, B.G. (2011). High platelet volume and increased risk of myocardial infarction: 39,531 participants from the general population. Journal of thrombosis and haemostasis : JTH 9, 49-56.
87. Koupenova, M., Kehrel, B.E., Corkrey, H.A., and Freedman, J.E. (2016). Thrombosis and platelets: an update. European heart journal.
88. Chu, S.G., Becker, R.C., Berger, P.B., Bhatt, D.L., Eikelboom, J.W., Konkle, B., Mohler, E.R., Reilly, M.P., and Berger, J.S. (2010). Mean platelet volume as a predictor of cardiovascular risk: a systematic review and meta-analysis. Journal of thrombosis and haemostasis : JTH 8, 148-156.

89. Jackson, S.P. (2011). Arterial thrombosis--insidious, unpredictable and deadly. Nature medicine 17, 1423-1436.
90. Fischer, M., Broeckel, U., Holmer, S., Baessler, A., Hengstenberg, C., Mayer, B., Erdmann, J., Klein, G., Riegger, G., Jacob, H.J., et al. (2005). Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. Circulation 111, 855-862.
91. Garner, C., Tatu, T., Reittie, J.E., Littlewood, T., Darley, J., Cervino, S., Farrall, M., Kelly, P., Spector, T.D., and Thein, S.L. (2000). Genetic influences on F cells and other hematologic variables: a twin heritability study. Blood 95, 342-346.
92. Cardno, A.G., and Gottesman, II. (2000). Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. American journal of medical genetics 97, 12-17.
93. Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nature reviews Genetics 7, 385-394.
94. Lander, E.S. (1996). The new genomics: global views of biology. Science 274, 536-539.
95. Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science 273, 1516-1517.
96. Collins, F.S., Guyer, M.S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science 278, 1580-1581.
97. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. Science 296, 2225-2229.
98. International HapMap, C. (2003). The International HapMap Project. Nature 426, 789-796.
99. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.
100. McCarthy, M.I., and Hirschhorn, J.N. (2008). Genome-wide association studies: potential next steps on a genetic journey. Human molecular genetics 17, R156-165.
101. Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitziel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Fornage, M., et al. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. American journal of human genetics 94, 223-232.
102. Kanoni, S., Masca, N.G., Stirrups, K.E., Varga, T.V., Warren, H.R., Scott, R.A., Southam, L., Zhang, W., Yaghootkar, H., Muller-Nurasyid, M., et al. (2016). Analysis with the exome array identifies multiple new independent variants in lipid loci. Human molecular genetics 25, 4094-4106.
103. Liu, C., Kraja, A.T., Smith, J.A., Brody, J.A., Franceschini, N., Bis, J.C., Rice, K., Morrison, A.C., Lu, Y., Weiss, S., et al. (2016). Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. Nature genetics 48, 1162-1170.
104. Surendran, P., Drenos, F., Young, R., Warren, H., Cook, J.P., Manning, A.K., Grarup, N., Sim, X., Barnes, D.R., Witkowska, K., et al. (2016). Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. Nature genetics 48, 1151-1161.
105. Wessel, J., Chu, A.Y., Willems, S.M., Wang, S., Yaghootkar, H., Brody, J.A., Dauriz, M., Hivert, M.F., Raghavan, S., Lipovich, L., et al. (2015). Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nature communications 6, 5897.

106. Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y., Schurmann, C., Highland, H.M., et al. (2017). Rare and low-frequency coding variants alter human adult height. Nature 542, 186-190.
107. Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132, 631-644.
108. Pilia, G., Chen, W.M., Scuteri, A., Orru, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS genetics 2, e132.
109. Hoffman, M., Blum, A., Baruch, R., Kaplan, E., and Benjamin, M. (2004). Leukocytes and coronary heart disease. Atherosclerosis 172, 1-6.
110. Boos, C.J., and Lip, G.Y. (2007). Assessment of mean platelet volume in coronary artery disease - what does it mean? Thrombosis research 120, 11-13.
111. Nieswandt, B., Kleinschnitz, C., and Stoll, G. (2011). Ischaemic stroke: a thrombo-inflammatory disease? The Journal of physiology 589, 4115-4123.
112. Ebrahim, S., and Davey Smith, G. (2008). Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? Human genetics 123, 15-33.
113. Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Holm, H., Ding, E.L., Johnson, T., et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380, 572-580.
114. Do, R., Willer, C.J., Schmidt, E.M., Sengupta, S., Gao, C., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., et al. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat Genet 45, 1345-1352.
115. Zeng, S.M., Yankowitz, J., Widness, J.A., and Strauss, R.G. (2001). Etiology of differences in hematocrit between males and females: sequence-based polymorphisms in erythropoietin and its receptor. The journal of gender-specific medicine : JGSM : the official journal of the Partnership for Women's Health at Columbia 4, 35-40.
116. McLaren, C.E., Barton, J.C., Gordeuk, V.R., Wu, L., Adams, P.C., Reboussin, D.M., Speechley, M., Chang, H., Acton, R.T., Harris, E.L., et al. (2007). Determinants and characteristics of mean corpuscular volume and hemoglobin concentration in white HFE C282Y homozygotes in the hemochromatosis and iron overload screening study. American journal of hematology 82, 898-905.
117. Lin, J.P., O'Donnell, C.J., Jin, L., Fox, C., Yang, Q., and Cupples, L.A. (2007). Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study. American journal of hematology 82, 605-610.
118. Menzel, S., Jiang, J., Silver, N., Gallagher, J., Cunningham, J., Surdulescu, G., Lathrop, M., Farrall, M., Spector, T.D., and Thein, S.L. (2007). The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans. Blood 110, 3624-3626.
119. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S., and Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33, 177-182.
120. Lettre, G. (2012). The search for genetic modifiers of disease severity in the beta-hemoglobinopathies. Cold Spring Harbor perspectives in medicine 2.
121. Sankaran, V.G., Ludwig, L.S., Sicinska, E., Xu, J., Bauer, D.E., Eng, J.C., Patterson, H.C., Metcalf, R.A., Natkunam, Y., Orkin, S.H., et al. (2012). Cyclin D3 coordinates the cell cycle during

differentiation to regulate erythrocyte size and number. Genes & development 26, 2075-2087.
122. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature 480, 201-208.
123. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369-375.
124. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476-482.
125. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature methods 7, 248-249.
126. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4, 1073-1081.
127. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., 3rd, et al. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome biology 10, R130.
128. Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. Nature reviews Genetics 10, 184-194.
129. Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
130. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology 28, 1045-1048.
131. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190-1195.
132. Paul, D.S., Albers, C.A., Rendon, A., Voss, K., Stephens, J., van der Harst, P., Chambers, J.C., Soranzo, N., Ouwehand, W.H., and Deloukas, P. (2013). Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. Genome research 23, 1130-1141.
133. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research 40, D930-934.
134. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome research 22, 1748-1759.
135. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M., et al. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. American journal of human genetics 91, 794-808.
136. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Human molecular genetics 22, 2529-2538.

