Université de Montréal

The genetics of red blood cell density, a biomarker of clinical severity in sickle cell disease.

par

Yann Ilboudo

Programme de bio-informatique

Faculté de médecine

Mémoire présenté à la Faculté des études supérieures et postdoctorales

en vue de l'obtention du grade de M.Sc.

en bio-informatique

Déc. 2016

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé:

The genetics of red blood cell density, a biomarker of clinical severity in sickle cell disease.

présenté par
Yann Ilboudo

a été évalué par un jury composé des personnes suivantes :

Sylvie Hamel, Ph. D.
président-rapporteur

Guillaume Lettre, Ph. D.
directeur de recherche

Yves Pastore, MD.
membre du jury

Résumé

L'anémie falciforme est l'une des maladies du sang les plus répandues chez l'homme. Les complications liées à la maladie sont systémiques. Influant virtuellement tous les organes du corps, cette affection provoque des crises de douleurs imprévisibles et aigües dont les complications mènent parfois à la mort. Le processus à travers lequel un globule rouge sain prend la forme d'une faucille est bien décrit dans la littérature; sous désoxygénation, l'eau et les solutés se retirent des globules rouges, la concentration d'hémoglobine S augmente et nous donne des globules rouges denses et déshydratés qui par la suite deviennent falciformes. Les traitements d'aujourd'hui sont pour la plupart expérimentaux et coûteux. De plus, leurs efficacités à long terme varient d'un patient à l'autre. Il est donc impératif de trouver un biomarqueur qui est à la fois abordable et qui améliore la santé des malades de façon systématique. La densité des globules rouges est un biomarqueur largement ignoré par la communauté médicale dans le contexte de la drépanocytose. Aborder l'étude de la sévérité de cette maladie en se concentrant sur la densité des globules rouges nous met en position d'identifier des traitements pour réhydrater les érythrocytes et leur rendre leur forme originale de disque biconcave. Plusieurs études cliniques et physiologiques se sont penchées sur ce biomarqueur sans explorer le volet génétique. Nous avons cherché à éclaircir cet aspect en menant une étude d'association pangénomique et en examinant les séquences exomiques d'individus avec des mesures de densité extrême. Notre étude d'association pangénomique n'a pas conduit à la découverte de nouveau loci, probablement parce que la taille de notre échantillon, et donc notre puissance statistique, était limitée. En revanche, à travers notre approche de priorisation, nous avons découvert un marqueur intronique qui contrôle l'expression d'*ATP2B4*, la protéine principale de transport de calcium dans les hématies. Notre séquençage exomique a identifié deux mutations rares faux-sens chez un même patient; l'une dans *ATP1B2*, un transporteur de Na+/K+, et l'autre dans *SPTB*, le gène du β-spectrin. Ces mutations expliqueraient pourquoi ce patient a le pourcentage de densité le plus élevé parmi tous nos patients séquencés, et pourquoi il vit avec plusieurs complications de la maladie. Finalement, nous avons localisé une mutation faux-sens rare chez deux patients avec un indice élevé de densité de globule rouge, dans *PIEZO1*, le canal ionique mécano-sensitif. La mutation est prédite délétère par deux algorithmes de prédiction de fonction protéique.

Mots-clés: Analyse pangénomique, séquençage d'exome, anémie falciforme, densité des globules rouges, hydratation des hématies, eQTL.

.

Abstract

Sickle cell disease is one of most common blood disorder amongst human. The complications associated with the disease are systemic. They damage virtually all the organs of the body, causing severe, unpredictable pain episodes, which repercussions can eventually lead to death. The process through which a biconcave, healthy red blood cell assumes a crescent-shape is well described in the literature; under deoxygenation, as water and solutes leave erythrocytes the concentration of hemoglobin S increases thus giving us dense dehydrated cells and subsequently sickled cells. Today's current therapies are for the most part experimental, costly, and vary widely in their long-term effectiveness from patient to patient. There is, therefore, a pressing need, to identify a biomarker that is cost-effective and provides positive health outcomes to patients. The density of red blood cell is a biological indicator largely ignored by the medical community in sickle cell disease. Exploring erythrocytes density can facilitate the development of new therapies by targeting channels to rehydrate cells back to their normal shape. Clinical and physiological characterizations of this phenotype exist in many studies, but the genetic characterization is absent. We attempted to elucidate the genetic underpinning of this phenotype, by conducting a genome-wide scan, and examining the whole-exome sequences of individuals with extreme red blood cell density. Our genome-wide association study did not highlight any new loci due to our limited statistical power reflected by the cohort's small sample size. However, our prioritization approach highlighted an intronic variant that controls the expression of *ATP2B4*, the main calcium pump in erythrocytes. Our whole-exome sequencing experiment pointed out two rare missense mutations in the same patient; one in *ATP1B2*, a Na+/K+ transporter, and the other in *SPTB*, the β-spectrin gene. These variants could explain why he has the highest measured density of red blood cells amongst all of our sequenced patients, and why this person experiences several of the disease-related complications. Another rare missense mutation in two patients with elevated levels of dense cells was discovered in *PIEZO1*, the mechanosensitive ion channel. The mutation is predicted to be deleterious by both protein function prediction algorithms.

Keywords: Genome wide association, whole-exome sequencing, sickle cell disease, erythrocyte density, red blood cell hydration, eQTL.

Table of Content

List of tables

List of figures

List of abbreviations

1000G	1000 Genome
ASW	African ancestry in Southwest USA
ATP	Adenosine triphosphate
Ca	Calcium
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
CI	Confidence interval
Cl	Chloride
CO_2	Carbon dioxide
DNA	Deoxyribonucleic acid
DRBC	Dense dehydrated red blood cell
eQTL	Expression quantitative trait loci
GIH	Gujarati Indians in Houston, Texas
GWAS	Genome wide association study
HbA	Adult hemoglobin
HbF	Fetal hemoglobin
HbS	Hemoglobin S
HCT	Hematocrit
HGB	Hemoglobin
HU	Hydroxyurea
HS	Hereditary Spherocytosis
IBD	Identity by descent
IRS	Irreversibly sickled
JPT	Japanese in Tokyo, Japan
K	Potassium
kb	Kilo base
kDa	Kilo dalton
LD	Linkage disequilibrium
LWK	Luhya in Webuye, Kenya
MAF	Minor allele frequency
MCH	Mean cell hemoglobin
MCHC	Mean cell hemoglobin concentration
MCV	Mean corpuscular volume
MDS	Multi dimensional scaling
Mg	Magnesium
MKK	Maasai in Kinyawa, Kenya
MPV	Mean platelet volume
MXL	Mexican ancestry in Los Angeles, California
Na	Sodium
NIH	National institute of health
OR	Odds ratio
PCA	Principal component analysis

PP	Prolonged priapism
SP	Stuttering priapism
QC	Quality control
r^2	Imputation measure of quality of imputation
R^2	Variance explained
RA	Rare nonsense minor allele
RBC	Red blood cell
Retic	Reticulocyte
SCA	Sickle cell anemia
SCD	Sickle cell disease
SE	Standard error
SKAT	Sequence kernel association test
SNP	Single nucleotide polymorphism
TSI	Toscani in Italia
VT	Variable threshold
WBC	White blood cell count
WES	Whole-exome sequencing
YRI	Yoruba in Ibadan, Nigeria

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Guillaume Lettre, for his continuous support, encouragements, and enthusiasm throughout these intense two years. Thank you for giving me the opportunity to grow as a bio-informatician, a scientist overall, and for providing me exciting yet challenging projects. I could not have found a better supervisor. I would also like to thank John Rioux for sharing with me stories of his career path, which inspired me to work hard.

I want to thank our collaborators Frederic Galacteros, Carlos Brugnara, Pablo Bartolucci, for their depth and breath of knowledge, and their incisive contributions to our projects.

I am grateful to all members of the Lettre lab, for all the fun lab outings, lab meetings, lab lunches, you all participated in creating a collegial and stimulating work environment. I am fortunate to be a part of an amazing group who made this journey as enjoyable as possible. For their contribution, and insights into my work I would like thank Ken Sin Lo (my coding sensei master), Cecile Low-Kam (my statistics guru), Samuel Lessard, Nathalie Chami, and Melissa Beaudoin (all three geneticist mentors).

Last but not least, I would like to thank my fiancé, Marième Dembélé, my family members, Christiane, Jean-Pierre, Andy, and Giulia for your unwavering love and support throughout my studies. Finally, also like to thank my friends, for their patience and for bearing with me when I canceled several get-togethers, showed up late, or missed events because of my homeworks or labwork, you guys are the best.

1. Introduction

1.1 Sickle Cell Disease Historical Background

Sickle cell disease (SCD) was first described more than a 100 years ago in Occidental literature, by a cardiologist named James B. Herrick while tending to a dental student who complained about chest pain. In 1910, he published what is considered today the first report in a medical journal describing red blood cells with an odd shape, as seen in **Figure 1**, which he called “sickle-shaped cells”[1,2]. In 1927, E. Vernon Hahn and Elizabeth Biermann Gillespie[3] were the first to discover the relationship between red blood cells and low oxygen. Three years later, Scriver and Waugh determined that in the absence of oxygen, red blood cells become sickled[4]. About twenty years later, in 1948, Janet Watson was the first scientist to elucidate the protective role of fetal hemoglobin (HbF)[5] on the disease noticing that newborns with the disorder did not display any of the known complications. That same year, award winning Nobel scientist, Linus Pauling[6] called “Sickle Cell Anemia, a Molecular Disease” in Science , where he explained that the sickling phenomena originated from abnormal hemoglobin (HbS) which differed from normal hemoglobin. The following year James V. Neel[7] uncovered the recessive model of inheritance. The last two historical landmarks of SCD occurred in the 1950s. In 1956, Anthony Allison discovered the link between the protective effect of the sickle cell trait and malaria[8]. The second one happened two years later when Vernon Ingram confirmed that the abnormal hemoglobin (HbS) differed from normal adult hemoglobin (HbA) by a single amino acid which replaced a glutamic acid by a valine amino acid at position 6 of the β-globin subunit of hemoglobin[9]. Although previous reports described the process between red blood cell deoxygenation and sickling, Ferrone et colleagues[10] cemented our appreciation of the process by explaining that the abnormal hemoglobin polymerizes under deoxygenation thus disrupting the shape of erythrocytes. This body of discoveries contributed to our understanding of the molecular basis of sickle cell disease and constitutes the foundation of future investigations of SCD in the 21st century.

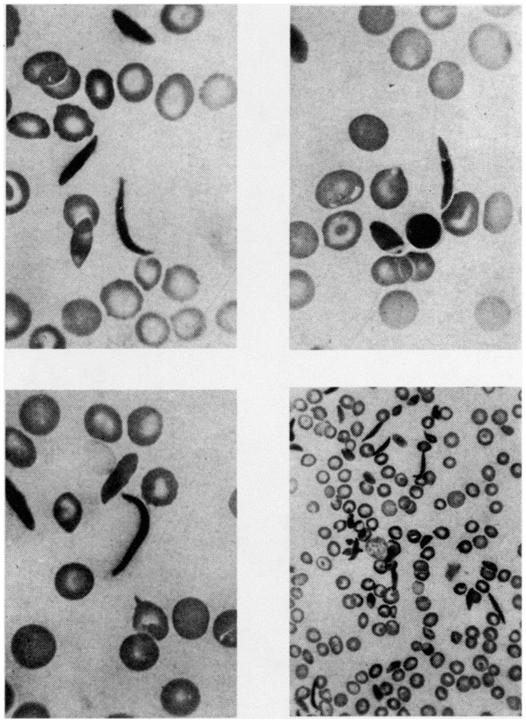


Figure 1. Blood Smear of Sickle Cell

The image was copied from J.B. Herrick (1910).

1.2 SCD Burden in Today's Society

The two main inherited hemoglobinopathies are: sickle cell disease (SCD), and the thalassemia syndromes. On one hand, SCD made distinctive by red blood cell assuming a crescent shape as opposed to the normal biconcave disc-like shape (**Figure 2**). This change in anatomy will cause cells to clog up in blood vessels, small capillaries, and to have a shorter lifespan. As a result, problems such as kidney damage, stroke, acute pain, skin ulcers, infections, to name a few, will ensue. Thalassemia, on the other hand, is characterized by an imbalance in the synthesis of the globin chains. Several types of thalassemia exist; the most common ones are α and β thalassemia, which cause ailments such as enlargement of the spleen (splenomegaly), susceptibility to infections, and more.

In 2006, the World Health Organization (WHO) recognized SCD as a global health problem[11]. The increase in SCD awareness prompted public health organizations and health professionals to implement strategies to reduce infant mortality, which translated into a systematic prenatal screening, prescription of antibiotics, and vaccinations for children. However, these pediatric preemptive measures are more accessible to high-income countries compared to low-income countries where 1 in 2 neonates will not reach the age of 5[12]. Although the infant mortality has decreased in some parts of the world, projections indicate that the global burden of the disorder is set to increase, from over 300,000 newborns in 2010 to more than 400,000 by 2050 (**Figure 3**)[13]. Most of this increase in birth will be attributed to the African continent, which accounts for 70% of the world's cases of SCD. This growth in population from Africa is evidence for the argument that malaria endemic countries have the highest disease prevalence[14].

While the highest incidence of the trait is attributable to Africa, WHO's survey from 2011 found that worldwide 35 million individuals carried a mutant allele of the disease (i.e. individuals that are heterozygous for the sickle cell disease mutation). Indeed, we can find individuals of Hispanic (South America, Central America, and parts of the Caribbean), Mediterranean (such as Greece, Turkey, and Italy), Indian and Arab descent with the sickle cell trait. Studies attributes the occurrence of the gene in those populations to migration, which introduced the allele in non-malaria endemic regions, and to selective pressure of malaria, which increased the survival of individuals who lived to pass on their genes[15-18].

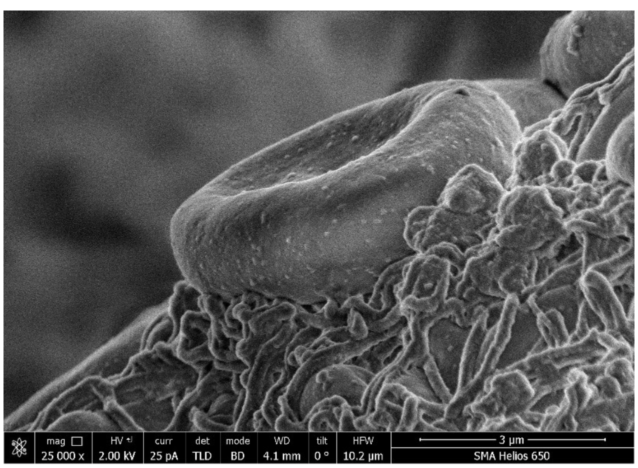


Figure 2. Scanning Electron Micrograph of Normal Erythrocyte

Retrieved as is from Wikipedia.

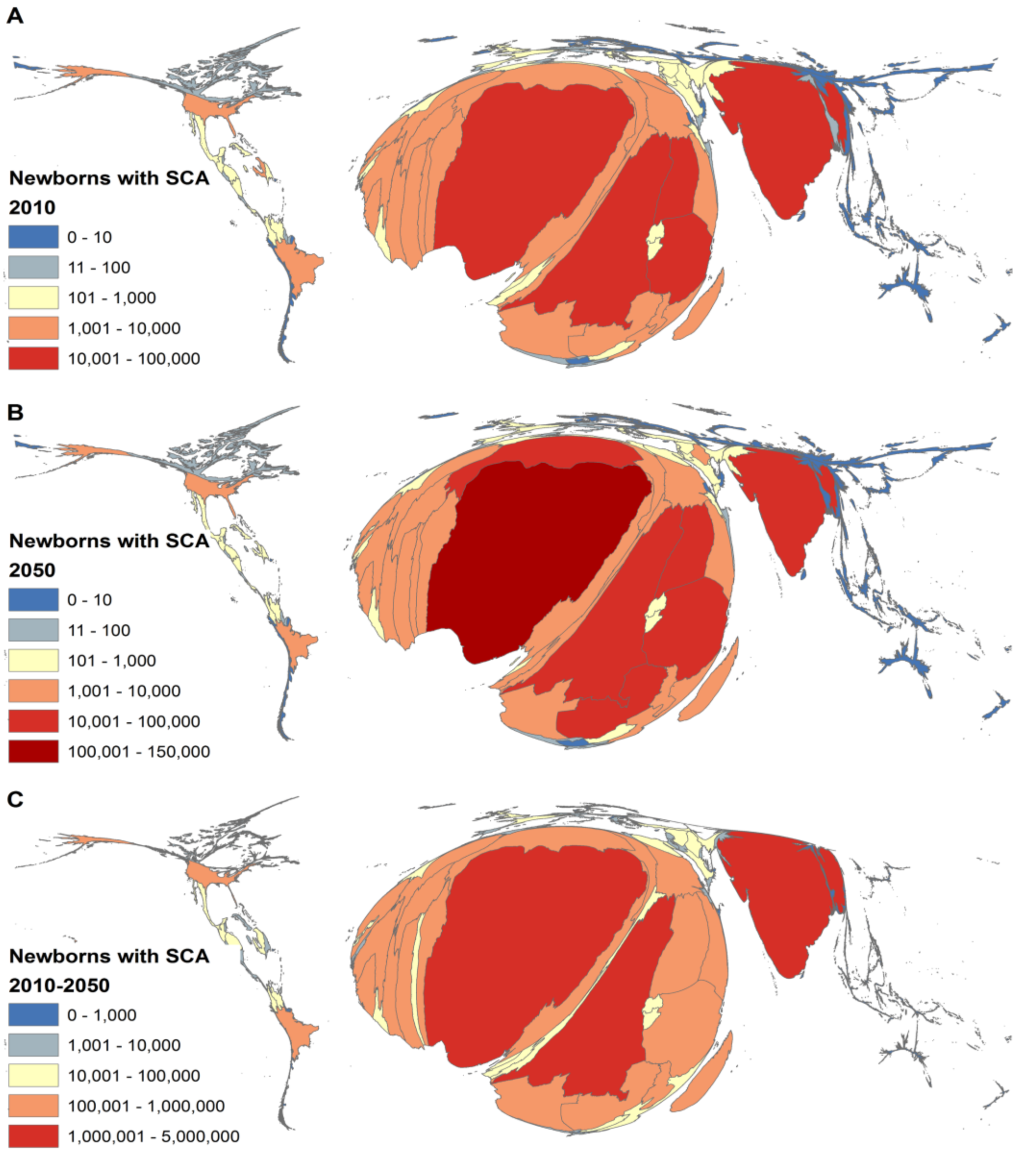


Figure 3. World's Distribution of Infants with SCA

As seen in Piel et al. (2013).

1.3 Red Blood Cell and Hemoglobin

In mammals, red blood cells are biconcave and disc-shaped. Hemoglobin protein molecule depicted in **Figure 4** is a 64 kDa complex with four polypeptide chains; two β polypeptide chains, and two α polypeptide chains, held together by non-covalent bonds[19]. The hemoglobin tetramer, also known as HbA, or adult hemoglobin, is the most predominant in humans. Each globin chain contains a heme group in which the iron atom binds to oxygen as red blood cells pass through the lungs and releases it once in peripheral tissues. Carbon dioxide (CO_2) is then loaded for a return trip to the lungs where it is exchanged for oxygen. Two different gene clusters encode the α-globin and the β-globin families. The α-globin locus on chromosome 16 contains from 5' to 3' the embryonic ζ-globin gene and two adult α-globin genes. The β-globin locus resides on chromosome 11 and contains from 5' to 3' the embryonic gene (also known as ε-gene), two fetal γ-globin genes, and the adult genes, δ and β genes. Each of these genes combines to become different hemoglobin tetramer form during various stages of development (embryonic, fetal, and adult life). **Figure 5** illustrates the relative levels of expression of the different globins over time during pregnancy on the y-axis and the organs responsible for blood cell production (erythropoiesis) on the x-axis. The ε-globin and ζ-globin genes responsible for embryonic hemoglobin are produced during the early maturation stage of red blood cells (erythroblast) in the yolk sack[20]. As the fetal liver becomes the site of erythropoiesis, red blood cells become more and more mature, with α and γ genes taking over the previous embryonic globin genes. At the time of birth in humans, the bone marrow replaces the fetal liver as the site of erythropoiesis[21].

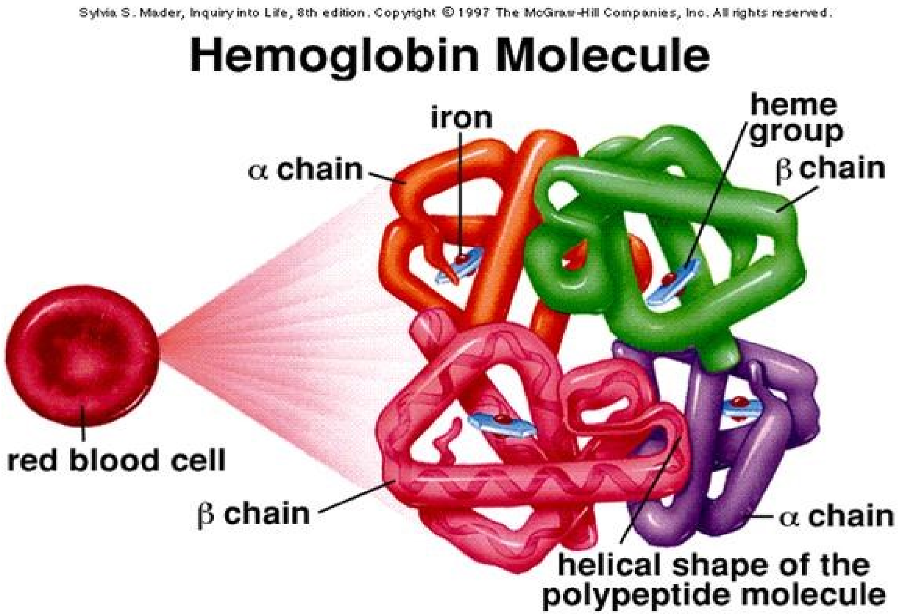


Figure 4. Hemoglobin Molecule

Copied from the book Inquiry into Life.

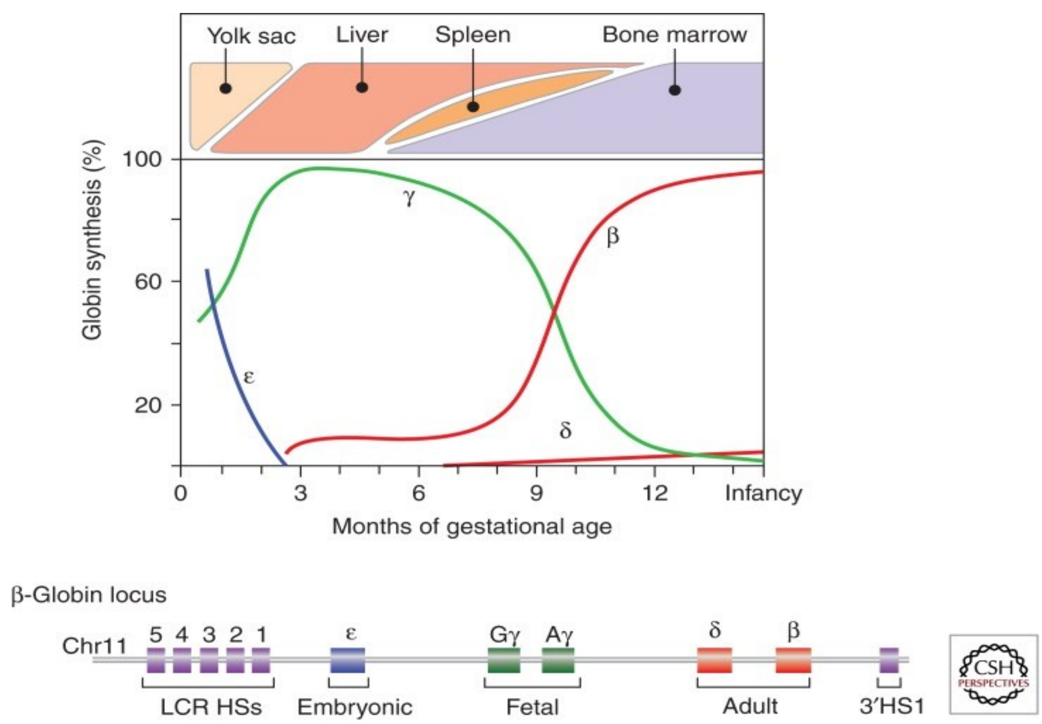


Figure 5. Globin Gene Expression during Development

Duplicated from Sankaran VG et. al (2013).

