Université de Montréal

Multi-omics approaches to sickle cell disease heterogeneity

Par Mitikbeta Yann Tanguy Ilboudo

Département de biochimie et médecine moléculaire
Faculté de médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Ph. D.
en Bio-informatique

October 2022

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Multi-omics approaches to sickle cell disease heterogeneity

Presenté par

Mitibketa Yann Tanguy Ilboudo

a été évaluée par un jury composé des personnes suivantes :

Sarah Gagliano Ph. D.
présidente-rapporteuse

Guillaume Lettre,
directeur de recherche

Martin Smith
membre du jury

Yohan Bossé,
examinateur externe

Martin Sauvageau,
représentant de la doyenne de la FES

Résumé

La drépanocytose est une maladie causée par une seule mutation dans le gène de la bêta-globine. Les complications liées à la maladie se manifestent sur le plan génétique, épigénique, transcriptionnel, et métabolique. Les approches intégratives des technologies de séquençage à haut-débit permettent de comprendre le mécanisme pathologique et de découvrir des thérapies en lien avec la maladie. Dans cette thèse, j'intègre divers jeux de données omiques et j'applique des méthodes statistiques pour élaborer de nouvelles hypothèses et analyser les données.

Dans les deux premières études, je combine les résultats des études d'association pangénomique d'hémoglobine fœtale (HbF) et des globules rouges denses déshydratés (DRBC) avec l'expression génique, l'interaction chromatinienne, les bases de données relatives aux maladies et les cibles médicamenteuses sélectionnées par des experts. Cette approche intégrative a révélé trois nouveaux loci sur le chromosome 10 (*BICC1*), le chromosome 19 (*KLF1*) et le chromosome 22 (*CECR2*) comme régulateurs de l'HbF. Pour l'étude sur la densité de globules rouges, quatre cibles médicamenteuses (*BCL6*, *LRRC32*, *KNCJ14* et *LETM1*) ont été identifiées comme des modulateurs potentiels de la sévérité.

Dans la troisième étude, j'intégre la métabolomique à la génomique pour établir une relation causale entre la L-glutamine et les crises douleurs en utilisant la randomisation mendélienne. En outre, nous avons identifié 66 biomarqueurs pour 6 complications liées à la drépanocytose et le débit de filtration glomérulaire estimé (DFGe). Enfin, dans la dernière étude j'ai appliqué une approche de clustering aux métabolites que j'ai ensuite combiné aux données de génotype. J'ai découvert des changements métabolomiques mettant en évidence des familles de métabolites impliqués dans les dysfonctionnements rénaux et hépatiques, en plus de confirmer le rôle d'une classe d'acides gras dans la formation en faucille des globules rouges. Ce travail met en évidence l'importance des approches multi-omiques pour découvrir de nouveaux mécanismes biologiques et étudier les maladies humaines.

Mots clés : Drépanocytose, hémoglobine fœtale, études d'association pangénomique, études d'association à l'échelle de l'exome, clustering de métabolites.

Abstract

Sickle cell disease is a monogenic disorder caused by a point mutation in the beta-globin gene. The complications related to the disease are characterized by a broad spectrum of distinct genetic, epigenetic, transcriptional, and metabolomic states. Integrative high-throughput technologies approaches to sickle cell disease pathophysiology are crucial to understanding complications mechanisms and uncovering therapeutic interventions. In this thesis, I integrate various omics datasets and apply statistical methods to derive new hypotheses and analyze data.

I combine genome-wide association studies results of fetal hemoglobin (HbF) and dehydrated dense red blood cells (DRBC) with gene expression, chromatin interaction, disease-relevant databases, and expert-curated drug targets. This integrative approach revealed three novel loci on chromosome 10 (*BICC1*), chromosome 19 (*KLF1*) and chromosome 22 (*CECR2*) as key modulators of HbF. For DRBC, four drug targets (*BCL6*, *LRRC32, KNCJ14,* and *LETM1*) were identified as potential severity modifiers.

Using mendelian randomization, I integrated metabolomics with genomics in the third study to establish a potential causal relationship between L-glutamine and painful crisis. Additionally, we identified 66 biomarkers for 6 SCD-related complications and estimated glomerular filtration rate (eGFR). Finally, the last study applied a clustering framework to metabolites which I then combined with genotypes. I found specific metabolomics changes highlighting families of metabolites involved in renal and liver dysfunction and confirming the role of a class of fatty acids in red blood cell sickling. This work highlights the importance of multi-omics approaches to unearth new biology and study human diseases.

Keywords: Sickle cell disease, fetal hemoglobin, genome-wide association studies, exome-wide association studies, metabolite clustering

Table of Contents

List of Figures

List of Tables

List of abbreviations

1000G 1000 Genomes
CAAPA Consortium on Asthma among African-ancestry Populations in the Americas
CI Confidence interval
DRBC Dehydrated dense red blood cell
eQTL Expression quantitative trait loci
FDA Food and Drug Administration
GWAS Genome-wide association study
HAPMAP Haplotype map
HbA Adult hemoglobin
HbF Fetal hemoglobin
HbS Hemoglobin S
HCT Hematocrit
HGB Hemoglobin
HPFH Hereditary persistence of fetal hemoglobin
HU Hydroxyurea
HS Hereditary Spherocytosis
HSPCs Hemangleoietic Stem and Progenitor Cells
IBD Identity by descent
IV Instrumental variable
IVW Inverse variance weighted
kb Kilobase
kDa Kilo dalton
LD Linkage disequilibrium
MAF Minor allele frequency
MCH Mean cell hemoglobin
MCHC Mean cell hemoglobin concentration
MCV Mean corpuscular volume
MR Mendelian randomization
MSCV Mean sphered cell volume
PC Principal component analysis
QC Quality control
r^2 Imputation measure of the quality of imputation
R^2 Variance explained
RBC Red blood cells
RDW Red cell distribution width
SCA Sickle cell anemia
SCD Sickle cell disease
SD standard deviation
SE Standard error
SKAT Sequence kernel association test
SNP Single nucleotide polymorphism
MR Mendelian randomization
VT Variable threshold
WES Whole-exome sequencing

"The greatest deception men suffer is from their own opinion."
-Leonardo da Vinci

"Il faut savoir douter où il faut, se soumettre où il faut, croire où il faut"
– Blaise Pascal

Acknowledgements

My deepest gratitude goes to Guillaume Lettre, whose mentorship, advice, support, and trust have been invaluable throughout my doctoral studies. From our initial meeting, I could not have foreseen the profound impact you would have on shaping my career and future. You provided me with unparalleled opportunities and projects, allowing me the freedom to grow and mature as a scientist, for which I am profoundly grateful. Our meetings were always a source of inspiration and innovation, sparking fresh thoughts and ideas. My thanks are beyond measure.

I would like to express my appreciation to Julie Hussin, my thesis godmother. Our earnest discussions about navigating careers in academia and industry were enlightening. Your willingness to share personal experiences has been instrumental in guiding me through the milestones of the Ph.D. journey.

I extend my thanks to my thesis committee members, John D Rioux and Adrian Serohijos. Your advice and mentorship have been pivotal in honing my scientific thinking, presentation skills, and diligent work ethic.

I'm grateful to Melissa Beaudoin and Ken for their invaluable contributions. Melissa, your intellect, spirit, and kindness are a gift to our lab. Your expertise in the technical aspects of biochemistry, particularly related to cell lines, CRISPR, and base-editing protocols, has been immensely helpful. Ken, your guidance in writing pipelines, attention to detail, and high expectations have significantly improved my coding and communication skills.

I'm thankful to our national and international collaborators, namely Allison E Ashley-Koch, Melanie E Garrett, Marylin Tellen, Carlo Brugnara, Pablo Bartolucci, and Seth Alper.

A heartfelt thanks to all the sickle cell disease patients who participated in our research studies. Your bravery and resilience are truly inspirational.

I would also like to express my gratitude to Elaine Meunier for handling all administrative aspects efficiently and providing unwavering support. You are indeed a beacon within the department.

I extend my appreciation to all lab members, past and present. Each of you has inspired and enriched me in unique ways, and I have greatly enjoyed getting to know you.

To my beloved family, my dear Marieme, and our treasures, Noah, Moïse and Joanna, your continual joy and happiness are my greatest comfort. My parents, Christiane and Jean-Pierre, my sister, and my brother, your unwavering support throughout my studies has been my foundation.

I also wish to acknowledge the team at Compute Canada for their diligent work, and the musicians, Drake, Tiken Dja Fakoly, Cesaria Evora, Big Sean, Niska, Booba, Elown's. Kiff no beat, DUDEN J, who provided the soundtrack to my long days and nights of writing. Lastly, my gratitude to David Goggins for his insightful advice.

Lastly, I dedicate this thesis to both of my siblings Dolobsom Andy Renaud Ilboudo, Wendpanga Giulia Cindy Laura Ilboudo.

Chapter 1: Introduction

The ensuing chapter offers a comprehensive overview of sickle cell disease, including its symptoms, the underlying pathophysiological abnormalities, existing treatments, landmark discoveries since its initial identification, and an exploration of the disease modifiers examined in this thesis.

SCD historical background

Sickle cell disease (SCD) was first described in Western literature in 1910 by a cardiologist named James B Herrick while tending to a medical student complaining about chest pain[1]. However, records dating from the 1870s describe the disease in African literature as "children who come and go" due to the high infant mortality[2]. Herrick is the originator of the term "sickled-shaped" to report the odd appearance of red blood cells in patients (**Figure 1**). Seventeen years later, Han and Gillespie proved that anoxia resulted in red blood cell (RBC) sickling[3]. A couple of years later, Scriver and Waugh demonstrated in vivo that hypoxia leads to RBC sickling. This body of evidence led Linus Pauling, Nobel laureate scientist, to intimate that altered hemoglobin might be at the origin of the sickling phenomena[4].

In 1949, a gel electrophoresis experiment distinguished between hemoglobin in sickle cell disease and normal hemoglobin in healthy individuals[4,5]. The next year James V. Neel discovered the autosomal recessive model of inheritance[6]. Around the same time, Janet B Watson hypothesized that fetal hemoglobin (HbF)[7] could be protective against SCD complications since newborns didn't display any symptoms until six months of age. In 1956, a pioneering study showed the protective effect of carrying the sickle cell trait against malaria[8]. Two years later, Vernon Ingram and his team confirmed that the difference between sickled hemoglobin (HbS) and the healthy adult hemoglobin (HbA) is a consequence of a single amino acid which replaces a glutamic acid by valine at the 6th position of the beta globin subunit in hemoglobin molecule[9]. Finally, another landmark discovery of the 20th century in sickle cell disease by Ferrone and colleagues[10] showed that under hypoxic conditions HbS forms polymers which distort normal bi-concave erythrocytes and turn them into rigid crescent erythrocytes. This body of evidence enhanced our appreciation of the molecular basis of sickle cell disease and propelled further discoveries in the 21st century.

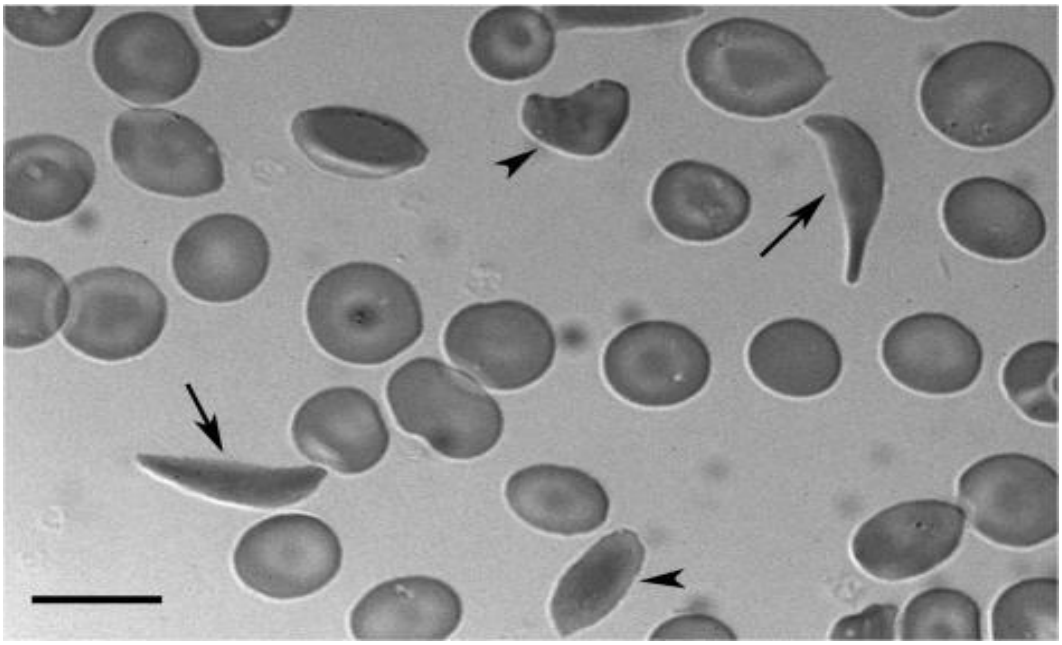


Figure 1 Sickle Erythrocytes. Peripheral blood smear from a patient with SCD obtained during a routine clinic visit. The smear shows classical sickle-shaped (arrows) and various other misshaped erythrocytes (arrowheads). The image was obtained from an air-dried smear using differential interference contrast (DIC) microscopy with an Olympus BX61WI work station equipped with a LUMPlanFI ×60 numerical aperture 0.90 ∞ objective (Olympus) and a CoolSnap HQ camera (6.6 μm^2 pixel, 1,392 × 1,040 pixel format) (Roper Scientific). Scale bar: 10 μm.

SCD Burden

Sickle cell disease and thalassemia are the two most common hemoglobinopathies. Thalassemia is characterized by a change in the ratio of α-globin to β-globin chain production due to a genetic mutation in either alpha or beta globin genes. This imbalance in the two globin proportions will cause the precipitation of the α-globin subsequently leading to ineffective erythropoiesis, and hemolysis[11]. Sickle cell disease (SCD) is caused by a point mutation in the 6^{th} position of the β-chain replacing a glutamic amino acid for a valine amino acid. This single amino acid change predisposes hemoglobin to polymerize under hypoxic conditions. Once the hemoglobin polymerizes, red blood cells become rigid, crescent-shaped (**Figure 1**), they damage endothelial walls, cause anemia, pain, and stroke. Individuals with sickle cell trait show no severe symptoms. Carriers of the trait present with one mutant allele of the β-globin gene and one normal resulting in HbAS. However, when both β-globin carry the sickle mutation, we then call the condition sickle cell anemia.

The gene burden is spread out throughout the world, with an elevated rate in sub-Saharan Africa, the Mediterranean (mostly Greece), parts of India, and the Middle East. The incidence rate seems to correlate with regions of the world where malaria is endemic since the sickle mutation is protective against severe forms of malaria. Global burden account for at least 7% of African American population[12,13], one-third of Sub-Saharan Africa[14], at most 6% of the Latino population[12,15], up to 13% in India[16,17], between 0.2% to 27% of the Middle East[18,19], at most 10% of Greeks[20,21] and 4 to 10% of Caribbean countries[22,23]. A 17-year prospective study by the University of Michigan published in 2019 identified that 80% of newborns with the sickle cell trait were from African ancestry, 7.4% from Arab ancestry (7.4%) and 7% from white ancestry (7%)[24].

Although the protective pathophysiology acquired from the mutation remains uncertain[25], the 'malaria hypothesis' formulated by Anthony Allison and Haldane is a perfect model for natural selection and balanced polymorphism[26]. There is an estimated 300,000 newborn each year globally affected by sickle cell anemia[27]. Countries such as Nigeria, India and the Democratic Republic of the Congo see the highest birth rate. Finally, while improvements in areas such as prophylactic treatment, and newborn screening improved life expectancy, challenges related to poverty, malnutrition, malaria, and routine vaccination show a 90% mortality rate amongst children younger than 5 years of age[28].

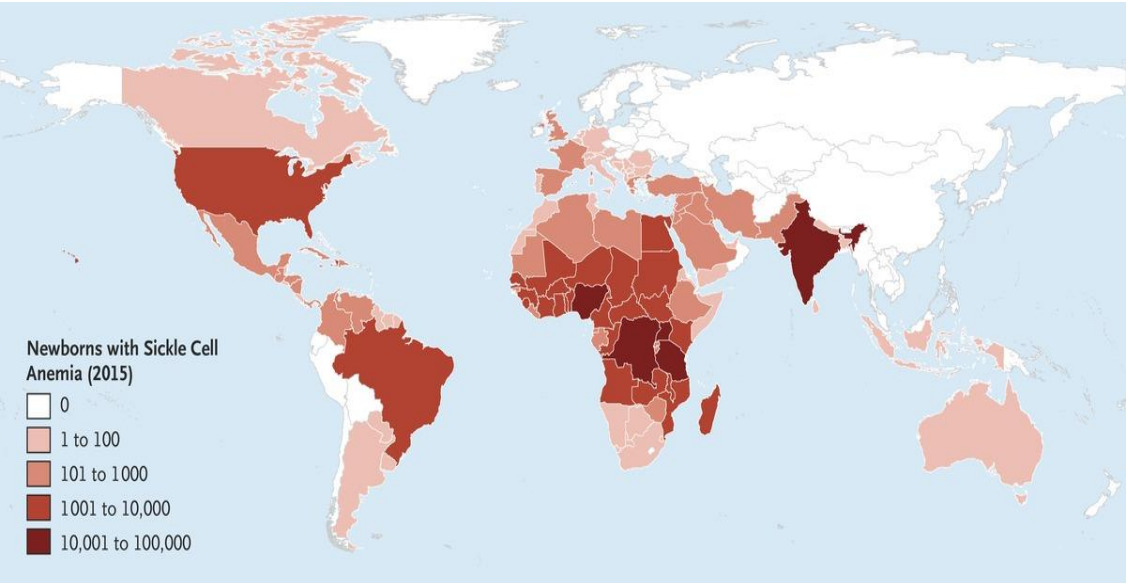


Figure 2. Number of newborns with sickle cell Anemia in each country in 2015. As seen in Piel *et al*. (2017).

SCD and Malaria

Historical perspective

Malaria is a deadly disease transmitted via a bite of a female *Anopheles spp*. An additional mechanism of transmission includes blood transfusions, sharing of contaminated needles, and organ transplantation[29,30]. The first accounts of the disease can be found more than four millenniums ago, in Chinese scrolls, European memoires[31]. Potent anti-malarial drugs such as artemisinin and quinine have been present within different civilisations[32]. The discovery of the parasite *Plasmodium Falciparum* resulted in a Nobel prize for Charles Louis Alphonse Laveran[31]. To date, we count 5 species of *Plasmodium* parasite (*P. vivax, P. falciparum, P. malariae, P. ovale, and P. knowlesi.*) which can infect humans[33]. There are however over 150 species which could infect and cause malaria in other vertebrates[34-38].

Pathophysiology

Severe forms of malaria, its pathogenicity, the role of the parasite, and its host interactions are described in detail elsewhere[39-41]. Consequences of infection by *P.falciparrum* can be divided into three clinical outcomes groups: triumvirate of cerebral malaria (CM), respiratory distress, and severe malarial anemia[42].

Burden and protection

Malaria is an endemic disease which spreads throughout most of the tropics. The world health organization (WHO) documented a quarter billion cases and more than half a million deaths from malaria in 2020[43]. The 2019 coronavirus pandemic is the culprit for the rise of malaria-related death[44]. The global disease burden of malaria is 8% according to recent WHO reports. Africa alone accounts for 95% of the global burden, with Southeast Asia, and the Eastern Mediterranean regions accounting for 2% each. The main determinants of malaria include the lifespan, the number of human bites, and the number of female anopheline vectors. In fact, an equation can predict the rate of transmission of malaria[45,46]

The life cycle of the deadly form of malaria, the *Plasmodium falciparum* malaria, includes several stages. A stage in which the parasite resides within the female (Anopheles) mosquito, and the human stage includes stages within the liver, and the red blood cells. [47,48] The genetic protection in sickle cell disease originates from the red blood cell stage. Indeed, studies found that the sickle cell trait is protective at 90% against severe forms (cerebral and anemic) and complications of the disease, and at 60% against hospitalization-related malaria infection[47,49]. The pathophysiology of sickle cell trait protection against malaria is not fully understood. Three main

hypotheses exist: increased sickling of red blood cells[50], weakened parasite growth during vascular sequestration[51], and the parasite's genotype[52]. Several additional hemoglobin mutations are protective against malaria including β^C, β^{SC}, β^E, α- and β-thalassemia[53-57] (**Figure 3**). Beyond the hemoglobinopathies listed before, erythrocyte membrane defects (e.i., elliptocytosis, and ovalocytosis), Dantu blood group, enzymatic deficiencies (e.i., glucose-6-phosphate dehydrogenase (*G6PD*), and glycophorins have been associated with resistance to infection[58-61].

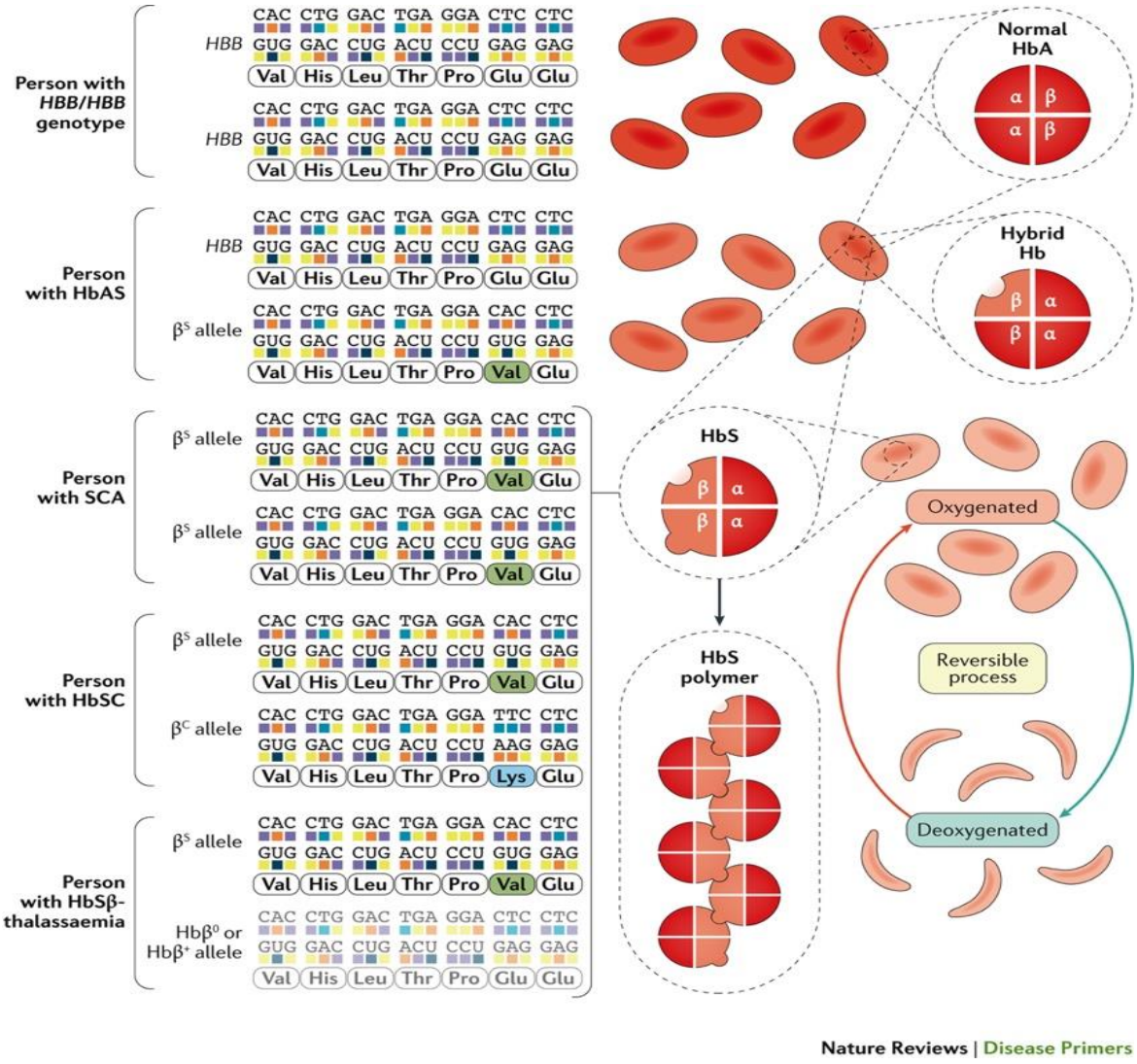


Figure 3. Genetic alterations in HBB. Copied from Kato et al 2018

Normal haemoglobin A (HbA) is formed by two α-globin subunits and two β-globin subunits, the latter of which are encoded by HBB. The sickle Hb (HbS) allele, β^S, is an HBB allele in which an adenine-to-thymine substitution results in the replacement of glutamic acid with valine at position 6 in the mature β-globin chain. Sickle cell disease (SCD) occurs when both HBB alleles are mutated and at least one of them is the β^S allele. Deoxygenated (not bound to oxygen) HbS can polymerize, and HbS polymers can stiffen the erythrocyte. Individuals with one β^S allele have the sickle cell trait (HbAS) but not SCD; individuals with sickle cell anaemia (SCA), the most common SCD genotype, have two β^S alleles (β^S/β^S). Other relatively common SCD genotypes are possible. Individuals with the HbSC genotype have one β^S allele and one HBB allele with a different nucleotide substitution (HBB Glu6Lys, or β^C allele) that generates another structural variant of Hb, HbC. The β^C allele is mostly prevalent in West Africa or in individuals with ancestry from this region HbSC disease is a condition with generally milder haemolytic anaemia and less frequent acute and chronic complications than SCA, although retinopathy and osteonecrosis (also known as bone infarction, in which bone tissue is lost owing to interruption of the blood flow) are common occurrences. The β^S allele combined with a null HBB allele ($Hb\beta^0$) that results in no protein translation causes $HbS\beta^0$-thalassaemia, a clinical syndrome indistinguishable from SCA except for the presence of microcytosis (a condition in which erythrocytes are abnormally small). The β^S allele combined with a hypomorphic HBB allele ($Hb\beta^+$; with a decreased amount of normal β-globin protein) results in $HbS\beta^+$-thalassaemia, a clinical syndrome generally milder than SCA owing to low-level expression of normal HbA. Severe and moderate forms of HbSβ-thalassaemia are most prevalent in the eastern Mediterranean region and parts of India, whereas mild forms are common in populations of African ancestry. Rarely seen compound heterozygous SCD genotypes include HbS combined with HbD, HbE, HbO^{Arab} or Hb Lepore (not shown).

Pathophysiology

Intraerythrocytic hemoglobin S (HbS) deoxygenation in tissues with elevated oxygen request produces a hydrophobic motif in the deoxygenated HbS tetramer[62]. As a result, HbS chains on different tetramer bind to each other to hide the hydrophobic motifs. This in turn, will cause the formation of long polymers, which will distort red blood cells (RBC) into sickle cell shape[63] (**Figure 4A**). Erythrocytes then become more rigid and sticky to endothelial walls. Repeated cycles of sickling and hemolysis, together with inflammation, result in severe and persistent organ damage (**Figure 4B**). Tissue ischemia combined with vaso-occlusion is responsible for acute chest syndrome, acute pain, and avascular necrosis[64]. Whereas hemolysis-related endothelial dysfunction leads to complications such as leg ulcers, priapism, stroke, and pulmonary hypertension[65]. While homozygote experience more severe symptoms at an early age compared to heterozygotes (i.e., HbSC), painful episodes and splenic infarction are common across genotypes[66]. Added characteristics linking these complications include concomitance of other globin genetic variants such as α-thalassemia and fetal hemoglobin expression[66]. All documented sickle cell disease genotypes and their features were catalogued by Rees *et al* (2010)[62].

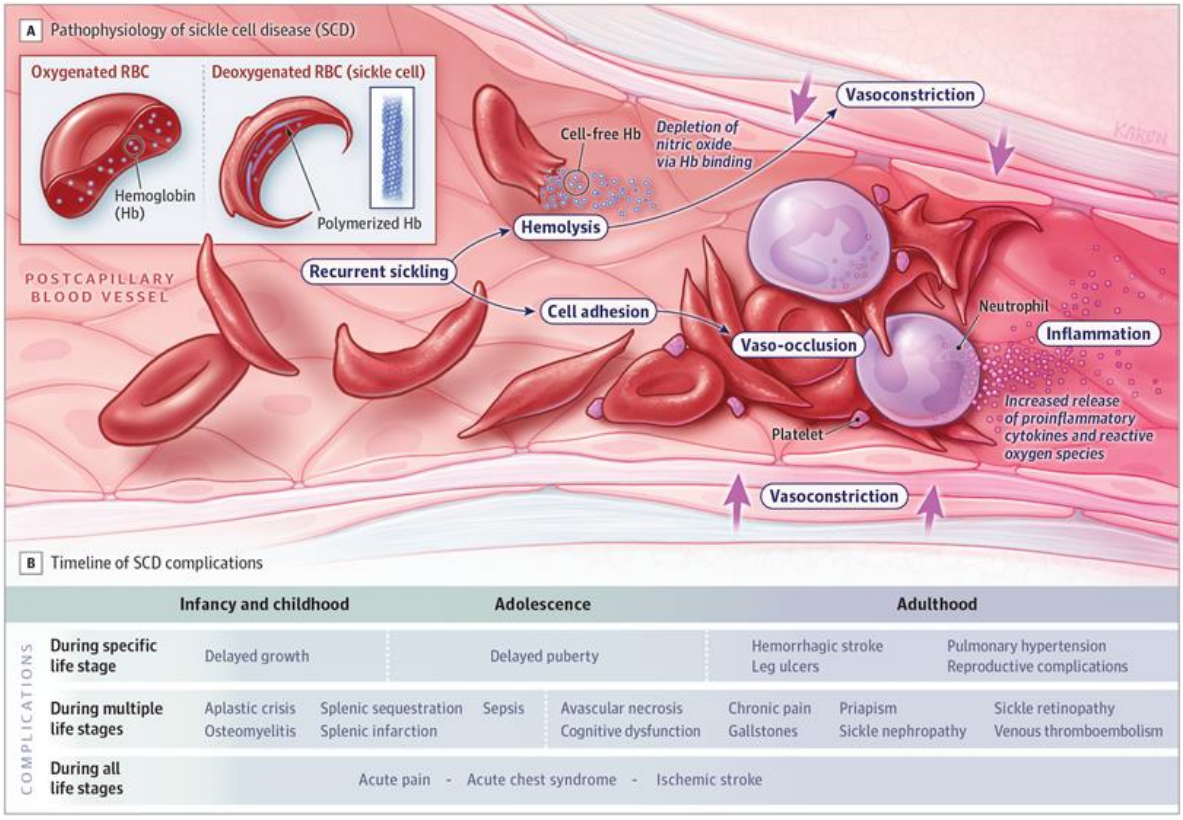


Figure 4. Pathophysiology and complications of sickle cell disease. Copied from Kavanagh et al 2022

A, HbS polymerizes when deoxygenated, inducing recurrent red blood cell (RBC) sickling and hemolysis. The sickled RBCs interact with white blood cells and platelets on vascular endothelium via adhesion molecules which leads to vaso-occlusion. The free Hb and heme released from RBC hemolysis trigger endothelial dysfunction due to depletion of nitric oxide and resultant vasoconstriction. The dual processes of vaso-occlusion and endothelial dysfunction activate inflammatory responses, via increased cytokines and reactive oxygen species, which perpetuates further vaso-occlusion.

B, The morbidity of SCD is progressive throughout the life span. Early on, most complications occur in acute recurrent episodes. Additionally, growth and puberty are delayed due to the increased metabolic demands secondary to ongoing hemolytic anemia. In adulthood, organ damage is prominent in addition to acute complications. Acute pain episodes, acute chest syndrome, and ischemic stroke can occur at any life stage.

Complications of SCD

Sickle cell disease is a multisystem illness. The disorder affects practically every organ in the body (**Figure 5**). SCD-related complications can be classified as chronic or acute.

Acute complications include:

Acute pain episode: painful crises are a hallmark of the disease. They are described as acute and continuous bone pain[67]. Often requiring emergency care, these painful crises are the consequence of blood vessel occlusion. The vaso-occlusive crises implicate tissue ischemia and inflammation. Many individuals can be triggered by cold, dehydration, wind, low humidity, alcohol, and stress.

Experiencing pain can start as early as six months of age and will stay throughout life. The site of pain can include the back, extremities, and abdomen. In young children, dactylitis (acute pain in the hands or feet) may be the most common location of the pain.

Acute chest syndrome (ACS) is established as the presence of liquid in the chest combined with pleuritic chest pain, fever, hypoxemia, or tachypnea[68]. ACS may result from pneumonia, fat embolism, or thromboembolism[69].

Fever is life-threatening in children as it may be a sign of sepsis since splenic removal is a common procedure for SCD patients, which makes them susceptible to infection. In fact, prior to penicillin prophylaxis and vaccines, one in two SCD children died from infections related to S pneumoniae and H influenza[70].

Children with SCD experience strokes at an alarming rate. In fact, about 10% of children with sickle cell anemia and β-thalassemia experience strokes[71]. Generally treated by RBC exchange transfusion, acute ischemic stroke in children is a consequence of vaso-occlusive events in major cerebral arteries[72].

A decrease in hemoglobin from baseline by 2g/dL or greater is considered acute anemia. Key features of this complication include upper body pain, enlarged spleen, thrombocytopenia (< 150 000/microL), increased reticulocyte count, and decreased hemoglobin[73]. The principal culprit of acute anemia is splenic sequestration, followed by aplastic crisis[74]. The incidence rate of acute anemia varies between 7% to 30% [68].

Pigmented stones will form in the gallbladder of sickle cell disease patients due to RBC hemolysis. About three out of four adult SCD individuals will experience this condition[75]. Among young adults, it is estimated that 43% will suffer from gallstone disease. Removal of the gallbladder (cholecystectomy) or laparoscopic is recommended for individuals presenting symptoms.

Priapism, characterized by an erection lasting over 4 hours, affects 40% of males in childhood[76,77]. Irregular levels of nitric oxide combined with RBC hemolysis prevent smooth muscle relaxation (**Figure 4**), leading to congestion of blood in the penis. Intermittent priapism (lasting less than three hours) and priapism can lead to scarring and erectile impotence[76].

Chronic complications include:

Chronic pain occurs between 30% to 40% of adolescents and adults. Research shows that the pain is neuropathic (results in nerve damage) and nociceptive (caused by damage to tissue, the

pain is sharp, throbbing, and aching)[78]. This pain is experienced even in the absence of vaso-occlusive events.

Avascular necrosis, also called aseptic necrosis or osteonecrosis, can result from bone ischemia of the shoulders, spine, or hips. The complication affects 10 to 32% of SCD patients[79]. Bone fractures are linked to avascular necrosis. Bone marrow infarction results in the death of hematopoietic cells, lower RBC production, and brings about anemia. Bone marrow infarction can lead to life-threatening pulmonary fat embolism.

Sickle cell disease can cause retinopathy which manifests as retinal artery occlusion, retinal detachment, and vitreal hemorrhage[80-82]. Interestingly, in contrast to other complications, SCD retinopathy is more severe for hemoglobin SC individuals compared to individuals carrying other genotypes[80,81,83]. It's noteworthy that eye disorders can be observed in nearly all adults.

Kidney trauma in sickle cell disease is referred to as nephropathy. Kidney injury is multifactorial and is often diagnosed when serum creatinine levels exceed standard threshold[84-86]. Hyperfiltration, defined as an eGFR greater than 180 mL/min/1.73m^2 is widespread and is seen early in children with SCD[87]. Microalbuminuria, albuminuria, and red blood sickling can lead to glomerular, and tubular damage.

Leg ulcers result from vaso-occlusion of the skin. While the pathophysiology of the complications is not well characterized, we know that blood flow, inflammation, endothelial dysfunction, slow rate of healing, and thrombosis play a role in the complications. Although the incidence in the US population is around 10%, it can be higher than 70%, as seen in Jamaica[88]. Highlighting the fact that individuals living in tropical regions are more affected.

Additional complications include pulmonary hypertension[65,89,90], cardiomyopathy[91-93], heart failure, asthma[93] and pregnancy-related[94], and infection-relations[95-98] complications which are beyond the scope of this review and are documented elsewhere[63,99].

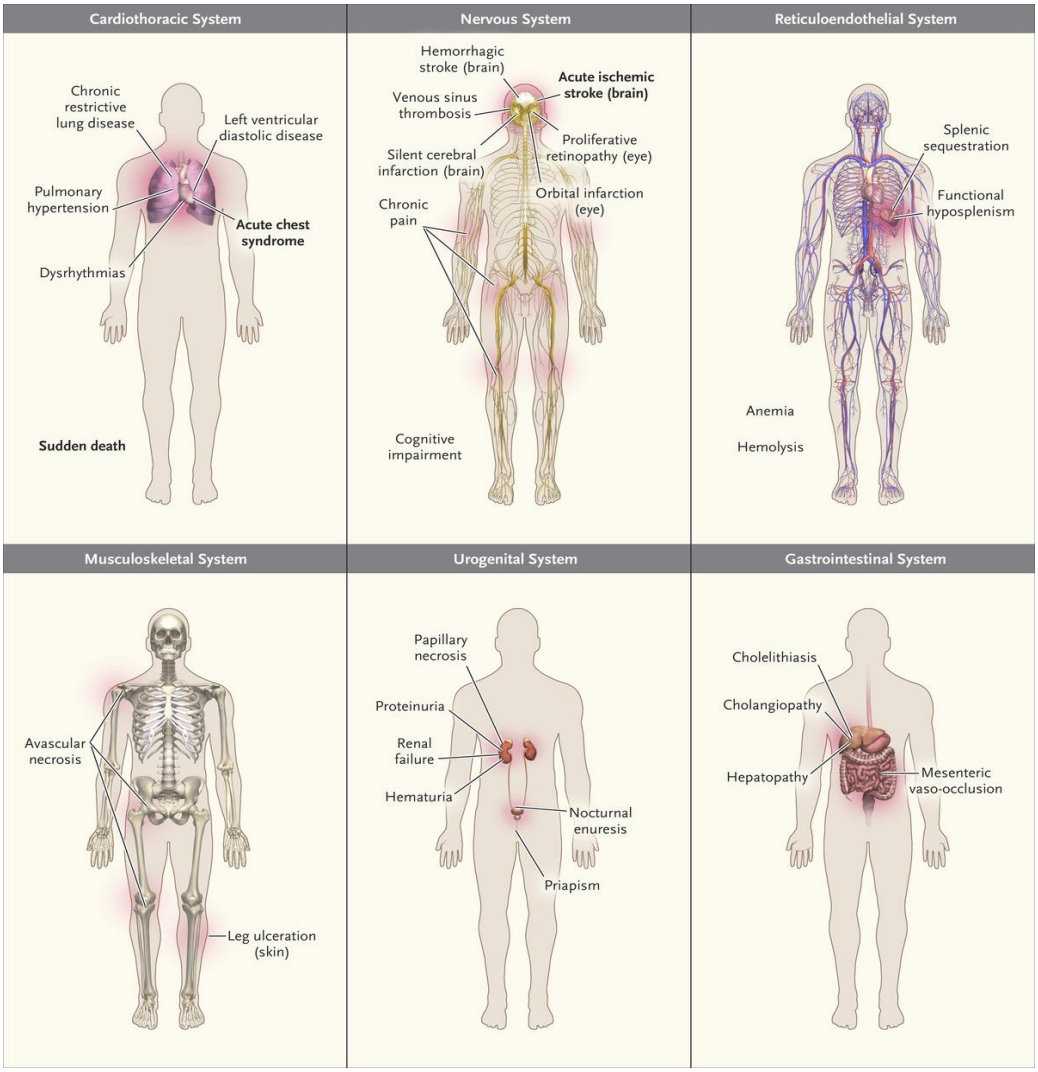


Figure 5. **Common complications of sickle cell disease**. Data are from Rees et al. 2010 and Serjeant. et al 1996 Acute complications are shown in boldface type. Image copied from Piel et al. 2017

Therapies in SCD

SCD is a century-old disease, yet, the lack of cost-effective therapies is jarring. Inadequate research funding thwarts high-quality randomized control trials (RCTs), and stifles the generation of substantial evidence. In an open letter published in the JAMA network, Farooq *et al*. (2020) conducted a comparative analysis of research productivity between cystic fibrosis and sickle cell disease studies. The authors confirmed that the 10-fold disparity in funding for cystic fibrosis compared to sickle cell disease contributed to the development of effective therapeutics for cystic

fibrosis patients[72]. For sickle cell disease, only three well-established therapies are recognized: hydroxycarbamide, blood transfusion, and hematopoietic stem cell transplantation[100].

Hydroxycarbamide, also known as hydroxyurea is a ribonucleotide reductase inhibitor, that leads to several physiological consequences including an increase in HbF levels and a decrease in leucocyte count[101]. The drug was approved by the FDA in 1998 and by the European Medicines Agency in 2007 for the treatment of adult SCD patients. The benefits of the compound included a reduction in vaso-occlusive crises, hospitalization, and mortality. In addition to confirming its safety[102], claims of reduced fertility, and risk of carcinogenicity were disproved in prospective studies[101,103]. While usage of hydroxyurea is elevated in high-income countries, upwards of 60% in certain cohorts, usage in low-income countries is close to inexistent[104,105].

Blood transfusion improves blood circulation in capillaries by reducing the number of circulating sickled red blood cells. Therefore, inflammatory damage and endothelial injury are significantly decreased. Additionally, in individuals at risk for stroke, frequent transfusion prevents vaso-occlusive crises and strokes. Several adverse events exist, including iron overload, immune response to a foreign antigen (alloimmunization), and hemolytic transfusion reaction.

Hematopoietic stem cell transplantation in SCD is curative. However, it requires a human-leucocyte antigen (HLA)-matched family donor. Given that the procedure is costly it is estimated that just 2,000 individuals have undergone the procedure, and more than 90% of the patients survived[106,107]. While hematopoietic stem cell transplantation from the bone marrow (**Figure 6**) cures SCD, the short supply of HLA-match donor [108] limits the impact of the therapy.

Since 2017 several new drug treatments have become available to SCD patients. Indeed the FDA approved L-glutamine[109], crizanlizumab[104], and voxelotor[110] to treat SCD. L-glutamine is an oral amino acid with the ability to reduce oxidative stress on RBC, therefore reducing sickling and RBC stickiness to the endothelial walls. Results from the clinical trial showed that L-glutamine reduces pain crises by 25%, and hospitalization by 33%. Crizanlizumab is a monoclonal antibody which prevents P-selectin, an adhesion molecule implicated in vaso-occlusion. The P-selectin inhibitor reduced pain crises from about 3 per year to 1.63. Voxelotor is a compound which

decreases hemoglobin polymerization rate and hemolysis by promoting HbS binding to oxygen. The RCT of Voxelotor increased hemoglobin level by 1.0g/dL (51%) compared to placebo[110].

Experimental therapies (**Figure 6**), such as gene therapy, revolve around two strategies. One strategy employs HbAT87Q (a synthetic hemoglobin with antisickling properties) injected into a patient's stem cell and infused back into the patient after chemotherapy. While this therapy yielded a significant increase in hemoglobin and a reduction of pain crisis and vaso-occlusive events, 2 out of the 35 patients developed dysplastic features along with anemia. Another strategy is to increase HbF expression by decreasing the expression of *BCL11A* (a γ-globin repressor which regulates HbF levels)[111]. Patients who received this therapy saw their fetal hemoglobin level increase by 2-fold in 18 months and experienced no pain crisis[112]. Other gene modifying therapies such as CRISPR/Cas9 aim to correct the mutated hemoglobin gene[113]. Finally, pyruvate kinase, proinflammatory cytokine inhibitor[114], and blockers of cellular adhesion[115] are experimental approaches to treat the disease.

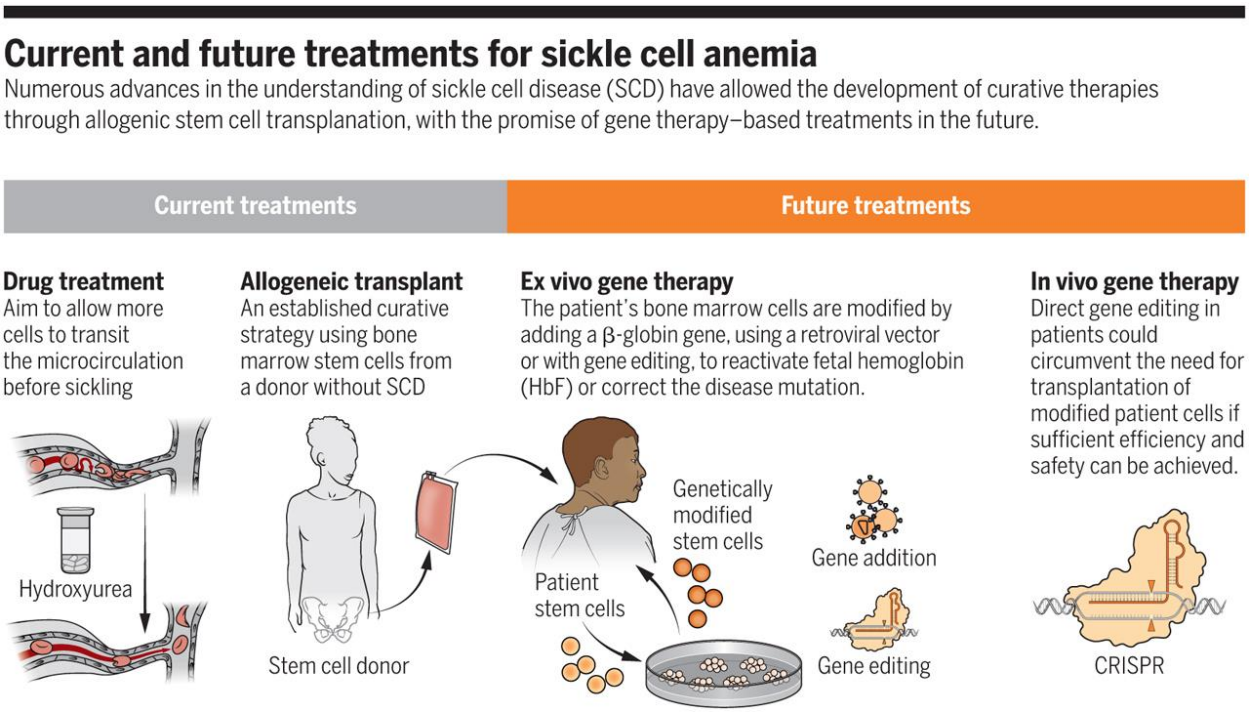


Figure 6. Current and future treatments for sickle cell anemia. Copied from Tisdale et al 2020.

Fetal hemoglobin in SCD

To understand the role of fetal hemoglobin in SCD, one must first understand the role of hemoglobin in humans. As the oxygen-carrying molecule, hemoglobin plays a vital role in humans. The tetramer most predominant in humans is adult hemoglobin (HbA), composed of two alpha chains and β-globin chains (**Figure 7**). A gene cluster on chromosome 16 contains from 5' to 3' the embryonic ζ-globin gene, and the two adults (α- globin) genes. On chromosome 11, we find the β-globin cluster with the embryonic gene (also known as ε-gene), two fetal γ-globin genes, and the adult genes, δ, and β genes. Different combination of these genes results in different hemoglobin tetramer that are expressed at various stages of human development (embryonic (ζ_2 ε_2), fetal (α_2 γ_2), and adult life (α_2 β_2; α_2 δ_2)) through the locus control region (LCR) (**Figure 7**). Fetal hemoglobin is the predominant form of hemoglobin during the last two trimesters of pregnancy.

In SCD, the polymerization of abnormal hemoglobin has some drastic consequences in patients. Red blood cells become dehydrated; they stick to the vasculature and ultimately lead to organ damage. X-ray crystallography studies have shown how fetal hemoglobin inhibits sickling[116]. Fetal hemoglobin's antisickling properties stem from a mutation replacing Gln for a

Thr at position 87 causing the lessening of the hydrophobic reaction. Resulting in a reduce co-polymerations.

The work conducted by Tom Maniatis and George Stamatoyannopoulos has shaped our current understanding of the globin switch. After birth, *BCL11A*, and *ZBTB7A* initiate the switch by blocking *LCR* γ-globin transcription. This causes fetal hemoglobin expression to be repressed, and adult hemoglobin becomes the main form of hemoglobin. While some levels of HbF in healthy humans exist, they are minute (< 1% of total hemoglobin). HbA represents 97% of the total hemoglobin, with the remaining contribution from the minor form of hemoglobin (HbA2)[117]. In 1949, scientist Janet Watson uncovered the link between the protective effect of HbF and SCD complications. She observed that newborn didn't show any complications until 6 months of life. In 1994, epidemiological advancements brought to life a large prospective study on the role of fetal hemoglobin in sickle cell disease. The study followed 3,764 SCD individuals for close to seven decades from their date of birth[118]. The key result showed that higher levels of fetal hemoglobin confer greater survival rate. In addition to extending life expectancy, fetal hemoglobin was found to decrease painful crises, other co-morbidities (i.e., acute chest syndrome, and osteonecrosis).

The advent of genetics, and genomics propelled our understanding of HbF in modern days. Common variation in twin studies revealed that HbF levels are hereditary ranging between 60 to 90% of heritability. When genome-wide association studies (GWAS) came onto the scene, they completely transform the field. They showed that alleles at *BCL11A*, *HBSL1-MYB,* and *HBB* were causally associated with higher concentration of fetal hemoglobin[119,120]. Together they explain about 50% of the heritability of fetal hemoglobin (HbF). Several functional studies have cemented *BCL11A* as the key regulator of fetal hemoglobin levels. However, mouse knockout study and CRISPR-Cas9 mutagenesis also showed that *ZBTB7A* and *ZNF410* are implicated in γ-globin gene repression[11]. To date, the largest GWAS of fetal hemoglobin across three ancestries adds up to 28,279 individuals with 3,963 SCD individuals[121]. A new locus on chromosome 2, *BACH2,* was identified as a regulator of HbF switching. *BCL11A* has become the main target for gene therapy in SCD. Whether it's targeting *BCL11A* mRNA, *BCL11A* erythroid enhancer, or *BCL11A* binding site in γ-globin gene promoters, through CRISPR-Cas9 or a base editing approach the outcome is to raise HbF level by repressing *BCL11A* or mimic HPFH mutation. In fact, an ongoing clinical trial for SCD and β-thalassemia by CRISPR and VERTEX therapeutics (CTX001), which targets the enhancer of *BCL11A* is showing promising results[11]. While

identifying new loci associated with fetal hemoglobin can uncover patient-to-patient variability, studying the mechanisms of red cell hydration can also provide additional insights.

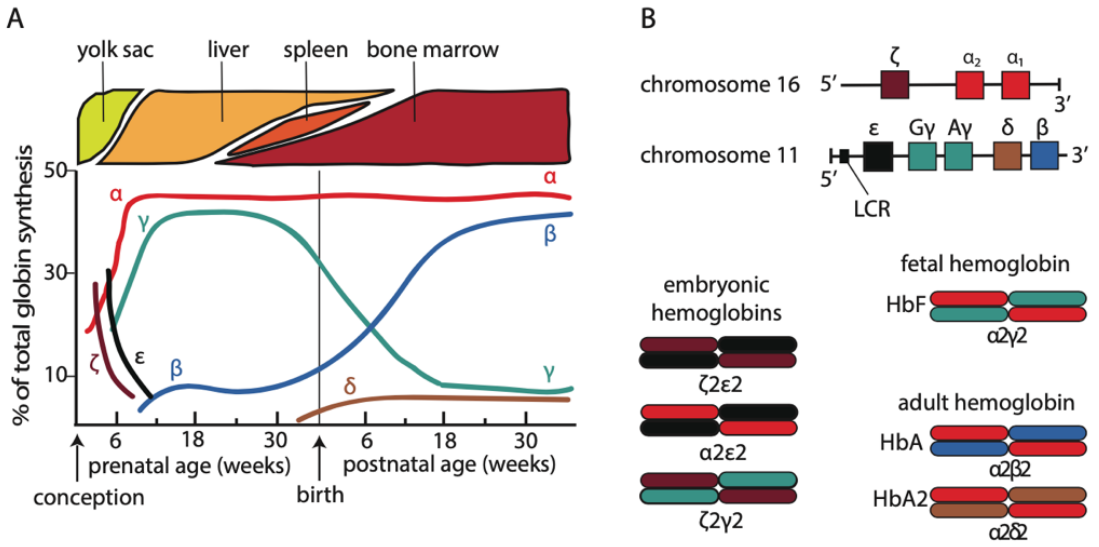


Figure 7 Globin switch in humans. A) Expression of human globins throughout development: hemoglobin switching occurs two times in human; one around the 6th week of gestation, when embryonic globins silence and HbE is replaced by HbF. The second switching occurs after birth, when HbF is almost entirely replaced by HbA. (B) Schematic representation of human globin loci and hemoglobins.
Copied from Maria Mikropoulou, An shRNA Screen for the Discovery of Suppressors of Fetal Hemoglobin, 2016

Red blood cell hydration in SCD

The rate of hemoglobin polymerization and the sickling phenomena are important to red cell hydration. Eaton et colleagues, showed that while increasing Hb S concentration promotes sickling, red cell hydration reduces hemoglobin concentration. Although therapeutic intervention targeting the swelling of sickle red blood cell showed a reduction in polymerisation in vivo, this approach was not explored further in a clinical setting[122].

Dehydrated, dense red blood cells (DRBC) are defined as cells with increased mean corpuscular hemoglobin (MCHC) levels and decreased mean cell volume (MCV). DRBC are a distinctive feature of SCD patients and can explain the patient-to-patient heterogeneity. In fact, a study of ~500 patients found that higher levels of DRBC raises the risk of renal dysfunction, leg ulcer and priapism[123]. Additionally, increased DRBC levels are protective against the malaria parasite *Plasmodium falciparum* [124,125]. Limited information about the genetic of red cell density exists. A candidate gene approach and a genome-wide scan in 374 SCD patients failed to identify strong association due to the lack of power[126]. However, a nominal association with a SNP mapping to

an *ATP2B4* enhancer, the main calcium channel pump in red blood cell was uncovered. Prior functional experiments found this enhancer to be causally implicated in modulating red cell volume and protecting against malaria infection[127].

To date, three key pathways are believed to be implicated in regulating red blood cell hydration in SCD patients. One of these pathways involves the potassium-chloride co-transporter also known as the KCC pathway. This transporter intermittently collaborates with the Gardos channel[128-130] to move solutes (potassium, water, and chloride) in or outside of red blood cells. KCC has four isoforms *SLC12A4*/*KCC1*, *SLC12A5*/*KCC2*, *SLC12A6*/*KCC3*, and *SLC12A7*/*KCC4*, which are all present in human erythrocytes. Knockout models and molecular characterization of all its isoforms (KCC1, KCC2, KCC3, KCC4) shed light on its role. Of interest, knockout mouse KCC3 (-/-) [131] 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 (-/-) [132] lead to deafness and tubular acidosis, and KCC2 (-/-) [133] is lethal just after birth due to respiratory failures.

The second pathway is the Gardos channel. The gene encodes the potassium calcium-activated channel subfamily N member 4 (*KCNN4*). In hypoxia, the red blood cell membrane allows extracellular calcium in and chloride out. This displacement of solutes has been linked to exacerbating the sickling process and thus the vaso-occlusive pathology[129,134]. Finally, a phase III clinical trial of 144 people targeting the Gardos channel with Senicapoc (ICA-17043) [135] proved beneficial to SCD patients. Several blood parameters, namely, reticulocyte counts, hematocrit levels, DRBC, and hemoglobin levels, were improved. Unfortunately, since the trial aimed at reducing the number of painful crises, the drug didn't move to the next phase of development. Finally, studies performed in mice and humans showed the promise and drugability of sickled cells.

Lastly, and still a topic of ongoing debate, the deoxygenation-induced fluxes *(P v*sickle *, Piezo-1)* pathway. Permeable to Na+, K^+, Rb^+, Ca^{2+} and Mg^{2+}, and not to Zn^{2+} or Mn2+, the pathway is qualified as a monovalent, non-selective, and divalent cation conductance. The mechano-sensitive ion channel, *PIEZO1*, appears to be the main channel for this pathway. Although some research shows that *PIEZO1* is an important key to red cell volume homeostasis, knockout studies in the mouse and zebrafish showed a mild effect on MCHC.

Metabolites in SCD

Metabolites represent the integration of gene expression, protein interaction, other regulatory processes, and the environment. They are, therefore, ideal for understanding and tracking by-products of the physiological progression of diseases. Measuring metabolites in SCD patients can uncover the heterogeneity of sickle cell disease.

In SCD, a handful of studies have leveraged metabolomic methodologies to tackle the patients' clinical heterogeneity. Zhang *et al*. discovered elevated adenosine levels in the blood of both SCD patients and transgenic mice. Their study showed that higher adenosine levels exacerbated sickling, hemolysis, and organ damage[136]. Furthermore, the same research group discovered that sphingosine-1-phosphate (S1P) and 2,3-bisphosphoglycerate (2,3-BPG) blood concentrations are elevated in SCD patients and mice. This elevation leads to the re-programming of glycolysis and exacerbates the severity of the disease[137,138] . Finally, Darghouth *et al*. conducted a comprehensive profiling of the metabolome in red blood cells (RBCs) from both healthy individuals and SCD patients. They identified a range of metabolites that underscore distinctions between the two groups, particularly in glycolysis, membrane turnover, as well as glutathione and nitric oxide metabolism[139]. Although exciting, these pioneering metabolomic studies were performed in a limited number of SCD patients (N = 14–30) and did not take advantage of MR methodology to address causality.

More recently, L-glutamine received FDA approval. RBC from SCD patients have high oxidative stress and a compromised ability to counteract free radicals due to a low ratio of the reduction-oxidation (redox) co-factor nicotinamide adenine dinucleotide (NAD) and its reduced form ([NADH]:[NAD^++NADH]) [140]. L-glutamine is one of the most abundant amino acids in the human body, and its role in protein synthesis is required to synthesize NAD. Treatment with L-glutamine increases the NAD redox ratio and reduces adhesion of sickle RBC to endothelial cells, a hallmark of vaso-occlusive painful crises[109,141]. Finally, the Lands' cycle is a significant biochemical pathway regulating the composition of erythrocyte membranes by utilizing two enzymes: lysophospholipid acyltransferases and phospholipase A2 (*PLA2*). In sickle cell disease, an excessive amount of *PLA2* activity can modify the composition of erythrocyte membranes, increasing their lysophospholipid content and leading to sickling and inflammation[142].

Methods for high dimensional molecular data analysis.

Recent advancements in high-throughput technologies have revolutionized medicine. They enhanced our comprehension of biology (i.e., interpretation of GWAS results from variant to function), and enabled better disease risk stratification and biomarker discovery in Type 2 diabetes and osteoarthritis[143,144]. The integration of various omics technologies, such as genomics, transcriptomics, and metabolomics, among others, can provide a more comprehensive understanding of disease processes. However, all forms of omics data require computational preprocessing before any meaningful biological insights can be derived. Typically, this preprocessing and the requisite quality control measures precede any form of statistical analysis.

Pre-processing

Next-generation sequencing (NGS) and mass spectrometry (MS) technology tend to introduce random variability in the data. If this variability is not properly accounted for, it can distort genuine biological signals, leading to potential issues in downstream analysis. Preprocessing and quality assurance are, therefore, crucial steps for data analysis. The origin of these errors depends on the bioassay and its biochemical properties. In DNA sequencing, for example, most errors arise from inaccurate base calls sequencing. Different companies have different biases; Illumina sequencers seem to have a bias for substitutions, while long-range sequencers like PacBio have a bias for insertions and deletions at the homopolymeric regions. Therefore, scoring quality bases and discarding poorer ones is an important first step in DNA sequencing analysis. Other sources of errors include adaptor contamination, duplicated reads, overrepresented sequence, and more. Many algorithms (FASTQC, Trimmomatic, CutAdapt, TrimGalore)[145] exist to perform quality assessments and correct these errors.

The quantification step transforms the quality-assessed raw data into quantitative values describing the abundance of the genetic variants, transcripts, or proteins. The purpose of this step is to ascertain the relative abundance of genetic variants, transcripts, or metabolites. For sequencing data, a typical procedure at this stage is to align and map reads to a reference genome with tools such as BWA or bowtie2[146,147].

This last step of preprocessing is a normalization step. For this thesis, we had access only to the raw data the metabolomics dataset. Therefore, our example for normalization is more applicable to the metabolomics datasets. To compare samples with each other and account for unwanted variability, the data needs to be normalized. For example, metabolites from different

families (fatty acids vs amino acids vs carbohydrates) are profiled using different ionization techniques. This will yield metabolite values with vastly different ranges. By normalizing, using inverse rank normal transformation, we can make the data more uniform and thus easier to contrast.

Statistical genetics approaches

Genome-wide association studies (GWAS)

Genetic recombination and linkage disequilibrium

The transmission of genes to the subsequent generation is determined by tightly wound DNA segments referred to as chromosomes. A healthy human possesses a total of 46 chromosomes, with each parent contributing half. During the formation of gametes, a phenomenon known as homologous recombination can occur, causing segments of DNA to be interchanged between each pair of chromosomes.[148].

Because of genetic recombination, alleles located within physical proximity on a chromosome have a higher likelihood of being inherited together. Thus, on a population scale, two alleles are more correlated to each other the closer they are to one another. This idea often confounds GWAS, as it can be challenging to know if a variant is causally associated with a phenotype or if it is located nearby the causal variant which is often inherited together. **Figure 8** exemplifies the difficulty in distinguishing which among the highly correlated variants is the actual causal variant. Methods for fine-mapping and LD clumping are discussed in other sections.

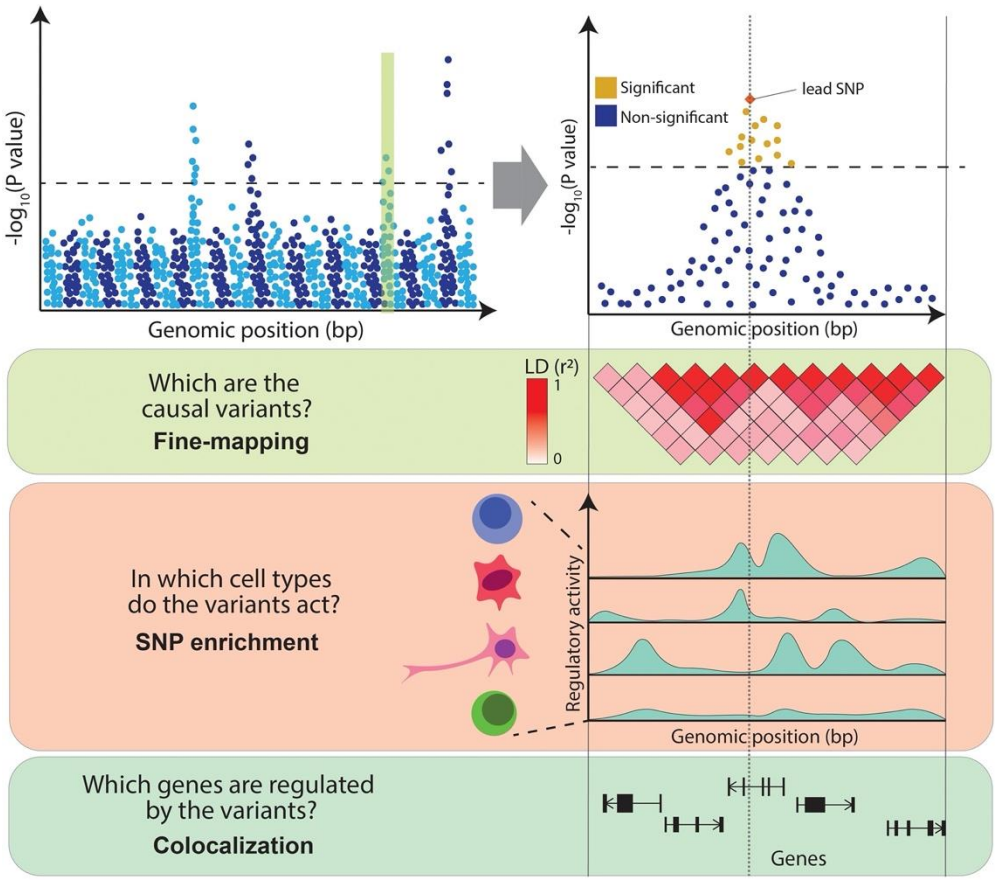


Figure 8. Challenges in interpreting GWAS associations. Copied from Cano-Gamez et al (2020) From the top: Manhattan plot illustrates the association between genetic variants and a trait (e.g., a disease) at a genome-wide level (left panel) and within an example locus (right panel). Variants above the dotted line represent genome-wide significant associations. The panels below illustrate the main challenges in interpreting GWAS associations: high LD between variants (encoded in shades of red), variable levels of regulatory activity of the genomic regions across cell types (peaks of different heights represent different levels of activity of chromatin marks) and multiple genes within the associated locus.

Regression models

GWAS are a powerful tool for dissecting the genetic architecture of a trait. Genetic associations refer to the association test between a SNP and a trait. The traits can be categorical (i.e., having or not having multiple sclerosis) or quantitative (i.e., height, weight, high-density lipoprotein cholesterol). For a given trait, the association will be significant if the disease frequency varies according to the genotype. This type of analysis highlights the region of the genome which influences the trait under consideration. For a quantitative trait, linear regression is employed to generate a test statistic, while for categorical traits, logistic regression is employed. SNP are tested individually with adjustment for covariates which can include age, weight, ancestry related

principal components which could be confound the phenotype of interest. The equation for a quantitative association is as follows:

$$E[y]=\alpha+x_c\beta_c+x\beta \quad (1.1)$$

$$y \sim N(\mu, \sigma^2) \quad (1.2)$$

y is a vector of phenotype values, x_c represents a matrix of covariate values across individuals, x is the genotypes of a given individual encoded as 0 for homozygous reference allele, 1 for heterozygous alternate allele, or 2 homozygous alternate allele. The intercept, the effect sizes of the genotype and covariates to be estimated are represented by α, β, βc respectively. For categorical association (binary traits) the equation is as follows:

$$E[y] = \frac{1}{1+e^{-(\alpha+x_c\beta_c+x\beta+\varepsilon)}} \quad (1.3)$$

$$y \sim B(p) \quad (1.4)$$

y is a vector of 0s, and 1s. with equation 1.1 being modified to fit a sigmoid function for logistic regression. A linear mixed model must be utilized, taking into account adjustments for sample relatedness using a genetic relatedness matrix (GRM). Since the GRM varies among samples - given that some are more related than others - a random effect model can sufficiently account for this variance.

Homoscedascity and heteroscedasticity

The homoscedasticity is a statistical concept that refers to the assumption that the variability of error terms, or residuals, is the same across all levels of an independent variable. This means that the spread or dispersion of your data is consistent across your dataset. In contrast, heteroscedasticity occurs when the size or spread of the residuals varies at different levels of the

independent variables. This can violate the assumptions of linear regression and may result in inefficient or biased parameter estimates. To ensure accurate results, it's important to check for homoscedasticity when performing regression analysis. If the assumption is violated, other methods, such as transforming the dependent variable or using a heteroscedasticity-consistent standard error estimate, might be more appropriate.

Normal distribution of residuals

The residuals of the model are assumed to be normally distributed, in the context of my analysis the dependent variable (phenotype to be tested) is transformed to be normally distributed by an inverse-normal quantile transformation. This assists in maintaining a normal distribution of residuals, assuming there isn't a significant departure from the homoscedasticity assumption. As previously discussed, this is improbable since the variance explained by any single variant being tested is low.

Representation of GWAS results

The standard presentation of GWAS results includes two types of p-value plots: Manhattan plots and quantile-quantile (QQ) plots. Manhattan plots depict the p-values of the entire GWAS on a genomic scale (**Figure 9b**). The p-values are arranged in genomic order by chromosome and position on the chromosome along the x-axis. The value on the y-axis displays the -log10 of the p-value (which is equal to the number of zeros after the decimal point plus one). Since genetic variants are locally correlated due to infrequent genetic recombination, significant p-values tend to form high spikes on the Manhattan plot (red dots on chromosome 2 of **Figure 9b**), giving the graph the appearance of a Manhattan skyline.