137. Ganesh, S.K., Zakai, N.A., van Rooij, F.J., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet 41, 1191-1198.
138. Gudbjartsson, D.F., Bjornsdottir, U.S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G.M., Thorleifsson, G., Helgadottir, H., Steinthorsdottir, V., Stefansson, H., et al. (2009). Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nature genetics 41, 342-347.
139. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet 42, 210-215.
140. Lo, K.S., Wilson, J.G., Lange, L.A., Folsom, A.R., Galarneau, G., Ganesh, S.K., Grant, S.F., Keating, B.J., McCarroll, S.A., Mohler, E.R., 3rd, et al. (2011). Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. Human genetics 129, 307-317.
141. Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., Lichtner, P., Klopp, N., Stephens, J., Watkins, N.A., et al. (2009). A genome-wide association study identifies three loci associated with mean platelet volume. Am J Hum Genet 84, 66-71.
142. Qayyum, R., Snively, B.M., Ziv, E., Nalls, M.A., Liu, Y., Tang, W., Yanek, L.R., Lange, L., Evans, M.K., Ganesh, S., et al. (2012). A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. PLoS genetics 8, e1002491.
143. Soranzo, N., Spector, T.D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature genetics 41, 1182-1190.
144. Soranzo, N., Rendon, A., Gieger, C., Jones, C.I., Watkins, N.A., Menzel, S., Doring, A., Stephens, J., Prokisch, H., Erber, W., et al. (2009). A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. Blood 113, 3831-3837.
145. Chambers, J.C., Zhang, W., Li, Y., Sehmi, J., Wass, M.N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M.I., Peltonen, L., et al. (2009). Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nat Genet 41, 1170-1172.
146. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., et al. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS genetics 7, e1002108.
147. Crosslin, D.R., McDavid, A., Weston, N., Nelson, S.C., Zheng, X., Hart, E., de Andrade, M., Kullo, I.J., McCarty, C.A., Doheny, K.F., et al. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Human genetics 131, 639-652.
148. Okada, Y., Hirota, T., Kamatani, Y., Takahashi, A., Ohmiya, H., Kumasaka, N., Higasa, K., Yamaguchi-Kabata, Y., Hosono, N., Nalls, M.A., et al. (2011). Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. PLoS genetics 7, e1002067.
149. Li, J., Glessner, J.T., Zhang, H., Hou, C., Wei, Z., Bradfield, J.P., Mentch, F.D., Guo, Y., Kim, C., Xia, Q., et al. (2013). GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Human molecular genetics 22, 1457-1464.

150. Nalls, M.A., Couper, D.J., Tanaka, T., van Rooij, F.J., Chen, M.H., Smith, A.V., Toniolo, D., Zakai, N.A., Yang, Q., Greinacher, A., et al. (2011). Multiple loci are associated with white blood cell phenotypes. PLoS genetics 7, e1002113.
151. Shameer, K., Denny, J.C., Ding, K., Jouni, H., Crosslin, D.R., de Andrade, M., Chute, C.G., Peissig, P., Pacheco, J.A., Li, R., et al. (2013). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Human genetics.
152. Okada, Y., Takahashi, A., Ohmiya, H., Kumasaka, N., Kamatani, Y., Hosono, N., Tsunoda, T., Matsuda, K., Tanaka, T., Kubo, M., et al. (2011). Genome-wide association study for C-reactive protein levels identified pleiotropic associations in the IL6 locus. Human molecular genetics 20, 1224-1231.
153. Monda, K.L., Chen, G.K., Taylor, K.C., Palmer, C., Edwards, T.L., Lange, L.A., Ng, M.C., Adeyemo, A.A., Allison, M.A., Bielak, L.F., et al. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. Nat Genet 45, 690-696.
154. N'Diaye, A., Chen, G.K., Palmer, C.D., Ge, B., Tayo, B., Mathias, R.A., Ding, J., Nalls, M.A., Adeyemo, A., Adoue, V., et al. (2011). Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry. PLoS genetics 7, e1002298.
155. Reich, D., Nalls, M.A., Kao, W.H., Akylbekova, E.L., Tandon, A., Patterson, N., Mullikin, J., Hsueh, W.C., Cheng, C.Y., Coresh, J., et al. (2009). Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS genetics 5, e1000360.
156. Greenburg, A.G. (1996). Pathophysiology of anemia. The American journal of medicine 101, 7S-11S.
157. Skoda, R. (2007). The genetic basis of myeloproliferative disorders. Hematology Am Soc Hematol Educ Program 2007, 1-10.
158. Oh, S.T., and Gotlib, J. (2010). JAK2 V617F and beyond: role of genetics and aberrant signaling in the pathogenesis of myeloproliferative neoplasms. Expert review of hematology 3, 323-337.
159. Vannucchi, A.M., Pieri, L., and Guglielmelli, P. (2011). JAK2 Allele Burden in the Myeloproliferative Neoplasms: Effects on Phenotype, Prognosis and Change with Treatment. Therapeutic advances in hematology 2, 21-32.
160. Hobbs, C.M., Manning, H., Bennett, C., Vasquez, L., Severin, S., Brain, L., Mazharian, A., Guerrero, J.A., Li, J., Soranzo, N., et al. (2013). JAK2V617F leads to intrinsic changes in platelet formation and reactivity in a knock-in mouse model of essential thrombocythemia. Blood 122, 3787-3797.
161. Sankaran, V.G., Xu, J., and Orkin, S.H. (2010). Advances in the understanding of haemoglobin switching. Br J Haematol 149, 181-194.
162. Sankaran, V.G., Lettre, G., Orkin, S.H., and Hirschhorn, J.N. (2010). Modifier genes in Mendelian disorders: the example of hemoglobin disorders. Annals of the New York Academy of Sciences 1214, 47-56.
163. Platt, O.S., Brambilla, D.J., Rosse, W.F., Milner, P.F., Castro, O., Steinberg, M.H., and Klug, P.P. (1994). Mortality in sickle cell disease. Life expectancy and risk factors for early death. N Engl J Med 330, 1639-1644.
164. Platt, O.S., Thorington, B.D., Brambilla, D.J., Milner, P.F., Rosse, W.F., Vichinsky, E., and Kinney, T.R. (1991). Pain in sickle cell disease. Rates and risk factors. N Engl J Med 325, 11-16.

165. Castro, O., Brambilla, D.J., Thorington, B., Reindorf, C.A., Scott, R.B., Gillette, P., Vera, J.C., and Levy, P.S. (1994). The acute chest syndrome in sickle cell disease: incidence and risk factors. The Cooperative Study of Sickle Cell Disease. Blood 84, 643-649.
166. Thein, S.L., and Craig, J.E. (1998). Genetics of Hb F/F cell variance in adults and heterocellular hereditary persistence of fetal hemoglobin. Hemoglobin 22, 401-414.
167. Menzel, S., Garner, C., Gut, I., Matsuda, F., Yamaguchi, M., Heath, S., Foglio, M., Zelenika, D., Boland, A., Rooks, H., et al. (2007). A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. Nat Genet 39, 1197-1199.
168. Thein, S.L., Menzel, S., Peng, X., Best, S., Jiang, J., Close, J., Silver, N., Gerovasilli, A., Ping, C., Yamaguchi, M., et al. (2007). 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.
169. Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V.G., Chen, W., Usala, G., Busonero, F., Maschio, A., Albai, G., et al. (2008). Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proc Natl Acad Sci U S A 105, 1620-1625.
170. Galarneau, G., Palmer, C.D., Sankaran, V.G., Orkin, S.H., Hirschhorn, J.N., and Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet 42, 1049-1051.
171. Lettre, G., Sankaran, V.G., Bezerra, M.A., Araujo, A.S., Uda, M., Sanna, S., Cao, A., Schlessinger, D., Costa, F.F., Hirschhorn, J.N., et al. (2008). 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.
172. Nuinoon, M., Makarasara, W., Mushiroda, T., Setianingsih, I., Wahidiyat, P.A., Sripichai, O., Kumasaka, N., Takahashi, A., Svasti, S., Munkongdee, T., et al. (2009). A genome-wide association identified the common genetic variants influence disease severity in beta(0)-thalassemia/hemoglobin E. Human genetics.
173. Sankaran, V.G., Menne, T.F., Xu, J., Akie, T.E., Lettre, G., Van Handel, B., Mikkola, H.K., Hirschhorn, J.N., Cantor, A.B., and Orkin, S.H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science 322, 1839-1842.
174. Xu, J., Peng, C., Sankaran, V.G., Shao, Z., Esrick, E.B., Chong, B.G., Ippolito, G.C., Fujiwara, Y., Ebert, B.L., Tucker, P.W., et al. (2011). Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing. Science 334, 993-996.
175. Bauer, D.E., Kamran, S.C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., Shao, Z., Canver, M.C., Smith, E.C., Pinello, L., et al. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science 342, 253-257.
176. Hardison, R.C., and Blobel, G.A. (2013). Genetics. GWAS to therapy by genome edits? Science 342, 206-207.
177. http://omim.org/ (accessed 2013/11/18).
178. Yusen, R.D., Christie, J.D., Edwards, L.B., Kucheryavaya, A.Y., Benden, C., Dipchand, A.I., Dobbels, F., Kirk, R., Lund, L.H., Rahmel, A.O., et al. (2013). The Registry of the International Society for Heart and Lung Transplantation: Thirtieth Adult Lung and Heart-Lung Transplant Report--2013; focus theme: age. In The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation. pp 965-978.
179. Michels, V.V., Driscoll, D.J., and Miller, F.A., Jr. (1985). Familial aggregation of idiopathic dilated cardiomyopathy. The American journal of cardiology 55, 1232-1233.