1.4 SCD and Malaria

1.4.1 Historical Perspective

Malaria is a potentially fatal disease caused by a protozoan parasite infection of red blood cells. Depictions of the disease symptoms date back more than four millenniums ago, in Chinese medical scrolls, Greek documents, Roman writings, and Spanish missionaries memoires[22]. Since then remedies existed to overcome the illness. In China for example, the Qinghao plant (*Artemisia annua*) dispensed to infected individuals is today known to contain artemisinin, an effective antimalarial drug particularly in combination with other drugs. In fact, 2015 Nobel Prize in Medicine was awarded to Youyou Tu for her work on an artemisinin-based drug, which completely cures sick individuals within 72 hours. Additionally, quinine, another potent antimalarial drug used today, was administered back in the 17th century in the form of a bark tree known as the *Peruvian bark* as a cure[23]. The discovery of the parasite, at the time named *Plasmodium Oscillaria* and later renamed *Plasmodium Falciparum*, came from Nobel-prize winning French surgeon, Charles Louis Alphonse Laveran[24]. His discovery was groundbreaking because it was the first time a eukaryotic pathogen was observed in human cells. Following his findings, in 1886, Italian scientist Camilo Golgi, found that two other species of the parasite caused variable symptoms of malaria[22]. His fellow countrymen, Giovanni Batista Grassi, and Raimondo Filetti ten years afterwards set out to name each of them; *Plasmodium vivax*, *Plasmodium malariae*[25]. Simultaneously to Ronald Ross' discovery that the *Plasmodium relictum* is transmitted by mosquitoes in birds causing avian malaria, a joined effort between Italian scientists, lead by Giovanni Batista Grassi, observed that the Anopheles mosquitoes act as a vector for the *P. falciparum, P. vivax*, and *P. malariae* in humans. The scientific inquiries on the parasite life cycle, pathophysiology, therapies, complications related to the disease, and resistance of β^S trait carriers against the Plasmodium stemmed from these breakthroughs. Finally, to date, at least 150 species of the Plasmodium genus have been discovered to cause malaria in other vertebrates[26-30], two of which in humans: *Plasmodium ovale* and *Plasmodium knowlesi.*[25]

1.4.2 Pathophysiology

The symptoms of the disease intensify from parasite to parasite. The *P. Falciparum* causes the most severe clinical manifestations because of its role in the lifecycle in the human body. Complications resulting from the infection, include fever, nausea, vomiting, headache, and in severe cases seizures, pulmonary embolism, jaundice and renal failure, convulsions, uremia, and acidosis.[31] The pathogenesis of the *Plasmodium* in humans has been extensively documented by Ricardo T. Gazzinelli et al[32] and more recently by Samuel Crocodile Wassmer et al[33].

1.4.3 Burden and Protection

Despite recent claims of malaria mortality reduction (from 21% to 57%)[34,35,36] over the past decade, the disease continues to claim hundreds of thousands of lives each year and remains the leading cause of death in developing countries. With estimates ranging between 207 and 214 million new cases every year, the infection is most frequent in Africa (88%), southeast Asia (7%) and the Mediterranean region (2%) [35,36]. The high prevalence of malaria in Africa can be attributed to the selective pressure, which caused germline mutations to confer a survival advantage against the disease. In fact, since Anthony Allison's discovery of the protective capability of the sickle cell trait against malaria, studies have attempted to uncover the exact biomolecular mechanism through which the β-globin mutation hinders the parasite life cycle in the human body. Several genetic determinants are thought to shield from the parasite; genotypes β^A/β^E, β^A/β^C, β^C/β^C and β^C, β^E mutations prevent the parasite multiplications through its interaction with low oxygen or reduce adherence to erythrocytes[37,38]. Additionally, two erythrocyte enzymes deficiency; G6PD (glucose-6-phosphate dehydrogenase deficiency**)**, and PKLR (pyruvate kinase) are thought to lessen the density of the *Plasmodium*, the first one through oxidative stress while the second one through inhibition of its replication *in vitro*.[39] Moreover, National Institute of Health (NIH) group lead by Miller LH found that platelet glycoprotein 4**,** also known as CD36, sequester the plasmodium parasite inside erythrocytes and compromises its immune system[40]. Several other hypothesis and mechanism involving heme-oxygenase 1 (HMOX1), or the interactions between higher levels of carbon monoxide and hemoglobin S[41] (**Figure 6**), and other molecules were put forward, without necessarily being confirmed in humans on a large scale[42-44] .

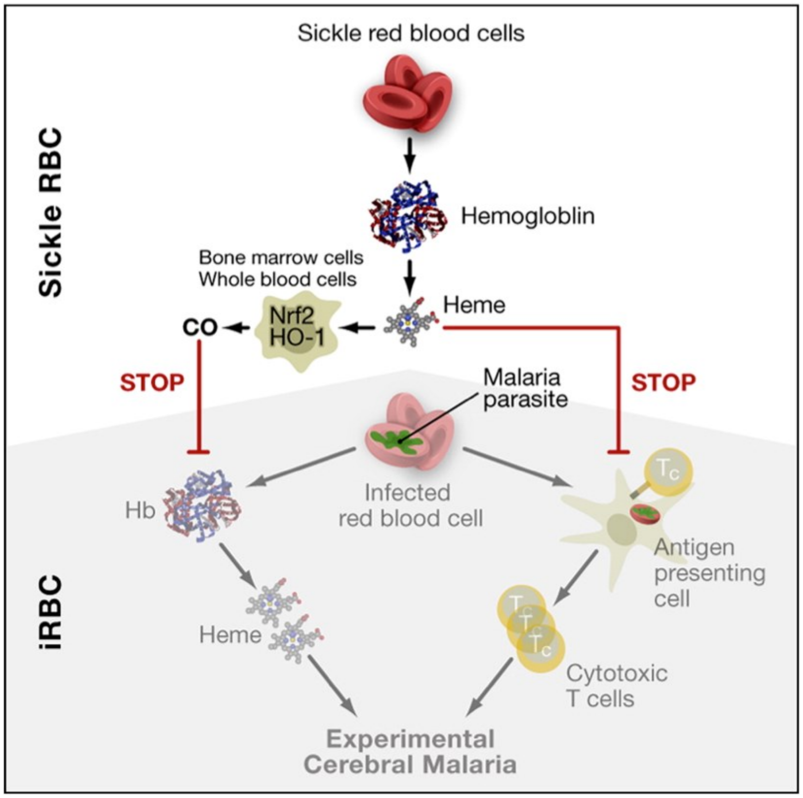


Figure 6. Biomolecular Model of Malaria Protection

Copied from Ferreira et al. (2011)

1.5 Pathophysiology of Sickled Cells

Hemoglobin S (HbS) polymerization is the chief and initial phenomenon in SCD pathophysiology. Upon the removal of the oxygen atom and dehydration of erythrocytes, the hemoglobin molecule becomes sticky and starts to form rod-like structures. Erythrocytes then become rigid, dehydrated and eventually assume a sickle shape. While the polymerization process is initially reversible, after multiple cycles of sickling, the blood cells become irreversibly sickled (IRS). These cycles, in turn, have some significant effects on red cell membrane structure, function, and adherence to the vascular endothelium, which will lead to the trapping of red blood cells and leucocytes in small capillary beds. According to Rees et al.[45], and as seen in **Figure 7**, there are two major pathways involved in the manifestations of the complications related to SCD. The first one is a direct consequence of cells being trapped in blood vessels and of the endothelium becoming sticky. This obstruction of blood vessels, and increase adhesiveness of the endothelium causes complications such as acute pain, nephropathy, inflammation, and pulmonary hypertension to name a few. The second pathway is the result of red blood cells bursting and therefore releasing hemoglobin, which will bind nitric oxide (NO)[46]. Hemolysis is the source of complications such as leg ulcer, priapism, stroke and chest pain[45]. The sickling rate in erythrocytes is directly correlated with the intracellular concentration of HbS, which can be reduced by the presence of fetal hemoglobin (HbF) as seen in individuals with the hereditary persistency of fetal hemoglobin (HPFH) genotype. Indeed, different alleles other than the β-globin mutation can influence HbS concentration and either raise or lower the polymerization rate. We find three main SCD genotypes. The first one is sickle cell anemia (SCA) groups together all individuals with homozygous β-globin mutation (β^S/β^S) which depending on reports and populations' ethnicity can vary from 36.4% to 95.7%[47]. The second genotype consists of heterozygous (β^S/β^C) with allelic frequency ranging from 3.6% to 92.2% depending on ethnicity and reports. The final genotype consists of β^S and β-thalassemia (β^S/β^0) which is widespread mostly in Arabs (~28%), and in Indians (~30% and ~63%), but has a low frequency in Africans (~0.7%)[47]. Other genotypes of the disease can occur, as described in **Table1** below.

Severe sickle-cell disease	Characteristics
HbS/S (β6Glu>Val/ β6Glu>Val); sickle-cell anaemia	The most common form of sickle-cell disease
HbS/β^0 thalassaemia	Most prevalent in the eastern Mediterranean region and India[48]
HbS/OArab (β6Glu>Val/ β121Glu>Lys)	Reported in north Africa, the Middle East, and the Balkans; relatively rare[48]
HbS/D Punjab (β6Glu>Val/ β121Glu>Gln)	Predominant in northern India but occurs worldwide[48]
HbS/C Harlem (β6Glu>Val/ β6Glu>Val/ β, β73Asp>Asn)	Electrophoretically resembles HbSC, but clinically severe; double mutation in β-globin gene; very rare[49]
HbC/S Antilles (β6Glu>Lys/ β6Glu>Val, β23ValIle)	Double mutation in β-globin gene results in severe sickle-cell disease when co-inherited with HbC; very rare[50]
HbS/Quebec-CHORI (β6Glu>Val/ β87Thr>Ile)	Two cases described; resembles sickle-cell trait with standard analytical techniques[51]
Moderate sickle-cell disease	
HbS/C (β6Glu>Val/ β6Glu>Lys)	25-30% cases of sickle-cell disease in populations of African origin[52]
Moderate HbS/ β^+ thalassaemia	Most cases in the eastern Mediterranean region; 6-15% HbA present[48]
HbA/S Oman (β^A/ β6Glu>Val, β121Glu>Lys)	Dominant form of sickle-cell disease caused by double mutation in β-globin gene; very rare[50]
Mild sickle-cell disease	
Mild HbS/ β^{++} thalassaemia	Mostly in populations of African origin; 16-30% HbA present[48]
HbS/E (β6Glu>Val/ β26Glu>Lys)	HbE predominates in southeast Asia and so HbSE uncommon, although frequency is increasing with population migration[53]
HbA/Jamaica Plain (β^A/ β6Glu>Val, β68Leu/Phe)	Dominant form of sickle-cell disease; double mutation results in Hb with low oxygen affinity; one case described[54]
Very mild sickle-cell disease	
HbS/HPFH	Group of disorders caused by large deletions of the β-globin gene complex; typically 30% fetal haemoglobin[48]
HbS/other Hb variants	HbS is co-inherited with many other Hb variants, and symptoms develop only in extreme hypoxia

Table 1. Different Types of Sickle Cell Disease

Copied as is from Rees DC, et al (2010). Genotypes that have been reported to cause sickle-cell disease are listed. All include at least one copy of the β^S allele, in combination with one or more mutations in the β-globin gene. HbS=sickle haemoglobin. HbA=haemoglobin variant A. HbE=haemoglobin variant E. Hb=haemoglobin.

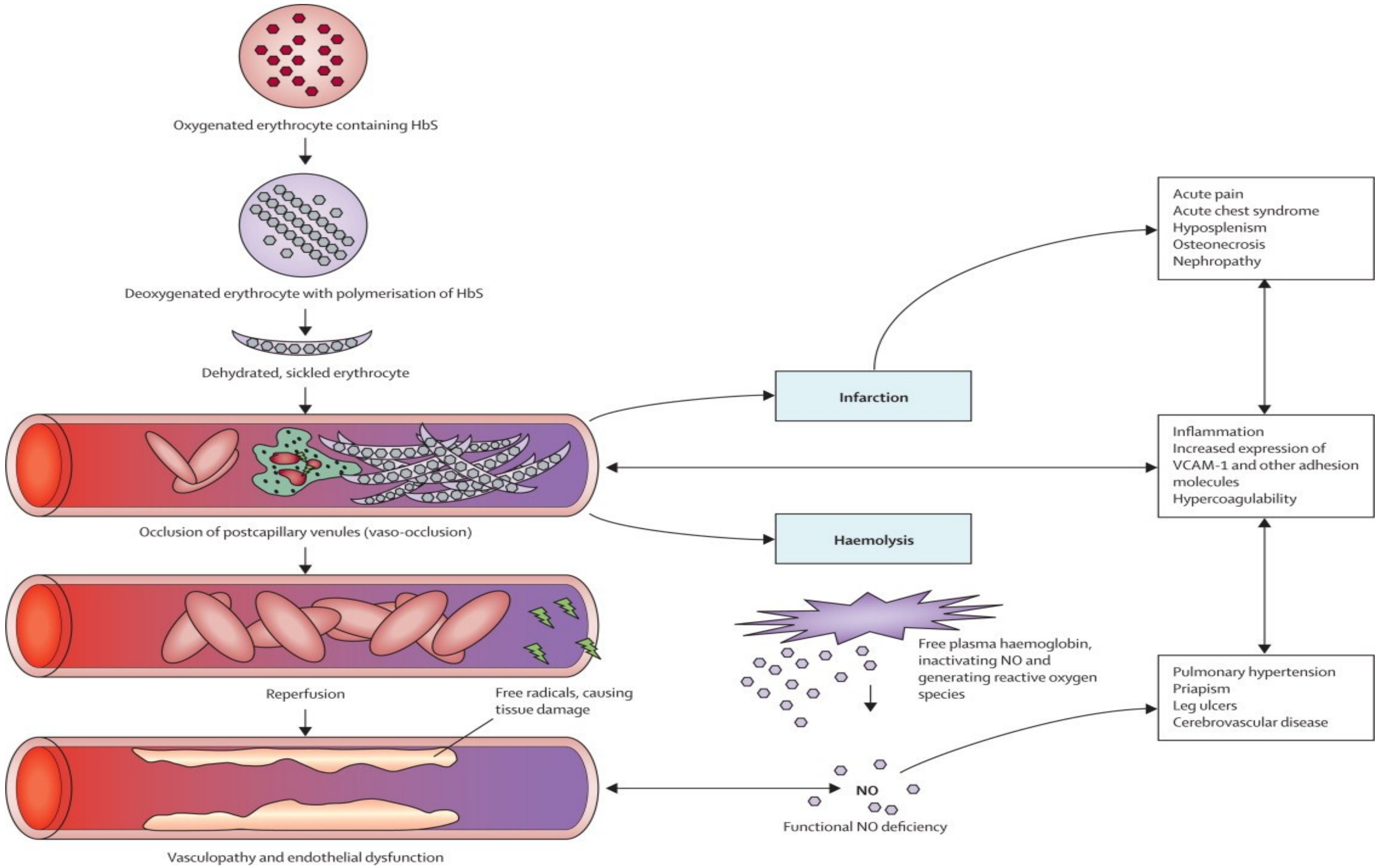


Figure 7. Pathophysiology of Sickle Cell Disease

Copied from Rees DC, et al (2010)

1.6 Complications of Sickle Cell Disease

As mentioned before, SCD is a systemic disorder affecting multiple organs. These include the cardiovascular system (chronic myocardial insufficiency), the digestive system (chronic liver disease, gallbladder dysfunction), the urinary system (nephropathy), the nervous system (cerebral infarction), the male reproductive system (priapism), the ocular system (retinopathy), the skin system (leg ulcers), the blood system (splenectomy), and the respiratory system (acute chest syndrome). The National Heart Lung and Blood Institute website at the National Institute of Health identifies 18 major inter-related complications to SCD[55]. Although children show no sign of the disease until they are 5 to 6 months old due to HbF's protective effect, some of the first most common complications newborns are likely to experience include: dactylitis (aching and swelling of hands and feet), anemia (manifested by fatigue), and jaundice (yellow coloring of skin and eyes due to hemolysis of cells)[56]. Below are the four overarching complications, each with the specific organs they affect and their manifestations:

- Acute Pain
    - One of the major complications and a sure sign of SCD is the acute pain episode [18,57]. Pain crises are sudden, unpredictable, and are attributable to erythrocytes being entangled in blood vessels, thus, reducing the supply of oxygen to tissue organs. The pain episodes, can therefore be felt at any location on the body, and have been described as intense, agonizing, excruciating aches that can require hospitalization.
    - Acute chest syndrome is another well-characterized complications of SCD[58,59]. It ensues from vaso-occlusion of erythrocytes in the lungs, which prevents the provision of oxygen to capillaries, which will, in turn become damaged. Acute chest syndrome's symptoms resemble those of pneumonia and often lead to hospitalization of the individual. This complication is so severe that it is the leading cause of deaths in adults**.**
    - An hour long lasting painful erection without sexual arousal (priapism) is a frequent complication in males with SCD. Priapism persisting more than three hours is called prolonged priapism (PP), whereas when there are intermittent relapsing attacks of lasting two to six hours it is known as stuttering priapism (SP). Severity or recurrence of either SP or PP can end in penile fibrosis and

impotence.

- Chronic Pain
  - Bone and joint-related complications in SCD patients are debilitating. Usually widespread amongst teenagers, both osteoporosis, and bone marrow expansion are common[60]**.** Osteonecrosis results from bone infarction, and is one of the leading causes of chronic pain in adults with devastating effect on the quality of life. Bone marrow infarction is a result of an acceleration of the production of blood cells (hematopoiesis), which can lead to reticulocytopenia (also known as aplastic crises, the decrease production of reticulocytes), a release of immature leukocytes and erythrocyte in the blood (known as leukoerythroblastic anemia).
  - The advent of adolescence and the presence of severe anemia may enable the rate of leg ulcers[61,62]. These ulcers are more often reported in patients with β^S/β^S genotype than with β^S/β^C (22% and 9% respectively)[63]. Plus, lifetime occurrence of ulcers can vary from patient to patient. 50% of individuals will experience leg ulcer once in their lifetime, another 25% will experience them once to twice a year over several years, and finally the remaining 25% will experience leg ulcers chronically with multiple relapses[64]. Due to inflammation, scarring, and infection that accompany the complications, leg ulcers can be very painful. While high levels of HbF improve the condition, low hematocrits count (HCT), with increase hemolysis and manhood are additional risk factors for developing leg ulcers[65]. Moreover, a Jamaican cohort study of 225 patients showed that the occurrence of leg ulcers is increased at 18 years of age[66].

- Vascular Disorders
  - Brain related complications also manifest themselves in SCD patients. These can be broken down into two types of complications: clinical strokes and silent strokes. Clinical strokes, which are due to a loss of the blood circulation to an area of the brain and cause tissue damage, are called ‘clinical’ strokes because their onset is noticeable and identifiable when they occur. Symptoms include seizures, single-

sided weakness or numbness/tingling, loss of balance, vision, and slurring of speech. One severe form of clinical stroke is known as intracerebral hemorrhage, which is the result of the breakage of an aneurysm that can lead to sudden death. The second form of brain related complications is the silent strokes, also known as silent brain infarct. These are a temporary loss of blood flow with unnoticeable symptoms of stroke causing brain lesions. In term of incidence, 35% of individuals will be affected by one of the types of brain vasculopathy, with 10 to 15% of them being under the age of 10[67]. Transcranial Doppler (TCD) screening is an effective method for cerebral vasculopathy[68].

  - Eye problems or retinopathies are common for SCD patients. Goldberg et al. [69] have divided a classification system to assess the progression of retinal complications in SCD. Stage 1 consists of a simple peripheral inadequate blood supply with arterial occlusion. Stage 2 features the degradation of ocular capillaries, with a benign change in vasculature near the retina. Stage 3 consists of the formation of functional microvascular networks with red blood cell perfusion known as neovascularization. Finally, stage 4, marks the leakage of blood into the areas in and around the vitreous humor of the eye, which can lead to a detachment of the retina.

- Organ insufficiencies and other complications
  - Renal complications or nephropathy are well characterized in SCD. Often times, renal manifestations are due to kidney's tubular malfunction, thus causing improper acid excretion, inappropriate uric acid elimination, and inefficient potassium regulation. It is estimated that 18% of SCD patients will experience renal failure over their lifetime. The presence of blood in the urine (hematuria) occurs more often in individuals with the β^S/β^C genotypes as opposed to β^S/β^S. Hematuria results from the death of renal papillary tissue, but can also be due to the formation of stony mass (calculi) in the body, tumor, or infection[70,71]. Additionally, the presence of protein in urine (proteinuria), the presence of albumin in urine (albuminuria), excess of uric acid in the blood and attacks of gout are the consequence of glomerular dysfunction, an acquired dysfunction in SCD.

- The other major category causing complications is severe anemia. Different degrees of anemia affect individuals with SCD. However, severe anemia, which manifests itself mainly in infants less than ten years of age, is life-threatening and can be caused by acute splenic sequestration crisis. The crisis presents itself as an enlargement of the spleen that is due to considerable drop in hemoglobin levels brought about by acute blood entrapment within the splenic tissue[72]. Moreover, aplastic crisis can also lead to severe anemia. The main cause of this predicament is parvovirus B19 infection, which in SCD patients causes a disruption of red blood cell production leading to severe anemia[73].
- Irrespective of the patients' age, infections are a recurrent theme for SCD patients. With the activity of the spleen being compromised early in their lifetime, patients are more at risk of contracting deadly bacterial infections[74,75]. These can cause blood infection (*septicemia)*, lung infection *(pneumonia),* infection of the membrane covering the brain and spinal cord (*meningitis)*, and bone infection (*osteomyelitis*).

1.7 Known biomarkers of severity in SCD

According to the NIH Biomarkers Definitions Working Group definition from 1998[76], biomarkers (a contraction of the word biological and markers), is defined as an unbiased observation or measure which can be used as an indicator of a diseased or natural biological process or drug response. Their critical role in biomedical research stems from their impact on enhancing drugs effectiveness, and their relevance in helping understand basic science research. Based on the currently available reviews of biomarkers in SCD [77,78], these indicators are categorized based on the pathophysiology of SCD, some are more functional than others (i.e., biomarkers of red cell rigidity vs. total hemoglobin), and some are interrelated (i.e., red cell survival and reticulocyte count). **Figure 8**, illustrates the physiological pathways of SCD starting from hemoglobin polymerization. The red boxes show the originator processes with red blood cell density (DRBC) highlighted in red inside the red cell hydration biomarkers box. The bright green boxes are the outcomes of interrelated processes. This summary points to the fact that using DRBC as a biomarker for SCD positions us early enough in the course of sickling that we may be able to reverse it. We believe that as more and more reseachers explore this biomarker and initiate clinical trials, it will become more central in the efforts to treat SCD.