The QQ plot is a visual representation of the divergence between observed p-values and the null hypothesis. It works by arranging the observed p-values for each SNP in descending order and plotting them against the expected values derived from a theoretical $\chi 2$-distribution. If the observed values match the expected values, then all data points would lie near the middle line between the x-axis and the y-axis, which is the null hypothesis (as shown in the light gray line in **Figure 9a**). However, if some observed p-values are much more significant than expected under the null hypothesis, the data points will shift towards the y-axis, as depicted in **Figure 9a**. If there

is an early deviation of the observed from the expected (as seen in **Figure 9a**), it implies that many moderately significant p-values are more significant than expected under the null hypothesis.

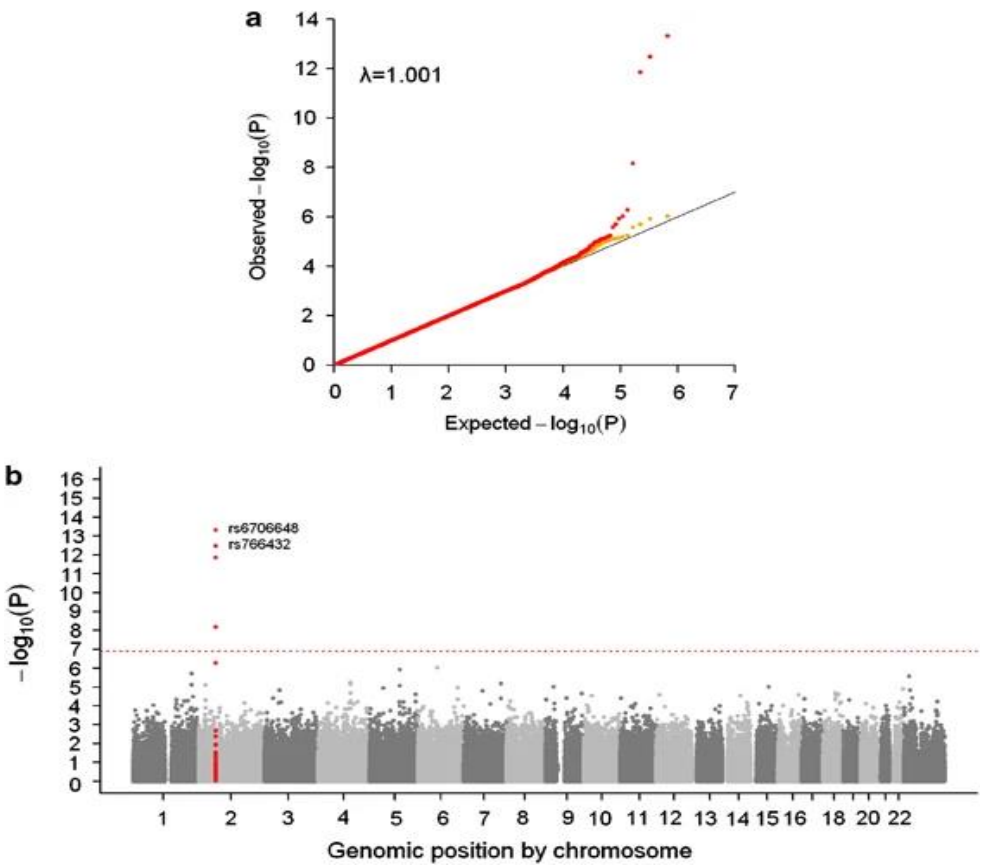


Figure 9. GWAS of proportion of F-cells in SIT Trial cohort (a sickle cell disease cohort). Copied from Bhatnagar et al (2011). Summary of the genome-wide association results of the proportion of F-cells in the SIT Trial cohort. (a) Q–Q plot of the observed versus the expected P-values from an additive genetic model for the entire set of 660 740 SNPs (red), and after removing genome-wide significant and their neighboring±100 kb region SNPs (yellow). (b) Manhattan plot for F-cells association results plotted against the position on each chromosome. The red color peak on chromosome 2 corresponds to the BCL11A region (±100 kb SNPs from rs6706648) and the red horizontal line represents a permutation-based genome-wide significant threshold (P-value <1.27 × 10−7).

Multiple testing

In the frequentist approach, statistical tests of a null hypothesis are considered significant if the p-value of association is below a pre-set threshold, typically 5%. If the statistical model's assumptions are correct, this method would wrongly reject the null hypothesis (yield a false positive result) in 5% of cases. In a GWAS analysis, each SNP is tested independently for an association, resulting in many simultaneous tests, and increasing the total number of false positives if the 5% p-value threshold is maintained. To minimize false positive results, I adjusted

the 5% p-value threshold by dividing it by the number of effective independent tests, which is also known as the Bonferroni correction. In a GWAS analysis, testing a high number of imputed variants means that many of these variants are highly correlated. Thus, the number of effective independent tests is less than the total number of variants tested for association, and the number of effective independent tests varies considerably depending on the MAF threshold. Studies that include many rare variants will carry out more independent tests since rare variants are less likely to be in LD with nearby variants. Alternative approaches have been proposed to estimate genome-wide significance threshold.

Conditional analysis

Conditional analysis allows to identify how many GWAS signals are present at a given locus and which variants are representative of the signal. Conditional analyses are not able to confirm the causal nature of a variant, only with follow-up experiment can this be done. We are, however, able to identify a credible set of variant with methods such as FINEMAP[149], CAVIAR[150], PAINTOR[151] and SuSIE[152] which give us a list of variants likely to contain the causal variant.
We can group conditional analysis into two categories, those with summary statistics GWAS like GTCA[153] and those with participant level genotypic information. Methods like GTCA rely on reference populations to perform the conditional analysis, whereas with conditional analyses without summary statistics rely on the LD pattern from the participant genotyped. This approach adjusts for regional association signal based on a set of variants in the locus by including the lead variant as a covariate in the regression model. In situations where several association signals are present, forward stepwise selection is performed until no associations remain[154].

History of GWAS in SCD

The first GWAS of sickle cell disease dates back to the discovery of *BCL11A* and its contribution to HbF levels in 2007[155,156], and pain crises in 2008[119]. We reviewed the important association studies in fetal hemoglobin, in metabolites, and red blood cell hydration in previous sections. In this section, we considered GWAS studies in SCD patients with replication. Since then, many mutations have been successfully identified including with bilirubin (*UGT1A*) which increases the risk of gallbladder disease in SCD[157,158]. SNPs in four genes were retained as modifier of the risk of stroke (*TGFBR3*, *TEK*, *ANXA2*, *ADCY9*) were identified as risk factors for strokes[159]. Two genes *APOL1*, and *HMOX1* have been confirmed as major regulator of kidney disease[160,161].

Further analyses should investigate rare variants or common variants with weaker effect size to find new associations with SCD complications, but also integrate the GWAS results with additional omics to gather a more complete picture of the role of these variants. A detailed review of known sickle cell disease genetic modifier replicated or not examining multiple SCD relevant phenotypes was recently published by Pincez *et al* (2022)[162].

Interpretation of GWAS

With the advent of cost-effective massively parallel genotyping arrays able to genotype millions of SNPs, cataloguing of human genetic variation in projects such as TOPMed[163], and robust statistical framework[164-166], we have witnessed a considerable increase in the number of association studies (**Figure 10**)[167]. As the number of samples increases, there is a greater need to interpret the results to create new scientific discoveries. We can group the main challenges of GWAS interpretation into three categories; (1) understanding the biological mechanism of the SNP and the tissue specificity, (2) confidently linking a SNP of interest to a gene through which it is acting, (3) estimating causal link between a risk factor and the disease risk.

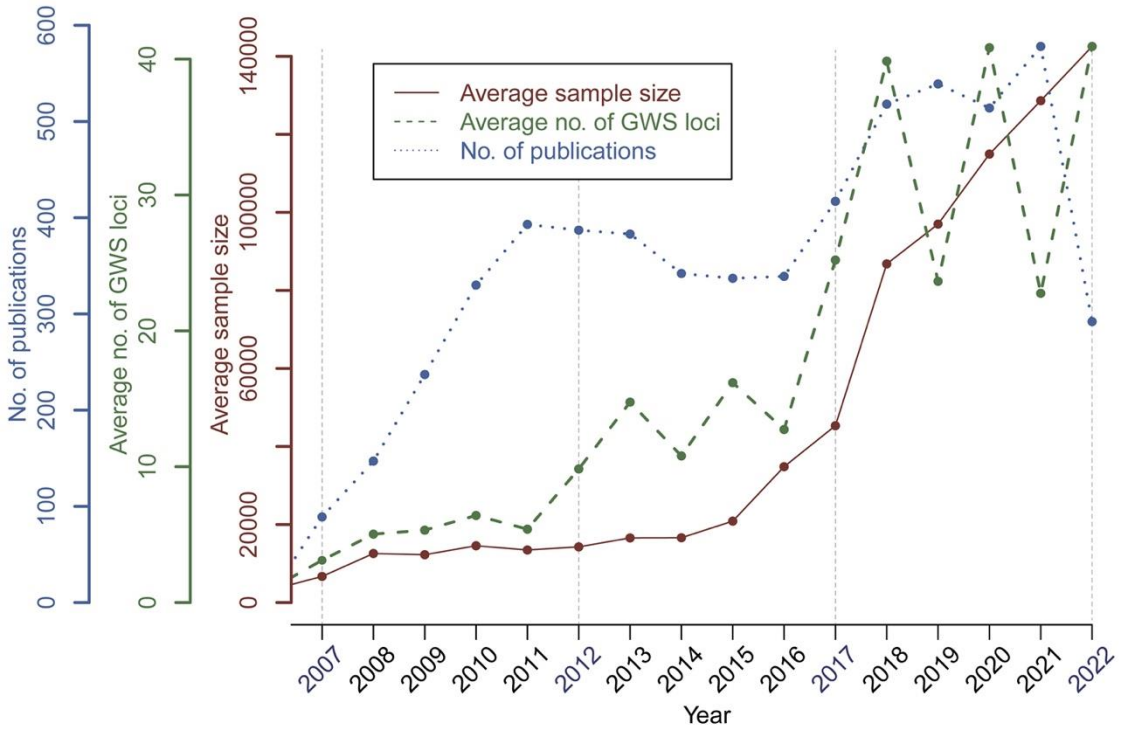


Figure 10. Average sample size and average number of genome-wide significant (GWS) loci per publication for each year during the 15 years history of GWAS discoveries. The data were extracted from 5,771 GWAS publications that used a genome-wide genotyping array and shared their summary statistics on GWAS Catalog before November 8, 2022. Copied from Abdellaouiet al (2023).

(1) Understanding genetic associations with colocalization analysis. Genetic colocalization analysis is a method used to investigate whether the same underlying causal variant influences different phenotypes that share a genetic association in the same location. In my research, I aim to gain a better understanding of the genetic associations related to various SCD-related haematological phenotypes. In this thesis, I don't employ colocalization, and more detailed review of how such method works can be found by Zuber *et al* (2022)[168].

(2) Although GWAS can provide important information about genes that contribute to a specific phenotype, it can be difficult to determine which gene is specifically responsible for the effect of an associated variant. This is because the variant may be located close to or overlapping multiple genes, or it may be located far away from the nearest gene. This presents a challenge in identifying the biological mechanism that is responsible for the associations identified by GWAS, particularly since it may not be clear which tissues are affected by the mechanism and how it leads to changes in the phenotype. To fully understand the biological mechanism behind genetic associations, it is important to identify the gene that is being modulated. There are two primary approaches to identifying the probable mediating gene(s) for a genetic variant.

- Mapping a variant within proximal distance from a gene can be done using an annotation software such as variant effect predictor (VEP)[169] :

- Integrating GWAS loci with additional omics datasets, gene expression, protein expression, open chromatin regions. Since a variant can overlap with multiple genes, some of which may not be expressed in a relevant tissue, it is advisable to combine genetic mapping with integrative analyses.

(3) Mendelian randomization (MR)

Mendelian randomization is a statistical epidemiology method that allows the estimation of a causal relationship between an exposure (i.e., HDL cholesterol) and an outcome (cardiovascular disease) using genetic variants across populations. Mendelian randomization relies on the natural, random assortment of genetic variants during meiosis. This is such that some individuals are naturally assigned alleles at birth that affect disease risks while others are not. MR mimics randomized clinical trials as it harnesses the random allocation of parental alleles when passed

on to their offspring (**Figure 11**). Therefore, alleles are randomly and independently distributed in the population and free from potential confounders. One of the success stories of MR is that of LDL-cholesterol and the risk of a heart attack[170]. Several R packages implementing MR exist and allow the sensitivity analysis to be performed [171-173].

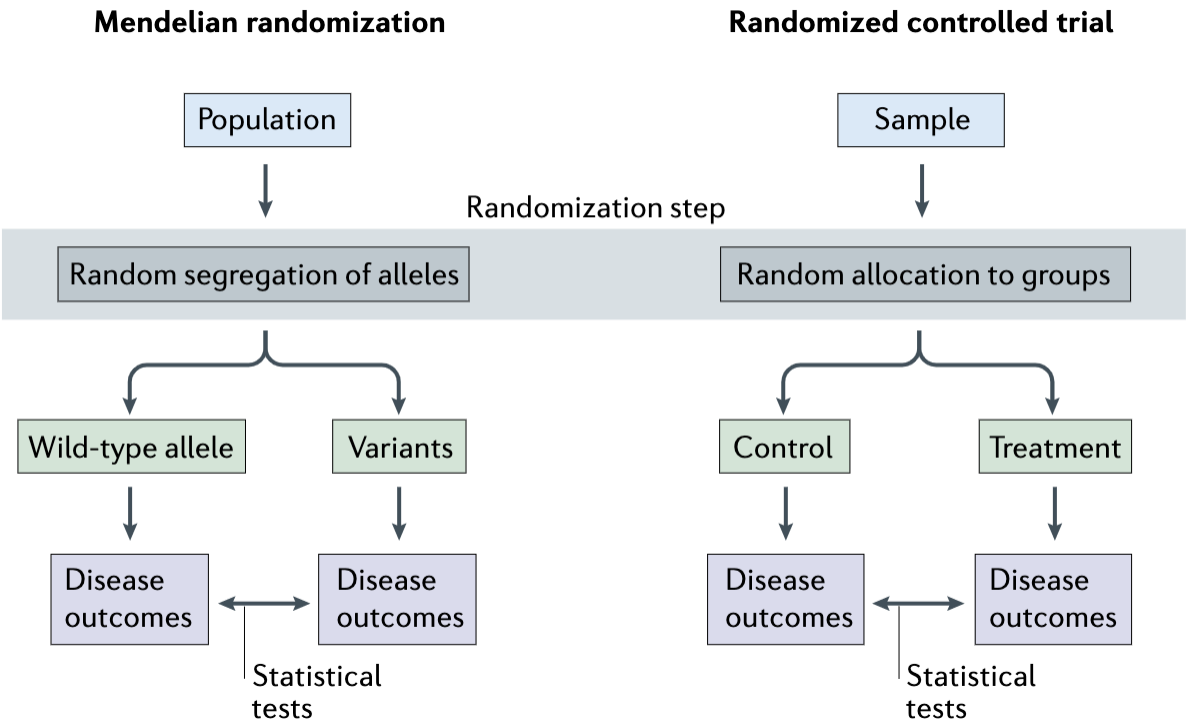


Figure 11 An overview of MR studies. Copied from Sanderson et al 2022. This overview compares and contrasts the parallels between Mendelian randomization (MR) and randomized controlled trials (RCTs). In MR, randomization is due to the random allocation of alleles. This conceptualization was originally based on between-sibling variation, where allocation of alleles is random and not dependent on population-level variation (see also BOX 1). Inference from MR in this way relies on the assumption of gene–environment equivalence — that a change in the exposure caused by genetic variation has the same effect on the outcome as a change in that exposure caused by environmental factors.

(3.1) Instrumental variables

Instrumental variables (IVs) are independent genetic variants associated with the exposure of interest. Inverse variance weighted (IVW) MR model is used to assess the causal association between the exposure and outcome of interest. The following three assumptions must hold for a genetic variant to be a valid IV[174] (**Figure 12**):

1. The variant must be strongly associated with the exposure.

2. The variant must be independent of any measured or unmeasured confounding factors which influence both the exposure and outcome.

3. The variant must not influence the outcome through any pathway other than the chosen exposure, often termed the 'exclusion restriction criterion'.

Assumption 1) can be verified by looking at SNPs' p-value. However, assumptions 2) and 3) are more difficult to ascertain because they rely on factors which may not be measured. If, for instance, a genetic variant affects another unknown factor that also influences the outcome, assumption 2) would be violated. Likewise, if the genetic variant is linked with modifications in an unmeasured confounding factor that affects both the outcome and exposure, it could create an apparent causal link between the exposure and the outcome. Since we are analyzing complex traits assumption 3) is hardly ever valid; variants usually have pleiotropic effects, which implies they affect various traits and phenotypes. In the context of IVW analysis assumption 3) is relaxed to assume 'balanced pleiotropy' between all instrumental variables. Balanced pleiotropy implies that the total amount of pleiotropy among all instrumental variables should equal zero. One way to test this is to use a funnel plot for qualitative assessment or MR-Egger, an extended version of the IVW model, for quantitative analysis, which allows the regression line intercept to vary.

(3.2) Mendelian randomization models

Overall, many MR models exist, each of which differed slightly in estimating the causal effect between the exposure and outcome. In this thesis I employed IVW, median based, mode based and MR egger regression:

Inverse variance weighted

IVW MR begins by calculating a Wald ratio between exposure and outcome for each instrumental variable:

$$\hat{\theta}_j = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}} \tag{1.5}$$

Where j is an index over all instrumental variables, $\hat{\beta}_{Y_j}$ is the estimated effect size of variant j on the outcome, and $\hat{\beta}_{X_j}$ is the estimated effect size of that variant on the exposure. As part of a

GWAS analysis, the determination of the effect sizes of variants on a phenotype (whether exposure or outcome) is conducted. These ratio of association estimates are meta-analyzed for each variant to estimate the overall causal association as IVW follows:

$$\hat{\theta}_{IVW} = \frac{\Sigma_j \hat{\beta}_{X_j}^2 se(\hat{\beta}_{Y_j})^{-2} \hat{\theta}_j}{\Sigma_j \hat{\beta}_{X_j}^2 se(\hat{\beta}_{Y_j})^{-2}} \quad (1.6)$$

The individual ratios are weighted by their associated uncertainty $\hat{\beta}_{X_j}^2 se(\hat{\beta}_{Y_j})^{-2}$. This is such that a variant with a small effect on the exposure and highly uncertain influence on the outcome is a down-weighted in the overall causal estimate.

Other models of MR exist and can be grouped into four categories: weak instrument robust methods (i.e., MR RAPS, NOME adjustments), outlier selection and removal (i.e., weighted mode, MR LASSO, Steiger), outlier adjustment (MR PRESSO, MR Robust, MRMix), estimation adjustment (i.e., MR Egger, multivariable MR, Bayes MR) and environmental control adjustment (i.e., MR GxE, MR GENIUS)[174].

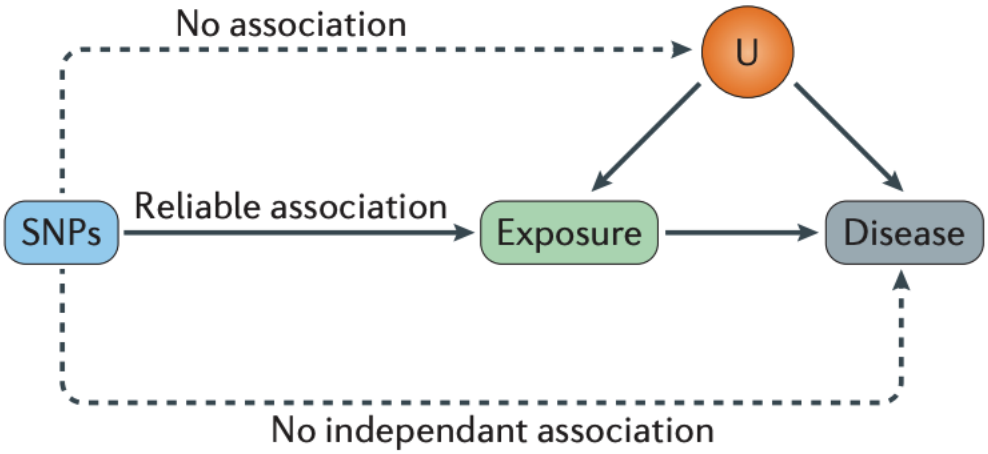


Figure 12. The three principles of instrumental variable analysis: the instrumental variable (in this case a genetic variant either in isolation or in combination with other variants) must associate with the exposure; the instrumental variable must not associate with confounders that are either known or unknown (U); and there is no pathway from the single nucleotide polymorphism (SNP) to disease that does not include the exposure of interest. This figure is a schematic representation and should not be interpreted as a formal directed acyclic graph.

Association-based approaches

Data-driven clustering

Multiple data driven approaches exist for clustering gene expression data. We can categorize clustering algorithms into four distinct groups: graph-based clustering, representative-based clustering, hierarchical clustering, and density-based clustering. Clustering algorithms based on graphs utilize structures reminiscent of graphs, such as K-nearest-neighbor graphs or affinity graphs, to assemble genes that exhibit strong connections within this graph-like framework. Techniques rooted in representation iterate the process of enhancing both cluster assignments and a representative entity (like the centroid) for each cluster. Hierarchical clustering strategies craft a hierarchical structure for all genes in the expression matrix. Meanwhile, density-based methodologies identify modules by scrutinizing areas with concentrated high density. It's important to recognize that certain clustering methods integrate elements from multiple categories, showcasing a blend of approaches. A team of researcher in Belgium, Saelens *et al.* 2018, has found that all clustering approaches except for the density-based clustering perform

the similarly. In our metabolomics experiment, we wanted to ask which metabolites cluster together based on their profiles and how these clusters can inform us about SCD pathophysiology. Graph-based methods such as weighted gene co-expression network analysis (WGCNA), multiscale embedded gene co-expression network analysis (MEGENA), and bipartite clustering represent a few data-driven methods that can generate insights into molecular data generated on populations of individuals[175-179]. The objective of these methods is to identify clusters or networks of molecular features, such as metabolites, that are more similarly related to each other than to any other features under consideration. For computational efficiency, most of these methods assume that a given gene can only be represented in one cluster across all the samples. WGNCA, with over 9,000 citations as of August 2021 according to Google Scholar, and ranked in the top 1% in the academic field of computer science, is an extensively utilized clustering method. WGCNA offers a versatile toolkit for investigating network module structures, quantifying the associations between genes and modules through module membership details, probing the interconnections among different modules using eigengene networks, and arranging genes or modules in a prioritized order, such as based on their relevance to a specific sample trait. Additionally, WGCNA can facilitate the formulation of hypotheses that are amenable to empirical validation using external datasets. As an illustrative instance, WGCNA might propose a potential correlation between a module, potentially representing a hypothetical pathway, and the outcome of a particular disease, thereby presenting opportunities for further substantiation. This approach starts with computing the correlation between all metabolite pairs accounting for the correlation of metabolites with one another. Then to magnify and improve the signal-to-noise ratio, the correlation matrix is raised to a given power (Beta). The Beta is selected in such a way that the adjacency matrix (correlation matrix raised to the power of Beta) follows a scale-free topology. With a few metabolites having a high number of connections. Then, the adjacency matrix is transformed into a topological overlap map (TOM), which captures both direct and indirect interactions. Hierarchical clustering is then applied to the TOM to organize the metabolites into modules. Generally, as part of the module identification procedure applied to the hierarchical cluster tree, random colors are assigned to signify each cluster. Metabolites that do not fit into any module (indicating the gene is not expressed or does not co-vary well enough with any other metabolite) are assigned to a "grey module" (metabolites that did not meet the module inclusion threshold). **Figure 13** recapitulates the steps.

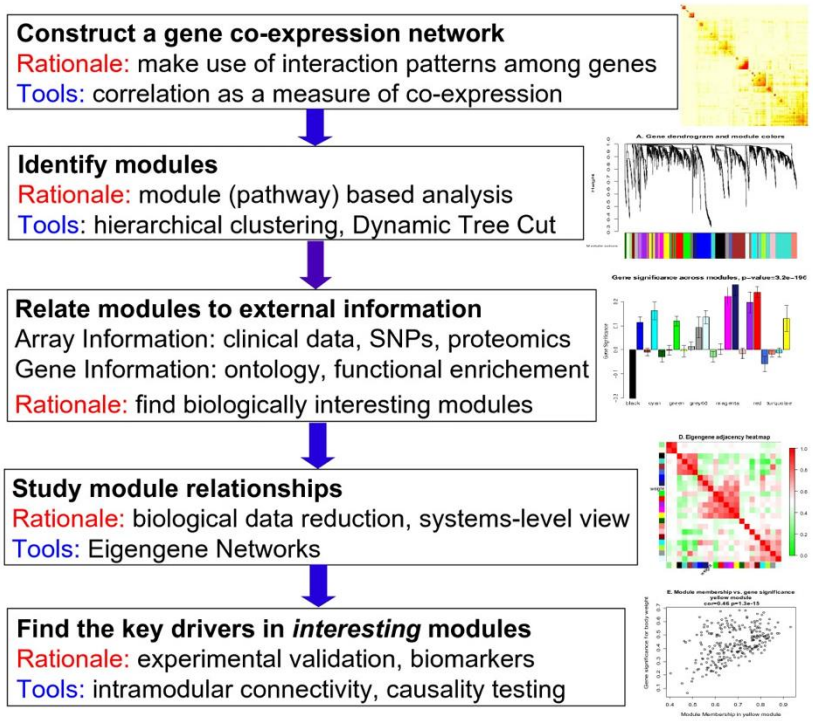


Figure 13. Overview of WGCNA methodology. Copied from Langfelder, P. et al (2008). This flowchart presents a brief overview of the main steps of Weighted Gene Co-expression Network Analysis.

Leveraging multi-omics approaches

Array-based genotyping

Genotyping array technologies are based on probes which target known single-nucleotide polymorphism[180]. Even though different technologies employ different assays, they all hybridize probes to select a SNP. Different colors enable the detection of A/T or G/C or of probe mismatch. A larger number of arrays are being commercialized each year, each looking at specific regions of the genome. Even though these SNP array target between 0.001% to 0.1% of the genome[181], their utility is of high value. This is because they rely on linkage disequilibrium (LD). LD describes the role of meiotic recombination events on inherited haplotypes. As humans evolved and recombination events occurred, regions with fewer recombination events led to structure between parts of the genome that correlated more with each other than it expected by chance[182]. This concept can be leveraged to strategically genotype SNPs that tag LD blocks. These blocks can then be used to impute other bases, although they were not directly typed[183]. Detailed genotype quality control steps have been reviewed elsewhere[184].

Imputation of genetic variants

Imputation allows the prediction of genotypes which were not directly typed. Imputation leverages the LD pattern amongst variants from a reference population to infer the genotypes. To ensure the imputation is performed correctly, the reference population and the imputed population must have similar ancestry. The first step of imputation is phasing (**Figure 14**), which estimates the contiguous regions of DNA with little evidence of recombination along the genome in the reference population. Once those regions are estimated, haplotypes from the reference population are used to impute missing genotypes in the sample population. The information (INFO) score metric, which ranges between 0 and 1 (with 0 meaning there is a lot of complete uncertainty about the genotype while 1 mean no uncertainty about the genotype imputation), is used to estimate the accuracy of the imputed genotypes. Using a variant's info score we can determine the power of association test[185]. Examples of reference panels including African ancestry populations include 1000G, TOPMed, HAPMAP, and CAAPA.

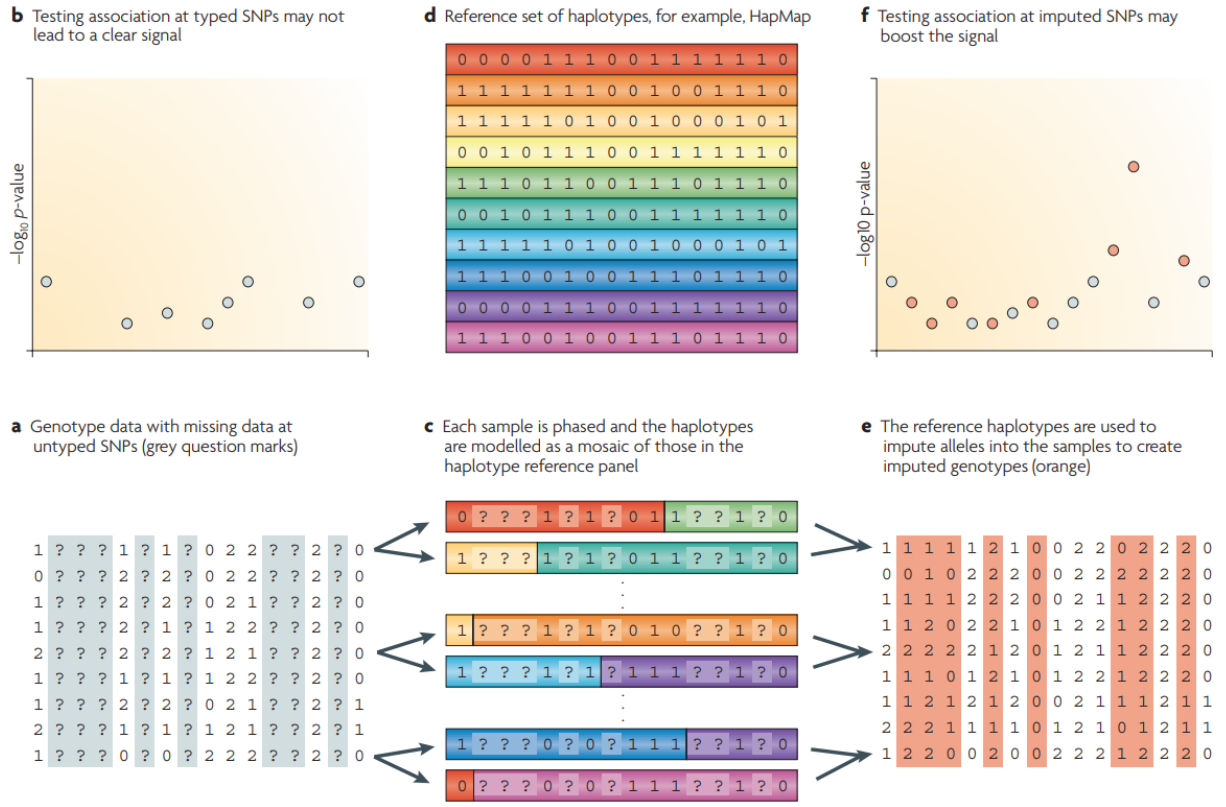


Figure 14. Schematic of steps for imputation of genotype data to estimate missing variants. Copied from Marchini *et al* 2010. a) Genotype data from the sample population with missing genotypes represented by question marks. b) Testing for an association signal using genotype data alone results in no association peak. c) Using a set of reference haplotypes from d) genotype data is phased to determine haplotypes present at each position along the genome. Three phased individuals are represented in the figure, each genome is a mosaic of haplotypes from the reference population. d) Reference haplotypes are defined from whole genome sequencing of a population with similar ancestry to the sample population. e) Missing variants in the genotyped sample population are estimated using the imputation procedure, with imputed variants highlighted in orange. f) In this example, testing for association of genotyped and imputed SNPs results in an association signal which was not identified before.

RNA-sequencing

RNA sequencing is considered a breakthrough for studying RNA. In 1965, Robert Holley *et al* were the first to publish a sequence of 77 alanine transfer RNA[186] which required 7 years to characterize. While we had to wait more than 20 years for the first automatic DNA sequencer to be commercially available, the advent of methods such as polymerase chain reaction (PCR), inverse transcriptase, and Taq polymerase allowed the first study using cDNA to sequence 397 nucleotides known as expressed sequence tags to be published[187]. The process involves converting RNA into cDNA and then sequencing the cDNA using short read sequencing RNA-seq allows for a direct estimation of the number of RNA molecules for a given gene. This process quantifies the expression levels of transcripts and generates tens of millions of reads which are

then assembled and aligned to a reference genome. Several pros and cons exist for the method[188]. One advantage is the fact we can easily detect different isoforms, ascertain circular RNA, long non-coding RNA, and bacterial genes[189]. Various tools are available for assessing differential expression, as well as visualizing and interpreting RNA-seq data[190,191]. The first step, mapping of reads, is often performed with Bowtie, HISAT2, STAR, TopHat[192-196]. The second step counts the number of raw reads aligned using featureCounts[197] or HTseq[198]. Given that the raw number of reads can be biased by the size of the transcript, the reads number are normalized, Fragment per kilobase of exon per million mapped reads (FPKM) or transcripts per million (TPM)[197]. Software tools such as Cufflinks, DeSEQ2 or RSEM are employed to normalize counts to FPKM or TPM and compute differential expression.

Metabolomics

Metabolomics is the study of metabolites, the intermediary products of metabolism. Metabolites are usually measured after multiple steps of quenching and extractions[199], using either nuclear magnetic resonance (NMR) based approaches or mass spectrometry methods in tandem with liquid chromatography (LC) or gas chromatography (GC)[200]. Metabolites are small molecules which reflect an individual's physiological processes. Samples can originate from feces, saliva, tissue, urine, blood, and breath. Data collection can be performed using a targeted approach, where metabolites are compared to internal standards, or belong to a pre-selected family of metabolites (amino acids, bile acids, or lipids). Gathering data can also be performed through an untargeted approach. Depending on whether the analysis requires hypothesis generation or specific hypotheses are being tested, supervised or unsupervised analyses can be performed. Tools such as MetaboAnalyst, WGCNA, and databases such as KEGG, and HMDB available for metabolomics analyses are reviewed in great details here[201,202].

Research questions and outline of the thesis

Sickle cell disease (SCD) is a genetic disorder caused by a point mutation in the -globin gene. It affects more than 300,000 newborns each year, with half of them tragically succumbing before reaching the age of five years. The complications related to the disease are systemic as they affect multiple organ systems. While some treatments exist to cure the disease, the clinical heterogeneity remains a puzzle. We decided to look at the genetics of fetal hemoglobin (HbF), red cell density (DRBC), and metabolites to unravel the heterogeneity in SCD.

Hypothesis: (1) The remaining variation in fetal hemoglobin could explain the remaining heterogeneity in sickle cell disease.
(2) Identifying red blood cell density loci will inform on therapeutic agents complementary to HbF's.
(3) In 2017, the FDA approved L-glutamine, which became the first metabolite approved for treating sickle cell disease. We hypothesize that metabolomics combined with genomics can aid in identifying drug targets for SCD patients.

General objective: Identify new genes and molecules that influence SCD heterogeneity.

Specific objectives: First, I applied stepwise conditioning on the three-main HbF loci (*BCL11A*, *HBSL1-MYB*, *HBB*) (Chapter 2) to discover new FH regulatory loci. Then, I evaluate drug targets controlling RBC hydration (Chapter 3) prioritizing GWAS and ExWAS genetic variation. Then, I estimate the causal role of L-glutamine in painful crises in SCD using Mendelian randomization (Chapter 4). Moreover, I used weighted gene co-expression network analysis (WGCNA) to identify metabolite clusters influencing RBC traits (Chapter 5). Finally, I explored the role of rare coding in hereditary spherocytosis (Annex E), and I characterized the effect of a common gain of function mutation on red blood cell density (Annex F).

Expected impact:

-Provide insights into the genetic complexity of SCD disease modifiers.

-Estimate the causal relationship between metabolites and the SCD-related diseases.

-Inform on the therapeutic interventions to treat SCD-related complications.

Chapter 2: Multi-ancestry meta-analysis identifies three novel loci associated with fetal hemoglobin levels

The following article is intended to be submitted in the journal, *American Journal of Human Genetics*. In this article, to identify novel genetic regulators of HbF levels, we combined association results at 24,272,278 variants ("combined" minor allele frequency (MAF) ≥1%) from 5,903 European-ancestry individuals from the SardiNIA Study[203] and 3,740 SCD participants, mostly of African descent. Because of the genetic heterogeneity in these populations, we used PCs as covariates and opted to analyze each study individually. For the meta-analysis, we used MR-MEGA, which was developed to account for ancestry differences to maximize discovery power in GWAS [204]. Additionally, we performed whole exome sequencing (WES) association testing in 1,354 SCD patients. In this study, the GWAS section involved conducting genotype quality control, phenotype harmonization and normalization, imputation, performing the conditioning analysis on each SCD cohort, running the cohort-specific GWAS and finally the meta-analysis. I performed all the previously described steps, except for the meta-analysis which was performed by Ken Sin Lo (co-author). He also generated the plots for the meta-analysis, and the UMAP. In the WES section, I conducted the quality control of the exome sequences, the exome-wide association analysis, the phenotype normalization, and plotted all the results.

Multi-ancestry meta-analysis identifies three novel loci associated with fetal hemoglobin levels

Yann Ilboudo[1,2,*], Nicolas Brosseau[1,2,*], Ken Sin Lo[1,2], Mélissa Beaudoin[1,2], Florian Wünnemann[1,2], Pablo Bartolucci[4], Frédéric Galactéros[4], Philippe Joly[5,6], Allison E. Ashley-Koch, Marilyn J. Telen, Swee Lay Thein, Carlo Sidore[3], Francesco Cucca[3] Abdullah Kutlar, Carlo Brugnara, Guillaume Lettre[1,2]

Affiliations:

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

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

[3]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy.

[4]Red Cell Genetic Disease Unit, Hôpital Henri-Mondor, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Est, IMRB - U955 - Équipe no 2, Créteil, France

[5]Unité Fonctionnelle 34445 'Biochimie des Pathologies Érythrocytaires', Laboratoire de Biochimie et Biologie Moléculaire Grand-Est, Groupement Hospitalier Est, Hospices Civils de Lyon, Bron, France

[6]Laboratoire Inter-Universitaire de Biologie de la Motricité (LIBM) EA7424, Equipe 'Biologie Vasculaire et du Globule Rouge', Université Claude Bernard Lyon 1, Comité d'Universités et d'Etablissements (COMUE), Lyon, France

Correspondence:

Guillaume Lettre

Montreal Heart Institute

5000 Bélanger Street

Montreal, Quebec, Canada

514-376-3330 (ext. 2657)

guillaume.lettre@umontreal.ca

ABSTRACT

Sickle cell disease (SCD) affects millions of individuals worldwide. It affects more than 300,000 newborns each year, half of whom will die before five. The complications related to the disease are systemic as they affect multiple organ systems. SCD patients with elevated fetal hemoglobin (HbF) levels present better clinical course and survival, therefore understanding the genetics of HbF could offer curative therapies. Genome-wide association studies yielded the first insights into humans' genic modulation of HbF. In fact, across different ancestries, loci at *BCL11A, HBSL1-MYB, and HBB* account for at least 50% of the variability of fetal hemoglobin levels. Leveraging the power of genome-wide association studies, we investigated the role of rare and common genetic variants regulating HbF levels in 9,643 imputed individuals of African and European ancestry and in 1,354 exomes from SCD patients. To identify novel genetic regulators of HbF levels, we combined association results at 24,272,278 variants from 5,903 European-ancestry individuals from the SardiNIA Study and 3,740 SCD participants, mostly of African descent. Because of the genetic heterogeneity in these populations, we used PCs as covariates and opted to analyze each study individually. For the meta-analysis, we used MR-MEGA, which was developed to account for ancestry differences to maximize discovery power in GWAS. We also focused our downstream analyses on association results conditioned on genotypes at the known HbF loci (*BCL11A*, *HBS1L-MYB*, β-globin). This strategy identified three loci that reached genome-wide significance ($P \leq 5 \times 10^{-8}$): chromosome 10 in *BICC1*, chromosome 19 near *KLF1* and chromosome 22 between *CECR2* and *SLC25A18*. The SardiNIA study previously reported an association between a rare intronic variant in *NFIX* on chromosome 19 (rs183437571) and HbF levels. This variant is not associated with HbF levels in SCD patients (MAF=0.0052, P=0.66), and is also not in linkage disequilibrium with the sentinel variant (rs4804210) that we identified in our multi-ancestry HbF meta-analysis (r^2 in European- ancestry populations = 0.027; r^2 in African-ancestry populations = 0). For the whole-exome association study, no variants identified reached exome-wide significance levels. However, our qualitative analysis highlighted a variant in *DMNT1,* which could be considered a hereditary persistence of fetal hemoglobin (HPFH) mutation. This GWAS represents the second to the largest meta-analysis of HbF (N=9,643) and the largest for SCD patients (N=3,704). It is also the largest whole-exome study in SCD patients. The highlighted loci could lead to therapeutic drug targets and therapies to lessen the severity of SCD.

INTRODUCTION

β-hemoglobinopathies are established anemias caused by single mutations or short deletions[205] in the adult β-globin gene. β-thalassemia and sickle cell disease (SCD) are the most common hemoglobinopathies. SCD results from a structural alteration (glutamine-valine substitution at codon 6) in the β-globin protein. Individuals affected by SCD suffer chronic and acute complications, leading to organ damage and death. Worldwide it is estimated that over a quarter billion individuals carry the mutation, and upwards of 300,000 new babies are born every year[206]. The clinical heterogeneity of the disease remains a puzzle as some individuals are asymptomatic while others experience severe and devasting forms of the disease. Observational studies showed the benefits of fetal γ-globin gene expression in ameliorating clinical manifestation in both disorders[118]. Therefore, understanding how the γ-globin gene is silenced can enable the development of new therapeutic interventions.

In 2007 and 2008, genome-wide association studies unearthed the first loci of fetal hemoglobin[119,155,156]. These reports pointed to three regions of the human genome: *BCL11A* on 2p, the intergenic region *HBSL1-MYB* on 6q, and *HBB* on 11p. These three loci account for at least 50% of the genetic variation of fetal hemoglobin across populations[11]. Since then, association analyses in healthy Europeans unearthed proteins encoding nuclear factor I X, *NFIX*[207] and glutathione-specific gamma-glutamylcyclotransferase 2, *CHAC2*[208]. Additionally, DNA methyltransferase 1 (*DNMT1*), previously reported to epigenetically silence γ-globin through methyltransferease maintenance[209,210], was recently identified in an exome-wide association study of Chinese β-thalassemia patients[211].

Outside of association studies, functional studies in erythroid progenitors or mice identified novel HbF regulators. These include Kruppel-like factor (*KLF1*), Pokemon *ZBTB7A*, RNA-binding protein *LIN28B*, heme-regulated inhibitor *HRI* (also known as *EIF2AK1*), and zinc-finger protein (ZNF) 410, *ZNF410* [212-216].

Although the prevalence of SCD is elevated in African patients, the population continues to be underrepresented. The genetic diversity in African-ancestry individuals could provide unique insights into the therapeutic target and better predictive tools for clinical outcomes. We, therefore, decided to pursue a meta-analysis of fetal hemoglobin in African SCD (N=3,740) patients and healthy Sardinian patients.

METHODS

Study Participants

The study aimed to identify and characterize common and rare DNA polymorphisms associated with fetal hemoglobin (HbF) levels. This meta-analysis comprised 9,338 participants: a detailed description of the participating cohorts is provided in **Supplementary Table 1**. The whole-exome sequence analysis contained 1,354 participants. Sample collections and procedures were in accordance with the committees' institutional and national ethical standards, and proper informed consent was obtained. The project was approved by the Montreal Heart Institute Ethics Committee (project #09-1137).

DNA genotyping and quality-control steps

Participants from the studies were genotyped on different genotyping arrays and at other locations. The Cooperative Study of Sickle Cell Disease (CSSCD), the Tanzania cohort and the SardiNIA study were described elsewhere [203,217]. In addition, DNA samples of participants from GEN-MOD, Mondor/Lyon, the Multicenter Study of Hydroxyurea in Sickle Cell Anemia (MSH), the Adult Sickle Cell Center at Georgia Health Sciences University (GHSU), and the Jamaica Sickle Cell Cohort Study (JSCCS) were genotyped on the Illumina Infinium HumanOmni2.5Exome-8v1.1 array at the Montreal Heart Institute Pharmacogenomics Center. We performed quality control using PLINK[218], removing SNPs with Hardy-Weinberg $P<1\times10^{-7}$ and genotyping rate <90%. After quality control, we imputed genotyped using reference AFR haplotypes from TOPMed Freeze5 GRCh38/hg38 and Minimac4 (v1.2.4) as implemented on the TOPMed imputation server[163]. For downstream analyses, we only considered variants with imputation quality Rsq >0.3.

Whole-exome DNA sequencing and quality-control steps

Study samples

We combined five cohorts (GEN-MOD, the Duke University Outcome Modifying Genes (OMG), the CSSCD, Differential Response to Hydroxyurea and Incidence of Stroke in Sickle Cell Disease (CIP, dbGaP Study Accession: phs000691.v2.p1), and Mondor/Lyon sickle cohort) sequenced at different time points and using different sequencing captures. We modelled our quality control steps after the Exome Aggregation Consortium (ExaC)[219].

Alignment and BAM processing, base quality recalibration, variant calling, and variant-quality score recalibration (VQSR) were performed using GATK best practices pipeline.
Sample quality control and selection utilized a common set of SNPs to ascertain ancestry concordance onto the 1000 Genomes dataset. Sample relatedness was calculated through kinship matrices in KING[220]. Variants were annotated with Variant Effect Predictor version 101. VEP's plugin (LOFTEE, SpliceAI, SIFT, Polyphen2, and MaxEnt)[169] were then used to predict deleteriousness. Finally, we retained high-quality variants as defined in EXAc's pipeline[219]. Any downstream analyses considered 985,119 high-quality variants.

Genetic association analyses (GWAS)

Within each cohort, we corrected fetal hemoglobin (HbF) levels for age, sex and β-globin genotypes (if appropriate in sickle cell disease (SCD) patients). We then normalized the residuals using inverse normal transformation to create HbF z-scores. HbF was measured in SCD patients not taking hydroxyurea. To condition the known HbF regulators (genetic variants at chr2-*BCL11A*, chr6-*HBS1L-MYB* and chr11-β-globin), we regressed out genotypes from HbF-associated variants at these loci from the HbF z-scores (using a stepwise approach with variants within a 1-Mb window centered on the strongest association signal). We calculated association statistics using linear regression (imputation dose, additive model) implemented in rvtests[166] and the first ten principal components as covariates. We used MR-MEGA to perform multi-ancestry meta-analyses[204]. To fine-map association results to calculate posterior inclusion probabilities (PIP) and create 95% credible sets, we used the approximate Bayes factor method as described previously[221].

Exome-wide association analyses (ExWAS)

Study population and phenotypes

The downstream analysis focused on SNPs with minor allele frequency (MAF) <1% (as calculated based on SCD individuals in this whole exome sequencing study) and excluded variants and samples with low genotyping rate (<95%). Additionally, SNPs deviating from Hardy-Weinberg equilibrium ($P<1x10^{-7}$) were not kept. Fetal hemoglobin percentage (% HbF) was measured using high-performance liquid chromatography across all cohorts. For the association test, %HbF levels were quantile normalized.

Single-variant association analysis

A total of 1,354 individuals of recent African ancestry and admixed African were included in association testing. We performed single-variant associations correcting for age, sex, the first ten principal components, the kinship matrix and the different sequencing captures. All analyses were performed using RVtests (v.20171009) [166]

Gene-level analysis

We employed three strategies to aggregate variants for gene-level association testing. First, for a given SNP, if at least one out of the seven algorithms (PolyPhen2 HumDiv and HumVar, LRT, MutationTaster, LOFTEE, SpliceAI, SIFT, and MaxEnt) predicted a variant as 'deleterious', we labelled the mask as broad. If all seven algorithms predicted the variant as 'deleterious,' we labelled the mask as 'strict'. The last strategy considered just loss-of-function variants as predicted by LOFTEE. Two statistical tests were considered for each mask: an adaptive burden test, which aggregates rare variants based on optimal frequency cut-off (VT), and SKAT, a bidirectional approach that includes SNPs with variable effect size and direction. Gene-level associations were conducted using rareMETALS_7.1[222]. rareMETALS outputs gene-level summary statistics including number of sites considered, top variant, p-value and the combined estimated effect size, more details are available on the tutorial website[223]. Only variants (MAF < 1%; allelic frequency is based on SCD individuals sequenced in this project) annotated as missense, nonsense, essential splice site, and frameshift indel were kept for gene-based analysis.

RESULTS

Genetic diversity among SCD participants

The SCD participants originated from seven studies based in the USA, Jamaica, France and Tanzania (**Supplementary Table 1**). To characterize the genetic diversity of these individuals, we combined their genotype information at 98,176 common SNPs with data from the 1000 Genomes Project. We performed dimension reduction analyses using principal component (PC) and uniform manifold approximation and projection (UMAP) analyses. As expected, the PC analysis revealed that most SCD participants in our experiment are aligned along an African-European axis of genetic variation on PC1 (**Fig. 1A**). The UMAP representation, generated using the first five PCs, provided more resolution and allowed us to identify participants from the East-African Tanzania SCD cohort, who overlap with the Luhya in Webuye (Kenya) population from the 1000 Genomes Project (**Fig. 1B-C**)[224]. We also identified a small number of SCD participants that cluster with South Asians or admixed Americans from the 1000 Genomes Project (**Fig. 1 B-C**).

Multi-ancestry meta-analysis for HbF levels

To identify novel genetic regulators of HbF levels, we combined association results at 24,271,278 variants ("averaged" minor allele frequency (MAF) ≥1%) from 5,903 European-ancestry individuals from the SardiNIA Study[203] and 3,740 SCD participants, mostly of African descent (**Supplementary Table 1**). Because of the genetic heterogeneity described above, we used PCs as covariates and opted to analyze each study individually. For the meta-analysis, we used MR-MEGA, which was developed to account for ancestry differences to maximize discovery power in GWAS[204]. We also focused our downstream analyses on association results conditioned on genotypes at the known HbF loci (*BCL11A*, *HBS1L-MYB*, β-globin; **Methods**) (**Figure 2**). This strategy identified three loci that reached genome-wide significance ($P \leq 5 \times 10^{-8}$): chromosome 10 in *BICC1*, on chromosome 19 near *KLF1* and chromosome 22 between *CECR2* and *SLC25A18* (**Table 1**). The SardiNIA study previously reported an association between a rare intronic variant in *NFIX* on chromosome 19 (rs183437571) and HbF levels[203]. This variant is not associated with HbF levels in SCD patients (MAF=0.0052, P=0.66), and is also not in linkage disequilibrium

with the sentinel variant (rs4804210) that we identified in our trans-ethnic HbF meta-analysis (r^2 in European-ancestry populations = 0.027; r^2 in African-ancestry populations = 0).

To characterize these three loci, we calculated posterior inclusion probabilities and generated 95% credible sets, thus fine-mapping the *BICC1*, *KLF1* and *CECR2/SLC25A18* loci to six, 29 and three variants, respectively (**Supplementary Table 2**). We annotated these variants using whole-blood expression quantitative trait loci (eQTL) results, open chromatin data and previous results from blood-cell traits GWAS (**Supplementary Table 2**). On chromosome 10, three of the *BICC1* variants were eQTL for *TFAM*, a gene located ~115-kb upstream, whereas a variant found in the intergenic sequence between *CECR2* and *SLC25A18* on chromosome 22 was an eQTL for *ATP6V1E1* and *BCL2L13*. The variants on chromosome 19 were whole-blood eQTLs for at least ten genes, including *DNASE2* and *FARSA*, but notably not the erythroid transcriptional regulator *KLF1* (**Supplementary Table 2**). None of the 38 fine-mapped variants overlapped DNAse1 hypersensitive sites previously characterized in fetal and adult human erythroblasts[225]. The variants near *BICC1* and *CECR2/SLC25A18* have not previously been implicated in GWAS for hematological traits. However, the HbF-associated variant at the *KLF1* locus was previously identified in trans-ethnic meta-analyses of several RBC traits (mean cell volume, mean cell hemoglobin, mean cell hemoglobin concentration, RBC count, RBC distribution width)[221].

Whole-exome DNA sequencing (WES) in SCD participants

To determine if rare coding variants modulate HbF levels, we analyzed available WES data from 1,354 SCD participants (**Methods**). We focused on variants with a MAF ≤1% and performed gene-based testing using two methods (VT and SKAT) and three variant selection strategies (broad, strict, and loss-of-function (LoF) as defined in **Methods**). Across these analyses, we found no genes that reached statistical significance after accounting for the number of genes tested ($\alpha=3.1\text{x}10^{-6}$, Bonferroni correction for 15,913 genes) (**Supplementary Figure 2** and **Supplementary Table 4**). Based on transcriptomic or proteomic experiments, we restricted our analyses to genes expressed in human erythroblasts or mature erythrocytes. However, we did not detect any of statistical signal's enrichment (**Supplementary Figure 3-5**)[127,226,227]. This negative result suggested limited statistical power, given our sample size.

Rare mutations can cause the hereditary persistence of fetal hemoglobin (HPFH) condition and have been identified at the β-globin locus as well as in the *KLF1* and *DNMT1* genes[211,228,229]. Therefore, we inspected rare (MAF ≤1%) coding variants in 83 genes that have been implicated in the γ- to β-globin gene switch during development (**Figure 3** and **Supplementary Table 3**) [210,230] We found 22 variants in 16 genes that are carried by SCD patients with mean HbF levels that are at least two standard deviations away from the mean HbF after correction for age, sex and β-globin genotypes (**Supplementary Table 3**).

Given the recent report of HPFH mutations in *DNMT1*[211] , we were particularly intrigued by the discovery of a novel missense variant in this gene (p.Gly95Ser) in an SCD participant with baseline HbF levels of 28.9% (2.36 standard deviations above the mean). Our review of the patient's medical records indicated that the high HbF levels were stable (additional HbF values of 27.7% and 28.8% taken three years apart), that the patient was not treated with hydroxyurea and that the patient was largely clinically asymptomatic. Furthermore, the high HbF levels phenotype was not explained by the inheritance of common alleles at *BCL11A*, *HBS1L-MYB* and *β-globin* associated with HbF levels (the patient's normalized HbF polygenic score was 0.65 standard deviation below the mean). These observations are consistent with the clinical benefits associated with the inheritance of a rare HPFH mutation.

DISCUSSION

This genome-wide association analysis identified three novel genetic loci associated with fetal hemoglobin levels. While not confirmed through replication, to our knowledge, this is the second largest genome-wide association study of HbF. Our findings confirm the genetic association of *BCL11A*, *HBSL1-MYB*, and *HBB* as primary loci of fetal hemoglobin across populations. Our discovery of the new loci rs4433524 in *BICC1,* rs4804210 in *KLF1*, and rs116175381 in *CECR2* remains to be replicated. While the transcription factor, Krüppel-like factor 1, is a well-established gene in the sickle cell disease literature[214,229], this is the first report of the gene in a GWAS.

We identified a missense singleton in *DNMT1* (G95S) predicted to be deleterious by both SIFT and Polyphen in an otherwise healthy individual with high baseline levels of fetal hemoglobin. However, little is known about this association between the *DNMT1* variant and sickle cell disease traits (pain crises, stroke). Further functional characterization in $CD34^+$ from patients and engineered HuDEP-2 mutant cells is required to demonstrate its role in γ-globin expression and thus an intracellular increase in HbF.

While the high presence of SCD individuals was an important strength of this study, our sample size for both GWAS and exome-wide study considerably limits our ability to identify new loci associated with HbF. Second, the lack of an independent cohort with fetal hemoglobin levels and matching ancestry backgrounds leads to caution when interpreting associations in this study. Third, the added diversity leads to ancestry-specific associations, thus making it difficult to pinpoint the causal variants. Finally, although our conditional analysis aimed at eliminating the association signal from the *BCL11A*, *HBSL1-MYB* and *HBB* to boost the signal from other regions of the genome, we observed residual associations after conditioning on *HBB*.

We present three novel loci associated with fetal hemoglobin and show that a rare missense mutation in *DNMT1* could result in an HPFH mutation. These results highlight the need for larger datasets with more SCD individuals. Furthermore, therapeutic intervention in those genes could increase the amount of fetal hemoglobin while simultaneously reducing the amount of sickle hemoglobin in their blood and potentially alleviating the condition.

URLs

Supplementary Table 3

Supplementary Table 4

FUNDING STATEMENT

This work was funded by the Canadian Institutes of Health Research (PJT #156248), the Canada Research Chair Program, DDCF, and Biogen/Sanofi.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank all participants who contributed data to this study.

Table 1. Novel genome-wide significant association results for fetal hemoglobin (HbF) levels. We combined HbF association results across seven sickle cell disease (SCD) studies (African-ancestry) and the SardiNIA study (European-ancestry) using MR-MEGA. We also provide the conditional fixed-effect HbF association results per ancestral group. Association results are conditional on genotypes at the known HbF loci (BCL11A, HBS1L-MYB, β-globin) (**Methods**). The variants' coordinates are on build GRCh38. N, sample size; EA, effect allele; Beta (SE), effect size and standard error in standard deviation units; NA, not available; 1000G_Eur, European-ancestry individuals from the 1000 Genomes Project.

Variant	Trans-ethnic meta-analysis (N=9,643)	SCD-only (N=3,740)			SardiNIA (N=5,903)			Note
	Conditional P-value	EA frequency (EA)	Beta (SE)	Conditional P-value	EA frequency (EA)	Beta (SE)	Conditional P-value	
10_58728559_G_A	6.17×10^{-9}	0.38 (A)	0.0805 (0.0193)	3.05×10^{-5}	0.86 (A)	0.09189 (0.0204)	6.77×10^{-6}	rs4433524, *BICC1*
19_12879166_A_G	4.50×10^{-8}	0.54 (A)	-0.0627 (0.019)	9.89×10^{-4}	0.77 (A)	-0.08175 (0.01703)	1.63×10^{-6}	rs4804210, *DNASE2*, *KLF1*
22_17559810_C_T	3.98×10^{-8}	0.054 (T)	-0.2192 (0.0416)	1.36×10^{-7}	NA			rs116175381, *CECR2*, *SLC25A18* (monomorphic in 1000G_Eur)

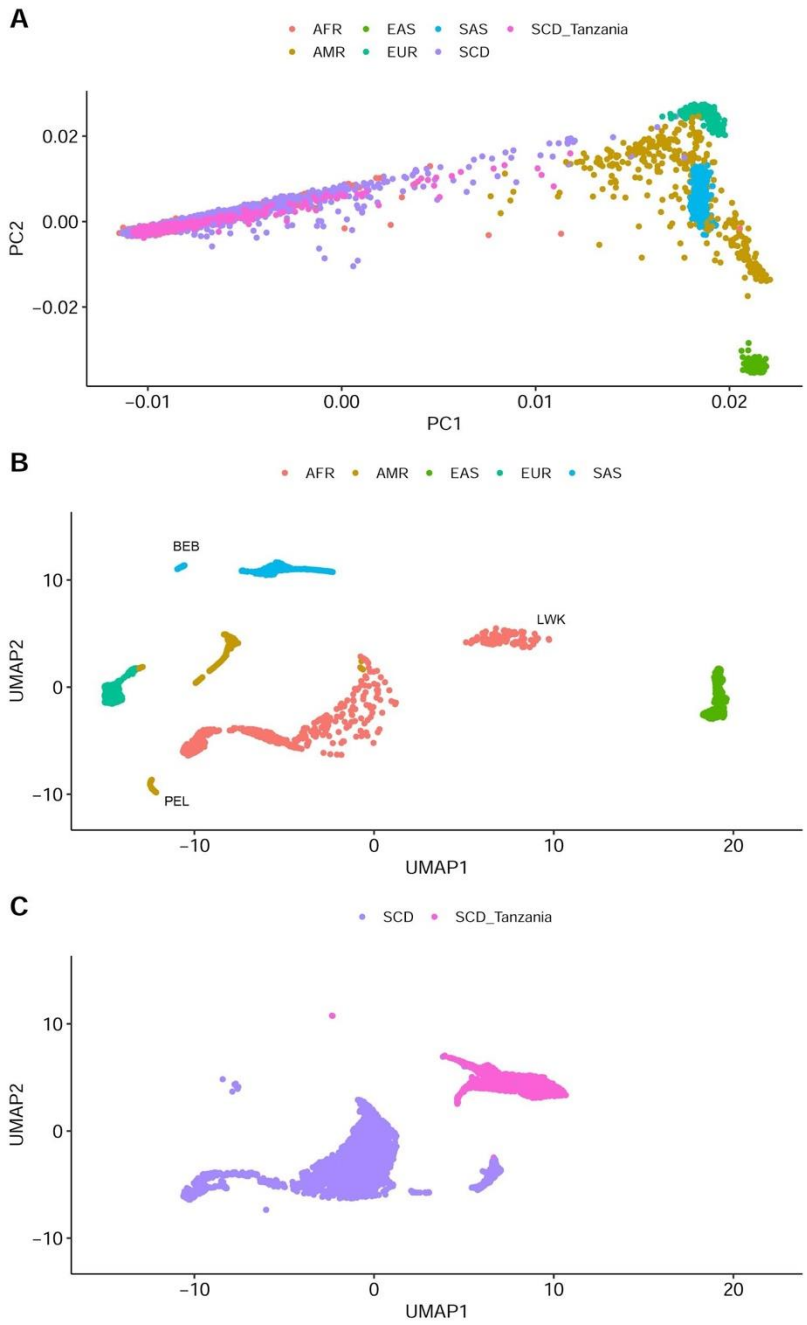


Figure 15. Study design GWAS. (**A**) Map of principal component (PC) analysis of 98,176 common variants in a merged dataset that includes all individuals from the 1000 Genomes Project and 3,740 SCD participants. (**B-C**) We applied uniform manifold approximation and projection (UMAP) on the first five PCs calculated on the merged dataset. To simplify visualization and interpretation, we present individuals from the 1000 Genomes Project in (**B)** and SCD participants in **C**. AFR, African ancestry; AMR, admixed American; EAS, East-Asian ancestry; EUR, European ancestry; SAS, South-Asian ancestry; BEB, Bengali from Bangladesh; PEL, Peruvians from Lima; LWK, Luhya in Webuye (Kenya).

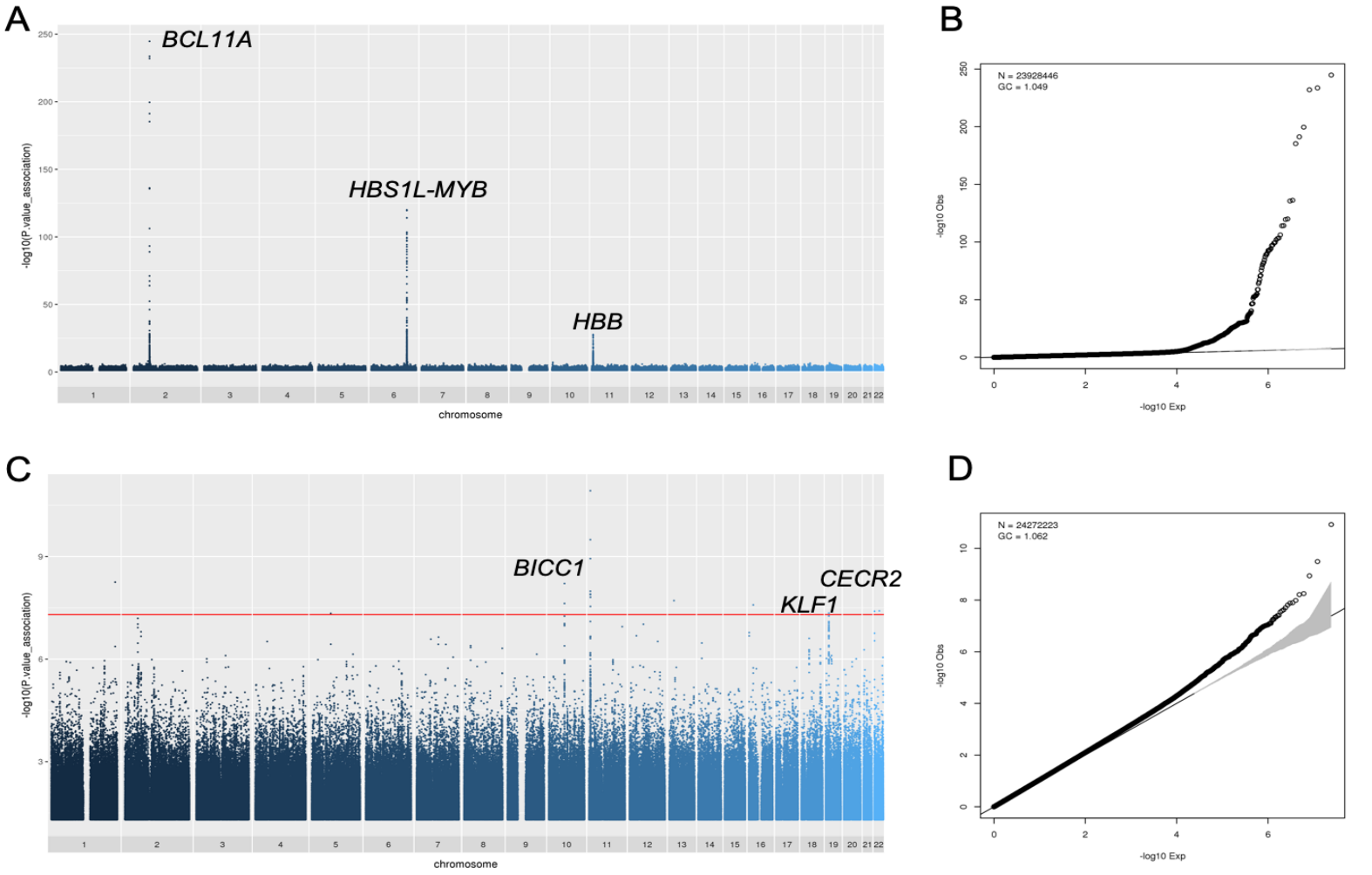


Figure 16. Trans-ethnic genome-wide association studies (GWAS) for fetal hemoglobin levels (HbF) in 5,903 Sardinians and 3,740 African-ancestry sickle cell disease (SCD) patients. (**A**) Manhattan and (**B**) QQ plots for the non-conditional HbF trans-ethnic meta-analysis highlight the known associations at the BCL11A, HBS1L-MYB and HBB loci. (**C**) Manhattan and (**D**) QQ plots for the HbF trans-ethnic meta-analysis conditioned on associated variants at BCL11A, HBS1L-MYB and HBB. We found genome-wide association results ($P<5x10^{-8}$, minor allele frequency (MAF) ≥0.01) on chromosomes 10 (BICC1), 19 (KLF1) and 22 (CECR2). Our conditional analyses did not remove all association signals at the HBB locus on chromosome 11. Further, we also found rare variants (MAF <0.01) with $P <5x10^{-8}$ (horizontal red line in **C**), but we did not investigate them further because of lower imputation quality. N is the number of tested variants in B and D, and GC is the genomic inflation factor.