180. Goerss, J.B., Michels, V.V., Burnett, J., Driscoll, D.J., Miller, F., Rodeheffer, R., Tajik, A.J., and Schaid, D. (1995). Frequency of familial dilated cardiomyopathy. European heart journal 16 Suppl O, 2-4.
181. Mestroni, L., Rocco, C., Gregori, D., Sinagra, G., Di Lenarda, A., Miocic, S., Vatta, M., Pinamonti, B., Muntoni, F., Caforio, A.L., et al. (1999). Familial dilated cardiomyopathy: evidence for genetic and phenotypic heterogeneity. Heart Muscle Disease Study Group. J Am Coll Cardiol 34, 181-190.
182. Kimura, A. (2011). Contribution of genetic factors to the pathogenesis of dilated cardiomyopathy: the cause of dilated cardiomyopathy: genetic or acquired? (genetic-side). Circ J 75, 1756-1765.
183. Ackerman, M.J., Priori, S.G., Willems, S., Berul, C., Brugada, R., Calkins, H., Camm, A.J., Ellinor, P.T., Gollob, M., Hamilton, R., et al. (2011). HRS/EHRA expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies. Europace : journal of the European Society of Cardiology 13, 1077-1109.
184. Gollob, M.H., Blier, L., Brugada, R., Champagne, J., Chauhan, V., Connors, S., Gardner, M., Green, M.S., Gow, R., Hamilton, R., et al. (2011). Recommendations for the use of genetic testing in the clinical evaluation of inherited cardiac arrhythmias associated with sudden cardiac death: Canadian Cardiovascular Society/Canadian Heart Rhythm Society joint position paper. The Canadian journal of cardiology 27, 232-245.
185. Limongelli, G., Elliott, P., Charron, P., Mogensen, J., and McKeown, P.P. (2012). Approaching genetic testing in cardiomyopathies. ESC Council for Cardiology Practice.
186. Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J., et al. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. The New England journal of medicine 363, 2220-2227.
187. Sankaran, V.G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J.A., Beggs, A.H., Sieff, C.A., Orkin, S.H., Nathan, D.G., Lander, E.S., et al. (2012). Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. J Clin Invest 122, 2439-2443.
188. Mestroni, L., Maisch, B., McKenna, W.J., Schwartz, K., Charron, P., Rocco, C., Tesson, F., Richter, A., Wilke, A., and Komajda, M. (1999). Guidelines for the study of familial dilated cardiomyopathies. Collaborative Research Group of the European Human and Capital Mobility Project on Familial Dilated Cardiomyopathy. European heart journal 20, 93-102.
189. Moreau, C., Vezina, H., Yotova, V., Hamon, R., de Knijff, P., Sinnett, D., and Labuda, D. (2009). Genetic heterogeneity in regional populations of Quebec--parental lineages in the Gaspe Peninsula. American journal of physical anthropology 139, 512-522.
190. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics 30, 97-101.
191. (2013). R: A language and Environment for Statistical Computing; R Core Team. R Foundation for Statistical Computing**.** Vienna, Austria. In. (R core Team.
192. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81.
193. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311.
194. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A., and Genomes Project, C. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

195. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.
196. Yancy, C.W., Jessup, M., Bozkurt, B., Butler, J., Casey, D.E., Jr., Drazner, M.H., Fonarow, G.C., Geraci, S.A., Horwich, T., Januzzi, J.L., et al. (2013). 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 62, e147-239.
197. Villard, E., Perret, C., Gary, F., Proust, C., Dilanian, G., Hengstenberg, C., Ruppert, V., Arbustini, E., Wichter, T., Germain, M., et al. (2011). A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy. European heart journal 32, 1065-1076.
198. Brauch, K.M., Karst, M.L., Herron, K.J., de Andrade, M., Pellikka, P.A., Rodeheffer, R.J., Michels, V.V., and Olson, T.M. (2009). Mutations in ribonucleic acid binding protein gene cause familial dilated cardiomyopathy. J Am Coll Cardiol 54, 930-941.
199. Wells, Q.S., Becker, J.R., Su, Y.R., Mosley, J.D., Weeke, P., D'Aoust, L., Ausborn, N.L., Ramirez, A.H., Pfotenhauer, J.P., Naftilan, A.J., et al. (2013). Whole exome sequencing identifies a causal RBM20 mutation in a large pedigree with familial dilated cardiomyopathy. Circulation Cardiovascular genetics 6, 317-326.
200. Hishiya, A., Kitazawa, T., and Takayama, S. (2010). BAG3 and Hsc70 interact with actin capping protein CapZ to maintain myofibrillar integrity under mechanical stress. Circulation research 107, 1220-1231.
201. Norton, N., Li, D., Rieder, M.J., Siegfried, J.D., Rampersaud, E., Zuchner, S., Mangos, S., Gonzalez-Quintana, J., Wang, L., McGee, S., et al. (2011). Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy. American journal of human genetics 88, 273-282.
202. Feldman, A.M., Begay, R.L., Knezevic, T., Myers, V.D., Slavov, D.B., Zhu, W., Gowan, K., Graw, S.L., Jones, K.L., Tilley, D.G., et al. (2014). Decreased Levels of BAG3 in a Family With a Rare Variant and in Idiopathic Dilated Cardiomyopathy. Journal of cellular physiology.
203. Campbell, N., Sinagra, G., Jones, K.L., Slavov, D., Gowan, K., Merlo, M., Carniel, E., Fain, P.R., Aragona, P., Di Lenarda, A., et al. (2013). Whole exome sequencing identifies a troponin T mutation hot spot in familial dilated cardiomyopathy. PloS one 8, e78104.
204. Akalin, N., Zietkiewicz, E., Makalowski, W., and Labuda, D. (1994). Are CpG sites mutation hot spots in the dystrophin gene? Human molecular genetics 3, 1425-1426.
205. Gerull, B., Atherton, J., Geupel, A., Sasse-Klaassen, S., Heuser, A., Frenneaux, M., McNabb, M., Granzier, H., Labeit, S., and Thierfelder, L. (2006). Identification of a novel frameshift mutation in the giant muscle filament titin in a large Australian family with dilated cardiomyopathy. J Mol Med (Berl) 84, 478-483.
206. Norton, N., Robertson, P.D., Rieder, M.J., Zuchner, S., Rampersaud, E., Martin, E., Li, D., Nickerson, D.A., Hershberger, R.E., National Heart, L., et al. (2012). Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. Circulation Cardiovascular genetics 5, 167-174.
207. Pugh, T.J., Kelly, M.A., Gowrisankar, S., Hynes, E., Seidman, M.A., Baxter, S.M., Bowser, M., Harrison, B., Aaron, D., Mahanta, L.M., et al. (2014). The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. Genetics in medicine : official journal of the American College of Medical Genetics.