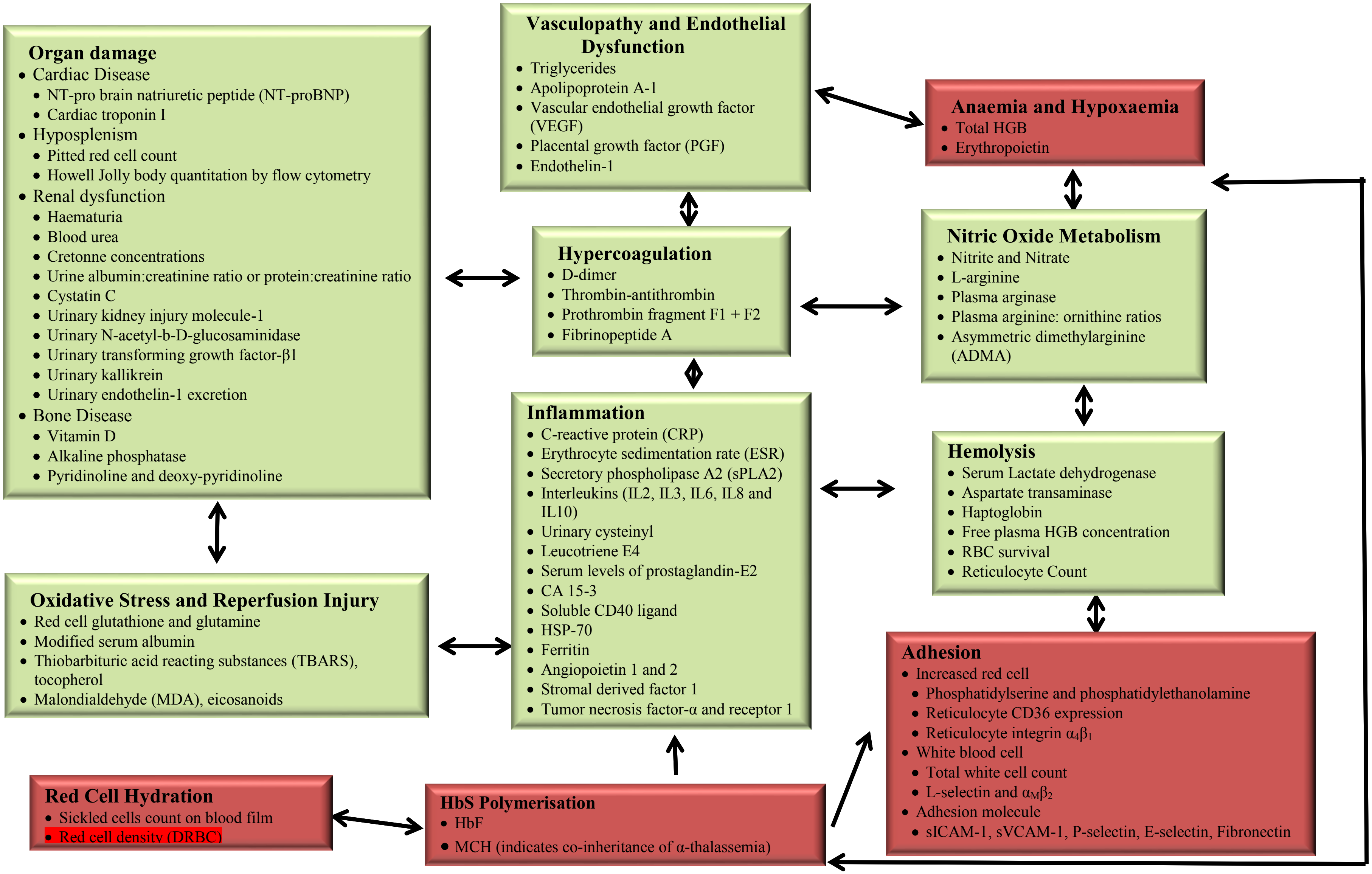


Figure 8. Sickle Cell Disease Biomarkers

Adapted from Rees, D. C. and J. S. Gibson (2012).

1.8 Therapies in SCD

Although gene therapy shows promising signs[79,80], the treatment remains in its early stages of development and is considered to be out of reach for most. Therefore, as of today, there is no official and readily available cure for individuals with SCD.

1.8.1 Hydroxyurea (HU)

HU is the treatment of choice in sickle cell disease, its goal is to stimulate the production of HbF in patients[81-83]. Remaining the only FDA approved drug for the therapy of the disease, it has been shown to reduce acute chest syndrome episodes, the number of hospitalizations and of blood transfusions in all age groups[84-88]. Moreover, it has been proven successful in decreasing mortality in a group of severely affected patients after a 5 to 10 years' follow-up[88]. Nevertheless, the drug is still considered as a half way measure since there are no clear benefits on the risk of stroke[89], priapism, renal complications, as well as pulmonary and cardiac insufficiencies. In addition, several severe side effects exist; myelotoxicity (which is the decrease in production of cells responsible for providing immunity, carrying oxygen, and/or those responsible for normal blood clotting (thrombocytes)), leg ulcers, and low sperm count (oligospermia)[90-92].

1.8.2 Bone Marrow Transplant

Bone marrow transplantation is another accepted therapy in SCD, yielding 80-90% disease free survival[93-96]. It is similar to receiving a blood transfusion, however, the cost of the procedure, the need to identify a sibling compatible donor (or human leucocyte antigen identical donor), the necessity of going through chemotherapy, and to take immune suppressor considerably limits the number of patients who can access the therapy.

1.9 Genome-Wide Association Studies (GWAS) of Sickle Cell Disease

GWAS are a powerful tool to explore the genetic architecture of human disease. In essence, genetic associations refer to the association test between a SNP and a trait. The traits can be categorical (e.i., having or not having type 2 diabetes) or quantitative (i.e., height, weight, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol). For a given trait, the association will be significant if the disease frequency varies according to the genotype. In other words, when testing a specific allele T, at an T/C biallelic SNP, we will find more cases than controls or a correlation with a quantitative trait for a significant association[97]. Thanks to the advent of cost-effective massively parallel genotyping arrays able to genotype upwards 2.4 millions SNP, and the cataloguing of human polymorphism in project such as the 1000 Genome project[98], the HapMap project[99], and more recently the Haplotyple Reference Consortium[100], we have witness a considerable increase in the number of association studies. As of 2014 the National Human Genome Research Institute Catalogue of published GWAS[101] indexes close to 2,000 curated publications, 12,000 SNPs for more than 200 traits. This profusion in genome wide scans is in stark contrast to 2005, at the eve of the completion of the human genome project[102,103] when just very few publications and loci were reported. GWAS confers a true advantage compare to linkage studies and candidate gene analysis, as they give the ability to interrogate the whole genome in a systematic manner[97].

One of the early successes of GWAS dates from 2005 with the identification of the *Complement Factor H* gene as a major risk factor for age-related macular degeneration[104-106] in studies of 146 patients with 90 cases, and about 100,000 SNPs. Nowadays, with the implementation of imputation in GWAS analysis pipelines, and extensive collaborations meta-analyses never conceived before are now undertaken. For example, the recent meta-analysis of blood traits in the UKBiobank[107] which consisted of a sample size of ~200,000 patients, with 29.5 million markers, and the one on body mass index which included over 300,000 individuals and tested over 2.5 millions SNPs[108].

In SCD research, early association studies that focused on acute chest syndrome, priapism, osteonecrosis, and pain crises yielded spurious results[109]. However, SNPs in *UGT1A1*[110] and *MYH9–APOL1*[111] were associated with gallstones, and renal failure respectively in SCD patients and replicated in other cohorts. The presence of replication for these findings provided the first loci for SCD complications. The majority of other successful GWAS in SCD implicated discoveries with HbF. Namely the link between HbF and *BCL11A*[112,113], and subsequently between HbF and *BCL11A*, *HBS1L-MYB*, and *HBB*[114-116] which together account for 50% of the heritability of fetal hemoglobin[117].

1.10 Density of Red Blood Cell in Sickle Cell Disease

1.10.1 Biology and Physiology of Red Cell Hydration

Before 1950 the role of moving solutes and water across red cell membrane was attributed to calcium-dependent potassium channels in general[118]. However, since the discovery of the Gardos channel, a calcium-activated potassium channel, our understanding of the importance of this channel on ion homeostasis has significantly improved. In fact, today, our comprehension of the transport mechanism involved in erythrocyte osmoregulation extent to designing pharmaceutical drugs to target these channels. Studies have found that in red cells both the volume, and the hemoglobin concentration are dependent on the cation, anion content, as well as the water amount. Cation content regulation involves two active and two passive transporter membrane proteins. The sodium-potassium-ATPase pump and the calcium-ATPase pump are the two ATP-dependent transporters that move sodium and calcium outside of erythrocytes either in collaboration with passive transporters or on their own. Although loss-of-function mutation and missense mutation have been reported to cause hemiplegic migraine type 2 and a type of Parkinson disease[119,120], scientific inquiries pinpointed the calcium-activated potassium channel to play an important role in the dehydration of erythrocytes in SCD[121-123]. Passive cation transporters rely on the external and internal concentration of potassium to become active, which why they are labeled co-transporters. The two co-transporters reviewed in human erythrocytes are the sodium-potassium-chloride co-transporter ($Na^+/K^+/Cl^-$), and the potassium-chloride (K^+/Cl^-) co-transporter. Reports indicates that the latter, the electroneutral co-transporter, plays an essential part in erythrocyte dehydration in SCD either by itself or in conjuction with the Gardos Channel. Knockout models and molecular characterization of all of its isoforms (KCC1, KCC2, KCC3, KCC4) shed light on its impact. Of interest, knockout mouse KCC3 (-/-)[124] results in dysfunctional cell volume regulation in neurons and kidney tubular cells, which is accompanied by a loss of hearing acuity, and neurological disorders. Additionally, knockout KCC4 (-/-)[125] lead to deafness and tubular acidosis, and KCC2 (-/-)[126] is lethal just after birth due to respiratory failures. Additional cation transporters include sodium-hydrogen (Na^+/H^+) exchanger, and sodium-magnesium (Na^+/Mg^+) exchanger. The Na^+/H^+ exchanger is crucial for the regulation of intracellular pH and cell volume[127], and was found to be up-regulated in mice with spherocytosis. The second cation exchanger, regulates magnesium content in erythrocytes and may be to incriminate for the low levels of magnesium found in SCD individuals. When it comes to the regulation of anion content the only protein known to be involved is the anion exchanger band 3[128,129]. Several functional roles have been identified for the protein, including; chloride-bicarbonate exchanger ($Cl^-/HCO3^-$), transporter of carbon dioxide (CO_2) from tissues to the lungs' alveoli, and stabilizer of red cell cytoskeleton. *In vitro* studies and DNA mutations analysis in this gene have been associated with severe health outcomes, such as spherocytosis, hemolytic anemia, splenectomy, and renal dysfunction[130-132]. Finally, water content moves freely across red blood cell membrane without the need of any energy input but it can be rushed through the water channel known as the aquaporin 1 (AQP-1)[133]. The water content generally depends on osmotic pressures, and solutes (Na^+, K^+, Cl^-) concentration. **Figure 9** describes the different channels controlling red cell hydration.

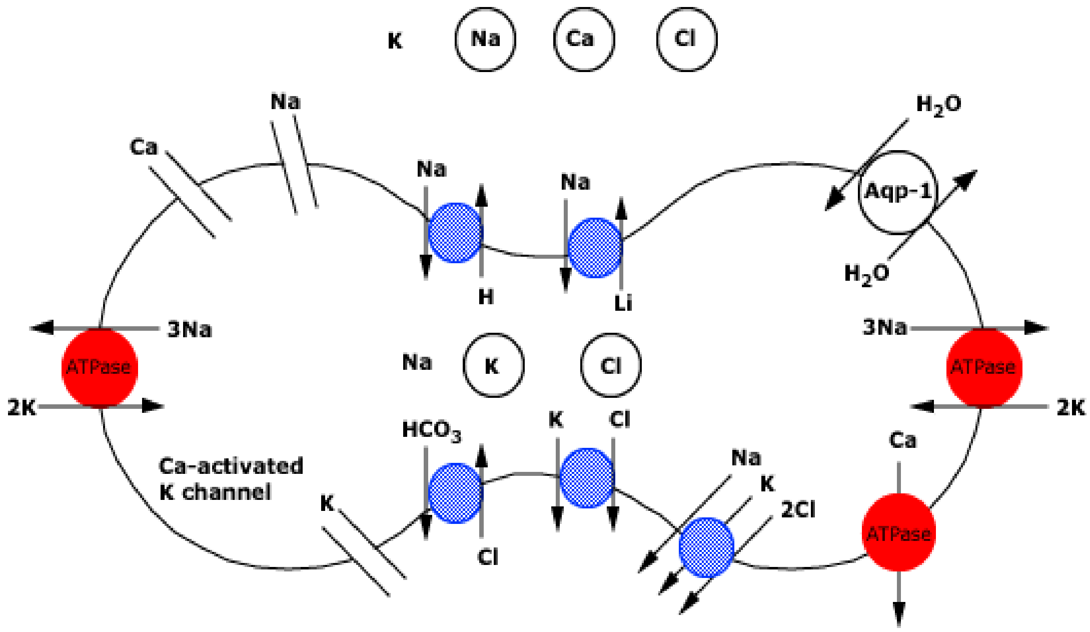


Figure 9. Control of Red Cell Hydration

Retrieved as is from UptoDate webpage on Control of red cell hydration. The figure is a schematic representation of the transport mechanisms regulating red cell hydration. The extracellular concentrations of sodium and calcium are higher than those within the cell, creating favorable gradients for entry, while the intracellular concentration of potassium is higher than that in the extracellular fluid, creating a favorable gradient for potassium exit by the K-Cl cotransporter or the calcium-activated (Gardos) potassium channel. The red transporters are active, the blue transporters are passive. Band 3 protein primarily functions as a Cl-HCO3 exchanger. Its primary physiological function is to facilitate CO2 transport from tissues to alveoli; it also plays an important role in defining red cell shape and membrane stability. Water movement passively follows that of cations and anions, or changes in tonicity of the red cell's environment. Transport of water can occur at a much faster rate via water channels (aquaporin-1, Aqp-1).

1.10.2 Sickle Cell Disease and Clinical Trials

To summarize red blood cells volume is dependent on osmotic pressures, which are dependent on water content, and solutes concentrations. In SCD, the literature shows that three primary pathways are involved in cell dehydration:

- The potassium-cloride co-transporter or KCC pathway, a key regulator in the dehydration of sickle red cells, playing a role alone or cooperatively with the Gardos channel[134-136]. When sickled cells are in contact with the renal medullary environment, they cause leakage of potassium, chloride, and water. KCC has four isoforms *SLC12A4/KCC1*, *SLC12A5/KCC2*, *SLC12A6/KCC3*, and *SLC12A7/KCC4*, which are all present in human erythrocytes.
- The Gardos channel, or KCNN4, activates under deoxygenation and sickling when the red cell membrane is more permeable to calcium. This will push chloride out of the membrane leading to further sickling. Some *in vitro* studies have shown that the calcium-dependent channel causes red blood cell to become irreversibly sickled and contribute to the vaso-occulusive process[137,138].
- The deoxygenation-induced pathway leads to dehydration because it causes the cell membrane to become permeable to calcium thus activating the Gardos Chanel. [139-141].

The density of red blood cell is a biomarker that captures the modification of intracellular HGB concentration and red cell dehydration. This biomarker is crucial in understanding the modulation of hemoglobin S in SCD and designing therapeutic drugs[142] to prevent dehydration or increase hydration. Indeed, Ishii et al[143] described the mechanism through which the calcium-activated potassium channel regulates the dehydration of erythrocytes. Upon activation the Gardos channel, which causes an increased intracellular calcium levels, potassium and water are forced out of the cell, therefore dehydrating it and raising HbS concentration. Bartolucci et al[144] who reported dense red blood as a cell with decreased water content, and increased MCH, provided empirical evidence of the role of DRBC in SCD complications. Their analysis on dense erythrocytes in ~500 SCD patients for the first time established the negative link between DRBC and complications such as renal dysfunction, priapism, and leg ulcer. Moreover, to strengthen the relationship between dense red blood cells and SCD, he showed that after six months of hydroxyurea usage the population of dense cells decreases by 34%. Both the Bartolucci's study and the role of the Gardos channel in red cell dehydration in the context of sickle cell provide evidence that the density of erythrocytes leads to complications, and that a specific protein could be involved in the process of SCD pathology. In fact, using antifungal drug, clotrimazole, in transgenic mice the Gardos channel was successfully inhibited leading to a reversal of dehydration and potassium loss[145,146]. Plus, a clinical investigation with the same therapeutic agent administered to 5 sickle cell anemia individuals was found to effectively inhibit cell dehydration and potassium loss[147]. Another trial consisted in giving oral supplement of magnesium pidolate[148] to 17 SCD patients. Although this pilot study resulted in an increase in cell volume, a decrease in hemoglobin concentration, and reduction in the rate of pain crises[149], it seems to raise the concentration of circulating hemoglobin. More recently a phase III clinical trial of 144 people with Senicapoc (ICA-17043)[150] a Gardos channel blocker was found to improve erythrocyte survival (e.i., hematocrit levels, and reticulocyte count), to reduce the number of dense red blood cells, and hemolysis. The trial didn't move to phase IV because it didn't have any impact on pain crises.

1.11 Research Objectives and Thesis Outline

Given the clinical heterogeneity of SCD, and that erythrocyte dehydration is a typical feature of the disorder, in this thesis, we attempted to identify the genetic factors contributing to the disease severity through red cell density. We hypothesize that DRBC, a precursor to red blood cell sickling, explains the clinical variability of SCD manifestations. We first performed a genome-wide association study to identify common variants with modest-to-weak effect size, prioritizing markers falling within erythroid enhancers, expression quantitative traits locus (eQTL) in candidate genes, and variants associated with MCHC. We then singled out variants based on their significance thresholds for further analysis. Finally, we sequenced the exomes of a subset of our cohort to identify rare variants in candidate genes with high penetrance using a variant scoring scheme and available bioinformatics annotations.

1.12 Thesis Outline

In Chapter 2, I provided a description of the GEN-MOD cohort detailing demographic information, sample size for hematological traits, and complications. Additionally, I reviewed the genotype imputation method, the normalization of DRBC, the quality control measures employed in the genotyping, and the whole-exome sequencing experiments. I also, provided an overview of the statistical methods used in genome-wide association testing. Finally, the chapter ends with a breakdown of the bioinformatics tools utilized to carry out the analyses presented in the thesis. Chapter 3 presents the methods, discussion as well as the association results of DRBC in sickle cell disease patients. Chapter 4, describes the methods, results and discussion of the whole-exome sequencing analysis. Finally, chapter 5 concludes the thesis summarizing and discussing all the key findings, and puts forward guidelines for future experiments.

2. Methods

2.1 Sample Size

The Genetic modifier study (GEN-MOD) is a cohort of African individuals from the West Indies, Sub-Saharan Africa, and Central Africa recruited in France as described in Bartolucci et al[144]. The cohort included 185 men, and 223 women, with a median age of 30 (± 9) (**Table 2** below describes the available blood traits, and complications). Our final dataset consisted of 403 patients selected for genotyping. One individual was removed during genotyping QC, and an additional twenty-eight that were dropped due to DRBC missing values.

Table 2. GEN-MOD Cohort Description

The table below describes the GEN-MOD cohort by sample size for hematological traits, and SCD-related complications. Additionally, it presents the median with standard deviation whenever possible.

Blood traits	**Sample Size**	**Median (± SD)**
DRBC	374	13.1 (± 8.6)
Basophils	407	1 (± 0.9)
Eosinophils	407	2 (± 2.8)
Hematocrit	407	26 (± 4.6)
Hemoglobin	399	8.8 (± 1.3)
HemoglobinA	408	0 (± 3)
HemoglobinA2	408	3.4 (± 0.7)
HemoglobinF	408	5.6 (± 4.8)
HemoglobinS	408	84.7 (± 6.0)
Lymphocytes	407	35 (± 10.7)
MCH	400	30 (± 4.1)
MCV	407	87 (± 10.2)
MPV	356	8.6 (± 1.04)
Metamyelmyel	392	0 (± 0.4)
Monocytes	407	7 (± 3.8)
NucRBC	402	1 (± 7.6)
Platelets	407	382 (± 120.8)
Polys	407	55 (± 11.32)
RBC	408	3.01 (± 0.78)
Reticulocytes	405	81 (± 46.9)
Whitebloodcount	407	10.2 (± 3.7)
Complications	**Sample Size (Control/Cases)**	-
Aseptic Necrosis	237/94	-
Leg ulcer	301/30	-
Cholecystectomy	207/201	-
Stroke	394/14	-
Priapism	107/42	-
Retinopathy	67/182	-

2.2 Phenotype Quality Control

To establish a link between genotype and phenotype in association testing it is common practice to use linear mixed models. As one of the main assumptions of these models is that the phenotype under consideration follows a normal distribution. When this is not the case, it is standard practice for phenotypes to be normalized[151]. Normalizations account for outliers and can involve one of the following: natural log, inverse normal, or square root. We can then evaluate confounding factors (age, gender, batch effect and more), fit them with a linear model to adjust for them. For DRBC we inverse normal transformed it after adjusting for age, and gender using custom R script. **Figure 10** below shows the before and after normalization of DRBC.

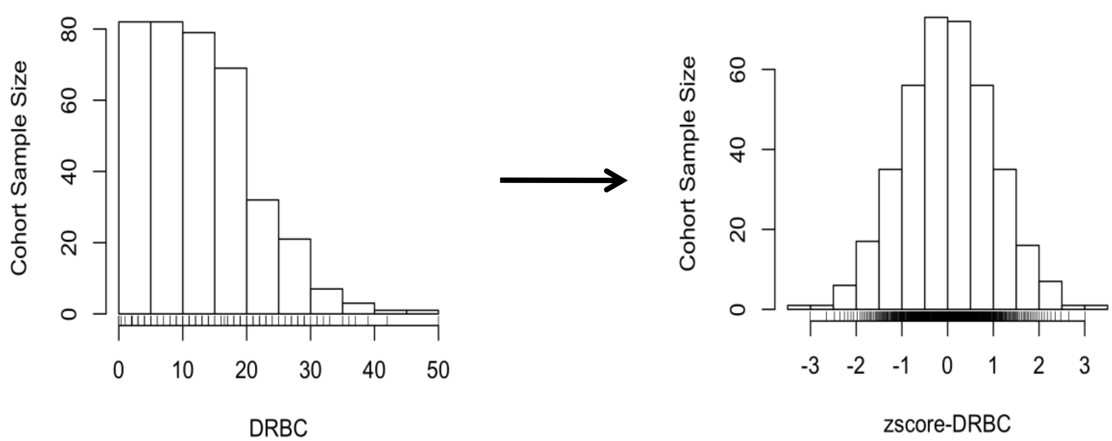


Figure 10. Dense Red Blood Cell Distribution Normalized

The left most histogram shows that the raw data follows a power law distribution prior to any transformation whereas the right most histogram shows that post transformation the data follows a normal distribution.