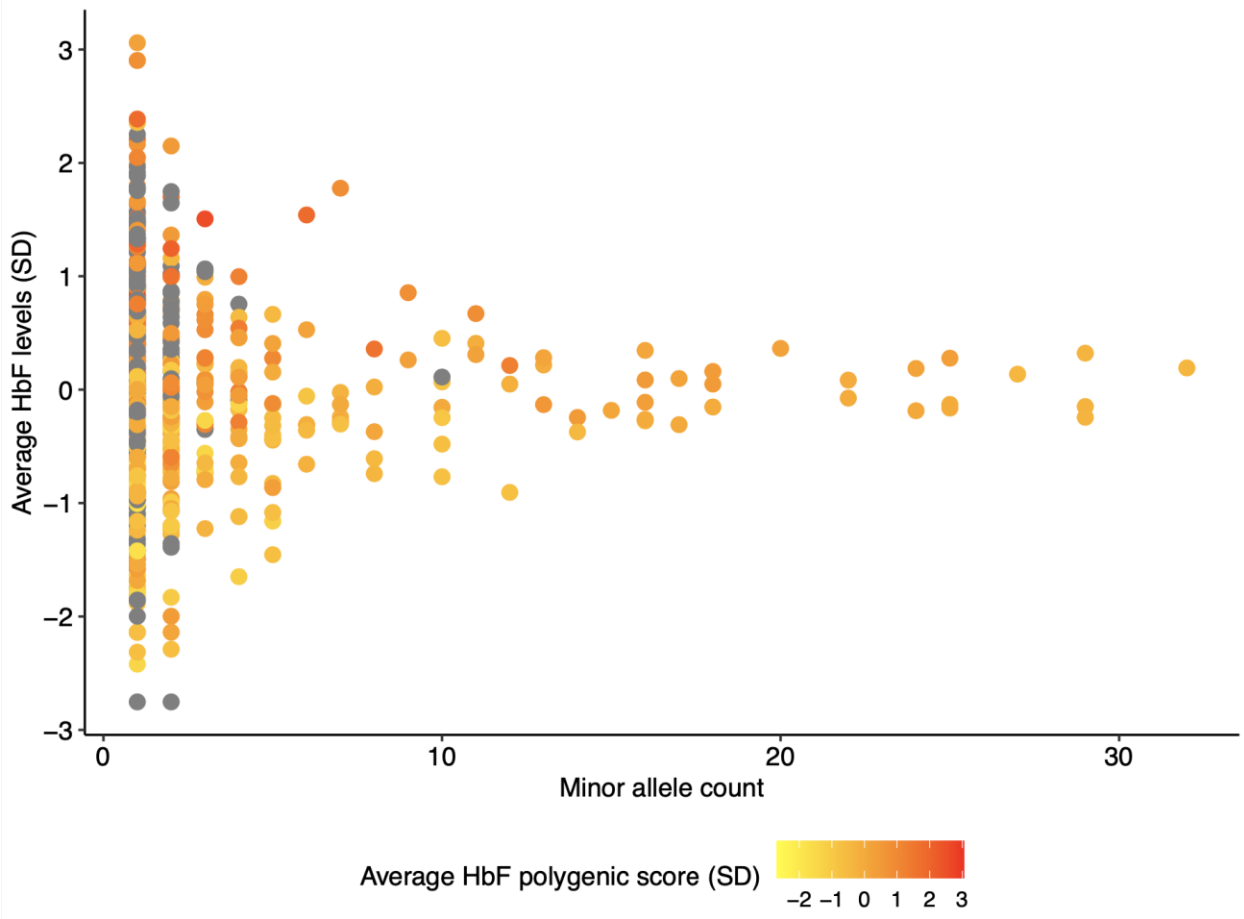


Figure 17. Rare coding variants identified in 1,354 sickle cell disease (SCD) patients by whole-exome sequencing in genes implicated in the γ-to-β globin switch. We only consider missense, nonsense, frameshift and essential splice site variants with a minor allele frequency <1% (corresponding to a minor allele count ≤32 (x-axis)). Average fetal hemoglobin (HbF) levels per variant are on the y-axis (in standard deviation units after correction for sex, age and β-globin genotypes). For each variant found in SCD patients with available genome-wide genotyping data, we averaged the normalized HbF polygenic score (in standard deviation units) calculated using known HbF common variants at BCL11A, HBS1L-MYB and β-globin (grey indicates missing genotyping data as not all sequenced individuals were also genotyped).

Chapter 3. Exome- and genome-wide association studies of red blood cell density in sickle cell disease patients

The following article is intended to be submitted to the journal, *British Journal of Hematology*. In this article to identify novel genetic regulators of dense red blood levels (DRBC). I performed the association tests of imputed genotypes of DRBC in 581 SCD patients. I then annotated the results and identified which ones were the most promising and needed to be replicated. Then I used whole-exome sequencing to identify rare coding variants regulating DRBC. I performed all the quality control procedures, the gene-based analysis with sequence kernel association test (SKAT) and variable threshold (VT) to detect associations of coding variants combined together.

Exome- and genome-wide association studies of red blood cell density in sickle cell disease patients

Yann Ilboudo[1,2], Pablo Bartolucci[3], Mélissa Beaudoin[2], Carlo Brugnara[6], Frederic Galactéros[3], Guillaume Lettre[1,2]

Affiliations

[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

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

Correspondence

Guillaume Lettre

Montreal Heart Institute

5000 Bélanger St

Montreal, Quebec, Canada, H1T 1C8

514-376-3330 ext. 2660

guillaume.lettre@umontreal.ca

ABSTRACT

Sickle cell disease (SCD) patients with a large fraction of dense and dehydrated red blood cells (DRBC) tend to have a severe clinical course**.** Previous genome-wide association studies and whole-exome sequencing in SCD patients highlighted two susceptibility loci (*PIEZO1* and *ATP2B4*) for DRBC. Yet, a large proportion of missing heritability remains to be investigated. This study sought to identify genetic variation associated with DRBC in SCD patients. We proposed a novel strategy for generating candidate genes susceptible to control red blood cell hydration and increase our statistical power. Exome-wide association analysis and a genome-wide association study was performed for upwards of 581 SCD patients. Likely causal variants were prioritized for replication. Additionally, we performed gene-based analyses to investigate the aggregate effect of multiple variants within the same genes or pathways. Although no signal reached the Bonferroni correction threshold, rs1228690182 in *SPTB* and rs372784283 in *SEC23B* showed the potential to modulate DRBC. The gene-based analysis pinpointed four drug targets, *BCL6*, *LRRC32, KNCJ14,* and *LETM1,* as potential regulators of DRBC. Our study identified several candidate genes and drug targets associated with DRBC, which has a role in erythrocyte osmotic regulation. Future studies are needed to validate these findings further and explore these genes and pathways as potential therapeutic targets in SCD.

INTRODUCTION

Sickle cell disease (SCD) describes a set of inherited disorders caused by a point mutation at the β-globin locus[1,2]. The disease touches millions of individuals, mainly in sub-Saharan Africa and southern Asia[3,4] where malaria is still endemic. Sickle cell anemia (SCA) is the most prevalent form of the disease, making up to about 3 in 4 cases of individuals of African ancestry. SCA originated from homozygosity of β^S allele (rs334), which replaces a hydrophilic glutamic amino acid for a hydrophobic valine at the 6th position of the β-globin chain[5,6]. Co-inheritance of β^S and other mutations result in other forms of SCD (e.g., SC, SβThal^0, SβThal$^+$)[1].

The disease is named after the 'sickle' shape the red blood cell assumes after deoxygenation. This is because the intracellular hemoglobin polymers become rigid and form rod-like structures which contort erythrocytes. Red cell dehydration exacerbates the rate of polymerization, which in turn increases intracellular hemoglobin concentration[7]. The polymerization rate negatively correlates to fetal hemoglobin level (HbF), which acts as an inhibitor and substitute to HbS[8,9]. One of the critical attributes of sickle cell disease is the existence of dense, dehydrated red blood cell density (DRBC). Previous studies linked a higher percentage of red blood cell density with end of organ complications and vaso-occlusive events[10].

Water, positively and negatively charged solutes are constantly tuned by the activity and interplay of erythrocyte transporters and channel systems[11,12]. This is illustrated by the sheer number of diseases associated with dysregulation of erythrocyte ion transport. In sickle cell disease, a major focus is applied on K-Cl cotransport (*KCC1, KCC2, KCC3, KCC4*) and the calcium-activated (Gardos) potassium channel (*KCNN4*). While no drugs reached the final phases of the clinical trial process, Gardos channel blockers such as charybdotoxin, clotrimazole, and senicapoc showed promising results.[13-15] Although identifying commercially viable drugs to block the K-Cl channel failed, magnesium ion and antioxidant N-acetylcysteine were identified as an inhibitor of the potassium chloride co-transport and dehydration.[16,17]

This study employed human genetics in more than 500 SCD patients to identify promising variants and drug targets associated with dense dehydrated red blood cells (DRBC). The objectives of the current study were: (i) to examine rare coding variants associated with DRBC identified in African ancestry individuals; (ii) to investigate the aggregate effect of multiple variants within the same genes; (iii) to explore the role of DRBC-scored rare coding DNA

sequences in SCD patients; and (iv) to integrate multiple omics datasets to prioritize common SNP associated with DRBC.

METHODS

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.

Samples and DNA genotyping

The GEN-MOD study, a cohort of sickle cell disease (SCD) patients recruited in Paris, France, has been described elsewhere[10]. 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[18] 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[19-21]. Briefly, the CSSCD is a multi-center research study conducted in the United States between 1978 and 1988 to investigate factors contributing to SCD complications. In total, enroll over 3,000 participants including children and adults who were followed for a 10-year window while receiving regular clinical tests, and laboratory tests. Genome-wide genotype data generated with the Illumina Human610-Quad array was available for 1,279 CSSCD participants. After quality control, we imputed genotyped using reference AFR haplotypes from TOPMed Freeze5 GRCh38/hg38 and Minimac4 (v1.2.4) as implemented on the TOPMed imputation server[22]. We restricted association testing to markers with an imputation r^2 >0.3.

Whole-exome summary statistics

The present study uses summary statistics data from the UK Biobank analysis of rare coding variants to hematological traits using gene collapsing data (https://app.genebass.org/). Karczewski and colleagues conducted a study involving 394,841 exomes from the UK Biobank to determine gene-based associations[231]. In their analysis, they performed a total of 75,767 group tests across 4,529 phenotypes. These tests included gene-based burden (mean), SKAT (variance), and SKAT-O (hybrid variance/mean) tests for various types of variants such as predicted loss of function (pLoF) variants, missense variants (including low-confidence pLoF variants and in-frame insertions or deletions [indels]), synonymous variants, and a combination of pLoF or missense variants (not displayed in https://app.genebass.org)[231]. More detailed information about the databases used can be found in the publications.

Whole-exome DNA sequencing and quality-control steps

GEN-MOD and Mondor/Lyon sickle cohorts were the two cohorts out of 5 SCD cohorts with DRBC values. Red blood cell density measures were performed using the phthalate distribution technique[10]. For both cohorts, we used Nimblegen SeqCap EZ Exome Capture and Nimblegen SeqCap Nimblegen MedExome kits to capture exons, and we sequenced DNA using the Illumina HiSeq4000 and the Illumina NovaSeq 6000 instruments with a paired-ends 2x100 base pairs protocol. We modeled our quality control steps after the Exome Aggregation Consortium (ExaC) using GATK (v4)[23].

Alignment and BAM processing

The paired-end sequence reads from exomes were aligned to the human genome reference (GRCh37p13/hg19) using bwa (v0.7.17) (BWA MEM)[24]. The default parameters were employed to generate alignment in the SAM format given paired-end reads. Examples of the commands are below:

```
bwa aln ref.fa short_read.fq > aln_sa.sai

bwa samse ref.fa aln_sa.sai short_read.fq > aln-se.sam
```

Base quality recalibration

The base quality scores were then recalibrated using GATK BaseRecalibrator and a list of known variant sites from dbSNP. The sequenced interval came from GENCODE. The new base quality scores were then applied using GATK ApplyBQSR but retaining the original base quality scores within the BAM.

Variant calling

Using GATK HaplotypeCaller, the recalibrated BAM file from the previous step was used to perform variant calling per sample. In addition, the output is in GVCF mode, which can be used for joint genotyping with multiple samples.

Variant-quality score recalibration (VQSR)

VQSR (GATK ApplyVQSR) is then applied, and the raw VCFs from the previous step are filtered to achieve a high degree of sensitivity and reduce false positives. The SNP VQSR model is trained using HapMap3.3 and 1KG Omni 2.5 SNP sites, and a 99.6% sensitivity threshold was applied to filter variants. Recalibration of insertions/deletions sites used Mills et *al*. 1KG gold standard and Axiom Exome Plus sites with a 95.0 % sensitivity threshold[25].

Sample quality control and selection

A common set of 7,000 SNPs were selected for principal component analysis. Samples with outlier heterozygosity were removed before principal component analysis (PCA). Individuals were clustered based on their ancestry's population groups as defined on 1000 Genomes dataset. They are color-coded based on their ancestry's populations. SCD sample relatedness was calculated using kinship matrices implemented in KING [26].

Variant annotation

We employed Variant Effect Predictor version 101 to annotate variants. We retrieved several protein prediction consequences info using VEP's plugin (LOFTEE, SpliceAI, SIFT, Polyphen2, and MaxEnt)[27] We queried Ensembl/GENCODE and RefSeq transcripts databases and restricted results to produce the most severe consequence per variant. Variants mapping to coding regions were kept for downstream analysis**.**

High Quality (HQ) variants

Variant sites were labeled as high-quality if they met the following criteria: (1) they were given a PASS filter status by VQSR, (2) at least 80% of the individuals in the dataset had a depth (DP) ≥10, and genotype quality (GQ) ≥20, (3) at least one individual was carrying the alternate allele with depth ≥10, and GQ≥20, and (4) the variant was not located in the ten 1-kb regions of the genome with the highest levels of multi-allelic variation. Once we applied the variant filtering criteria, 985,119 variants were left.

Statistical analyses

Single variant analysis

We performed a single-variant test by using linear regression as implemented in RVTESTS[28]. Continuous DRBC values were inverse normal transformed and corrected for age, sex, cohorts, sequencing batches, and the first ten principal components. We then identified the most promising variants ($P < 1 \times 10^{-4}$) and defined the statistical significance using Bonferroni correction and the set of exome-wide association significance threshold 1.6×10^{-7} (0.05/315,128) for single-variant analysis. Quantile–quantile and Manhattan plots were generated using R (V4.5.1, R Development Core Team).

Gene-based analysis

We used sequence kernel association test (SKAT)[29] and variable threshold (VT)[30] for our gene-based testing using rareMETALS(v.6.3)[31]. We focused our analysis on variants with minor allele frequency (MAF) <1% for gene-based testing. We ran three sets of gene-based analyses based on deleteriousness prediction. The broad set included variants predicted to be damaging by at

least one of the predictions PolyPhen2 HumDiv[32] and HumVar, LRT[33], MutationTaster[34], LOFTEE[35], SpliceAI[36], SIFT[37], and MaxEnt[38]. The strict set included variants predicted to be deleterious by all of the above algorithms. Finally, we focused on the predicted loss of function (pLoF). All genetic association analyses presented in this study were adjusted for the ten first principal components. We included all genes for which one or more variants were present. Bonferroni correction was employed to define the significance threshold for gene-based analysis [P_{broad} = 1.7 x 10^{-6}, 0.05/(14,949 genes x 2 tests); P_{strict}=9.3 x 10^{-6},0.05/(2,684 genes x 2 tests)].

Genome-wide association and functional prioritization of genetic variants

DRBC values were inverse normal transformed and adjusted for age, sex, and the first ten principal components when performing the linear regression using RVTESTS[28] in each cohort (**Descriptive cohort demographic Table1**). We then employed METAL[39] with the sample size scheme (the scheme uses p-value and direction of effect, weighted according to sample size) to combine summary statistics across cohorts.
Given the lack of genome-wide significant results, we sought to identify variants with relevant regulatory effects. We, therefore, explore annotations of variants with epigenomics features, gene expression, and proxy phenotypes to identify potential causal loci we would like to use for replication.

We also prioritized variants that map to enhancers validated via CRISPRi-FlowFISH[42] in K562 cells keeping the enhancer-gene pair with an ABC score >= 0.022. Finally, we queried the eQTLGen database[43] and GTEx[44] whole blood datasets to retrieve 339,493 cis-eQTL in 1,694 candidate genes. These genes were pre-selected based on seminal studies and supposed roles in erythrocyte hydration. **Supplementary Table 1** provides the inclusion criteria for these candidate genes, while the data table (**URLs)** lists the genes.

Qualitative analysis

Variant Scoring and correlation (WESvsc) to DRBC.

We used qualitative analysis to link the variants to DRBC. We queried all protein-truncating variants from 2,333 candidate genes, thus resulting in 23,189 variants. We calculated the average

raw DRBC and the inverse normal transformed DRBC across individuals carrying a given mutation for each variant. Finally, we kept all annotation information provided by the deleteriousness prediction software (i.e. Sift, polyphen, etc..), frequency across ancestry in ExAC, gnomAD, 1000 Genomes (1000G), position and type of the amino acid substitutions, and information of previously known phenotypes or PubMed publication (**Supplementary Table 3**). We then leveraged *Crosstalk*[45], an interactive graphics application, to dynamically interact and filter the scatter plot and the table on the variant scoring (**Figure 2.2**). This framework also allowed us to generate a browser-based interaction that can be shared in a typical HTML R Markdown output **(**See URLs : Candidate gene Crosstalk HTML visualisation). Our *WESvsc* allows for dynamical filtering on raw and normalized DBRC values. Additionally, the variant consequence (frameshift_variant, missense_variant, protein_altering_variant, splice_acceptor_variant, splice_donor_variant, start_lost) as provided from the VEP annotation can be selected. The user can select a specific gene or a set of genes to look at and the variant annotation (e.i., variant annotated as being involved in spherocytosis, MCHC GWAS, or malaria). Finally, the interactive visualization provides a table with anonymized sample ID, deleteriousness prediction provided by SIFT and Polyphen, and the variant's allelic frequency. R packages, Plotly[46] and d3scatter[47] generated the dynamic scatter plot, while DT[48] generated the dynamic table.

RESULTS

Whole-exome sequencing

Single variant analysis

In the single variant analysis, no SNP reached exome-wide significance levels. The quantile-quantile plots did not reveal a deviation from null when comparing the observed and expected P-values distribution. By annotating the 36 most significant variants with variants associated with MCHC in Chen, M. H. *et al*. (2020), an intronic variant rs201639174 at *ITFG3* stood out from the list since it's nominally significant (*P*=0.005) only in African ancestry individuals. *ITFG3,* also known as *FAM234A,* encodes the family with sequence similarity 234 member A is promising since variants within the gene were associated with various RBC traits phenotype, namely MCH, MCV, RBC count, Hemoglobin levels in healthy Africans[49,50], Europeans[51]. Additionally, the inframe deletion, rs775700353 at *PRDM2*, the intronic variants rs256412, rs371180559, and rs138514497 at *TRIO*, *KATNA1*, and *ACACB,* respectively, are expressed in the erythroid lineage[227], and erythroblast[127] and represent enzymatic drug targets originating from expert-curated pharmacological and medicinal chemistry literature, IUPHAR[232] (**Supplementary Table 3.2**).

Insights from gene-level analyses

We performed three sets of gene-based tests aggregating deleterious missense or splicing variants with MAF < 0.01. Unfortunately, no genes reached exome-wide significant results for any of the tested schemes. In **Table 3,** the most significant genes across all collapsing schemes using a lenient significance threshold of $P < 1 \times 10^{-3}$. Additionally, we cataloged all nominally significant associations ($P < 0.05$) in **Supplementary Table 3.3**.

All 11 of the 14 most associated genes are expressed in either erythroblast, erythroid lineage, red blood cell proteome, or a combination. The quantile-quantile plots from **Figure 3.1, C** (SKAT test) show an enrichment of drug target families labeled by IUPHAR as "other proteins". These enriched genes include *BCL6* ($P_{SKAT} = 3.3 \times 10^{-3}$; P_{VT}=0.051), *LRRC32* ($P_{SKAT} = 4.7 \times 10^{-3}$), encoding the B-cell lymphoma 6, and the Leucine-Rich Repeat Containing 32 respectively. In Figure 1, D (VT test) we four genes departing from the null: *LRRC32* ($P_{VT} = 5.8 \times 10^{-4}$), *KNCJ14*

(P_{SKAT}=0.02; P_{VT}=2.3 x 10^{-3}*), LETM1 (P_{SKAT} =4.5 x 10^{-3};P_{VT}=*9,7x10^{-4}*). KNCJ14*, which encodes a potassium inwardly-rectifying channel subfamily J member 14, is a voltage-gated ion channel. LETM1 is a mitochondrial calcium ion transmembrane transporter that encodes the leucine zipper and EF-hand containing transmembrane protein 1.

Qualitative analysis

Variant Scoring and correlation (WESvsc) to DRBC

Our variant scoring approach prioritized 23,189 nonsynonymous variants. Using the interactive visualization, we restricted variants with DRBC z-score > 2, we prioritized 227 variants in 203 genes (**Supplementary Table 3.4**). We also looked at variants with DRBC z-score < -1.5, we prioritized 1,769 variants in 1,018 genes. Since most of these variants are singleton, we queried top pLoF gene-based association results ($P < 2.5 \times 10^{-6}$) from the UKBiobank exome association statistics with mean cell hemoglobin concentration (MCHC), mean cell volume (MCV), hemoglobin concentration, red cell distribution width, and mean sphered cell volume (MSCV) published on AstraZeneca PheWAS Portal[52], and GeneBass[53].

We found a missense variant (rs372784283) carried by two individuals. The SNP maps to the SEC23 Homolog B (*SEC23B*) and is annotated as a spherocytic gene in OMIM. In the UK Biobank, the pLOF (P_{SKAT-O}= 1.73 x 10^{-11} ; P_{Burden}= 7.9 x 10^{-11}) is associated with mean sphered cell volume, and P_{DRBC} = 5.3 x 10^{-3}. Additionally, we found a missense in spectrin B (*SPTB*) rs1228690182; while not much is known about the variant, it is predicted to be deleterious by both SIFT and Polyphen. It is carried by one individual, lookup in the UK Biobank found that pLOF (P_{SKAT-O}= 1.04 x 10^{-12}; P_{Burden}=1.48 x 10^{-13}) is associated with mean sphered cell volume and P_{DRBC} = 0.042. *SPTB* is a gene linked to several diseases related to red blood cell hydration (pyropoikilocytosis, elliptocytosis, and spherocytosis). Another missense variant (rs144259338) carried by a single individual with a DRBC value of 50 (the largest DRBC value in the cohort) maps to the gene encoding ATP Binding Cassette Subfamily A Member 7 (*ABCA7*). The variant is associated with DRBC (P-value = 2.4 x 10^{-3}). The pLOF (P_{SKAT-O}= 5.81 x 10^{-7} ; P_{Burden}=2.4 x 10^{-6}) is associated with mean sphered cell volume; the gene is annotated by the expert database, IUPHAR[232], as a pharmacological classified as a transporter. Finally, a missense variant (rs767272040) mapping to the erythrocyte membrane protein band 4.1, *EPB41I,* is carried by an individual with a DRBC value of 32. The gene is a drug target belonging to the transporter family,

is well documented in the literature on red blood hydration, and is involved in Mendelian disorders such as pyropoikilocytosis, and elliptocytosis, spherocytosis. pLOF (P_{SKAT-O}= 9.18 x 10^{-15}; P_{Burden}=4.45 x 10^{-15}) is also associated with mean sphered cell volume and the P_{DRBC} = 0.01.

Looking at variant in individuals with low, dense red blood cells (z-score DRBC < -1.5), we found a missense variant (rs142161945) predicted to be deleterious by SIFT and Polyphen in the solute carrier family 4, anion exchanger, member 1 (Erythrocyte Membrane Protein Band 3, Diego Blood Group). *SLC4A1* is a well-characterized protein with roles in malaria, stomatocytosis, spherocytosis, pseudohyperkalemia, and xerocytosis and is a drug target annotated as a transporter. The pLOF (P_{SKAT-O}= 4.0 x 10^{-16}; P_{Burden}= 4.26 x 10^{-16}) is associated with MCHC and mean sphered cell volume. Another interesting finding is a variant mapping to the regulator of G protein signaling 11 (*RSG11*). The gene is found to be associated with a large MCHC GWAS[41] and annotated as a drug target in IUPHAR. The pLOF (P_{SKAT-O}= 2.66 x 10^{-11}; P_{Burden}4.35x 10^{-12}) is associated with MCHC, the DRBC GWAS is nominally associated with the variant. The individual carrying the mutation has a raw DRBC value of 0. Additionally, we found singletons in well-known erythrocyte hydration and morphology genes such as *PIEZO1*, *CD36,* and *EPB42*, with pLOF strongly associated with mean cell hemoglobin concentration and red blood cell width, mean sphered cell volume, respectively (**Supplementary Table 4**). Additionally, lesser-known genes implicated in erythrocyte volume regulation or morphology with pLOF < 2.5 x 10^{-6} included *MOK*, *SLC25A37*, *PTPRH, MTOR, CHEK2,* and *ACVRL1* associated with red blood cell width, mean cell volume, Mean sphered cell volume, mean cell volume, mean cell volume, and hemoglobin concentration respectively.

Insight from rs59446030-PIEZO1 inframe deletion

PIEZO1 is a mechanosensitive ion channel that functions as an undiscriminating cation channel in many tissues. Empirical evidence demonstrated the protein's role in sensing blood flow through vasculature[54] and red blood cell volume control[55]. Previous publications showed that rare gain-of-function (GOF) mutations[56] are involved in hereditary xerocytosis. In contrast, the common GOF-rs59446030[57-59] did not associate with hematological parameters or weakly influenced red cell hydration. We revisited the association of the GOF with DRBC in this cohort

of 581 individuals. We found no association between the allele and red blood cell density (*P-value* = 0.17).

Genetic association analysis of red blood cell density

After quality-control and genotype imputation, we performed a genome-wide association study (GWAS) between 34,766,316 million DNA sequence variants and red blood cell density (DRBC) in 374 sickle cell disease (SCD) patients from the GEN-MOD cohort and 199 SCD patients from Mondor-Lyon cohort (**Table 1**). **Table 5** presents results for loci and associated variants with $P_{DRBC} < 5 \times 10^{-7}$.

As part of the GWAS workflow, performing external replication with an independent cohort with matched ancestry and the same phenotype is crucial in validating the results. However, to our knowledge, no other SCD cohort fits the criteria previously described. Therefore, we decided to combine additional data points for a given SNP to ascertain its causal role in a given gene.

We implemented three strategies to increase the probability of finding robust genetic associations with DRBC. First, we considered variants mapping to erythroleukemia enhancers, defined by activity by contact (ABC) of enhancer-promoter[42]. Among the 36,346 cis-regulatory elements (CRE) tested, one variant mapping to four CREs (*NPM3*, *MGEA5*, *KCNIP2,* and *HPS6*) was more strongly associated ($P = 1.47 \times 10^{-6}$) with DRBC than would be expected by chance (**Figure 3**). Second, we retrieved 339,493 cis expression quantitative trait loci (eQTL) from 1,694 candidate genes selected because they encode proteins with direct or indirect effects on red blood cell hydration (**Supplementary Table 3.1**). 13 eQTL variants were associated with DRBC after Bonferroni correction $P=2.1 \times 10^{-5}$ (0.05/2,333 candidate genes) (**Table 5**). Additionally, while below the significance threshold, three eQTL SNPs from GTEx departed from the null hypothesis (**Figure 2.3**). rs16889330 in a mitochondrial transporter, *MTCH1* ($P= 3.1 \times 10^{-5}$), rs11772895 maps to a gene encoding the erythropoietin-producing hepatoma (Eph) receptor 1 (*EPHA1) P=* 4.0×10^{-5}), and rs12153855 in the immunophilin *FKBPL* ($P= 4.1 \times 10^{-4}$) (**Supplementary Table 4**). While promising because of the role in various pathophysiology of SCD (rheumatoid arthritis,

basophil count, or blood proteins)[51,60], these variants remained to be replicated in an independent SCD GWAS with matched ethnicity.

Our final strategy to prioritize variants was to exploit the physiological link between DRBC and MCHC. Mean cell hemoglobin concentration, MCHC in GEN-MOD (Pearson's r=0.62, P=9.4x10^{-41}), in Mondor-Lyon (Pearson's r=0.44, P=2.6 x 10^{-11}), in combined cohort (Pearson's r=0.57, P=3.7 x 10^{-51}) (**Supplementary Figure 5**) is correlated with DRBC. While the Pearson's r coefficient shows that we don't have a perfect positive linear relationship between MCHC and DRBC, we expect variants associated with RBC dehydration to increase MCHC. Additionally, hemoglobin concentration is a well-established factor influencing sickle hemoglobin (HbS) polymerization[40].

DISCUSSION

Analyzing DRBC potential drug targets highlights the auspicious opportunity and obstacles of putting exome-sequencing datasets in translational research. We confirmed the relationship between DRBC and MCHC (**Supplementary Figure 5**). We report the first exome-wide association analysis of DRBC in SCD patients. While rare single variants, gene-level associations, candidate genes didn't yield statistically significant results (**Figure 2.1**), we provide evidence that analyzing singletons, and nominally significant variant ($P < 0.05$) is a clinically relevant approach. Our analysis enhances the results of the published DRBC GWAS study. Finally, we highlight the interconnection between genetically associated SNPs with DRBC and those with MCHC in African individuals.

Phenotypic correlation accounting for age and sex for DRBC showed that the strongest positive correlation amongst blood parameters analyzed is with MCHC (*Pearson* $r_{\text{Henri-Mondor}}$=0.44, *Pearson* $r_{\text{Henri-Mondor}}$=0.62, *Pearson* r_{Combined}=0.57). We also note a negative and weak correlation between DRBC and fetal hemoglobin level (HbF) (*Pearson* $r_{\text{Henri-Mondor}}$=-0.21, *Pearson* $r_{\text{Henri-Mondor}}$=-0.16, *Pearson* r_{Combined}=-0.12). This shows the shared etiology between DRBC and MCHC and supports the rationale that therapeutic intervention for both DRBC and HbF could be complementary and synergistic for SCD patients. One limitation of using MCHC in SCD patients is that because of their anemic state, their MCHC value can be artificially elevated due to red

blood cell agglutination or opacification of the plasma or the presence of concomitance of alpha-thalassemia.

Because no marker reached the significance threshold for the exome study or the genome-wide association study, and we don't have replication samples, we employed candidate gene, enhancer, expression quantitative loci, MCHC proxy, qualitative analysis to prioritize further which SNPs warrant downstream or functional analysis. We highlight promising variants and gene-level associations for replication. Our qualitative research highlighted singletons that exacerbate DRBC levels. The first one is a missense mutation in SEC23 Homolog B (*SEC23B*), a mendelian disease gene known to impact red blood cell morphology. Predicted loss-of-function in large exomic datasets is strongly associated with mean sphered volume. The other missense DNA sequence, maps to a well characterize and known gene, spectrin B (*SPTB*) rs1228690182. The gene is thought to be a severity modifier in SCD[63], it responsible for the erythrocyte cytoskeletal stability, several mutation in the gene have been linked to spherocytosis, hereditary elliptocytosis, and neonatal hemolytic anemia[64]. The patient carrying this mutation has the largest DRBC value (50) recorded in the cohort. We found additional singleton in *PIEZO1*, *CD36*, and *EPB42* with strong pLoF associations thus providing strong grounds for functional studies follow-up. Finally, we showed that the gain-of-function mutation in *PIEZO1* did not association with DRBC ($P<$ 0.05). We suspect that the association we found was overestimated due to the winner's curse. Because the original discovery came from a smaller sample size, the effect of the association was overestimated, thus any attempt to replicate the results in a larger cohort failed to find the significant effect. While one major limitation for our study is our sample size, adding more sample correct the estimation.

A recent meta-analysis carried out on 15,171 participants of African ancestry identified 952 DNA sequence variants significantly associated with MCHC[41]. This query showed the presence of enrichment of MCHC variants, this is highlighted by an inflation of the test statistics and a departure from the null of all the SNPs (λ_{GC}=2.2, **Figure 3**). In fact, amongst these MCHC variants, we found an eQTL variant in eQTLGen for the post-glycosylphosphatidylinositol attachment to proteins 6 (*TMEM8A/ PGAP6)*, P_{DRBC}=1.6 x 10^{-5}, the same variant also associates with self-reported pleurisy in the UKBB. Finally, although all these associations require replication to be validated, one association seems likely causal as orthogonal data offer potential mechanistic insights. rs7634650 (P_{DRBC}=1.3 x 10^{-3}) maps to the downstream gene of the long intergenic non-protein coding RNA 885 (*LINC00885*). The DNA sequence variant maps to a cis-

regulatory (CRE) region as identified by Fulco *et al*.[233] which regulates *LINC00885*. The CRE, chr3:196,138,394-196,139,414 (hg38), that interacts with *LINC00885 (*ABC score: 0.0441). Additionally, previous studies identified rs7634650 to be associated with MCHC[61], and with MCV, MCH, and RBC[62] in non-anemic Europeans.

Consistent with our previous GWAS[65] on DRBC, 14 out of 25 SNPs previously identified remained significant. The direction of effect was different for 4 out of the 25 SNPs, and 7 could neither be genotyped nor imputed (they were considered missing). Surprisingly the intronic variant at *ATP2B4* (rs10751450) did not remain significant although it is strongly associated with MCHC in non-anemic Europeans ($P=5 \times 10^{-59}$)[62]. Exome-wide and genome-wide association studies exploring the role of MCHC include well over 100,000 individuals[221,231]. Therefore, our analysis with a little less than 600 samples is limited. Systematic collection of DRBC measurements during a regular patient visit or a hospital stay would enable much larger studies in SCD patients to be performed. Our results highlight the phenotypic and genetic overlap between DRBC and MCHC. We find genes associated with the structure of RBC cytoskeleton, with the osmotic regulation of intracellular ion and water content and with the cell surface-to-volume ratio. We extend previous reports on the role of red cell hydration and sickle cell disease as a therapeutic avenue to ameliorate SCD complications. Future studies should dissect on a functional level the associations between DRBC, and the variants highlighted in this study. Such functional studies could involve a pooled CRISPR-based perturbation, followed by single-cell RNA sequencing and cellular phenotyping of DRBC levels, or proxy phenotype such as MCHC.

URLs

- Expressed candidate genes
- Candidate gene Crosstalk HTML visualisation
-DRBC GWAS summary statistics

ACKNOWLEDGMENTS

We thank all participants for their contribution to this project. GL is funded by Biogen, the Canadian Institutes of Health Research (CIHR, MOP #123382), the Doris Duke Charitable Foundation, and the Canada Research Chair program. SLA is funded by the Doris Duke

Charitable Foundation. MT is funded by CIHR/Canadian Blood Services (MOP #3251163). Foundation.

AUTHOR CONTRIBUTIONS

YI and GL conceived and designed the experiments; YI performed the experiments; MB contributed DNA samples, clinical information, and expert knowledge; YI and GL analyzed the results; YI and GL wrote the manuscript with contributions from all authors.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

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

Phenotype	GEN-MOD (N=379)	Mondor-Lyon (N=202)
Males/females	185 / 223	74 / 128
Age, years	30 ± 9	36 ± 11
DRBC, %	12 ± 8.6	9 ± 8.4

Table 2. Top single variant association results with red blood cell density (DRBC) in 581 participants from GEN-MOD+Mondor-Lyon. We included in this table variants with P_{DRBC} <1x10^{-4} (**Methods**). Chr:Pos, genomic coordinates on build hg38; 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. GWAS results for MCHC association for individuals from African ancestry and all ancestry were retrieved from Chen, M. H., *et al*. (2020).[41]

rsID	CHROM: POS (hg38)	CHROM: POS:REF/ALT	AF	Beta(SE)	PVALUE	Consequence	SYMBOL	MCHC association(Chen *et al*_AfrAncestry)	MCHC association(Chen *et al*._AllAncestry)
-	chr16:20769977	chr16:20781299:C/T	0.006	1.99(0.411)	1.28E-06	intron_variant	ACSM3	NA	NA
-	chr16:20769975	chr16:20781297:G/T	0.006	1.982(0.411)	1.45E-06	intron_variant	ACSM3	NA	NA
rs9872688	chr3:169793568	chr3:169511356:G/C	0.19	1.107(0.241)	4.45E-06		LRRC34		
rs1042503	chr12:102852922	chr12:103246700:C/T	0.039	0.813(0.178)	4.76E-06	synonymous_variant	PAH	REF/ALT_AF=C/T_0.96; Beta(SE)=0.035 (0.033) ;Pval=0.285	REF/ALT_AF=C/T_0.626; Beta(SE)=0.012 (0.0163) ;Pval=0.465
rs151149890	chr17:64858810	chr17:62854928:G/A	0.057	0.549(0.122)	6.60E-06	missense_variant	LRRC37A3	REF/ALT_AF=G/A_0.952; Beta(SE)=-0.026 (0.033) ;Pval=0.442	NA
-	chr16:20769972	chr16:20781294:C/T	0.005	2.026(0.459)	1.01E-05	intron_variant	ACSM3	NA	NA
rs752049651	chr15:61967465	chr15:62259664:G/A	0.008	1.544(0.35)	1.05E-05	intron_variant	VPS13C	REF/ALT_AF=I/D_0.028; Beta(SE)=0.038 (0.056) ;Pval=0.498	NA
-	chr16:20769971	chr16:20781293:G/T	0.005	2.022(0.459)	1.06E-05	intron_variant	ACSM3	NA	NA
rs146746669	chr16:259446	chr16:309445:TG/T	0.123	-0.387(0.088)	1.09E-05	intron_variant	ITFG3	REF/ALT_AF=D/I_0.09; Beta(SE)=-0.04 (0.032) ;Pval=0.208	NA
rs34885736	chr4:173318553	chr4:174239704:C/T	0.024	0.768(0.175)	1.18E-05	synonymous_variant	GALNT7	REF/ALT_AF=C/T_0.975; Beta(SE)=0.001 (0.041) ;Pval=0.985	REF/ALT_AF=C/T_0.932; Beta(SE)=-0.0200 (0.0309) ;Pval=0.517
rs775700353	chr1:13778599	chr1:14105094:AGAG/A	0.004	1.989(0.459)	1.49E-05	inframe_deletion	PRDM2	NA	NA
rs148452011	chr1:208038505	chr1:208211850:G/A	0.011	1.334(0.31)	1.68E-05	intron_variant	PLXNA2	REF/ALT_AF=G/A_0.982; Beta(SE)=0.063 (0.052) ;Pval=0.222	REF/ALT_AF=A/G_0.0333; Beta(SE)=0.0152 (0.0447) ;Pval=0.733
rs34594998	chr17:5361621	chr17:5264916:C/T	0.238	0.275(0.064)	1.69E-05	synonymous_variant	RABEP1	REF/ALT_AF=C/T_0.771; Beta(SE)=-0.015 (0.015) ;Pval=0.306	REF/ALT_AF=T/C_0.0540; Beta(SE)=-0.00967 (0.0347) ;Pval=0.780
rs114826587	chr17:2041888	Chr17:1945182:G/A	0.184	0.292(0.068)	1.76E-05	intron_variant	DPH1	REF/ALT_AF=A/G_0.806; Beta(SE)= 0.010 (0.17);Pval=0.530	REF/ALT_AF=G/A_0 0.035; Beta(SE)= -0.0984 (0.045);Pval= 0.0292

rs201639174	chr16:259448	chr16:309447:G/C	0.124	-0.372(0.088)	2.18E-05	intron_variant	ITFG3	REF/ALT_AF=G/C_0.912; Beta(SE)=0.069 (0.025) ;Pval=0.005	REF/ALT_AF=C/G_0.000514; Beta(SE)=-0.106 (0.423) ;Pval=0.802
rs1277048565	chr4:68478849	chr4:69344567:G/T	0.01	-1.511(0.356)	2.19E-05	missense_variant	TMPRSS11E	NA	NA
rs1436127	chr12:95898506	chr12:96292284:C/A	0.121	0.357(0.084)	2.29E-05	intron_variant	CCDC38	REF/ALT_AF=C/A_0.893; Beta(SE)=-0.038 (0.021) ;Pval=0.063	REF/ALT_AF=C/A_0.957; Beta(SE)=0.0399 (0.0384) ;Pval=0.299
rs1218671775	chr4:68478848	chr4:69344566:G/T	0.01	-1.506(0.356)	2.32E-05	splice_acceptor_variant	TMPRSS11E	NA	NA
rs774787329	chr6:21595985	chr6:21596216:CG/C	0.008	1.178(0.279)	2.49E-05	3_prime_UTR	SOX4		
rs11868032	chr17:5380322	chr17:5283617:A/T	0.503	-0.228(0.054)	2.53E-05	intron_variant	RABEP1	REF/ALT_AF=A/T_0.506; Beta(SE)=-0.002 (0.013) ;Pval=0.847	REF/ALT_AF=A/T_0.375; Beta(SE)=0.00312 (0.0160) ;Pval=0.845
rs1272660186	chr4:68478842	chr4:69344560:C/T	0.01	-1.498(0.356)	2.58E-05	intron_variant	TMPRSS11E	NA	NA
rs764040709	chr12:40924544	chr12:41318346:C/T	0.002	2.231(0.531)	2.68E-05	intron_variant	CNTN1	NA	NA
rs752035515	chr4:68478847	chr4:69344565:A/T	0.01	-1.493(0.356)	2.73E-05	splice_acceptor_variant	TMPRSS11E	NA	NA
rs34489008	chr1:168183911	chr1:168153149:T/C	0.009	-1.28(0.309)	3.44E-05	synonymous_variant	TIPRL	REF/ALT_AF=T/C_0.995; Beta(SE)=-0.01 (0.104) ;Pval=0.921	NA
rs146436884	chr4:143528067	chr4:144449220:G/A	0.044	0.656(0.159)	3.60E-05	intron_variant	SMARCA5	REF/ALT_AF=G/A_0.959; Beta(SE)=0.017 (0.033) ;Pval=0.602	REF/ALT_AF=A/G_0.000470; Beta(SE)=0.0349 (0.562) ;Pval=0.950
rs371180559	chr6:149638574	chr6:149959710:G/A	0.003	1.858(0.457)	4.70E-05	intron_variant	KATNA1	NA	NA
rs61400567	chr10:29481553	chr10:29770482:C/T	0.014	0.955(0.235)	4.78E-05	intron_variant	SVIL	REF/ALT_AF=C/T_0.989; Beta(SE)=0.104 (0.064) ;Pval=0.108	NA
rs376602370	chr13:38850079	chr13:39424216:G/A	0.001	2.153(0.532)	5.14E-05	missense_variant	FREM2	NA	NA
rs200434209	chr1:168066489	chr1:168035727:A/C	0.012	-1.129(0.28)	5.43E-05	intron_variant	DCAF6	REF/ALT_AF=A/C_0.994; Beta(SE)=0.001 (0.088) ;Pval=0.989	NA
rs467960	chr21:41440964	chr21:42812891:C/T	0.363	-0.246(0.061)	5.63E-05	synonymous_variant	MX1	REF/ALT_AF=C/T_0.662; Beta(SE)=0.005 (0.013) ;Pval=0.684	REF/ALT_AF=C/T_0.627077; Beta(SE)=0.00073318 (0.0163065) ;Pval=0.964
rs17304212	chr2:178461008	chr2:179325735:C/G	0.038	0.61(0.152)	5.94E-05	missense_variant	DFNB59	REF/ALT_AF=C/G_0.962; Beta(SE)=0.053 (0.065) ;Pval=0.416	REF/ALT_AF=C/G_0.945; Beta(SE)=-0.0195 (0.0341) ;Pval=0.568
rs115309023	chr2:37049612	chr2:37276755:T/C	0.016	0.892(0.222)	6.15E-05	intron_variant	HEATR5B	REF/ALT_AF=T/C_0.987; Beta(SE)=-0.026 (0.058) ;Pval=0.655	NA
rs138514497	chr12:109264395	chr12:109702200:C/T	0.011	0.96(0.24)	6.22E-05	intron_variant	ACACB	REF/ALT_AF=C/T_0.989; Beta(SE)=0.01 (0.072) ;Pval=0.888	NA
rs11761299	chr7:122304473	chr7:121944527:T/C	0.616	0.219(0.056)	8.00E-05	5_prime_UTR_variant	FEZF1	REF/ALT_AF=T/C_0.405; Beta(SE)=-0.025 (0.014) ;Pval=0.071	REF/ALT_AF=T/C_0.391; Beta(SE)=-0.00678 (0.0168) ;Pval=0.686

rs1806265	chr17:5416724	chr17:5320044:T/A	0.515	-0.227(0.058)	8.02E-05	intron_variant	NUP88	REF/ALT_AF=T/A_0.504; Beta(SE)=-0.003 (0.013) ;Pval=0.791	REF/ALT_AF=T/A_0.373; Beta(SE)=0.000212 (0.0160856) ;Pval=0.989
rs116224881	chr1:58473955	chr1:58939627:T/G	0.011	1.012(0.257)	8.45E-05	missense_variant	OMA1	REF/ALT_AF=T/G_0.981; Beta(SE)=0.0004 (0.046) ;Pval=0.992	NA
rs62000369	chr9:88544477	chr9:91159392:G/A	0.066	-0.42(0.107)	9.05E-05	missense_variant	NXNL2	REF/ALT_AF=G/A_0.926; Beta(SE)=-0.014 (0.026) ;Pval=0.595	NA
rs74427615	chr11:102205880	chr11:102076611:C/T	0.011	-1.374(0.352)	9.60E-05	intron_variant	YAP1	NA	NA
rs256412	chr5:14394174	chr5:14394283:G/A	0.413	0.244(0.063)	9.74E-05	intron_variant	TRIO	REF/ALT_AF=G/A_0.575; Beta(SE)=0.006 (0.013) ;Pval=0.623	REF/ALT_AF=G/A_0.709; Beta(SE)=-0.00874 (0.0173) ;Pval=0.613

Table 3. 14 promising genes implicated by gene-based testing. These genes meet our criteria for statistical significance: (1) gene-based $P<1\times10^{-3}$ (a sub exome-wide threshold). For each gene, we provide P-values for the four different gene-based tests applied. Abbreviations: Strict: mask for which all prediction algorithms agree on variant deleteriousness. Broad: mask for which at one of the prediction algorithm predicts the variant to be deleterious. N. var SKAT/VT: Number of variants included in the gene-based test when performing SKAT test/VT test. Pvalue SKAT/VT: P-value for SKAT/VT test. Top SNP pval: P-value for the most significant variant within the gene. Top_SNP_hg38: chromosome, base pair (build hg38), reference allele, effect allele, and allele frequency for the most significant SNP in the gene-based test. Notes: additional information about the gene.

	Strict				Broad				
Symbol	**N. var SKAT/VT**	**Pvalue SKAT/VT**	**Top_SNP_hg38**	**Top_SNP_pval**	**N. var SKAT/VT**	**Pvalue SKAT/VT**	**Top_SNP_pval**	**Top_SNP_hg38**	**Notes**
CTNNAL1	2/2	0.16/0.16	9:108972757_T/G_0.00250713	0.05	10/10	7.7×10^{-4}/8.5×10^{-4}	5.0×10^{-4}	9:106210476_G/T_0.00507176	Expressed in erythroid lineage[66] and erythroblast[67]
NCAPG2	1/1	0.017/0.017	7:158680085_C/T_0.00187807	0.02	12/12	$0.020/6.8 \times 10^{-5}$	0.017	7:158887394_C/T_0.00187807	Expressed in erythroid lineage[66], erythroblast[67], and red blood cell[68]
APP	11/	3.5×10^{-4}/3.5×10^{-4}	21:26090072_C/T_0.00172732	3.4×10^{-4}	6/5	$0.030/5.4 \times 10^{-4}$	3.4×10^{-4}	21:24717758_C/T_0.00172732	Expressed in erythroid lineage[66], erythroblast[67], and red blood cell[68]
TMPRSS2	1/1	0.012/0.012	21:41479275_A/C_0.000870913	0.012	5/3	$0.22/6.7 \times 10^{-4}$	0.012	21:40107348_A/C_0.000870913	
SDR39U1	2/2	3.02×10^{-3}/4.8×10^{-4}	14:24440870_A/T_0.000860585	2.6×10^{-3}	10/9	$0.016/1.8 \times 10^{-4}$	2.6×10^{-3}	14:23971661_A/T_0.000860585	Expressed in erythroid lineage[66], erythroblast[67], and red blood cell[68]
TLR10	1/1	7.6×10^{-4}/7.6×10^{-4}	4:38773384_T/C_0.00349402	7.5×10^{-4}	12/5	0.14/0.71	7.5×10^{-4}	4:38771763_T/C_0.00349402	Gene expressed in erythroid lineage[66] and is a catalytic drug receptor[69]
IKBKAP	1/1	0.87/0.87	9:108874894_C/A_0.000823381	0.087	18/16	7.3×10^{-4}/0.28	5.1×10^{-4}	9:108929786_T/C_0.00505629	Expressed in erythroid lineage[66], erythroblast[67], and red blood cell[68]
CEP89	NA	NA	NA	NA	6/4	$0.064/8.3 \times 10^{-4}$	2.4×10^{-3}	19:32933522_T/C_0.000862088	Expressed in erythroid lineage[66] and erythroblast[67]
AP5S1	NA	NA	NA	NA	8/6	$0.027/9.3 \times 10^{-4}$	5.6×10^{-3}	20:3823925_G/T_0.00172419	Expressed in erythroid lineage[66] and erythroblast[67]
RSPO3	NA	NA	NA	NA	1/1	5.5×10^{-4}/5.5×10^{-4}	5.5×10^{-4}	6:127155351_C/T_0.00181695	

C6orf58	NA	NA	NA	NA	2/1	5.1×10^{-4}/0.041	9.6×10^{-4}	6:127590263_C/G_0.00127644	
NRXN2	NA	NA	NA	NA	6/4	0.03/2.8×10^{-4}	2.7×10^{-3}	11:64667255_C/T_0.000860585	Expressed in erythroid lineage[66]
LETM1	NA	NA	NA	NA	2/2	4.5×10^{-3}/9.8×10^{-4}	0.013	4:1849198_C/A_0.00270105	Expressed in erythroid lineage[66], erythroblast[67], red blood cell[68], and drug transporter[69]
LRRC32	NA	NA	NA	NA	7/7	4.7×10^{-3}/5.8×10^{-4}	8.1×10^{-3}	11:76659650_C/G_0.00281724	Expressed in erythroid lineage[66], erythroblast[67] and other drug protein[69]

Table 4. Association between the common PIEZO1 deletion allele and red blood cell density in sickle cell disease patients. Association was corrected for age, sex, and the first ten principal components. The direction of the effect is given for the functional PIEZO1 deletion allele (rs572934641; E756del): BETA and SE (SE) for RBC density.

CHROM	POS	REF	ALT	AF	ALT_EFFSIZE	P-VALUE
16	88800372	GTCC	G	0.215927	0.106	0.17

Table 5. Top genome-wide association results of red blood cell density in 573 sickle cell disease individuals. Meta-analysis results were generated with METAL; cohort-level summary statistics came from RVTEST. Inverse-normal transformed DRBC was regressed for age, sex, and each cohort's first ten principal components. The table includes variants with $P_{DRBC} < 5 \times 10^{-7}$, in addition to $P_{DRBC\_eQTL\_candidate\_gene} < 2.1 \times 10^{-5}$.Abbreviations: CHR_BP_A1_A2; chromosome_basepair_allele1_allele2, AF; allele frequency for A1, Zscore: the magnitude and the direction of effect, P; pvalue, Beta; effect size, HM; Henri-Mondor cohort. Coordinates on build hg38.

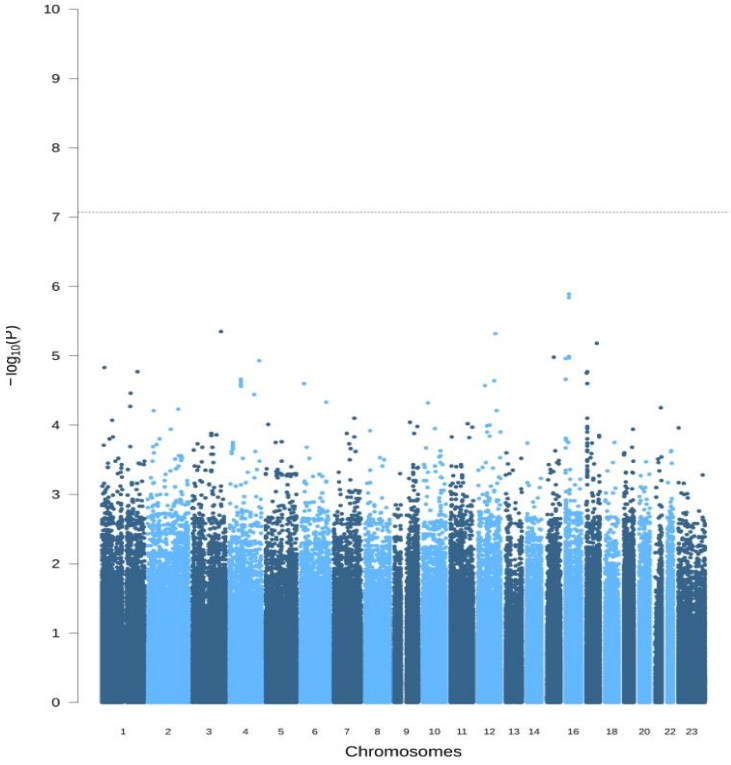
	Meta-analysis			HM			GENMOD						
CHR_BP_A1_A2	**AF**	**Zscore**	**P**	**AF**	**Beta**	**P**	**AF**	**Beta**	**P**	**Consequences**	**SYMBOL**	**rsID**	**Notes**
chr5_124385732_A_T	0.58	5.3	1.5×10^{-7}	0.6	0.26	0.011	0.57	0.35	3.3×10^{-6}	intron_variant,non_coding_transcript_variant	LINC01170	rs11241738	
chr17_73362505_T_A	0.99	-5.2	1.7×10^{-7}	0.01	1.32	0.0033	0.01	1.30	1.51×10^{-5}	intron_variant_intron_variant,non_coding_transcript_variant	SDK2	rs139516980	
chr5_124383129_T_A	0.42	-5.2	2.0×10^{-7}	0.6	0.26	0.010	0.57	0.35	4.6 x 10-6	intron_variant,non_coding_transcript_variant	LINC01170	rs7737676	
chr5_124383128_C_T	0.42	-5.2	2.0×10^{-7}	0.6	0.26	0.010	0.57	0.35	4.6×10^{-6}	intron_variant,non_coding_transcript_variant	LINC01170	rs67407299	
chr10_101655396_T_G	0.11	-5.2	2.3×10^{-7}	0.12	-0.62	4.4×10^{-5}	0.10	-0.42	6.3×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs73351221	
chr18_66306564_T_G	0.44	5.1	3.0×10^{-7}	0.48	0.25	0.014	0.42	0.35	5.4×10^{-6}	intergenic_variant	NA	rs7245198	
chr10_101659949_C_A	0.89	5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs28565478	
chr10_101664450_G_A	0.89	5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs56970368	
chr10_101693207_T_A	0.89	5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4_NA	rs12240593	
chr10_101672663_G_C	0.89	5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant_downstream_gene_variant	FBXW4	rs7095907	

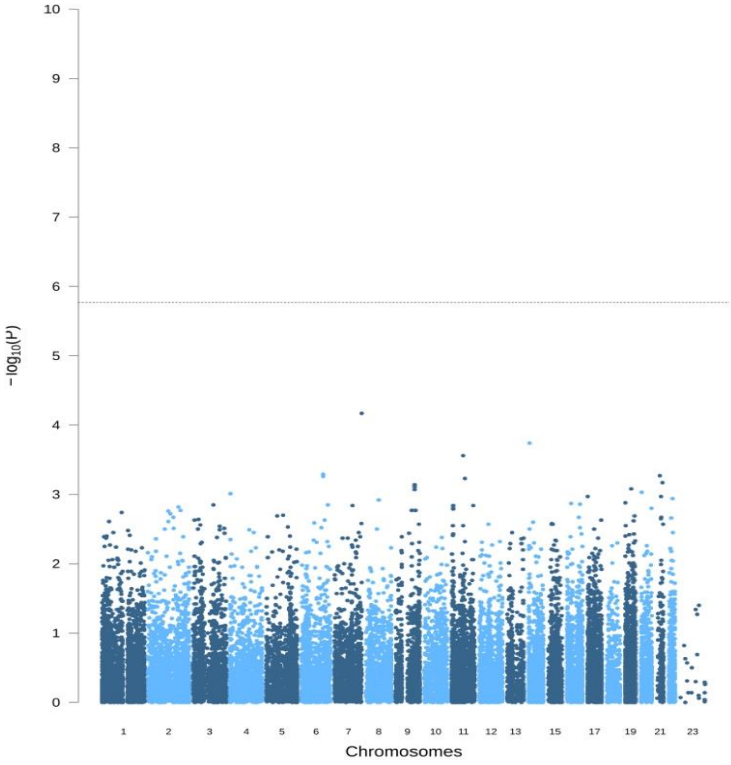
chr10_101691139_G_A	0.89	5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4_NA	rs12249989	
chr10_101683534_G_A	0.89	5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs12263004	
chr10_101686330_T_C	0.11	-5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs73351280	
chr10_101682838_GA_G	0.89	5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	NA	
chr10_101692238_A_G	0.11	-5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4_NA	rs60239868	
chr10_101666740_A_T	0.11	-5.1	3.1×10^{-7}	0.12	-0.63	3.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs73351242	
chr1_29541498_G_T	0.99	-5.2	3.4×10^{-7}	0.01	2.29	0.00095	0.01	1.58	9.6×10^{-5}	intergenic_variant	NA	rs191013203	
chr10_101766483_G_T	0.91	5.2	3.5×10^{-7}	0.11	-0.56	0.00040	0.09	-0.48	2.0×10^{-4}	upstream_gene_variant_intron_variant,non_coding_transcript_variant_downstream_gene_variant	LOC105378458_LOC105378457_FGF8	rs73338885	
chr2_217896096_G_A	0.94	5.1	3.6×10^{-7}	0.05	-0.96	7.5×10^{-5}	0.06	-0.53	6.5×10^{-4}	intron_variant_upstream_gene_variant_downstream_gene_variant	MIR6809_TNS1	rs115665349	
chr10_101653726_C_T	0.89	5.1	3.7×10^{-7}	0.12	-0.62	4.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs73351217	
chr10_101657095_C_T	0.89	5.1	3.7×10^{-7}	0.12	-0.62	4.4×10^{-5}	0.10	-0.42	9.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs12241580	
chr6_36957882_C_T	0.89	5.1	3.9×10^{-7}	0.12	-0.47	0.0023	0.10	-0.50	5×10^{-5}	intron_variant	PI16	rs76367788	
chr10_101764560_G_T	0.90	5.05	4.3×10^{-7}	0.11	-0.56	0.00040	0.09	-0.46	2.4×10^{-4}	upstream_gene_variant_intron_variant,non_coding_transcript_variant_downstream_gene_variant	LOC105378458_LOC105378457	rs12244840	
chr10_101937533_A_G	0.086	-5.0	4.6×10^{-7}	0.09	-0.52	0.0010	0.08	-0.52	1.2×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	ARMH3	rs1173759800	
chr10_101678011_A_G	0.11	-5.0	4.9×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	1.2×10^{-3}	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4	rs73351268	
chr10_101682560_T_C	0.11	-5.0	4.9×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	1.2×10^{-3}	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs7068187	
chr10_101678587_A_G	0.11	-5.0	5.0×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	1.2×10^{-3}	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4	rs386747080	

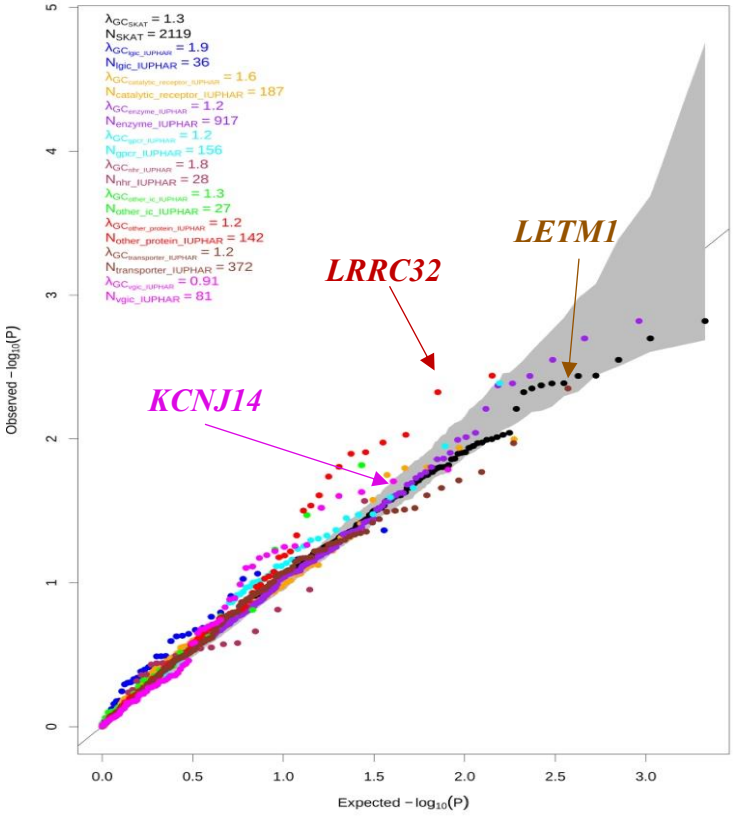
chr10_101677839_A_G	0.11	-5.0	5.0×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	1.2×10^{-3}	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4	rs73351267	
chr10_101678545_T_C	0.11	-5.0	5.0×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	0.0012	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4	rs73351271	
chr10_101688442_G_A	0.89	5.0	5.0×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	0.0011	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs73351283	
chr10_101678929_TATAGA_T	0.89	5.0	5.0×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	0.0011	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4	NA	
chr10_101693147_G_A	0.89	5.0	5.0×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	0.0011	intron_variant_intron_variant,non_coding_transcript_variant_upstream_gene_variant	FBXW4_NA	rs12262279	
chr10_101674286_A_G	0.11	-5.0	5.0×10^{-7}	0.12	-0.6	4.4×10^{-5}	0.10	-0.40	0.0011	intron_variant_intron_variant,non_coding_transcript_variant	FBXW4	rs112372206	
eQTL Genes													
chr17_74193463_C_A	0.99	4.58	4.7×10^{-6}	0.01	-1.15	0.05	0.01	-1.65	2.3×10^{-5}	intergenic_variant	NA	rs12603773	GPRC5C_eQTL_Gen
chr1_85342012_A_G	0.35	4.38	1.2×10^{-5}	0.4	0.22	0.03	0.33	0.3	1.5×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	DDAH1_NA	rs233072	DDAH1_eQTL_Gen
chr1_85339672_G_A	0.65	-4.38	1.2×10^{-5}	0.4	0.22	0.03	0.33	0.3	1.5×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	DDAH1_NA	rs233068	DDAH1_eQTL_Gen
chr1_85330874_G_T	0.65	-4.38	1.2×10^{-5}	0.4	0.22	0.03	0.33	0.3	1.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	DDAH1_NA	rs233133	DDAH1_eQTL_Gen
chr1_85333599_G_A	0.65	-4.38	1.2×10^{-5}	0.4	0.22	0.03	0.33	0.3	1.4×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	DDAH1_NA	rs233061	DDAH1_eQTL_Gen
chr1_85359790_G_C	0.65	-4.35	1.4×10^{-5}	0.4	0.22	0.03	0.33	0.29	1.7×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	DDAH1_NA	rs233097	DDAH1_eQTL_Gen
chr1_85340322_C_T	0.66	-4.33	1.5×10^{-5}	0.39	0.21	0.04	0.32	0.3	1.3×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	DDAH1_NA	rs233071	DDAH1_eQTL_Gen
chr8_141287645_G_A	0.93	-4.31	1.6×10^{-5}	0.08	0.35	0.06	0.07	0.56	7.0×10^{-5}	intron_variant_intron_variant,non_coding_transcript_variant	SLC45A4_NA	rs11781174	SLC45A4_eQTL_Gen
chr6_13749344_A_T	0.02	4.31	1.7×10^{-5}	0.02	0.84	0.02	0.01	1.16	2.9×10^{-4}	intergenic_variant	NA	rs1410919279	SIRT5_eQTL_Gen
chr6_13748210_T_C	0.02	4.31	1.7×10^{-5}	0.02	0.84	0.02	0.01	1.16	2.9×10^{-4}	intergenic_variant	NA	rs474233	SIRT5_eQTL_Gen

chr1_85357361_G_C	0.65	-4.29	1.8×10^{-5}	0.39	0.21	0.04	0.32	0.3	1.5×10^{-4}	intron_variant_intron_variant,non_coding_transcript_variant	DDAH1_NA	rs233091	DDAH1_eQTL_Gen
chr10_102678808_C_T	0.6	-4.29	1.8×10^{-5}	0.4	0.29	0.0039	0.4	0.26	1.4×10^{-4}	intron_variant	ARL3	rs7077678	SFXN2_eQTL_Gen
chr16_386808_G_A	0.8	4.31	1.6×10^{-5}	0.19	-0.49	1.6×10^{-4}	0.2	-0.24	9.8×10^{-3}	intron_variant_upstream_gene_variant_intron_variant,non_coding_transcript_variant_non_coding_transcript_exon_variant	LOC100134368_PGAP6_LOC105371036_NA	rs780300251	TMEM8A_eQTL_Gen & MCHC_GWAS_AFR

Figure 18. DRBC GWAS and exome associations. a) Manhattan plot of single variant association results with red blood cell density (DRBC) in 581 sickle cell disease patients. Bonferroni threshold line at 8.6×10^{-8} (0.05/ 580,149). b) Manhattan plot of gene-based results with red cell density in sickle cell patients. Bonferroni threshold line at 1.7×10^{-6} (0.05/ (14,949 x 2)) c) QQplot of gene-based results prioritized according to drug targets as identified by IUPHAR, and genes expressed in erythroid lineage or erythrocytes for SKAT test. d) QQplot of gene-based results prioritized according to drug targets identified by IUPHAR and genes expressed in erythroid lineage or erythrocytes for VT test. For both c and d, black dots represent all the markers together, blue dots represent markers ligand-gated ion channel, orange dots represent catalytic receptors, purple-colored dots represent enzymes, cyan-colored dots represent G protein-coupled receptors, maroon dots represent nuclear hormone receptors, green dots represent other ion channels, red dots represent other protein targets, burgundy dots represent transporters, and magenta dots represent voltage-gated ion channels. The grey area corresponds to the 95% confidence interval. λ_{GC}, genomic inflation factor.