208. Yusen, R.D., Christie, J.D., Edwards, L.B., Kucheryavaya, A.Y., Benden, C., Dipchand, A.I., Dobbels, F., Kirk, R., Lund, L.H., Rahmel, A.O., et al. (2013). The Registry of the International Society for Heart and Lung Transplantation: Thirtieth Adult Lung and Heart-Lung Transplant Report--2013; focus theme: age. The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation 32, 965-978.
209. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.
210. Popowicz, G.M., Schleicher, M., Noegel, A.A., and Holak, T.A. (2006). Filamins: promiscuous organizers of the cytoskeleton. Trends in biochemical sciences 31, 411-419.
211. Selcen, D. (2010). Myofibrillar myopathies. Current opinion in neurology 23, 477-481.
212. van der Flier, A., and Sonnenberg, A. (2001). Structural and functional aspects of filamins. Biochimica et biophysica acta 1538, 99-117.
213. van der Ven, P.F., Obermann, W.M., Lemke, B., Gautel, M., Weber, K., and Furst, D.O. (2000). Characterization of muscle filamin isoforms suggests a possible role of gamma-filamin/ABP-L in sarcomeric Z-disc formation. Cell motility and the cytoskeleton 45, 149-162.
214. Vorgerd, M., van der Ven, P.F., Bruchertseifer, V., Lowe, T., Kley, R.A., Schroder, R., Lochmuller, H., Himmel, M., Koehler, K., Furst, D.O., et al. (2005). A mutation in the dimerization domain of filamin c causes a novel type of autosomal dominant myofibrillar myopathy. American journal of human genetics 77, 297-304.
215. Kley, R.A., Hellenbroich, Y., van der Ven, P.F., Furst, D.O., Huebner, A., Bruchertseifer, V., Peters, S.A., Heyer, C.M., Kirschner, J., Schroder, R., et al. (2007). Clinical and morphological phenotype of the filamin myopathy: a study of 31 German patients. Brain : a journal of neurology 130, 3250-3264.
216. Shatunov, A., Olive, M., Odgerel, Z., Stadelmann-Nessler, C., Irlbacher, K., van Landeghem, F., Bayarsaikhan, M., Lee, H.S., Goudeau, B., Chinnery, P.F., et al. (2009). In-frame deletion in the seventh immunoglobulin-like repeat of filamin C in a family with myofibrillar myopathy. European journal of human genetics : EJHG 17, 656-663.
217. Luan, X., Hong, D., Zhang, W., Wang, Z., and Yuan, Y. (2010). A novel heterozygous deletion-insertion mutation (2695-2712 del/GTTTGT ins) in exon 18 of the filamin C gene causes filaminopathy in a large Chinese family. Neuromuscular disorders : NMD 20, 390-396.
218. Furst, D.O., Goldfarb, L.G., Kley, R.A., Vorgerd, M., Olive, M., and van der Ven, P.F. (2013). Filamin C-related myopathies: pathology and mechanisms. Acta neuropathologica 125, 33-46.
219. Valdes-Mas, R., Gutierrez-Fernandez, A., Gomez, J., Coto, E., Astudillo, A., Puente, D.A., Reguero, J.R., Alvarez, V., Moris, C., Leon, D., et al. (2014). Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy. Nature communications 5, 5326.
220. Brodehl, A., Ferrier, R.A., Hamilton, S.J., Greenway, S.C., Brundler, M.A., Yu, W., Gibson, W.T., McKinnon, M.L., McGillivray, B., Alvarez, N., et al. (2016). Mutations in FLNC are Associated with Familial Restrictive Cardiomyopathy. Human mutation 37, 269-279.
221. Begay, R.L., Tharp, C.A., Martin, A., Graw, S.L., Sinagra, G., Miani, D., Sweet, M.E., Slavov, D.B., Stafford, N., Zeller, M.J., et al. (2016). FLNC Gene Splice Mutations Cause Dilated Cardiomyopathy. JACC Basic to translational science 1, 344-359.
222. Ortiz-Genga, M.F., Cuenca, S., Dal Ferro, M., Zorio, E., Salgado-Aranda, R., Climent, V., Padron-Barthe, L., Duro-Aguado, I., Jimenez-Jaimez, J., Hidalgo-Olivares, V.M., et al. (2016). Truncating FLNC Mutations Are Associated With High-Risk Dilated and Arrhythmogenic Cardiomyopathies. Journal of the American College of Cardiology 68, 2440-2451.

223. Assomull, R.G., Prasad, S.K., Lyne, J., Smith, G., Burman, E.D., Khan, M., Sheppard, M.N., Poole-Wilson, P.A., and Pennell, D.J. (2006). Cardiovascular magnetic resonance, fibrosis, and prognosis in dilated cardiomyopathy. Journal of the American College of Cardiology 48, 1977-1985.
224. Wu, K.C., Weiss, R.G., Thiemann, D.R., Kitagawa, K., Schmidt, A., Dalal, D., Lai, S., Bluemke, D.A., Gerstenblith, G., Marban, E., et al. (2008). Late gadolinium enhancement by cardiovascular magnetic resonance heralds an adverse prognosis in nonischemic cardiomyopathy. Journal of the American College of Cardiology 51, 2414-2421.
225. Wu, T.J., Ong, J.J., Hwang, C., Lee, J.J., Fishbein, M.C., Czer, L., Trento, A., Blanche, C., Kass, R.M., Mandel, W.J., et al. (1998). Characteristics of wave fronts during ventricular fibrillation in human hearts with dilated cardiomyopathy: role of increased fibrosis in the generation of reentry. Journal of the American College of Cardiology 32, 187-196.
226. Nazarian, S., Bluemke, D.A., Lardo, A.C., Zviman, M.M., Watkins, S.P., Dickfeld, T.L., Meininger, G.R., Roguin, A., Calkins, H., Tomaselli, G.F., et al. (2005). Magnetic resonance assessment of the substrate for inducible ventricular tachycardia in nonischemic cardiomyopathy. Circulation 112, 2821-2825.
227. Iles, L., Pfluger, H., Lefkovits, L., Butler, M.J., Kistler, P.M., Kaye, D.M., and Taylor, A.J. (2011). Myocardial fibrosis predicts appropriate device therapy in patients with implantable cardioverter-defibrillators for primary prevention of sudden cardiac death. Journal of the American College of Cardiology 57, 821-828.
228. Ulker, P., Sati, L., Celik-Ozenci, C., Meiselman, H.J., and Baskurt, O.K. (2009). Mechanical stimulation of nitric oxide synthesizing mechanisms in erythrocytes. Biorheology 46, 121-132.
229. Jiang, N., Tan, N.S., Ho, B., and Ding, J.L. (2007). Respiratory protein-generated reactive oxygen species as an antimicrobial strategy. Nat Immunol 8, 1114-1122.
230. Schnabel, R.B., Baumert, J., Barbalic, M., Dupuis, J., Ellinor, P.T., Durda, P., Dehghan, A., Bis, J.C., Illig, T., Morrison, A.C., et al. (2010). Duffy antigen receptor for chemokines (Darc) polymorphism regulates circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators. Blood 115, 5289-5299.
231. Colin, Y., Le Van Kim, C., and El Nemer, W. (2014). Red cell adhesion in human diseases. Curr Opin Hematol 21, 186-192.
232. Whelihan, M.F., and Mann, K.G. (2013). The role of the red cell membrane in thrombin generation. Thromb Res 131, 377-382.
233. Brugnara, C. (2003). Iron deficiency and erythropoiesis: new diagnostic approaches. Clin Chem 49, 1573-1578.
234. Huang, Y.L., Hu, Z.D., Liu, S.J., Sun, Y., Qin, Q., Qin, B.D., Zhang, W.W., Zhang, J.R., Zhong, R.Q., and Deng, A.M. (2014). Prognostic value of red blood cell distribution width for patients with heart failure: a systematic review and meta-analysis of cohort studies. PLoS One 9, e104861.
235. Nada, A.M. (2015). Red cell distribution width in type 2 diabetic patients. Diabetes Metab Syndr Obes 8, 525-533.
236. Zalawadiya, S.K., Zmily, H., Farah, J., Daifallah, S., Ali, O., and Ghali, J.K. (2011). Red cell distribution width and mortality in predominantly African-American population with decompensated heart failure. J Card Fail 17, 292-298.
237. Zalawadiya, S.K., Veeranna, V., Panaich, S.S., and Afonso, L. (2012). Red cell distribution width and risk of peripheral artery disease: analysis of National Health and Nutrition Examination Survey 1999-2004. Vasc Med 17, 155-163.