2.3 Genotyping Quality Control

The genotyping quality control (QC) is essential for any association analysis, and it has been extensively reviewed by Ziegler et al.[152] and by Teo et al.[153]. The three important steps to apply to genotyping QC are SNP quality assurance, sample quality assurance, and population stratification assessment. Looking at the SNP quality assurance involves, filtering out autosomal variants with a genotyping call rate of less than 95%, variants with minor allele frequency (MAF) less than 1% or 5%. Additionally, weeding out correlated variants identified by LD threshold, and SNPs out of Hardy-Weinberg Equilibrium (HWE *P-value* < 1e-7) improves the quality of the dataset (variants out of HWE violate the assumption that allele frequency and genotype frequency are predictable). Sample quality assurance requires filtering out cryptically related individuals, erroneously labeled gender, individuals with a missing call rate less than 95% and individuals with over or under heterozygosis rate. Finally, investigating the population stratification involves identifying individuals that fall outside of their expected ethnicity when comparing them to another population samples dataset through principal components analysis (PCA). In this work, multidimensional scaling (MDS) available in PLINKv1.07[154] was used when comparing the GEN-MOD cohort to the HapMap3[99] samples (**Figure 11**) because we wanted to make sure we could identify population substructure based on genotypic distances.

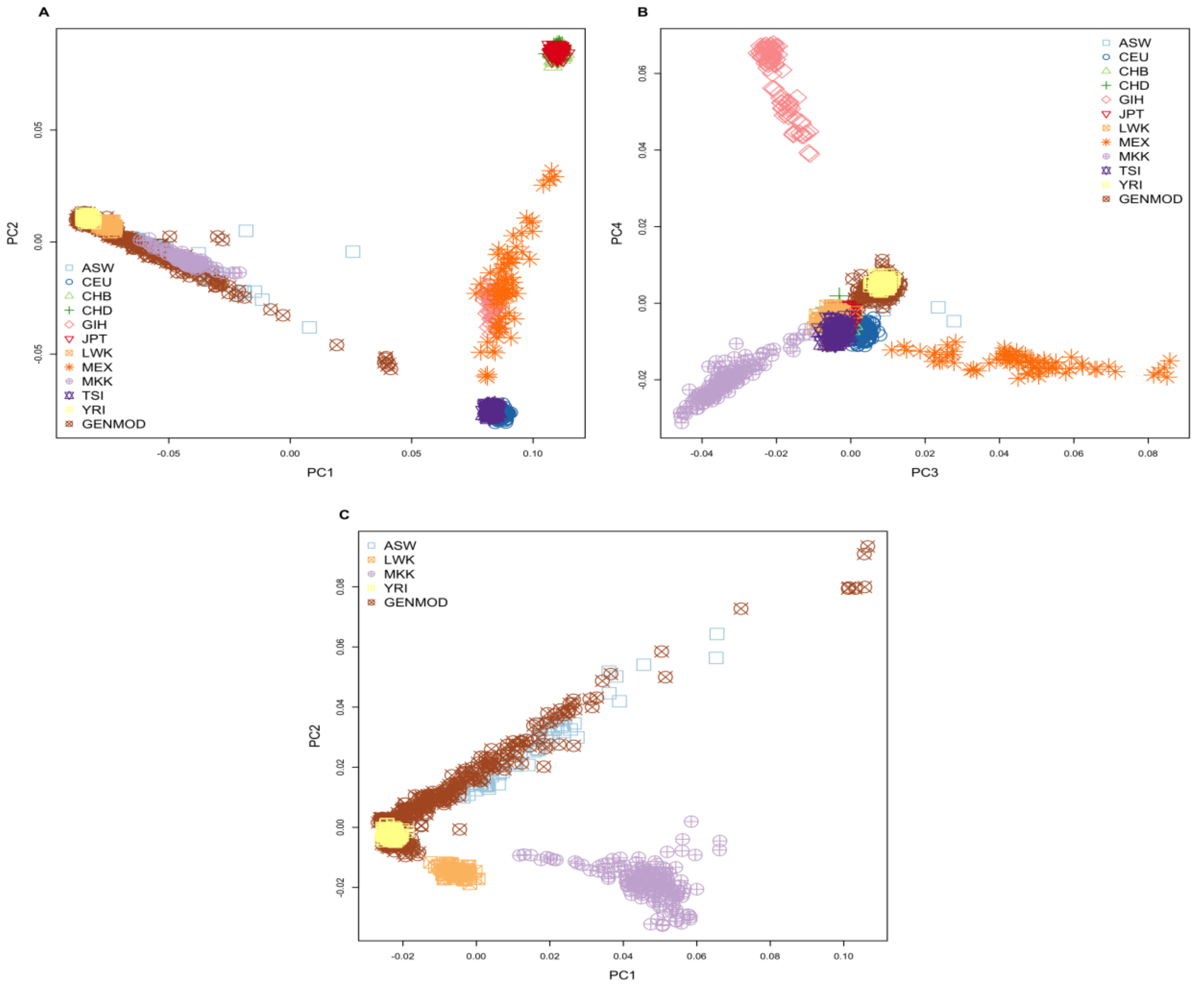


Figure 11. PCA of GEN-MOD against HapMap3

(A) First principal component versus second principal component shows that the GEN-MOD cohort aligns as expected with individuals from African ancestry. (B) Third principal component versus fourth principal component shows again that the GEN-MOD aligns with individuals of African ancestry. (C) First principal component versus second principal component only for individuals from African ancestry, shows the difference in ethnicity amongst Africans.

2.4 Whole-Exome Sequencing Quality Control

Whole-exome sequencing requires extensive pre-processing, and quality control measure to reduce false positives and improve the sensitivity of variant calling. Multiple tools and pipelines for whole-exome sequencing exist and have been reviewed by Bao et al[155]. Additionally, different manufacturers provide their own recommendations for cleaning, and processing their data, yet they all involve the same steps:

- Quality control which entails the removal of low-quality reads, PCR primers, adaptors, duplicates and other contaminants
- Mapping reads onto a reference genome
- Targeted sequencing quality control
- Quality control of mapped reads
- Post-alignment processing
- Variant calling

2.5 Imputation

Genotype imputation is the statistical method to infer genotypes that were not directly typed[156]. The existing implementations (BEAGLE, IMPUTE, and Minimac[157-159]) rely on hidden Markov model to predict untyped markers using both correlation information of typed markers and the reference panel. Imputation allows researchers to analyze markers that were not previously available in their study, and therefore represent significant cost saving. In a comparative analysis that looks at the differences in computation between tools mentioned above, pre-phasing[160] haplotypes was found to have a dramatic improvement on imputation speed for all three methods. In decreasing order of imputation speed, the methods are BEAGLE, Minimac, and IMPUTE[161]. Yet, looking only at factors specific to sensitivity such as concordance (percentage that an observed SNP genotype is identical after imputation), imputation quality score (IQS[162]; concordance adjusted for probability of consensus), and average r^2 [156] (correlation between the imputed genotype and the observed), identified Minimac and IMPUTE2 as best the performer[163]. Our dataset increased from ~2.5 million to 31 million SNP after we carried phasing with SHAPEIT[164] and imputation with Minimac[165] using haplotypes from the 1000 Genomes Project.

2.6 Statistical Methods for Association Testing

2.6.1 Power & Replication

Statistical power in GWAS provides the likelihood of observing a true association. Researchers attempt to optimize or increase their study power by increasing their sample size, focusing on allele frequency and effect size thresholds, and reducing the existing correlation (LD) between tested SNPs such that only independent variants are tested[166]. Testing multiple SNPs increases the odds of observing a significant association just by chance, this is known as the 'multiple burden' hypothesis. Geneticists consider that all associations with a *P-value* $< 5 \times 10^{-8}$ (i.e. a Bonferroni correction for the number of independent loci in the human genome) are deemed genome-wide significant[166]. However, several additional multiple testing corrections exist some more stringent than others[166-168]. Proper GWAS study design does not only rely on power estimation calculation, but also on replication. In the early days of GWAS, the lack of replication led to several markers and loci to be erroneously reported. It is now an imperative to publication to have replication data for the most promising association results in an independent cohort. Given that DRBC is a phenotype rarely measured in clinical studies, replication analysis is not readily available, representing the main limitation of our study. We calculated our power of association for single variant test based on the non-centrality parameter of the chi-squared distribution[166] using a custom Rscript. **Figure 12** below shows the various power curves for N=374 (DRBC sample size in GEN-MOD), and N=1000 (hypothetical sample size if additional samples with DRBC measures were added) with $\alpha = 5 \times 10^{-8}$ at minor allele frequencies of 10%, 25%, and 50%. Based on the power estimation **Figure 12 A** we have 70% > power for variants with MAF = 25%, N=374, beta = 0.5, and 90% > power for variants with MAF = 25%, N=1000, beta = 0.5 (**Figure 12 B**).

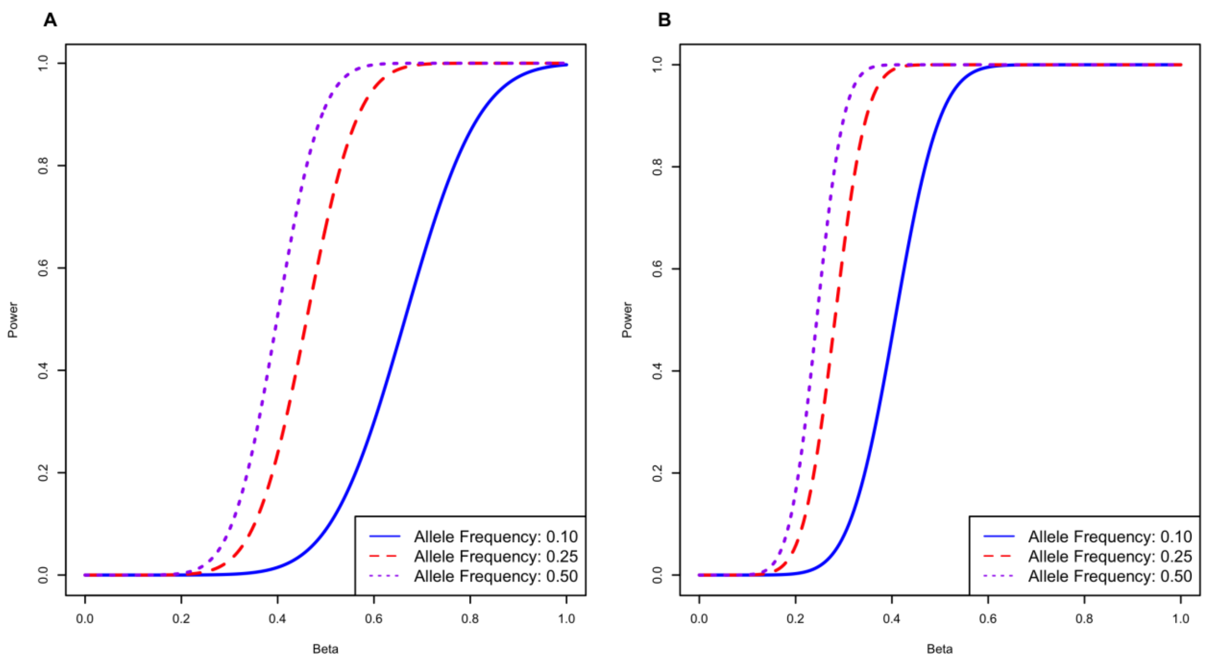


Figure 12. Power Estimation for Association Test of DRBC

This figure shows the power calculations on y-axis according to different beta (effect sizes) on x-axis with each curves representing different allelic frequencies, given that $\alpha = 5 \times 10^{-8}$. In figure A, we assumed that the sample size N = 374 (DRBC sample size in GEN-MOD). In figure B, we assumed that the sample size N = 1000 (if we added more samples with DRBC values).

2.6.2 Single Variant Testing

Depending on the nature of the phenotype, association testing relies on two types of tests. For quantitative traits, the standard is to use a generalized linear model, often an analysis of variance (ANOVA) to test whether or not there is a difference in means in any of the genotype group for the trait of interest. When the phenotype is dichotomous, a logistic regression or a contingency table (i.e., chi-square test) is used. In a contingency table, we assess the difference in genotype frequency between cases and controls. While in the logistic regression the same is accomplished estimates of effect size (odds ratio) can be generated, and covariates can be added to the model. Test statistics remain the same regardless of the disease mode of transmission (additive, dominant, recessive, or multiplicative), however, interpretation of results differ[169]. For our analysis, we tested SNPs individually in an additive model using linear regression given that DRBC is a quantitative trait.

2.6.3 Gene-Based Testing

Markers with low minor allele frequency are not suitable for single variant test because of the reduce power of association[170]. One approach to overcome this challenge - the one we used with DRBC - is to focus on nonsynonymous variation that annotate to the same gene, and then testing them using a burden test or a quadratic test[171]. Given that they test different assumptions, we used them both in our study. On one hand burden tests assume that all variants are causal and have the same direction of effect (VT[172]). On the contrary, quadratic tests retain power even when variants are not in the same direction or necessarily causal (SKAT [173]).

2.7 Bioinformatics Analyses

2.7.1 Bioinformatics Software

Table 3. Summary of Bioinformatics Tools

The table below provides a summary of the tools, source and usage for all the analysis, and quality control steps.

Software Name	Source	Usage
Bedtools	Quinlan AR et al.[174] https://github.com/arq5x/bedtools2	WES QC
BWA	Li H et al.[175] http://maq.sourceforge.net/	WES QC
checkVCF	https://github.com/zhanxw/checkVCF	Genotype QC
GATK	McKenna A et al[176] https://github.com/broadinstitute/gatk	WES QC
Minimac	Das S et al[165] http://genome.sph.umich.edu/wiki/Minimac3	Genotype Imputation
Picard	http://broadinstitute.github.io/picard/	WES QC
PLINK	Chang CC et al[177] https://www.cog-genomics.org/plink2	Genotype QC & Analysis
PLINK/SEQ	https://atgu.mgh.harvard.edu/plinkseq/	Genotype QC
Python	Custom script: https://github.com/yilboudo	GWAS Analysis/WES Analysis
R	Custom script: https://github.com/yilboudo	GWAS Analysis/WES Analysis
Raremetals	Liu D.J et al[178] http://genome.sph.umich.edu/wiki/RareMETALS	Genotype Association Testing
Rvtest	Zhang X et al[179] http://zhanxw.github.io/rvtests/	Genotype Association Testing
SHAPEIT	Delaneau O et al[180-182] https://shapeit.fr/	Genotype Imputation
vcflib	https://github.com/vcflib/vcflib#vcflib	WES Analysis
VCFtools	http://vcftools.sourceforge.net/	Genotype QC
VEP	McKenna et al[176] http://www.ensembl.org/info/docs/tools/vep/	Variant Annotation

2.7.2 Genotyping QC and Imputation

All the individuals were genotyped on the Illumina Infinium HumanOmni2.5Exome-8v1.1 array. PLINK v1.07 was used to remove poorly genotyped variants and samples. Relatedness and duplicate samples were assessed through identity-by-descent calculation (IBD). The following parameters were used: --geno, --mind, --hardy, --maf, --check-sex, --indep 50 5 2, --IBD, --genome. Population stratification was calculated using PLINKv1.07's --cluster --mds-plot 10 with HapMap3[99] as a reference. The files were converted to variant call format file (VCF) with PLINK/SEQ and inspected for strand alignment issues using checkVCF package. After splitting the VCF file by chromosomes with vcftools v0.1.11, we phased and imputed each file with SHAPEITv2.790 and Minimac3 (v1.0.11) against 1000 Genome phase 3 haplotypes (version 5) as the reference panel. Subsequently each VCF file was then filtered out with a custom python script to include variant with $r^2 > 0.3$.

2.7.3 GWAS Analysis, Prioritization, and VEP annotation

We derived the association summary statistics for our GWAS using RVTests (v.20140416)[179] with default options, correcting for age, sex and the first 10 principal components. A custom python script was used to select variants falling within enhancer regions[183,184]. Variants annotation was performed with VEP default script, and afterward aggregated per gene symbol and allelic frequency. Additionally, a Python script was used to distinguish nonsynonymous variants with the following consequences: splice_acceptor, splice_donor_variant, stop_gained, frameshit_variant, stop_lost, start_lost, protein_altering_variant, missense_variant, coding_sequence_variant. We then derived the association summary statistics with rareMETALS(v.6.3)[178]. Identification of proxy variants with LD > 0.8 in the 1000 Genomes European population was performed with PLINK v1.09. The following parameters were used: --r2, --ld-snp-list, --ld-window-kb 1000, --ld-window 99999, --ld-window-r2 0.8.

2.7.4 Whole-Exome Sequencing Analysis

We aligned the reads to the human reference genome (version GRCh37/hg19) with BWA default parameters. Thus generating sequence alignment map files (SAM) which we merged into a single file. Duplicates were marked and removed with Picard with a validation stringency set to lenient. With GATK default parameter, we then proceeded to define intervals to target for local realignment. We then performed the local realignment of reads around indels, fixing mate pair information. Again, with default options recalibration and realignment steps were also carried out with GATK, while for the depth of coverage we set omitBase to true, and minimum mapping quality (mmq) and minimum mapping quality to 9. Gene coverage was calculated at -ct 1 -ct 5 -ct 10 -ct 20 -ct 30. Finally, variant calling was performed with the same tool, using default parameters, and then annotated with VEP. Before converting the numerical representation of genotypes provided by the GT field in VCF files to a human-readable genotype format with vcflib's option vcfgenotypes all the variants were sorted. Keeping only those annotated as nonsynonymous, and with a gene symbol corresponding to our candidate genes. We later used a custom python script to identify individuals carrying a mutation. The subset of SNPs identified was reannotated with VEP this time querying RefSeq transcripts, gene symbol identifiers, exome aggregation consortium (ExAC) allele frequencies, SIFT and PolyPhen prediction and score. Finally, each variant was assigned a score for each hematological trait analyzed. This was accomplished with a python script that computed the average for a hematological trait across individuals carrying the mutation.

3. Genome-wide Association Study of Erythrocyte Density in Sickle Cell Disease Patients

3.1 Author Contribution

This article is in preparation and meant to be published in the American Journal of Hematology. Yann Ilboudo and Guillaume Lettre conceived and designed the statistical and bioinformatics experiments. Yann Ilboudo performed the statistical and bioinformatics experiments. Seth L. Alper, Pablo Bartolucci, Carlo Brugnara, Frederic Galactéros, and Alicia Rivera contributed DNA samples, clinical information, and expert knowledge. Yann Ilboudo and Guillaume Lettre analyzed the results, and wrote the manuscript with contributions from all authors. Josepha-Clara Sedzro and Marie Trudel performed mouse matings, genotyping and isolation of bone marrow cells.

3.2 Affiliations

Yann Ilboudo[1,2], Pablo Bartolucci[3], Alicia Rivera[4], Josepha-Clara Sedzro[5], Mélissa Beaudoin[2], Marie Trudel[5], Seth L. Alper[4], Carlo Brugnara[6], Frederic Galactéros[3], Guillaume Lettre[1,2]

[1]Faculty of Medecine, Program in Bioinformatics, Université de Montréal, Montreal, Quebec, Canada

[2]Montreal Heart Institute, Montreal, Quebec, Canada

[3]Red Cell Genetic Disease Unit, Hôpital Henri-Mondor, Assistance Publique–Hôpitaux de Paris (AP-HP), Université Paris Est IMRB - U955 - Equipe n°2, Créteil, France

[4]Division of Nephrology and Vascular Biology Research Center, Beth Israel Deaconess Medical Center, Boston, USA Department of Medicine, Harvard Medical School, Boston, USA

[5]Faculty of Medecine, Department of Medecine and Department of Biochemistry, Université de Montréal, Montreal, Quebec, Canada

[6]Department of Laboratory Medicine, Boston Children's Hospital, Boston, Massachusetts, USA

3.3 ABSTRACT

Deoxy-hemoglobin S polymerization into rigid fibers is the direct cause of the clinical sequelae observed in sickle cell disease (SCD) patients. The rate of polymerization of deoxygenated sickle hemoglobin is determined primarily by intracellular hemoglobin concentration, itself dependent on the amount of sickle hemoglobin and on red blood cell (RBC) volume. Dense, dehydrated RBC are observed in SCD patients, and their number correlates with hemolytic parameters and complications such as renal dysfunction, leg ulcers and priapism. In order to identify new genes and biological pathways involved in RBC hydration in SCD, we performed the first genome-wide association study for dense RBC (DRBC) in 374 homozygous SCD patients. We did not find genome-wide significant results among the 31 million DNA sequence variants tested, indicating that variants that modulate DRBC have modest-to-weak effects. A secondary analysis demonstrated nominal association of a variant associated with mean corpuscular hemoglobin concentration in non-anemic individuals with DRBC in SCD patients (P=0.003). This intronic variant controls the expression of *ATP2B4*, the main calcium pump in erythrocytes. We showed that *Atp2b4* is not differentially expressed in the bone marrow of SCD mice when compared to control mice. Our study highlights *ATP2B4* as a promising target to modulate RBC hydration in SCD patients.

3.4 INTRODUCTION

Sickle cell disease (SCD) is one of the most common monogenic diseases in the world. It is caused by a single mutation in the gene that encodes the beta-chain of hemoglobin. Despite this genetic homogeneity, SCD patients are characterized by extreme clinical heterogeneity, ranging in presentation from benign mild anemia to devastating cerebrovascular events. Studies of the natural history of this blood disorder have improved clinical care such that most SCD patients in North America and Europe can now expect to reach middle age. Despite this progress, the life expectancy and quality-of-life of SCD patients is reduced, treatment options remain limited, and no widely accessible curative therapy is available. Moreover, universal genetic screening and improved care for SCD have been slow to reach the sub-Saharan region in which resides the vast majority of SCD patients.

Results of seminal observational, epidemiological, biochemical, and genetic experiments have led to the emergence of fetal hemoglobin (HbF) as a key genetic modifier of severity in SCD[185]. The beneficial effects of hydroxyurea (HU), the only drug currently approved to treat SCD, are mediated in part by increasing HbF production. Dense, dehydrated erythrocytes are a hallmark of SCD patients, and red blood cell density (DRBC) has been investigated as a potential modifier of patient-to-patient clinical variability in SCD. Patients with elevated numbers of dense erythrocytes are expected to have clinical courses of greater severity, because the intracellular concentration of sickle hemoglobin (HbS) influences its rate of polymerization after deoxygenation[139]. Indeed, a study carried out in ~500 SCD patients showed that higher DRBC was associated with increased risk of leg ulcer, priapism, and renal dysfunction[144]. Interestingly, DRBC is only partially correlated with HbF, suggesting that therapeutic modulation of DRBC could further reduce complications when combined with HbF-stimulating agents such as HU.

Several ion transporters and channels can control directly or indirectly RBC hydration (and thus density)[186]. Senicapoc, a selective inhibitor of the calcium-activated potassium Gardos channel, was shown in a mouse model of SCD to reduce the number of DRBC[187]. A phase III clinical trial of senicapoc in SCD patients similarly decreased the number of dense red blood cells, but failed to reduce the number of painful vaso-occlusive crises[150]. Strong interest nonetheless persists in the pursuit of identifying novel drug targets, inhibition of which would selectively re-hydrate erythrocytes in SCD patients. Evidence of the pathologic importance of dehydration in SCD erythrocytes continues to accumulate[188]. Human genetics can provide an unbiased approach to discover the role of proteins and biological pathways in RBC hydration. In this article, we describe results from the first genome-wide association study (GWAS) to identify DNA sequence variants associated with DRBC in SCD patients.

3.5 METHODS

3.5.1 Ethics Statement

Informed consent was obtained for all participants in accordance with the Declaration of Helsinki. This project was also reviewed and approved by the Montreal Heart Institute Ethics Committee and the different recruiting centers.