10
9
8
7
6
5
4
3
2
1
0
$-\log_{10}(P)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 23
Chromosomes


10
9
8
7
6
5
4
3
2
1
0
$-\log_{10}(P)$
1 2 3 4 5 6 7 8 9 10 11 12 13 15 17 19 21 23
Chromosomes


$\lambda_{GC_{SKAT}}$ = 1.3
N_{SKAT} = 2119
$\lambda_{GC_{lgic\_IUPHAR}}$ = 1.9
N_{lgic_IUPHAR} = 36
$\lambda_{GC_{catalytic\_receptor\_IUPHAR}}$ = 1.6
$N_{catalytic\_receptor\_IUPHAR}$ = 187
$\lambda_{GC_{enzyme\_IUPHAR}}$ = 1.2
N_{enzyme_IUPHAR} = 917
$\lambda_{GC_{gpcr\_IUPHAR}}$ = 1.2
N_{gpcr_IUPHAR} = 156
$\lambda_{GC_{nhr\_IUPHAR}}$ = 1.8
N_{nhr_IUPHAR} = 28
$\lambda_{GC_{other\_ic\_IUPHAR}}$ = 1.3
$N_{other\_ic\_IUPHAR}$ = 27
$\lambda_{GC_{other\_protein\_IUPHAR}}$ = 1.2
$N_{other\_protein\_IUPHAR}$ = 142
$\lambda_{GC_{transporter\_IUPHAR}}$ = 1.2
$N_{transporter\_IUPHAR}$ = 372
$\lambda_{GC_{vgic\_IUPHAR}}$ = 0.91
N_{vgic_IUPHAR} = 81
LETM1
LRRC32
KCNJ14
Observed $-\log_{10}(P)$
Expected $-\log_{10}(P)$
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 1 2 3 4 5

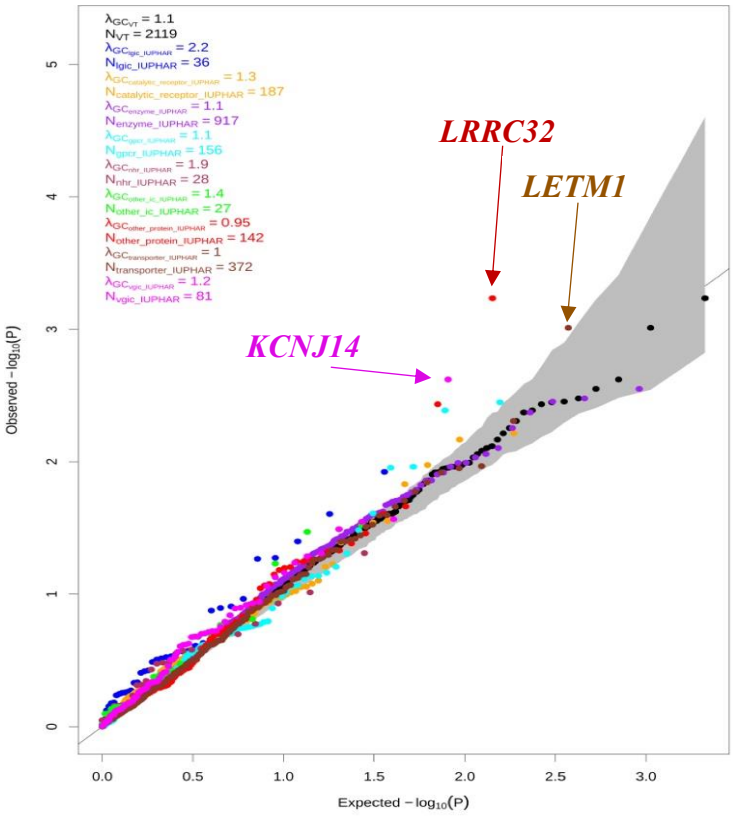
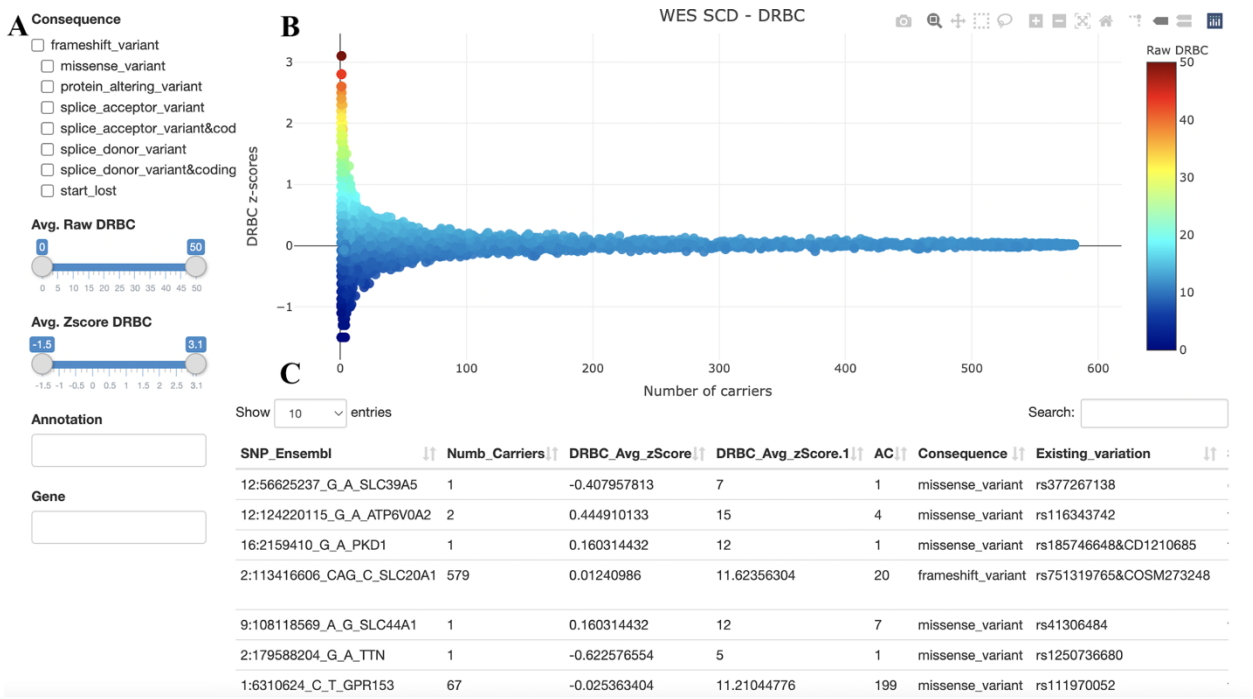
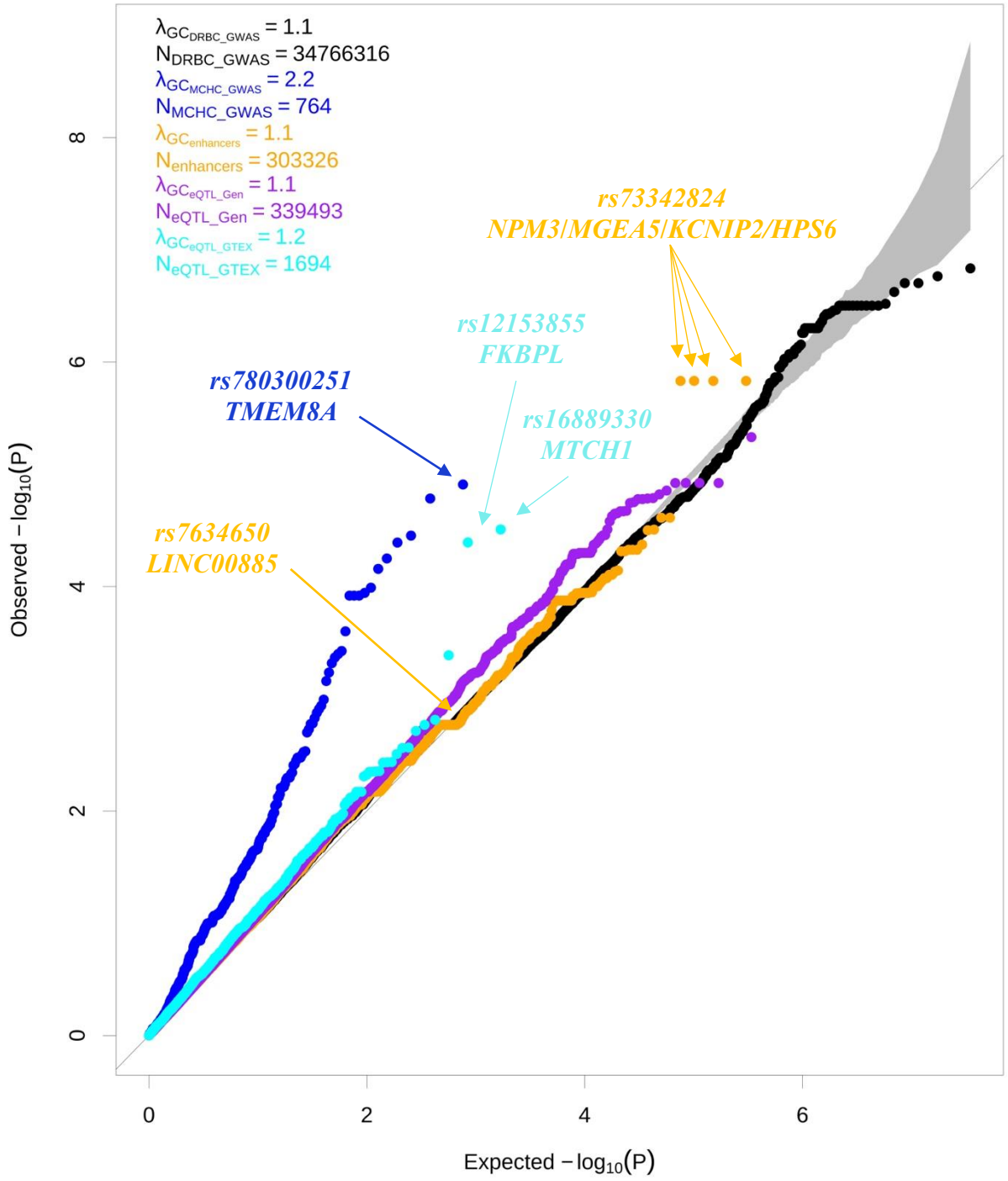

$\lambda_{GC_{VT}}$ = 1.1
N_{VT} = 2119
$\lambda_{GC_{lgic\_IUPHAR}}$ = 2.2
N_{lgic_IUPHAR} = 36
$\lambda_{GC_{catalytic\_receptor\_IUPHAR}}$ = 1.3
$N_{catalytic\_receptor\_IUPHAR}$ = 187
$\lambda_{GC_{enzyme\_IUPHAR}}$ = 1.1
N_{enzyme_IUPHAR} = 917
$\lambda_{GC_{gpcr\_IUPHAR}}$ = 1.1
N_{gpcr_IUPHAR} = 156
$\lambda_{GC_{nhr\_IUPHAR}}$ = 1.9
N_{nhr_IUPHAR} = 28
$\lambda_{GC_{other\_ic\_IUPHAR}}$ = 1.4
$N_{other\_ic\_IUPHAR}$ = 27
$\lambda_{GC_{other\_protein\_IUPHAR}}$ = 0.95
$N_{other\_protein\_IUPHAR}$ = 142
$\lambda_{GC_{transporter\_IUPHAR}}$ = 1
$N_{transporter\_IUPHAR}$ = 372
$\lambda_{GC_{vgic\_IUPHAR}}$ = 1.2
N_{vgic_IUPHAR} = 81
LRRC32
LETM1
KCNJ14
Observed $-\log_{10}(P)$
Expected $-\log_{10}(P)$
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 1 2 3 4 5

Figure 19. WES variant scoring interface. a. Users can select eight types of variant consequences from top to bottom. Sliders are provided to filter variants scores for both raw and normalized DRBC values. Users can directly specify gene(s) of interest (Gene Symbol). Lastly, different annotations (malaria, spherocytosis, MCHC GWAS, etc.…) can be selected. b Correlation between the carrier frequency of the coding variant in 2,333 RBC volume candidate genes and mean normalized DRBC in 581 patients. (c), dynamic table with variants position, number of carriers, raw and normalized DRBC, rsID, SIFT and Polyphen deleteriousness prediction, SampleID of individual carrying the mutation, Allele frequency, GWAS p-value, annotation, and gene symbol.



SNP_Ensembl	Numb_Carriers	DRBC_Avg_zScore	DRBC_Avg_zScore.1	AC	Consequence	Existing_variation
12:56625237_G_A_SLC39A5	1	-0.407957813	7	1	missense_variant	rs377267138
12:124220115_G_A_ATP6V0A2	2	0.444910133	15	4	missense_variant	rs116343742
16:2159410_G_A_PKD1	1	0.160314432	12	1	missense_variant	rs185746648&CD1210685
2:113416606_CAG_C_SLC20A1	579	0.01240986	11.62356304	20	frameshift_variant	rs751319765&COSM273248
9:108118569_A_G_SLC44A1	1	0.160314432	12	7	missense_variant	rs41306484
2:179588204_G_A_TTN	1	-0.622576554	5	1	missense_variant	rs1250736680
1:6310624_C_T_GPR153	67	-0.025363404	11.21044776	199	missense_variant	rs111970052

Figure 20 Quantile-quantile plot of red blood cell density in 573 sickle cell disease patients. All variants (black), variants mapping to erythroid enhancers (orange), expression quantitative trait loci variant (eQTL) for 2,333 candidate genes implicated in red blood cell hydration from GTEx (purple) from eQTLGen (light blue), and markers associated with mean corpuscular hemoglobin concentration (MCHC) from previous genome-wide association studies (navy blue). The grey area corresponds to the 95% confidence interval. λ_{GC}, genomic inflation factor.


λGC_DRBC_GWAS = 1.1
N_DRBC_GWAS = 34766316
λGC_MCHC_GWAS = 2.2
N_MCHC_GWAS = 764
λGC_enhancers = 1.1
N_enhancers = 303326
λGC_eQTL_Gen = 1.1
N_eQTL_Gen = 339493
λGC_eQTL_GTEX = 1.2
N_eQTL_GTEX = 1694
rs73342824
NPM3/MGEA5/KCNIP2/HPS6
rs12153855
FKBPL
rs780300251
TMEM8A
rs16889330
MTCH1
rs7634650
LINC00885
Observed −log10(P)
Expected −log10(P)
0
2
4
6
8

Chapter 4. Potential causal role of l-glutamine in sickle cell disease painful crises: A Mendelian randomization analysis

The presented article has been published in the journal, *Blood Cells, Molecules, and Diseases*. In this study, I employed MR (Mendelian randomization) to investigate the causal relationship between l-glutamine levels and painful crises in patients with sickle cell disease (SCD). Furthermore, I successfully identified 66 metabolites that exhibit associations with various SCD complications, such as gall bladder disease and renal dysfunction. This approach exemplifies the efficacy of integrating genetics and metabolomics to gain insights into the pathophysiology of SCD.

Potential causal role of l-glutamine in sickle cell disease painful crises: A Mendelian randomization analysis

Blood Cells, Molecules, and Diseases, Volume 86, February 2021, 102504,doi: 10.1016/j.bcmd.2020.102504

Yann Ilboudo[a,b] , Melanie E. Garrett[c] , Pablo Bartolucci[d] , Carlo Brugnara[e] , Clary B. Clish[f] , Joel N. Hirschhorn[f,g] , Frédéric Galactéros[d] , Allison E. Ashley-Koch[c] , Marilyn J. Telen[h] , Guillaume Lettre[a,b,*]

a Montreal Heart Institute, Montréal, Québec, Canada

b Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada

c Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC, USA

d Red Cell Genetic Disease Unit, Hôpital Henri-Mondor, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Est, IMRB - U955 - Equipe no 2, Créteil, France

e Department of Laboratory Medicine, Boston Children's Hospital, Boston, MA, USA

f Broad Institute, Cambridge, MA, USA

g Boston Children's Hospital, Boston, MA, USA

h Department of Medicine, Division of Hematology, Duke University Medical Center, Durham, NC, USA

* Corresponding author at: Montreal Heart Institute, 5000 Belanger Street, Montreal, Quebec, Canada. E-mail address: guillaume.lettre@umontreal.ca (G. Lettre).

ABSTRACT

In a recent clinical trial, the metabolite L-glutamine was shown to reduce painful crises in sickle cell disease (SCD) patients. To support this observation and identify other metabolites implicated in SCD clinical heterogeneity, we profiled 129 metabolites in the plasma of 705 SCD patients. We tested correlations between metabolite levels and six SCD-related complications (painful crises, cholecystectomy, retinopathy, leg ulcer, priapism, aseptic necrosis) or estimated glomerular filtration rate (eGFR), and used Mendelian randomization (MR) to assess causality. We found a potential causal relationship between L-glutamine levels and painful crises (N = 1278, odds ratio (OR) [95% confidence interval] = 0.68 [0.52–0.89], P = 0.0048). In two smaller SCD cohorts (N = 299 and 406), the protective effect of L-glutamine was observed (OR = 0.82 [0.50–1.34]), although the MR result was not significant (P = 0.44). We identified 66 significant correlations between the levels of other metabolites and SCD-related complications or eGFR. We tested these correlations for causality using MR analyses and found no significant causal relationship. The baseline levels of quinolinic acid were associated with prospectively ascertained survival in SCD patients, and this effect was dependent on eGFR. Metabolomics provide a promising approach to prioritize small molecules that may serve as biomarkers or drug targets in SCD.

INTRODUCTION

Sickle cell disease (SCD) is one of the most common Mendelian diseases in the world, affecting millions of patients living in Sub-Saharan Africa and the Indian sub-continent[26]. In the United States, > 100,000 individuals, mostly of African descent, live with SCD, and healthcare costs associated with SCD management and treatment are substantial[234]. Although fundamentally a disease of the blood – caused by mutations in the β-globin gene HBB – SCD is characterized by systemic and debilitating complications, such as painful crises, stroke, pulmonary hypertension and kidney failure. Unfortunately, there are no robust prognostic biomarkers to predict who will develop which complications, and when. SCD treatment still relies primarily on chronic blood transfusions and hydroxyurea (HU), a drug that acts partly by raising the concentration of anti-sickling fetal hemoglobin (HbF)[235].

Progress in gene therapies and genome editing technologies now offer realistic hope of developing a cure for SCD[236]. However, these complex clinical interventions are unlikely to benefit most SCD patients worldwide in the short term. Therefore, we need to continue searching for novel biomarkers and drug targets for SCD. Recently, the US Food and Drug Administration approved three new molecules to treat SCD (L-glutamine, crizanlizumab-tmca and voxelotor). In a double-blind phase 3 clinical trial, L-glutamine was shown to reduce the number of painful crises over a 48-week period[109]. The emergence of L-glutamine as a therapy was based on decades of work investigating the role of oxidative stress in SCD pathophysiology [237]. Red blood cells (RBC) from SCD patients have high oxidative stress and a compromised ability to counteract free radicals due to a low ratio of the reduction-oxidation (redox) co-factor nicotinamide adenine dinucleotide (NAD) and its reduced form ([NADH]:[NAD++NADH])[140]. L-glutamine is one of the most abundant amino acids in the human body and in addition to its role in protein synthesis, is required to synthesize NAD. Treatment with L-glutamine increases the NAD redox ratio and reduces adhesion of sickle RBC to endothelial cells, a hallmark of vaso-occlusive painful crises[141,238].

Metabolites, like L-glutamine, are small molecules that result from the activities of endogenous enzymes[239]. The development of high throughput mass spectrometry-based methodologies makes it possible to profile 100–1000s of metabolites in human biospecimens. Such metabolomic studies have been used to identify metabolite signatures of diseases, but also to pinpoint specific metabolites that may have prognostic and/or therapeutic values[240]. Metabolite

levels are variable between individuals (in disease, but also in health) and large genetic studies – termed metabolite genome-wide association studies (mGWAS) – have identified 1000s of genetic variants that control them. Besides providing an opportunity to characterize the biological pathways that control metabolite levels, these genetic discoveries become powerful instruments for Mendelian randomization (MR) studies. MR uses genetic variants to determine the effect of genetically modulated phenotypes on disease outcome[241]. MR mimics randomized clinical trial as it harnesses the random allocation of parental alleles when they are passed on to their offspring. As a consequence, the alleles are independently distributed in the population and free from potential confounders[242,243]. For instance, it has been possible to show using MR that tobacco smoking, as opposed to other confounders such as socioeconomic status, causes lung cancer by demonstrating that a genetic variant associated with smoking heaviness and located in the nicotinic acetylcholine receptor subunit genes is also associated with lung cancer only through its effect on smoking habits[244]. Other MR studies have validated many drug targets for various human diseases (e.g. statins that lower LDL-cholesterol levels to reduce coronary artery disease (CAD) risk), but have also been useful to rule out many biomarkers as potential causal factors (e.g. HDL-cholesterol or C-reactive protein for CAD) [170,245,246].

In SCD, only a limited number of studies have used metabolomic approaches to tackle clinical heterogeneity. Zhang *et al*. discovered increased adenosine levels in blood from SCD patients and transgenic mice: they showed that higher adenosine levels exacerbated sickling, hemolysis and organ damage[136]. Additionally, the same group found that sphingosine-1-phosphate (S1P) and 2,3-bisphosphoglycerate (2,3-BPG) blood concentrations are elevated in SCD patients and mice, which results in the re-programming of the glycolysis program and enhanced disease severity [137,138]. Finally, Darghouth *et al*. profiled the metabolome of RBC from healthy individuals and SCD patients and identified several metabolites that highlight differences between the two groups in glycolysis, membrane turnover, and glutathione and nitric oxide metabolism[139]. Although exciting, these pioneering metabolomic studies were performed in a limited number of SCD patients (N = 14–30) and did not take advantage of MR methodology to address causality.

To prioritize metabolites that may be important biomarkers or drug targets for SCD, we profiled 129 known metabolites, including L-glutamine, in the plasma of 705 SCD patients. First, we used MR to test the causal relationship between plasma L-glutamine levels and painful crises. Second, we tested the association between all measured metabolites and SCD-related

complications and combined these results with previous mGWAS findings to perform MR studies. Finally, we tested if baseline plasma metabolite levels were associated with survival in our SCD cohorts. Our results highlight the value of combining genetic and metabolomic strategies to disentangle the complex pathophysiology of SCD.

2. SUBJECTS AND METHODS

2.1. Study participants

Sample collections and procedures were in accordance with the institutional and national ethical standards of the responsible committees and proper informed written consent was obtained. The Genetic Modifier (GEN-MOD), the Cooperative Study of Sickle Cell Disease (CSSCD), and the Duke University Outcome Modifying Genes (OMG) cohorts have been described elsewhere [126,247,248]. Plasma samples were collected during steady state in outpatient visit over 3 weeks from treatment for vaso-occlusive crisis in OMG, and over 4 weeks for GEN MOD. In particular for GEN-MOD, a dedicated research assistant validated all clinical information. Demographic and clinical information for each SCD cohort is available in Table 1. For both GEN-MOD and OMG, patients were not fasting when plasma was collected, and diet information is not available.

2.2. Metabolomics profiling

Plasma metabolites were profiled using two complimentary liquid chromatography tandem mass spectrometry (LC-MS) methods. Amino acids, amino acid metabolites, acylcarnitines, and other cationic polar metabolites were measured using a Nexera X2 U-HPLC (Shimadzu Corp.) coupled to a Q Exactive Hybrid Quadrupole Orbitrap Mass Spectrometer (Thermo Fisher Scientific). Plasma samples (10 µL) were prepared via protein precipitation, with the addition of 9 volumes of acetonitrile/methanol/formic acid (74.9:24.9:0.2; v/v/v) containing stable isotope-labeled, quality control internal standards (valine-d8, Sigma-Aldrich; St. Louis, MO; and phenylalanine-d8, Cambridge Isotope Laboratories; Andover, MA). The samples were centrifuged (10 min, 9000×g, 4 °C), and the supernatants were injected directly onto a 150 × 2 mm, 3 µm Atlantis HILIC column (Waters). The column was eluted isocratically at a flow rate of 250 µL/min with 5% mobile phase A (10 mM ammonium formate and 0.1% formic acid in water) for 0.5 min followed by a linear gradient to 40% mobile phase B (acetonitrile with 0.1% formic acid) over 10 min. MS analyses were carried out using electrospray ionization in the positive ion mode using full scan analysis over 70–800 m/z at 70,000 resolution and 3 Hz data acquisition rate. Other MS settings were: sheath gas 40, sweep gas 2, spray voltage 3.5 kV, capillary temperature

350 °C, S-lens RF 40, heater temperature 300 °C, microscans 1, automatic gain control target 1e6, and maximum ion time 250 ms. Raw data were processed using TraceFinder software (Thermo Fisher Scientific; Waltham, MA) for supervised, targeted extraction of data from a subset of lipids and Progenesis QI (Nonlinear Dynamics; Newcastle upon Tyne, UK). Organic acids, sugars, purines, pyrimidines, and other anionic polar metabolites were measured using an ACQUITY UPLC (Waters Corp, Milford MA) coupled to a 5500 QTRAP triple quadrupole mass spectrometer (AB SCIEX, Framingham MA). Plasma samples (30 µL) were extracted using 120 µL of 80% methanol containing 0.05 ng/µL inosine15 N4, 0.05 ng/µL thymine-d4, and 0.1 ng/µL glycocholate-d4 as quality control internal standards (Cambridge Isotope Laboratories, Inc., Tewksbury MA). The samples were centrifuged (10 min, 9000×g, 4 °C) and the supernatants (10 µL) were injected directly onto a 150 × 2.0 mm Luna NH2 column (Phenomenex, Torrance CA). The column was eluted at a flow rate of 400 µL/min with initial conditions of 10% mobile phase A (20 mM ammonium acetate and 20 mM ammonium hydroxide (Sigma-Aldrich) in water (VWR)) and 90% mobile phase B (10 mM ammonium hydroxide in 75:25 v/v acetonitrile/methanol (VWR)) followed by a 10 min linear gradient to 100% mobile phase A. The ion spray voltage was −4.5 kV, the source temperature was 500 °C, and multiple reaction monitoring (MRM) settings for each metabolite were determined using authentic reference standards. Raw data were processed and visually reviewed using MultiQuant software (AB SCIEX, Framingham MA).

2.3. Metabolomics pre-processing

We removed metabolites with > 20% missing values. We imputed missing metabolite values using the k-nearest neighbors algorithm[249] as implemented in the R package impute. We log10-transformed metabolite values, and applied batch effect correction based on metabolites' dates of extraction, the types of ionization, and whether they were obtained from targeted or untargeted approaches. Finally, we applied batch effect correction based on the year of profiling, since sample collection occurred within a 3 years span (2015–2017). We conducted all batch effect correction using combat[250]. Using a linear model, we then derived residuals correcting for age, sex, SCD genotypes, HU usage, and cohort affiliation. Supplementary Fig. 1 summarizes the design of our metabolomic experiment. Although we captured many unknown metabolites, which we used as part of the quality-control steps, this study focuses on the 129 known

metabolites that were available in both GEN-MOD and OMG. All metabolite levels that we measured in this study are in arbitrary units as we did not perform absolute quantification.

2.4. Metabolite levels association with SCD complications, eGFR, or survival in GEN-MOD and OMG

All statistical tests performed in this study, including models and covariates, are thoroughly described in **Supplementary Table 1**. We implemented a permutation procedure that considers the correlation between metabolite levels to test the association between metabolite levels and SCD complications (painful crises, aseptic necrosis, cholecystectomy (gall bladder removal), retinopathy, priapism, leg ulcer, survival), estimated glomerular filtration rate (eGFR, calculated using the chronic kidney disease epidemiology collaboration (CKD-EPI) equation[251]) or to predict the risk of prospectively ascertained death (survival). For eGFR, we chose the CKD-EPI rather than MDRD equation in order to properly assess high GFR values; this approach allows for the ascertainment of hyperfiltration, which is often observed in SCD patients[247,252]. We did not measure cystatin C levels, and are not aware of any data on SCD cohorts with concomitant measures of GFR and cystatin C. We randomly permuted the phenotype of interest and computed 100,000 P-values (for each metabolite) in a linear or a logistic model. We then stored the smallest P-value out of the 100,000, and obtained the adjusted/permutated P-value (P_{perm}) by comparing the number of times the permutated P-values are smaller than the observed P-values:

$$P_{perm} = \frac{(b+1)}{(m+1)}$$

where b is the number of times P_{perm} is greater or equal than Pobs, and m the number of permutations. The procedure was implemented in the R statistical package.

Genetic association study in the CSSCD DNA genotyping and genotype imputation in the CSSCD have been described in detail elsewhere[126]. We restricted our analysis to markers with imputation quality r2 > 0.3 and minor allele frequency (MAF) > 1%. We removed the effect of sex and age on batch effect corrected metabolites levels, and used inverse normal transformation to normalize the residuals. We used RvTests (v20171009)[166] to test the association between genotype dosage and the various phenotypes: we used logistic regression models for painful

crises or cholecystectomy, and linear regression models for bilirubin and eGFR. All statistical models are defined in **Supplementary Table 1**. For eGFR, we did not correct for age and sex because the eGFR-EPI equation takes them both into account.

2.5. Genetic association study in the CSSCD

DNA genotyping and genotype imputation in the CSSCD have been described in detail elsewhere [25]. We restricted our analysis to markers with imputation quality r2 > 0.3 and minor allele frequency (MAF) > 1%. We removed the effect of sex and age on batch effect corrected metabolites levels, and used inverse normal transformation to normalize the residuals. We used RvTests (v20171009)[166] to test the association between genotype dosage and the various phenotypes: we used logistic regression models for painful crises or cholecystectomy, and linear regression models for bilirubin and eGFR. All statistical models are defined in Supplementary Table 1. For eGFR, we did not correct for age and sex because the eGFR-EPI equation takes them both into account.

2.6 Mendelian randomization

2.6.1. Instrument identification

Because of our reduced sample size, we selected instruments for MR analyses from large published mGWAS carried out in healthy individual of European ancestry. We identified metabolite-associated variants from the published meta-analysis of KORA and TwinsUK (N = 6056 + 1768, 529 metabolites), as well as the whole-genome sequence metabolite association study in TwinsUK (N = 1960, 644 metabolites) [253,254]. We focused on these publications because they are the two largest published mGWAS to date. We selected sub-genome wide significant mGWAS SNPs ($P < 1 \times 10{-}5$) in order to maximize the phenotypic variance explained, and tested two MR models. The first MR model included all sub-genome wide significant SNPs as valid instruments. For the second MR model, we removed pleiotropic SNPs from the first model if they were associated with other metabolites at a Bonferroni corrected $P < 0.05$ threshold when considering the number of SNPs in model 1. Pleiotropic SNPs were identified by querying Phenoscanner[255].

2.6.2 Instrument pruning

We employed PLINK1.9v5.2[218] to identify independent SNP within 5-Mb window and linkage disequilibrium (LD) r 2 < 0.01 in the CSSCD. This provided us with a list of pseudo-independent variants.

2.6.3. Analysis

We used a two-sample MR approach to test the causal link between metabolites and SCD-related phenotypes. We retrieved association results (effect sizes, standard errors) between instruments and SCD-related phenotypes from the large and clinically well characterized CSSCD. All MR analyses were performed in R version 3.5.1 with the TwoSampleMR package (v0.4.22)[171]. We used a multiplicative random effect inverse variance-weighted (IVW) method in each MR analysis. For the analysis of plasma L-glutamine levels and painful crises, we tested 2 models and defined statistical significance using a Bonferroni corrected threshold of $\alpha \leq 0.025$. All other analyses were exploratory and statistical significance was set at nominal $\alpha \leq 0.05$. Additionally, we computed the weighted median[256], which selects the median MR estimate as the causal estimate, and MR-Egger[257], which allows the intercept to vary freely and therefore estimates the amount of horizontal pleiotropy, for all the analyses. Moreover, we utilized MR-PRESSO (Pleiotropy Residual Sum and Outlier)[258] to estimate the presence of horizontal pleiotropic bias and to calculate causal estimate adjusted for outliers for all reported results. Finally, we assess the validity of our significant results by conducting additional tests for horizontal pleiotropy, including Cochran's Q statistic, MREgger intercept test of deviation from the null, and I^2 heterogeneity statistic[241].

2.7. Genetic risk scores (GRS)

Using PLINK1.9v5.2[218], we calculated the genetic risk scores for L-glutamine and 3-ureidopropionate in CSSCD, GEN-MOD and OMG. Effect size estimates from the two large mGWAS referenced in the MR analysis served as weights. Detailed description of the logistic and linear models employed for CSSCD, GEN-MOD, and OMG for inverse normal transformed

GRS association with painful crisis, eGFR, L-glutamine, and 3-ureidopropionate is available in **Supplementary Table 1**. We performed principal component analysis in PLINK using 1000 Genomes Project populations as reference.

2.8. Data sharing statement

The CSSCD genetic dataset is available on the database of Genotypes and Phenotypes (dbGaP: https://www.ncbi.nlm.nih.gov/gap/)The GEN-MOD and OMG data are available upon requests to the authors.

3. RESULTS

3.1. Plasma metabolites in SCD patients

To identify metabolites that may be useful to predict or treat SCD complications, we measured plasma values of 129 known metabolites in 705 patients from the GEN-MOD and OMG cohorts (**Supplementary Fig. 1** and **Table 1**). Although our metabolomic experiment was performed at the same center, it was run in three batches so we applied stringent quality-control and batch effect correction filters to avoid confounding (Methods and **Supplementary Fig. 2**). The two main classes of metabolites that we measured were amino acids (33%) and lipids (30%) (**Supplementary Fig. 3** and **Supplementary Table 2**).

3.2. Mendelian randomization supports a potential causal link between L-glutamine and SCD painful crises

L-glutamine therapy in SCD was previously shown to improve the NAD redox ratio, although this effect was not detected in a recent clinical trial[109,259]. Because we measured plasma L-glutamine levels as part of our metabolomic experiment, we were interested to test association between its plasma levels and SCD-related complications or other clinically relevant parameters. In GEN-MOD and OMG, we found no evidence of association between plasma L-glutamine levels and SCD complications, including painful crises (**Table 2**). However, plasma L-glutamine levels were nominally associated with several hematological traits measured at baseline,

including reduced hemoglobin concentration and RBC count (**Table 2**). For SCD complications, interpretation of these results is challenging because clinical events occurred before plasma L-glutamine was measured, and this one time metabolomic measure may not reflect life-long endogenous exposure to L-glutamine. For these reasons, we sought to further test the relationship between L-glutamine and SCD painful crises using MR.

Instrument strength plays a critical role in the validity of MR analyses. Although we measured plasma L-glutamine levels in 705 SCD patients, we wanted to take advantage of existing and well-powered mGWAS for the selection of the best metabolite associated SNPs to use as MR instruments[253,254]. However, these mGWAS were carried out in Europeans, whereas SCD patients in our cohorts are of African-descent (**Supplementary Fig. 4**), raising the question whether we could use SNPs found in Europeans as MR instruments for phenotypes observed in African ancestry SCD patients. To validate this strategy, we tested the well known causal link between bilirubin levels in serum and gallstones leading to surgical removal of the gallbladder (cholecystectomy), a complication often observed in SCD patients[260,261]. From a GWAS of serum bilirubin levels in 9464 individuals of European ancestry, we selected 10 SNPs as MR instruments[262]. In the large and well characterized CSSCD (**Table 1**), we tested the association between these SNPs and bilirubin levels or cholecystectomy, and replicated the strong association between these phenotypes and the UGT1A1 locus (**Supplementary Table 3**). The two-sample inverse variance-weighted (IVW) MR analysis confirmed that high bilirubin levels causes gallbladder disease in SCD: a one standard deviation increase in genetically-controlled bilirubin levels was associated with a 6-fold increase in the risk of cholecystectomy in the CSSCD (odds ratio (OR) [95% confidence interval] = 6.0 [2.8–17.0], P_{IVW} = 1.9 × 10−6) (**Supplementary Table 4**).

From the available mGWAS results[253,254], we identified 51 SNPs associated with plasma L-glutamine levels at $P < 5 \times 10{-}5$ that were available in the CSSCD genetic dataset. Single variant and polygenic trait score association results are available in Supplementary Table 5[126]. Using these 51 SNPs as instruments in a two-sample IVW MR analysis, we did not detect a causal association between L-glutamine and painful crises (Model 1: OR = 0.81 [0.63–1.00], P_{IVW} = 0.086) (**Fig. 1**). When we excluded 24 pleiotropic SNPs (Methods) and repeated the analysis with the remaining 27 SNPs, the MR association with painful crises was significant: a one standard deviation increase in genetically-controlled L-glutamine levels was associated with a 32% reduction in the risk of painful crises in the CSSCD (Model 2: OR = 0.68 [0.52–0.89], P_{IVW} =

0.0048) (**Fig. 1**). MR analyses using the sensitivity tests MR-Egger and weighted-median did not yield significant associations for Model 2, suggesting insufficient statistical power for these tests[241] or potential residual horizontal pleiotropy (**Supplementary Table 6**). We repeated the MR analysis in the GEN-MOD and OMG cohorts: although the direction of the effect of the GEN-MOD+OMG meta-analysis indicated a protective effect of L-glutamine on painful crises (OR = 0.82 [0.54–1.34]), the result was not significant (P_{IVW} = 0.54), presumably due to limited power given the smaller sample size (**Supplementary Table 6**). In secondary MR analyses, we found no evidence of causal associations between L-glutamine SNPs and several other SCD complications (**Supplementary Table 6**).

To determine if the 27 SNPs in model 2 capture the known L-glutamine biology, we annotated the nearby genes using ProGeM[263] (**Supplementary Table 7**). Although genes at many loci remain to be characterized, ProGeM highlighted strong candidate genes: (1) *DBT* (rs524219) encodes an enzyme involved in the synthesis of glutamate, a precursor to L-glutamine [44], (2) NADSYN1 (rs10431159), which encodes glutamine-dependent NAD(+) synthetase-1, has been implicated in L-glutamine synthesis through a reaction with nitrogen[264,265], (3) *SLC38A8* (rs12447776) is a sodium-coupled neutral amino acid transporter with a preference for glutamine [266,267], (4) PPA2 (rs4699183) plays a role in the detection of glutamine levels[268], and (5) *AADAT* (rs138354882) is involved in the conversion of glutamine to glutamate through the production of α-ketoglutarate[269].

3.3. Potential causal link between 3-ureidopropionate and kidney function in SCD

We tested 6 SCD-related complications as well as eGFR against the levels of the 129 known metabolites measured in GEN-MOD and OMG. In total, we found 66 metabolites with Pperm ≤ 0.05, including 62 metabolites associated with eGFR, two metabolites associated with painful crises and two metabolites associated with cholecystectomy (**Fig. 2** and **Supplementary Table 8**). There was a strong association between eGFR and creatinine levels, which serves as an internal control given that we use this metabolite to calculate eGFR. Most of these metabolites have never been linked to SCD and future work could therefore test if they represent potential novel biomarkers of disease severity. Previous metabolomic studies had found high levels of adenosine, S1P and 2,3-BPG in the blood of SCD patients[136-138]. Unfortunately, we did not measure S1P and 2,3-BPG in our experiment. We could retrieve adenosine levels in a subset of

patients (N = 404), but we are cautious with interpretation since plasma adenosine has a short half-life and can be generated during blood extraction. Within the limitations of our experimental design, we observed higher adenosine levels in SCD patients with painful crises and cholecystectomy, although the results were not significant (**Supplementary Table 9**).

Using the same strategy as for L-glutamine, we derived MR instruments for 48 of the 66 metabolites identified in the pairwise analyses with SCD phenotypes; there were no significant mGWAS variants for the remaining 18 metabolites. Across these 48 tests, we identified a single nominally significant association in our two-sample MR analyses involving eGFR and 3-ureidoproprionate levels (see URL for all available MR results, including sensitivity tests). In a European mGWAS[254], we retrieved 22 SNPs associated with 3-ureidoproprionate levels, including 16 that were not pleiotropic (**Supplementary Table 10**). Our results indicate that a one standard deviation increase in genetically controlled 3-ureidopropionate levels was associated with improved eGFR of 0.07 mL/min per 1.172 m2 ($P_{IVW\text{-}model1} = 8.7 \times 10{-}4$; $P_{IVWmodel2} = 9.7 \times 10{-}4$) (Fig. 3 and Supplementary Table 11). The sensitivity analyses did not allow us to exclude the possibility of confounding due to horizontal pleiotropy (Fig. 3 and Supplementary Table 11). Furthermore, we could not replicate the MR result in GENMOD and OMG, indicating that larger SCD cohorts are needed to confirm the potential causal link between 3-ureidoproprionate and eGFR (**Supplementary Table 11**).

Again, we used ProGeM to annotate the genes located near the 16 non-pleiotropic SNPs used in the 3-ureidoproprionate MR analyses[263] (**Supplementary Table 12**). This metabolite has been less studied, but we found that one of the variants, rs11704820, maps to an intron of the beta-ureidopropionase (*UPB1*) gene. Otherwise, we retrieved few genes that have been implicated in kidney functions and renal disease: (1) *WDR72* (rs555045773) has been associated with eGFR variation[270] and (2) LPIN1 (rs78734409 and rs71394795) is linked to myoglobinuria, which leads to renal failure due to the accumulation of creatine kinase in the kidneys[271-273].

3.4. Predicting survival status using baseline metabolite levels

Given the clinical heterogeneity that characterizes this disease, being able to predict which SCD patients will follow a severe clinical course could be extremely useful. This is particularly true for more invasive therapeutic options (e.g. gene therapy) or in settings where resources to treat SCD are limited, such as Sub-Saharan Africa. Thus, we decided to explore the prognostic value

of plasma metabolites in SCD. As discussed above, the data currently available in GEN-MOD and OMG are largely retrospective. However, we could prospectively ascertain SCD severity using a simple definition based on survival status during the follow-up period (**Table 1** and Methods). We identified 10 metabolites that were nominally associated with survival status, but only quinolinic acid remained significant after permutations to account for the number of tests performed (**Table 2**). For all 10 metabolites, increased levels were associated with increased risk of death, and for all but cytosine levels, the effect on survival was mediated by an association with eGFR. Quinolinic acid is a product of the kynurenine pathway, which also metabolizes the amino acid tryptophan.

4. DISCUSSION

While the cause of SCD has been known for over a century, treatment options are limited and it is extremely difficult to predict which patients will have a more severe presentation of the disease. To continue to address these challenges, we performed the largest using a targeted approach in 705 participants. We also measured 1985 unknown metabolites as part of our experiment, but they were not considered in our analyses except during the pre-processing quality-control steps. Our effort was motivated by recent successes using this metabolomic approach to find new prognostic biomarkers and potential drug targets for human diseases[274]. In fact, a recent study showed using MR the causal impact of L-glutamine on RBC, mean corpuscular hemoglobin (MCH), and mean corpuscular hemoglobin concentration (MCHC) in healthy Europeans[275].

Among the 62 eGFR associated metabolites, we noted 11 acylcarnitines that were inversely correlated with eGFR. It is known that acylcarnitines accumulate in the plasma of patients with renal disease due to reduced clearance of esterified carnitine moieties, probably mediated by the renal tubular carnitine transporter OCTN2[276]. Our eGFR analyses also identified metabolites from the tryptophan metabolism (kynurenic acid, kynurenine) and the choline derivatives tubulointerstitial dysfunction and have been associated with incident chronic kidney disease[277]. We found two metabolites from secondary bile acid metabolism – deoxycholic acid glycine conjugate and taurodeoxycholic acid – that were associated with cholecystectomy. This observation is consistent with the fact that bile salts contribute to gallstones and gallbladder disease. Finally, we identified phosphocreatine and pyroglutamic acid as two metabolites

associated with SCD painful crises. Little is known about a role for phosphocreatine in pain biology, although its levels are increased in skeletal muscles of fibromyalgia patients[278]. Pyroglutamic acid belongs to the glutathione metabolism pathway and its levels were elevated in SCD patients with painful crises. Because high pyroglutamic acid levels are sometimes observed in individuals who chronically use acetaminophen[279], this association might therefore reflect pain management as opposed to a causal link between pyroglutamic acid and pain in SCD patients. This conclusion is supported by the fact that our MR analyses did not causally implicate pyroglutamic acid in painful crises.

By combining metabolite profiles with mGWAS results, we could use MR methods to test causality between metabolites and SCD-related complications. Using this strategy, we identified in the large CSSCD a promising causal relationship between plasma L-glutamine levels and painful crises, which provides independent evidence consistent with recent results from a phase 3 clinical trial[109]. Given that the CSSCD was initiated in the ~1980s and that L-glutamine was only recently approved, it supports the idea that new drugs targeting known pathophysiological mechanisms (e.g. increased oxidative stress) could yield effective SCD therapeutic options. Our analyses also highlighted 3-ureidoproprionate, an intermediate in the metabolism of uracil, as a potential positive modulator of eGFR. Interpretation of this result is difficult because little is known about this metabolite and the result was not replicated in additional SCD patients. Mutations in the gene UPB1, which encodes the enzyme that transforms 3-ureidopropionate into beta-alanine, cause beta-ureidopropionase deficiency, a rare monogenic disease characterized by high plasma levels of 3-ureidoproprionate [280]. Only a few patients with this disease have so far been described and they presented mostly with neurologic development issues. However, there is no report of abnormal glomerular filtration rate or other kidney defects in these patients. We propose that future MR replication in independent SCD cohorts and animal studies could be extremely useful to investigate the possible role of 3-ureidoproprionate in regulating kidney functions, and in particular whether raising 3-ureidopropionate levels could improve glomerular filtration rate in SCD patients.

Our study presents with a few limitations, especially as it relates to some of the MR assumptions and the ability to detect a causal effect between L-glutamine and SCD painful crises. First, our statistical power to detect heterogeneity due to confounding in our MR analyses and to replicate our main findings was limited because there are few large, well-characterized and

genotyped SCD cohorts available. Second, some of the SNPs used as MR instruments may be pleiotropic beyond the filtering that we applied (i.e. have an effect on multiple unknown metabolites and other phenotypes) so that it is not possible to rule out an effect on SCD complications that is independent from the tested metabolites. For these two reasons, it will be important to replicate the L-glutamine painful crises MR analyses in independent large SCD cohorts when they become available. Third, we used MR instruments derived from mGWAS performed in Europeans to test for causality in African ancestry SCD patients. There have been many reports on the transferability (or lack thereof) of GWAS findings across ancestries[281]. We used the well-known relationship between bilirubin levels and gallbladder disease to show that our approach can work. However, it is likely that having access to large mGWAS results in African ancestry populations would provide better instruments, and may lead to the identification of additional causal links between metabolites and SCD phenotypes by MR. Finally, we measured metabolite levels in plasma, but their levels in RBC could have provided complementary information (in particular for L-glutamine).

One characteristic of our study is that we measured metabolites in SCD cohorts that have mostly collected retrospective clinical data. One exception is information on the survival status of the participants. Using a simple linear model, we found a significant association between prospectively ascertained survival status and baseline quinolinic acid levels. This association was mediated by eGFR, consistent with our previous observation that quinolinic acid levels correlate with rapid renal function decline in SCD patients[247]. In the future, it will be interesting and important to test whether metabolites predict other complications in prospective SCD cohorts. In conclusion, our results motivate future experiments to integrate metabolite profiles and other orthogonal omics datasets (e.g. genetics) to build better predictors of SCD-related complications and overall severity.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bcmd.2020.102504.

CRediT authorship contribution statement

A.E.A.K., M.J.T and G.L. conceived this study. C.B.C. performed plasma metabolites profiling. Y.I. performed Mendelian randomization analyses. Y.I. and M.G. performed genetic analyses and genetic risk association analyses. Y.I. performed statistical analyses. G.L. supervised this work. Y.I. and G.L. wrote the manuscript with input from all authors.

Acknowledgments

We thank all participants for their contribution to this project. We also thank Adil Harroud and Brent Richards for advices on the Mendelian randomization analyses. G.L. is funded by the Canadian Institutes of Health Research (CIHR, PJT #156248), the Doris Duke Charitable Foundation, and the Canada Research Chair program. GEN-MOD sample and data collection were supported by NIH grant HL- 68922. A.A-K. M.J.T. and establishment and analysis of the OMG cohort have been funded by NHLBI (R01HL68959, HL79915, HL70769, HL87681).

URL

All Mendelian randomization results are available at:
http://www.mhi-humangenetics.org/dataset/MR_Analysis_SCD_everything.html

Declaration of competing interest

The authors have declared that no conflict of interest exists.

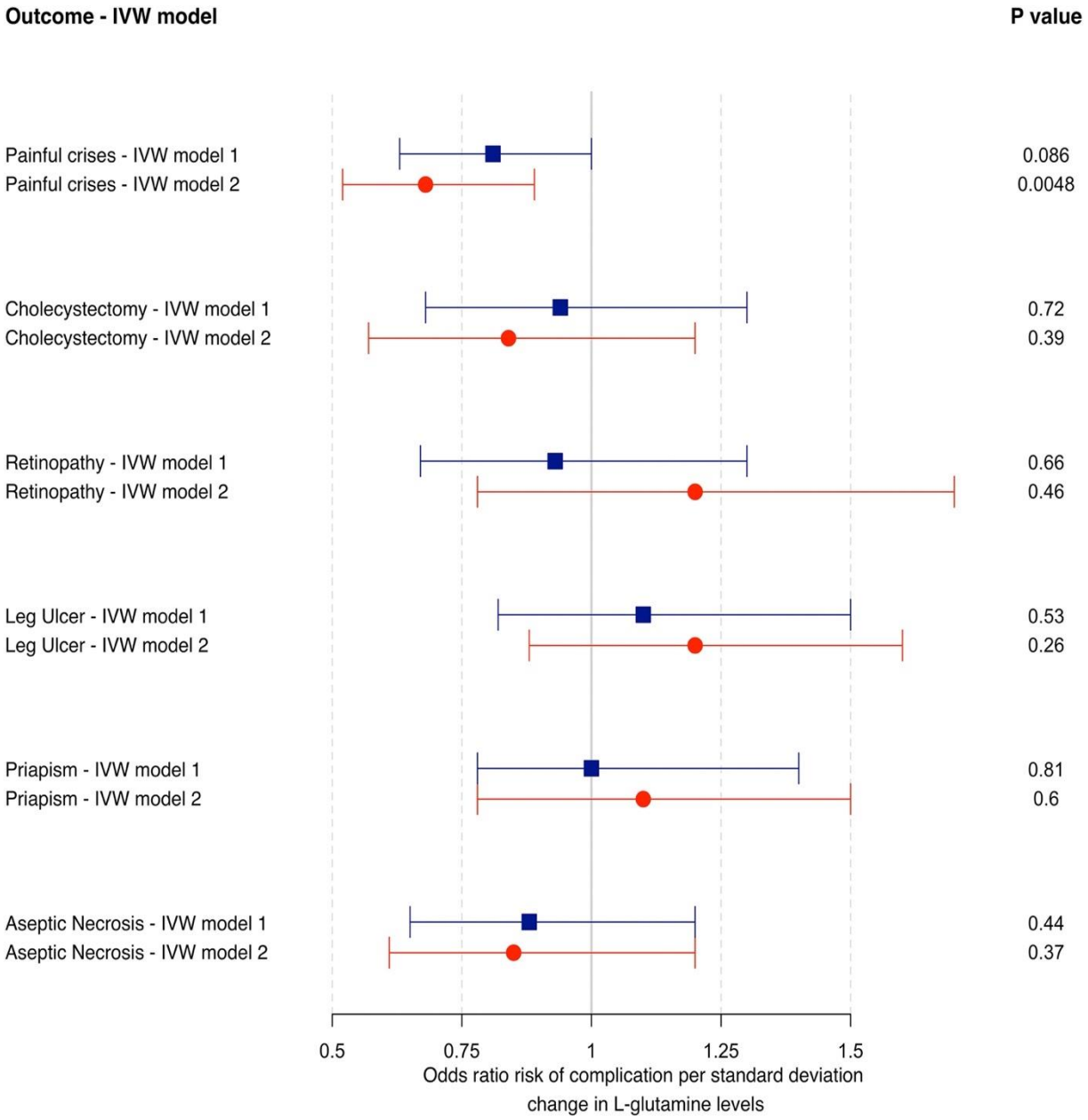
Table1. Demographics and clinical information. Sickle cell disease patients from three cohorts were included in this study. For the CSSCD, all data are prospective and we only considered patients with genome-wide genotyping data available. For GEN-MOD and OMG, all data were collected at baseline and are retrospective, except survival which is prospective. 1 Painful crises in GEN-MOD and OMG are defined as crises requiring hospitalization which was dichotomized (individuals with ≥1 painful crises in the last 12 months are assigned as cases, while individuals with no painful crisis are assigned as controls). In the CSSCD, painful crises are defined as painful episodes requiring emergency room visits, and we dichotomized the data as no crisis (control) or at least one crisis (case) during the follow-up period. For all quantitative variable, we provide mean ± standard deviation. LDH, lactate dehydrogenase; RBC, red blood cell; MCH: mean corpuscular hemoglobin; MCV: mean corpuscular volume; eGFR, estimated glomerular filtration rate.

Characteristic	GEN-MOD	OMG	CSSCD
Sex (male/female)	222/184	163/136	616/662
Age (year)	31 ± 9	35 ± 14	14 ± 12
β-globin genotypes (HbSS/ HbS β^0-thal/ HbSS α-thal/ HbSC/ HbS β+)	406/0/0/0/0	255/12/0/23/9	883/0/395/0/0
Painful crises (cases/controls)[1]	150/180	161/128	194/907
Leg ulcer (cases/controls)	30/300	79/214	185/970
Cholecystectomy (cases/controls)	200/206	173/72	152/932
Aseptic necrosis (cases/controls)	94/236	93/194	164/991
Priapism (cases/controls)	41/116	55/236	96/460
SCD retinopathy (cases/controls)	182/67	65/210	274/292
SCD survival (cases/controls)	19/384	35/91	44/1235
Bilirubin (mg/dL)	3.5± 2.1	2.9 ± 2.4	3.3 ± 2.2
eGFR (mL/min per 1.172 m^2)	143.6± 22.8	126.0 ± 40.3	165.3 ± 47.0
Hemoglobin (g/dL)	8.8 ± 1.3	8.2 ± 1.8	8.4 ± 1.3
Hematocrit (%)	25.8 ± 4.5	25.3 ± 5.7	24.8 ± 4.03
Lactate dehydrogenase (units/L)	400.4 ± 144.7	326.8 ± 240.2	451.6 ± 244.7
MCH (pg)	29.3 ± 4.1	32.0 ± 4.8	29.8 ± 3.1
MCV (fL)	87.0 ± 10.2	92.2 ± 12.7	89.2 ± 8.5
RBC count (x10^6 cells/μL)	3.0 ± 0.80	2.9 ± 0.80	2.8 ± 0.57

Table 2. Associations between L-glutamine plasma levels and sickle cell disease (SCD)-related complications and other clinically relevant phenotypes. In participants from the GEN-MOD and OMG cohorts, we tested the association between L-glutamine levels measured in plasma and SCD-related complications or clinically relevant blood-based biomarkers. Dichotomous phenotypes were analyzed using logistic regression while correcting for age, sex, hydroxyurea (HU) usage, SCD genotypes and cohort affiliation. Quantitative phenotypes were corrected for age, sex, HU usage, SCD genotypes and cohort affiliation. They were inverse normal-transformed before being tested for association using linear regression. Odds ratio and effect sizes (Beta) are given per standard deviation change in plasma L-glutamine levels. LDH, lactate dehydrogenase; RBC, red blood cell; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; eGFR, estimated glomerular filtration rate; LDH, lactate dehydrogenase; CI, confidence interval; SE, standard error.

Complications	N	Odds ratio	95% CI	P-value
Painful crises	619	1.06	(0.90-1.24)	0.52
Survival	529	1.01	(0.75-1.35)	0.79
Aseptic necrosis	617	0.97	(0.97-1.16)	0.76
Cholecystectomy	651	1.06	(0.90-1.25)	0.45
Leg ulcer	623	1.09	(0.88-1.35)	0.44
Priapism	448	1.11	(0.88-1.4)	0.39
Retinopathy	524	0.99	(0.82-1.18)	0.88
Renal Parameter	**N**	**Beta**	**SE**	**P-value**
eGFR	702	-0.067	0.036	0.067
Blood Parameter	**N**	**Beta**	**SE**	**P-value**
Bilirubin	585	0.10	0.041	0.010
Hematocrit	697	-0.08	0.035	0.019
Hemoglobin	685	-0.098	0.035	0.0048
LDH	579	0.078	0.039	0.044
MCH	626	0.067	0.035	0.053
MCV	697	0.07	0.032	0.03
RBC	698	-0.11	0.033	7.1×10^{-4}

Figure 21. Mendelian randomization (MR) analysis of plasma L-glutamine with sickle cell disease (SCD) painful crises. Forest plot of MR evaluating the causal relationship between plasma L-glutamine levels and painful crises in SCD patients. Effect sizes and standard errors of 51 variants associated with plasma L-glutamine were retrieved from large European mGWAS. Associations statistics between these 51 variants and SCD complications were calculated in the large prospective and well-characterized CSSCD. In model 1, we considered all 51 SNPs as instruments, whereas model 2 only included 27 variants not associated with other metabolites (Methods). The MR effect size estimates and 95% confidence



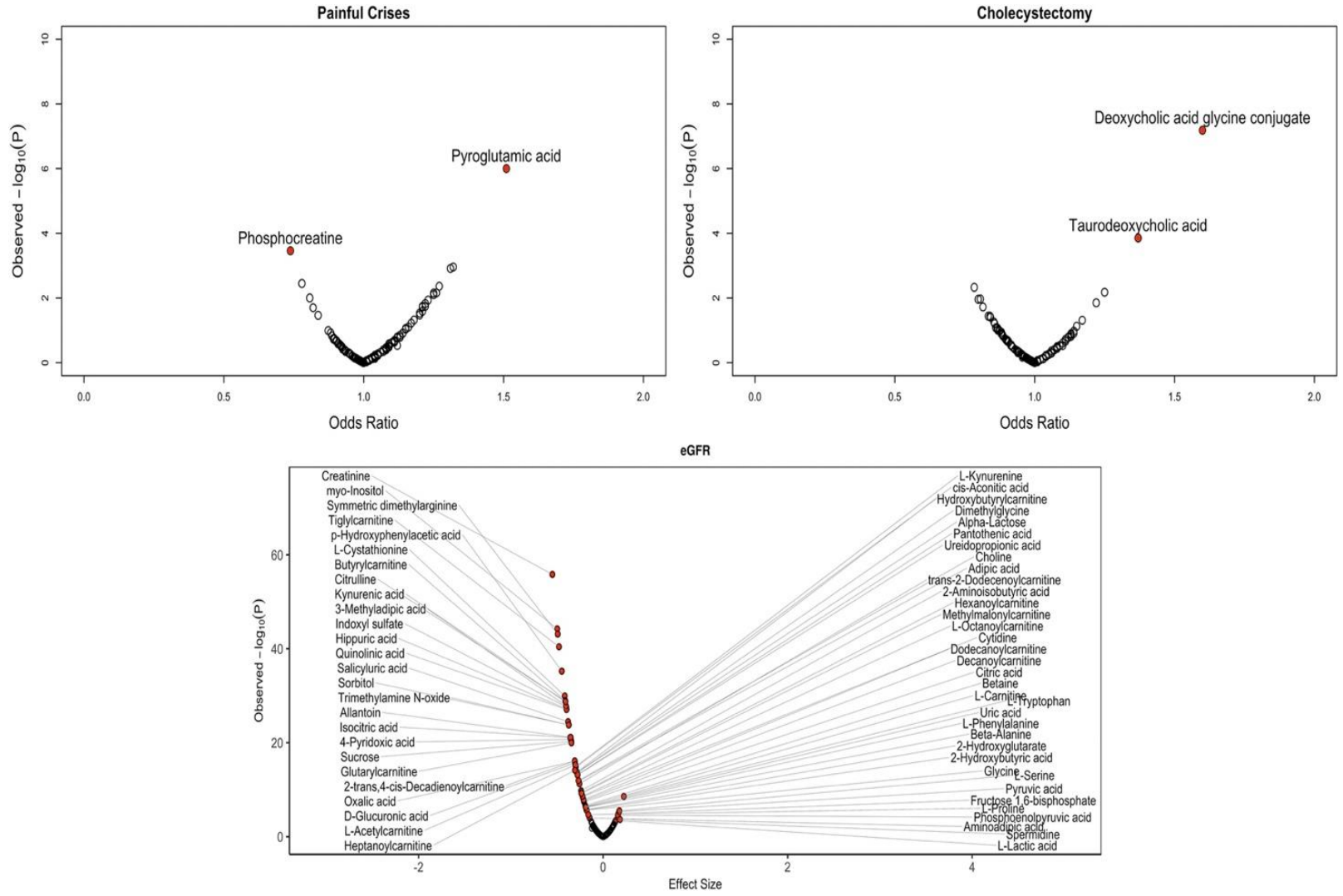
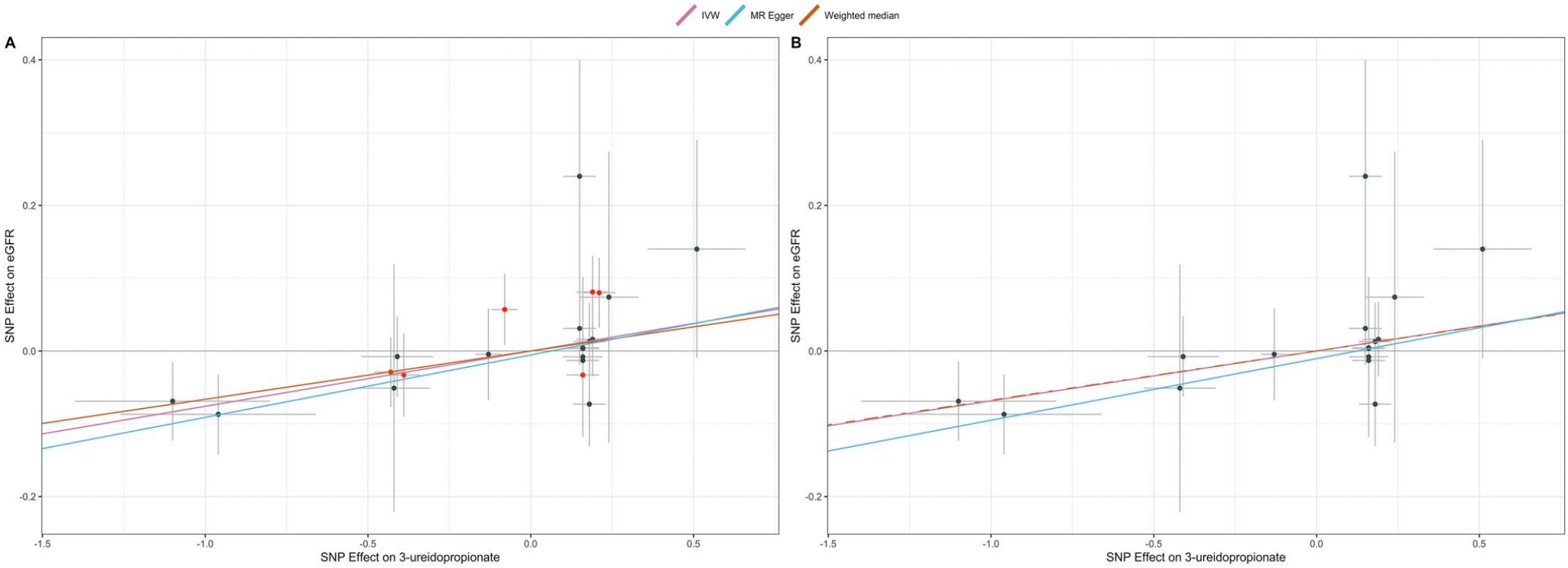


Figure 22 Known metabolites associated with SCD complications and estimated glomerular filtration rate (eGFR) in GEN-MOD and OMG. We tested 129 metabolites against clinical complications by logistic regression (linear regression for quantitative eGFR). On the x-axis, we report odd ratios (effect sizes for eGFR) in metabolite standard deviation units, whereas the y-axis presents the observed analytical P-values. Red circles highlight metabolites with Pperm < 0.05 calculated using 100,000 permutations. In total, we found 2 metabolites for painful crises, 2 metabolites with cholecystectomy, and 62 metabolites for eGFR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Figure 23. 3-Ureidopropionate causally influences estimated glomerular filtration rate (eGFR) in sickle cell disease (SCD) patients. (A) Mendelian randomization (MR) plot comparing the effects of SNPs on 3-ureidopropionate in Europeans (retrieved from large European mGWAS) (x-axis) and eGFR in SCD patients (CSSCD) (y-axis). The slope of each line corresponds to the MR effect for each method (inverse variance-weighted (IVW), MR-Egger or weighted median). Data are expressed as effect sizes with 95% confidence intervals. SNPs in red are pleiotropic. (B) Same as A, except that we removed pleiotropic variants. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.


IVW
MR Egger
Weighted median
A
B
SNP Effect on eGFR
SNP Effect on 3-ureidopropionate
0.4
0.2
0.0
-0.2
-1.5
-1.0
-0.5
0.0
0.5

Chapter 5: Integrating metabolomics with GWAS reveals novel insights into the liver and kidney dysfunction in sickle cell disease patients

The article presented is intended to be published to the journal, *Scientific Reports*. In this article, employing both targeted and untargeted approaches we profiled the plasma of 706 SCD patients using liquid chromatography tandem mass spectrometry. The cohort included 406 French patients (GEN-MOD cohort) of recent African descent and 300 African Americans (OMG cohort) from southeastern US. In total, we measured the levels of 233 known and 1,880 unknown metabolites. I constructed 66 modules containing at least 7 metabolites per module using the clustering framework weighted correlation network analysis (WGCNA). I found a module strongly associated with increased risks of gall bladder removal. Additionally, I retrieved another module of metabolites strongly correlated with a measure of kidney function, namely estimated glomerular filtration rate (eGFR). Finally, I performed a GWAS for each of the 39 most robust modules, which resulted in two modules strongly associated with SNPs (FDR < 0.05). I obtained one module with multiple SNPs significantly associated ($P < 8.0 \times 10^{-10}$) near the gene encoding for hepatic triglyceride lipase (*LIPC*).

Integrating metabolomics with GWAS reveals novel insights into the liver and kidney dysfunction in sickle cell disease patients.

Yann Ilboudo[1], Melanie Garrett[2], Aurelie Guilbault[1], Allison Ashley-Koch[2], Marilyn Telen[3], Guillaume Lettre[4]

Affiliations

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

[2]Center for Human Disease Modeling, Duke University Medical Center, Durham, North Carolina, United States of America

[3]Department of Medicine, Division of Hematology, Duke University Medical Center, Durham, North Carolina, United States of America.

[4]Montreal Heart Institute, Montréal, Québec, Canada; Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada

Correspondence

Guillaume Lettre

Montreal Heart Institute

5000 Bélanger St

Montreal, Quebec, Canada, H1T 1C8

514-376-3330 ext. 2660

guillaume.lettre@umontreal.ca

Keywords: GWAS, sickle cell disease, metabolites, PheWAS, WGCNA

ABSTRACT

Sickle cell disease (SCD) is a monogenic disease caused by a mutation in the β-globin gene. The complications related to the disease are systemic as they impact multiple organ systems. Our goal in this study was to identify metabolome changes contributing to SCD-related severity. Employing both targeted and untargeted approaches, we profiled the plasma of 706 SCD patients using liquid chromatography-tandem mass spectrometry. The cohort included 406 French patients of recent African descent and 300 African Americans from the southeastern US. We applied weighted gene correlation network analysis (WGCNA) algorithms to account for correlations among metabolites and identify specific metabolomic clusters associated with SCD-related complications, blood indices and renal function. Finally, we incorporated genetic data from 15 million SNPs into the clusters to identify the biological pathways implicated by the unknown metabolites. We constructed 66 networks containing at least seven metabolites per network. We found a group of metabolites strongly associated with increased risks of gall bladder removal. That module contained four known metabolites involved in bile acid metabolism, including glycocholate, glycodeoxycholate, taurocholate, and taurodeoxycholate. Additionally, we found another group of metabolites strongly correlated with estimated globular filtration rate (eGFR). The cluster implicated molecules in the carboxylic acid metabolic process and the purine pyrimidine metabolism. We performed a GWAS for each of the 39 most robust modules, which resulted in two modules strongly associated with SNPs at genome-wide level (FDR < 0.05). We found that one of these two modules was significantly associated ($P < 8.0 \times 10^{-10}$) with multiple SNPs near the gene encoding for the hepatic triglyceride lipase (*LIPC*). This association reiterates findings on SCD severity and the impaired Land's cycle. Functional or computational experiments to identify the unknown metabolites could yield a better understanding of how these metabolic pathways play a role in organ dysfunction and then can be exploited in therapeutic intervention.

INTRODUCTION

Sickle cell disease is a debilitating genetic disorder caused by a single mutation at the beta-globin gene (Glu6Val). The mutation induces the formation of long polymers under hypoxic conditions. A build-up of these rod-like polymers leads to the deformation of red blood cells leading to vaso-occlusion, anemia, and eventually organ damage. This inherited blood disorder is one of the most widespread monogenic diseases in humans. Most cases impact individuals of African, Indian, and Arab descent[1-3]. The disease is predominantly present in West Africa, where at least 25% of the population carries the mutation. Countries such as Nigeria and the Democratic Republic of Congo[4] have the most considerable burden. Metabolites represent the integration of gene expression, protein interaction, other regulatory processes, and the environment. Therefore, they are ideal for understanding and tracking by-products of the physiological progression of diseases. Also, metabolomics can yield insights into ground-breaking therapies targeting specific pathways or enzymes[5,6]. Therefore, measuring metabolites in SCD patients can help us uncover the pathophysiological mechanisms and treatments for sickle cell disease.

Metabolomics studies in transgenic knockout mice with human sickle hemoglobin and sickle cell disease[7] implicated two metabolic pathways and six families of metabolites. The studies showed a pronounced increase in lipids, amino acids, nucleotides, fatty acids, carbohydrates, xenobiotics, glycolysis-related molecules (2,3-bisphosphoglycerate), and the pentose phosphate pathway[8-10]. Subsequent studies by Darghouth *et al*. confirmed the mouse study results[11]. By comparing the metabolome of 24 healthy individuals with that of 28 SCD patients, they found that higher levels of 2,3-bisphosphoglycerate (2,3-BPG)) and a reduction of glutathione are characteristics of SCD progression. Moreover, the discovery of two independent pathways modulating sickle severity further improved our understanding of the disease. On the one hand, the damaged Land's cycle[10], responsible for the plasma membrane's stability, can be a therapeutic target since inhibiting the phospholipase A2 (*PLA2*) enzyme through a small interfering RNA reduces the sickling phenomena[12].

On the other hand, targeting the CD73–ADORA2B and SphK1–S1P–S1PR1–IL-6 signaling cascades will reduce adenosine levels which will, in turn, reduce inflammation, sickling, and painful crises events. An FDA-approved compound targeting this signaling pathway proved beneficial in SCD mice as it reduced morbidity rate, lowered systemic inflammation, and

lessened tissue damage[13]. While more recent studies linked metabolites such as asymmetric dimethylarginine, quinolinic acid, and L-glutamine to estimated globular filtration rate (eGFR)[14] and painful crises[15,16], other analyses consistently implicate the role of Land's cycle, S1P[17], and the arginine and glycolysis pathways[18,19]. Finally, Mitapivat[20], a red blood cell-specific pyruvate kinase targeting the 2,3-BPG and glutathione pathway, is currently under phase 3 clinical trial as a potential therapeutic for SCD.