238. Patel, K.V., Semba, R.D., Ferrucci, L., Newman, A.B., Fried, L.P., Wallace, R.B., Bandinelli, S., Phillips, C.S., Yu, B., Connelly, S., et al. (2010). Red cell distribution width and mortality in older adults: a meta-analysis. J Gerontol A Biol Sci Med Sci 65, 258-265.
239. Patel, H.H., Patel, H.R., and Higgins, J.M. (2015). Modulation of red blood cell population dynamics is a fundamental homeostatic response to disease. Am J Hematol 90, 422-428.
240. Whitfield, J.B., and Martin, N.G. (1985). Genetic and environmental influences on the size and number of cells in the blood. Genet Epidemiol 2, 133-144.
241. Evans, D.M., Frazer, I.H., and Martin, N.G. (1999). Genetic and environmental causes of variation in basal levels of blood cells. Twin Res 2, 250-257.
242. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum Mol Genet 22, 2529-2538.
243. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dube, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nature genetics 46, 629-634.
244. Tajuddin, S.M., Schick, U.M., Eicher, J.D., Chami, N., Giri, A., Brody, J.A., Hill, W.D., Kacprowski, T., Li, J., Lyytikainen, L.P., et al. (2016). Large-Scale Exome-wide Association Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases. American journal of human genetics 99, 22-39.
245. Eicher, J.D., Chami, N., Kacprowski, T., Nomura, A., Chen, M.H., Yanek, L.R., Tajuddin, S.M., Schick, U.M., Slater, A.J., Pankratz, N., et al. (2016). Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 Individuals. American journal of human genetics 99, 40-55.
246. Grove, M.L., Yu, B., Cochran, B.J., Haritunians, T., Bis, J.C., Taylor, K.D., Hansen, M., Borecki, I.B., Cupples, L.A., Fornage, M., et al. (2013). Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PloS one 8, e68095.
247. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.
248. Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E., Magi, R., Ferreira, T., Fall, T., Graff, M., Justice, A.E., et al. (2014). Quality control and conduct of genome-wide association meta-analyses. Nature protocols 9, 1192-1212.
249. Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O.L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. Nature genetics 46, 200-204.
250. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 89, 82-93.
251. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. American journal of human genetics 86, 832-838.
252. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190-2191.
253. Lessard, S., Beaudoin, M., Benkirane, K., and Lettre, G. (2015). Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors. Genome Med 7, 1.

254. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511.
255. Zhang, X., Gierman, H.J., Levy, D., Plump, A., Dobrin, R., Goring, H.H., Curran, J.E., Johnson, M.P., Blangero, J., Kim, S.K., et al. (2014). Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. BMC genomics 15, 532.
256. Astner, I., Schulze, J.O., van den Heuvel, J., Jahn, D., Schubert, W.D., and Heinz, D.W. (2005). Crystal structure of 5-aminolevulinate synthase, the first enzyme of heme biosynthesis, and its link to XLSA in humans. EMBO J 24, 3166-3177.
257. Halpain, S., and Dehmelt, L. (2006). The MAP1 family of microtubule-associated proteins. Genome Biol 7, 224.
258. Takei, Y., Kikkawa, Y.S., Atapour, N., Hensch, T.K., and Hirokawa, N. (2015). Defects in Synaptic Plasticity, Reduced NMDA-Receptor Transport, and Instability of Postsynaptic Density Proteins in Mice Lacking Microtubule-Associated Protein 1A. J Neurosci 35, 15539-15554.
259. Mukherjee, S. Multivariate analysis of whole exome sequence data identifies rare variants with pleiotropic effects on obesity-related metabolic traits in 31,000 participants of the Regeneron Genetics Center – Geisinger MyCode collaborative project – DiscovEHR. American Society of Human Genetics (ASHG) Conference 2015.
260. Consortium, G.T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648-660.
261. Dehghan, A., Dupuis, J., Barbalic, M., Bis, J.C., Eiriksdottir, G., Lu, C., Pellikka, N., Wallaschofski, H., Kettunen, J., Henneman, P., et al. (2011). Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123, 731-738.
262. Global Lipids Genetics, C., Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nature genetics 45, 1274-1283.
263. Taylor, K.C., Lange, L.A., Zabaneh, D., Lange, E., Keating, B.J., Tang, W., Smith, N.L., Delaney, J.A., Kumari, M., Hingorani, A., et al. (2011). A gene-centric association scan for Coagulation Factor VII levels in European and African Americans: the Candidate Gene Association Resource (CARe) Consortium. Hum Mol Genet 20, 3525-3534.
264. de Vries, P.S., Chasman, D.I., Sabater-Lleal, M., Chen, M.H., Huffman, J.E., Steri, M., Tang, W., Teumer, A., Marioni, R.E., Grossmann, V., et al. (2016). A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. Hum Mol Genet 25, 358-370.
265. Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F., Chia, K.S., Dimas, A.S., Hassanali, N., et al. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nature genetics 43, 984-989.
266. Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nature genetics 45, 580-585.
267. Pradel, L.C., Vanhille, L., and Spicuglia, S. (2015). [The European Blueprint project: towards a full epigenome characterization of the immune system]. Medecine sciences : M/S 31, 236-238.
268. Aulchenko, Y.S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I.M., Pramstaller, P.P., Penninx, B.W., Janssens, A.C., Wilson, J.F., Spector, T., et al. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41, 47-55.

269. Kullo, I.J., Ding, K., Shameer, K., McCarty, C.A., Jarvik, G.P., Denny, J.C., Ritchie, M.D., Ye, Z., Crosslin, D.R., Chisholm, R.L., et al. (2011). Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. Am J Hum Genet 89, 131-138.
270. Dobbeling, U. (1997). The effects of cyclosporin A on V(D)J recombination activity. Scand J Immunol 45, 494-498.
271. Zarza, R., Alvarez, R., Pujades, A., Nomdedeu, B., Carrera, A., Estella, J., Remacha, A., Sanchez, J.M., Morey, M., Cortes, T., et al. (1998). Molecular characterization of the PK-LR gene in pyruvate kinase deficient Spanish patients. Red Cell Pathology Group of the Spanish Society of Haematology (AEHH). Br J Haematol 103, 377-382.
272. Valentini, G., Chiarelli, L.R., Fortin, R., Dolzan, M., Galizzi, A., Abraham, D.J., Wang, C., Bianchi, P., Zanella, A., and Mattevi, A. (2002). Structure and function of human erythrocyte pyruvate kinase. Molecular basis of nonspherocytic hemolytic anemia. J Biol Chem 277, 23807-23814.
273. Van Sligtenhorst, I., Ding, Z.M., Shi, Z.Z., Read, R.W., Hansen, G., and Vogel, P. (2012). Cardiomyopathy in alpha-kinase 3 (ALPK3)-deficient mice. Vet Pathol 49, 131-141.
274. Gu, Y., Nakamura, T., Alder, H., Prasad, R., Canaani, O., Cimino, G., Croce, C.M., and Canaani, E. (1992). The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to Drosophila trithorax, to the AF-4 gene. Cell 71, 701-708.
275. Guastadisegni, M.C., Lonoce, A., Impera, L., Di Terlizzi, F., Fugazza, G., Aliano, S., Grasso, R., Cluzeau, T., Raynaud, S., Rocchi, M., et al. (2010). CBFA2T2 and C20orf112: two novel fusion partners of RUNX1 in acute myeloid leukemia. Leukemia 24, 1516-1519.
276. Serbanovic-Canic, J., Cvejic, A., Soranzo, N., Stemple, D.L., Ouwehand, W.H., and Freson, K. (2011). Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its unexpected role in iron uptake. Blood 118, 4967-4976.
277. Okumura, N., Tsuji, K., and Nakahata, T. (1992). Changes in cell surface antigen expressions during proliferation and differentiation of human erythroid progenitors. Blood 80, 642-650.
278. Kiefer, C.R., and Snyder, L.M. (2000). Oxidation and erythrocyte senescence. Curr Opin Hematol 7, 113-116.
279. Nicholson, A.C., Han, J., Febbraio, M., Silversterin, R.L., and Hajjar, D.P. (2001). Role of CD36, the macrophage class B scavenger receptor, in atherosclerosis. Ann N Y Acad Sci 947, 224-228.
280. Elbers, C.C., Guo, Y., Tragante, V., van Iperen, E.P., Lanktree, M.B., Castillo, B.A., Chen, F., Yanek, L.R., Wojczynski, M.K., Li, Y.R., et al. (2012). Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. PLoS One 7, e50198.
281. Ayodo, G., Price, A.L., Keinan, A., Ajwang, A., Otieno, M.F., Orago, A.S., Patterson, N., and Reich, D. (2007). Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am J Hum Genet 81, 234-242.
282. Aitman, T.J., Cooper, L.D., Norsworthy, P.J., Wahid, F.N., Gray, J.K., Curtis, B.R., McKeigue, P.M., Kwiatkowski, D., Greenwood, B.M., Snow, R.W., et al. (2000). Malaria susceptibility and CD36 mutation. Nature 405, 1015-1016.
283. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am J Hum Genet 89, 368-381.
284. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.

285. Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F., et al. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet 47, 1272-1281.
286. Schmidt, V.A., Scudder, L., Devoe, C.E., Bernards, A., Cupit, L.D., and Bahou, W.F. (2003). IQGAP2 functions as a GTP-dependent effector protein in thrombin-induced platelet cytoskeletal reorganization. Blood 101, 3021-3028.
287. Hennekens, C.H., Dyken, M.L., and Fuster, V. (1997). Aspirin as a therapeutic agent in cardiovascular disease: a statement for healthcare professionals from the American Heart Association. Circulation 96, 2751-2753.
288. Sutcliffe, P., Connock, M., Gurung, T., Freeman, K., Johnson, S., Kandala, N.B., Grove, A., Gurung, B., Morrow, S., and Clarke, A. (2013). Aspirin for prophylactic use in the primary prevention of cardiovascular disease and cancer: a systematic review and overview of reviews. Health technology assessment 17, 1-253.
289. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide Association Study of Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans. American journal of human genetics 98, 229-242.
290. Shameer, K., Denny, J.C., Ding, K., Jouni, H., Crosslin, D.R., de Andrade, M., Chute, C.G., Peissig, P., Pacheco, J.A., Li, R., et al. (2014). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Human genetics 133, 95-109.
291. Kim, Y.K., Oh, J.H., Kim, Y.J., Hwang, M.Y., Moon, S., Low, S.K., Takahashi, A., Matsuda, K., Kubo, M., Lee, J., et al. (2015). Influence of Genetic Variants in EGF and Other Genes on Hematological Traits in Korean Populations by a Genome-Wide Approach. BioMed research international 2015, 914965.
292. Oh, J.H., Kim, Y.K., Moon, S., Kim, Y.J., and Kim, B.J. (2014). Genome-wide association study identifies candidate Loci associated with platelet count in koreans. Genomics & informatics 12, 225-230.
293. Nurnberg, S.T., Rendon, A., Smethurst, P.A., Paul, D.S., Voss, K., Thon, J.N., Lloyd-Jones, H., Sambrook, J.G., Tijssen, M.R., HaemGen, C., et al. (2012). A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. Blood 120, 4859-4868.
294. Johnson, A.D. (2011). The genetics of common variation affecting platelet development, function and pharmaceutical targeting. Journal of thrombosis and haemostasis : JTH 9 Suppl 1, 246-257.
295. Chami, N., Chen, M.H., Slater, A.J., Eicher, J.D., Evangelou, E., Tajuddin, S.M., Love-Gregory, L., Kacprowski, T., Schick, U.M., Nomura, A., et al. (2016). Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. American journal of human genetics 99, 8-21.
296. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics 88, 76-82.
297. Johnson, A.D., Yanek, L.R., Chen, M.H., Faraday, N., Larson, M.G., Tofler, G., Lin, S.J., Kraja, A.T., Province, M.A., Yang, Q., et al. (2010). Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. Nature genetics 42, 608-613.

298. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 46, 310-315.
299. Morrison, A.C., Fu, Y.P., O'Donnell, C.J., the Cohorts for, H., Aging Research in Genomic Epidemiology Consortium Subclinical, A., and Group, C.H.D.W. (2016). Variants in ANGPTL4 and the Risk of Coronary Artery Disease. The New England journal of medicine 375, 2303.
300. Eicher, J.D., Wakabayashi, Y., Vitseva, O., Esa, N., Yang, Y., Zhu, J., Freedman, J.E., McManus, D.D., and Johnson, A.D. (2016). Characterization of the platelet transcriptome by RNA sequencing in patients with acute myocardial infarction. Platelets 27, 230-239.
301. Consortium, C.A.D., Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., Thompson, J.R., Ingelsson, E., Saleheen, D., Erdmann, J., et al. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nature genetics 45, 25-33.
302. Fehrmann, R.S., Jansen, R.C., Veldink, J.H., Westra, H.J., Arends, D., Bonder, M.J., Fu, J., Deelen, P., Groen, H.J., Smolonska, A., et al. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS genetics 7, e1002197.
303. Sandomenico, A., Monti, S.M., Marasco, D., Dathan, N., Palumbo, R., Saviano, M., and Ruvo, M. (2009). IgE-binding properties and selectivity of peptide mimics of the FcvarepsilonRI binding site. Molecular immunology 46, 3300-3309.
304. Mackay, G.A., Hulett, M.D., Cook, J.P., Trist, H.M., Henry, A.J., McDonnell, J.M., Beavil, A.J., Beavil, R.L., Sutton, B.J., Hogarth, P.M., et al. (2002). Mutagenesis within human FcepsilonRIalpha differentially affects human and murine IgE binding. Journal of immunology 168, 1787-1795.
305. Cook, J.P., Henry, A.J., McDonnell, J.M., Owens, R.J., Sutton, B.J., and Gould, H.J. (1997). Identification of contact residues in the IgE binding site of human FcepsilonRIalpha. Biochemistry 36, 15579-15588.
306. Garman, S.C., Kinet, J.P., and Jardetzky, T.S. (1999). The crystal structure of the human high-affinity IgE receptor (Fc epsilon RI alpha). Annual review of immunology 17, 973-976.
307. Granada, M., Wilk, J.B., Tuzova, M., Strachan, D.P., Weidinger, S., Albrecht, E., Gieger, C., Heinrich, J., Himes, B.E., Hunninghake, G.M., et al. (2012). A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. The Journal of allergy and clinical immunology 129, 840-845 e821.
308. Page, C., and Pitchford, S. (2014). Platelets and allergic inflammation. Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology 44, 901-913.
309. Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. Jama 312, 1880-1887.
310. Nurden, A.T., Pillois, X., Fiore, M., Heilig, R., and Nurden, P. (2011). Glanzmann thrombasthenia-like syndromes associated with Macrothrombocytopenias and mutations in the genes encoding the alphaIIbbeta3 integrin. Seminars in thrombosis and hemostasis 37, 698-706.
311. Obinata, D., Takayama, K., Urano, T., Murata, T., Ikeda, K., Horie-Inoue, K., Ouchi, Y., Takahashi, S., and Inoue, S. (2012). ARFGAP3, an androgen target gene, promotes prostate cancer cell proliferation and migration. International journal of cancer 130, 2240-2248.
312. Kartberg, F., Asp, L., Dejgaard, S.Y., Smedh, M., Fernandez-Rodriguez, J., Nilsson, T., and Presley, J.F. (2010). ARFGAP2 and ARFGAP3 are essential for COPI coat assembly on the Golgi membrane of living cells. The Journal of biological chemistry 285, 36709-36720.