3.5.2 Samples and DNA Genotyping

The GEN-MOD study, a cohort of sickle cell disease (SCD) homozygous patients recruited in Paris, France, has been described elsewhere[144]. 408 GEN-MOD participants, for whom red blood cell density (DRBC) was measured at baseline using the phthalate density-distribution technique, were available for our genetic investigation. The DNA of the GEN-MOD participants was genotyped on the Illumina Infinium HumanOmni2.5Exome-8v1.1 array at the Montreal Heart Institute Pharmacogenomics Center. We used PLINK[154] and other custom scripts to control the quality of the genotyping dataset: we excluded samples and markers with genotyping success rate <95%, markers out of Hardy-Weinberg Equilibrium (P<1x10^{-7}) and markers with extreme (high or low) heterozygosity. We performed multidimensional scaling (MDS) in PLINK, anchoring these results on projections obtained using reference populations from the 1000 Genomes Project, to detect and remove (after visual inspection) population outliers. The Cooperative Study of Sickle Cell Disease (CSSCD) has been described extensively elsewhere[189-191]. Genome-wide genotype data generated with the Illumina Human610-Quad array was available for 1,279 CSSCD participants. We conducted genotype imputation using Minimac3 (v1.0.11)[165] and reference haplotypes from phase 3 of the 1000 Genomes Project. We restricted association testing to markers with an imputation r^2 >0.3.

3.5.3 Statistical analyses

The descriptive statistics of the participants analyzed in this study are presented in **Table 1**. Continuous phenotypes (DRBC and mean corpuscular hemoglobin concentration (MCHC)) were adjusted for sex and age, and the residuals were normalized using inverse normal transformation. Because low MCHC can be confounded by the thalassemia trait, we excluded from the analyses participants with α-thalassemia or a mean corpuscular hemoglobin (MCH)

<26 pg. We used linear regression for association testing between single variants and continuous traits, as implemented in RVtests (v.20140416)[179]. We used Sequence Kernel Association Test (SKAT)[173] and Variable Threshold (VT)[172] for our gene-based testing using rareMETALS(v.6.3)[178]. For gene-based testing, we focused our analysis on genotyped variants with minor allele frequency (MAF) <5%. We ran two sets of gene-based analyses: broad set (missense, nonsense, splice-site, frameshift and stop codon) and strict set (all of the above except missense variants). All genetic association analyses presented in this study were adjusted for the ten first principal components. Furthermore, we applied a genomic control correction to the DRBC GWAS results.

We defined genome-wide significance as $\alpha=5\text{x}10^{-8}$ and $\alpha=2.5\text{x}10^{-6}$ for single-variant and gene-based tests, respectively. In the post-hoc prioritization analyses (see below), we considered 12,360 erythroid enhancers ($\alpha=4\text{x}10^{-6}$ after Bonferroni correction) or expression quantitative trait loci (eQTL) for 66 candidate genes ($\alpha=8\text{x}10^{-4}$ after Bonferroni correction). For the 84 variants previously associated with MCHC by GWAS, and their linkage disequilibrium (LD) proxies, we highlighted variants with nominal significance ($\alpha=0.05$) given the strong prior probability of these loci contributing to RBC hydration.

3.5.4 Genetic and functional prioritization of genetic variants

Given the limited statistical power offered by our sample size, we sought to prioritize variants using independent genetic and functional genomic information. In GEN-MOD, DRBC is strongly correlated with MCHC (Pearson's $r=0.63$, $P=7\text{x}10^{-41}$, **Supplementary Figure 1**). Although MCHC is not a perfect proxy for DRBC, variants associated with RBC dehydration are expected to result in increased MCHC. Since hemoglobin concentration is one of the major factors influencing sickle hemoglobin (HbS) polymerization,[139] we tested the association of the top DRBC variants ($P_{\text{DRBC}}<1\text{x}10^{-6}$) for association with MCHC in GEN-MOD and the CSSCD. We also tested whether the variants associated with MCHC in a large genome-wide association study (GWAS) of European-ancestry non-anemic individuals[107] are associated with DRBC in SCD participants from GEN-MOD. For this lookup, we considered not only the sentinel MCHC GWAS variants, but also all variants in strong LD ($r^2>0.8$) in European populations from the 1000 Genomes Project.

We also prioritized variants that map to erythroid enhancers defined using DNAse I hypersentive sites and histone tail modifications[183]. Finally, we queried the GTEx database[192] to retrieve eQTL for 66 candidate genes. These genes were pre-selected based on their known and suspected roles in erythrocyte hydration. **Supplementary Table 1** lists these candidate genes and rationales for their inclusion in the study.

3.5.5 RNA extraction and qPCR

The protocols for in vivo mouse experiments were reviewed and approved by the IRCM Animal Care Committee (ACC #2014-27), which follow the regulations and requirements of the Canadian Council on Animal Care (CACC). Transgenic SAD sickle cell disease mice have been backcrossed for 49th generation on C57Bl/6J inbred mouse and were genotyped by hemoglobin analysis for the presence of human globin chains[193]. Adult male SAD (n=3) and control (n=3) bone marrow cells were obtained from femur flushed with PBS and 1% fetal bovine serum. Bone marrow cells were then centrifuged at 1400rpm for 5 minutes and flash frozen on liquid nitrogen.

We extracted RNA from mice bone marrow using the RNeasy Plus mini kit from Qiagen. RNA quality and concentration were measured by Agilent RNA 6000 Nano II assays (Agilent Technologies) on an Agilent 2100 Bioanalyzer and purity was assessed by Nanodrop. We reverse transcribed 1µg of total RNA using random primers and the MultiScribe Reverse Transcriptase from Applied Biosystems. We performed qPCR analysis using Platinum SYBR Green qPCR SuperMix-UDG (Life Technologies) on the CFX384 (Biorad) with the following thermal profile: 10 minutes at 95°C, and 40 cycles of: 30 seconds at 95°C, 30 seconds at 55°C and 45 seconds at 72°C following by a melt curve. Expression levels were measured and normalized in relation to the expression levels of the reference gene hypoxanthine-guanine phosphoribosyltransferase (*HPRT*) and ribosomal S16 using the ΔΔCq method[194] and the geNorm software. We obtained a M value of 0.871. For *Atp2b4*, we used Quantitect primer assay from Qiagen (QT00252532). The primer sequences are: *S16* forward (5’-AGGAGCGATTTGCTGGTGTGG-3’) and reverse (5’-GCTACCAGGGCCTTTGAGATG-3’); *Hprt* forward (5’-CAGCGTCGTGATTAGCGATG-3’) and reverse (5’-

CAGAGGGCCACAATGTGATGG-3’).

3.6 RESULTS

3.6.1 Genome-wide association study of red blood cell density

After quality-control and genotype imputation, we performed a genome-wide association study (GWAS) between ~31 million DNA sequence variants and red blood cell density (DRBC) in 374 sickle cell disease (SCD) patients from the GEN-MOD cohort (**Table 1**). Although our single variant analysis was adjusted using principal components, we noted a modest inflation of the test statistics (λ_{GC}=1.1, **Figure 1**). For this reason, we corrected the test statistics using genomic controls. **Table 2** presents results for loci and associated variants with P_{DRBC} <5x10^{-6}. Gene-based testing focused on directly genotyped coding variants with minor allele frequency (MAF) <5% identified no significant association with DRBC.

The gold standard validation of genetic association studies requires replication of the initially observed associations for the same phenotype and variant in an independent cohort. Unfortunately, we are unaware of any SCD cohorts of sufficient size to replicate our DRBC genetic results. For this reason, we explored the use of mean corpuscular hemoglobin concentration (MCHC) as a surrogate phenotype. DRBC and MCHC are highly correlated in SCD patients (**Supplementary Figure 1**), and high DRBC and MCHC each can reflect erythrocyte dehydration. Thus, a variant associated with DRBC might be predicted also to associate with MCHC.

First, we tested the association between the top variants associated with DRBC in GEN-MOD and MCHC in GEN-MOD. As expected for two correlated traits tested in the same individuals, several variants are associated with both DRBC and MCHC in GEN-MOD (**Table 2**). As an independent validation step, we performed the MCHC analysis in the Cooperative Study of Sickle Cell Disease (CSSCD). After excluding participants with α-thalassemia, which may independently affect MCHC, we identified 584 CSSCD participants with baseline MCHC and genotype data available. Only one of the 15 variants tested with P_{DRBC} <5x10^{-6} in GEN-MOD had a P_{MCHC} <0.05 and consistent direction of effect in the CSSCD: this variant, rs59264502, is common (MAF=46%) and intergenic (**Table 2**).

3.6.2 Variant prioritization

We implemented three strategies to increase the probability of finding robust genetic associations with DRBC. First, we considered variants mapping to erythroid enhancers, as defined by DNAse I hypersensitive sites and histone modifications[183]. Among the 12,360 regulatory elements tested, we found no variants more strongly associated with DRBC than would be expected by chance (**Figure 1**). Second, we retrieved from the GTEx resource[192] expression quantitative trait loci (eQTL) for 66 candidate genes, selected because they encode proteins with direct or indirect effects on red blood cell hydration (**Supplementary Table 1**). Three of these genes had eQTLs that were also associated with DRBC in SCD patients from GEN-MOD (at $P_{\mathrm{DRBC}} < 8\times10^{-4}$, Bonferroni correction for 66 genes), although none were significantly associated with MCHC in the CSSCD (**Table 2**). These three promising variants control the expression of the Mg^{2+} transporter *SLC41A3*, cytoskeletal protein *SPTB* (beta spectrin), and mechanosensitive cation channel *PIEZO1*.

Our final strategy to prioritize variants was to exploit the physiological link between DRBC and MCHC. We reasoned that some variants previously associated with MCHC by GWAS could also influence DRBC. A recent meta-analysis carried out in 173,480 participants of European ancestry identified 84 DNA sequence variants robustly associated with MCHC[107]. To accommodate ethnicity difference, we retrieved DRBC results for these 84 variants as well as for all variants in strong linkage disequilibrium (LD, $r^2 > 0.8$ in European-ancestry individuals from the 1000 Genomes Project). This query highlighted eight variants with $P_{\mathrm{DRBC}} < 0.05$ (**Table 3**).

One of these eight variants, rs1203972, is located near the α-globin locus on chromosome 16. This is promising since the presence of α-thalassemia is associated with fewer DRBC[144], although it is unknown whether this specific SNP is in LD with an α-thalassemia mutation. The most common cause of α-thalassemia in individuals of African ancestry is a 3.7-kb deletion that encompasses one of the genes (*HBA2*) encoding the α-chain of hemoglobin. Analyses of whole-genome sequence data from African populations in the 1000 Genomes Project showed this deletion is in LD with rs13335629[195]. However, rs1203972 and rs13335629 are not in LD in GEN-MOD (r^2=0.02), nor is rs13335629 associated with DRBC (P_{DRBC}=0.24).

3.6.3 *ATP2B4* and DRBC in SCD patients

The second interesting result arising from this analysis of MCHC-associated SNPs in the DRBC GWAS data is an intronic SNP at the *ATP2B4* locus. *ATP2B4*, also known as *PMCA4*, encodes the main calcium pump of erythrocytes. We recently showed that this SNP, rs10751450, strongly associated with MCHC in European populations[107] and with malaria susceptibility in African populations[196], is an erythroid-specific eQTL for *ATP2B4* (Samuel Lessard and G.L., unpublished). We compared *Atp2b4* expression in bone marrow of normal mice and SAD mice[193], a well-established mouse model of sickle cell disease (**Figure 2**). *Atp2b4* was not differentially expressed (*t*-test P=0.68). Next, we tested a potential differential impact on red cell volume in normal and SAD mice of *PMCA4* inhibition by aurintricarboxylicacid (ATA).

3.7 DISCUSSION

By genotyping 403 GEN-MOD individuals, we ran the first genome-wide scan of DRBC. We didn't identify new genomic loci, whether when testing variants one at time, or when testing them as a collection in a gene. Notwithstanding the lack of samples for replication, we leveraged the strong connection between DRBC and MCHC and investigated the association of MCHC in GEN-MOD and CSSCD. This analysis yielded only one significant DNA sequence variant in both cohorts with matching effect size directions, therefore constituting our only pseudoreplication of DRBC variants.

To further elucidate genetic modulator of DRBC, we prioritized variants evaluating erythroid specific enhancers, and eQTLs in candidate genes. Additionally, we cross-referenced DRBC associations results in GEN-MOD to those found in a large meta-analysis of non-anemic Europeans for MCHC. We note that our candidate gene approach, and cross-reference analysis provided promising results implicating mutations in genes previously reported to have a functional impact on dense red cell physiology. The exact impact of these variations in SCD remains suggestive, awaiting replication and functional characterization as we did with *PMCA4*.

The genome-wide association of DRBC can provide a window into a better comprehension of SCD severity, and broadly the osmotic regulation in red blood cell. In addition to the need for molecular characterization of our promising findings, the need for additional samples to unearth DRBC loci are current limitations of this study. From a clinical standpoint, while the Senicapoc[150] trial failed, it showed that targeting red cell transporter channels can effectively rehydrate red blood cells and reduce the rate of some SCD-related complications. In the same line of thought, identifying DRBC susceptibility loci can inform us on additional strategies to rehydrate red blood cell in SCD patients with the goal of eliminating all complications.

3.8 ACKNOWLEDGMENTS

We thank all participants for their contribution to this project. G.L. is funded by Biogen, the Canadian Institutes of Health Research (CIHR, MOP #123382), the Doris Duke Charitable Foundation, and the Canada Research Chair program. S.L.A. is funded by the Doris Duke Charitable Foundation. M.T. is funded by CIHR/Canadian Blood Services (MOP #3251163).

3.9 CONFLICT OF INTEREST

The authors declare no competing financial interests.

Table 1. Descriptive statistics of the GEN-MOD and CSSCD sickle cell disease participants analyzed in this study. For continuous variables, we provide the mean ± standard deviation and the number of participants with available data. NA, not available

Phenotype	GEN-MOD (N=408)	CSSCD (N=1279)
Males/females	185 / 223	616 / 663
Age, years	30 ± 9	13 ± 12
DRBC, %	13.1 ± 8.6	NA
MCHC, g/dL	34.5 ± 1.8	34.6 ± 1.16

Table 2. Top single variant association results with red blood cell density (DRBC) in 374 participants from GEN-MOD. We included in this table variants with P_{DRBC} <5x10^{-6} or variants that are expression quantitative trait loci (eQTL) for candidate genes in GTEx and have a P_{DRBC} <8x10^{-4} (**Methods**). Chr:Pos, genomic coordinates on build hg19; REF/ALT, reference and alternate alleles; AF, frequency of the alternate allele; BETA/SE, effect size (for the alternate allele) and standard error in standard deviation units.

rsID	Chr:Pos	REF/ALT	GEN-MOD, DRBC (N=374)			GEN-MOD, MCHC (N=317)		CSSCD, MCHC (N=584)			Gene	Annotation
			AF	BETA (SE)	P-value	BETA (SE)	P-value	AF	BETA (SE)	P-value		
Top association results												
rs4234795	4:7210802	A/G	0.94	-0.84 (0.15)	1.99x10^{-7}	-0.46 (0.17)	0.0062	0.94	0.03 (0.13)	0.80	*SORCS2*	intron
rs9714060	3:195487476	A/G	0.43	-0.39 (0.08)	7.43x10^{-7}	-0.11 (0.08)	0.19	0.44	-0.01 (0.07)	0.93	*MUC4*	intron
rs146893001	9:112181617	T/C	0.01	-2.04 (0.4)	1.29x10^{-6}	-1.33 (0.6)	0.028	0.004	-0.06 (0.48)	0.90	*PTPN3*	intron
rs7216169	17:5219511	C/T	0.22	0.45 (0.09)	1.36x10^{-7}	0.25 (0.1)	0.0087	0.22	0.05 (0.08)	0.54	*RABEP1*	intron
rs543023132	6:155973785	GTTTT/G	0.02	-1.54 (0.3)	1.37x10^{-6}	-0.68 (0.36)	0.061	0.022	-0.15 (0.21)	0.47	-	intergenic
rs144995469	14:57199082	C/T	0.03	-1.15 (0.23)	1.48x10^{-6}	-0.32 (0.26)	0.22	0.033	0.36 (0.18)	0.039	-	intergenic
rs74989317	21:35296139	T/A	0.04	-0.99 (0.2)	1.53x-10^{-6}	-0.42 (0.2)	0.041	0.045	-0.21 (0.15)	0.17	-	regulatory
rs73108077	20:30006859	T/C	0.06	-0.83 (0.17)	1.75x10^{-6}	-0.22 (0.2)	0.27	0.063	-0.07 (0.13)	0.58	*DEFB122*	downstream
rs114402357	13:22493635	C/T	0.01	2.03 (0.4)	1.78x10^{-6}	1.1 (0.45)	0.015	0.016	0.28 (0.25)	0.26	-	intergenic
rs77141833	1:159825190	T/C	0.03	-1.12 (0.22)	1.80x10-^{6}	-0.44 (0.28)	0.12	0.032	0.08 (0.18)	0.66	*VSIG8*	intron
rs62015549	15:71671418	C/T	0.01	-2.44 (0.49)	1.89x10^{-6}	-1.07 (0.73)	0.15	0.015	-0.25 (0.26)	0.33	*THSD4*	intron
rs76513454	1:218861569	G/C	0.01	-2.17 (0.43)	1.97x10^{-6}	-0.93 (0.73)	0.20	NA	NA	NA	-	intergenic
rs139628543	2:239053045	A/C	0.06	0.75 (0.15)	1.99x10^{-6}	0.22 (0.17)	0.19	0.05	0.08 (0.14)	0.59	*KLHL30*	intron
rs59264502	13:106846272	AT/A	0.46	0.37 (0.08)	2.39x10^{-6}	0.21 (0.08)	0.011	0.47	0.14 (0.06)	0.030	-	intergenic
rs147900370	1:115552925	A/C	0.04	-0.92 (0.18)	2.44x10^{-6}	-0.35 (0.21)	0.090	0.038	-0.09 (0.17)	0.59	-	intergenic

eQTL for candidate genes												
rs62270871	3:125672365	G/A	0.51	0.33(0.07)	2.60×10^{-5}	0.14 (0.08)	0.11	0.47	0.02 (0.07)	0.71	*ALG1L*	intron; eQTL for *SLC41A3*
rs146977005	14:65305030	G/GA	0.28	-0.33(0.09)	5.5×10^{-4}	-0.10 (0.09)	0.26	0.75	0.11 (0.07)	0.15	*SPTB*	intron; eQTL for *SPTB*
rs8048714	16:88809773	G/C	0.72	-0.3(0.08)	7.3×10^{-4}	0.08 (0.1)	0.45	0.25	0.14 (0.07)	0.056	*PIEZO1*	intron; eQTL for *PIEZO1*

Table 3. Top association results between variants previously associated with mean corpuscular hemoglobin concentration (MCHC) in non-anemic European-ancestry individuals and red blood cell density in 374 sickle cell disease patients. We included in this table variants with nominal P_{DRBC} <0.05. Chr:Pos, genomic coordinates on build hg19; REF/ALT, reference and alternate alleles; AF, frequency of the alternate allele; BETA/SE, effect size (for the alternate allele) and standard error in standard deviation units.

rsID	Chr:Pos	REF/ALT	GEN-MOD, DRBC (N=374)			Gene	Annotation
			AF	BETA (SE)	P-value		
rs144514173	1:205246482	TTTTG/T	0.108	0.37 (0.12)	0.0029	*TMCC2*	downstream
rs10751450	1:203650945	C/T	0.643	-0.25 (0.08)	0.0031	*ATP2B4*	intron
rs148303943	6:16263455	T/C	0.85	-0.32 (0.11)	0.0057	*GMPR*	intron
rs11421513	6:13901073	G/GT	0.689	-0.23 (0.08)	0.0074	-	intergenic
rs1203972	16:283232	T/C	0.658	-0.22 (0.08)	0.0082	*LUC7L*	upstream
rs201794926	8:145710909	G/GA	0.491	0.18 (0.07)	0.021	*PPP1R16A*	intron
rs34514965	19:13071559	T/TG	0.832	0.21 (0.1)	0.043	*GADD45GIP1*	upstream
rs5875087	6:26118437	CA/C	0.906	-0.27 (0.13)	0.045	*HIST1H2BC*	intron

Figure 1. Distribution of genome-wide association results with red blood cell density (DRBC) in 374 sickle cell disease patients. We present results for all imputed markers (pink), markers that map to erythroid enhancers (purple), markers that are expression quantitative trait loci (eQTL) for 66 candidate genes implicated in red blood cell hydration (brown), and markers associated with mean corpuscular hemoglobin concentration (MCHC) from previous genome-wide association studies (green). The grey area corresponds to the 95% confidence interval. λ_{GC}, genomic inflation factor.