In this study, we want to inquire how groups of metabolites contribute to SCD pathophysiology. First, employing weighted gene co-expression network analysis (WGCNA)[21], we constructed networks of metabolites by calculating the dissimilarity coefficients and subsequently performing by hierarchical clustering. Second, we calculated the correlation between these networks and SCD-relevant complications. Finally, we integrated the genotypic information into the metabolite clusters by performing genome-wide association analyses on robust modules.

MATERIALS AND METHODS

Ethics statement

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

Samples and DNA genotyping

The Genetic Modifier (GEN-MOD) and the Duke University Outcome Modifying Genes (OMG) were recruited in France and the southeastern USA, respectively. Both cohorts' genotyping, imputation, and plasma collection were described elsewhere[14,15,22].

Metabolite profiling, pre-processing, and quality control

Metabolomic profiling employed tandem liquid chromatography with mass spectrometry. Sample preparation, data acquisition, metabolite identification, bath effect correction, metabolite imputation, and normalization are described elsewhere[15]. Employing both targeted and untargeted approaches, we profiled the plasma of 706 SCD patients; the cohort included 406 French patients (GEN-MOD) of recent African descent and 300 African Americans (OMG) from the southeastern US. We measured the levels of 233 known and 1,880 unknown metabolites. We considered metabolites as unknown when they could not be reliably measured and quantified and whose identity could not be determined or is absent from profiling libraries. Additionally, we developed an R package VIQCing (visualization, imputation, quality control, wgcna functions) implemented in the R language. Finally, we documented all the functions and methods employed (**see URLs**).

Weighted gene co-expression network analysis

We analyzed both targeted and untargeted metabolites for the weighted gene co-expression analysis (WGCNA). The smallest soft threshold with an adjusted $R^2 > 0.85$ was 7 (**Supplementary Figure 2**). We, therefore, chose to calculate the adjacency score between any

seven metabolites within a sample set. We then computed the topological overlap matrix (TOM), which we converted to a distance matrix (adjacency matrix) by subtracting the matrix values from 1. Finally, we identified the modules by applying hierarchical clustering to the adjacency matrix. We set minModuleSize 7, mergeCutHeight 0.25, deepSplit 2, networkType "signed"for the WGCNA analysis. We, therefore, generated 66 modules. We employed the WGCNA (v1.12.0)[21], implemented in R, to perform all the previously mentioned analyses.

Identification of modules associated with clinical phenotypes

Module-clinical associations were determined using logistical and linear regression for six categorical (leg ulcers, retinopathy, cholecystectomy, priapism, aseptic necrosis, survival) and twelve continuous variables (mean cell volume, mean corpuscular hemoglobin concentration, mean corpuscular hemoglobin, hematocrit, platelets, reticulocyte count, fetal hemoglobin, red blood cell count, white blood cell count, lactate dehydrogenase, bilirubin, and estimated globular filtration rate). Continuous traits were ranked inverse normal transformed. We tested the association of the clinical phenotypes with the module eigengene - the first principal component - of metabolites within a module. All associations were adjusted for age, sex, SCD genotypes, and hydroxyurea usage. We considered significant all modules with a *P-value* < 4.5×10^{-5} (Bonneferoni 0.05/ (66 x 18 (number of modules times the number clinical phenotypes)).

Robustness analysis, genome-wide association, and phenome-wide association study

To assess the reproducibility of the networks, we performed a robustness analysis in which we randomly sampled half of the individuals in the cohort. We then generated modules and calculated their preservation statistics (module interconnections, separability, and Z statistic). We repeated this analysis eight times and kept all modules with a Z-score statistic greater than 10 (the threshold by which modules are considered reproducible, distinct, tightly connected, and robust). As a result, we kept 39 robust modules.

In terms of genotypes, we restricted our analysis to markers with imputation quality r^2 >0.3 and minor allele frequency (MAF) >1%. Upon performing the association test, we removed the effect of age, sex, ancestry (first ten principal components), SCD genotype, and hydroxyurea usage on inverse normal transformed principal component 1 (PC1) of each 39 modules. We used RvTests (v20171009)[23] to test the association between genotype dosage and PC1 in GEN-MOD. Next,

we used SNPTESTv2[24] to test the association between genotype dosage and PC1 in OMG. We employed METAL[25] to then meta-analyze the summary statistics from each cohort. Over 15 million variants resulted from the final step. Finally, we performed the PheWAS analysis for blood indices employing the variant level approach implemented in the variant annotation method pointing to interesting regulatory effects (VAMPIRE)[26] and the gwasATLAS[27] website.

RESULTS

Gene co-expression modules associated with blood traits and SCD complications

After quality control, we generated 2,113 metabolites from 688 SCD patients. Quality control steps were performed as described in[15]. By employing WGCNA on 2,113 metabolites, we constructed 66 modules, among which the number of metabolites ranged from 7 to 304 (**Supplementary Table2**) (**Figure 1 A**). We then tested the association of modules with eleven blood traits, one renal parameter, and six SCD-related complications. We found 43 modules with $P < 4.2 \times 10^{-5}$ (0.05/(number of module times the traits tested)). One module, Lightcyan1, was significantly ($OR = 2$, $P = 1.5 \times 10^{-6}$) and specifically associated with cholecystectomy. Upon further inspection, we found that the module includes four known metabolites (glycocholate, glycodeoxycholate, taurocholate, and taurodeoxycholate), all involved in bile acid metabolism, and 20 unknown metabolites (**Supplementary Table 2**, **Figure 2 A**). Plus, we noticed a strong association between the magenta module and eGFR levels (*beta* = -0.6, *P-value* = 1.5 x 10^{-61}). The magenta cluster contains 33 known metabolites and 114 unknown metabolites (**Supplementary Table 3)**. Among the known metabolites, we find well-documented compounds such as citrulline, creatinine, and dimethylarginine[28,29]. We also found novel metabolites such as quinolinic acid[14] and N2,N2-Dimethylguanosine[30]. Additionally, close to the significance threshold, Coral1 associated with leg ulcers ($OR = 0.6$, $P = 7.4 \times 10^{-5}$). No other module significantly associated with SCD-complications.

We noticed that the same module could be associated with several continuous traits (i.e., the Lightcyan module is associated with, among others, MCV, hematocrit, and eGFR). In contrast, some modules are more precise (i.e., the magenta and the brown modules are specific to eGFR) (**Figure 1 B & C**). Due to the specificity and significance of association of the magenta module

with eGFR, we explored whether the module associated with renal decline in a subset of the OMG cohort. We found an inverse relationship, although not significant, between the metabolites within the magenta and renal decline in OMG cohort (**Figure 3**).

We found several modules containing unknown metabolites. For example, the lightcyan module contains 78 metabolites, 3 of which are known (**Supplementary Tables 2 and 3**). We further prioritized the associations of modules with blood traits by looking at the relationship between module membership and metabolite significance. We considered as strong an association when the correlation coefficient is above 0.7 and the significance *P-value* < 0.05/(number of metabolites in the module times 18 (the number of traits)). The robustness analysis based on connectivity strength confirmed the association between cholecystectomy and lightcyan1 and the association between eGFR and the magenta module (**Figure 2**). Furthermore, while the grey module (the module that receives metabolites that could not be confidently assigned to any clusters) was associated with eGFR, the module strength connectivity showed weak correlation and significance (r=0.18, p-value=0.0039) (**Supplementary Figure 10**).

Genome-wide association study identifies lipid-related a pathway linked to hematocrit and red blood cell count

Several of the modules constructed contain both known and unknown metabolites. Since combining high-throughput genotyping data with metabolomics can reveal the functional genetic loci linked to unknown metabolites[31,32], we performed a GWAS for PC1 for each of the 39 most robust modules (**Supplementary Figure 14**). As shown in **Figure 4**, we found that the darkorange and darkred modules were significantly associated ($P < 8.0 \times 10^{-10}$) with multiple SNPs near the genes encoding for hepatic triglyceride lipase (*LIPC*) and the embryonic ectoderm development (*EED*), respectively (**Table 1**). The darkorange module contains six known metabolites and fifteen unknown metabolites. The known metabolites are all phosphatidylethanolamines (PE(16:0/18:2(9Z,12Z)), PE(16:0/20:4(5Z,8Z,11Z,14Z)),PE(16:0/20:4(5Z,8Z,11Z,14Z)),PE(18:2(9Z,12Z)/20:4(5Z,8Z,11Z,14Z)),PE(18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))), belonging to the class of glycerophospholipids. The darkred module, on the other hand, contains 22 metabolites, all unknowns. The SNPs for both modules are in strong linkage disequilibrium ($R^2 > 0.9$) in the African ancestry population.

Phenome-wide association analysis (PheWAS) for *LIPC* and *EED* SNPs

We performed a PheWAS for the top SNP identified in **Table 1**. Given they are all in LD, we queried the most significant SNPs associated with both modules. For the darkorange module we queried rs1800588, and for the darkred cluster we queried rs12277271. We focused our query on seven red blood cell indices (HCT-Hematocrit; HGB-Hemoglobin Concentration; MCH-Mean Corpuscular Hemoglobin; MCHC- Mean Corpuscular Hemoglobin Concentration; MCV-Mean Corpuscular Volume; RBC-Red Blood Cell Count; RDW-Red Blood Cell Distribution Width) from three recent articles on the GWAS of hematological traits across ancestry[33-35]. We also queried SNPs across 3302 human phenotypes (https://atlas.ctglab.nl/)[27] with P-value < 5×10^{-5}. We found four significant eQTL associations for rs1800588 implicating *LIPC* with hematocrit levels, red blood cell counts, and no other blood indices (**Supplementary Data Table1**). We also found an enrichment of associations with lipid-related metabolic traits (HDL-related traits, LDL-related traits, and cholesterol-related traits) (**Supplementary Data Table 1 & 2**). For rs12277271, we found no association with any of the blood indices. However, two associations were related to asthma (**Supplementary Data Table 3**).

DISCUSSION

To the best of our knowledge, this is the first study to explore the interplay between the genetic, blood traits, and complications by integrating GWAS data into metabolomics in SCD. We applied the WGCNA framework, a network approach, to identify and integrate genotypes and modules of coregulated metabolites that correlate with SCD-related clinical endpoints. Specifically, this includes metabolites such as glycocholate, glycodeoxycholate, taurocholate, and taurodeoxycholate, which associate with cholecystectomy and have been shown to increase post-surgery[36,37]. Additionally, metabolites associated with eGFR in the magenta module seem to correlate with renal decline in a subset of SCD patients. While the p-value is not significant, most likely due to sample size (N=82), the correlation coefficient (beta=-0.87) suggests a link between this metabolite cluster and kidney function. Several other modules are strongly associated with blood traits and eGFR; they, however, encapsulated primarily unknown metabolites (e.i, the brown cluster association with eGFR, the bisque4 cluster association with white blood cell count, and the lightyellow association with reticulocyte count). Thus, rendering

their interpretation more difficult. Further investigation is warranted to identify these metabolites employing in silico[38] or in vivo[39] approaches.

Plus, combining metabolomics with GWAS data allowed the exploration of red blood cell mechanisms in sickle cell disease. Our findings support the evidence for the role of glycerophospholipids SNP that influences *LIPC* expression, rs1800588 (Supplementary Data Table1). Plus, this finding echoes the contribution of phospholipase A_2 (*PLA2*) and lysophospholipid acyltransferases (*LPLATs*) to the impaired Lands' cycle in mice and cultured human erythrocytes[12]. In fact, our HMDB annotation of the known darkorange metabolites showed that they are all associated with several phospholipases A_2 (i.e., *PLA2G5; PLA2G2F; PLA2G4A*) (**Supplementary Table 3**). Other studies on cholesterol and lipoproteins in SCD showed an association between increased triglyceride/HDL-C ratio and endothelial dysfunction[40]. Moreover, triglycerides were lower in SCD patients compared to healthier individuals[41,42].

This study reveals the power of clustering approaches in combining omics datasets in a biologically relevant fashion. It also shows that network-based methods are helpful in generating hypotheses to pinpoint the most relevant component of highly dimensional datasets. Another advantage of the network approach is performing GWAS on principal components (PC). PCs enable us to summarize information from multiple highly correlated traits (like metabolites). Furthermore, GWAS on PC scores can decrease type 1 error rate by avoiding multiple testing[43,44]. Finally, combining traits through PC scores could discover regions missed by individual phenotypes.

There are some limitations to this study. While our sample size is large (N=688) within the context of sickle cell disease and African ancestry individuals, it is far from the large metabolite analysis performed in European ancestry individuals ($N > 7{,}000$)[45]. Also, only 11% of the metabolite profiled could be considered as known. This limited the interpretability and identification of novel pathways. Plus, including only known metabolites in our WGCNA analysis would prevent us from generating multiple clusters, thus limiting our scope. Finally, our clustering results require replication in an independent SCD cohort. The absence of a well-powered metabolomics study in another SCD cohort makes replication difficult, if not impossible. Accordingly, the present results should be interpreted with caution.

This study correlated metabolites clusters to SCD complications, blood traits, and renal function, in addition to integrating these clusters with genotypes to detect underlying pathways/genes.

Our results suggest a possible contribution of phospholipase A_1 to Land's impaired cycle and highlight the potential of integrative omics analyses to resolve the complexity of disease biology.

URLs

R package for running QC and WGCNA analysis: https://github.com/yilboudo/VIQCing

R markdown WGCNA analysis: WGCNA Rmarkdown SCD

SUPPLEMENTARY TABLE

-S1.SCD_WGCNA_module_membership

-S2.metabolite complication association results

-S3.metabolite blood trait association results

Supplementary data table: **Supplementary Table 1-3**

ACKNOWLEDGMENTS

We thank all participants for their contribution to this project. G.L. is funded by Sanofi, the Canadian Institutes of Health Research (CIHR, MOP #123382), and the Canada Research Chair program.

AUTHOR CONTRIBUTIONS

G.L. and M.T conceived and designed the experiments; Y.I. performed bioinformatics and statistical analysis; M.B. contributed DNA samples, clinical information, and expert knowledge; Y.I. and M.G. analyzed the results; Y.I. wrote the manuscript with contributions from all authors. Y.I., A.G., M.G. wrote R code.

CONFLICT OF INTEREST

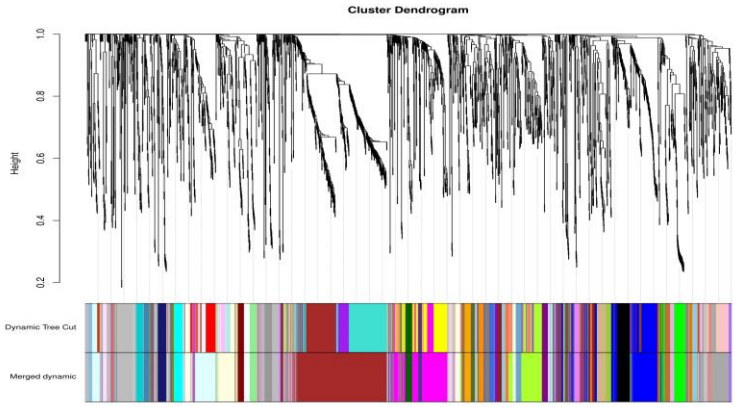
The authors declare no competing financial interests.

Table 1. Genome-wide significant hits. Association results of modules and genotype data were performed for GENMOD and OMG (*P-value* < 5 x 10^{-8}). We tested each 39 robust modules against genotypes accounting for sex, hemoglobin genotype, and hydroxyurea status within the OMG cohort. Within GENMOD, we also tested each 39nrobust module against genotype accounting for age and sex. We then meta-analyzed results from both cohorts using METAL. Abbreviations: rsID: SNPs ID, CHR: chromosome; POS: base pair position; REF: reference allele; ALT: alternate allele; Z-score: effect size resulting from meta-analysis; PVAL: P-value for association of SNP; FDR PVAL: false discovery rate of P-value of SNP.

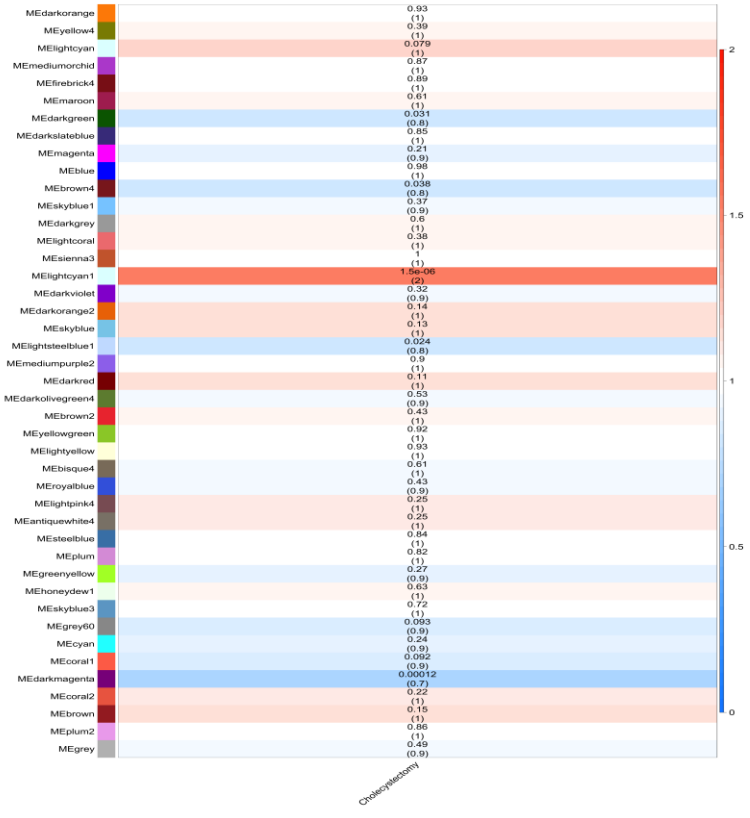
Nearest gene symbol	rsID	CHR	POS(hg37)	POS(hg38)	REF/ALT	N	Z-score	PVAL	FDR PVAL	Module
ALDH1A2, *LIPC*	rs1800588	15	58723675	58431476	T/C	637	-6.28	3.48 x 10^{-10}	0.0025	Darkorange
ALDH1A2,*LIPC*	rs2070895	15	58723939	58431740	A/G	637	-6.27	3.70 x 10^{-10}	0.0025	Darkorange
ALDH1A2,*LIPC*	rs1077835	15	58723426	58431227	A/G	637	6.18	6.61 x 10^{-10}	0.0025	Darkorange
ALDH1A2,*LIPC*	rs1077834	15	58723479	58431280	T/C	637	6.18	6.61 x 10^{-10}	0.0025	Darkorange
EED	rs12277271	11	85881955	86170913	T/C	637	-6.0	1.99 x 10^{-09}	0.015	Darkred
EED	rs12271958	11	85882691	86171649	A/C	637	6.0	1.99 x 10^{-09}	0.015	Darkred

Figure 24 The weighted gene correlation network analysis (WGCNA) for 688 SCD patients. (A) Dendrogram of all metabolites clustered based on dissimilarity measurement (1-TOM). The color band shows the results obtained from the automatic single-block analysis. In total, 66 metabolite modules were constructed. (B) top module trait association with blood traits. (C) top module trait association with complications. Each row corresponds to a module eigenmetabolite, column to a blood trait/complication. Each cell contains the corresponding beta/odds ratio and p-value. The table is color-coded by correlation according to the color legend.

A)


Cluster Dendrogram
Height
1.0
0.8
0.6
0.4
0.2
Dynamic Tree Cut
Merged dynamic

B)


MEdarkorange
MEyellow4
MElightcyan
MEmediumorchid
MEfirebrick4
MEmaroon
MEdarkgreen
MEdarkslateblue
MEmagenta
MEblue
MEbrown4
MEskyblue1
MEdarkgrey
MElightcoral
MEsienna3
MElightcyan1
MEdarkviolet
MEdarkorange2
MEskyblue
MElightsteelblue1
MEmediumpurple2
MEdarkred
MEdarkolivegreen4
MEbrown2
MEyellowgreen
MElightyellow
MEbisque4
MEroyalblue
MElightpink4
MEantiquewhite4
MEsteelblue
MEplum
MEgreenyellow
MEhoneydew1
MEskyblue3
MEgrey60
MEcyan
MEcoral1
MEdarkmagenta
MEcoral2
MEbrown
MEplum2
MEgrey
Cholecystectomy
2
1.5
1
0.5
0

C)

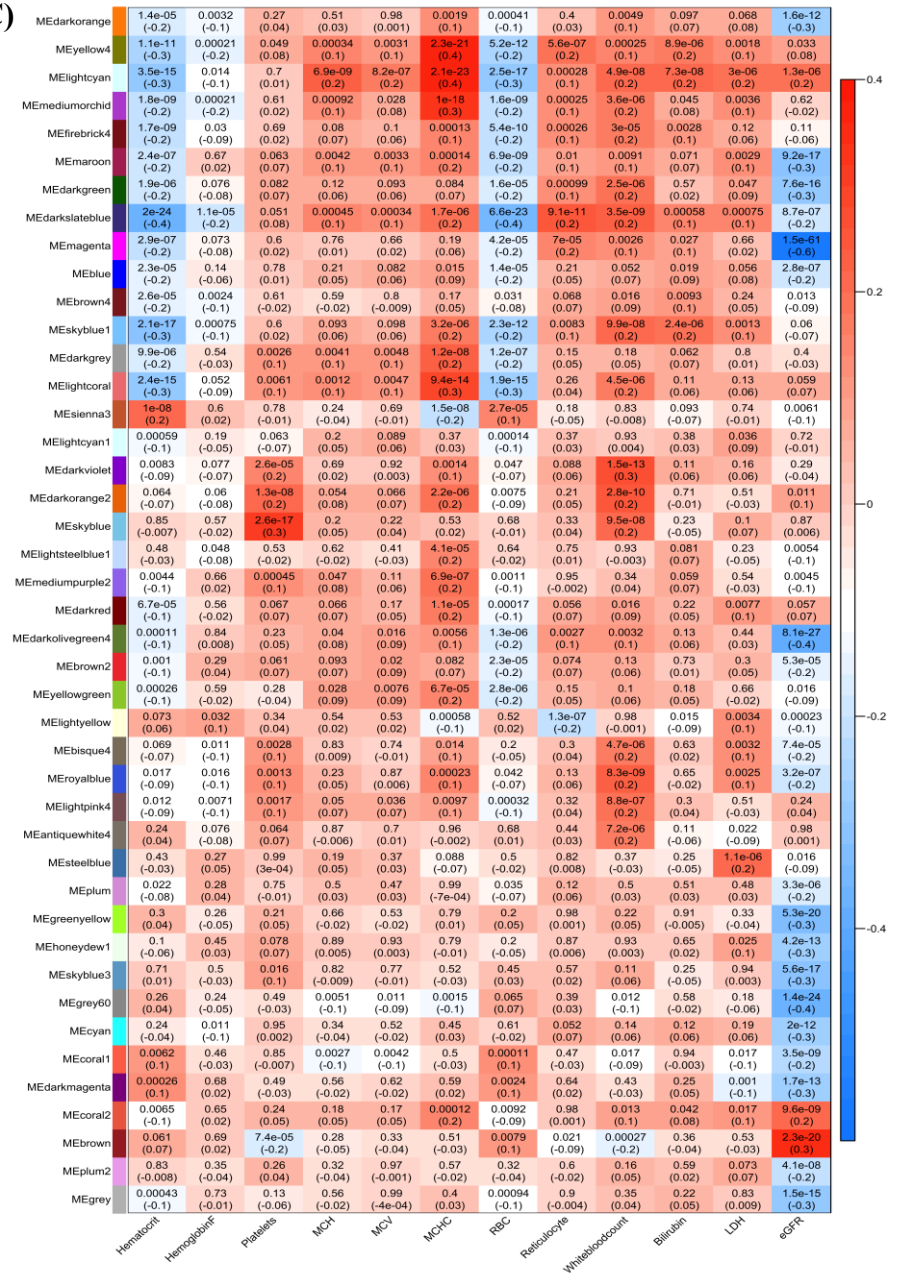

MEdarkorange
MEyellow4
MElightcyan
MEmediumorchid
MEfirebrick4
MEmaroon
MEdarkgreen
MEdarkslateblue
MEmagenta
MEblue
MEbrown4
MEskyblue1
MEdarkgrey
MElightcoral
MEsienna3
MElightcyan1
MEdarkviolet
MEdarkorange2
MEskyblue
MElightsteelblue1
MEmediumpurple2
MEdarkred
MEdarkolivegreen4
MEbrown2
MEyellowgreen
MElightyellow
MEbisque4
MEroyalblue
MElightpink4
MEantiquewhite4
MEsteelblue
MEplum
MEgreenyellow
MEhoneydew1
MEskyblue3
MEgrey60
MEcyan
MEcoral1
MEdarkmagenta
MEcoral2
MEbrown
MEplum2
MEgrey
Hematocrit
Hemoglobin
Platelets
MCH
MCV
MCHC
RBC
Reticulocyte
Whitebloodcount
Bilirubin
LDH
eGFR
0.4
0.2
0
-0.2
-0.4

Figure 25 Identification of modules related to the clinical red blood cell traits and SCD complications. (A-B) Scatter plots of metabolite significance (*MS*) for the risk of cholecystectomy and eGFR levels vs. lightcyan1and magenta module membership (MM). (C-D) Visualization of the 30 most highly connected metabolites network connections the lightcyan1 and magenta modules whose topological overlap is above the thresholds of 0.02.

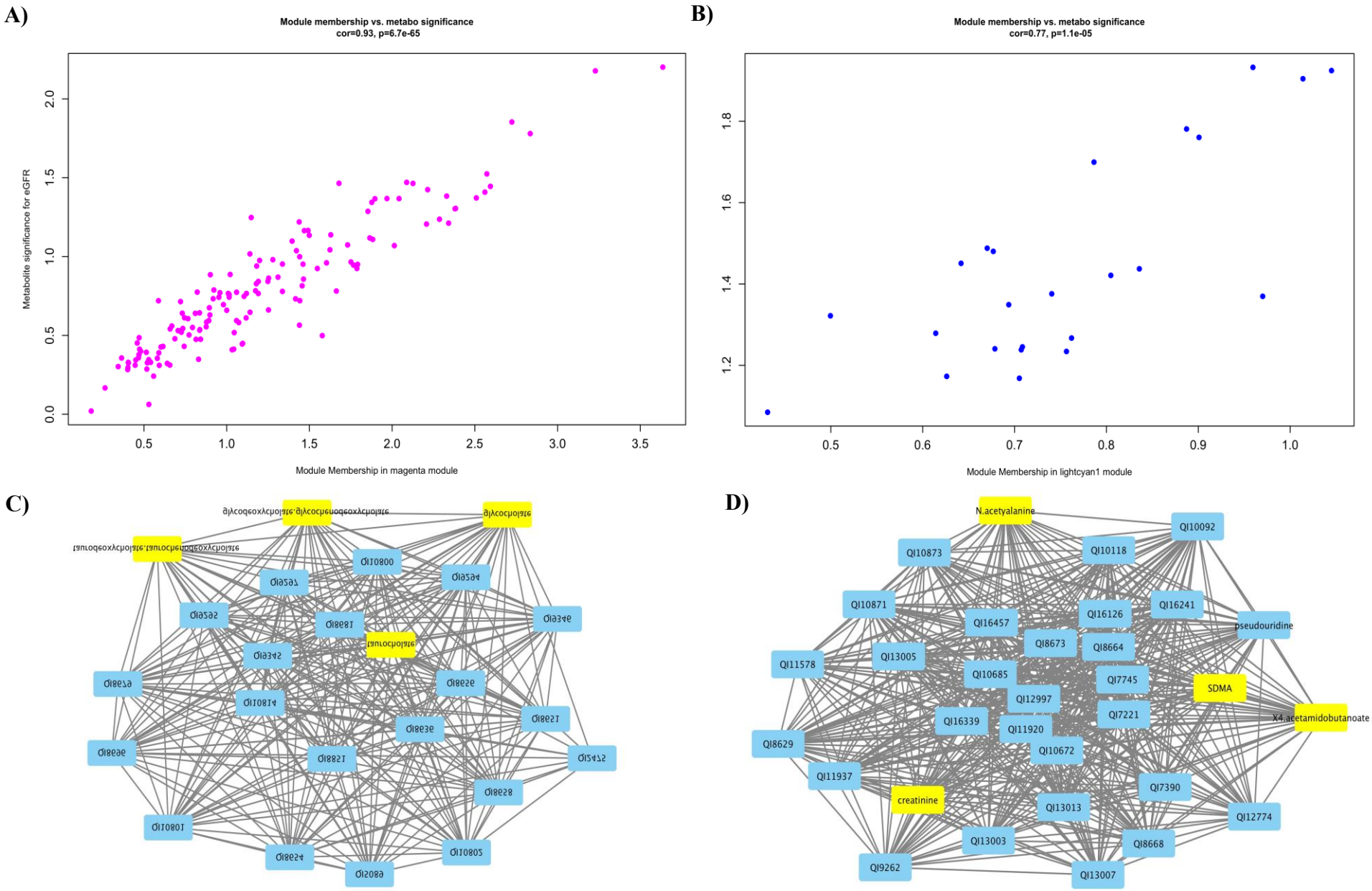

A)
Module membership vs. metabo significance
cor=0.93, p=6.7e-65
Metabolite significance for eGFR
Module Membership in magenta module
B)
Module membership vs. metabo significance
cor=0.77, p=1.1e-05
Module Membership in lightcyan1 module
C)
D)
N.acetylalanine
QI10092
QI10873
QI10118
QI10871
QI16241
QI16457
QI16126
pseudouridine
QI8673
QI8664
QI11578
QI13005
QI10685
QI7745
SDMA
QI12997
QI16339
QI7221
X4.acetamidobutanoate
QI11920
QI8629
QI10672
QI11937
QI7390
creatinine
QI13013
QI12774
QI13003
QI8668
QI9262
QI13007

Figure 26 Scatter plot of the rapid decline of kidney function and PC1 of the magenta module in 82 OMG patients. The association between rapid decline in kidney function and with magenta module highlights that the higher the values of the magenta module, the steeper the renal decline.

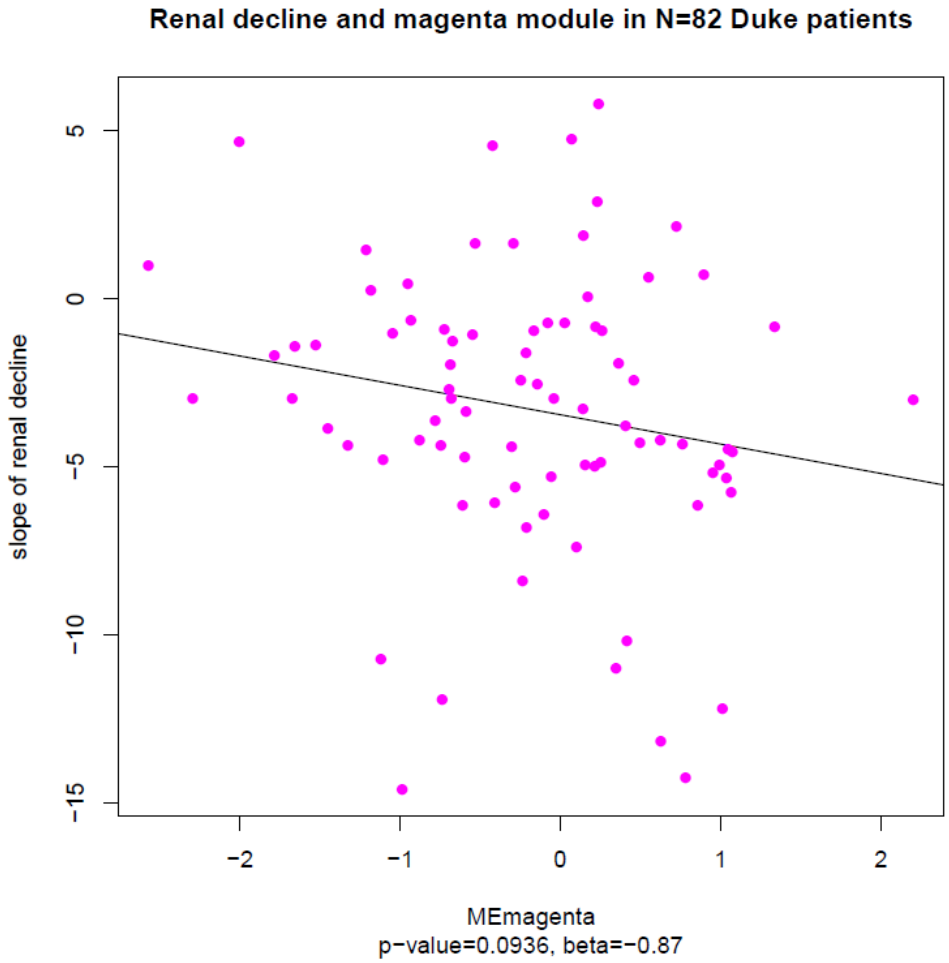
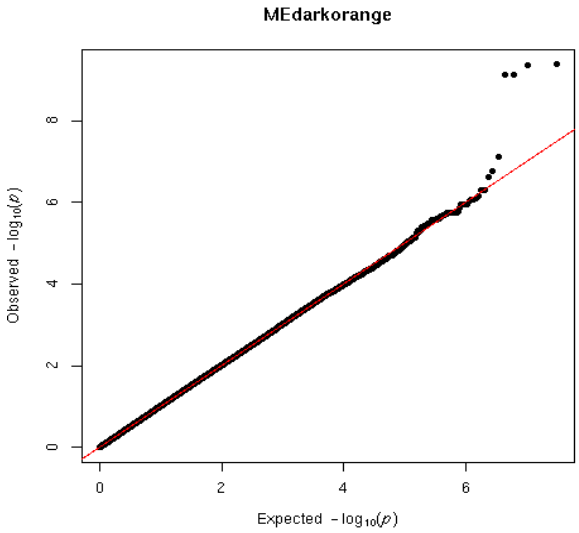
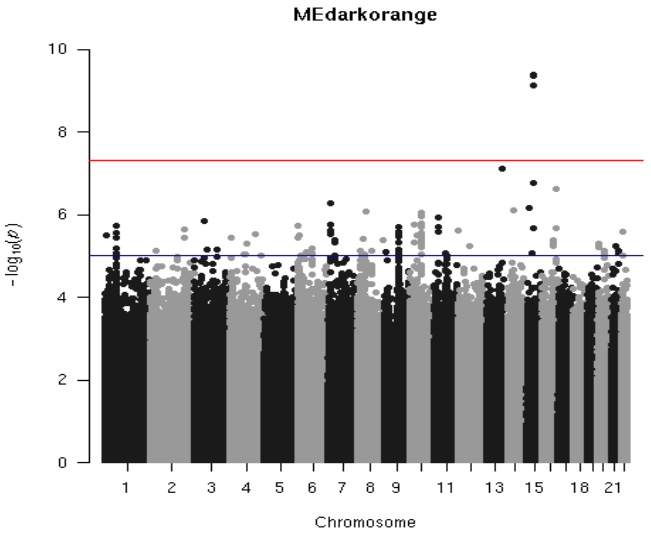
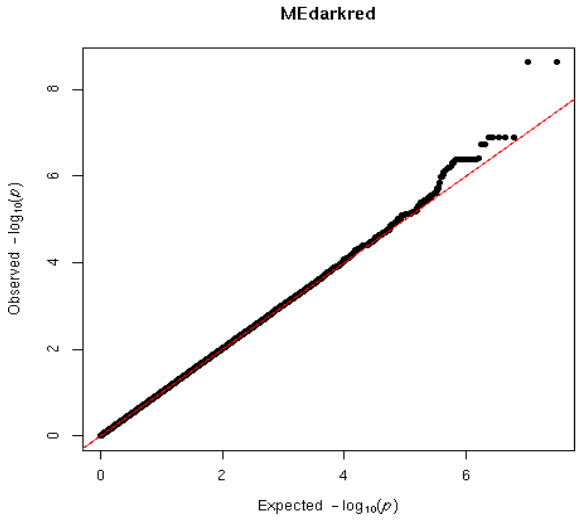
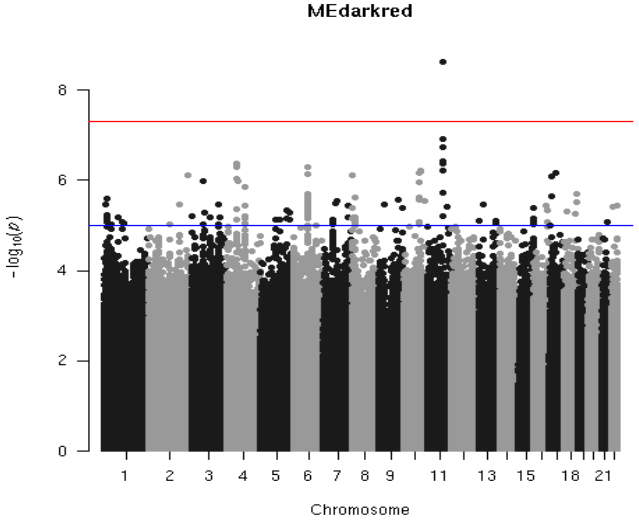


Figure 27 Module GWAS manhattan and QQplot. A & B Manhattan and QQ plot for darkorange module GWAS. C & D Manhattan and QQ plot for darkred module GWAS. The significant cut-off is *P*-value $<5 \times 10^{-8}$ (red line in Manhattan plot), the sub genome-wide cut off is P-value $< 5 \times 10^{-6}$ (blue line in Manhattan plot). Associations for each module PC1 were performed within each cohort (GENMOD and OMG). GWAS in GENMOD employed a linear regression implemented in RVTEST adjusting for age and sex. GWAS in OMG used a linear regression implemented in SNPTEST adjusting for age, sex, hemoglobin genotype, and hydroxyurea usage status. Meta-analysis of summary statistics combined over 15 million SNPs with MAF > 1% and employed METAL.


MEdarkorange
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
2
4
6
8


MEdarkorange
$-\log_{10}(p)$
Chromosome
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
11
13
15
18
21


MEdarkred
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
2
4
6
8


MEdarkred
$-\log_{10}(p)$
Chromosome
0
2
4
6
8
1
2
3
4
5
6
7
8
9
11
13
15
18
21

Chapter 6. Discussion

Summary of thesis

With three new FDA-approved drugs in a three-year span[99], 210 actively recruiting clinical trials in sickle cell diseases as of October 2022 on clinicaltrials.gov, optimism surrounding gene therapy[282], and recent drug repurposing research[283,284], sickle cell disease research has garnered a lot of keen interest. However, more integrative studies combining multiple omics in large samples with causal frameworks can uncover novel disease mechanisms.

In this thesis, I performed the largest GWAS both for imputed genotypes and exome sequencing on fetal hemoglobin (Chapter 2). My analysis includes a conditional meta-analysis of 9 SCD cohorts and one healthy Europeans cohort (SardiNIA), and an exome-wide association study (ExWAS) of recent African ancestry and admixed African SCD individuals. I demonstrated how melding disease-relevant transcriptomics, chromatin interactions and human genetics data can aid to identify likely causal variant in association studies.

In relation to the density of red blood cells (Chapter 3), I conducted an ExWAS, and an analysis of rare variants. These results were then collated with gene expression, established drug target genes, and association analysis from the UK BioBank to prioritize association signals relevant to traits related red blood cells density.

In Chapter 4, through integrating metabolomics with genomics, I contribute the largest Mendelian randomization study on metabolites and SCD complications to dated. This study on the causal relationship of metabolite in SCD, proposes the effector role of l-glutamine in painful crises and underscores 2 biomarkers of painful crises, 2 biomarkers of gallbladder dysfunction, and 62 biomarkers of kidney functions.

Finally, in Chapter 5, I developed an R package to perform quality control on metabolomics datasets and include wrapper functions for leveraging the clustering framework, WGCNA. This allowed me to uncover a liver-specific and a kidney-specific metabolite network with key drivers involved in bile acid regulation and in rapid kidney function decline. Additionally, this clustering approach implicated a family of lipids previously documented to exacerbate the sickling process.

Limitations of the thesis

Sample size and replication

While sickle cell disease is a monogenic disorder, the patient-to-patient heterogeneity remains a puzzle. Factors such as fetal hemoglobin, DRBC, and metabolites which impact this heterogeneity are complex traits. As a result, multiple noncoding variants modulate these factors making the biological implications of these associations are often unclear, and requiring functional follow-up. Another limitation of this study is the lack for reproducibility. While Chapters 2, 3, and 5 represent the most extensive studies of their kind to date, they lack the replication which is deemed as the gold standard in a well-executed GWAS study. Moreover, the maximum sample size across all my studies involving SCD patients is 3,704, which is relatively modest when compared to the larger GWAS studies conducted in 2022, encompassing millions of samples. Greater sample sizes offer increased statistical power, enabling the detection of smaller effect sizes. To compensate for this shortage of replication and samples, I consulted results from proxy phenotypes and performed a conditional analysis to reduce the variance of known loci, thereby enhancing the association signal from other loci. One recommendation for future studies would be to create of large consortia to sequence individuals at scale. Such an approach has been successful in the past for other disorders[221,285].

Identification of the causal mechanism

Progress in lowering the cost of sequencing, and computation enabled a rapid increase in the amount of biological data that can be generated. As a results GWAS have been used to generate hypothesis for various phenotypes. Yet, identifying the causal signal require sifting through hundreds of variants which could be responsible for the association. As a results, data intensive, and intricate multi-omics functional perturbation follow-up are often necessary to establish causality and gain novel biological insights[127,286,287]. To improve our understanding of complex phenotypes and especially in sickle cell disease research we need to collect higher resolution, multi-omic, multi-time point, multi-tissue, multi-state data set. While Mendelian randomization can also be used to establish causality between an exposure and an outcome, MR relies on several assumptions, including the validity of instrumental variables and the absence of pleiotropy, which may affect the validity of the results. Finally, MR results are not generalizable to other populations or ethnic groups.

Computational considerations

The field of biomedical and life sciences research is confronted with a significant challenge, namely the lack of reproducibility. One of the underlying issues is the inadequate sharing of protocols, code, and data, which hampers exact replication[288]. However, efforts are being made to address these challenges, and there is a growing demand for more open-source research practices. Platforms like bitbucket and GitHub have made it easier for programmers to share and track changes in real time. To facilitate faster dissemination of research, platforms such as arXiv and bioRxiv were established in 1991 and 2013, respectively, enabling researchers to share their work before formal acceptance in peer-reviewed journals[289]. Data sharing, on the other hand, presents more complex considerations, as generating a dataset can take years, and relinquishing it for public access before publication requires careful consideration[290]. Initiatives like the GWAS Catalog, PGS Catalog, and the GTEx[291-293] consortium have allowed data to be downloaded, but an embargo on publication remained until the group had the opportunity to analyze and publish the data[294]. However, even data sharing has limitations, as complete datasets are often not shared, which can hinder the ability to address specific research questions.

Future omics studies in SCD

As more and more investment is being poured into sickle cell disease research, new drug targets and improved understanding of the disease will come to light.

Improved reference genome and structural variations

Employing the new reference genome[295], or the pangenome[296] or exploring the missing heritability in copy number variants[297] in SCD could bring about new fascinating biology. Increased variant coverage, improved representation of population diversity, and annotation of variants can results in ameliorated fine-mapping of genomic loci. Therefore, facilitating new discoveries, and personalized approaches for treatments.

With the same line of thinking, copy number variation (CNV), variable Number of tandem repeats (VNTRs), and transposable elements are genomic variations that can be utilized to gain insights into the genetic architecture and mechanisms underlying SCD. Any of these structural variations can impact the expression, protein function, and regulatory mechanisms related to SCD pathology.

Multiplexed functional assays

Multiplexed functional assays offer the opportunity to integrate diverse omics technologies, enabling a comprehensive assessment of how genetic variants impact protein function, cellular processes, and disease-related pathways in sickle cell disease (SCD)[298]. By targeting genes involved in erythropoiesis and red blood cell hydration (such as *BCL11A*, *KLF1*, and *SPTB*) simultaneously, we can gain valuable insights into their collective influence on SCD pathology. Conducting assays that examine enzymes identified through metabolomics experiments or utilizing protein-protein interaction assays to investigate proteins affecting red blood cell sickling propensity can lead to the discovery of novel therapeutic agents and shed light on previously unknown mechanisms.

In a recent study, a combination of single-cell perturbation in primary hematopoietic stem and progenitor cells (HSPCs) and chromosome conformation capture (3C) was employed to explore the role of specific mutations controlling fetal hemoglobin (HbF) levels[286]. This innovative functional assay highlighted the collaborative impact of multiple functional elements carrying these mutations on HbF expression. Furthermore, another perturbation screen, utilizing base editors in HSPCs, was conducted to identify non-coding variants that modulate HbF expression[299].

Predictive models

So far most of the omics integration available in the literature combine two or three complementary layers of omics data with each. However, several machine learning approaches exist to integrate more layers and therefore enable new discoveries[300]. In SCD a drug repurposing assay employed a automate image machine learning algorithm to characterize the red blood cell morphology[301]. Around 21 of the identified compounds exhibit potential as drug candidates based on their inhibitory concentrations and comparison with free concentrations of oral drugs in human serum. Furthermore, the therapeutic potential of each compound can be predicted based on measurements of sickling times in individuals with varying severity of sickle syndromes.

Additionally, another study developed a deep learning system for detecting sea fan neovascularization, a sign of proliferative sickle cell retinopathy, from ultra-widefield color fundus photographs. The research highlights the limited adherence to screening for vision-threatening retinopathy in patients with sickle cell hemoglobinopathy. The deep learning system

achieved high sensitivity (97.4%) and specificity (97.0%) for detecting sea fan neovascularization, demonstrating its potential to expand access to rapid retinal evaluations and identify patients at risk of vision loss from this condition[302].

Conclusion

Therapeutic approaches involving gene editing, such as CRISPR or base editing, require substantial additional time to ensure their safety in human applications. Moreover, the accessibility of these advanced therapies to low-income countries, where the burden of sickle cell disease (SCD) is highest, may take decades or longer to achieve. Considering the prevalence of organ damage in SCD patients later in life, the ideal proposition would be the development of safe small molecule solutions that can synergistically work with other drugs. Disease research, including the study of biology, is intricate and multifaceted. Focusing on individual components within a complex system may yield partial insights into the problem at hand. Therefore, it is crucial for future studies to encompass the various components and their interactions, thereby advancing our understanding of biology and diseases as a whole.

Annex A: Supplementary Information for Multi-ancestry meta-analysis identifies 3 novel loci associated with fetal hemoglobin levels

The article presented is in preparation to be submitted to the *American Journal of Human Genetics*. In this article, to identify novel genetic regulators of HbF levels, we combined association results at 24,272,278 variants ("combined" minor allele frequency (MAF) ≥1%) from 5,903 European-ancestry individuals from the SardiNIA Study [203] and 3,740 SCD participants, mostly of African descent. Because of the genetic heterogeneity from these populations, we used PCs as covariates and opted to analyze each study individually. For the meta-analysis, we used MR-MEGA, which was developed to account for ancestry differences to maximize discovery power in GWAS [204]. Additionally, we performed whole exome sequencing association testing in 1,354 SCD patients. The annex includes cohort description, all the variants prioritization, and annotations I performed.

SOM: Multi-ancestry meta-analysis identifies 3 novel loci associated with fetal hemoglobin levels

Authors: Yann Ilboudo[1,2,*], Nicolas Brosseau[1,2,*], Ken Sin Lo[1,2], Mélissa Beaudoin[1,2], Florian Wünnemann[1,2], Pablo Bartolucci[4], Frédéric Galactéros[4], Philippe Joly[5,6], Allison E. Ashley-Koch, Marilyn J. Telen, Swee Lay Thein, Carlo Sidore[3], Francesco Cucca[3] Abdullah Kutlar, Carlo Brugnara, Guillaume Lettre[1,2]

Affiliations:

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

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

[3]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy.

[4]Red Cell Genetic Disease Unit, Hôpital Henri-Mondor, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Est, IMRB - U955 - Équipe no 2, Créteil, France

[5]Unité Fonctionnelle 34445 'Biochimie des Pathologies Érythrocytaires', Laboratoire de Biochimie et Biologie Moléculaire Grand-Est, Groupement Hospitalier Est, Hospices Civils de Lyon, Bron, France

[6]Laboratoire Inter-Universitaire de Biologie de la Motricité (LIBM) EA7424, Equipe 'Biologie Vasculaire et du Globule Rouge', Université Claude Bernard Lyon 1, Comité d'Universités et d'Etablissements (COMUE), Lyon, France

*Authors contributed equally to this manuscript.

Correspondence:

Guillaume Lettre

Montreal Heart Institute

5000 Bélanger Street

Montreal, Quebec, Canada

514-376-3330 (ext. 2657)

guillaume.lettre@umontreal.ca

Supplementary Table 1. Descriptive statistics of each cohort included in the study. N, sample size; SD, standard deviation; NA, not available.

Cohort	N_{GWAS} (male/female)	N_{WES} (male/female)	Age (mean±SD)	%HbF (mean±SD)	Reference
Cooperative Study of Sickle Cell Disease (CSSCD)	1,132 (593/539)	116/128	14.6±12.8	6.5±4.4	[248]
Multicenter Study of Hydroxyurea (MSH)	57 (34/23)	NA	28.5 ± 6.8		[303]
Jamaica Sickle Cell Cohort Study (JSCCS)	89 (41/48)	NA	NA	5.5±4.2	[304]
GEN-MOD	406 (184/222)	183/223	25.1±13.2	6.8±5.0	[126]
Mondor/Lyon	324 (120/202)	120/201	34.6±11.8	7.8±6.3	[305]
Georgia Health Sciences University (GHSU)	186 (95/91)	NA	31.4±10.6	NA	
Tanzania	1,213 (575/638)	NA	13.3 ± 7.4	5.6±4.3	[224]
SardiNIA	5,903 (2512/3391)	NA	43.5 ±17.6		[203]
Duke University Outcome Modifying Genes (OMG)	299 (136/163)	3/10	36.0±12.3	8.3±8.9	[163]
CIP: Differential response to hydroxyurea and incidence of stroke in sickle cell disease	NA	371 (224/147)	8.8±4.4	10.2±6.1	[306]

Supplementary Table 2. Functional annotations of 95% credible set variants.

Locus	Variant	MAF (CSSCD?)	MAF (SardiNIA)	Conditional P-value (N=9643)	PIP	Non-conditional P-value (N=10,045)	rsID	VEP functional annotation (most severe)	eQTLgen	GTEx (whole blood)	GWAS catalog
BICC1	10_58728559_G_A	0.39	0.39	6.16697E-09	0.6171	3.73057E-06	rs4433524	intron	TFAM		
	10_58715963_C_T	0.40	0.86	2.36271E-08	0.1611	7.26799E-06	rs3816114	non-coding transcript exon variant (FAM133CP)			
	10_58727996_T_C	0.33	0.86	5.5759E-08	0.0683	1.32839E-05	rs2393491	intron (BICC1)			
	10_58745172_A_G	0.34	0.86	9.14764E-08	0.0416	1.03888E-05	rs10826229	intron (BICC1)	TFAM		
	10_58745524_C_G	0.34	0.86	9.76649E-08	0.039	8.42413E-06	rs12764407	intron (BICC1)	TFAM		
	10_58744436_T_A	0.35	0.86	1.05544E-07	0.0361	1.88484E-05	rs11006248	intron (BICC1)			
KLF1	19_12879166_A_G	0.42	0.23	4.50221E-08	0.1128	6.3249E-07	rs4804210	intron (DNASE2)			
	19_12877067_A_G	0.45	0.23	5.5158E-08	0.0921	1.12642E-06	rs2418568	intron (DNASE2)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	IFR
	19_12876791_T_C	0.41	0.23	8.02442E-08	0.0633	1.06843E-06	rs10404876	intron (DNASE2)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	MCH
	19_12877614_T_C	0.62	0.23	8.31126E-08	0.0611	3.77894E-06	rs2085466	intron (DNASE2)	DNASE2,	DNASE2,	

									FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	FARSA	
	19_12864182_T_C	0.62	0.22	9.04784E-08	0.0561	1.35175E-05	rs1985646	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2, FARSA	
	19_12871461_G_T	0.40	0.23	1.02115E-07	0.0497	1.21746E-06	rs2242513	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12874762_A_G	0.49	0.23	1.03488E-07	0.0491	1.26619E-06	rs11085822	synonymous (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12871984_A_G	0.42	0.23	1.14794E-07	0.0442	1.04088E-06	rs2242514	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12872658_C_T	0.40	0.23	1.23931E-07	0.041	1.28344E-06	rs2242516	intron (MAST1)	DNASE2,	DNASE2	

									FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A		
	19_12849889_T_C	0.62	0.23	1.29422E-07	0.0392	4.18725E-06	rs8099965	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12851458_G_A	0.44	0.23	1.39211E-07	0.0365	1.50003E-06	rs10407116	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12851834_A_G	0.44	0.23	1.44096E-07	0.0352	1.48636E-06	rs7250751	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12837124_T_C	0.62	0.21	1.47647E-07	0.0344	5.1723E-06	rs4804737	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12852329_T_C	0.45	0.23	1.5187E-07	0.0334	1.25491E-06	rs1078264	splice region variant	DNASE2,	DNASE2	

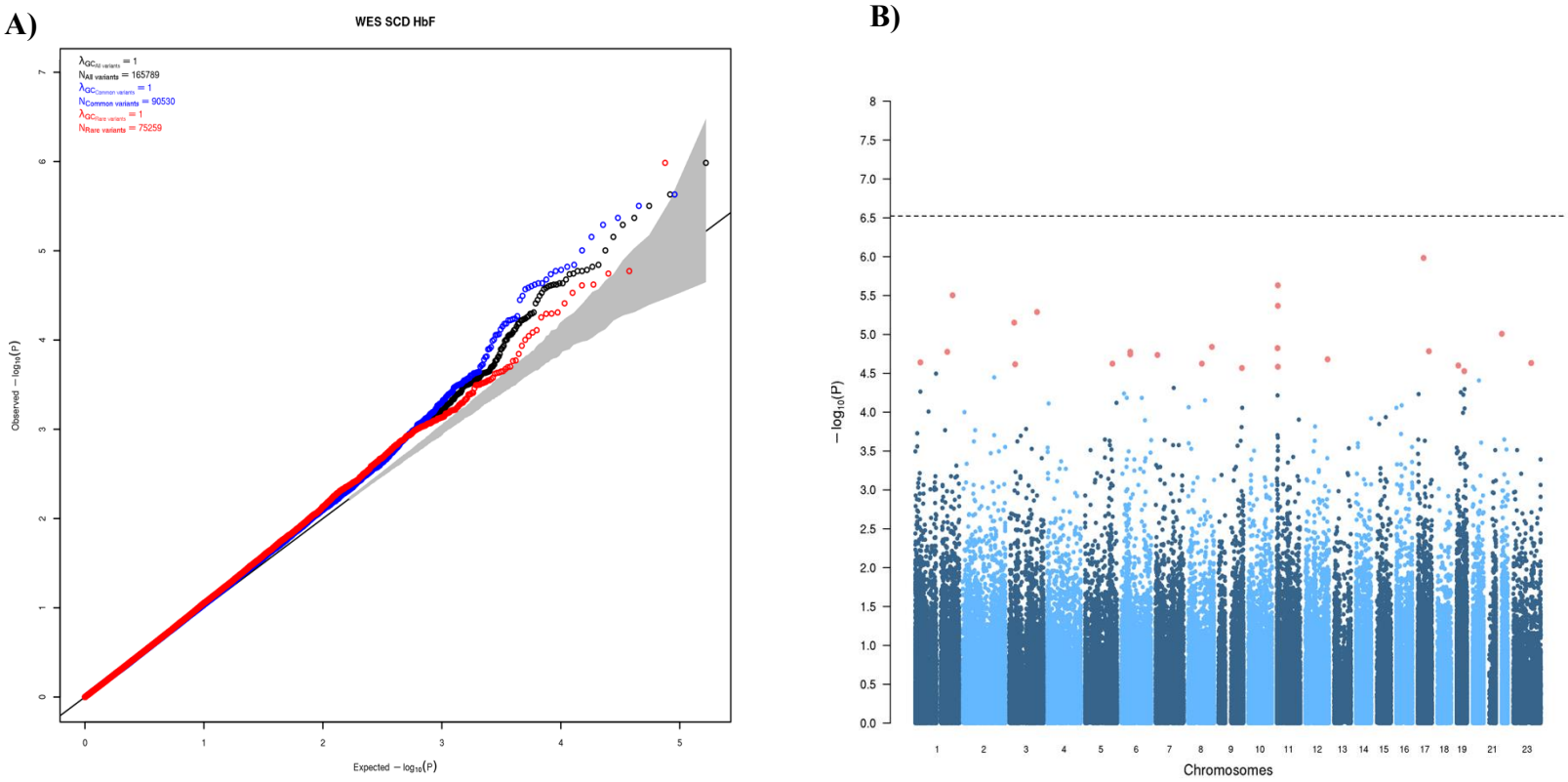
								(MAST1)	FARSA, WDR830S, RAD23A, CRYZ, PIGK, DNAJB4		
	19_12844730_A_C	0.62	0.23	2.04335E-07	0.0249	1.13174E-05	rs6511843	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12847883_T_C	0.62	0.23	2.12423E-07	0.0239	6.95395E-06	rs2290688	synonymous (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12847036_A_G	0.62	0.23	0.000000219	0.0232	5.42175E-06	rs10426080	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12847546_T_C	0.62	0.23	2.26306E-07	0.0224	6.94976E-06	rs2290689	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12838050_G_T	0.62	0.21	2.50966E-07	0.0202	9.00915E-06	rs8111370	intron (MAST1)	DNASE2, FARSA,	DNASE2	

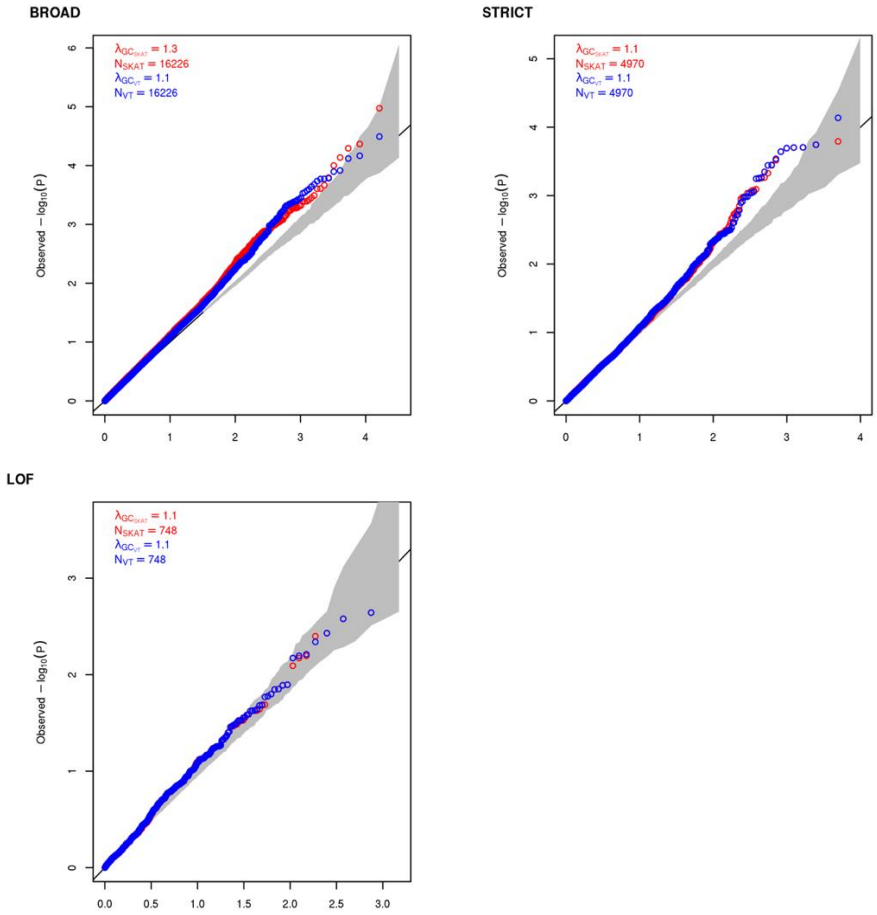
									WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A		
	19_12848817_G_C	0.44	0.23	3.97718E-07	0.0128	0.000001563	rs8105643	non-coding transcript exon variant (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12837648_A_G	0.57	0.21	4.04527E-07	0.0126	1.35282E-05	rs10424001	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12846728_A_G	0.40	0.23	4.24607E-07	0.012	4.64234E-06	rs7259590	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12846300_G_A	0.40	0.23	4.38568E-07	0.0116	4.71454E-06	rs1124820	intron (MAST1)	LRRC14, PPP1R16A, DNASE2, CPSF1, VPS28, CTD-2517M22.14, RPL8, RP11-457M11.5, ARHGAP39, TONSL	DNASE2	

	19_12853349_A_G	0.39	0.23	6.25379E-07	0.0081	6.08336E-06	rs4570988	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, JUNB	DNASE2	
	19_12863822_G_T	0.36	0.23	6.84184E-07	0.0074	7.89099E-06	rs4614850	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12864785_G_A	0.35	0.23	7.00371E-07	0.0073	9.12888E-06	rs2242512	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12865667_AAAC_A	0.39	0.02	7.12322E-07	0.0071	0.0836445	rs146361171	intron (MAST1)			
	19_12868306_T_A	0.39	0.23	7.39002E-07	0.0069	6.81825E-06	rs1810363	intron (MAST1)	DNASE2, ATP6V1D, MPP5, FARSA, WDR830S, RAD23A, PLEK2, HOOK2, JUNB, RNASEH2A	DNASE2	
	19_12852771_T_C	0.39	0.23	8.02752E-07	0.0063	5.85434E-06	rs10423124	intron (MAST1)	DNASE2, FARSA, WDR830S, RAD23A, HOOK2, JUNB,	DNASE2	

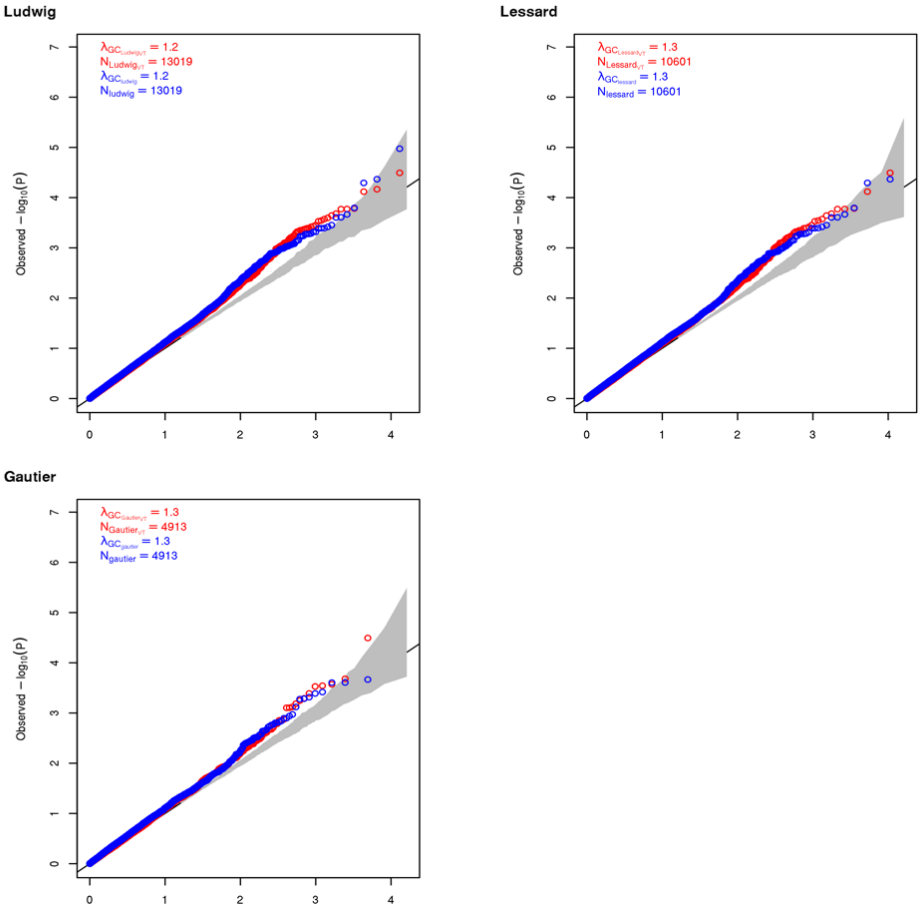
									RNASEH2A		
CECR2, SLC25A18	22_17559810_C_T	0.06	NA	3.98E-08	0.7411	2.06E-06	rs116175381	intergenic			
	22_17559209_A_G	0.09	0.11	1.75E-07	0.1347	0.000137029	rs885971	intergenic	ATP6V1E1, BCL2L13		
	22_17557429_G_A	0.06	NA	2.92E-07	0.1011	9.30E-06	rs115000541	3'UTR variant (CECR2)			

Supplementary Figure 1. Single variant association whole exome variants of fetal hemoglobin in SCD. A) QQplot of fetal hemoglobin stratified by allelic frequency. Blue points are common variants (MAF < 5%), red points are rare variants (MAF < 1%) and black points are all the points. B) Manhattan plot of HbF, we found the exome-wide threshold at $P < 3.0x10^{-7}$.