313. Weimer, C., Beck, R., Eckert, P., Reckmann, I., Moelleken, J., Brugger, B., and Wieland, F. (2008). Differential roles of ArfGAP1, ArfGAP2, and ArfGAP3 in COPI trafficking. The Journal of cell biology 183, 725-735.
314. Begonja, A.J., Pluthero, F.G., Suphamungmee, W., Giannini, S., Christensen, H., Leung, R., Lo, R.W., Nakamura, F., Lehman, W., Plomann, M., et al. (2015). FlnA binding to PACSIN2 F-BAR domain regulates membrane tubulation in megakaryocytes and platelets. Blood 126, 80-88.
315. Li, T., Shi, Y., Wang, P., Guachalla, L.M., Sun, B., Joerss, T., Chen, Y.S., Groth, M., Krueger, A., Platzer, M., et al. (2015). Smg6/Est1 licenses embryonic stem cell differentiation via nonsense-mediated mRNA decay. The EMBO journal 34, 1630-1647.
316. Butler, M., Morel, A.S., Jordan, W.J., Eren, E., Hue, S., Shrimpton, R.E., and Ritter, M.A. (2007). Altered expression and endocytic function of CD205 in human dendritic cells, and detection of a CD205-DCL-1 fusion protein upon dendritic cell maturation. Immunology 120, 362-371.
317. Cao, L., Shi, X., Chang, H., Zhang, Q., and He, Y. (2015). pH-Dependent recognition of apoptotic and necrotic cells by the human dendritic cell receptor DEC205. Proceedings of the National Academy of Sciences of the United States of America 112, 7237-7242.
318. Catani, L., Sollazzo, D., Ricci, F., Polverelli, N., Palandri, F., Baccarani, M., Vianelli, N., and Lemoli, R.M. (2011). The CD47 pathway is deregulated in human immune thrombocytopenia. Experimental hematology 39, 486-494.
319. Olsson, M., Bruhns, P., Frazier, W.A., Ravetch, J.V., and Oldenborg, P.A. (2005). Platelet homeostasis is regulated by platelet expression of CD47 under normal conditions and in passive immune thrombocytopenia. Blood 105, 3577-3582.
320. Yamao, T., Noguchi, T., Takeuchi, O., Nishiyama, U., Morita, H., Hagiwara, T., Akahori, H., Kato, T., Inagaki, K., Okazawa, H., et al. (2002). Negative regulation of platelet clearance and of the macrophage phagocytic response by the transmembrane glycoprotein SHPS-1. The Journal of biological chemistry 277, 39833-39839.
321. Qayyum, R., Becker, L.C., Becker, D.M., Faraday, N., Yanek, L.R., Leal, S.M., Shaw, C., Mathias, R., Suktitipat, B., and Bray, P.F. (2015). Genome-wide association study of platelet aggregation in African Americans. BMC genetics 16, 58.
322. Lewis, J.P., Ryan, K., O'Connell, J.R., Horenstein, R.B., Damcott, C.M., Gibson, Q., Pollin, T.I., Mitchell, B.D., Beitelshees, A.L., Pakzy, R., et al. (2013). Genetic variation in PEAR1 is associated with platelet aggregation and cardiovascular outcomes. Circulation Cardiovascular genetics 6, 184-192.
323. Eicher, J.D., Xue, L., Ben-Shlomo, Y., Beswick, A.D., and Johnson, A.D. (2016). Replication and hematological characterization of human platelet reactivity genetic associations in men from the Caerphilly Prospective Study (CaPS). Journal of thrombosis and thrombolysis 41, 343-350.
324. Stone, J.C. (2011). Regulation and Function of the RasGRP Family of Ras Activators in Blood Cells. Genes & cancer 2, 320-334.
325. Kauskot, A., Vandenbriele, C., Louwette, S., Gijsbers, R., Tousseyn, T., Freson, K., Verhamme, P., and Hoylaerts, M.F. (2013). PEAR1 attenuates megakaryopoiesis via control of the PI3K/PTEN pathway. Blood 121, 5208-5217.
326. Karpatkin, S. (1978). Heterogeneity of human platelets. VI. Correlation of platelet function with platelet volume. Blood 51, 307-316.
327. Bertin, A., Mahaney, M.C., Cox, L.A., Rogers, J., VandeBerg, J.L., Brugnara, C., and Platt, O.S. (2007). Quantitative trait loci for peripheral blood cell counts: a study in baboons. Mammalian genome : official journal of the International Mammalian Genome Society 18, 361-372.

328. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics 45, 1238-1243.
329. Panova-Noeva, M., Schulz, A., Hermanns, M.I., Grossmann, V., Pefani, E., Spronk, H.M., Laubert-Reh, D., Binder, H., Beutel, M., Pfeiffer, N., et al. (2016). Sex-specific differences in genetic and nongenetic determinants of mean platelet volume: results from the Gutenberg Health Study. Blood 127, 251-259.
330. Sloan, A., Gona, P., and Johnson, A.D. (2015). Cardiovascular correlates of platelet count and volume in the Framingham Heart Study. Annals of epidemiology 25, 492-498.
331. Daly, M.E. (2011). Determinants of platelet count in humans. Haematologica 96, 10-13.
332. Moreno-Ayala, R., Schnabel, D., Salas-Vidal, E., and Lomeli, H. (2015). PIAS-like protein Zimp7 is required for the restriction of the zebrafish organizer and mesoderm development. Developmental biology 403, 89-100.
333. Peng, Y., Lee, J., Zhu, C., and Sun, Z. (2010). A novel role for protein inhibitor of activated STAT (PIAS) proteins in modulating the activity of Zimp7, a novel PIAS-like protein, in androgen receptor-mediated transcription. The Journal of biological chemistry 285, 11465-11475.
334. Liu, Y., Lee, J.W., and Ackerman, S.L. (2015). Mutations in the microtubule-associated protein 1A (Map1a) gene cause Purkinje cell degeneration. The Journal of neuroscience : the official journal of the Society for Neuroscience 35, 4587-4598.
335. Ganesh, S.K., Tragante, V., Guo, W., Guo, Y., Lanktree, M.B., Smith, E.N., Johnson, T., Castillo, B.A., Barnard, J., Baumert, J., et al. (2013). Loci influencing blood pressure identified using a cardiovascular gene-centric array. Human molecular genetics 22, 1663-1678.
336. Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S., et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nature genetics 41, 666-676.
337. Bochenek, G., Hasler, R., El Mokhtari, N.E., Konig, I.R., Loos, B.G., Jepsen, S., Rosenstiel, P., Schreiber, S., and Schaefer, A.S. (2013). The large non-coding RNA ANRIL, which is associated with atherosclerosis, periodontitis and several forms of cancer, regulates ADIPOR1, VAMP3 and C11ORF10. Human molecular genetics 22, 4516-4527.
338. Gomes, A.L., Carvalho, T., Serpa, J., Torre, C., and Dias, S. (2010). Hypercholesterolemia promotes bone marrow cell mobilization by perturbing the SDF-1:CXCR4 axis. Blood 115, 3886-3894.
339. Su, Y., Wang, Z., Yang, H., Cao, L., Liu, F., Bai, X., and Ruan, C. (2006). Clinical and molecular genetic analysis of a family with sitosterolemia and co-existing erythrocyte and platelet abnormalities. Haematologica 91, 1392-1395.
340. Wang, Z., Cao, L., Su, Y., Wang, G., Wang, R., Yu, Z., Bai, X., and Ruan, C. (2014). Specific macrothrombocytopenia/hemolytic anemia associated with sitosterolemia. American journal of hematology 89, 320-324.
341. Murphy, A.J., Bijl, N., Yvan-Charvet, L., Welch, C.B., Bhagwat, N., Reheman, A., Wang, Y., Shaw, J.A., Levine, R.L., Ni, H., et al. (2013). Cholesterol efflux in megakaryocyte progenitors suppresses platelet production and thrombocytosis. Nature medicine 19, 586-594.
342. Arimura, T., Ishikawa, T., Nunoda, S., Kawai, S., and Kimura, A. (2011). Dilated cardiomyopathy-associated BAG3 mutations impair Z-disc assembly and enhance sensitivity to apoptosis in cardiomyocytes. Human mutation 32, 1481-1491.