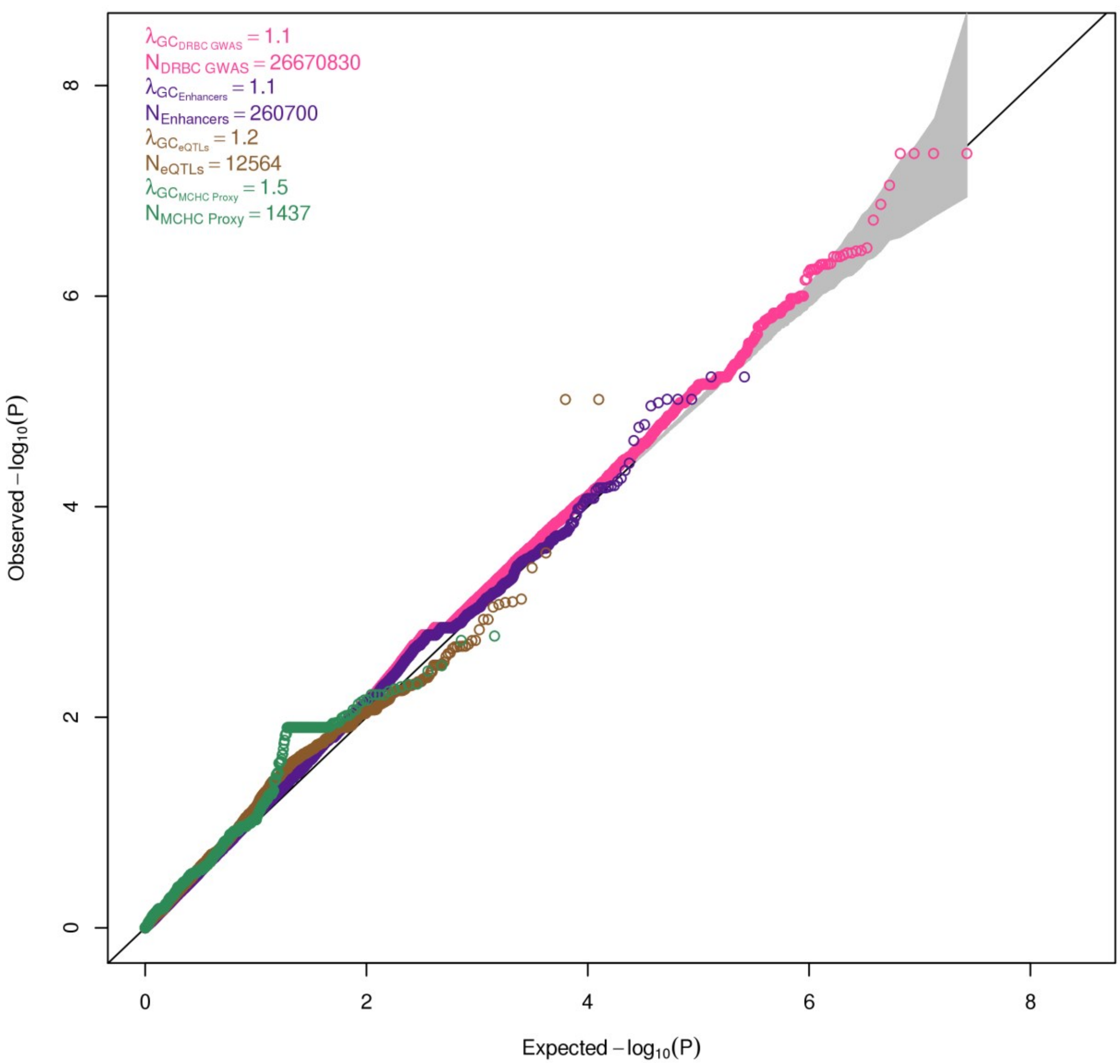
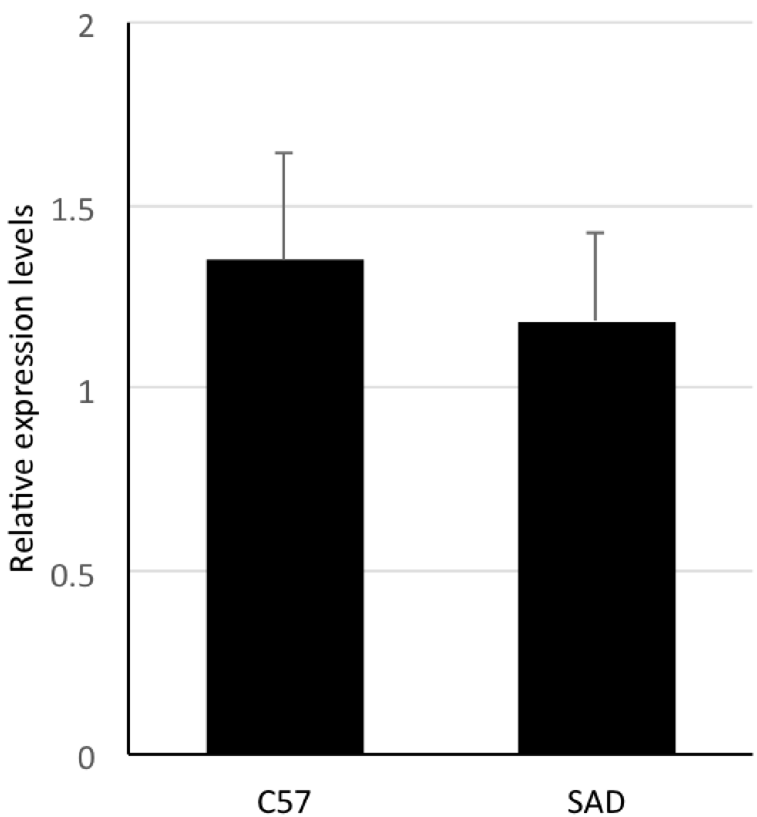
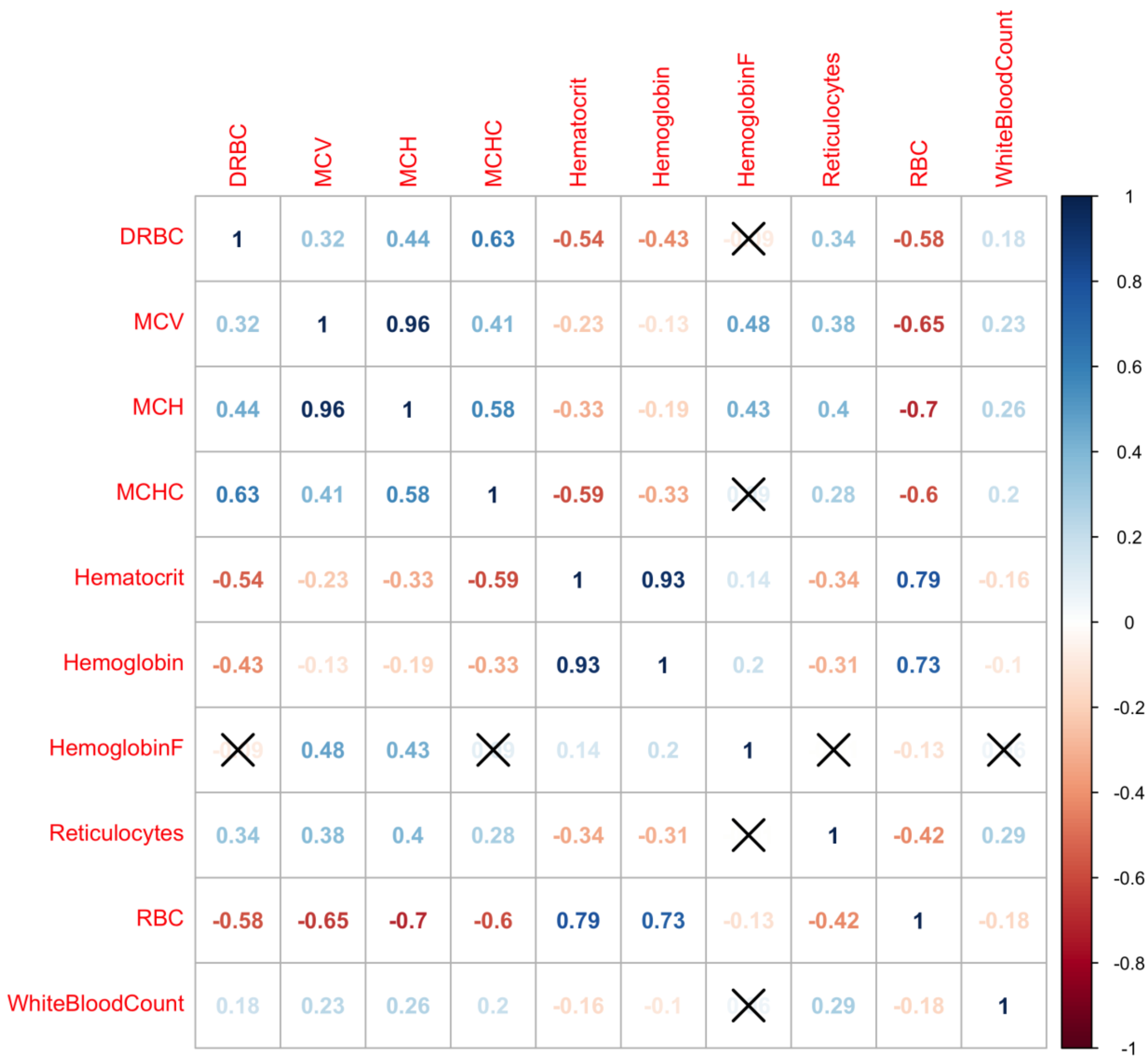


Figure 2. *Atp2b4* expression levels in the bone marrow of normal mice (C57) or a mouse model of sickle cell disease (SAD). RNA was extracted from both femurs of three C57 and three SAD mice. Data show mean ± standard error of the mean. *Atp2b4* is not differentially expressed between the bone marrow of C57 and SAD mice (*t*-test *P*=0.68).



Supplementary Figure 1. Correlations (Pearson's *r*) between hematological parameters corrected for age, and sex in up to 408 patients with sickle cell disease from the GEN-MOD cohort. Numbers in blue and red indicate positive and negative correlations, respectively. Cells with an "X" are non-significant correlations ($P \geq 0.05$).



Supplementary Table 1. List of candidate genes with a potential role in red blood cell hydration.

Official Gene Symbol	Gene Name	Rationale
ABCB6	ATP-binding cassette, sub-family B (MDR/TAP), member 6 (Langereis blood group)	Mutations in this nominal pyrrole/hemin transporter cause autosomal dominant pseudohyperkalemia and can cause hereditary xerocytosis with transient perinatal edema
ABCG5	ATP-binding cassette, sub-family G (WHITE), member 5	Mutations in this sterol hemitransporter cause hereditary sitosterolemia with stomatocytosis
ABCG8	ATP-binding cassette, sub-family G (WHITE), member 8	Mutations in this sterol hemitransporter cause hereditary sitosterolemia with stomatocytosis
ANK1	Ankyrin 1, Erythrocytic	Ankyrin - LOF mutations are the most common cause of hereditary spherocytosis
ANO1	Anoctamin 1, calcium activated chloride channel	ANO1/TMEM16A knockout murine RBC exhibit reduced Ca^{2+}-activated anion current
ANO6	Anoctamin 6	Ca^{2+}-activated anion channel with phospholipid flippase activity – Loss-of-function mutations cause Scott Syndrome
ATP1A1	ATPase, Na^+/K^+ transporting, alpha 1 polypeptide	Na^+/K^+ -ATPase alpha1 subunit
ATP1A2	ATPase, Na^+/K^+ transporting, alpha 2 polypeptide	Na^+/K^+ -ATPase alpha2 subunit
ATP1B1	ATPase, Na^+/K^+ transporting, beta 1 polypeptide	Na^+/K^+ -ATPase beta1 subunit
ATP1B2	ATPase, Na^+/K^+ transporting, beta 2 polypeptide	Na^+/K^+ -ATPase beta2 subunit
ATP2B1	ATPase, Ca^{2+} transporting, plasma membrane 1	PMCA1 is one of the calcium ATPases of the RBC membrane
ATP2B4	ATPase, Ca^{2+} transporting, plasma membrane 4	PMCA4 is one of the calcium ATPases of RBC membrane. It inhibited by vanadate; binding to maitotoxin elicits cation channel activity
CACNA1A	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	P/Q-type voltage-gated Ca^{2+} channel implicated by omega-agatoxin blockade of LPS-stimulated Ca^{2+} entry into RBC
CR1	Complement component (3b/4b) receptor 1 (Knops blood group)	Complement receptor 1 liganding elevates RBC calcium and triggers phosphorylation cascades
EPB41	Erythrocyte membrane protein band 4.1	Band 4.1 erythroid isoform - loss-of-function mutations cause elliptocytosis
EPB42	Erythrocyte membrane protein band 4.2	Band 4.2 LOF mutations cause spherocytosis in Japanese
FXYD2	FXYD domain containing ion transport regulator 2	Na^+/K^+ -ATPase gamma subunit.- also part of or regulator of a renal distal tubular Mg^{2+} channel
GRIN1	Glutamate receptor, ionotropic, N-methyl D-aspartate 1	NMDA receptors (NMDAr) have been defined pharmacologically in RBC membrane. Liganding elevates Ca^{2+} and promotes ATP release and can promote shape change. Antagonist memantine is under consideration for clinical trial in sickle disease
GRIN2A	Glutamate receptor, ionotropic, N-methyl D-aspartate 2A	NMDAr subunit
GRIN2B	Glutamate receptor, ionotropic, N-methyl D-aspartate 2B	NMDAr subunit
GRIN2	Glutamate receptor, ionotropic, N-methyl D-aspartate 2C	NMDAr subunit
GRIN2D	Glutamate receptor, ionotropic, N-methyl D-aspartate 2D	NMDAr subunit
KCNK5	Potassium channel, subfamily K, member 5	Two-pore domain (K2P) potassium channel in RBC proteome
KCNK6	Potassium channel, subfamily K, member 6	Two-pore domain (K2P) potassium channel in RBC proteome
KCNN4	Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4	Gardos Channel KCa3.1, gain-of-function mutations cause hereditary xerocytosis
KEL	Kell blood group, metallo-endopeptidase	Unknown function, mutant in neuro-acanthocytosis syndrome
NOX4	NADPH oxidase 4	Candidate contributor to oxidative damage to RBC membrane
NOX5	NADPH oxidase	Candidate contributor to oxidative damage to RBC membrane
P2RX7	Purinergic receptor P2X, ligand-gated ion channel, 7	Deoxygenation releases ATP from normal RBC and (to greater degrees) from sickle RBC, and RBC P2X7 may act as both cation channel and ATP permease
PANX1	Pannexin 1	Pannexin 1 (homologous connexin gap junction hemichannels) is a pH-gated permease for mid-range solutes, including ATP
PANX2	Pannexin 2	Relative of PANX1
PANX3	Pannexin 3	Relative of PANX1
PIEZO1	Piezo-type mechanosensitive ion channel component 1	Gain-of-function mutations in this mechanosensitive Ca-permeable cation channel cause autosomal dominant hereditary xerocytosis
PIEZO2	Piezo-type mechanosensitive ion channel component 2	Sensory neuron-predominant homolog of PIEZO1, expression in RBC uncertain
PKD2	Polycystic kidney disease 2 (autosomal dominant)	Ca^{2+} -permeable cation channel, expression in RBC uncertain - otherwise widely expressed as part of endoplasmic reticulum Ca release mechanism in addition to plasma membrane location

PRDX1	Peroxiredoxin 1	Redox regulators binding to RBC membrane and likely acting on membrane proteins, including transporters and channels
PRDX2	Peroxiredoxin 2	Relative of PRDX1
RHAG	Rh-associated glycoprotein	This component of Rh antigen has missense mutations associated with overhydrated stomatocytosis with increased RBC cation permeability
SLC12A4	Solute carrier family 12 (potassium/chloride transporter), member 4	KCC1; gain-of-function mutation associated with RBC dehydration and sickle disease exacerbation in mouse model
SLC12A6	Solute carrier family 12 (potassium/chloride transporter), member 6	KCC3; predominant K-Cl cotransporter of murine RBC
SLC12A7	Solute carrier family 12 (potassium/chloride transporter), member 7	KCC4
SLC2A1	Solute carrier family 2 (facilitated glucose transporter), member 1	Missense mutation of GLUT1 can cause overhydrated stomatocytosis with increased RBC cation permeability
SLC41A1	Solute carrier family 41 (magnesium transporter), member 1	Candidate Na^+/Mg^+ exchanger - intracell Mg regulates KCC activity
SLC41A2	Solute carrier family 41 (magnesium transporter), member 2	Candidate Na^+/Mg^+ exchanger
SLC41A3	Solute carrier family 41, member 3	Candidate Na^+/Mg^+ exchanger
SLC4A1	Solute carrier family 4 (anion exchanger), member 1 (Diego blood group)	Band 3 is the major intrinsic protein of RBC - Overhydrated stomatocytosis mutations associated with increased RBC cation leak
SPTA1	Spectrin, alpha, erythrocytic	Alpha-spectrin - LOF mutations cause herededitary spherocytosis
SPTB	Spectrin, beta, erythrocytic	Beta-spectrin - LOF mutations cause hereditary spherocytosis
STOM	Stomatin	Stomatin (STOM) protein deficiency is associated with some forms of hereditary spherocytosis; but STOM knock out without RBC phenotype in mouse.
STOML1	Stomatin (EPB72)-like 1	STOM-related protein
STOML2	Stomatin (EPB72)-like 2	STOM-related protein
STOML3	Stomatin (EPB72)-like 3	Stomatin-related protein whose oligomerization modulates PIEZO2 activity
TRPC1	Transient receptor potential cation channel, subfamily C, member 1	In murine RBC, one of the Ca^{2+}entry pathways
TRPC3	Transient receptor potential cation channel, subfamily C, member 3	Hetero-oligomerizes with TRPC6
TRPC6	Transient receptor potential cation channel, subfamily C, member 6	In murine RBC, one of the RBC Ca^{2+}entry pathways
TRPM2	Transient receptor potential cation channel, subfamily M, member 2	Mg^{2+} -permeable cation chanzyme expressed widely
TRPM4	Transient receptor potential cation channel, subfamily M, member 4	Transient receptor potential family cation channel, RBC expression uncertain
TRPM7	Transient receptor potential cation channel, subfamily M, member 7	Mg^{2+} cation permease/chanzyme
TRPV1	Transient receptor potential cation channel, subfamily V, member 1	Transient receptor potential family cation channel activated by vanilloids, expression in RBC uncertain
TRPV4	Transient receptor potential cation channel, subfamily V, member 4	Transient receptor potential family cation channel activated by hypotonic swelling, expression in RBC uncertain
TRPV5	Transient receptor potential cation channel, subfamily V, member 5	Ca-selective channel of distal convoluted tubule - not known to be expressed in RBC
TRPV6	Transient receptor potential cation channel, subfamily V, member 6	Ca-selective channel of enterocytes - not known to be expressed in RBC
TUSC3	Tumor suppressor candidate 3	Candidate Mg transporter of unknown mechanism and RBC expression
VDAC1	Voltage-dependent anion channel 1	Nonspecific large-pore channel of mitochondrial inner membrane has also been reported in RBC membrane
VDAC2	Voltage-dependent anion channel 2	VDAC1 homolog
VDAC3	Voltage-dependent anion channel 3	VDAC1 homolog

4. Whole-Exome Sequencing of Sixty-Four Patients with Sickle Cell Disease

4.1. Motivation

As mentioned earlier, manifestations of SCD-related complications become noticeable as soon as the protective effects of HbF dwindle. Whole-exome sequencing is a powerful method to identify rare coding mutations with moderate-to-large effect size, and to personalize therapies for patients[197,198]. We present a report of whole-exome sequencing in 64 patients from the GEN-MOD cohort, connecting mutations in candidate genes to SCD-related complications.

4.2. Methods

Cohort selection. 64 homozygous (β^S/β^S) patients were selected from the GEN-MOD cohort[144] at the top 10% and bottom 10% of HbF distribution from another study on HbF.

Variant Annotation and Selection. We annotated all the variants, using Variant Effect Predictor (VEP) from Ensembl[199], querying allele frequency across all ethnicities, and for the African/African American (AFR) population from Exome Aggregation Consortium (ExAC)[200]. Additionally, we included protein prediction and scores from SIFT and PolyPhen[201,202] which assess the probable impact of a nonsynonymous mutation on the protein function. We then selected autosomal variants annotated as nonsynonymous and which gene symbol corresponded to a subset of our candidate gene (**Supplementary Table1**).

Variant Scoring and Association to DRBC. To link the variants to the hematological traits, we calculated the average HBG, HbF, Retic, HCT, WBC, MCV, MCH, and MCHC across individuals carrying a mutation. We did the same for DRBC which we inversed normal transformed correcting for age and sex.

Targeted Exon Capture and Whole Exome Sequencing. Targeted regions of genomic DNA were captured using NimbleGen SeqCap EZxome V3.0 solution-based capture system as specified by the company's protocol. This was performed at the Montreal Heart Institute, pharmacogenomics center. The captured, purified, and amplified libraries targeting the exomes from SCD patients were sequenced on Illumina HiSeq with paired-end sequencing at 100bp read length.

Exome Sequence QC Analysis. Sequencing reads were aligned to the human genome (version hg 19/build 37) using BWA mem 0.5.9a software[175]. We followed the most current best practices recommendations that exist for GATK[176,203] to complete variant calling, recalibration, and to remove duplicates.

4.3. Results and Discussion

4.3.1 Cohort description

Our analysis included individuals with extreme DRBC values (**Figure 14**). However, 9 out of 64 selected patients had missing DRBC measures. Although they remained in our exome sequencing analysis, they were not used when computing the hematological traits' averages across individuals per variant.

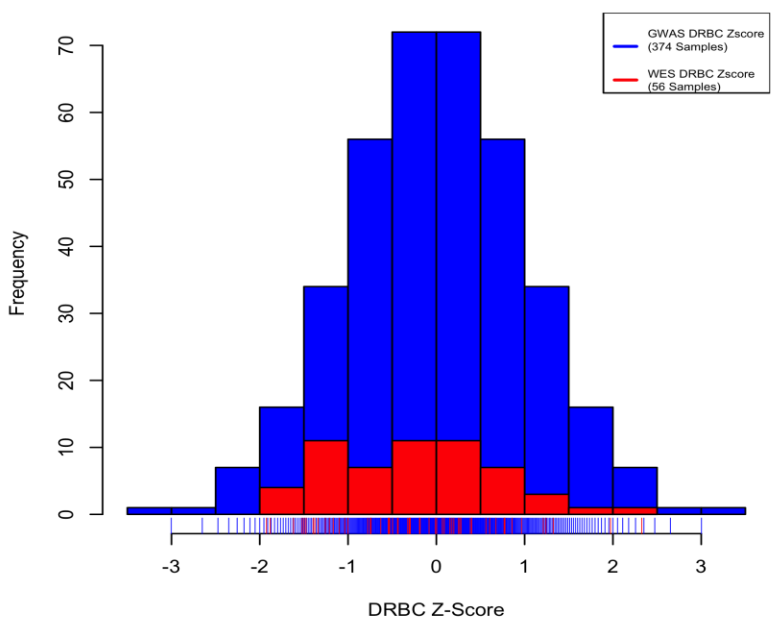


Figure 14 Whole-Exome Sequencing Dense Red Blood Cell Distribution in GEN-MOD

4.3.2 Data Mining Variant Annotation and Correlation

Top DRBC z-score. Post variant annotation, we prioritized 297 nonsynonymous variants that map to our candidate genes (**Supplementary Table1**) for further analysis. We visually correlated all these variants against the nine hematological traits mentioned earlier (**Figure 15**). The purpose of this representation is to single out mutations carried by individuals that could explain the severity of the disease. Overall, **Figure 15** shows that the more the variants are shared amongst all the individuals the more the DRBC z-score is 0. Meaning that several of the shared mutations don't explain the disease severity. However, as we move towards increasing or decreasing DRBC z-score values we can see that hematological traits increase or decrease depending on the correlation between DRBC and the blood trait (**Supplementay Figure 1**). The approach is most evident for MCV, MCHC, HCT and RETIC, as opposed to the other traits.

Focusing on variants with DRBC z-score > 1, we found 7 rare DNA sequences (**Table 4**), that could explain the role of dense cells in SCD severity. A patient with a history of leg ulcers, priapism, gallbladder removal, and sceptic necrosis was identified with two mutations in different genes. One of them in *ATP1B2*, a heterozygous replacement of A to a G (rs531342420) substituting a Gln (CAG) for an Arg(CGG) amino acid at the 108 position. The other mutation predicted to be damaging by both SIFT and Polyphen, replaced of a G for an A in *SPTB*, changing an Arg (CGG) to Trp (TGG) amino acid at the 44 position. The patient in which these variations were uncovered has the highest DRBC measure (37%) of the analyzed cohort. According to study by Bartolucci et al[144], denser erythrocyte lead to an increased incidence of SCD-related pathology, particularly skin ulcer, priapism, and renal dysfunction. Plus, although the in silico prediction of the *ATP1B2* missense change is non-pathogenic, the gene is also known as AMOG (Ca^{2+}-dependent adhesion molecule of glia) expressed abundantly in the brain and the retina. The protein binds to retinoschisin, a retinal degeneration gene for X-linked human juvenile retinoschisis.

Another patient with a medical history of stage 1 or 2 retinopathy, and priapism was identified with two rare mutations predicted to be deleterious by both prediction algorithms (**Table 4**), in *SLC12A7* and *SPTA1*. The mutant allele is found in the solute carrier KCC4 gene. It's an heterozygous replacement of G to an A (rs146681871). It changes Ser (TCG) to Leu (TTG) at the 84 position. The other mutant allele is found in the protein that encodes the spectrin alpha, erythrocytic 1, and substitutes an A for a T (rs146681871), changing Leu (CTG) to Gln (CAG) at the 1565 position.

Additionally, three rare DNA sequence variations were identified in *SLC12A7* (rs139369204, COSM4127004), *TRMP7* (rs202245737), and *ABCB6* (rs113159519) genes in the same individual with a history of gall bladder removal and stage 1 or 2 of retinopathy. None of the previously mentioned mutations were predicted to be detrimental by our prediction algorithms. Finally, the same SNP (rs34246477) in *PIEZO1*, the mammalian mechanosensory protein, was predicted to be deleterious by SIFT but benign by PolyPhen and had two different carriers.