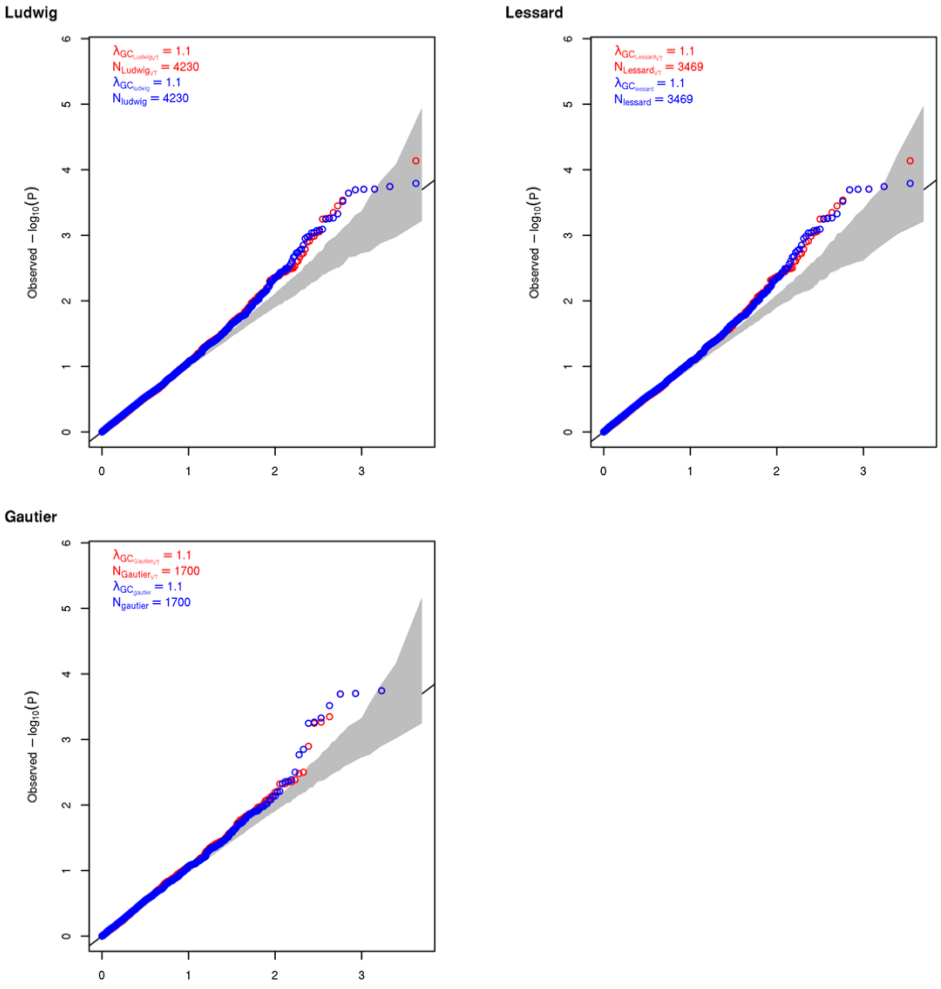




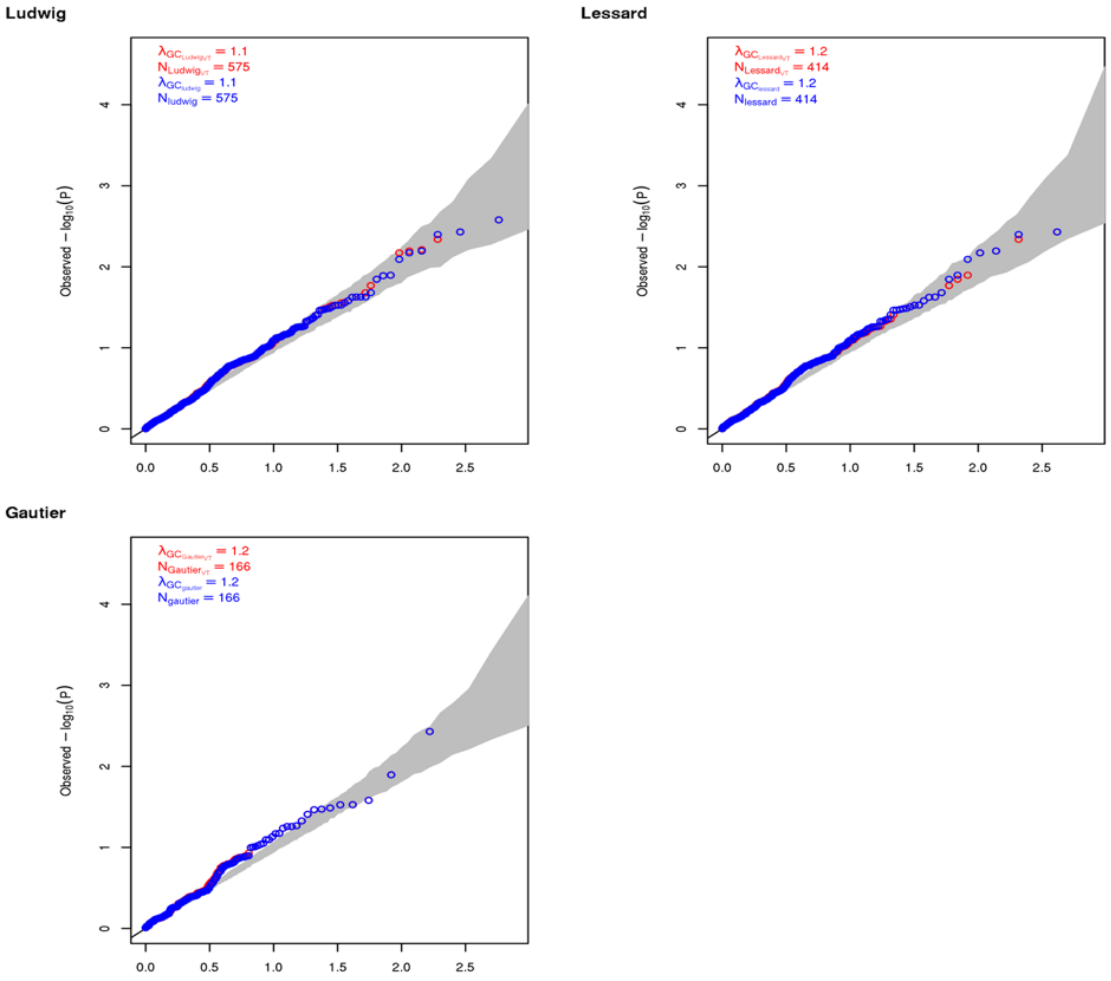
Supplementary Figure 2. Gene-based QQplot of fetal hemoglobin in SCD patients employing three burden test approaches: Broad, strict and LOF selecting variants with MAF < 1%.



Supplementary Figure 3. Gene-based QQplot of fetal hemoglobin in SCD employing broad mask, and filtered according to three different expression dataset MAF < 1%. Ludwig, Gautier and Lessard represent the expression datasets. Lessard, S et al (2017), Gautier, E. F. et al. (2016), Ludwig, L. S. et al. (2019).



Supplementary Figure 4. Gene-based QQplot of fetal hemoglobin in SCD employing strict mask, and filtered according to three different expression dataset MAF < 1%. Ludwig, Gautier and Lessard represent the expression datasets. Lessard, S et al (2017), Gautier, E. F. et al. (2016), Ludwig, L. S. et al. (2019).



Supplementary Figure 5. Gene-based QQplot of fetal hemoglobin in SCD patients employing strict LoF, and filtered according to three different expression dataset MAF < 1%. Ludwig, Gautier and Lessard represent the expression datasets. Lessard, S et al (2017), Gautier, E. F. et al. (2016), Ludwig, L. S. et al. (2019).

Annex B. Supplementary Information for Exome- and genome-wide association studies of red blood cell density in sickle cell disease patients

The following article is intended to be submitted to the journal, *British Journal of Hematology*. In this article to identify novel genetic regulators of dense red blood levels (DRBC). I performed association tests of imputed genotypes of DRBC in 581 SCD patients. I then annotated the results, and identified which ones are the most promising and need to be replicated. Then I used whole-exome sequencing to identify rare coding variants regulating DRBC. I performed gene-based analysis with sequence kernel association test (SKAT) and variable threshold (VT) to detect associations of coding variants put together. The annex includes a table describing where to find the large supplementary tables (URL links), the description of their content. It also includes supplementary figures I generated for various variant prioritization approaches.

Tables of content for supplementary tables:

List	N	N*	Reference/Criteria of inclusion
MCHC GWAS	29	23	All coding variants associated (pval < $5x10^{-8}$) with MCHC in both trans-ancestry and african population. Chen *et al* (2020)
OMIM	133	108	OMIM (acc. Feb 23 2022) searched terms 'cryohydrocytosis', 'elliptocytosis',' 'malaria', 'pseudohyperkalemia', 'pyropoikilocytosis', 'spherocytosis', 'stomatocytosis',' xerocytosis'. Downloaded as gene-map tables. Rasmussen, S. A.*et al* (2020)
RBC enzymopathies	16	14	All gene identified in Luzzatto, L. (2021)
IUPHAR	3,089	2,233	Download complete target and family list (tsv). Harding, S. D. *et al.(2022)*

Supplementary Table 1 Data source for candidate genes[1] For the candidate gene approach, we selected protein-coding genes with unambiguous mapping to current approved gene symbols. N; total numbers of protein coding genes according to the inclusion criteria. N* column represents the subset of genes expressed in erythrocyte and erythroblast according to the following references Lessard, S *et al* (2017), Gautier, E. F. *et al*. (2016), Ludwig, L. S. *et al*. (2019).

Supplementary Table 2 Single-variants association nominally significant [1] For the candidate gene approach, we selected protein-coding genes with unambiguous mapping to current approved gene symbols. N; total numbers of protein coding genes according to the inclusion criteria. N* column represents the subset of genes expressed in erythrocyte and erythroblast according to the following references[2-4].

Supplementary Table 3 Gene-based association nominally significant [1] For the candidate gene approach, we selected protein-coding genes with unambiguous mapping to current approved gene symbols. N; total numbers of protein coding genes according to the inclusion criteria. N* column represents the subset of genes expressed in erythrocyte and erythroblast according to the following references[2-4].

Supplementary Table 4 Nominally significant genome wide association of DRBC ($P_{DRBC\ Meta\text{-}analysis} < 0.01$). Meta-analysis, and cohort specific (GENMOD & Henri-Mondor) association results of DRBC in 573 SCD patients. All base pair position are provided on hg38 build. Abbreviations:

SNP: chromosome_basepair_allele1,allele2; REF: reference allele; ALT: alternate allele; N_INFORMATIVE_HM: sample size for GWAS in Henri-Mondor cohort; AF_HM: Allele frequency for A1 for GWAS in Henri-Mondor; ALT_EFFSIZE_HM: Effect size for A1 allele for GWAS in Henri-Mondor; PVALUE_HM: P-value for GWAS in Henri-Mondor cohort.

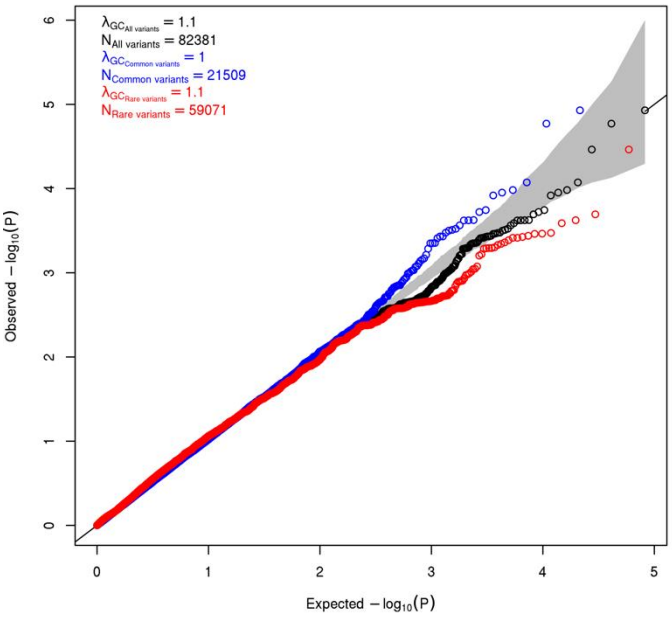
SNP: chromosome_basepair_allele1,allele2; REF: reference allele; ALT: alternate allele; N_INFORMATIVE_GENMOD: sample size for GWAS in GENMOD cohort; AF_HM: Allele frequency for A1 for GWAS in GENMOD; ALT_EFFSIZE_ GENMOD: Effect size for A1 allele for GWAS in GENMOD; PVALUE_ GENMOD: P-value for GWAS in GENMOD cohort.

CHR_BP_A1_A2: chromosome_basepair_allele1,allele2; CHR: chromosome; BP: base pair; A1: allele 1; A2: Allele2; Freq1: weighted average frequency allele 1 in meta-analysis; FreqSE: frequency corresponding standard error for allele frequency estimate; MinFreq: min frequency; MaxFreq: max frequency; Weight: the sum of the individual study weights (typically, N) for this marker; Zscore : the combined z-statistic for this marker; P-value : meta-analysis p-value; Direction: summary of effect direction for each study, with one '+' or '-' per study; HetISq: heterogeneity I^square; HetChiSq: chi-squared statistic in simple test of heterogeneity; HetDf; heterogeneity degrees of freedom; HetPVal: heterogeneity p-value; P-

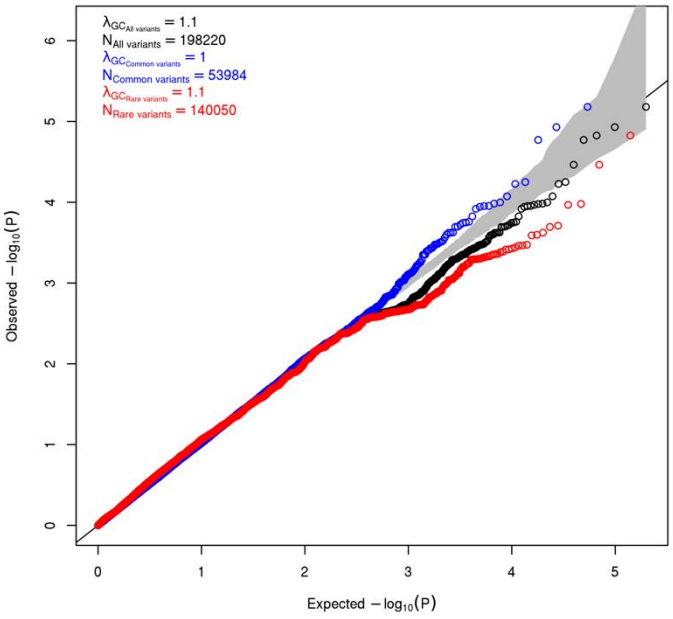
value for heterogeneity statistic; Tracks: Prioritization annotation; -Consequence: VEP calculated variant consequences (all possible consequences per transcript were retrieved); Genes: VEP gene symbol (all possible gene symbol per transcript were retrieved); rsID: Existing rsID; Disease: Disease associated with SNP according to DisGeNET (v7.0)[8]; pmid: pubmed ID for disease associated with SNP according to DisGeNET (v7.0)[8]

Supplementary Figure 1. Single-variants association genes expressed in erythrocyte (A) and erythroblast (B & C) according to the following references Lessard, S *et al* (2017), Gautier, E. F. *et al*. (2016), Ludwig, L. S. *et al*. (2019).

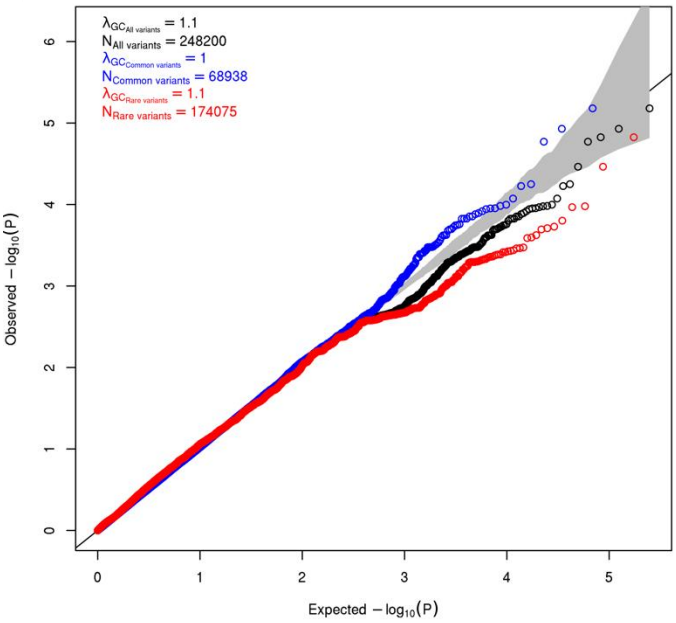
A)



B)

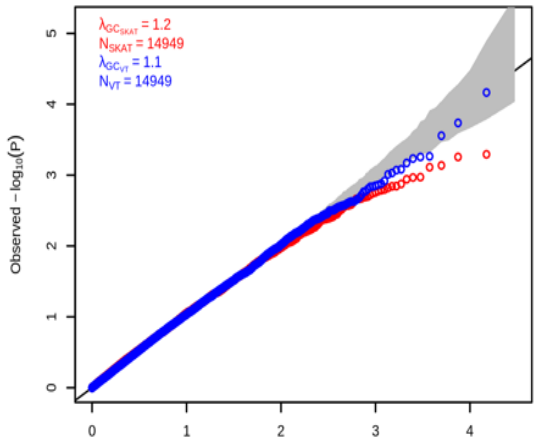


C)

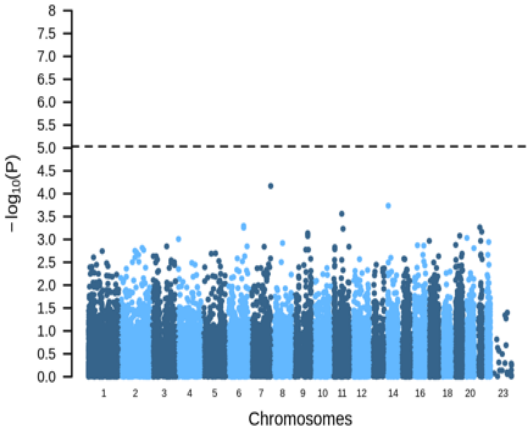


Supplementary Figure 2. QQplot and manhatan plot for gene-based test for different mask. A & B for broad collapsing scheme (at least one prediction algorithm considers the variant to be deleterious) $P_{\text{significance level}}=1.7 \times 10^{-6}$. B & C for strict collapsing scheme (all prediction algorithms consider the variant to be deleterious) $P_{\text{significance level}}=9.3 \times 10^{-6}$. D & E for variant predicted to be loss of function (pLoF) 3.3×10^{-5}. For all the plots, MAF < 1%. Two tests were employed for each analyses: sequence kernel association test (SKAT) & variable threshold (VT).

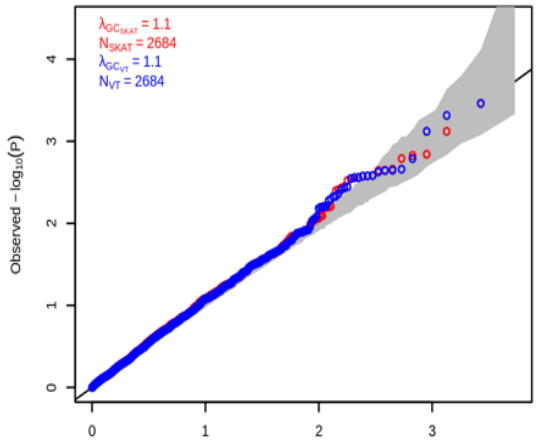
A)


λGCSKAT = 1.2
NSKAT = 14949
λGCVT = 1.1
NVT = 14949
Observed −log10(P)

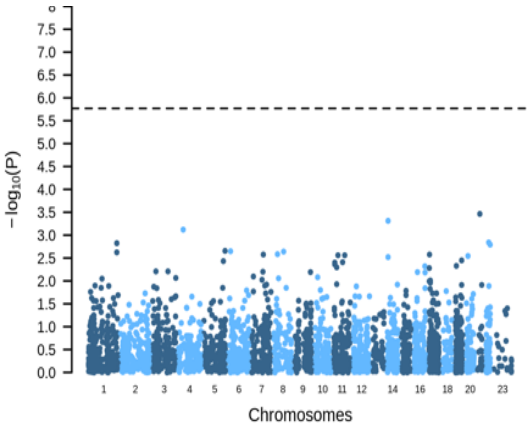
B)


−log10(P)
Chromosomes

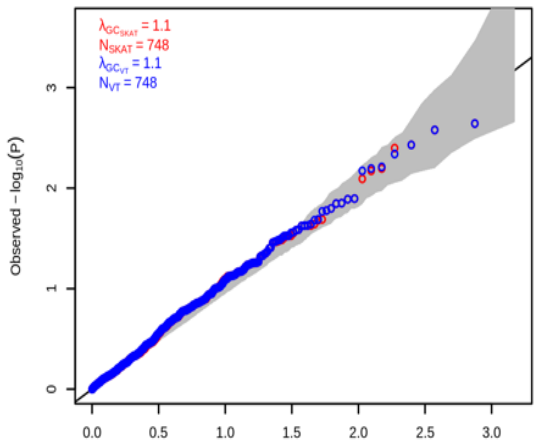
C)


λGCSKAT = 1.1
NSKAT = 2684
λGCVT = 1.1
NVT = 2684
Observed −log10(P)

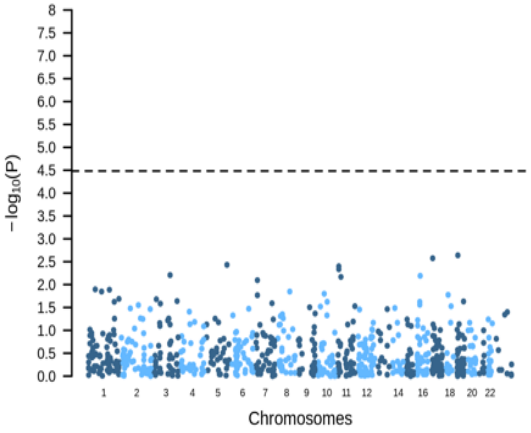
D)


−log10(P)
Chromosomes

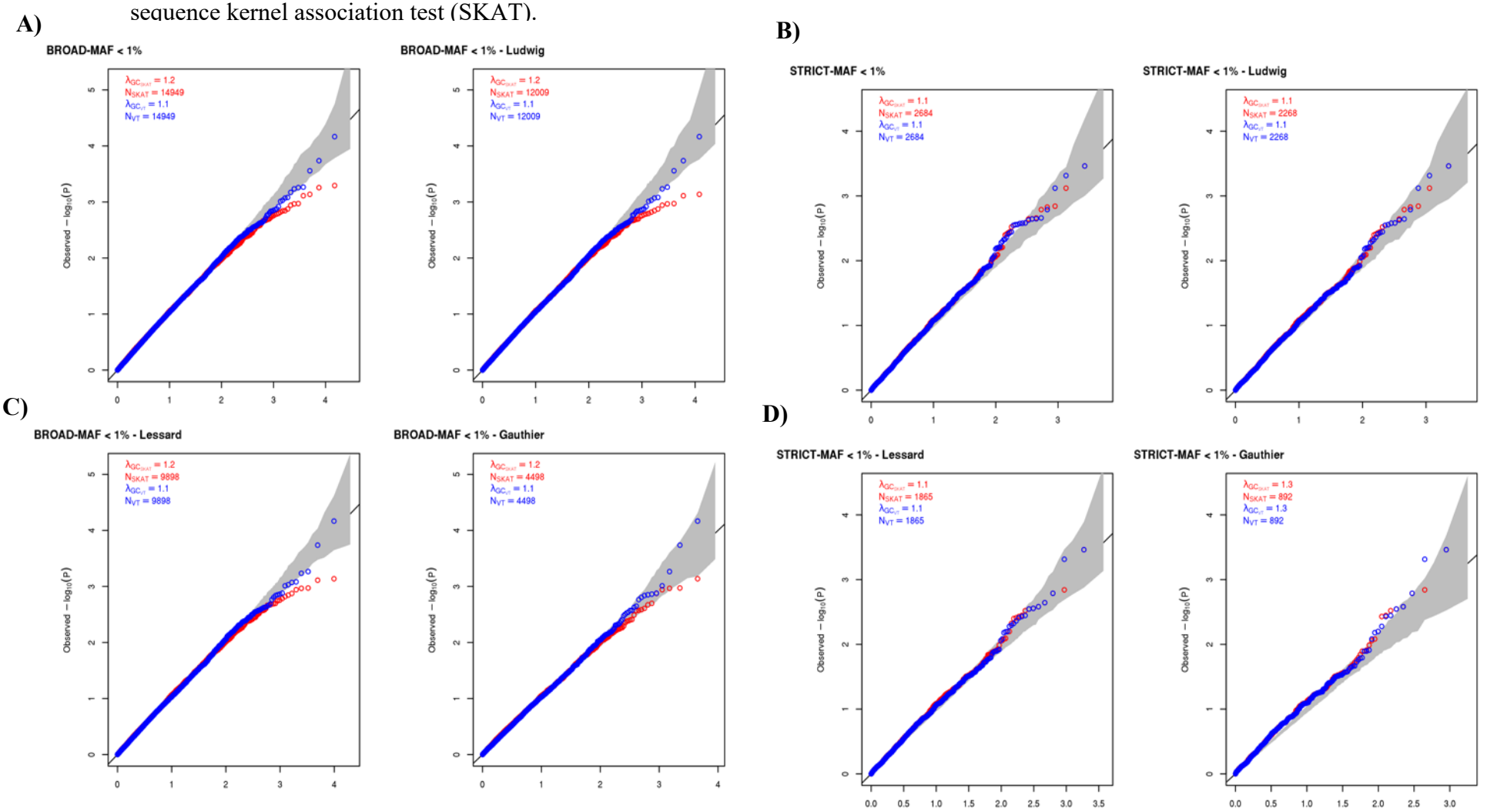
E)


λGCSKAT = 1.1
NSKAT = 748
λGCVT = 1.1
NVT = 748
Observed −log10(P)

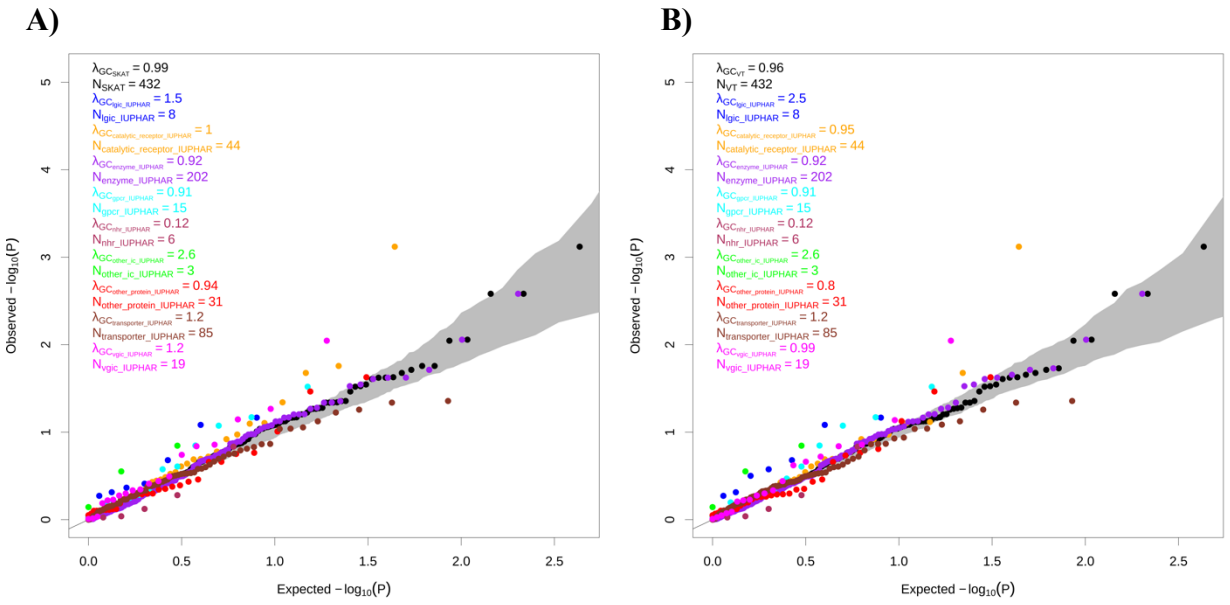
F)


−log10(P)
Chromosomes

Supplementary Figure 3. Gene-based associations genes filtered based on expression profile in in erythrocyte (A) and erythroblast (B & C) according to the following references Lessard, S *et al* (2017), Gautier, E. F. *et al*. (2016), Ludwig, L. S. *et al*. (2019).,and based two variant collapsing schemes (BROAD & STRICT). Blues point represent association with variable threshold (VT), while red points are sequence kernel association test (SKAT).

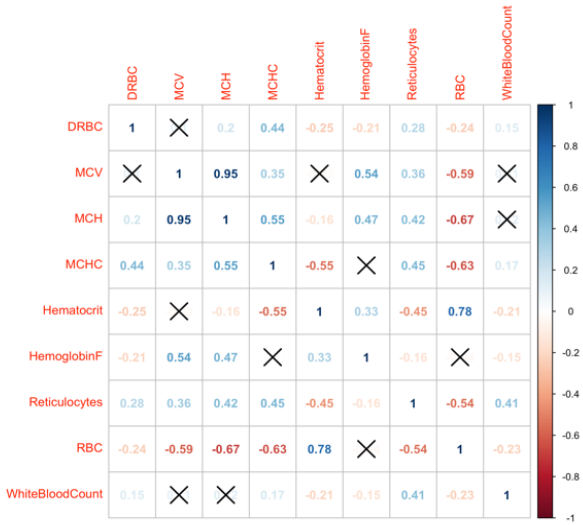


Supplementary Figure 4. Candidate gene approach looking at different IUPHAR genes families including variants in which all predictions algorithms agreed on deleteriousness (strict mask). A is for SKAT, B for VT. Blue dots represent markers ligand-gated ion channel (l), orange dots represent catalytic receptors, purple-colored dots represent enzymes, cyan dots represent G protein-coupled receptors, maroon dots represent nuclear hormone receptor, green represent other ion channels, red dots represent other protein targets, burgundy dots represent transporters, and magenta dots represent voltage gated ion channels. Abbreviations: lgic_IUPHAR: ligand-gated ion channel, gpcr_IUPHAR: G protein coupled receptor, nhr_IUPHAR: nuclear hormone receptor, other_ic_IUPHAR: other ion channel, vgic_IUPHAR: voltage gated ion channel.

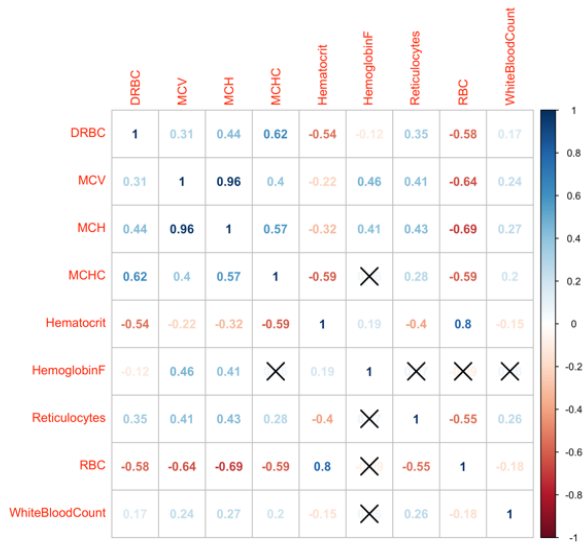


Supplementary Figure 5. Correlation matrices for DRBC and other hematological traits. A) correlation matrix in the Mondor-Lyon cohort. B) correlation matrix in GEN-MOD. C) correlation matrix in both Mondor-Lyon and GEN-MOD Positive correlations are displayed in blue, while negative correlation are in red. Color intensity are proportional to the correlation coefficients. The correlations coefficients are Pearsons' r coefficient. X symbol on a given cell means that the correlation is not significant (Pval > 0.05). Abbreviations: DRBC; dense dehydrated red blood cell. MCV; mean cell volume. MCH; mean cell hemoglobin. MCHC; mean cell hemoglobin concentration. HemoglobinF; fetal hemoglobin. RBC; red blood cell count. All blood traits were normalized and corrected for age and sex.

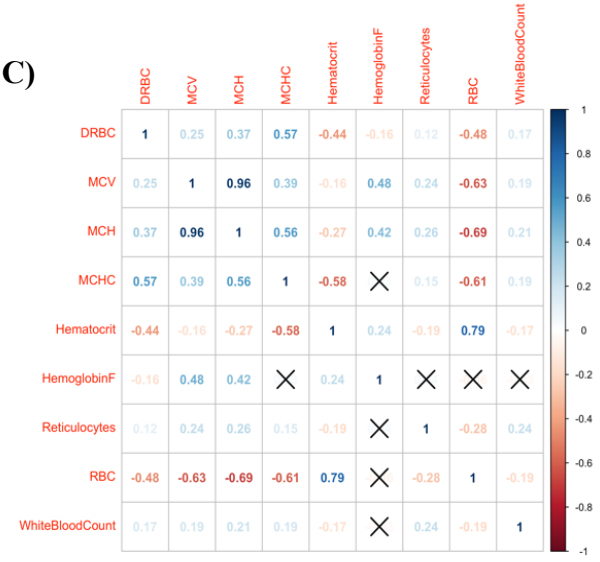
A)


DRBC MCV MCH MCHC Hematocrit HemoglobinF Reticulocytes RBC WhiteBloodCount
DRBC 1 X 0.2 0.44 -0.25 -0.21 0.28 -0.24 0.15
MCV X 1 0.95 0.35 X 0.54 0.36 -0.59 X
MCH 0.2 0.95 1 0.55 -0.16 0.47 0.42 -0.67 X
MCHC 0.44 0.35 0.55 1 -0.55 X 0.45 -0.63 0.17
Hematocrit -0.25 X -0.16 -0.55 1 0.33 -0.45 0.78 -0.21
HemoglobinF -0.21 0.54 0.47 X 0.33 1 -0.16 X -0.15
Reticulocytes 0.28 0.36 0.42 0.45 -0.45 -0.16 1 -0.54 0.41
RBC -0.24 -0.59 -0.67 -0.63 0.78 X -0.54 1 -0.23
WhiteBloodCount 0.15 X X 0.17 -0.21 -0.15 0.41 -0.23 1
1 0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1

B)


DRBC MCV MCH MCHC Hematocrit HemoglobinF Reticulocytes RBC WhiteBloodCount
DRBC 1 0.31 0.44 0.62 -0.54 -0.12 0.35 -0.58 0.17
MCV 0.31 1 0.96 0.4 -0.22 0.46 0.41 -0.64 0.24
MCH 0.44 0.96 1 0.57 -0.32 0.41 0.43 -0.69 0.27
MCHC 0.62 0.4 0.57 1 -0.59 X 0.28 -0.59 0.2
Hematocrit -0.54 -0.22 -0.32 -0.59 1 0.19 -0.4 0.8 -0.15
HemoglobinF -0.12 0.46 0.41 X 0.19 1 X X X
Reticulocytes 0.35 0.41 0.43 0.28 -0.4 X 1 -0.55 0.26
RBC -0.58 -0.64 -0.69 -0.59 0.8 X -0.55 1 -0.18
WhiteBloodCount 0.17 0.24 0.27 0.2 -0.15 X 0.26 -0.18 1
1 0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1

C)


DRBC MCV MCH MCHC Hematocrit HemoglobinF Reticulocytes RBC WhiteBloodCount
DRBC 1 0.25 0.37 0.57 -0.44 -0.16 0.12 -0.48 0.17
MCV 0.25 1 0.96 0.39 -0.16 0.48 0.24 -0.63 0.19
MCH 0.37 0.96 1 0.56 -0.27 0.42 0.26 -0.69 0.21
MCHC 0.57 0.39 0.56 1 -0.58 X 0.15 -0.61 0.19
Hematocrit -0.44 -0.16 -0.27 -0.58 1 0.24 -0.19 0.79 -0.17
HemoglobinF -0.16 0.48 0.42 X 0.24 1 X X X
Reticulocytes 0.12 0.24 0.26 0.15 -0.19 X 1 -0.28 0.24
RBC -0.48 -0.63 -0.69 -0.61 0.79 X -0.28 1 -0.19
WhiteBloodCount 0.17 0.19 0.21 0.19 -0.17 X 0.24 -0.19 1
1 0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1

]

Supplementary Table 5. Comparing association results identified in previous DRBC GWAS of 374 SCD patients[9] with current GWAS of 573 SCD patients. Column names, and abbreviations are the same as those described in Supp.Table 4. Out of 25 SNPs identified in the previous DRBC GWAS Ilboudo, Y. *et al.* (2017). 7 were missing (not genotyped or imputed), and 14 remained significant, of which 10 were both significant and had the same direction of effect as in the published study. Surprisingly the intronic variant at *ATP2B4* (rs10751450) did not stay significant although is strongly associates with MCHC in non-anemic Europeans (P=5 x 10^{-59}) Astle, W. J. *et al (2016)*. Abbreviations. Consequences: VEP calculated variant consequences (all possible consequences per transcript were retrieved); Genes: VEP gene symbol (all possible gene symbol per transcript were retrieved); rsID: Existing rsID. Beta(SE); regression effect size (standard error); HM: Henri-Mondor.

			Published				Meta-analysis		HM	GENMOD
rsID	**Consequence**	**Genes**	**Beta(SE)**	**Pvalue**	**SYMBOL**	**Consequence**	**Zscore**	**P-value**	**Beta ; Pvalue**	**Beta ; Pvalue**
rs4234795	intron_variant	SORCS2	-0.84(0.15)	2.0×10^{-7}	SORCS2	Intron	-4.435	9.2×10^{-6}	-0.051 ; P=0.81	-0.79 ; P=1.1×10^{-7}
rs77141833	upstream_gene_variant_intron_variant,non_coding_transcript_variant_downstream_gene_variant_intron_variant_intron_variant,NMD_transcript_variant	VSIG8_LOC107985216_SNHG28_NA	−1.12(0.22)	1.8×10^{-6}	VSIG8	Intron	-3.4	0.00077	-0.089 ; P=0.68	-0.71 ; P= 1.1×10^{-4}
rs146977005	intron_variant	SPTB	−0.33(0.09)	5.5×10^{-4}	SPTB	Intron_eQTL_for_SPTB	-1.5	0.13	0.20 ; P=0.079	-0.28 ; P=1.6×10^{-3}
rs5875087	intron_variant	H2BC4	−0.27(0.13)	0.045	HIST1H2BC	Intron	-0.36	0.73	0.37 ; P=0.017	-0.27 ; P=0.029
rs147900370	intergenic_variant	NA	−0.92(0.18)	2.4×10^{-6}	–	Intergenic	-2.6	0.0094	0.33 ; P=0.15	-0.76 ; P=2.0×10^{-5}
rs114402357	intergenic_variant	NA	2.03(0.4)	1.8×10^{-6}	–	Intergenic	-3.6	0.00028	-0.057 ; P=0.93	1.6 ; P = 5.2×10^{-6}
rs7216169	intron_variant_intron_variant,non_coding_transcript_variant	RABEP1	0.45(0.09)	1.4×10^{7}	RABEP1	Intron	-4.4	1.2×10^{-5}	0.12 ; P=0.29	0.40 ; P = 3.2×10^{-6}
rs1203972	upstream_gene_variant_downstream_gene_variant	LUC7L_FAM234A_NA	−0.22(0.08)	0.0082	LUC7L	Upstream	-2.2	0.027	0.0025 ;P=0.98	-0.21 ; P=6.1×10^{-3}
rs146893001	intron_variant_upstream_gene_variant	PTPN3	-2.04(0.4)	1.3×10^{-6}	PTPN3	Intron	-4.2	2.9×10^{-5}	-0.83 ;P=0.15	-1.5 ; P=3.8×10^{-5}
rs148303943	intron_variant_intron_variant,non_coding_transcript_variant_non_coding_transcript_exon_variant_upstream_gene_variant	GMPR_NA	−0.32(0.11)	0.0057	GMPR	Intron	-2.0	0.042	0.12 ;P=0.49	-0.32 ; P=0.0026
rs76513454	intergenic_variant	NA	-2.17(0.43)	2.0×10^{-5}	–	Intergenic	2.9	0.0033	-0.051 ;P=0.95	-1.8 ; P=3.3×10^{-4}
rs10751450	intron_variant_upstream_gene_variant	ATP2B4	−0.25(0.08)	0.0031	ATP2B4	Intron	1.6	0.11	0.15 ;P=0.20	-0.23 ; P=3.3×10^{-3}
rs62015549	intron_variant	THSD4	−2.44(0.49)	1.9×10^{-6}	THSD4	Intron	2.0	0.040	-0.27 ; P=0.46	-1.3 ; P=0.045
rs74989317	intron_variant_upstream_gene_variant_intron_variant,non_coding_transcript_variant_downstream_gene_variant	LINC00649_RN7SL740P_NA	−0.99(0.2)	1.5×10^{-6}	LINC00649	Intron	2.2	0.026	0.43 ;P=0.032	-0.78 ; P=1.5×10^{-5}
rs34514965	upstream_gene_variant	GADD45GIP1_DAND5	0.21(0.1)	0.043	GADD45GIP1	Upstream	1.5	0.13	-0.012 ; P=0.92	0.18 ; P = 0.05

rs11421513	intergenic_variant	NA	−0.23(0.08)	0.0074	–	Intergenic	-1.6	0.1004	0.074 ;P=0.50	-0.21 ; P=0.012
rs8048714	intron_variant_downstream_gene_variant_non_coding_transcript_exon_variant	NA_LOC100289580_LOC339059_PIEZO1	−0.3(0.08)	0.00073	PIEZO1	Intron_eQTL_for_PIEZO1	2.4	0.016	0.048 ; P0.67	-0.27 ; P=9.9 x 10^{-4}
rs73108077	upstream_gene_variant_downstream_gene_variant	DEFB121_DEFB122	−0.83(0.17)	1.8 x 10^{-6}	DEFB122	Downstream	-3.7	2.6 x 10^{-3}	-0.041 ; P=0.82	-0.73 ; P=1.3 x 10^{-5}
rs144995469	upstream_gene_variant	NA	-1.15(0.23)	1.5 x 10^{-6}	-	Intergenic	3.8	1.2 x 10^{-3}	0.12 ; P=0.69	-1.1 ; P=4.4 x 10^{-7}
rs9714060	NA	NA	−0.39(0.08)	7.4 x 10^{-7}	MUC4	Intron	NA	NA	NA	NA
rs543023132	NA	NA	−1.54(0.3)	1.4 x 10^{-6}	–	Intergenic	NA	NA	NA	NA
rs139628543	NA	NA	0.75(0.15)	2.0 x 10^{-6}	KLHL30	Intron	NA	NA	NA	NA
rs62270871	NA	NA	0.33(0.07)	2.6 x 10^{-5}	ALG1L	Intron_eQTL_for_SLC41A3	NA	NA	NA	NA
rs144514173	NA	NA	0.37(0.12)	0.0029	TMCC2	Downstream	NA	NA	NA	NA
rs201794926	NA	NA	0.18(0.07)	0.021	PPP1R16A	Intron	NA	NA	NA	NA

Annex C: Supplementary Information for Integrating metabolomics with GWAS reveals novel insights into the liver and kidney dysfunction in sickle cell disease patients.

The article presented is intended to be published in the journal, *Frontiers in Genetics*. In this article, employing both targeted and untargeted approaches we profiled the plasma of 706 SCD patients using liquid chromatography tandem mass spectrometry. The cohort included 406 French patients (GEN-MOD cohort) of recent African descent and 300 African Americans (OMG cohort) from the southeastern US. In total, we measured the levels of 233 known and 1,880 unknown metabolites. I constructed 66 modules containing at least 7 metabolites per module using clustering framework weighted correlation network analysis (WGCNA). I found a module strongly associated with increased risks of gallbladder removal. Additionally, I retrieved another module of metabolites strongly correlated with a measure of kidney function, namely estimated glomerular filtration rate (eGFR). Finally, I performed a GWAS for each of the 39 most robust modules, which resulted in two modules strongly associated with SNPs (FDR < 0.05). I obtained one module with multiple SNPs significantly associated ($P < 8.0 \times 10^{-10}$) near the gene encoding for hepatic triglyceride lipase (*LIPC*). The supplementary material includes analysis on data preprocessing; namely network construction, network topology for various soft-thresholding powers, visualization of networks associated SCD complications and traits, intramodular association for SCD complications, and traits, and finally QQplots of the associations between robust networks of metabolites and genotypes.

Integrating metabolomics with GWAS reveals novel insights into the liver and kidney dysfunction in sickle cell disease patients.

Yann Ilboudo[1], Melanie Garrett[2], Aurelie Guilbault[1], Allison Ashley-Koch[2], Marilyn Telen[3], Guillaume Lettre[4]

Affiliations

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

[2]Center for Human Disease Modeling, Duke University Medical Center, Durham, North Carolina, United States of America

[3]Department of Medicine, Division of Hematology, Duke University Medical Center, Durham, North Carolina, United States of America.

[4]Montreal Heart Institute, Montréal, Québec, Canada; Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada

Correspondence

Guillaume Lettre

Montreal Heart Institute

5000 Bélanger St

Montreal, Quebec, Canada, H1T 1C8

514-376-3330 ext. 2660

guillaume.lettre@umontreal.ca

Keywords: GWAS, sickle cell disease, metabolites, Phe-WAS, WGCNA

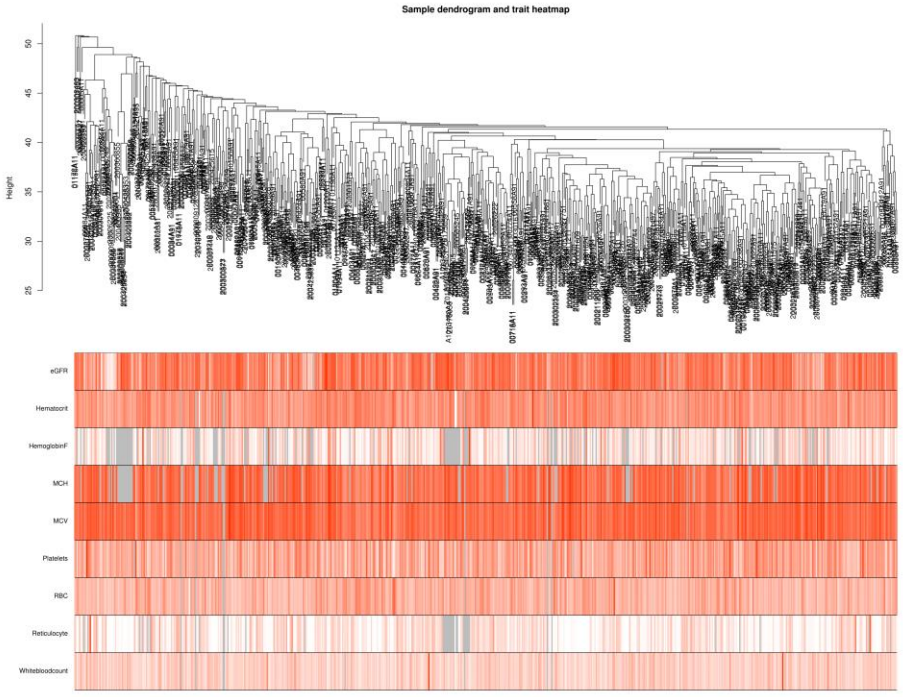
Supplementary Data Table 1. Phenome-wide analysis results for rs1800588 using VAMPIRE[307]. Abbreviations: Pheno: phenotype; BIOS_ceQTL_AA: eQTL allele from BIOS[308]; BIOS_ceQTL_pvalue : eQTL pvalue from BIOS[308]; eQTLGen_ceQTL_AA: eQTL allele from eQTLGen[309]; eQTLGen_ceQTL_FDR: eQTL false discovery rate from eQTLGen[309]; NESDA_ceQTL_AA: eQTL allele from NESDA[310]; NESDA_ceQTL_Beta: eQTL effect size from NESDA[310]; NESDA_ceQTL_FDR: eQTL false discovery rate from NESDA[310]; DGN bulk_ceQTL_beta: eQTL effect size from DGN[311]; DGN bulk_ceQTL_tstat: eQTL tstat from DNG bulk[311]; DGN bulk_ceQTL_pvalue: eQTL pvalue from DNG bulk[311]; DGN bulk_ceQTL_FDR: eQTL false discovery rate fron DNG bulk[311]; Westra_teQTL_AA: eQTL allele from Westra[312]; Westra_teQTL_FDR: eQTL false discovery rate from Westra[312]; Westra_teQTL_zscore: eQTL zscore from Westra[312]; Anno_cat: Annotation category

Supplementary Data Table 2. Phenome-wide analysis results for rs1800588 using gwasATLAS[313]. Abbreviations: ID: identification; PMID: PubMed identification; N: sample size; EA: effect allele; NEA: non-effect allele.

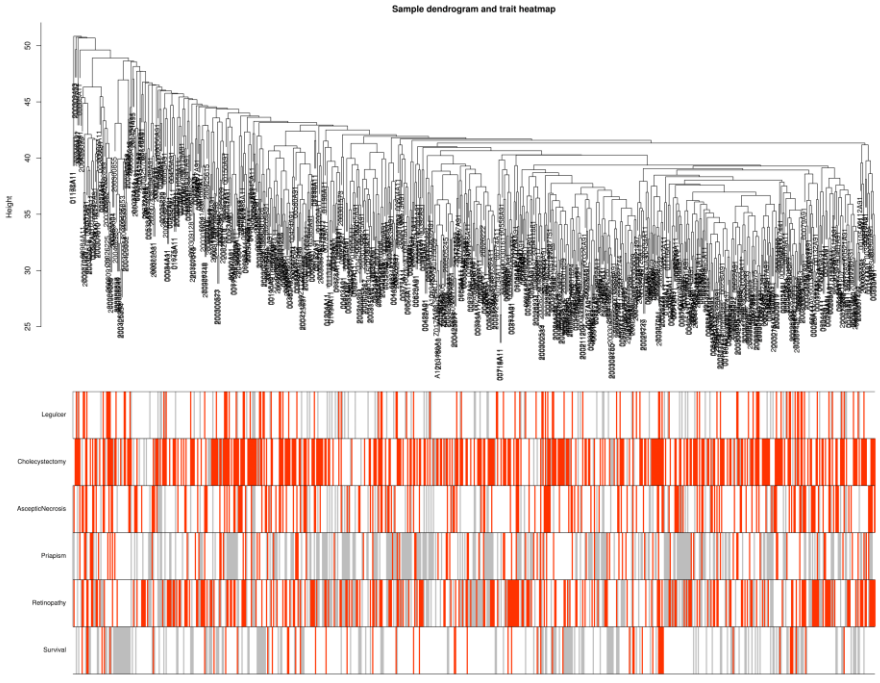
Supplementary Data Table 3. Phenome-wide analysis results for rs12277271 using gwasATLAS[313]. Abbreviations: ID: identification; PMID: PubMed identification; N: sample size; EA: effect allele; NEA: non-effect allele.

Supplementary Figure 1. Clustering dendrogram of 688 samples and heat map. A) Clustering samples based on their Euclidean distance with a heatmap of blood traits. B) Clustering samples based on their Euclidean distance with a heatmap of complications. White and red indicate "NO" and "YES," respectively. Missing data are represented in grey.

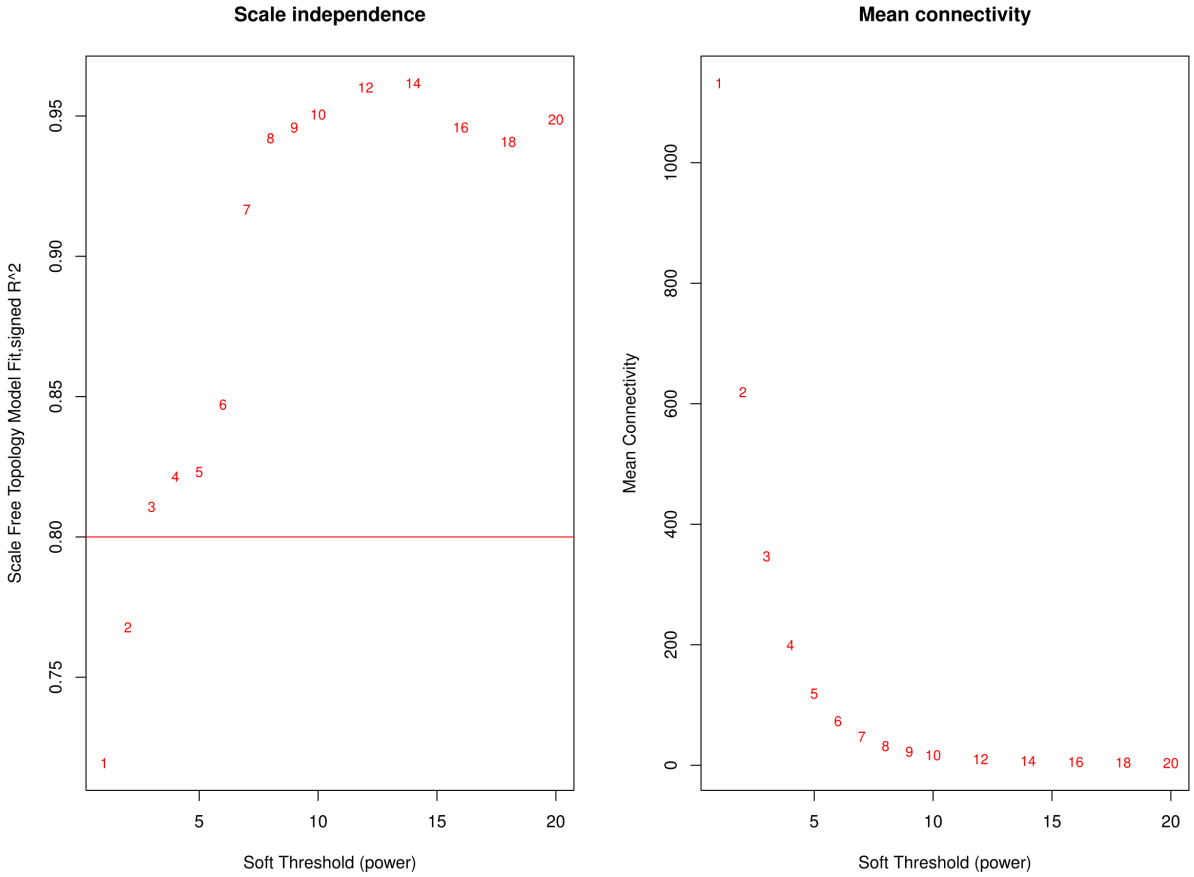
A)



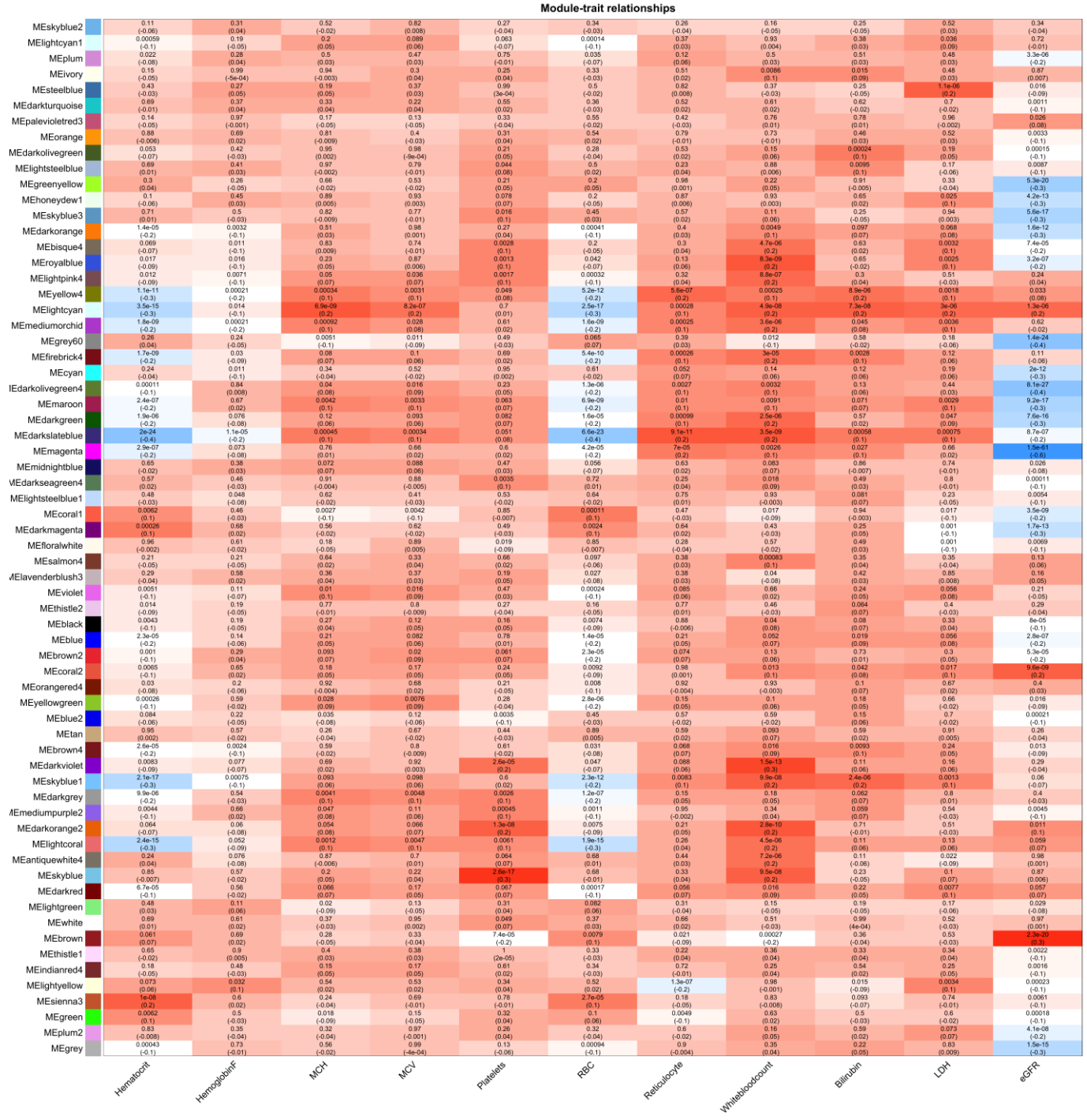
B)



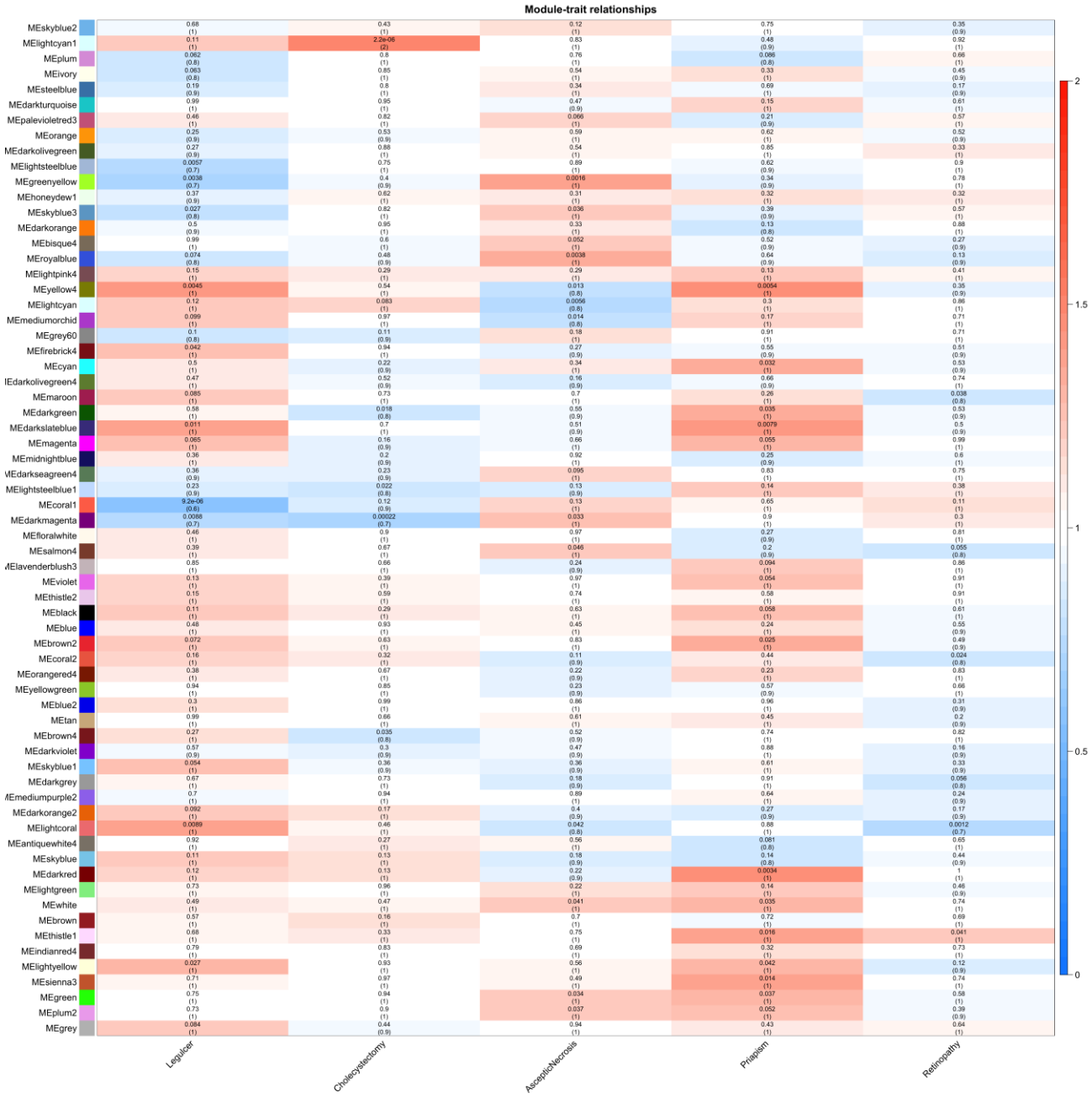
Supplementary Figure 2. Analysis of metabolomic network topology for various soft-thresholding powers. The scale-free fit index (*y*-axis) as a function of the soft-thresholding power (*x*-axis) and (B) the mean connectivity (degree, *y*-axis) as a function of the soft-thresholding power (*x*-axis).



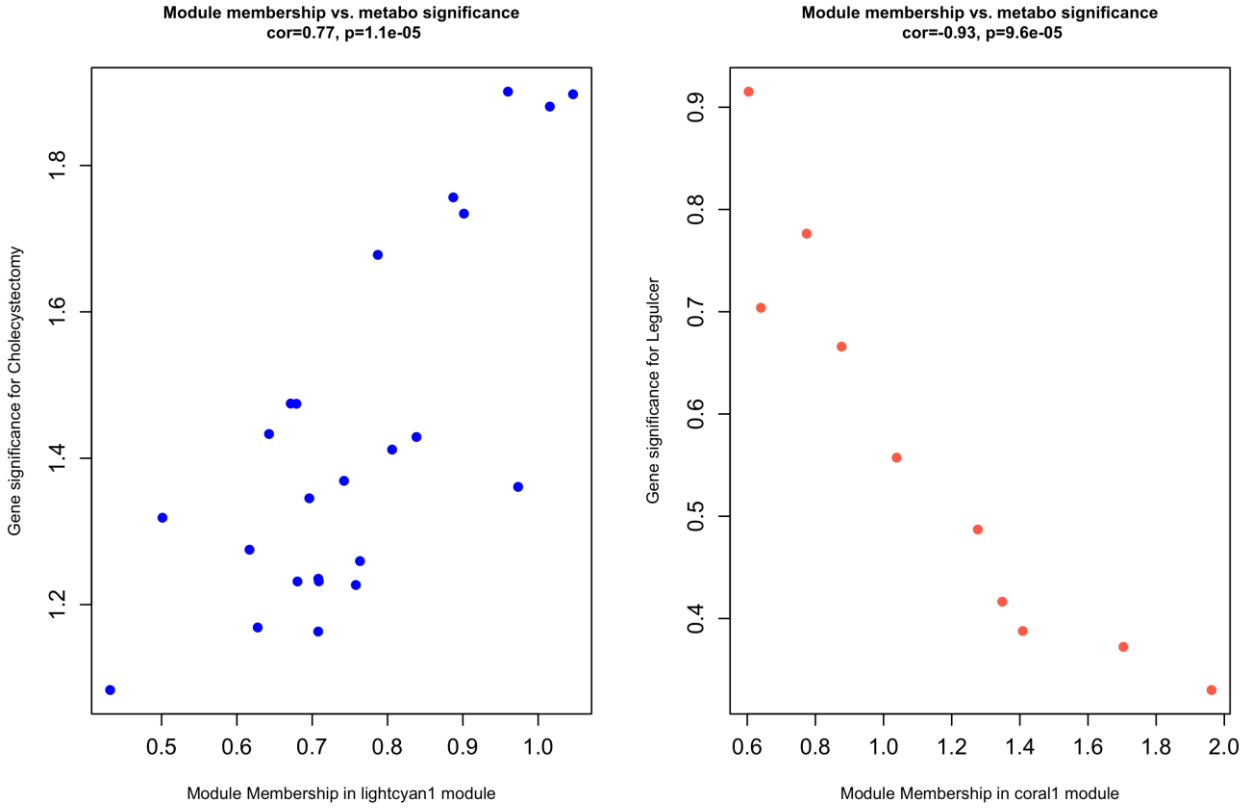
Supplementary Figure 3. Module association with blood traits. The weighted gene correlation network analysis (WGCNA) for 688 SCD patients. Dendrogram of all metabolites clustered based on the measurement of dissimilarity (1-TOM). The color band shows the results obtained from the automatic single-block analysis. In total, 66 metabolite modules were constructed. Each row corresponds to a module eigenmetabolite, column to a blood trait. Each cell contains the corresponding beta and p-value. The table is color-coded by correlation according to the color legend.



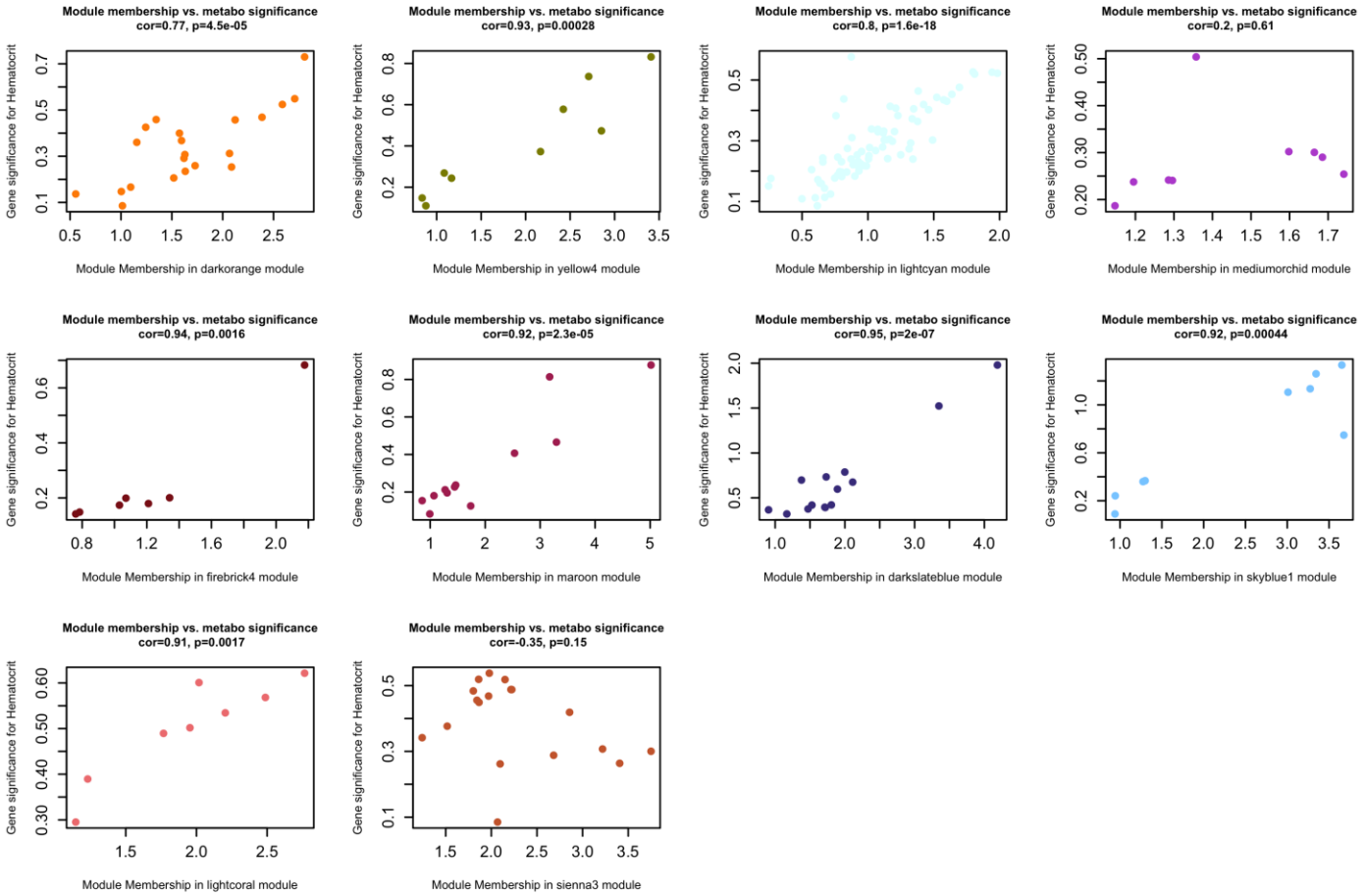
Supplementary Figure 4. Module metabolites association with complications. The weighted gene correlation network analysis (WGCNA) for 688 SCD patients. Dendrogram of all metabolites clustered based on the measurement of dissimilarity (1-TOM). The color band shows the results obtained from the automatic single-block analysis. In total, 66 metabolite modules were constructed. Each row corresponds to a module eigenmetabolite, column to complications. Each cell contains the corresponding beta and p-value. The table is color-coded by correlation according to the color legend.