343. Beatham, J., Romero, R., Townsend, S.K., Hacker, T., van der Ven, P.F., and Blanco, G. (2004). Filamin C interacts with the muscular dystrophy KY protein and is abnormally distributed in mouse KY deficient muscle fibres. Human molecular genetics 13, 2863-2874.
344. Holmes, W.B., and Moncman, C.L. (2008). Nebulette interacts with filamin C. Cell motility and the cytoskeleton 65, 130-142.
345. Guergueltcheva, V., Peeters, K., Baets, J., Ceuterick-de Groote, C., Martin, J.J., Suls, A., De Vriendt, E., Mihaylova, V., Chamova, T., Almeida-Souza, L., et al. (2011). Distal myopathy with upper limb predominance caused by filamin C haploinsufficiency. Neurology 77, 2105-2114.
346. Lowe, T., Kley, R.A., van der Ven, P.F., Himmel, M., Huebner, A., Vorgerd, M., and Furst, D.O. (2007). The pathomechanism of filaminopathy: altered biochemical properties explain the cellular phenotype of a protein aggregation myopathy. Human molecular genetics 16, 1351-1358.
347. Hershberger, R.E. (2008). Cardiovascular genetic medicine: evolving concepts, rationale, and implementation. Journal of cardiovascular translational research 1, 137-143.
348. Beaulieu, C.L., Majewski, J., Schwartzentruber, J., Samuels, M.E., Fernandez, B.A., Bernier, F.P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., et al. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. Am J Hum Genet 94, 809-817.
349. Whiffin, N. (2016). Using high-resolution variant frequencies to empower clinical genome interpretation. bioRxiv.
350. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine : official journal of the American College of Medical Genetics 17, 405-424.
351. Walsh, R., Thomson, K.L., Ware, J.S., Funke, B.H., Woodley, J., McGuire, K.J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J.C., et al. (2017). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genetics in medicine : official journal of the American College of Medical Genetics 19, 192-203.
352. Splawski, I., Timothy, K.W., Tateyama, M., Clancy, C.E., Malhotra, A., Beggs, A.H., Cappuccio, F.P., Sagnella, G.A., Kass, R.S., and Keating, M.T. (2002). Variant of SCN5A sodium channel implicated in risk of cardiac arrhythmia. Science 297, 1333-1336.
353. Olson, T.M., Michels, V.V., Ballew, J.D., Reyna, S.P., Karst, M.L., Herron, K.J., Horton, S.C., Rodeheffer, R.J., and Anderson, J.L. (2005). Sodium channel mutations and susceptibility to heart failure and atrial fibrillation. JAMA : the journal of the American Medical Association 293, 447-454.
354. Pankratz, N. (2016). Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. Nature genetics 48, 867-876.
355. Polfus, L.M., Khajuria, R.K., Schick, U.M., Pankratz, N., Pazoki, R., Brody, J.A., Chen, M.H., Auer, P.L., Floyd, J.S., Huang, J., et al. (2016). Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative GFI1B Splice Variants in Human Hematopoiesis. American journal of human genetics 99, 481-488.
356. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415-1429 e1419.

357. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature 526, 68-74.
358. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82-90.
359. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Y, A.R., H, K.F., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics 48, 1443-1448.
360. Jain, D., Hodonsky, C.J., Schick, U.M., Morrison, J.V., Minnerath, S., Brown, L., Schurmann, C., Liu, Y., Auer, P.L., Laurie, C.A., et al. (2017). Genome-Wide Association of White Blood Cell Counts in Hispanic/Latino Americans: The Hispanic Community Health Study/Study of Latinos. Human molecular genetics.
361. Young, A.I., Wauthier, F., and Donnelly, P. (2016). Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. Nature communications 7, 12724.
362. Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature reviews Genetics 12, 628-640.
363. Bielczyk-Maczynska, E., Serbanovic-Canic, J., Ferreira, L., Soranzo, N., Stemple, D.L., Ouwehand, W.H., and Cvejic, A. (2014). A loss of function screen of identified genome-wide association study Loci reveals new genes controlling hematopoiesis. PLoS genetics 10, e1004450.
364. Nakamura, S., Takayama, N., Hirata, S., Seo, H., Endo, H., Ochi, K., Fujita, K., Koike, T., Harimoto, K., Dohda, T., et al. (2014). Expandable megakaryocyte cell lines enable clinically applicable generation of platelets from human induced pluripotent stem cells. Cell stem cell 14, 535-548.
365. Lu, S.J., Feng, Q., Park, J.S., and Lanza, R. (2010). Directed differentiation of red blood cells from human embryonic stem cells. Methods in molecular biology 636, 105-121.
366. Alasoo, K., Martinez, F.O., Hale, C., Gordon, S., Powrie, F., Dougan, G., Mukhopadhyay, S., and Gaffney, D.J. (2015). Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription. Scientific reports 5, 12524.
367. Lopez, D. (2008). Inhibition of PCSK9 as a novel strategy for the treatment of hypercholesterolemia. Drug news & perspectives 21, 323-330.
368. Zemojtel, T., Kohler, S., Mackenroth, L., Jager, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Science translational medicine 6, 252ra123.
369. Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N.P., Roos, C., Hirschhorn, J.N., Berglund, G., Hedblad, B., Groop, L., et al. (2008). Polymorphisms associated with cholesterol and risk of cardiovascular events. The New England journal of medicine 358, 1240-1249.
370. Brautbar, A., Ballantyne, C.M., Lawson, K., Nambi, V., Chambless, L., Folsom, A.R., Willerson, J.T., and Boerwinkle, E. (2009). Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the Atherosclerosis Risk in Communities study. Circulation Cardiovascular genetics 2, 279-285.

371. Goldstein, B.A., Knowles, J.W., Salfati, E., Ioannidis, J.P., and Assimes, T.L. (2014). Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. Frontiers in genetics 5, 254.
372. Plenge, R.M., Scolnick, E.M., and Altshuler, D. (2013). Validating therapeutic targets through human genetics. Nature reviews Drug discovery 12, 581-594.
373. Dotson, W.D., Douglas, M.P., Kolor, K., Stewart, A.C., Bowen, M.S., Gwinn, M., Wulf, A., Anders, H.M., Chang, C.Q., Clyne, M., et al. (2014). Prioritizing genomic applications for action by level of evidence: a horizon-scanning method. Clinical pharmacology and therapeutics 95, 394-402.
374. Evaluation of Genomic Applications in, P., Prevention Working, G., Evaluation of Genomic Applications in, P., and Prevention Working, G. (2013). Recommendations from the EGAPP Working Group: does genomic profiling to assess type 2 diabetes risk improve health outcomes? Genetics in medicine : official journal of the American College of Medical Genetics 15, 612-617.
375. Camilleri, E., Jacquin, L., Paganelli, F., and Bonello, L. (2011). Personalized antiplatelet therapy: review of the latest clinical evidence. Current cardiology reports 13, 296-302.
376. Raghavan, S., and Vassy, J.L. (2014). Do physicians think genomic medicine will be useful for patient care? Personalized Medicing 11, 8.