Other Noteworthy Mutations

We shed light on four previously reported pathogenic mutations associated with hereditary spherocytosis (HS). In fact, mining the patient's medical history, revealed a prior of a combination of at least two of the following five complications; leg ulcer, gall blader removal, retinopathy, priapism or aseptic necrosis. Two SNPs in *ANK1* gene known as the ankyrin Brüggen mutation (rs2304877, COSM3982542, COSM3982541), and other one known as the Tubarao mutation (rs35213384)[204,205]. Additionally, two others missense mutations known as the Montefiore (CM930673, rs45562031, BGMUT_178) and the Tuscaloosa (CM920621, rs28931583)[206-208] were in *SLC41A* gene. Finally, a SNP (rs145343957) in *PKD2* the autosomal dominant polycystic kidney disease gene was found in an individual who exhibited four of the complications mentioned earlier. By and large, these mutations are linked to blood disorder other than SCD, they affect red cell membrane deformability, so these results could potentially highlight the concomitance of SCD and HS in GEN-MOD a phenomena previously reported in SCD patients[209].

Future Directions

Sequencing additional individuals with extreme DRBC values will help confirm our current findings, and unearth additional findings. Although, functional experiments need to be undertaken to ascertain the causal link between these DNA sequence variations and their impact on SCD complications, and on red cell density, more prioritization of variants through functional annotation can be undertaken. Based on the inconsistencies of SIFT and PolyPhen algorithms[210], it would be of interest to lookup additional severity prediction tools (e.i., FATHMM-MKL, MutationTaster2, Mutation Assessor, PROVEAN, CADD) and then identify their consensus, and their differences in term of prediction[210]. Finally, as we sequence the exome of more individuals we will gain enough statistical power to run association testing on this dataset.

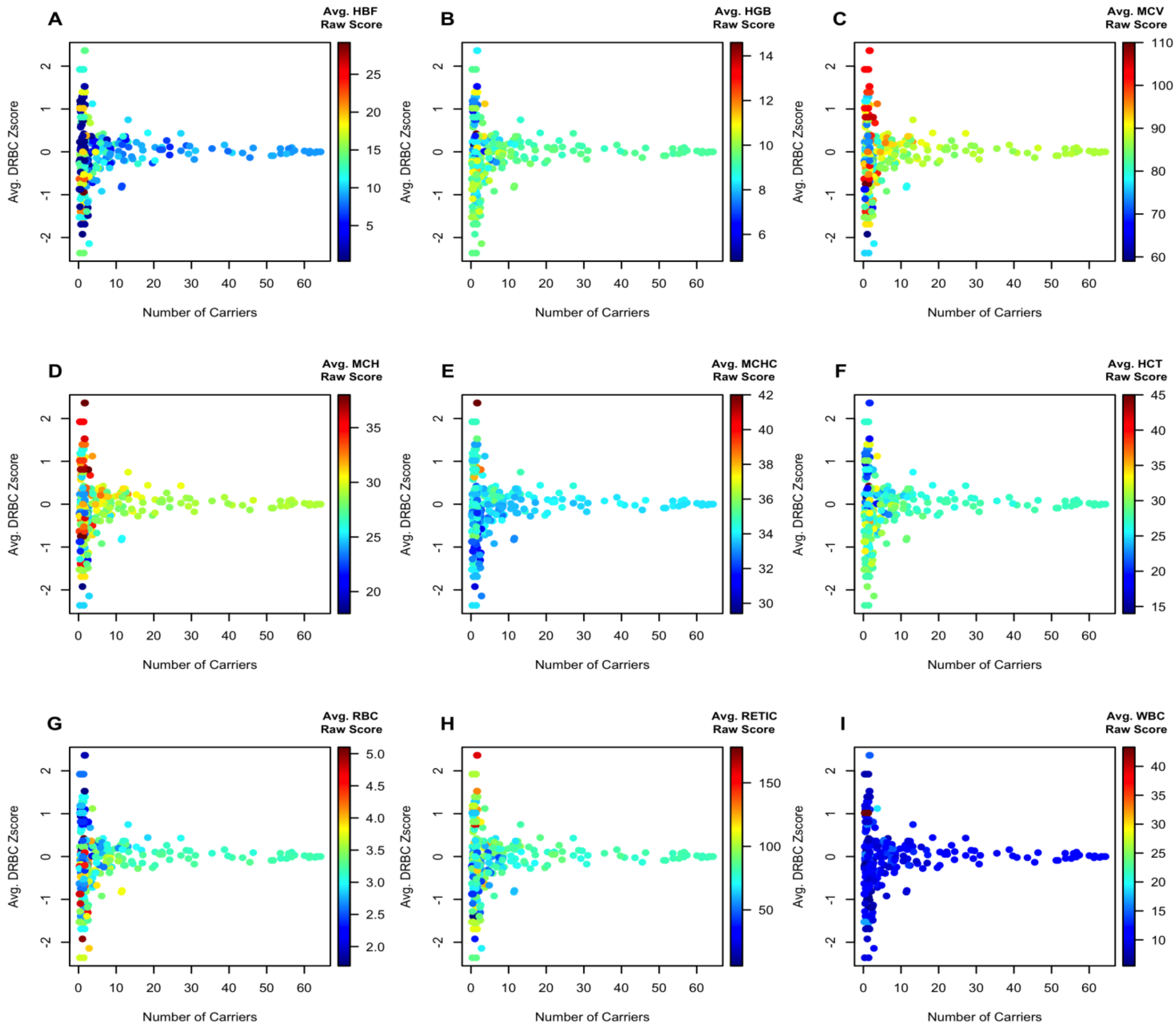


Figure 15. Visual Correlation of the z-score of Dense Red Blood Cell to Hematological Traits

Variants color-coded based on gradient of the average hematological traits. The x-axis represents the number of individuals with carrying a mutation and the y-axis shows the average z-score DRBC for each mutation.

SNP	N Carriers	Avg. zDRBC	Avg. DRBC	REF/ALT	Gene Symbol	Existing variation	SIFT/PolyPhen	Annotation Details	ExAC AFR MAF
chr17:7557240	1	2.36	37	A/G	ATP1B2	rs531342420	tolerated(0.26)/benign(0.004)	Gln108Arg [exon3]	T:0, G:9.6e-05
chr14:65289683	1	2.36	37	G/A	SPTB	-	deleterious(0)/probably_damaging(1)	Arg44Trp [exon1]	-
chr16:88792725	2	1.52	24	G/A	PIEZO1	rs34246477	deleterious(0.04)/benign(0.079)	Ala1312Val [exon27]	A:0.011
chr15:50884381	1	1.52	24	A/G	TRPM7	rs202245737	tolerated_low_confidence(0.08)/possibly_damaging(0.447)	Ser1351Pro [exon26]	G:0.0013
chr2:220081152	1	1.52	24	G/C	ABCB6	rs113159519	tolerated(0.19)/benign(0.005)	Leu302Val [exon4]	C:0.0011
chr5:1065523	1	1.52	24	C/T	SLC12A7	rs139369204	tolerated(0.1)/probably_damaging(0.991)	Arg771Gln [exon18]	T:0.00070
chr5:1093739	1	1.92	25	G/A	SLC12A7	rs146681871	deleterious(0.01)/probably_damaging(0.999)	Ser84Leu [exon3]	A:0.0020
chr1:158589121	1	1.92	25	G/A	SPTA1		deleterious(0.01)/benign(0.212)	Arg2141Trp [exon45]	-
chr1:158612244	1	1.92	25	A/T	SPTA1	rs202217097	deleterious(0)/probably_damaging(1)	Leu1565Gln [exon33]	T:0.0032
chr1:158612618	1	1.92	25	C/G	SPTA1	rs143779235	tolerated(1)/benign(0)	Ala1531Pro [exon32]	G:0.0032
chr8:41557033	4	0.097	12	C/T	ANK1	rs34523608	tolerated(0.73)/benign(0.001)	Arg832Gln [exon23]	T:0.020
chr1:158592847	22	0.051	11	G/A	SPTA1	rs78394850	deleterious(0)/probably_damaging(0.965)	Arg2016Cys [exon43]	A:0.14
chr1:158592901	22	0.051	11	C/G	SPTA1	rs77877855	deleterious(0.01)/possibly_damaging(0.882)	Ala1998Pro [exon43]	G:0.15
chr8:41566438	4	0.20	13	C/T	ANK1	rs2304877	deleterious(0.01)/benign(0.012)	Arg619His [exon17]	T:0.054
chr8:41552213	4	0.21	13.5	G/A	ANK1	rs35213384	tolerated(0.12)/benign(0.01)	Thr1075Ile [exon28]	A:0.050
chr17:42338993	1	-1.70	2	C/T	SLC4A1	rs45562031	tolerated(0.45)/benign(0.173)	Glu40Lys [exon4]	T:0.0028, T:0.0028
chr17:42335888	1	-0.87	0	G/C	SLC4A1	rs28931583	tolerated(0.14)/possibly_damaging(0.549)	Pro327Arg [exon10]	C:0.0023
chr4:88989102	1	-0.091	12	G/A	PKD2	rs145343957	tolerated(0.09)/probably_damaging(0.934)	Ser804Asn [exon13]	A:0.0080
chr17:42328598	8	-0.018	7.25	C/T	SLC4A1	rs5026	tolerated(0.48)/benign(0.014)	Val862Ile [exon19]	T:0.10
chr17:42334822	1	0.09	10	C/T	SLC4A1	rs45568837	deleterious(0)/probably_damaging(1)	Glu508Lys [exon13]	T:0.0048

Table 4. Top DRBC Missense Mutations From WES

The table provides the top DRBC missense mutation: SNP: The position of the mutation in the genome; N Carriers: number of individuals carrying a mutation; SNP, Avg. zDRBC: average z-score across individuals carrying a mutation; Avg. DRBC, average DRBC across individuals carrying a mutation; REF, the allele from the reference human genome; ALT, the non-reference allele observed in the sample; Gene Symbol, symbol of the protein sequence; Existing Variation, rsID, SIFT, prediction algorithm determines whether an amino acid substitution affects protein function; PolyPhen, prediction algorithm determines whether an amino acid substitution affects protein function; Annotation Details, details on the amino acid substitution and position; ExAC AFR MAF, minor allele frequency for the African/African American populations from the Exome Aggregation Consortium.

5. Discussion

5.1 Aims

The scope of this thesis was to provide empirical evidence of the genetic determinants of DRBC. The objectives were to identify common variants of small-to-moderate effect, in addition to rare variants with moderate-to-large effect associated with DRBC. Identifying the genetic modulator of DRBC could help identify erythrocyte channels and other pathway that could be targeted for therapies in SCD.

We conducted a genome wide association study of 374 patients testing ~31 million imputed markers for association with DRBC. We prioritized variants looking at erythrocytes specific enhancers, and eQTLs for candidate genes involved in erythrocyte osmotic regulation. Plus, we sought to exploit the correlation between DRBC and MCHC, by retrieving variations in the largest GWAS of blood traits in non-anemic individuals. Additionally, we sequenced the exomes of a subset of individuals with extreme DRBC values looking for a relationship between genes involved with moving solutes and water across erythrocytes membrane to DRBC, and then to SCD-related complications. Contrary to the GWAS analysis, the aim was to find rare coding variants that could explain that increased levels of DRBC lead to increase severity of SCD.

5.2 Significance of results

Single-variants association tests, and gene-based tests did not yield any markers that reached genome-wide association threshold. Our prioritization strategy highlighted *ATP2B4* the main calcium pump in erythrocytes, but functional experiments did not provide conclusive results for the role of plasma membrane calcium pump in SCD severity. Our WES experiment suggests that there is a direct relationship between DRBC and SCD. Indeed, we found rare DNA sequences mutations in *ATP1B2, PIEZO1, SPTA1* and *SLC12A7* in patients with multiple acute complications. Also, we found carriers of well characterized hereditary spherocytosis mutations in *ANK1* and *SLC41A*. This indicates the presence of a concomitance of spherocytosis and sickle cell in our cohort, which is a rare but previously reported event. Amongst our total sample size other individuals with more extreme DRBC values are present. Validating the exome sequencing approach by adding more samples at both end of the distribution tails would cement our findings and potentially open the door for the identification of specific pathways or transporters that can be targeted for therapies. Generally, our research findings suggest that DRBC should be more often recorded in routine blood test in SCD patients the same way MCV, MCH, and RETIC are. Systematically measuring DRBC in patients might be a better biomarker and indicator of severity than any other ones mentioned in the context of SCD. Clinical trials targeting transporter channels like the Gardos channel for Senicapoc (ICA-17043) showed that modulating erythrocytes volume might be the most cost-effective and reliable approach to a cure for the vast majority of SCD individuals.

5.3 Strength and limitations

Our scientific inquiry has two major strength; first we are pioneering the genetic investigation on the density of red blood cell in SCD. Therefore, our work can aid future discoveries or clinical trials to guide their efforts away from negative results or towards new exciting questions. Our second strength is the consistency of our analysis pipelines. I developed automated routines for the prioritizations of the variants, and the visualization of summary statistics. Therefore making this work easily reproducible, and verifiable by others.

In terms of limitations, the first one is our small sample size. This greatly limits our ability to replicate our current findings, which are essential for validating our results. The fact that DRBC is rarely recorded during full blood work exacerbates the challenge of performing a replication. The second limitation is our need for statistical power to identify common and rare variants with small effect size. In fact, according Zuk et al[211], well-powered GWAS discovery sample sizes for common and rare variants associations test should have at least 25,000 cases with equally large replication sample size. This recommendation for half of 100,000 samples for both discovery and replication is more plausible today by joining large consortium such as the one on height, or obesity than it was 10 years ago. However, in the context of SCD, to date the largest GWAS analyzed fetal hemoglobin[116] and was composed of less than 2,000 individuals.

5.4 Recommendations

Based on the limitations mention earlier, the most straightforward recommendation is to increase our sample size. As our power calculation showed, adding enough patients to reach a sample size of 1,000 would significantly improve our ability to find loci attributable to DRBC. Therefore, providing empirical candidate genes for functional experiments to reverse dehydration or rehydrate cells in animal models and eventually clinical trials. Another recommendation that could complement our grasp SCD severity and DRBC would be to acquire red blood cell metabolism data. This would capture small molecule reactions that are potentially associated with DRBC. In fact, metabolomics experiments have shown promising results in other complex traits disorders such as obesity and type-2 diabetes. In our context, we could seek to capture metabolomics signature of erythrocyte density or of volume control to fine tune our understanding of red blood osmotic regulation as it pertains to SCD, with implications to other blood disorders.

Bibliography

1. Ingram V.M. Abnormal human haemoglobins. I. The comparison of normal human and sickle-cell haemoglobins by fingerprinting. . *Biochim Biophys Acta* 1958;28:539–545.
2. Herrick J.B. Peculiar elongated and sickle-shaped red blood corpuscules in a case of severe anemia. *Arch Int Med.* 1910;6:517–521.
3. Hanh EV, Gillespie EB. Sickle cell anemia. *Arch Int Med.* 1927;39:233.
4. Scriver J.R. WTR. Studies on a case of sickle cell anemia. . *Can Med Assoc J* 1930(23):375–380.
5. Watson J. The significance of the paucity of sickle cells in newborn Negro infants. *Am J Med Sci.* 1948;215(4):419-423.
6. Pauling L, Itano HA, et al. Sickle cell anemia, a molecular disease. *Science.* 1949;110(2865):543-548.
7. Neel JV. The Inheritance of Sickle Cell Anemia. *Science.* 1949;110(2846):64-66.
8. Allison AC. Sickle cells and evolution. *Scientific American.* 1956;195:87–94.
9. Ingram V.M. Abnormal human hemoglobins. The chemical difference between normal and sickle cell hemoglbins. *Biochimica et biophysica acta.* 1959;36:402-411.
10. Ferrone FA. Polymerization and sickle cell disease: a molecular view. *Microcirculation.* 2004;11(2):115-128.
11. Organization WH. *Fifty-ninth World Health Assembly: resolutions and decisions, annexes.* Geneva: World Health Organization;2006.
12. Grosse SD, Odame I, Atrash HK, Amendah DD, Piel FB, Williams TN. Sickle cell disease in Africa: a neglected cause of early childhood mortality. *Am J Prev Med.* 2011;41(6 Suppl 4):S398-405.
13. Piel FB, Hay SI, Gupta S, Weatherall DJ, Williams TN. Global burden of sickle cell anaemia in children under five, 2010-2050: modelling based on demographics, excess mortality, and

interventions. *PLoS Med.* 2013;10(7):e1001484.

14. Allison AC. Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J.* 1954;1(4857):290-294.

15. Piel FB, Patil AP, Howes RE, et al. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat Commun.* 2010;1:104.

16. Antonarakis SE, Orkin SH, Kazazian HH, Jr., et al. Evidence for multiple origins of the beta E-globin gene in Southeast Asia. *Proc Natl Acad Sci U S A.* 1982;79(21):6608-6611.

17. Pagnier J, Mears JG, Dunda-Belkhodja O, et al. Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc Natl Acad Sci U S A.* 1984;81(6):1771-1773.

18. Platt OS, Thorington BD, Brambilla DJ, et al. Pain in sickle cell disease. Rates and risk factors. *N Engl J Med.* 1991;325(1):11-16.

19. Schechter AN. Hemoglobin research and the origins of molecular medicine. *Blood.* 2008;112(10):3927-3938.

20. Sankaran VG, Menne TF, Xu J, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science.* 2008;322(5909):1839-1842.

21. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. *Cold Spring Harb Perspect Med.* 2013;3(1):a011643.

22. Cox FE. History of the discovery of the malaria parasites and their vectors. *Parasit Vectors.* 2010;3(1):5.

23. Malaria GH-DoPDa. The History of Malaria, an Ancient Disease. 2016; Malaria. Available at: https://www.cdc.gov/malaria/about/history/. Accessed December 20 2016.

24. Haas LF. Neurological stamp. Charles Louis Alphonse Laveran (1845-1922). *J Neurol Neurosurg Psychiatry.* 1999;67(4):520.

25. Schlagenhauf P. Malaria: from prehistory to present. *Infect Dis Clin North Am.* 2004;18(2):189-205.

26. Martinsen ES, Perkins SL, Schall JJ. A three-genome phylogeny of malaria parasites (Plasmodium and closely related genera): evolution of life-history traits and host switches. *Mol Phylogenet Evol.* 2008;47(1):261-273.

27. Hayakawa T, Culleton R, Otani H, Horii T, Tanabe K. Big bang in the evolution of extant malaria parasites. *Mol Biol Evol.* 2008;25(10):2233-2239.

28. Escalante AA, Freeland DE, Collins WE, Lal AA. The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. *Proc Natl Acad Sci U S A.* 1998;95(14):8124-8129.

29. Perkins SL, Schall JJ. A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *J Parasitol.* 2002;88(5):972-978.

30. Vargas-Serrato E, Corredor V, Galinski MR. Phylogenetic analysis of CSP and MSP-9 gene sequences demonstrates the close relationship of Plasmodium coatneyi to Plasmodium knowlesi. *Infect Genet Evol.* 2003;3(1):67-73.

31. White NJ, Pukrittayakamee S, Hien TT, Faiz MA, Mokuolu OA, Dondorp AM. Malaria. *The Lancet.* 2014;383(9918):723-735.

32. Gazzinelli RT, Kalantari P, Fitzgerald KA, Golenbock DT. Innate sensing of malaria parasites. *Nat Rev Immunol.* 2014;14(11):744-757.

33. Wassmer SC, Grau GE. Severe malaria: what's new on the pathogenesis front? *Int J Parasitol.* 2017;47(2-3):145-152.

34. Gething PW, Casey DC, Weiss DJ, et al. Mapping Plasmodium falciparum Mortality in Africa between 1990 and 2015. *N Engl J Med.* 2016;375(25):2435-2445.

35. Organization WH. Fact Sheet: World Malaria Report 2015. 2015; http://www.who.int/malaria/media/world-malaria-report-2015/en/.

36. (CDC) CFDC. Impact of Malaria. 2016; https://www.cdc.gov/malaria/malaria_worldwide/impact.html.

37. Cholera R, Brittain NJ, Gillrie MR, et al. Impaired cytoadherence of Plasmodium falciparum-infected erythrocytes containing sickle hemoglobin. *Proc Natl Acad Sci U S A.* 2008;105(3):991-996.

38. Taylor SM, Parobek CM, Fairhurst RM. Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis. *Lancet Infect Dis.* 2012;12(6):457-468.

39. Ayi K, Min-Oo G, Serghides L, et al. Pyruvate kinase deficiency and malaria. *N Engl J Med.* 2008;358(17):1805-1810.

40. Baruch DI, Ma XC, Pasloske B, Howard RJ, Miller LH. CD36 peptides that block cytoadherence define the CD36 binding region for Plasmodium falciparum-infected erythrocytes. *Blood.* 1999 94(6):2121-2127.

41. Ferreira A, Marguti I, Bechmann I, et al. Sickle hemoglobin confers tolerance to Plasmodium infection. *Cell.* 2011;145(3):398-409.

42. Aitman TJ, Cooper LD, Norsworthy PJ, et al. Malaria susceptibility and CD36 mutation. *Nature.* 2000;405(6790):1015-1016.

43. Fry AE, Ghansa A, Small KS, et al. Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. *Hum Mol Genet.* 2009;18(14):2683-2692.

44. Sambo MR, Trovoada MJ, Benchimol C, et al. Transforming growth factor beta 2 and heme oxygenase 1 genes are risk factors for the cerebral malaria syndrome in Angolan children. *PLoS One.* 2010;5(6):e11141.

45. Reiter CD, Wang X, Tanus-Santos JE, et al. Cell-free hemoglobin limits nitric oxide bioavailability in sickle-cell disease. *Nat Med.* 2002;8(12):1383-1389.

46. Rees DC, Williams TN, Gladwin MT. Sickle-cell disease. *Lancet.* 2010;376(9757):2018-2031.

47. Saraf SL, Molokie RE, Nouraie M, et al. Differences in the clinical and genotypic presentation of sickle cell disease around the world. *Paediatr Respir Rev.* 2014;15(1):4-12.

48. Serjean G, Serjeant B. *Sickle cell disease.* Oxford University Press; 3 edition; 2001.

49. Moo-Penn W, Bechtel K, Jue D, et al. The presence of hemoglobin S and C Harlem in an individual in the United States. *Blood.* 1975;46(3):363-367.

50. Nagel RL, Daar S, Romero JR, et al. HbS-oman heterozygote: a new dominant sickle syndrome. *Blood.* 1998;92(11):4375-4382.

51. Witkowska HE, Lubin BH, Beuzard Y, et al. Sickle cell disease in a patient with sickle cell trait and compound heterozygosity for hemoglobin S and hemoglobin Quebec-Chori. *N Engl J Med.* 1991;325(16):1150-1154.

52. Nagel RL, Fabry ME, MH S. The paradox of hemoglobin SC disease. *Blood Rev.* 2003(3):167-178.

53. Masiello D, Heeney MM, Adewoye AH, et al. Hemoglobin SE disease: a concise review. *Am J Hematol.* 2007;82(7):643-649.

54. Geva A, Clark JJ, Zhang Y, Popowicz A, Manning JM, Neufeld EJ. Hemoglobin Jamaica Plain—A sickling hemoglobin with reduced oxygen affinity. *N Engl J Med.* 2004;351(15):1532-1538.

55. NIH. What Are the Signs and Symptoms of Sickle Cell Disease? *Explore Sickle Cell Disease* 2016; https://www.nhlbi.nih.gov/health/health-topics/topics/sca/signs. Accessed Dec 1, 2016.

56. Bainbridge R, Higgs DR, Maude GH, Serjeant GR. Clinical presentation of homozygous sickle cell disease. *J Pediatr.* 1985;106(6):881-885.

57. Benjamin LJ DC, Jacox AK et al *Guideline for the management of acute and chronic pain in sickle cell disease.* Glenview, IL1999.