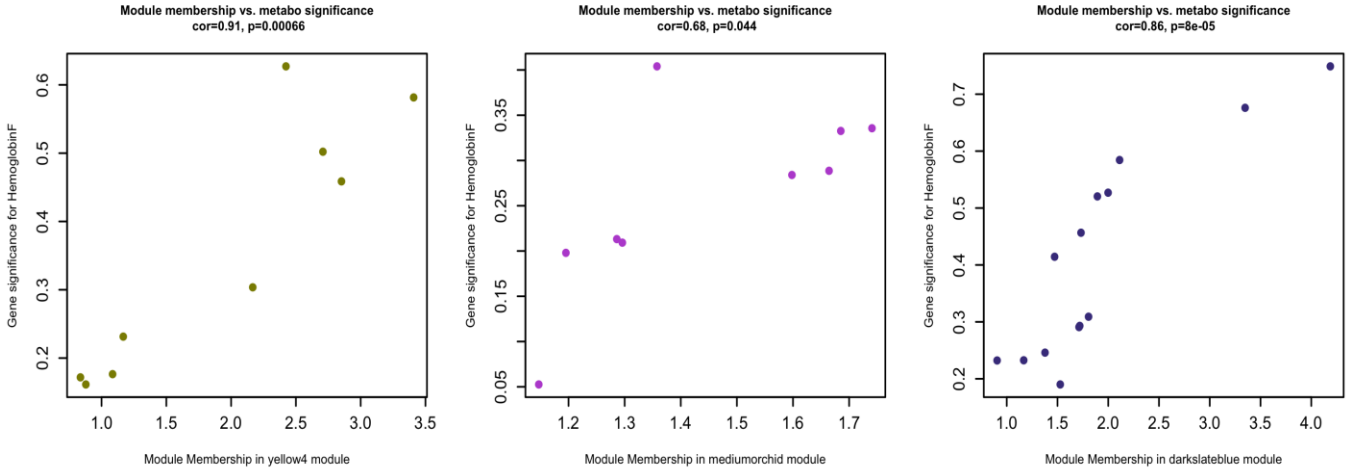
Supplementary Figure 5. Intramodular strength analysis of cholecystectomy and leg ulcer. The scatter plot of metabolite significance (*MS*) for cholecystectomy, leg ulcer, and *lightcyan1 and coral1* module membership (*MM*).



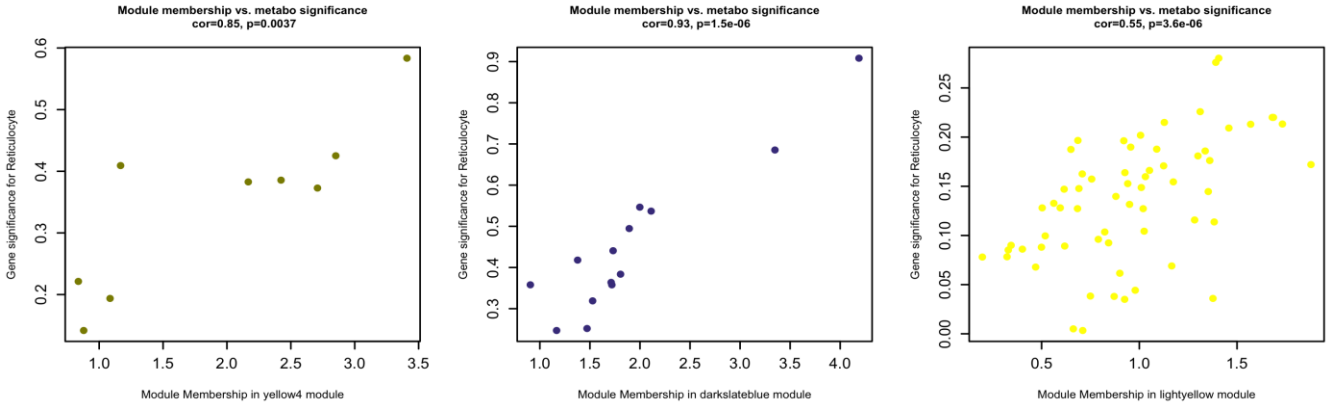
Supplementary Figure 6. Intramodular strength analysis of hematocrit. The scatter plot of metabolite significance (*MS*) for hematocrit levels and darkorange, yellow4, lightcyan, mediumorchid, firebrick4, maroon, darkslateblue, skyblue1, lightcoral, and sienna3 module membership (*MM*).



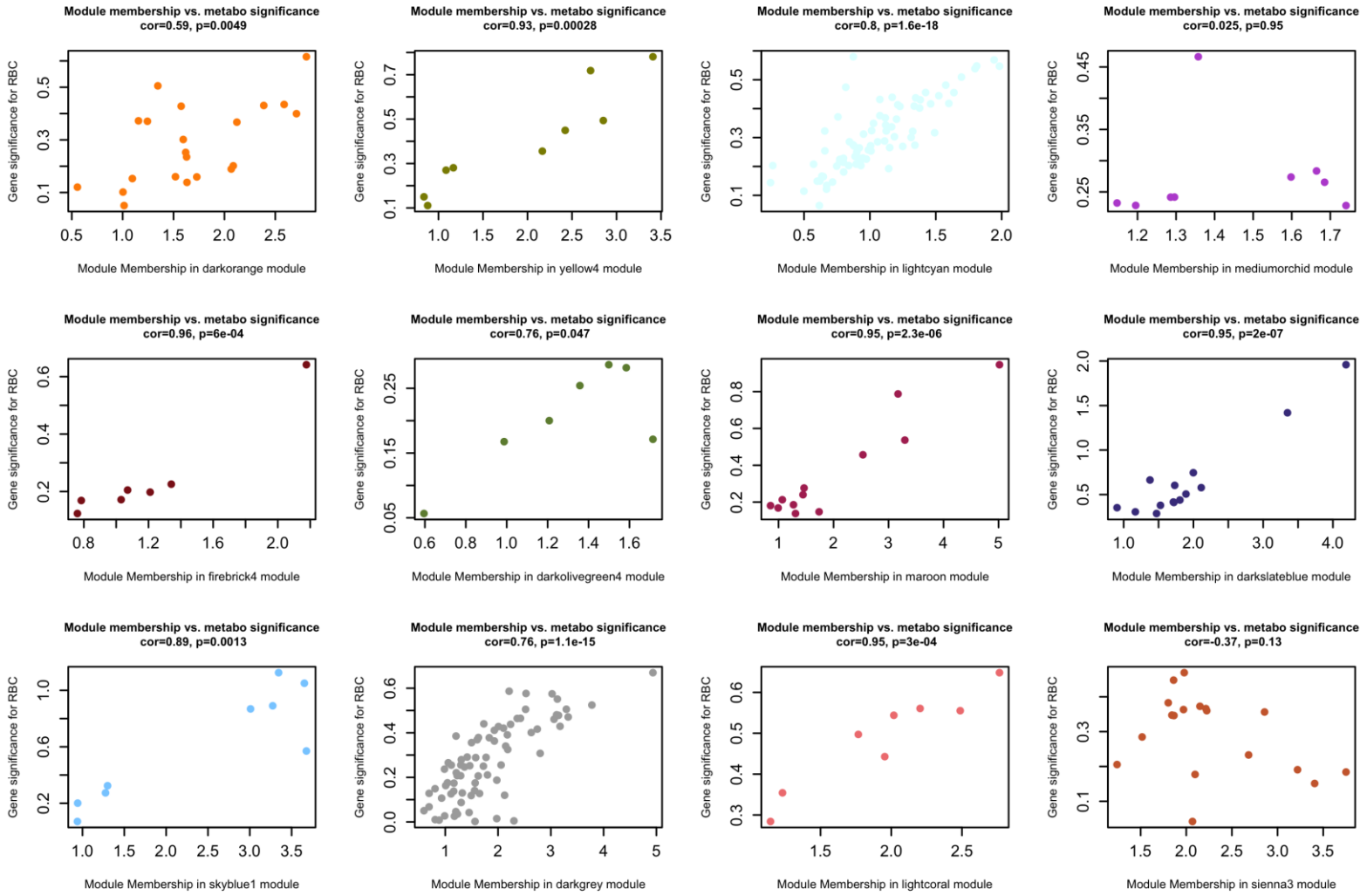
Supplementary Figure 7. Intramodular strength analysis of fetal hemoglobin. The scatter plot of metabolite significance (*MS*) for fetal hemoglobin levels, and yellow4, mediumorchid, darkslateblue, module membership (*MM*).



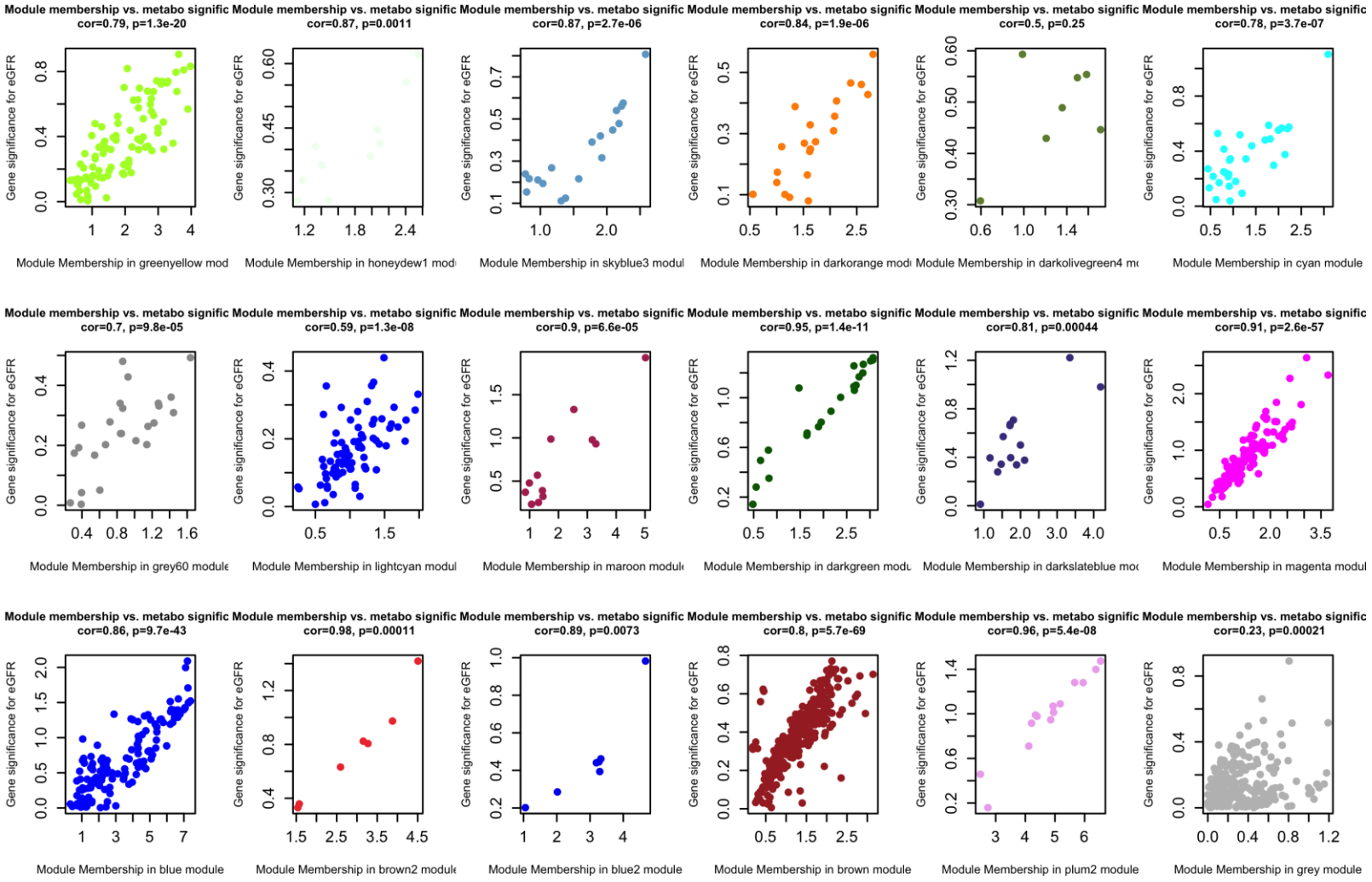
Supplementary Figure 8. Intramodular strength analysis of reticulocyte count. The scatter plot of metabolite significance (*MS*) for reticulocyte count levels and yellow4, darkslateblue, and lightyellow module membership (*MM*).



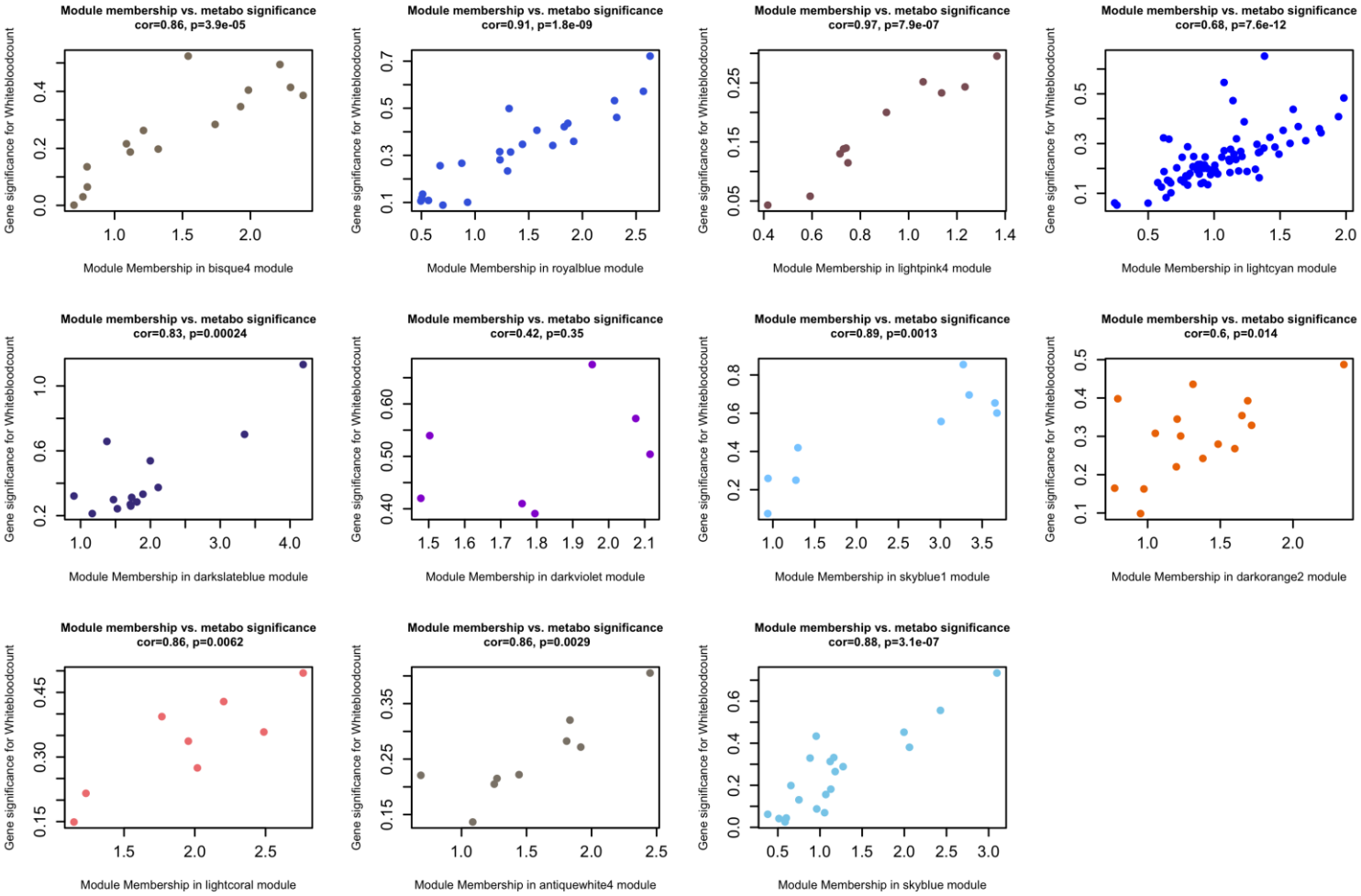
Supplementary Figure 9. Intramodular strength analysis of red blood cell count. The scatter plot of metabolite significance (*MS*) for red blood cell count, and darkorange, yellow4, lightcyan, mediumorchid, firebrick4, darkolivegreen4, maroon, darkslateblue, skyblue, darkgrey, lightcoral, sienna3 module membership (*MM*).



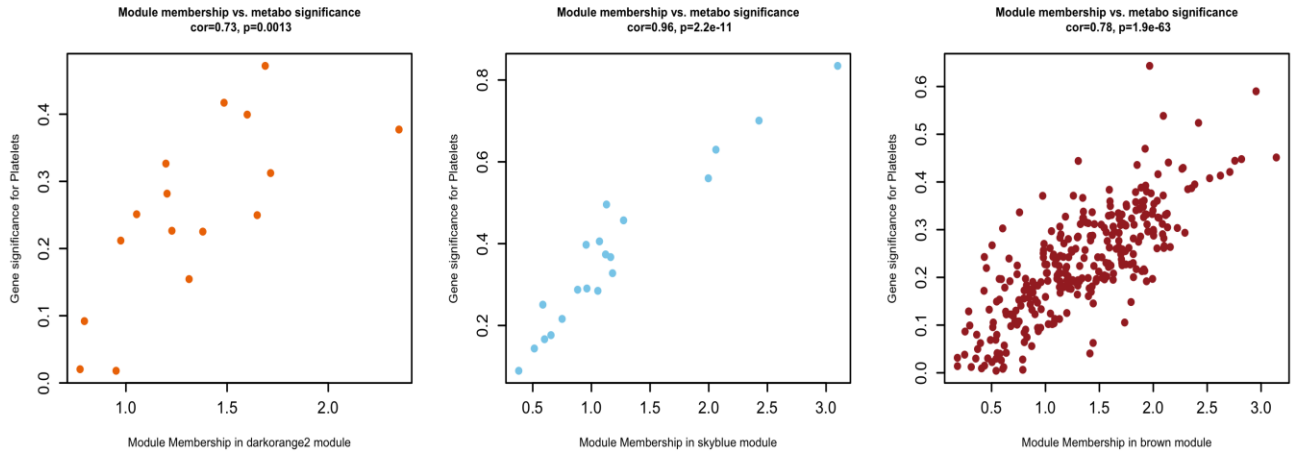
Supplementary Figure 10. Intramodular strength analysis of eGFR. The scatter plot of metabolite significance (*MS*) for eGFR, and plum, greenyellow, honeydew1, skyblue3, darkorange, royalblue, lightcyan, grey60, cyan, darkolivegreen4, maroon, darkgreen, darkslateblue, magenta, coral1, darkmagenta blue, coral2, brown, plum2, grey module membership(*MM*).



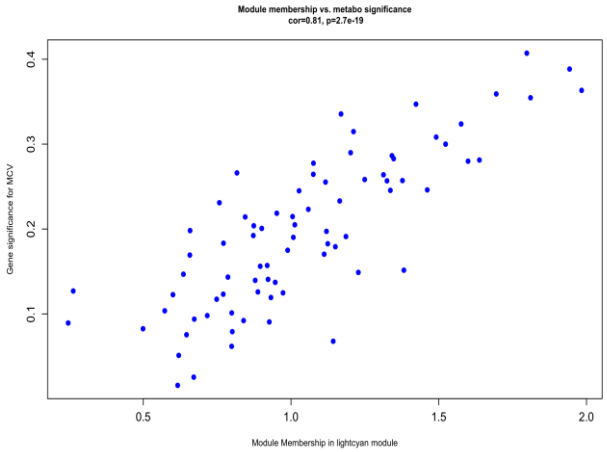
Supplementary Figure 11. Intramodular strength analysis of white blood cell count. The scatter plot of metabolite significance (*MS*) for white blood cell count, and bisque4, royalblue, lightpink4, lightcyan, darkorange2, skyblue1, darkslateblue, skyblue1, darkviolet, lightcoral, skyblue, lightcoral module membership (*MM*).

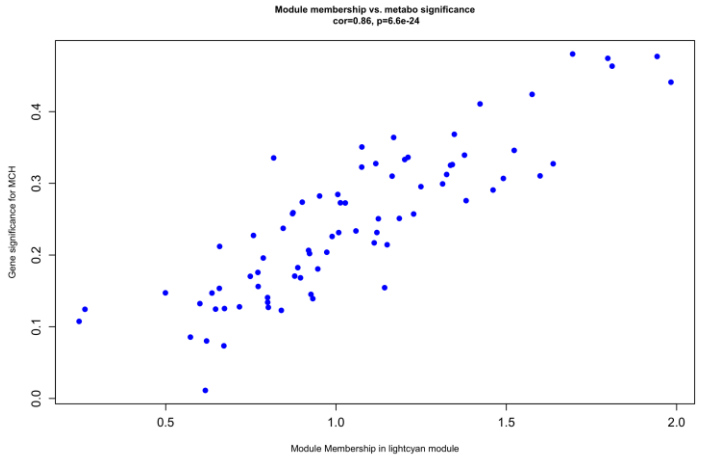


Supplementary Figure 12. Intramodular strength analysis of platelets. The scatter plot of metabolite significance (*MS*) for platelet, and darkorange2, skyblue, brown module membership (*MM*).

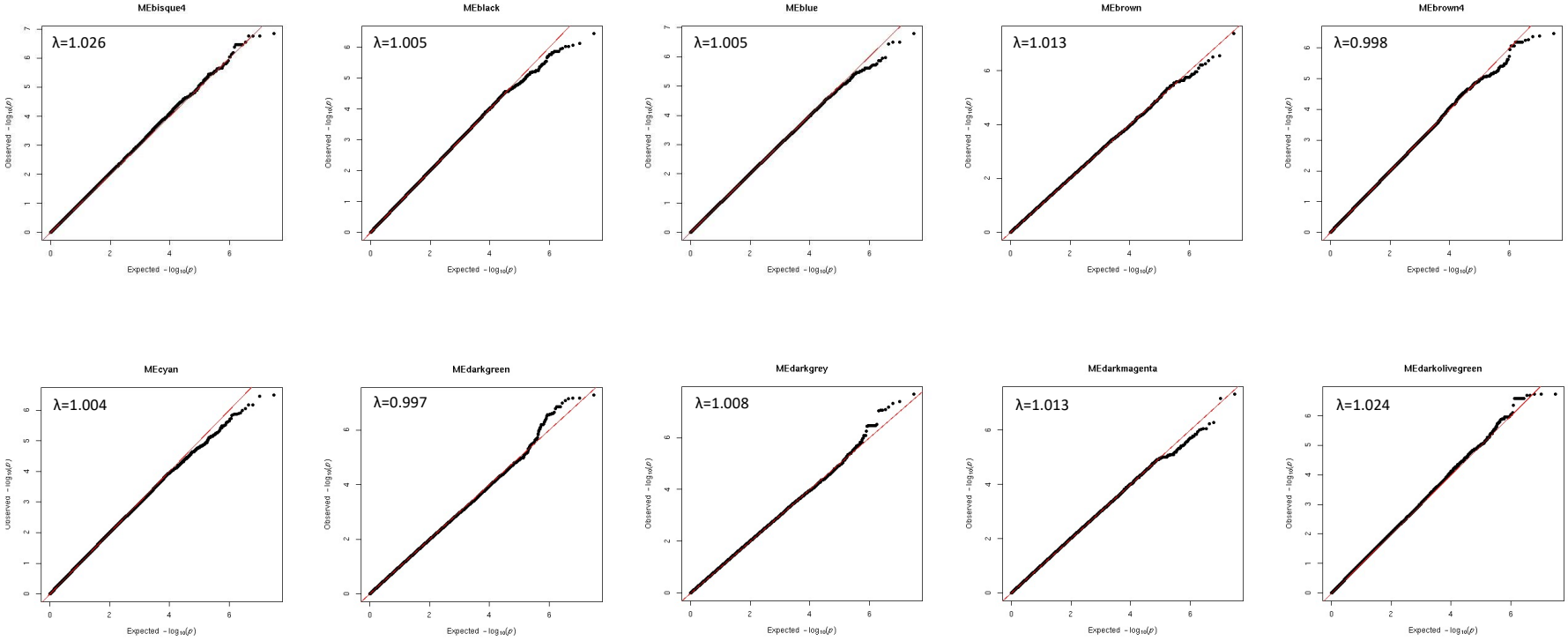


Supplementary Figure 13. Intramodular strength analysis of MCV & MCH. The scatter plot of metabolite significance (*MS*) for MCV & MCH and lightcyan module membership (*MM*).

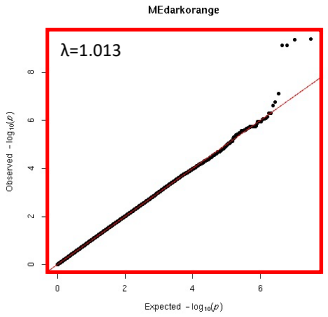

Module membership vs. metabo significance
cor=0.81, p=2.7e-19
Gene significance for MCV
Module Membership in lightcyan module
0.1
0.2
0.3
0.4
0.5
1.0
1.5
2.0

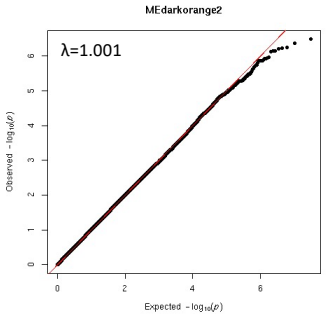

Module membership vs. metabo significance
cor=0.86, p=6.6e-24
Gene significance for MCH
Module Membership in lightcyan module
0.0
0.1
0.2
0.3
0.4
0.5
1.0
1.5
2.0

Supplementary Figure 14. Quantile-quantile plots of genotype module associations. Associations for each module PC1 were performed within each cohort (GENMOD and OMG). GWAS in GENMOD employed a linear regression implemented in RVTEST adjusting for age and sex. GWAS in OMG used a linear regression implemented in SNPTEST adjusting for age, sex, hemoglobin genotype, and hydroxyurea usage status. Meta-analysis of summary statistics combined over 15 million SNPs with MAF > 1% and employed METAL. Abbreviations: FDR; false discovery rate.

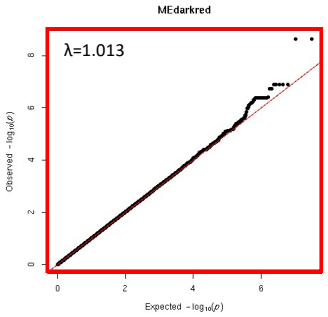

MEbisque4
λ=1.026
MEblack
λ=1.005
MEblue
λ=1.005
MEbrown
λ=1.013
MEbrown4
λ=0.998
MEcyan
λ=1.004
MEdarkgreen
λ=0.997
MEdarkgrey
λ=1.008
MEdarkmagenta
λ=1.013
MEdarkolivegreen
λ=1.024
Expected −log10(p)
Observed −log10(p)

FDR p<0.2

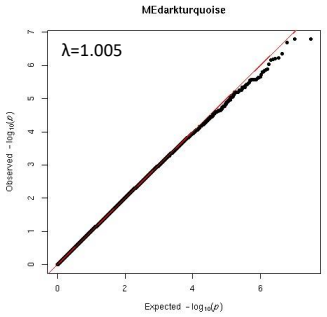


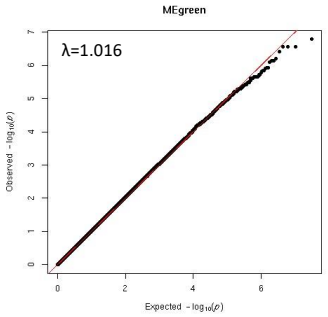


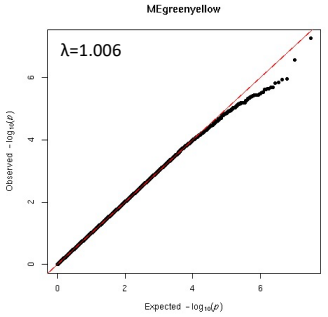
FDR p<0.05

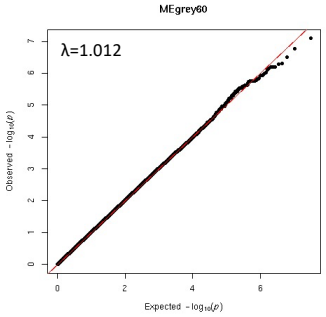


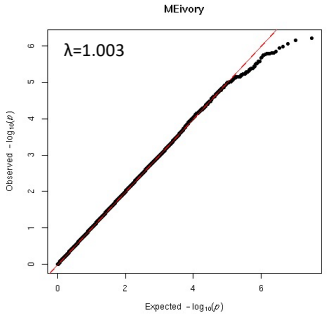
FDR p<0.05

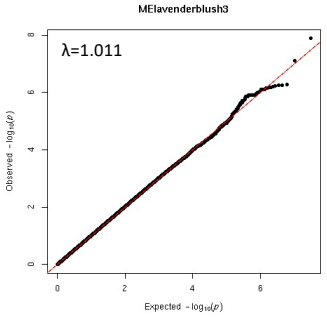




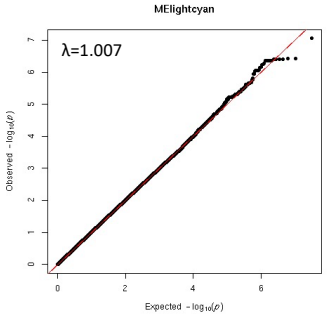


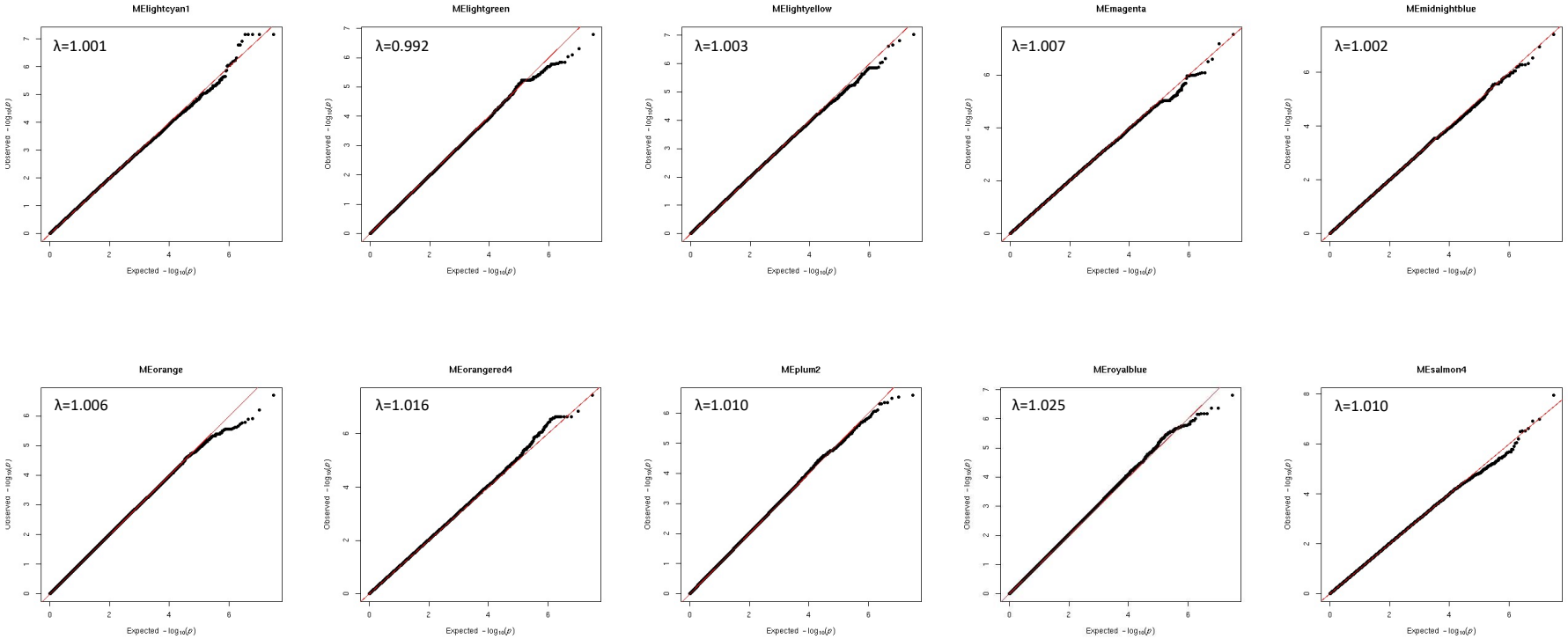




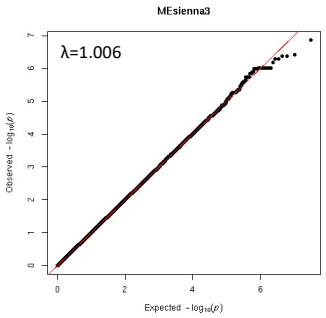


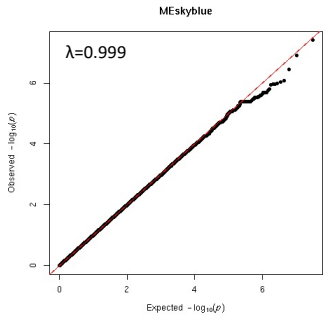
FDR p<0.2

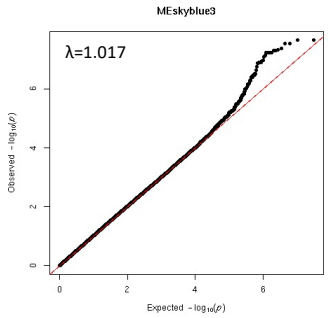



MElightcyan1
λ=1.001
MElightgreen
λ=0.992
MElightyellow
λ=1.003
MEmagenta
λ=1.007
MEmidnightblue
λ=1.002
MEorange
λ=1.006
MEorangered4
λ=1.016
MEplum2
λ=1.010
MEroyalblue
λ=1.025
MEsalmon4
λ=1.010
Observed −log10(p)
Expected −log10(p)

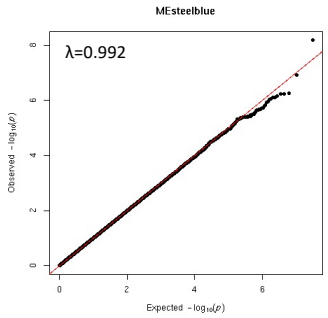
FDR p<0.2



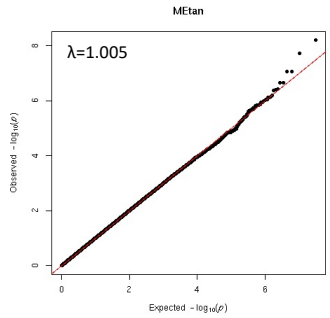




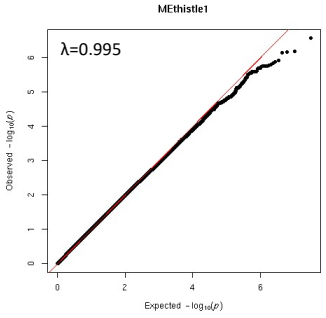
FDR p<0.1

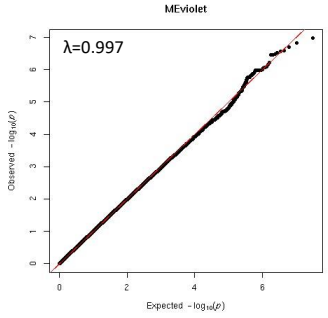


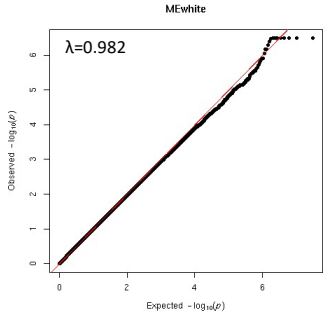
FDR p<0.1

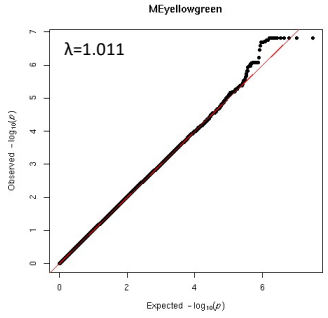


FDR p<0.1









FDR p<0.2

Annex D Supplementary Information for "Potential causal role of l-glutamine in sickle cell disease painful crises: A Mendelian randomization analysis

The article presented is published in the journal, *Blood Cells, Molecules, and Diseases*. In this article, we used MR to test the causal relationship between l-glutamine levels and painful crises in SCD patients. We identified 66 metabolites that are associated with SCD complications (e.g. gall bladder disease, renal dysfunction). Our approach illustrates the power to combine genetics and metabolomics to understand SCD pathophysiology.

SUPPLEMENTARY MATERIALS

Potential causal role of L-glutamine in sickle cell disease painful crises: a Mendelian randomization analysis

Yann Iboudo[1,2], Melanie E. Garrett[3], Pablo Bartolucci[4], Carlo Brugnara[5], Clary B. Clish[6], Joel N. Hirschhorn[6,7], Frédéric Galactéros[4], Allison E. Ashley-Koch[3], Marilyn J. Telen[8], Guillaume Lettre[1,2]

Affiliations

[1]Montreal Heart Institute, Montréal, Québec, Canada.

[2]Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada

[3]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, North Carolina, USA

[4]Red Cell Genetic Disease Unit, Hôpital Henri-Mondor, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Est, IMRB - U955 - Equipe no 2, Créteil, France

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

[6]Broad Institute, Cambridge, MA, USA

[7]Boston Children's Hospital, Boston, MA, USA

[8]Department of Medicine, Division of Hematology, Duke University Medical Center, Durham, North Carolina, USA

Corresponding author

Guillaume Lettre

Montreal Heart Institute

5000 Belanger Street

Montreal, Quebec, Canada

Tel: 514-376-3330

Fax: 514-593-2539

Email: guillaume.lettre@umontreal.ca.

Supplementary Table 3. Genetic associations for bilirubin levels and cholecystectomy. SNPs (effect alleles) are the lead variants in each gene region identified in a GWAS for bilirubin levels in European-ancestry individuals (Johnson, A. D. *et al*. (2009)). Association effect sizes with bilirubin levels in the CSSCD are in standard deviation units, and association effect sizes with bilirubin levels in Johnson et *al*. are in log-transformed bilirubin units. Association effect size with cholecystectomy in the CSSCD are log odds ratios from logistic regression analyses. These variants were selected because of their strong association ($P<2.0 \times 10^{-5}$) with bilirubin levels in Johnson et *al*. *These variants were identified in the Johnson et *al*. study, but because their minor allele frequency (MAF) <1% they were excluded from the Mendelian randomization analysis.

		Association result with bilirubin-associated variants (effect allele frequency, effect size (β), P-value, sample size)		
SNP	**Gene region**	**Bilirubin (Johnson *et al*)**	**Bilirubin (CSSCD)**	**Cholecystectomy Risk (CSSCD)**
rs6742078(T)	UGT1A1	(0.31, 0.23, P = 5e-324, N = 8988)	(0.44, 0.43, P = 9.7e-21, N = 930)	(0.43, 0.44, P = 0.00091, N = 1084)
rs12714207(T)	KRCC1	(0.33, -0.033, P = 5.3e-07, N = 8988)	(0.52, 0.047, P = 0.31, N = 930)	(0.53 -0.065, P = 0.63, N = 1084)
rs1986655(A)	intergenic	(0.15, -0.035, P = 2e-06, N = 8988)	(.96, 0.026, P = 0.82, N = 930)	(0.96, 0.33, P = 0.33, N = 1084)
rs12206204(T)/ rs113892814(A)*	histone cluster	(0.015, 0.15, P = 7.5e-07, N = 8988)	(0.0022, -0.27, P = 0.59, N = 930)	(0.00046/0.0028*, 1.8, P = 0.28, N = 1084)
rs9380833(T)	KCNK5	(0.027, 0.079, P = 1.6e-05, N = 8988)	(0.11, -0.022, P = 0.76, N = 930)	(0.11, 0.12, P = 0.55, N = 1084)
rs4236644(A)	SEMA3C	(0.27, -0.031, P = 2.1e-06, N = 8988)	(0.52, -0.065, P = 0.15, N = 930)	(0.51, 0.039, P = 0.77, N = 1084)
rs12337836(A)*	PRG-3, BAAT	(0.076, 0.053, P = 1.3e-05, N = 8988)	(0.0054, -0.14, P = 0.66, N = 930)	(0.0055, 0.03, P = 0.98, N = 1084)
rs16928809(A)*	SLC22A18	(0.096, 0.051, P = 1.1e-07, N = 8988)	(0.0038, 0.11, P = 0.77, N = 930)	(0.0037, 0.99, P = 0.35, N = 1084)
rs4149056(T)	SLCO1B1	(0.15, -0.053, P = 6.7e-13, N = 8988)	(0.978, -0.26, P = 0.11, N = 930)	(0.98, -0.23, P = 0.63, N = 1084)
rs4773330(A)	ARHGEF7	(0.12, -0.036, P = 7.7e-06, N = 8988)	(0.17, 0.03, P = 0.63, N = 930)	(0.17, 0.15, P = 0.41, N = 1084)
rs7173819(A)	intergenic	(0.12, 0.035, P = 1.2e-05, N = 8988)	(0.79, -0.021, P = 0.71, N = 930)	(0.79, 0.19, P = 0.26, N = 1084)
rs12923103(A)	intergenic	(0.32, 0.027, P = 1.4e-05, N = 8988)	(0.17, 0.037, P = 0.55, N = 930)	(0.17, 0.12, P = 0.49, N = 1084)
rs4410172 (C)	BC051727	(0.24, 0.026, P = 1.9e-05, N = 8988)	(0.097, -0.11, P = 0.16, N = 930)	(0.098, 0.21, P = 0.37, N = 1084)

Supplementary Table 4. Mendelian randomization (MR) results of bilirubin levels with cholecystectomy. As instruments for our MR analyses, we used SNPs identified by Johnson et *al*. in a genome-wide association study (GWAS) for bilirubin levels in 9,464 individuals of European ancestry(Johnson, A. D. *et al*. (2009)).. From the list of 15 bilirubin-associated variants from the GWAS, we kept 10 SNPs that were in linkage equilibrium (rs6742078, rs12714207, rs1986655, rs9380833, rs4236644, rs4149056, rs4773330, rs7173819, rs12923103, rs4410172). We analyzed cholecystectomy in 1,294 participants from the CSSCD. We performed a two-sample MR analysis using exposure effect sizes from the bilirubin GWAS of Johnson et *al*. (GWAS_beta). For the two-sample MR analysis, effect sizes of the MR are odds to develop cholecystectomy per 27% increase in bilirubin levels.

Method	Number of variants	Odds ratio (95% confidence interval) and P-value
Inverse variance-weighted (IVW)	10	6.0 (2.8-17) $P=1.9 \times 10^{-6}$
MR-Egger	10	7.0 (1.7-29) P=0.027
Weighted median	10	6.4 (2.1-20) $P=1.2 \times 10^{-3}$

Supplementary Table 5. Genetic associations between 51 L-glutamine-associated SNPs and painful crises. SNPs (effect alleles) are the lead variants in each gene region identified in mGWAS for L-glutamine levels in Europeans (P <5.0x10^{-5})[253,254] Shin, S. Y. *et al*. 2014, Long, T. *et al*. 2017. Association effect sizes with painful crises in the CSSCD and GEN-MOD are log odds ratios from logistic regression. *These variants were identified to be pleiotropic by Phenoscanner queries and were excluded from model 2 (**Methods**). At the bottom of the table, we also provide association results between polygenic trait scores (PTS) calculated using 51 L-glutamine-associated SNPs (PTS_{51SNPs}) or after excluding pleiotropic variants (PTS_{27SNPs}). For the PTS, the effect size is per PTS standard deviation units.

		Association result with L-glutamine-associated variants (effect allele frequency, effect size (beta or OR), P-value, sample size			
SNP (hg19)	**Gene region**	**L-glutamine (Shin *et al*/Long *et al*)**	**L-glutamine (Meta-analysis GEN-MOD-OMG)**	**Painful crises (CSSCD)**	**Painful crises (Meta-analysis GEN-MOD-OMG)**
rs524219(G)/1:100821493	CDC14A	(0.93, -0.28, P = 4.7e-06, N = 1958)	(0.02, -0.21, P= 0.33, N = 651)	(0.99 ,0.59, P = 0.24, N = 1101)	(0.98, -0.91, P = 0.22, N = 575)
rs11166473(T)/1:101010015*	GPR88	(0.92, -0.36, P = 9.3e-08, N = 1958)	(0.48, 0.01, P= 0.85, N = 651)	(0.47 ,-0.12, P = 0.31, N = 1101)	(0.49, -0.11, P = 0.56, N = 575)
rs2811981(A)/1:23950147*	MDS2	(0.89, -0.12, P = 3.8e-06, N = 1958)	(0.87, 0.09, P= 0.28, N = 651)	(0.88 ,-0.12, P = 0.49, N = 1101)	(0.87, -0.11, P = 0.69, N = 575)
rs3127550(A)/1:49431909*	AGBL4	(0.49, 0.19, P = 5.7e-06, N = 1958)	(0.18, -0.02, P= 0.77, N = 651)	(0.17 ,-0.021, P = 0.89, N = 1101)	(0.19, -0.52, P = 0.03, N = 575)
rs7394051(T)/10:122008026*	intergenic	(0.7, 0.0055, P = 2.2e-06, N = 7372)	(0.72, 0.06, P= 0.39, N = 651)	(0.76 ,-0.052, P = 0.7, N = 1101)	(0.72, 0.14, P = 0.51, N = 575)
rs10762121(T)/10:68708291	CTNNA3	(0.18, -0.0094, P = 3.1e-05, N = 1768)	(0.44, 0.04, P= 0.45, N = 651)	(0.4 ,-0.018, P = 0.88, N = 1101)	(0.44, 0.04, P = 0.84, N = 575)
rs11596604(A)/10:87141944*	intergenic	(0.024, 0.032, P = 2.5e-06, N = 1768)	(0, -0.43, P= 0.46, N = 401)	(0.01 ,0.16, P = 0.78, N = 1101)	(0, -2.5, P = 0.07, N = 325)
rs7078003(T)/10:99359412*	HOGA1	(0.17, 0.0087, P = 1.8e-06, N = 7372)	(0.11, -0.01, P= 0.93, N = 651)	(0.12 ,-0.12, P = 0.5, N = 1101)	(0.12, -0.21, P = 0.48, N = 575)
rs7131407(T)/11:128431418	ETS1	(0.13, 0.0082, P = 3.2e-05, N = 7372)	(0.16, 0.02, P= 0.82, N = 651)	(0.16 ,0.29, P = 0.074, N = 1101)	(0.16, 0.19, P = 0.46, N = 575)
rs10431159(A)/11:70967219	SHANK2	(0.65, -0.0047, P = 4.8e-05, N = 7372)	(0.1, 0.09, P= 0.29, N = 651)	(0.15 ,-0.061, P = 0.71, N = 1101)	(0.1, -0.06, P = 0.86, N = 575)
rs17666239(A)/12:47194757*	SLC38A4	(0.078, 0.0092, P = 1.7e-05, N = 7372)	(0.03, 0.1, P= 0.53, N = 651)	(0.035 ,0.24, P = 0.45, N = 1101)	(0.03, -0.69, P = 0.22, N = 575)
rs735246(A)/12:52547743	intergenic	(0.83, 0.0083, P = 4.8e-05, N = 7372)	(0.82, -0.05, P= 0.49, N = 651)	(0.81 ,-0.11, P = 0.42, N = 1101)	(0.82, -0.08, P = 0.74, N = 575)

rs774044(T)/12:56837979*	TIMELESS	(0.055, -0.015, P = 5.6e-07, N = 7372)	(0.05, -0.11, P= 0.4, N = 651)	(0.045 ,0.093, P = 0.74, N = 1101)	(0.05, -0.17, P = 0.7, N = 575)
rs7313455(A)/12:56853231*	MIP	(0.44, 0.0062, P = 2.1e-08, N = 7372)	(0.1, 0.11, P= 0.29, N = 651)	(0.13 ,-0.026, P = 0.88, N = 1101)	(0.06, 0.2, P = 0.63, N = 575)
rs2657879(A)/12:56865338*	GLS2	(0.82, 0.015, P = 6.1e-18, N = 7372)	(0.96, 0.19, P= 0.15, N = 651)	(0.94 ,0.25, P = 0.32, N = 1101)	(0.96, -0.4, P = 0.41, N = 575)
rs12232026(A)/12:56960766*	RBMS2	(0.12, -0.0097, P = 4.4e-06, N = 7372)	(0.17, 0.12, P= 0.1, N = 651)	(0.14 ,-0.32, P = 0.049, N = 1101)	(0.16, 0.12, P = 0.64, N = 575)
rs941893(T)/14:100542061*	EVL	(0.74, -0.0049, P = 4.1e-05, N = 7372)	(0.26, -0.01, P= 0.86, N = 651)	(0.27 ,0.013, P = 0.92, N = 1101)	(0.25, -0.08, P = 0.72, N = 575)
rs144325715(A)/14:95948631	intergenic	(0.021, -0.74, P = 5.3e-06, N = 1958)	(0.04, -0.12, P= 0.46, N = 651)	(0.044 ,0.25, P = 0.4, N = 1101)	(0.04, 0.42, P = 0.38, N = 575)
rs11636988(A)/15:26822814*	GABRB3	(0.57, -0.16, P = 6.5e-06, N = 1958)	(0.08, 0.14, P= 0.18, N = 651)	(0.13 ,-0.0023, P = 0.99, N = 1101)	(0.07, -0.25, P = 0.53, N = 575)
rs1910151(A)/15:38158699*	NA	(0.64, 0.0048, P = 2.2e-05, N = 7372)	(0.64, -0.05, P= 0.42, N = 651)	(0.64 ,0.11, P = 0.36, N = 1101)	(0.64, 0.34, P = 0.08, N = 575)
rs35150605(TAAC)/15:88047276	RP11-648K4.2	(0.25, -0.15, P = 6.3e-06, N = 1958)	(0.7, -0.03, P= 0.58, N = 651)	(0.3 ,0.037, P = 0.77, N = 1101)	(0.28, 0.09, P = 0.66, N = 575)
rs2560409(C)/16:24099496	PRKCB	(0.51, 0.26, P = 1.4e-06, N = 1958)	(0.84, 0.19, P= 0.01, N = 651)	(0.18 ,-0.12, P = 0.43, N = 1101)	(0.14, -0.41, P = 0.13, N = 575)
rs16977047(T)/16:27924612*	GSG1L	(0.7, -0.0056, P = 1.1e-06, N = 7372)	(0.78, -0.01, P= 0.84, N = 651)	(0.77 ,-0.11, P = 0.43, N = 1101)	(0.8, -0.12, P = 0.61, N = 575)
rs12447776(T)/16:84053027	SLC38A8	(0.019, 0.029, P = 4.1e-05, N = 1768)	(0.05, -0.09, P= 0.49, N = 651)	(0.042 ,0.057, P = 0.85, N = 1101)	(0.06, -0.19, P = 0.65, N = 575)
rs9912445(A)/17:37202603	LRRC37A11P	(0.79, 0.0052, P = 1.7e-05, N = 7372)	(0.71, 0.08, P= 0.17, N = 651)	(0.72 ,-0.17, P = 0.17, N = 1101)	(0.7, 0.3, P = 0.12, N = 575)
rs8069305(A)/17:53638670	CTD-2033D24.2	(0.68, -0.0047, P = 4.9e-05, N = 7372)	(0.14, 0.1, P= 0.18, N = 651)	(0.19 ,0.031, P = 0.84, N = 1101)	(0.16, 0.17, P = 0.49, N = 575)
rs4798682(G)/18:8738265	SOGA2	(0.74, 0.15, P = 5e-06, N = 1958)	(0.11, -0.15, P= 0.1, N = 651)	(0.87 ,-0.17, P = 0.33, N = 1101)	(0.89, 0.2, P = 0.52, N = 575)
rs73971292(C)/2:161821810	intergenic	(0.023, -0.65, P = 5.1e-06, N = 1958)	(0.97, 0.13, P= 0.43, N = 651)	(0.025 ,0.15, P = 0.69, N = 1101)	(0.02, -0.68, P = 0.3, N = 575)
rs780093(T)/2:27742603*	GCKR	(0.4, -0.0058, P = 1.9e-07, N = 7372)	(0.12, -0.07, P= 0.42, N = 651)	(0.14 ,0.15, P = 0.37, N = 1101)	(0.12, 0.03, P = 0.93, N = 575)
rs2199619(T)/2:28854958*	PLB1	(0.28, 0.0054, P = 5.2e-06, N = 7372)	(0.21, -0.11, P= 0.12, N = 651)	(0.22 ,-0.11, P = 0.45, N = 1101)	(0.19, -0.01, P = 0.97, N = 575)
rs992580(C)/20:15246472*	MACROD2	(0.44, 0.15, P = 6.2e-06, N = 1958)	(0.53, 0.0036, P= 0.95, N = 651)	(0.47 ,-0.0083, P = 0.94, N = 1101)	(0.47, -0.12, P = 0.52, N = 575)
rs6137021(A)/20:20375560*	RALGAPA2	(0.67, -0.0083, P = 4.5e-05, N = 1768)	(0.2, 0.17, P= 0.01, N = 651)	(0.23 ,-0.32, P = 0.023, N = 1101)	(0.18, 0.45, P = 0.05, N = 575)
rs2425059(C)/20:33912371*	UQCC1	(0.38, 0.15, P = 5.6e-06, N = 1958)	(0.4, -0.01, P= 0.91, N = 651)	(0.56 ,-0.12, P = 0.31, N = 1101)	(0.63, -0.35, P = 0.08, N = 575)
rs2948828(A)/3:124811942	SLC12A8	(0.56, 0.0053, P = 2.2e-05, N = 7372)	(0.75, -0.0033, P= 0.96, N = 651)	(0.75 ,0.056, P = 0.67, N = 1101)	(0.76, -0.19, P = 0.39, N = 575)
rs73168973(T)/3:151691599	intergenic	(0.21, -0.23, P = 4e-07, N = 1958)	(0.06, 0.03, P= 0.81, N = 651)	(0.079 ,-0.18, P = 0.42, N = 1101)	(0.06, -0.46, P = 0.28, N = 575)

rs4699183(A)/4:106444435	AC004066.2	(0.9, -0.27, P = 6.2e-06, N = 1958)	(0.93, 0.29, P= 0.0086, N = 651)	(0.92 ,0.082, P = 0.71, N = 1101)	(0.93, 0.51, P = 0.16, N = 575)
rs138354882(A)/4:171429817	intergenic	(0.028, -0.36, P = 4.5e-06, N = 1958)	(0.1, -0.06, P= 0.52, N = 651)	(0.09 ,0.24, P = 0.25, N = 1101)	(0.1, 0.04, P = 0.91, N = 575)
rs7667615(C)/4:182921992*	AC108142.1	(0.26, -0.17, P = 6.1e-06, N = 1958)	(0.81, -0.0035, P= 0.96, N = 651)	(0.17 ,-0.031, P = 0.84, N = 1101)	(0.18, 0.16, P = 0.52, N = 575)
rs542300(A)/6:12252237	intergenic	(0.51, -0.13, P = 7.2e-06, N = 1958)	(0.54, 0.07, P= 0.22, N = 651)	(0.54 ,-0.064, P = 0.59, N = 1101)	(0.54, 0.32, P = 0.07, N = 575)
rs9478369(A)/6:153269698*	intergenic	(0.25, 0.0049, P = 4.1e-05, N = 7372)	(0.56, 0.0019, P= 0.97, N = 651)	(0.51 ,0.052, P = 0.66, N = 1101)	(0.57, -0.14, P = 0.44, N = 575)
rs71569656(C)/6:22903627	RP1-209A6.1	(0.28, -0.14, P = 8.9e-06, N = 1958)	(0.92, -0.01, P= 0.91, N = 651)	(0.095 ,0.42, P = 0.04, N = 1101)	(0.07, -0.46, P = 0.23, N = 575)
rs2748991(T)/6:52596516	intergenic	(0.45, -0.007, P = 2.3e-06, N = 5604)	(0.33, 0.16, P= 0.0096, N = 651)	(0.34 ,-0.091, P = 0.45, N = 1101)	(0.33, -0.2, P = 0.34, N = 575)
rs1582256(C)/7:126634804	GRM8	(0.51, -0.18, P = 8.5e-06, N = 1958)	(0.5, -0.07, P= 0.19, N = 651)	(0.49 ,0.092, P = 0.42, N = 1101)	(0.5, 0.23, P = 0.21, N = 575)
rs767772939(CT)/7:128350744	FAM71F1	(0.096, 0.26, P = 4.1e-06, N = 1958)	(0.79, 0.01, P= 0.93, N = 651)	(0.82 ,-0.16, P = 0.29, N = 1101)	(0.8, 0.11, P = 0.66, N = 575)
rs17837468(A)/7:138309274	SVOPL	(0.88, 0.0088, P = 3.8e-05, N = 7372)	(0.74, 0.03, P= 0.58, N = 651)	(0.76 ,-0.14, P = 0.3, N = 1101)	(0.74, 0.17, P = 0.43, N = 575)
rs17152416(T)/7:25765238	AC003090.1	(0.76, 0.0049, P = 3.9e-05, N = 7372)	(0.7, 0.03, P= 0.64, N = 651)	(0.71 ,-0.18, P = 0.17, N = 1101)	(0.69, 0.24, P = 0.25, N = 575)
rs4722699(T)/7:27456500*	intergenic	(0.27, -0.0048, P = 3.8e-05, N = 7372)	(0.23, -0.05, P= 0.49, N = 651)	(0.2 ,-0.0076, P = 0.96, N = 1101)	(0.22, 0.07, P = 0.73, N = 575)
rs1799211(T)/7:76240677	UPK3B	(0.41, -0.0082, P = 4.2e-06, N = 1768)	(0.18, 0.04, P= 0.7, N = 401)	(0.24 ,0.037, P = 0.79, N = 1101)	(0.18, 0.31, P = 0.21, N = 325)
rs9314463(A)/8:2525166	RP11-134O21.1	(0.19, 0.27, P = 9.1e-07, N = 1958)	(0.45, -0.01, P= 0.83, N = 651)	(0.43 ,-0.073, P = 0.54, N = 1101)	(0.48, -0.05, P = 0.78, N = 575)
rs112508772(A)/9:140016354	snoU13	(0.025, 0.48, P = 7e-07, N = 1958)	(0.11, 0.14, P= 0.12, N = 651)	(0.056 ,-0.11, P = 0.67, N = 1101)	(0.13, 0.15, P = 0.58, N = 575)
rs7848854(C)/9:7766733*	intergenic	(0.11, 0.0083, P = 2.4e-05, N = 7372)	(0.02, -0.04, P= 0.85, N = 651)	(0.027 ,0.68, P = 0.06, N = 1101)	(0.02, -0.39, P = 0.58, N = 575)
PTS_{51SNPs}	NA	NA	(NA, 0.021, P=0.60, N=651)	(NA, -0.056, P=0.12, N=1101)	(NA, -0.022, P=0.80, N=575)
PTS_{27SNPs}	NA	NA	(NA, 0.025, P=0.53, N=651)	(NA, -0.081, P=0.021, N=1101)	(NA, 0.0036, P=0.97, N=575)

Supplementary Table 6. Mendelian randomization results for L-glutamine with sickle cell disease (SCD)-complications and estimated glomerular filtration rate (eGFR). For complications, estimates are odds ratios (95% confidence intervals) for the effect of a 1 standard deviation increase in L-glutamine. For eGFR (0.07 mL/min per 1.172 m^2), estimates are effect size (standard error) for the effect of a 1 standard deviation increase in L-glutamine. All 51 genetic variants that are associated with L-glutamine at $P<5x10^{-5}$ are included in Model 1 analyses (29 SNPs from Shin *et al.*, 22 SNPs from Long *et al.*). In Model 2, we only kept 27 variants that were not pleiotropic (12 from Shin *et al.*, 15 from Long *et al.*). IVW: inverse variance-weighted. In light grey, we present MR replication results for L-glutamine and painful crises in the smaller GEN-MOD and OMG cohorts.

Metabolite	Method	Painful crises (CSSCD)	Painful crises (GEN-MOD+OMG)	Cholecystectomy (CSSCD)	Retinopathy (CSSCD)	Leg ulcer (CSSCD)	Priapism (CSSCD)	Aseptic necrosis (CSSCD)	eGFR (CSSCD)
L-glutamine – Model 1	IVW	0.81 (0.63-1) P=0.086	0.77 (0.51- 1.16) P=0.21	0.94 (0.68-1.3) P=0.72	0.93 (0.67-1.3) P=0.66	1.1 (0.82-1.5) P=0.53	1 (0.78-1.4) P=0.81	0.88 (0.65-1.2) P=0.44	0.027 (0.057) P=0.64
	MR-Egger	0.76 (0.54-1.1) P=0.12	0.84 (0.5- 1.4) P=0.50	1.1 (0.76-1.7) P=0.52	0.85 (0.57-1.3) P=0.45	1.1 (0.75-1.6) P=0.66	1.1 (0.67-1.8) P=0.73	0.97 (0.65-1.4) P=0.88	0.061 (0.072) P=0.4
	Weighted median	0.77 (0.53-1.1) P=0.17	0.93 (0.55- 1.58) P=0.8	0.91 (0.58-1.4) P=0.67	0.82 (0.52-1.3) P=0.4	0.97 (0.62-1.5) P=0.88	1.1 (0.66-2) P=0.65	0.91 (0.59-1.4) P=0.67	0.015 (0.084) P=0.85
L-glutamine – Model 2	IVW	0.68 (0.52-0.89) P=0.0048	0.82 (0.5- 1.34) P=0.44	0.84 (0.57-1.2) P=0.39	1.2 (0.78-1.7) P=0.46	1.2 (0.88-1.6) P=0.26	1.1 (0.78-1.5) P=0.6	0.85 (0.61-1.2) P=0.37	-0.026 (0.079) P=0.74
	MR-Egger	0.74 (0.48-1.1) P=0.16	0.80 (0.42- 1.53) P=0.50	0.97 (0.59-1.6) P=0.92	1.1 (0.66-1.9) P=0.71	1.2 (0.75-1.9) P=0.46	1.2 (0.66-2.2) P=0.54	1 (0.64-1.6) P=0.96	0.012 (0.1) P=0.91
	Weighted median	0.73 (0.49-1.1) P=0.12	0.85 (0.44- 1.61) P=0.61	0.78 (0.46-1.3) P=0.36	1.2 (0.73-2.1) P=0.43	1.1 (0.66-1.8) P=0.72	1.2 (0.64-2.4) P=0.53	0.78 (0.46-1.3) P=0.37	-0.0049 (0.099) P=0.96

Supplementary Table 9. Associations between adenosine plasma levels and sickle cell disease (SCD)-related complications and eGFR. In participants from the GEN-MOD and OMG cohorts, we tested the association between adenosine levels measured in plasma and SCD-related complications and eGFR. Association results were tested per cohort and then meta-analyzed. Dichotomous phenotypes were analyzed using logistic regression while correcting for age, sex, hydroxyurea (HU) usage, SCD genotypes and cohort affiliation. Quantitative phenotypes were corrected for age, sex, HU usage, SCD genotypes. They were inverse normal-transformed before being tested for association using linear regression. Odds ratio and effect sizes (Beta) are given per standard deviation change in plasma L-glutamine plasma levels. eGFR, estimated glomerular filtration rate; CI, confidence interval; SE, standard error.

Complications	**N**	**Odds ratio**	**95% CI**	**P-value**
Painful crises	367	1.16	(0.94 -1.44)	0.17
Survival	267	1.08	(0.7 -1.67)	0.73
Aseptic necrosis	365	0.86	(0.68 -1.09)	0.21
Cholecystectomy	362	1.22	(0.97 -1.54)	0.09
Leg ulcer	370	0.86	(0.65 -1.13)	0.28
Priapism	289	1.10	(0.83 -1.48)	0.50
Retinopathy	384	1.2	(0.87 -1.65)	0.26
Renal Parameter	**N**	**Beta**	**SE**	**P-value**
eGFR	404	-0.01	0.05	0.85

Supplementary Table 10. Genetic associations between SNPs associated with 3-ureidopropionate and eGFR. SNPs (effect alleles) are the lead variants in each gene region identified in a GWAS for 3-ureidopropionate levels in Europeans (P <5.0x10^{-5}) (ref. Long, T. *et al*. 2017). Association effect size with eGFR in the CSSCD and GENMOD are beta-coefficient from linear regression in 0.07 mL/min per 1.172 m^2. *These variants were identified to be pleiotropic by Phenoscanner queries and were excluded from model 2 (**Methods**). At the bottom of the table, we also provide association results between polygenic trait scores (PTS) calculated using 22 3-ureidopropionate-associated SNPs (PTS_{22SNPs}) or after excluding pleiotropic variants (PTS_{16SNPs}). For the PTS, the effect size is per PTS standard deviation units.