58. Vichhinsky EP NL, Earles AN et al. . Causes and outcomes of the acute chest syndrome in sickle cell disease. . *N Engl J Med.* 2000(342):1855-1865.

59. Gladwin MT, Schechter AN, Shelhamer JH, Ognibene FP. The acute chest syndrome in sickle cell disease. Possible role of nitric oxide in its pathophysiology and treatment. *Am J Respir Crit*

Care Med. 1999;159(5 Pt 1):1368-1376.

60. Geraldo BdSJ, Daher EDF, Rocha FACd. Osteoarticular involvement in sickle cell disease. *Rev Bras Hematol Hemoter.* 2012;34(2):156-164.

61. Eckman JR. Leg ulcers in sickle cell disease. *Hematol Oncol Clin North Am.* 1996;10(6):1333-1344.

62. Koshy M, Entsuah R, Koranda A, et al. Leg ulcers in patients with sickle cell disease. *Blood.* 1989;74(4):1403-1408.

63. Minniti CP, Taylor JGt, Hildesheim M, et al. Laboratory and echocardiography markers in sickle cell patients with leg ulcers. *Am J Hematol.* 2011;86(8):705-708.

64. Minniti CP, Kato GJ. Critical Reviews: How we treat sickle cell patients with leg ulcers. *Am J Hematol.* 2016;91(1):22-30.

65. Minniti CP, Eckman J, Sebastiani P, Steinberg MH, Ballas SK. Leg ulcers in sickle cell disease. *Am J Hematol.* 2010;85(10):831-833.

66. Cumming V, King L, Fraser R, Serjeant G, Reid M. Venous incompetence, poverty and lactate dehydrogenase in Jamaica are important predictors of leg ulceration in sickle cell anaemia. *Br J Haematol.* 2008;142(1):119-125.

67. Ohene-Frempong K, Weiner SJ, Sleeper LA, et al. Cerebrovascular accidents in sickle cell disease: rates and risk factors. *Blood.* 1998;91(1):288-294.

68. Cherry MG, Greenhalgh J, Osipenko L, et al. The clinical effectiveness and cost-effectiveness of primary stroke prevention in children with sickle cell disease: a systematic review and economic evaluation. *Health Technol Assess.* 2012;16(43):1-129.

69. Goldberg MF. Classification and pathogenesis of proliferative sickle retinopathy. *Am J Ophthalmol.* 1971;71(3):649-665.

70. Amjad AI, Ali H, Appleman LJ, et al. Renal medullary carcinoma: case report of an aggressive malignancy with near-complete response to dose-dense methotrexate, vinblastine, Doxorubicin,

and Cisplatin chemotherapy. *Case Rep Oncol Med.* 2014;2014:615895.

71. Baig MA, Lin YS, Rasheed J, Mittman N. Renal medullary carcinoma. *J Natl Med Assoc.* 2006;98(7):1171-1174.

72. Brousse V, Elie C, Benkerrou M, et al. Acute splenic sequestration crisis in sickle cell disease: cohort study of 190 paediatric patients. *Br J Haematol.* 2012;156(5):643-648.

73. Naymagon L, Pendurti G, Billett HH. Acute Splenic Sequestration Crisis in Adult Sickle Cell Disease: A Report of 16 Cases. *Hemoglobin.* 2015;39(6):375-379.

74. Wright J, Thomas P, Serjeant GR. Septicemia caused by Salmonella infection: an overlooked complication of sickle cell disease. *J Pediatr.* 1997;130(3):394-399.

75. Zarkowsky HS GD, Gill FM et al. Bacteremia is sickle hemoglobinopathies. *J Pediatr.* 1986(109):579-585.

76. Biomarkers Definitions Working G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther.* 2001;69(3):89-95.

77. Damanhouri GA, Jarullah J, Marouf S, Hindawi SI, Mushtaq G, Kamal MA. Clinical biomarkers in sickle cell disease. *Saudi J Biol Sci.* 2015;22(1):24-31.

78. Rees DC, Gibson JS. Biomarkers in sickle cell disease. *Br J Haematol.* 2012;156(4):433-445.

79. Lettre G, Bauer D. Fetal Hemoglobon in sickle-cell disease: from genetic epidemiology to new therapeutic strategies. *Lancet.* 2016;387(2554-64).

80. Hoban MD, Orkin SH, Bauer DE. Genetic treatment of a molecular disorder: gene therapy approaches to sickle cell disease. *Blood.* 2016;127(7):839-848.

81. Platt OS, Orkin SH, Dover G, Beardsley GP, Miller B, Nathan DG. Hydroxyurea enhances fetal hemoglobin production in sickle cell anemia. *J Clin Invest.* 1984;74(2):652-656.

82. Dover G, RK H, JG M. Hydroxyurea induction of hemoglobin F production in sickle cell disease: relationship between cytotoxicity and F cell production. . *Blood* 1986;67:735.

83. Charache S, Dover GJ, Moyer MA, Moore JW. Hydroxyurea-induced augmentation of fetal

hemoglobin production in patients with sickle cell anemia. *Blood.* 1987;69(1):109-116.

84. Charache S, Barton FB, Moore RD, et al. Hydroxyurea and sickle cell anemia. Clinical utility of a myelosuppressive "switching" agent. The Multicenter Study of Hydroxyurea in Sickle Cell Anemia. *Medicine (Baltimore).* 1996;75(6):300-326.

85. Charache S, Terrin ML, Moore RD, et al. Effect of hydroxyurea on the frequency of painful crises in sickle cell anemia. Investigators of the Multicenter Study of Hydroxyurea in Sickle Cell Anemia. *N Engl J Med.* 1995;332(20):1317-1322.

86. Fung EB, Barden EM, Kawchak DA, Zemel BS, Ohene-Frempong K, Stallings VA. Effect of hydroxyurea therapy on resting energy expenditure in children with sickle cell disease. *J Pediatr Hematol Oncol.* 2001;23(9):604-608.

87. Ware RE, Eggleston B, Redding-Lallinger R, et al. Predictors of fetal hemoglobin response in children with sickle cell anemia receiving hydroxyurea therapy. *Blood.* 2002;99(1):10-14.

88. Steinberg MH, Barton F, Castro O, et al. Effect of hydroxyurea on mortality and morbidity in adult sickle cell anemia: risks and benefits up to 9 years of treatment. *JAMA.* 2003;289(13):1645-1651.

89. Adams RJ, McKie VC, Carl EM, et al. Long-term stroke risk in children with sickle cell disease screened with transcranial Doppler. *Ann Neurol.* 1997;42(5):699-704.

90. de Montalembert M, Begue P, Bernaudin F, Thuret I, Bachir D, Micheau M. Preliminary report of a toxicity study of hydroxyurea in sickle cell disease. French Study Group on Sickle Cell Disease. *Arch Dis Child.* 1999;81(5):437-439.

91. Kinney TR, Helms RW, O'Branski EE, et al. Safety of hydroxyurea in children with sickle cell anemia: results of the HUG-KIDS study, a phase I/II trial. Pediatric Hydroxyurea Group. *Blood.* 1999;94(5):1550-1554.

92. Best PJ, Daoud MS, Pittelkow MR, Petitt RM. Hydroxyurea-induced leg ulceration in 14 patients. *Ann Intern Med.* 1998;128(1):29-32.

93. Johnson FL, Look AT, Gockerman J, Ruggiero MR, Dalla-Pozza L, Billings FT, 3rd. Bone-marrow transplantation in a patient with sickle-cell anemia. *N Engl J Med.* 1984;311(12):780-783.

94. Vermylen C, Cornu G, Ferster A, et al. Haematopoietic stem cell transplantation for sickle cell anaemia: the first 50 patients transplanted in Belgium. *Bone Marrow Transplant.* 1998;22(1):1-6.

95. Vermylen C, Cornu G. Bone marrow transplantation for sickle cell disease. The European experience. *Am J Pediatr Hematol Oncol.* 1994;16(1):18-21.

96. Walters MC, Storb R, Patience M, et al. Impact of bone marrow transplantation for symptomatic sickle cell disease: an interim report. Multicenter investigation of bone marrow transplantation for sickle cell disease. *Blood.* 2000;95(6):1918-1924.

97. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet.* 2005;6(2):95-108.

98. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.

99. Consortium TIH. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007;449:851-862.

100. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet.* 2016;48(11):1443-1448.

101. Danielle Welter JM, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research.* 2014.

102. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860-921.

103. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science.*

2001;291(5507):1304-1351.

104. Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science.* 2005;308(5720):419-421.

105. Edwards AO, Ritter R, 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. *Science.* 2005;308(5720):421-424.

106. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005;308(5720):385-389.

107. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell.* 2016;167(5):1415-1429 e1419.

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

109. Lettre G. The search for genetic modifiers of disease severity in the beta-hemoglobinopathies. *Cold Spring Harb Perspect Med.* 2012;2(10).

110. Milton JN, Sebastiani P, Solovieff N, et al. A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. *PLoS One.* 2012;7(4):e34741.

111. Genovese G, Friedman DJ, Ross MD, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science.* 2010;329(5993):841-845.

112. Uda M, Galanello R, Sanna S, et al. 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.* 2008;105(5):1620-1625.

113. Menzel S, Garner C, Gut I, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet.* 2007;39(10):1197-1199.

114. Bae HT, Baldwin CT, Sebastiani P, et al. Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood.* 2012;120(9):1961-1962.

115. Bhatnagar P, Purvis S, Barron-Casella E, et al. Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J Hum Genet.* 2011;56(4):316-323.

116. Mtatiro SN, Singh T, Rooks H, et al. Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One.* 2014;9(11):e111464.

117. Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat Genet.* 2010;42(12):1049-1051.

118. Anthony D. Maher PWK. The Gárdos channel: a review of the Ca2+-activated K+ channel in human erythrocytes. *The International Journal of Biochemistry & Cell Biology.* 2003;35:1182–1197.

119. Segall L SR, Kaunisto MA, et al. . Kinetic alterations due to a missense mutation in the Na,K-ATPase alpha2 subunit cause familial hemiplegic migraine type 2. . *J Biol Chem* 2004;279(43692).

120. de Carvalho Aguiar P, Sweadner KJ, Penniston JT, et al. Mutations in the Na+/K+ -ATPase alpha3 gene ATP1A3 are associated with rapid-onset dystonia parkinsonism. *Neuron.* 2004;43(2):169-175.

121. Bookchin RM, Ortiz OE, Lew VL. Evidence for a direct reticulocyte origin of dense red cells in sickle cell anemia. *J Clin Invest.* 1991;87(1):113-124.

122. Lew VL, Ortiz OE, Bookchin RM. Stochastic nature and red cell population distribution of the sickling-induced Ca2+ permeability. *J Clin Invest.* 1997;99(11):2727-2735.

123. Lew VL, Etzion Z, Bookchin RM. Dehydration response of sickle cells to sickling-induced Ca(++) permeabilization. *Blood.* 2002;99(7):2578-2585.

124. Boettger T, Rust MB, Maier H, et al. Loss of K-Cl co-transporter KCC3 causes deafness, neurodegeneration and reduced seizure threshold. *EMBO J.* 2003;22(20):5422-5434.

125. Boettger T, Hubner CA, Maier H, Rust MB, Beck FX, Jentsch TJ. Deafness and renal tubular acidosis in mice lacking the K-Cl co-transporter Kcc4. *Nature.* 2002;416(6883):874-878.

126. Hubner CA, Stein V, Hermans-Borgmeyer I, Meyer T, Ballanyi K, Jentsch TJ. Disruption of KCC2 reveals an essential role of K-Cl cotransport already in early synaptic inhibition. *Neuron.* 2001;30(2):515-524.

127. Valles PG, Bocanegra V, Gil Lorenzo A, Costantino VV. Physiological Functions and Regulation of the Na+/H+ Exchanger [NHE1] in Renal Tubule Epithelial Cells. *Kidney Blood Press Res.* 2015;40(5):452-466.

128. Perrotta S, Borriello A, Scaloni A, et al. The N-terminal 11 amino acids of human erythrocyte band 3 are critical for aldolase binding and protein phosphorylation: implications for band 3 function. *Blood.* 2005;106(13):4359-4366.

129. Chu H, Breite A, Ciraolo P, Franco RS, Low PS. Characterization of the deoxyhemoglobin binding site on human erythrocyte band 3: implications for O2 regulation of erythrocyte properties. *Blood.* 2008;111(2):932-938.

130. Ribeiro ML, Alloisio N, Almeida H, et al. Severe hereditary spherocytosis and distal renal tubular acidosis associated with the total absence of band 3. *Blood.* 2000;96(4):1602-1604.

131. Cordat E. Unraveling trafficking of the kidney anion exchanger 1 in polarized MDCK epithelial cells. *Biochem Cell Biol.* 2006;84(6):949-959.

132. Toye AM, Williamson RC, Khanfar M, et al. Band 3 Courcouronnes (Ser667Phe): a trafficking mutant differentially rescued by wild-type band 3 and glycophorin A. *Blood.* 2008;111(11):5380-5389.

133. Carlo Brugnara M. Control of red blood cell hydration. 2016; http://www.uptodate.com/contents/control-of-red-blood-cell-hydration, 2016.

134. Joiner CH. Cation transport and volume regulation in sickle red blood cells. *The American journal of physiology.* 1993 264(2 Pt 1):C251-270.

135. Anthony J. McGoron CHJ, Mary B. Palascak, William J. Claussen and Robert S. Franco. Dehydration of mature and immature sickle red blood cells during fast oxygenation/deoxygenation cycles: role of KCl cotransport and extracellular calcium. *Blood* 2000 95:2164-2168.

136. Robert S. Franco HT, Mary Palascak and Clinton H. Joiner. The Formation of Transferrin Receptor-Positive Sickle Reticulocytes With Intermediate Density Is Not Determined by Fetal Hemoglobin Content. *Blood* 1997 90:3195-3203.

137. McGoron AJ, Joiner CH, Palascak MB, Claussen WJ, Franco RS. Dehydration of mature and immature sickle red blood cells during fast oxygenation/deoxygenation cycles: role of KCl cotransport and extracellular calcium. *Blood.* 2000;95(6):2164-2168.

138. Lew VL BR. Ion transport pathology in the mechanism of sickle cell dehydration. . *Physiol Rev* 2005; ;85(179).

139. Eaton WA, Hofrichter J. Sickle cell hemoglobin polymerization. *Adv Protein Chem.* 1990;40:63-279.

140. Sunshine HR, Hofrichter J, Eaton WA. Requirements for therapeutic inhibition of sickle haemoglobin gelation. *Nature.* 1978;275(5677):238-240.

141. Bridges KR, Barabino GD, Brugnara C, et al. A multiparameter analysis of sickle erythrocytes in patients undergoing hydroxyurea therapy. *Blood.* 1996;88(12):4701-4710.

142. Gallagher PG. Transporting down the road to dehydration. *Blood.* 2015;126(26):2775-2776.

143. Ishii TM, Silvia C, Hirschberg B, Bond CT, Adelman JP, Maylie J. A human intermediate conductance calcium-activated potassium channel. *Proc Natl Acad Sci U S A.* 1997;94(21):11651-11656.

144. Bartolucci P, Brugnara C, Teixeira-Pinto A, et al. Erythrocyte density in sickle cell syndromes is associated with specific clinical manifestations and hemolysis. *Blood.* 2012;120(15):3136-3141.

145. De Franceschi L, Saadane N, Trudel M, Alper SL, Brugnara C, Beuzard Y. Treatment with oral clotrimazole blocks Ca(2+)-activated K+ transport and reverses erythrocyte dehydration in transgenic SAD mice. A model for therapy of sickle cell disease. *J Clin Invest.* 1994;93(4):1670-1676.

146. Brugnara C, de Franceschi L, Alper SL. Inhibition of Ca(2+)-dependent K+ transport and cell dehydration in sickle erythrocytes by clotrimazole and other imidazole derivatives. *J Clin Invest.* 1993;92(1):520-526.

147. Brugnara C, Gee B, Armsby CC, et al. Therapy with oral clotrimazole induces inhibition of the Gardos channel and reduction of erythrocyte dehydration in patients with sickle cell disease. *J Clin Invest.* 1996;97(5):1227-1234.

148. De Franceschi L, Bachir D, Galacteros F, et al. Oral magnesium supplements reduce erythrocyte dehydration in patients with sickle cell disease. *J Clin Invest.* 1997;100(7):1847-1852.

149. De Franceschi L, Bachir D, Galacteros F, et al. Oral magnesium pidolate: effects of long-term administration in patients with sickle cell disease. *Br J Haematol.* 2000;108(2):284-289.

150. Ataga KI, Reid M, Ballas SK, et al. Improvements in haemolysis and indicators of erythrocyte survival do not correlate with acute vaso-occlusive crises in patients with sickle cell disease: a phase III randomized, placebo-controlled, double-blind study of the Gardos channel blocker senicapoc (ICA-17043). *Br J Haematol.* 2011;153(1):92-104.

151. Fusi N, Lippert C, Lawrence ND, Stegle O. Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat Commun.* 2014;5:4890.

152. Ziegler A. Genome-wide association studies: quality control and population-based measures. *Genet Epidemiol.* 2009;33 Suppl 1:S45-50.

153. Teo YY. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol.*

2008;19(2):133-143.

154. Purcell S NB, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. . *American Journal of Human Genetics.* 2007;81.

155. Bao R, Huang L, Andrade J, et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform.* 2014;13(Suppl 2):67-82.

156. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010;11(7):499-511.

157. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda).* 2011;1(6):457-470.

158. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet.* 2009;84(2):210-223.

159. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2010;34(8):816-834.

160. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet.* 2012;44(8):955-959.

161. Laughbaum A. Comparing BEAGLE, IMPUTE2, and Minimac Imputation Methods for Accuracy, Computation Time, and Memory Usage. 2013; Technology review. Available at: http://blog.goldenhelix.com/alaughbaum/comparing-beagle-impute2-and-minimac-imputation-methods-for-accuracy-computation-time-and-memory-usage/, 2016.

162. Lin P, Hartz SM, Zhang Z, et al. A new statistic to evaluate imputation reliability. *PLoS One.* 2010;5(3):e9697.

163. Hancock DB, Levy JL, Gaddis NC, et al. Assessment of genotype imputation performance using 1000 Genomes in African American studies. *PLoS One.* 2012;7(11):e50610.

164. O'Connell J, Gurdasani D, Delaneau O, et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* 2014;10(4):e1004234.

165. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284-1287.

166. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet.* 2014;15(5):335-346.

167. Wang SB, Feng JY, Ren WL, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep.* 2016;6:19444.

168. Johnson RC, Nelson GW, Troyer JL, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics.* 2010;11:724.

169. Lewis CM. Genetic association studies: design, analysis and interpretation. *Brief Bioinform.* 2002;3(2):146-153.

170. Lettre G. Rare and low-frequency variants in human common diseases and other complex traits. *J Med Genet.* 2014;51(11):705-714.

171. Lin WY, Lou XY, Gao G, Liu N. Rare variant association testing by adaptive combination of P-values. *PLoS One.* 2014;9(1):e85728.

172. Price AL, Kryukov GV, de Bakker PI, et al. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010;86(6):832-838.

173. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011;89(1):82-93.

174. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-842.

175. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.

Bioinformatics. 2009;25(14):1754-1760.

176. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303.

177. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.

178. Liu DJ, Peloso GM, Zhan X, et al. Meta-analysis of gene-level tests for rare variant association. *Nat Genet.* 2014;46(2):200-204.

179. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics.* 2016;32(9):1423-1426.

180. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2011;9(2):179-181.

181. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 2013;10(1):5-6.

182. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. *Am J Hum Genet.* 2013;93(4):687-696.

183. Xu J, Shao Z, Glass K, et al. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell.* 2012;23(4):796-811.

184. O'Leary NA WM, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, et, al. . Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. . *Nucleic Acids Res* 2016 44((D1)):D733-745

185. Lettre G, Bauer DE. Fetal haemoglobin in sickle-cell disease: from genetic epidemiology to new therapeutic strategies. *Lancet.* 2016;387(10037):2554-2564.

186. Gallagher PG. Disorders of red cell volume regulation. *Current opinion in hematology.* 2013;20(3):201-207.

187. Stocker JW, De Franceschi L, McNaughton-Smith GA, Corrocher R, Beuzard Y, Brugnara C. ICA-17043, a novel Gardos channel blocker, prevents sickled red blood cell dehydration in vitro and in vivo in SAD mice. *Blood.* 2003;101(6):2412-2418.

188. Li Q, Henry ER, Hofrichter J, et al. Kinetic assay shows that increasing red cell volume could be a treatment for sickle cell disease. *Proc Natl Acad Sci U S A.* 2017;114(5):E689-E696.

189. Gaston M, Smith J, Gallagher D, et al. Recruitment in the Cooperative Study of Sickle Cell Disease (CSSCD). *Control Clin Trials.* 1987;8(4 Suppl):131S-140S.

190. Gaston M, Rosse WF. The cooperative study of sickle cell disease: review of study design and objectives. *Am J Pediatr Hematol Oncol.* 1982;4(2):197-201.

191. Solovieff N, Milton JN, Hartley SW, et al. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5 ' olfactory receptor gene cluster. *Blood.* 2010;115(9):1815-1822.

192. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45(6):580-585.

193. Trudel M, De Paepe ME, Chretien N, et al. Sickle cell disease of transgenic SAD mice. *Blood.* 1994;84(9):3189-3197.

194. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. *Nature protocols.* 2008;3(6):1101-1108.

195. Chen Z, Tang H, Qayyum R, et al. Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum Mol Genet.* 2013;22(12):2529-2538.

196. Timmann C, Thye T, Vens M, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature.* 2012;489(7416):443-446.

197. Lacy JN, Ulirsch JC, Grace RF, et al. Exome sequencing results in successful diagnosis and

treatment of a severe congenital anemia. *Cold Spring Harb Mol Case Stud.* 2016;2(4):a000885.

198. Sankaran VG, Gallagher PG. Applications of high-throughput DNA sequencing to benign hematology. *Blood.* 2013;122(22):3575-3582.

199. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17(1):122.

200. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-291.

201. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7(4):248-249.

202. Kumar P HS, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* 2009;4(8):1073-1081.

203. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491-498.

204. Gallagher PG, Ferreira JD, Costa FF, Saad ST, Forget BG. A recurrent frameshift mutation of the ankyrin gene associated with severe hereditary spherocytosis. *Br J Haematol.* 2000;111(4):1190-1193.

205. Gallagher PG. Red cell membrane disorders. *Hematology Am Soc Hematol Educ Program.* 2005;13(8):13-18.

206. Rybicki AC QJ, Musto S, Rosen NL, Nagel RL, Schwartz RS. Human erythrocyte protein 4.2 deficiency associated with hemolytic anemia and a homozygous 40glutamic acid-->lysine substitution in the cytoplasmic domain of band 3 (band 3Montefiore). *Blood.* 1999;81(2155).

207. Alloisio N, Texier P, Vallier A, et al. Modulation of clinical expression and band 3 deficiency in hereditary spherocytosis. *Blood.* 1997;90(1):414-420.

208. Jarolim P, Palek J, Rubin HL, Prchal JT, Korsgren C, Cohen CM. Band 3 Tuscaloosa: Pro327----Arg327 substitution in the cytoplasmic domain of erythrocyte band 3 protein associated with

spherocytic hemolytic anemia and partial deficiency of protein 4.2. *Blood.* 1992;80(2):523-529.

209. Selcuk Duru N, Celkan T, Civilibal M, Ozbek NO, Basak AN, Elevli M. Coinheritance of sickle cell anemia and hereditary spherocytosis. *Pediatr Blood Cancer.* 2008;51(4):560-563.

210. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet.* 2015;24(8):2125-2137.

211. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A.* 2014;111(4):E455-464.