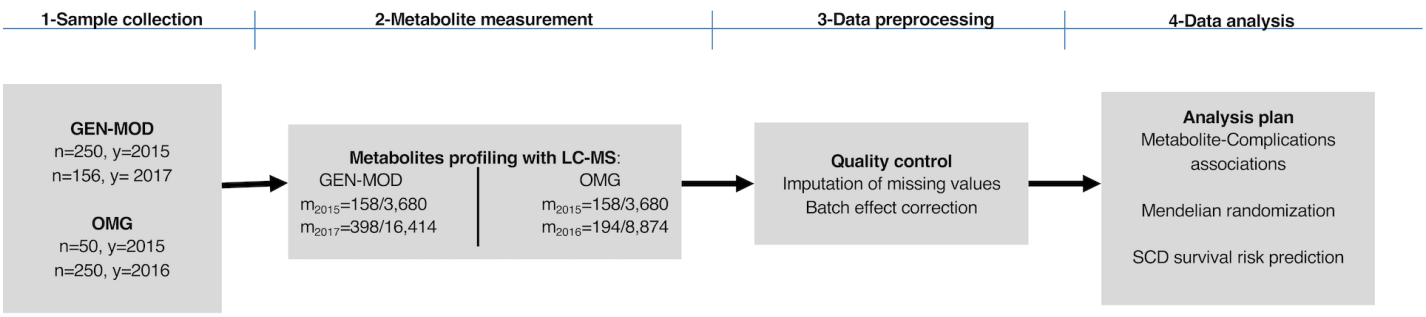
		Association result with 3-ureidopropionate -associated variants (effect allele frequency, effect size (beta), P-value, sample size			
SNP (hg19)	**Gene region**	**3-ureidopropionate (Long *et al*)**	**3-ureidopropionate (Meta-analysis GEN-MOD+OMG)**	**eGFR (CSSCD)**	**eGFR (Meta-analysis GEN-MOD-OMG)**
rs75277555(T)/1:92235001	TGFBR3	(0.072, 0.16, P = 4e-06, N = 1956)	(0.057, -0.093, P = 0.43, N = 651)	(0.047 ,-0.008, P = 0.94, N = 859)	(0.06, -0.24, P = 0.15, N = 640)
rs8013355(C)/14:52871418	intergenic	(0.32, 0.16, P = 1.7e-06, N = 1956)	(0.51, 0.012, P = 0.83, N = 651)	(0.48 ,-0.013, P = 0.78, N = 859)	(0.51, -0.05, P = 0.46, N = 640)
rs555045773(GA)/15:54319379	UNC13C	(0.15, -1.1, P = 6.2e-06, N = 1956)	(0.23, -0.040, P = 0.58, N = 651)	(0.26 ,-0.069, P = 0.2, N = 859)	(0.23, 0.15, P = 0.09, N = 640)
rs59930743(C)/17:3393604	ASPA	(0.22, 0.18, P = 2.8e-06, N = 1956)	(0.33, 0.030, P = 0.62, N = 651)	(0.29 ,0.013, P = 0.81, N = 859)	(0.33, 0.01, P = 0.87, N = 640)
rs56104151(T)/18:61129481	intergenic	(0.17, 0.15, P = 8.1e-06, N = 1956)	(0.35, 0.064, P = 0.27, N = 651)	(0.35 ,0.031, P = 0.53, N = 859)	(0.34, 0.08, P = 0.32, N = 640)
rs78734409(A)/2:12335302	AC096559.1	(0.038, 0.51, P = 3e-06, N = 1956)	(0.026, 0.080, P = 0.63, N = 651)	(0.025 ,0.14, P = 0.37, N = 859)	(0.03, -0.36, P = 0.1, N = 640)
rs71394795(GTTTA)/2:12355843	AC096559.1	(0.034, 0.19, P = 4.9e-07, N = 1956)	(0.71, -0.024, P = 0.70, N = 651)	(0.66 ,0.016, P = 0.76, N = 859)	(0.71, 0.06, P = 0.49, N = 640)
rs13427576(T)/2:56523619	CCDC85A	(0.12, 0.16, P = 9e-06, N = 1956)	(0.21, 0.036, P = 0.61, N = 651)	(0.22 ,0.0034, P = 0.95, N = 859)	(0.22, -0.1, P = 0.31, N = 640)
rs11704820(G)/22:24912248	UPB1	(0.43, 0.16, P = 4.3e-07, N = 1956)	(0.30, -0.14, P = 0.017, N = 651)	(0.32 ,0.0044, P = 0.93, N = 859)	(0.29, -0.01, P = 0.93, N = 640)
rs77020847(G)/3:150016629	intergenic	(0.013, 0.24, P = 2.1e-06, N = 1956)	(0.015, -0.13, P = 0.60, N = 651)	(0.014 ,0.074, P = 0.71, N = 859)	(0.01, 0.65, P = 0.05, N = 640)
rs6788347(C)/3:37754694*	ITGA9	(0.5, 0.19, P = 4e-06, N = 1956)	(0.53, 0.066, P = 0.25, N = 651)	(0.52 ,0.081, P = 0.1, N = 859)	(0.53, 0.06, P = 0.42, N = 640)

rs4698029(G)/4:10312798*	intergenic	(0.19, -0.39, P = 3.5e-17, N = 1956)	(0.23, -0.096, P = 0.17, N = 651)	(0.21 ,-0.033, P = 0.56, N = 859)	(0.22, 0.05, P = 0.56, N = 640)
rs13135526(A)/4:128932595*	C4orf29	(0.51, -0.08, P = 9.1e-06, N = 1956)	(0.60, -0.088, P = 0.12, N = 651)	(0.61 ,0.057, P = 0.25, N = 859)	(0.6, 0.01, P = 0.92, N = 640)
rs2725772(T)/4:140438033*	SETD7	(0.84, 0.21, P = 8.9e-08, N = 1956)	(0.50, -0.016, P = 0.78, N = 651)	(0.54 ,0.08, P = 0.093, N = 859)	(0.51, -0.05, P = 0.52, N = 640)
rs11735831(A)/4:9951591*	SLC2A9	(0.22, -0.43, P = 1.4e-23, N = 1956)	(0.46, 0.054, P = 0.343, N = 651)	(0.45 ,-0.029, P = 0.54, N = 859)	(0.46, 0.06, P = 0.45, N = 640)
rs33379(C)/5:171099634*	intergenic	(0.44, 0.16, P = 4e-06, N = 1956)	(0.83, -0.011, P = 0.88, N = 651)	(0.77 ,-0.033, P = 0.58, N = 859)	(0.83, -0.1, P = 0.33, N = 640)
rs76129636(G)/6:130978833	intergenic	(0.12, -0.41, P = 9.6e-06, N = 1956)	(0.24, -0.014, P = 0.83, N = 651)	(0.26 ,-0.0077, P = 0.89, N = 859)	(0.24, -0.13, P = 0.16, N = 640)
rs113133874(T)/6:33379903	PHF1	(0.11, -0.42, P = 7e-06, N = 1956)	(0.017, -0.062, P = 0.78, N = 651)	(0.021 ,-0.051, P = 0.76, N = 859)	(0.02, 0.15, P = 0.61, N = 640)
rs56286439(C)/7:14307105	DGKB	(0.2, -0.13, P = 4.8e-06, N = 1956)	(0.83, -0.005, P = 0.95, N = 651)	(0.82 ,-0.0045, P = 0.94, N = 859)	(0.83, -0.03, P = 0.79, N = 640)
rs2352451(G)/8:112781984	intergenic	(0.74, 0.18, P = 5.3e-06, N = 1956)	(0.75, 0.034, P = 0.59, N = 651)	(0.78 ,-0.073, P = 0.21, N = 859)	(0.75, -0.04, P = 0.67, N = 640)
rs7006208(A)/8:124157948	TBC1D31	(0.17, -0.96, P = 8e-06, N = 1956)	(0.25, 0.033, P = 0.61, N = 651)	(0.26 ,-0.087, P = 0.11, N = 859)	(0.25, 0.05, P = 0.56, N = 640)
rs74795659(A)/8:73633099	KCNB2	(0.068, 0.15, P = 9e-06, N = 1956)	(0.031, 0.13, P = 0.43, N = 651)	(0.023 ,0.24, P = 0.12, N = 859)	(0.03, -0.24, P = 0.28, N = 640)
PTS_{22SNPs}	NA	NA	(NA, 0.013, P=0.73, N=651)	(NA, 0.069, P=0.044, N=859)	(NA, -0.017, P=0.065, N=649)
PTS_{16SNPs}	NA	NA	(NA, 0.012, P=0.74, N=651)	(NA, 0.082, P=0.016, N=859)	(NA, -0.022, P=0.55, N=649)

Supplementary Table 11. Mendelian randomization results for 3-ureidopropionate with estimated glomerular filtration rate (eGFR) in sickle cell disease (SCD) patients. Estimates are effect size (standard error) in eGFR units (0.07 mL/min per 1.172 m^2) for the effect of a one standard deviation increase in genetically-controlled 3-ureidopropionate. All 22 genetic variants that are associated with 3-ureidopropionate at $P<5x10^{-5}$ are included in Model 1. In Model 2, we only kept 16 variants that were not pleiotropic. IVW: inverse variance-weighted. In light grey, we present MR replication results for 3-ureidopropionate and eGFR in the smaller GEN-MOD and OMG cohorts.

Metabolite	Method	eGFR (CSSCD)	eGFR (GEN-MOD)	eGFR (OMG)
3-ureidopropionate – Model 1	IVW	0.078 (0.023) P=0.00087	-0.093 (0.059) P=0.12	-0.076 (0.059) P=0.19
	MR-Egger	0.089 (0.048) P=0.077	-0.16 (0.089) P=0.082	-0.024 (0.090) P=0.79
	Weighted median	0.068 (0.045) P=0.13	-0.13 (0.076) P=0.65	0.027 (0.096) P=0.77
3-ureidopropionate – Model 2	IVW	0.07 (0.021) P=0.00097	-0.082 (0.068) P=0.22	-0.074 (0.068) P=0.28
	MR-Egger	0.088 (0.05) P=0.1	-0.14 (0.099) P=0.17	-0.057 (0.01) P=0.58
	Weighted median	0.068 (0.049) P=0.16	-0.12 (0.081) P=0.15	0.077 (0.11) P=0.49

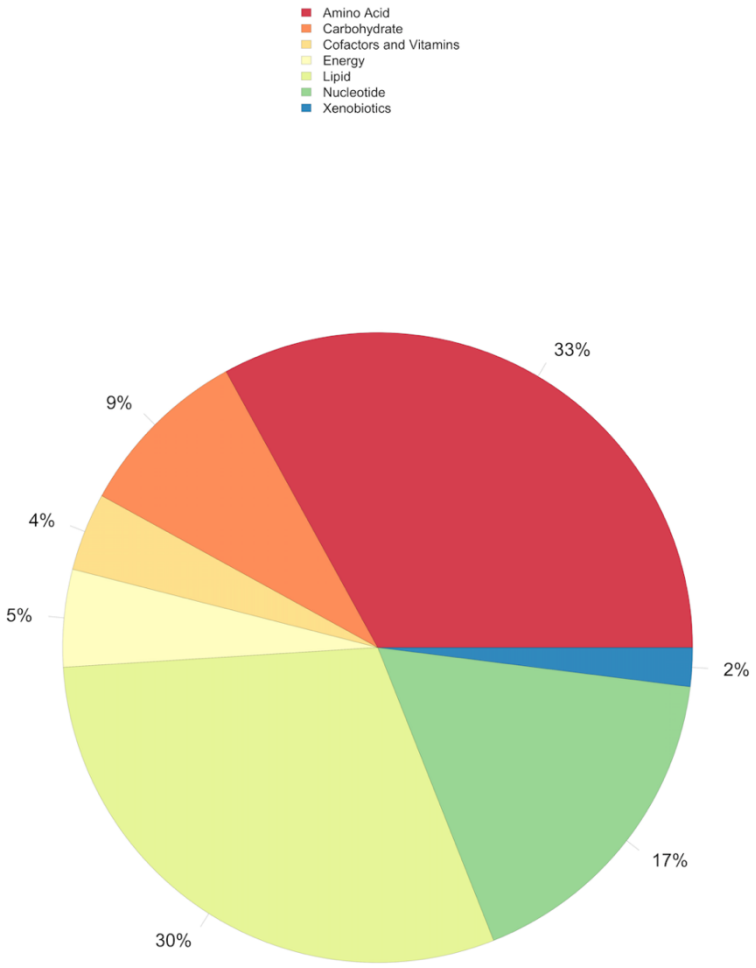
Supplementary Figure 1. Study design of the metabolomic study in sickle cell disease (SCD) patients. 250 GEN-MOD samples and 50 OMG samples were profiled in 2015, 250 OMG samples were profiled in 2016, and 156 GEN-MOD samples were profiled in 2017. Known/targeted and unknow/untargeted metabolites were measured using liquid-chromatography in tandem with mass spectrometry (LC-MS). Data preprocessing involved standard quality-control procedures, imputation of missing values, batch-effect correction and data scaling. Data analysis included association testing of known metabolites with SCD-related complications, Mendelian randomization, and SCD survival prediction using statistical modelling. n, number of patients included in the study; y, year during which metabolites were measured; m, number of targeted/untargeted metabolites measured in each year.



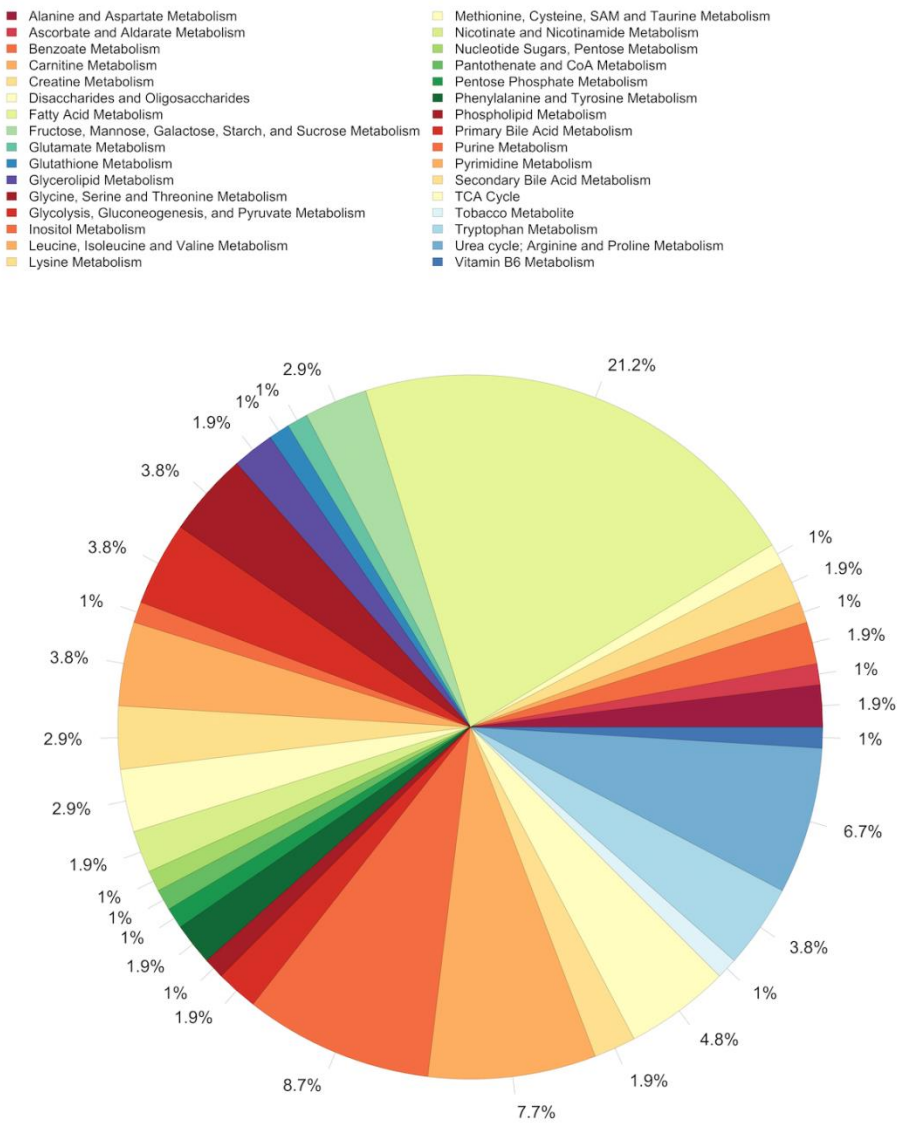
Supplementary Figure 2. Principal component analysis (PCA) of the metabolomic data before (**A**) and after (**B**) batch-effect correction using comBAT. Although the 3 different batches are clearly distinguishable before correction, comBAT pre-processing removes this effect. In each plot, the x- and y-axis represent the first and second principal components. The legend is the same for both plots.

Supplementary Figure 3. 129 known metabolites grouped in super-pathways (**A**) and pathways (**B**) based on criteria from the human metabolome database (HMDB).

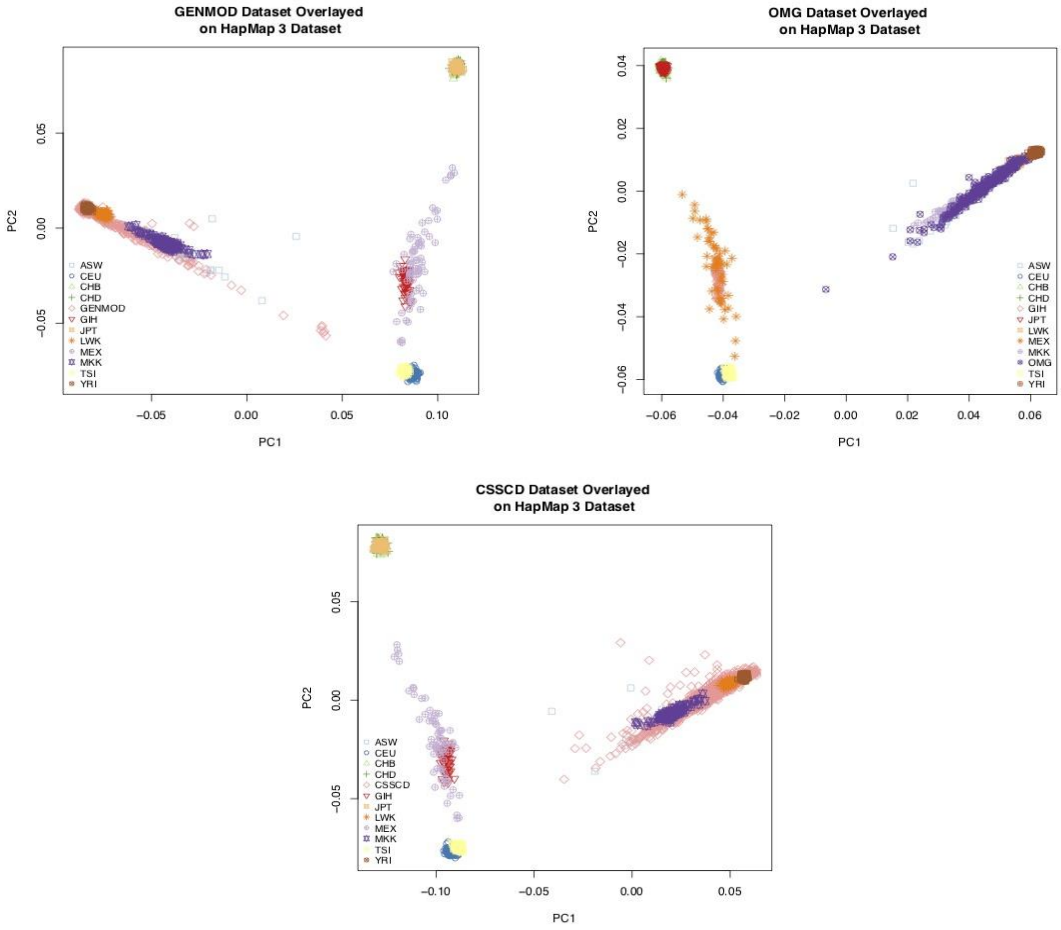
A



B


Alanine and Aspartate Metabolism
Ascorbate and Aldarate Metabolism
Benzoate Metabolism
Carnitine Metabolism
Creatine Metabolism
Disaccharides and Oligosaccharides
Fatty Acid Metabolism
Fructose, Mannose, Galactose, Starch, and Sucrose Metabolism
Glutamate Metabolism
Glutathione Metabolism
Glycerolipid Metabolism
Glycine, Serine and Threonine Metabolism
Glycolysis, Gluconeogenesis, and Pyruvate Metabolism
Inositol Metabolism
Leucine, Isoleucine and Valine Metabolism
Lysine Metabolism
Methionine, Cysteine, SAM and Taurine Metabolism
Nicotinate and Nicotinamide Metabolism
Nucleotide Sugars, Pentose Metabolism
Pantothenate and CoA Metabolism
Pentose Phosphate Metabolism
Phenylalanine and Tyrosine Metabolism
Phospholipid Metabolism
Primary Bile Acid Metabolism
Purine Metabolism
Pyrimidine Metabolism
Secondary Bile Acid Metabolism
TCA Cycle
Tobacco Metabolite
Tryptophan Metabolism
Urea cycle; Arginine and Proline Metabolism
Vitamin B6 Metabolism
21.2%
1%
1.9%
1%
1.9%
1%
1.9%
1%
6.7%
3.8%
1%
4.8%
1.9%
7.7%
8.7%
1.9%
1%
1.9%
1%
1%
1%
1.9%
2.9%
1.9%
2.9%
3.8%
1%
3.8%
1%
1.9%
2.9%

Supplementary Figure 4. Principal component analyses in the 3 sickle cell disease (SCD) cohorts analyzed in this study. We used reference populations from the 1000 Genomes Project to anchor these analyses. For all 3 SCD cohorts, most participants map to the European-African axis of variation, reflecting variable levels of admixture



Annex E A *Grammastola spatulata* mechanotoxin-4 (GsMTx4)-sensitive cation channel mediates increased cation permeability in human hereditary spherocytosis of multiple genetic etiologies

The article presented is published in the journal, *Hematologica*. In this article, whole-exome sequencing in SCD patients was employed to identify potential genetic variants responsible for the hereditary spherocytosis. I performed the quality control, variant calling, and the annotation of variants.

A Grammastola spatulata mechanotoxin-4 (GsMTx4)-sensitive cation channel mediates increased cation permeability in human hereditary spherocytosisof multiple genetic etiologies.

Haematologica. 2021 Oct 1; 106(10): 2759–2762**. doi: 10.3324/haematol.2021.278770**

David H. Vandorpe [1 *], Boris E. Shmukler [1 *], Yann Ilboudo [2] , Manoj Bhasin [3#] , Alicia Rivera [1] , Jay G. Wohlgemuth [4] , Jeffrey C. Dlott [5] , L. Michael Snyder [6] , Guillaume Lettre [2] , Carlo Brugnara [7] , Seth L. Alper [1,@]

1 Division of Nephrology and Vascular Biology Research Center, Beth Israel Deaconess Medical
Center and Department of Medicine, Harvard Medical School, Boston, MA 02215
2 Centre de Recherche, Institut de Cardologie de Montréal, Montréal, QC H1T1C8
3 Division of Integrative Medicine and Vascular Biology Research Center, Beth Israel Deaconess
Medical Center and Department of Medicine, Harvard Medical School, Boston, MA 02215
4 Quest Diagnostics, San Juan Capistrano, CA
5 Quest Diagnostics, Chantilly, VA
6 Quest Diagnostics, Marlborough, MA
7 Department of Laboratory Medicine, Boston Children's Hospital and Department of Pathology,Harvard Medical School, Boston, MA 02115.

*, Equal contributions
, Current address: Departments of Pediatrics and Biomedical Informatics, Emory University School of Medicine, Atlanta, GA
@ , Correspondence:
Seth L. Alper, MD-PhD
99 Brookline Ave. RN386

Boston, MA 02215
salper@bidmc.harvard .edu

Hereditary spherocytosis (HS) is the most common inherited hemolytic anemia among people of Northern European descent. HS is caused by mutations in genes encoding the erythroid cytoskeleton proteins ankyrin-1 (ANK1), b-spectrin (SPTB), and α-spectrin (SPTA1), the major intrinsic erythroid membrane protein and chloridebicarbonate exchanger, SLC4A1/band 3, and rarely EPB42/protein 4.2. These mutations lead to destabilization and progressive loss of red cell membrane lipids and surface area, and in some cases to destabilization of cytoskeletal-membrane attachment. The resulting red cells often exhibit normochromic or hyperchromic, mildmoderate microcytosis with increased incubated osmotic fragility and reduced deformability. Anemia and chronic hemolysis can be accompanied by hyperbilirubinemia and painful splenic enlargement. Splenectomy often provides symptomatic relief and attenuates anemia and hemolysis[314].

Increased erythroid cation permeability in HS was first reported by Harris and Prankerd (1953) and Bertles (1957), as subsequently cited by Zarkowsky *et al*.[315] Later reports of increased red cell cation permeability appeared in the setting of Southeast Asian ovalocytosis and cyrohydrocytosis in association with *SLC4A1* mutations, in the setting of overhydrated stomatocytosis in association with mutations in *RHAG, SLC2A1*, and *SLC4A1*, and in the setting of familial pseudohyperkalemia associated with *ABCB6* mutations[316]. Spherocytic mouse red cells genetically lacking EBP41 or EBP42, or haploinsufficient for *SLC4A1* exhibited enhanced Gardos channel activity and increased hemolysis in the presence of the Gardos inhibitor, clotrimazole[317], consistent with enhanced nonspecific cation permeability associated with these mouse HS models. Human HS red cells of diverse genotype were uniformly characterized by increased steady-state concentrations of fluorometrically measured intracellular [Ca^{2+}][318]. However, Petkova-Kirova *et al*.[319] recently reported that HS red cells of the same individuals had a spectrum of decreased, increased, or unchanged cation channel activities as measured by an automated whole cell patch clamp technique.

These studies led us to investigate whether HS red cells might be characterized by increased cation channel activity as detected by on-cell patch clamp analysis. We isolated DNA and RNA from whole blood of 13 patients with a clinical diagnoses of HS under protocols approved by Investigational Review Boards of Boston Children's Hospital and Beth Israel Deaconess Medical Center. The hematologic indices of the patients' red cells are presented in *Online Supplementary Table S1*. From the isolated total RNA, we generated complementary DNA (cDNA) for Sanger sequencing of *SLC4A1*. cDNA and/or genomic DNA (gDNA) from those patients lacking an evident *SLC4A1* mutation in blood cDNA was subjected to Sanger sequencing of *ANK1* and *SPTB*. Whole exome sequencing was reserved for gDNA from the six of 13 patients' samples that remained uninformative. Mutations detected by whole exome sequencing were subsequently confirmed by Sanger sequencing (Table 1). We found seven novel pathogenic mutations and one novel missense variant of very high predicted pathogenicity in previously identified HS genes among these patients with clinical diagnoses of HS. A subset of HS mutant red cells was subjected to on-cell patch clamp analysis (Figure 1). All cells in which stable gigohm seals were achieved exhibited substantial cation channel activity as compared to non-HS red cells. Mean cation channel unitary conductance among HS red cells was 26±2.1 pS (n=6 genotypes encompassing 16 cells; see Table 1). This increased activity, in the cases tested, was nearly completely inhibited by the mechanosensitive cation channel inhibitor, *Grammastola spatulata* mechanotoxin-4 (GsMTx4) at a concentration of 1 mM in the recording pipette (Table 1, Figure 1C). Non-HS red cells from healthy donors exhibited minimal channel activity (Figure 1C), as we had previously reported[320].

In this collection of HS patients, we found mutations in *SLC4A1, ANK1, SPTA1*, and *SPTB* (Table 1). Several patients exhibited mutations in *SLC4A1* previously reported in HS. HS2 was heterozygous for both HS mutant *Band 3 Lyon* (SLC4A1 R150X) and Band 3 Montefiore (*SLC4A1* E40K). Each mutation was undetected in cDNA but confirmed in gDNA, strongly suggesting that the mutant transcript carrying both mutations was a substrate of nonsense-mediated mRNA decay. Siblings HS3 and HS4 were each heterozygous for *Band 3 Bicetre* (*SLC4A1* R490C). HS5 was heterozygous for *Band 3 Osnabruck* (*SLC4A1* del663). HS6

was heterozygous for *Band 3 Prague* III (*SLC4A1* R870W), accompanied by the nonpathogenic Band 3 Memphis I (*SLC4A1* E56K).

Our HS patients also revealed a novel mutation in *SLC4A1* and several novel mutations in *ANK1* and *SPTB*, including a novel SPTB missense variant strongly predicted to be pathogenic (Table 1). The novel *SLC4A1* E68X mutation in HS1 was associated with nonsense-mediated decay, whereas the known rare *SLC4A1* R180H variant found on the other allele was detectable in both cDNA and gDNA. Four HS patients exhibited novel, heterozygous *ANK1* loss-of-function mutations, including *ANK1* E883Gfs32X in HS7 (accompanied by the known, rare *SPTA1* R1074H variant of uncertain significance), *ANK1* A1110del2 in HS8 (mutating a splice acceptor site), *ANK1* K1140Gfs86X in HS9 (accompanied by the SLC4A1/*Band 3 Memphis II* polymorphism) and *ANK1* E1289Gfs86X in HS10. Combined cDNA and gDNA sequencing indicated that mRNA encoding both *ANK1* mutations E883Gfs32X and E1289Gfs76X were substrates of nonsense-mediated decay, whereas the other two *ANK1* mutations underwent partial nonsense-mediated decay (Table 1).

Two HS patients were found to have novel heterozygous loss-of-function mutations in the *SPTB* gene encoding b-spectrin, *SPTB* G1450Rfs41X in HS12 and *SPTB* E1815Pfs90X in HS13 (Table 1). Both of these frameshift termination mutations encoded substrates of nonsensemediated decay. Patient HS11 exhibited compound heterozygosity for the novel, "probably damaging" missense variant *SPTB* R1255G (Polyphen-2 score 0.999) and the known non-pathogenic ANK1 R619H variant. The likely pathogenic *SPTB* R1255G substitution is located in the ninth of b-spectrin's 17 repeat domains, portions of which comprise a dimerization domain, a tetramerization domain, and the ankyrin-binding domain. Remarkably, the purified recombinant ninth b-spectrin repeat generated in *E. coli* was found to be more unstable (with a melting temperature of 20°C) than any other recombinant b-spectrin repeat polypeptide, each of which had melting temperatures ≥37°C,[321] demonstrating increased mutation-associated susceptibility to dysfunctional conformational change.

Table 1. Hereditary Spherocytosis (HS) accompanied by increased cation currents

Subject	Genetic diagnosis[a]	Ref.	Sanger sequence	NPo	NPo +GsMTx4	Unitary Conductance[a]	Fam Hx[b]	Osm Frg[c]	Tx/ Ac[d]	Spx/ Cx[e]
HS1 (M)	SLC4A1 E68X, c.202G>T w NMD[ns]	novel[f]	cDNA, gDNA	n.d.	n.d.	n.d.	n.a	n.a.	n.a.	n.a.
HS2 (F)	SLC4A1 R150X, c.448C>T, *Band 3 Lyon* w NMD[ns]	(g)	cDNA, gDNA	1.43 (n=2)	0.14 (n=1)	35 pS	+	n.a	–	+
HS3, HS4 (M,F; sibs)	SLC4A1 R490C, c.1648C>T, *Band 3 Bicetre*	(h)	cDNA	3.54 (n=1)	n.d.	28 pS	+	+	–	+
HS5 (F)	SLC4A1 M663del, c.1987-9del*, *Band 3 Osnabruck II*	(i)	gDNA	0.91±0.32 (n=4)	n.d.	25 pS	+	+	+	–
HS6 (F)	SLC4A1 R870W, c.2608C>T, *Band 3 Prague III*	(j)	cDNA	n.d.	n.d.	n.d.	+	+/–	–	–
HS7 (F)	ANK1 E883Gfs32X, c.2648delA, w NMD[ns]	novel[k]	cDNA, gDNA	n.d.	n.d.	n.d.	+	n.a.	+	+
HS8 (F)	A1110-Q1111del, c.3328-3333del6, w pNMD[pns] (Exon 28 mutant alters splice acceptor site)	novel	cDNA, gDNA	0.56±0.13 (n=4)	0.059 (n=1)	22.5 pS	+	+	+	+
HS9 (M)	ANK1 K1140Gfs86X, c.3416ins16, w pNMD[pns]	novel[l]	cDNA, gDNA	0.93±0.17 (n=4)	n.d.	25.8 pS	+	+	–	–
HS10 (F)	ANK1 E1289Gfs86X, c.3865dupG, w NMD[ns]	novel	cDNA, gDNA	n.d	n.d.	n.d.	+	+	+	+
HS11 (F)	SPTB R1255G, c.3763A>G	novel variant, likely pathogenic[m]	gDNA	1.26 (n=1)	0.042±0.02 (n=4)	21 pS	+	n.a.	–	–
HS12 (M)	SPTB G1450Rfs40X, c.4346dupG, w NMD[ns]	novel	cDNA, gDNA	n.d.	n.d.	n.d.	+	+	+	–
HS13 (M)	SPTB E1815Pfs90X, c.5443G>CC, w NMD[ns]	novel	cDNA, gDNA	n.d.	n.d.	n.d.	–	+	+	+

*cDNA numbering from initiator ATG of the open reading frame; SLC4A1: NP000333.1, NM000342.4; ANK1 isoform 1: NP065209, NM020476.2; SPTA1: NP003117, NM003126.4; SPTB erythrocyte isoform A: NP 001020029.1; var.1 NM001024858. [ns]complete nonsense-mediated decay (NMD); [pns]partial nonsense-mediated decay (pNMD); *deletion of any three consecutive nucleotides between c.1987-1992; n.d.: not done; n.a.: not available. [a]Unitary slope conductance measured in a single representative cell of specified genotype, without GsMTx4 in the pipette solution. [b]FamHx: family history of hereditary spherocytosis; [c]OsmFrg: results of osmotic fragility test; [d]Tx/AC: history of transfusion without or with aplastic crisis; [e]Spx/Cx: history of splenectomy or cholecystectomy. [f]Found with *SLC4A1* R180H, c.539C>A, rs147390654, MAF 0.0001-0.01 in different populations, detected in cDNA and gDNA. [g]Alloisio N *et al.* Blood 1996;88:1062-1069, conallelic with *SLC4A1* E40K, c.118G>A, *Band 3 Montefiore*. Rybicki AC *et al.* Blood 1998;81:2155-2165. Detected in cDNA and gDNA. [h]Dhermy D *et al.* Br J Haematol 1997;98:32-40. [i]Eber SW *et al.* Nat Genet 1996;13:214-218. [j]Jarolim P *et al.* Blood 1995;85:634-640, found together with benign variant *SLC4A1* K56E, c.166A>G, *Band 3 Memphis*. Yannoukakos D *et al.* Blood 1991;78:1117-1120. Detected in cDNA and gDNA. [k]Found together with SPTA1 R1074H, c.3221G>A, rs551084590, MAF 0.00004, detected in gDNA. [l]Found together with *SLC4A1* P854L, c.2561C>T, and K56E, c.166A>G, *Band 3 Memphis II*. Bruce LJ *et al.* J Biol Chem 1994;269:16155-16158. Detected in cDNA. [m]Found together with likely benign variant ANK1 R619H, c.1856G>A, rs2304877, *Ankyrin Bruggen*, Nakanishi H *et al.* Int J Hematol 2001;73:54-63. Detected in gDNA.

We assessed some of the HS mutants shown in Table 1 for cation channel activity in on-cell patches, preserving any regulatory components of the red cell cytosol and membrane cytoskeleton. Red cells from patients carrying the known SLC4 HS mutants R150X, R490C, and M663del each exhibited channel activity. Red cells from the patients carrying the novel HS-associated mutations ANK1 A1110del2 and ANK1 K1140Gfs86X also exhibited channel activity. In addition, red cells from the patient carrying the novel, predicted pathogenic VUS SPTB R1255G exhibited channel activity. The representative current trace from patient HS4 in Figure 1A with reversal potential of ~0 mV and unitary conductance of 21 pS (Figure 1B) is consistent with cation channel activity. Oncell patch recordings of red cells from patients HS2, HS4, HS5, HS8, HS9 and HS11, representing mutations in SLC4A1, ANK1, and SPTB, exhibited a mean unitary conductance of 26±2.1 pS.

In on-cell patch recordings of red cells from patients with the previously known SLC4A1 HS mutation R150X (HS2), the novel ANK1 mutation delA1110Q1111 (HS8), and the novel, rare predicted pathogenic SPTB variant R1255G (HS11), channel activity was also monitored under conditions in which the micropipette fluid included the mechanosensitive cation channel blocker, GsMTx4 (1 mM). Mean NPo of channel activity was 1.44±0.44 as measured in 16 cells representing six genotypes (Figure 1C, Table 1). The presence of 1 μM GsMTx4 in the recording pipette was associated with ~95% inhibition of channel activity, reducing mean NPo to 0.08±0.03 as measured in six cells representing three genotypes (Figure 1C, Table 1). The unitary conductance, reversal potential, and sensitivity of the current to inhibition by GsMTx-4 are each consistent with PIEZO1 mediation of, or contribution to, the measured cation channel activity in HS red cells. The increased membrane tension of the gigaseal inside the pipette[322] may unmask increased cation current in on-cell patches which might be less readily detected in whole cell patch recordings[318]. Interestingly, however, small increases in whole cell current were detected in some, if not all HS patients' red cells haploinsufficient for SPTB or for SPTA1[318].

Cation channel activity in the presence of pathogenic stomatocytosis mutations in transmembrane transporters such as SLC4A1, RHAG, GLUT1, and ABCB6 has been attributed either to direct cation permeation through the dysfunctional mutant membrane protein itself, or to direct or

indirect modulation of PIEZO1 activity[323]. However, the similar properties of the increased cation channel activities measured in the presence of pathogenic HS mutants of the cytoskeletal proteins b-spectrin and ankyrin very likely arise from direct or indirect modulation of PIEZO1 (and/or another unidentified cation channel), possibly by perturbations transmitted through one of the SLC4A1/Band3-nucleated macro-complexes[324]. This modulation might reflect PIEZO1 properties such as the lower hydrostatic pressure threshold for PIEZO1 activation in on-cell patches of actin cytoskeleton-depleted cellular blebs than in on-cell patches with intact cell cortex, and/or the inhibition by cytochalasin D of pressureactivated PIEZO1 in on-cell patches of normal cultured cells, and by glass probe-mediated cell indentation as measured by whole cell currents[322].

Our data suggest that PIEZO1 likely mediates or contributes a major fraction of the incremental cation permeability of HS red cells. Clarification of the relationships between apparent cytoskeletal modulation of erythroid PIEZO1 and PIEZO1 modulation by flow[325] and by modulation of lateral membrane tension via the ceramidesphingomyelin balance of the red cell membrane[326] will require further study.

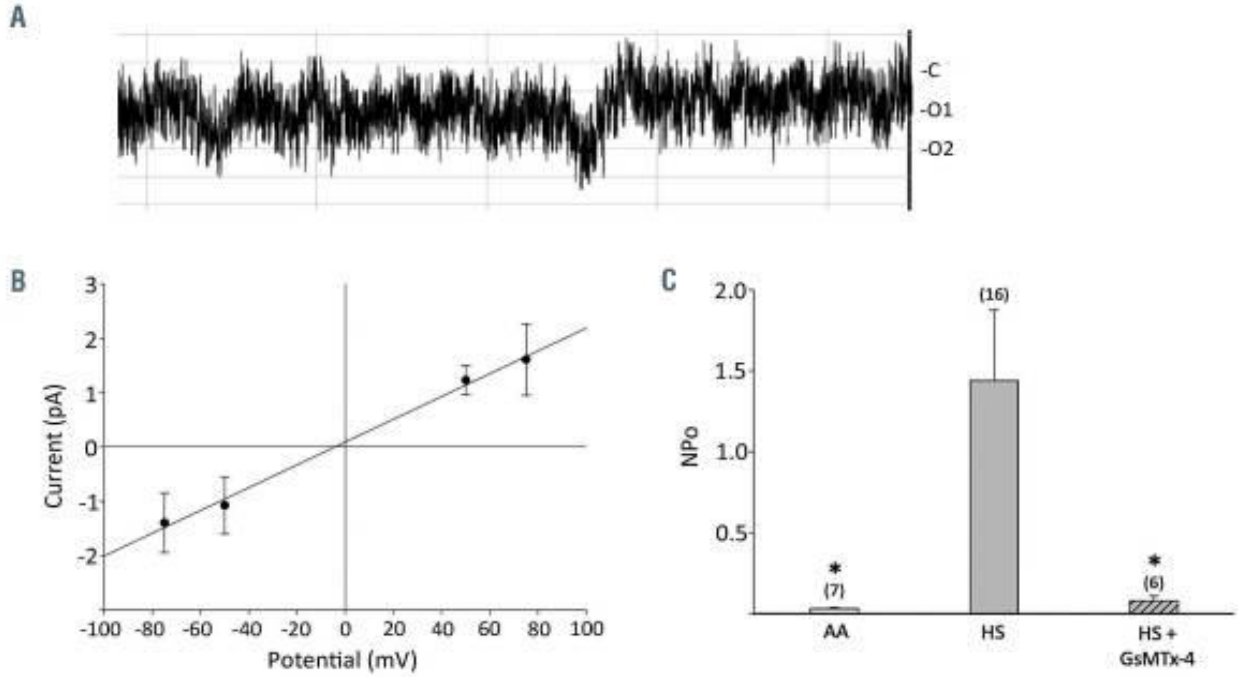
Acknowledgments

We thank Jeff Radcliff (Quest Diagnostics) for editorial suggestions.

Funding Statement

Funding: MB was supported by Beth Israel Deaconess Medical Center Core start-up funds. GL was supported by the Canadian Institutes of Health Research (PJT #156248). YI was supported by University of Montreal faculty of medicine merit scholarships for graduate and postdoctoral studies. SLA was supported by research funds from the Doris Duke Charitable Foundation and from Quest Diagnostics.

Figure 1. Electrophysiological properties of cation channels in hereditary spherocytosis red cells. (A) A representative current trace recorded at -Vp = -25mV from an on-cell patch recording from a red cell of patient HS11 with hereditary spherocytosis (HS) carrying the novel heterozygous SPTB missense variant R1255G (Polyphen-2 score 0.999). Identical bath and pipette fluid composition included (in mM) 140 NaCl, 4 KCl, 1 CaCl2, 1 MgCl2, 10 NaHEPES at a final pH of 7.40. On-cell patch currents were recorded by an Axopatch 200b amplifier and digitized by a Digidata 1440A A-D converter (Molecular Devices, Sunnyvale, CA, USA). Seal resistances were 6.0±1.0 GΩ (n=7) in non-HS cells, 5.0±0.8 GΩ (n=14) in HS cells without GsMTx4 in the pipette solution, and 4.8±1.0 GΩ (n=6) in HS cells with GsMTx4 in the pipette solution. Seal duration for recordings on HS cells unexposed to GsMTx4 was 18±11 min. Data were filtered at 500 Hz, digitized at 2 kHz by PClamp and analyzed offline by Clampfit (PCLAMP11, Molecular Devices). (B) Current-voltage relationship of HS11 red cell current measured in a representative on-cell patch, with unitary slope conductance of 21 pS. The current-voltage (I-V) relationship was generated in Clampex (PCLAMP 11, Axon Instruments) with the real-time control window in gap-free mode to record current traces of 10–30 s duration. Test potentials were selected in 25-50 mV increments ranging between a minimum of -100 mV to a maximum of +100 mV. (C) Summary data for NPo calculated from on-cell patch current traces of 5-10 s duration recorded in 16 cells from six HS mutant genotypes and in six cells from three mutant HS genotypes in the additional presence of GsMTx4 (1 μM) in the pipette fluid. NPo values recorded in seven non-HS red cells from four normal individuals (AA) are also shown. *P<0.05 for the t-test comparing normal to HS cells, and for the Mann-Whitney test comparing HS cells in the presence versus absence of GsMTx4.


A
-C
-O1
-O2
B
Current (pA)
Potential (mV)
C
NPo
(16)
(7)
(6)
AA
HS
HS +
GsMTx-4

A *Grammastola spatulata* mechanotoxin-4 (GsMTx4)-sensitive cation channel mediates increased cation permeability in human hereditary spherocytosis of multiple genetic etiologies

David H. Vandorpe,[1] Boris E. Shmukler,[1*] Yann Ilboudo,[2] Swati Bhasin,[3*] Beena Thomas,[3*] Alicia Rivera,[1] Jay G. Wohlgemuth,[4] Jeffrey S. Dlott,[4] L. Michael Snyder,[4] Colin Sieff,[5] Manoj Bhasin,[3°] Guillaume Lettre,[2] Carlo Brugnara[6] and Seth L. Alper[1]*

[1]*Division of Nephrology and Vascular Biology Research Center, Beth Israel Deaconess Medical Center and Department of Medicine, Harvard Medical School, Boston, MA, USA;* [2]*Montreal Heart Institute and Université de Montréal, Montréal, Québec, Canada;* [3]*Division of Integrative Medicine and Vascular Biology Research Center, Beth Israel Deaconess Medical Center and Department of Medicine, Harvard Medical School, Boston, MA, USA;* [4]*Quest Diagnostics, Seacaucus, NJ, USA;* [5]*Cancer and Blood Disorders Center, Dana-Farber Cancer Center and Boston Children's Hospital, and Department of Pediatrics, Harvard Medical School, Boston, MA, USA and* [6]*Department of Laboratory Medicine, Boston Children's Hospital and Department of Pathology, Harvard Medical School, Boston, MA, USA*

**DHV and BES contributed equally as co-first authors.*

°Current address: Departments of Pediatrics and Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA

Correspondence: SETH L. ALPER - salper@bidmc.harvard.edu

doi:10.3324/haematol.2021.278770

Supplemental Table 1 Legend.

Hematologic indices were measured by automated analysis on the ADVIA 2120 (Siemens). Bilirubin and LDH were measured by automated analysis on the COBAS (Roche).
For candidate gene sequencing, ≤ 1 mg total RNA prepared from whole blood of each patient (Rneasy Kit, Qiagen) was used for first strand cDNA synthesis (Retroscript, Ambion). Genomic DNA (gDNA) was also isolated from whole blood of each patient (Dneasy Blood and Tissue Kit, Qiagen). RT-PCR products (36-38 cycles) and/or genomic PCR products of the *SLC4A1*, *SPTB* and *ANK1* genes (36 cycles) were analyzed in 1% agarose gel. Fragments of expected size were excised, purifed (QIAquick Gel Extraction Kit, Qiagen) and subjected to Sanger sequencing. Pathogenic variants were discovered in samples of patients HS3-HS6, HS8, HS9, and HS11 (Table 1).
For whole exome sequencing, gDNA (1 ∝g) from whole blood of patient HS1 was fragmented by mechanical shearing (Covaris) to obtain fragments of length ~250 nt. DNA fragments were adenylated, adapter-ligated, and hybrid-captured, then processed for library preparation and paired-end sequencing using Novaseq 6000 at 100X coverage. Sequencing data was processed using a workflow including raw reads quality assessment by FastQC, adapter- and quality-trimming by Trimmomatic, alignment using BWA-MEM with hg19, post-alignment quality and removal of PCR duplicates by SAMtools and Picard-Tools (http://picard.sourceforge.net). Variants and indels were detected by GATK and annotated by ANNOVAR.

Exonic DNA fragments from 100 ng sheared gDNA from HS10 and HS12 were captured using the Illumina WES Nextera Kit. Exonic DNA fragments from HS2, HS7 and HS13 were captured using the Nimblegen SeqCap EZ Exome Capture Kit. Next-generation sequencing was conducted with an Illumina HiSeq4000 instrument using a paired-ends 2x100 base-pair protocol. Best practices pipeline recommendations for quality control and variant calling of reads were followed using the Genome Analysis Toolkit (GATK version 3.4-46) [1]. Sequenced reads were aligned to hg19 and analyzed. Varian calling was by GATK Haplotype Caller, and variant annotation was by Variant Effect Predictor (VEP)[2]. Mean target coverage was >91%, and >85% of bases were read at 10x coverage. Of 415,187 detected variants, 109,188 remained after selecting nonsynonymous variants with consequences matching the following terms: splice_acceptor, splice_donor_variant, stop_gained, frameshit_variant, stop_lost, start_lost, protein_altering_variant, missense_variant, coding_sequence_variant. Each pathogenic variant discovered by whole exome sequencing from patients HS1, HS2, HS7, HS10, HS12 and HS13 (Table 1) was subsequently validated by Sanger sequencing.

Supplemental Table 1. Hemolytic indices of patients.

Subject #	Genetic Dx	RBC (x 10^6)	Hb (g/dL)	Hct (%)	MCV (fL)	MCH (pg)	MCHC (g/dL)	HDW (g/dL)	RDW (%)	Retic (%)	Retic (x10^6/mL)	Bili (T/D) (mg/dL)	LDH (U/L)
HS1	SLC4A1 E68X SLC4A1 R180H	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
HS2	SLC4A1 R150X/E40K Band 3 Lyon in tandem w Band 3 Montefiore	3.75	12.5	34.7	92.7	33.5	36.1	3.38	15.2	5.0	0.189	1.1/0.2	294
HS3, HS4 (sibs)	SLC4A1 R490C sib1 Band 3 Bicetre sib2	3.88 3.00	11.8 10.0	32.3 27.3	83.2 81.4	30.4 29.8	36.5 36.6	3.83 3.83	20.1 19.6	7.5 9.7	0.365 0.326	2.1/0.3 1.5/0.2	287 398
HS5	SLC4A1 M663del Band 3 Osnabruck	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
HS6	SLC4A1 R870W/K56E Band 3 Prague 3	4.35	12.0	34.9	80.2	27.5	34.3	2.79	13.8	1.5	0.067	<0.1/<0.1	305
HS7	ANK1 E924X SPTA1 R1074H	4.47	12.9	34.3	76.7	28.8	37.5	4.37	18.5	2.3	0.105	0.8/0.2	208
HS8	ANK1 A1110-G1111del	5.03	12.8	34.8	69.3	25.5	36.8	4.44	18.3	3.0	0.151	0.6/0.1	240
HS9	ANK1 K1140Gfs.X87 SLC4A1 P854L/K56E	3.4	8.8	24.6	72.4	25.9	35.8	4.0	17.7	5.6	0.190	0.8/0.2	294
HS10	ANK1 E1289Gfs86X	3.69	11.6	30.2	81.9	31.4	38.4	4.68	19.5	10.0	0.369	2.4/0.2	370
HS11	SPTB R1255G ANK1 R619H	3.42	11.1	31.0	90.4	32.4	35.9	3.48	13.9	4.8	0.140	1.2/0.2	197
HS12	SPTB G1450Rfs41X	4.20	12.5	33.1	78.9	29.7	37.7	4.41	18.6	10.8	0.454	2.2/0.2	384
HS13	SPTB E1815AfsX	4.29	12.1	32.5	75.8	28.3	37.3	4.19	17.5	10.0	0.429	2.0/0.3	361
Normal range		3.92-4.72	11.0-12.8	31.5-36.8	76.8-83.3	26.8-29.4	34.2-35.7	2.75-3.21	13.2-14.5	0.8-2.0	0.029-0.080	0.3-0.0 - 1.2-0.4	110-295

Annex F. A common functional PIEZO1 deletion allele associates with red blood cell density in sickle cell disease patients

The article presented is published in the journal, *American Journal of Hematology*. In this article, I characterized the role of an inframe deletion in the gene PIEZO1 in SCD patients. I performed the quality control, variant calling, and annotation for the whole-exome sequence data. I additionally performed the association tests between the deletion and red blood cell density.

A common functional in-frame *PIEZO1* deletion allele associates with red blood cell density in sickle cell disease patients.[2]

American Journal of Hematology, 2018 Nov;93(11):E362-E365, Epub 2018 Sep 9

https://doi.org/10.1002/ajh.25245

Yann Ilboudo,[1,2] Pablo Bartolucci,[3] Melanie E. Garrett,[4,5] Allison Ashley-Koch,[4,5] Marilyn Telen,[4,5] Carlo Brugnara,[6] Frédéric Galactéros,[3] Guillaume Lettre[1,]

Affiliations

[1]Faculty of Medicine, 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]Center for Human Disease Modeling, Duke University Medical Center, Durham, North Carolina, United States of America

[5]Department of Medicine, Division of Hematology, Duke University Medical Center, Durham, North Carolina, United States of America

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

Correspondence

Guillaume Lettre

Montreal Heart Institute

5000 Bélanger St

Montreal, Quebec, Canada, H1T 1C8

514-376-3330 ext. 2660

guillaume.lettre@umontreal.ca

Number of words: 1048 (max 1200 intro-discussion)

Number of table: 1

Number of figure: 1

Key point

- A common functional allele in the gene *PIEZO1* associates with red blood cell density in sickle cell disease patients.

To the Editor:

PIEZO1 encodes a large mechanosensitive cation channel expressed in multiple cell types, including red blood cells (RBCs). In humans, rare gain-of-function mutations in *PIEZO1* cause hereditary xerocytosis (HX), characterized by RBC dehydration and anemia. Recently, Ma *et al*. identified an in-frame *PIEZO1* deletion allele (rs572934641) that is common in individuals of African ancestry[124]. In vitro, the deletion increased PIEZO1 inactivation time, mimicking other gain-of-function mutations found in HX patients. RBCs from nine healthy African Americans heterozygotes for rs572934641 were dehydrated when compared to erythrocytes from noncarriers[124].

RBC dehydration has been implicated in the clinical variability observed in patients with sickle cell disease (SCD), a multiorgan disorder caused by mutations in the β-globin gene. Increased RBC density, a hallmark of SCD, is independently correlated with hemolysis, priapism, leg ulcer, and renal dysfunction in patients[123]. Here, we investigated the association between the common functional *PIEZO1* deletion allele and RBC density, hemolytic parameters, estimated glomerular filtration rate (eGFR), and clinical complications in three large SCD cohorts. Our results indicate that common genetic variation in *PIEZO1* regulates RBC density in SCD patients, and thus represents one of many factors that influence clinical severity in this heterogenous blood disorder.

This project was reviewed and approved by the Montreal Heart Institute Ethics Committee and the different recruiting centers. Informed consent was obtained for all participants in accordance with the Declaration of Helsinki. The GEN-MOD cohort, the Cooperative Study of Sickle Cell Disease (CSSCD), and the Duke University Outcome Modifying Genes (OMG) cohort have been described elsewhere[126,248]. RBC density was measured using the phthalate density distribution technique in GEN-MOD participants[123]. DNA genotyping, quality-control, and genotype imputation using haplotypes from phase 3 of the 1000 Genomes Project were described previously[126,248]. We used Nimblegen SeqCap EZ Exome Capture kit to capture exons and we sequenced DNA using the Illumina HiSeq4000 instrument and a paired-ends 2x100 base pairs protocol. Whole-exome sequencing (WES) analysis was performed using the Genome Analysis Toolkit (GATK version 3.4-46). We followed best practices pipeline recommendations for reads

quality control and variant calling. All statistical analyses are described in the Supporting Information Methods.

rs572934641 maps to a complex DNA sequence region in *PIEZO1* with multiple TCC trinucleotide repeats. We inspected high-coverage WES data for 247 SCD patients from GEN-MOD (80% of targeted sequences covered at ≥80X) and identified at least four in-frame alleles. The most frequent allele has two TCC repeats and a frequency of 74.9%. The allele with one TCC repeat, which corresponds to the deletion allele (E756del) characterized recently[124], has a frequency of 22.9%. We also found two rarer alleles with zero (allele frequency = 0.6%) or three (allele frequency = 1.6%) repeats. In 226 SCD patients from GEN-MOD with phenotype and WES-derived genotypes available, the one repeat allele (E756del) was associated with increased RBC density ($P = .043$, **Table 1**). To increase our sample size, we imputed this common *PIEZO1* deletion allele in 375 GEN-MOD participants (including 149 additional SCD patients without WES data available) using reference haplotypes from phase 3 of the 1000 Genomes Project. The imputation quality metric was excellent (rsq_hat = 0.94) and concordance with genotypes from WES data was high (Supporting Information **Figure 1**). In this data set of 374 SCD patients, the association between RBC density and imputation-derived genotypes for the *PIEZO1* E756del allele was stronger and explained ~2.5% of the phenotypic variation ($P = .0039$, Table **1**). Thus, we provide in vivo evidence that RBCs from SCD patients that carry the *PIEZO1*-E756del allele are dehydrated[124].

We previously reported an association between RBC density in SCD patients and a regulatory DNA variant (rs10751450) within an erythroid enhancer at the *ATP2B4* locus[126,127]. *ATP2B4* encodes the main RBC calcium pump and erythroid cells with a deletion of the enhancer have increased intracellular Ca^{2+} concentration[127]. Increased intracellular calcium can activate the Gardos channel, leading to potassium efflux and dehydration. Because a gain-of-function mutation in *PIEZO1* could similarly result in excess calcium entry into RBCs, we tested if genotypes at *ATP2B4*-rs10751450 and *PIEZO1*-rs572934641 interacted to control RBC density in SCD patients. Both variants were independently associated with RBC density ($P_{ATP2B4} = .0016$ and $P_{PIEZO1} = .0034$ in a multivariate model), but we detected no evidence of interaction on RBC density ($P_{\text{interaction}} = .53$). We calculate a polygenic RBC density score by combining alleles from *PIEZO1*-rs572934641, *ATP2B4*-rs10751450, as well as rare missense variants

in *PIEZO1*, *ATP2B4,* and the Gardos channel gene *KCNN4* (Supporting Information **Table 1**). This polygenic score was strongly associated with RBC density ($P = 2.4 \times 10^{-4}$, 6.3% of variance explained, $N = 226$ SCD patients) (**Figure 1**).

Finally, we asked if the *PIEZO1* deletion allele was associated with SCD-related clinical phenotypes that are correlated with RBC density[327]. To maximize our sample size, we analyzed phenotype-genotype associations from 402 GEN-MOD participants, as well as 1081 and 552 SCD patients from the CSSCD and OMG study, respectively. Imputation quality for the common *PIEZO1* deletion allele in the CSSCD (rsq_hat = 0.90) and OMG study (rsq_hat = 0.93) was sufficiently high for association testing. Given the sample size of our study design, we saw no evidence of association between the *PIEZO1* E756del allele and tested SCD clinical phenotypes (**Table 1**).

In conclusion, we provide in vivo evidence that a common *PIEZO1* deletion allele is associated with RBC dehydration in SCD patients. Further, a simple polygenic score considering genetic variants at key RBC hydration genes (*PIEZO1*, *ATP2B4*, *KCNN4*) improves the association with RBC density. As for fetal hemoglobin[328], we anticipate that larger studies of RBC density genetics will provide new insights into SCD clinical heterogeneity.

ACKNOWLEDGMENTS

We thank all participants for their contribution to this project, and Nassima Djouder and Jugurtha Berkenou for help with the recruitment of GEN-MOD. G.L. is funded by the Canadian Institutes of Health Research (PJT #156248), the Doris Duke Charitable Foundation, and the Canada Research Chair program.

AUTHOR CONTRIBUTIONS

Y.I. and G.L. conceived and designed the experiments; Y.I. and M.E.G. performed the experiments; P.B., A.A.K., M.T., C.B., and F.G. contributed DNA samples, clinical information, and expert knowledge; Y.I., M.E.G. and G.L. analyzed the results; Y.I. and G.L. wrote the manuscript with contributions from all authors.

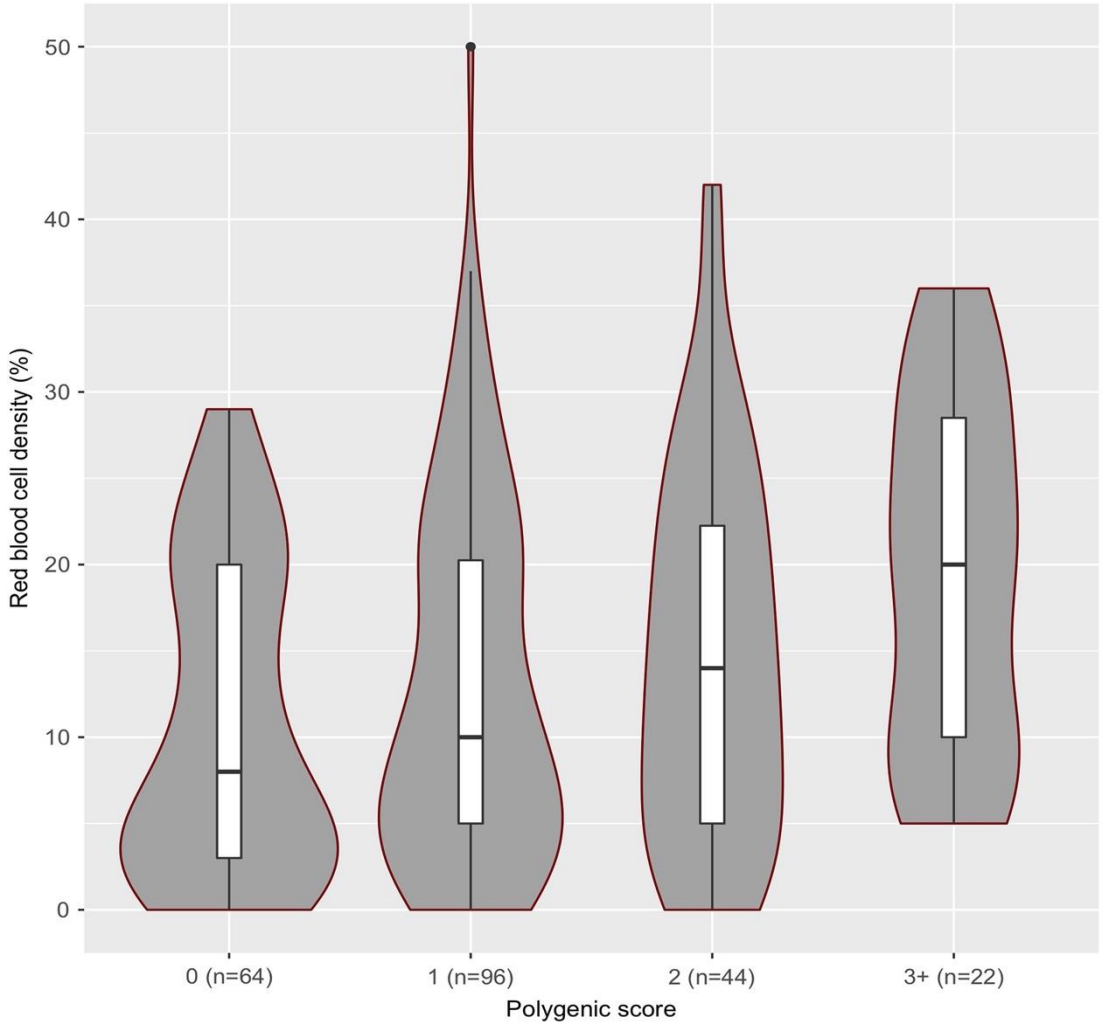
DISCLOSURE OF CONFLICTS OF INTEREST

The authors declare no competing financial interests.

Table 1. Association between the common *PIEZO1* deletion allele and clinical phenotypes in sickle cell disease patients. For each phenotype, the direction of the effect is given for the functional *PIEZO1* deletion allele (rs572934641; E756del): BETA and SE (SE) for RBC density, estimated glomerular filtration rate (eGFR) and hemolytic parameters are in SD units; odds ratio and 95% confidence interval for priapism and leg ulcer. The frequency of the imputed *PIEZO1* E756del allele is 18% in GENMOD ($N = 402$), 16% in CSSCD ($N = 1081$), and 16% in OMG ($N = 552$)

Phenotype	Sample size	Effect size	P-value
WES-derived genotypes			
RBC density	226	0.302 (0.15)	0.043
Imputation doses			
RBC density	374	0.289 (0.010)	0.0039
Bilirubin	1769	-0.0072 (0.042)	0.86
Lactate dehydrogenase	1657	0.0035 (0.048)	0.94
eGFR	865	0.045 (0.009)	0.49
Leg ulcer	294 cases / 1006 controls	1.01 (0.77 – 1.3)	0.96
Priapism	195 cases / 391 controls	0.803 (0.56 – 1.2)	0.22

Figure1. Polygenic score and dense red blood cell (RBC) density in 226 sickle cell disease patients from GEN-MOD. To generate the polygenic score, we counted the number of RBC density-increasing allele at *PIEZO1*-rs572934641 and *ATP2B4*-rs10751450, as well as the number of rare damaging nonsynonymous alleles in *PIEZO1*, *ATP2B4*, and *KCNN4*. The violin plots show the probability density of RBC density per polygenic score group. For the boxplots, the horizontal lines show the median, the boxes show the interquartile ranges (IQR), and the whiskers represent 1.5 times the IQR. We grouped together patients with 3 or 4 DRBC-increasing alleles (group 3+)



SUPPLEMENTAL MATERIALS

A common functional *PIEZO1* deletion allele associates with red blood cell density in sickle cell disease patients

Yann Ilboudo,[1,2] Pablo Bartolucci,[3] Melanie E. Garrett,[4,5] Allison Ashley-Koch,[4,5] Marilyn Telen,[4,5] Carlo Brugnara,[6] Frédéric Galactéros,[3] Guillaume Lettre[1,2]

Affiliations

[1]Faculty of Medicine, 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]Center for Human Disease Modeling, Duke University Medical Center, Durham, North Carolina, United States of America

[5]Department of Medicine, Division of Hematology, Duke University Medical Center, Durham, North Carolina, United States of America

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

Correspondence

Guillaume Lettre

Montreal Heart Institute

5000 Bélanger St

Montreal, Quebec, Canada, H1T 1C8

514-376-3330 ext. 2660

guillaume.lettre@umontreal.ca

Supplemental Method

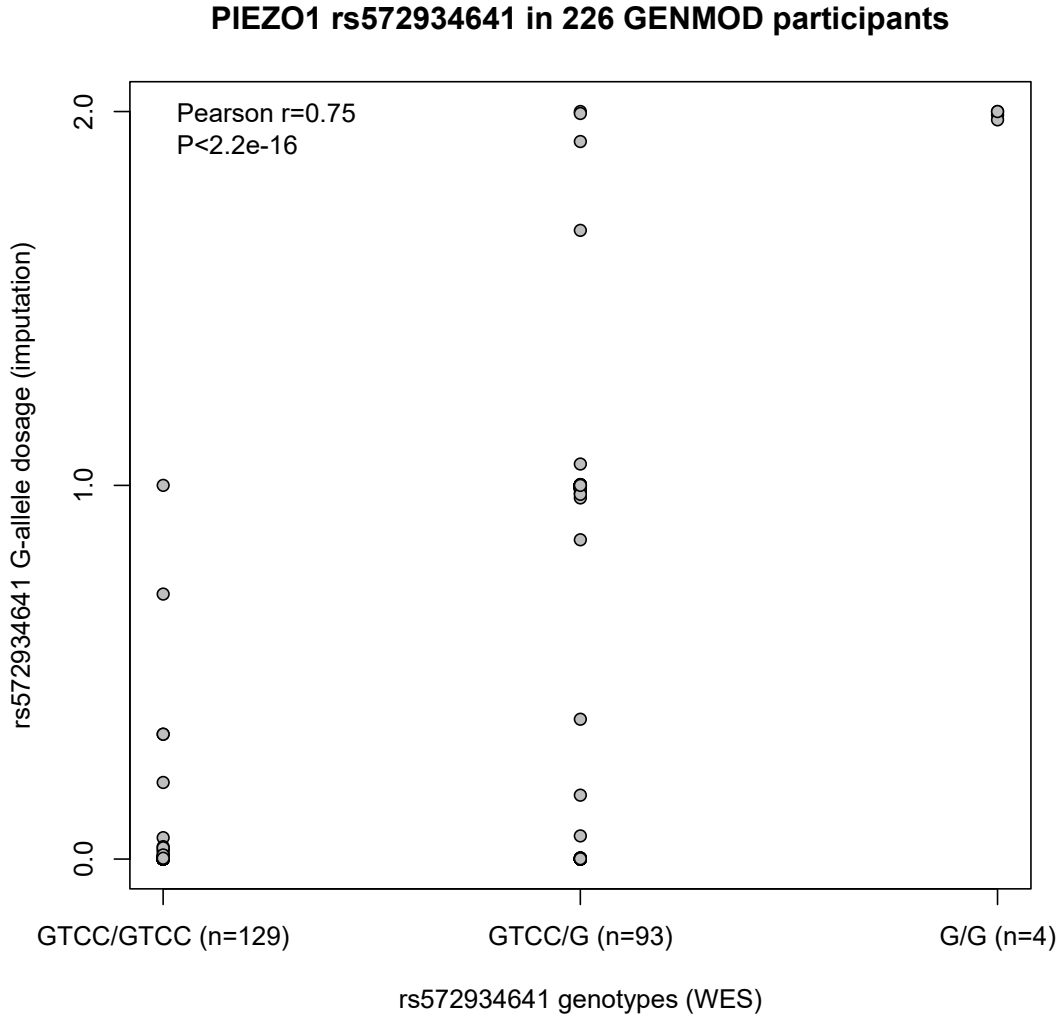
Statistical analyses

RBC density was available only in GEN-MOD, whereas all other baseline phenotypes (bilirubin, lactate dehydrogenase, estimated glomerular filtration rate (eGFR), leg ulcer, priapism) were available in all three SCD cohorts. RBC density, bilirubin, and lactate dehydrogenase levels were adjusted for age and sex, and the residuals were normalized using inverse normal transformation. eGFR, calculated via the CKD-EPI equation,[251] was normalized using inverse normal transformation. For leg ulcer and priapism, we focused on participants 18 years of age or older. Due to disease etiology, the association with priapism was restricted to men and was adjusted for age, while the association with leg ulcer was adjusted for age and sex. We used linear or logistic regression implemented in RVTESTS (GEN-MOD and CSSCD)[166] or SNPTEST (OMG)[329] to test the association between the *PIEZO1* deletion allele (additive model: testing the number of the functional allele described by Ma et *al*.[124] vs. all other *PIEZO1* alleles at this genomic position) and quantitative or dichotomous phenotypes, respectively. All analyses were adjusted for the first 10 (GEN-MOD and CSSCD) or two (OMG) principal components. Association results were combined with rareMETALS.[330] We used the R statistical package for all other analyses. For the polygenic score, we counted the number of RBC density-increasing alleles (no weights) at *PIEZO1*-rs572934641 and *ATP2B4*-rs10751450, and the number of rare damaging non-synonymous variants in *PIEZO1*, *ATP2B4*, and *KCNN4*. We tested the association between the polygenic score and RBC density by linear regression. P-values <0.05 were considered significant.

Supplemental Table 1. Rare non-synonymous variants identified in *PIEZO1*, *ATP2B4*, and *KCNN4* identified by whole-exome DNA sequencing in 226 sickle cell disease patients. Genomic positions are on build 37/hg19 of the human genome. REF and ALT are the reference and alternate allele; MAF is the minor allele frequency; DRBC_Avg_zScore are the normalized levels of dense red blood cells (in standard deviation units) in carriers of each rare variant.

VariantID	rsID	Chr	Pos	#Carriers	DRBC_Avg zScore	REF/ ALT	MAF	Annotation	Gene
16:88804653	-	16	88804653	1	-1.395	G/A	0.002024	missense	*PIEZO1*
16:88803073	rs759627248	16	88803073	2	-0.167	A/G	0.004049	missense	*PIEZO1*
16:88786920	rs761049480	16	88786920	1	0.033	G/A	0.002024	missense	*PIEZO1*
16:88782153	rs144035770	16	88782153	5	0.830	G/A	0.01	missense	*PIEZO1*
16:88786879	rs547409918	16	88786879	1	0.888	G/A	0.002024	missense	*PIEZO1*
16:88802710	-	16	88802710	1	1.456	G/T	0.002024	missense	*PIEZO1*
1:203691752	rs767342392	1	203691752	1	-0.956	G/A	0.002024	missense	*ATP2B4*
1:203680051	rs143539533	1	203680051	1	-0.577	A/C	0.002024	missense	*ATP2B4*
1:203696548	rs148156799	1	203696548	1	-0.532	C/T	0.004049	missense	*ATP2B4*
1:203652443	-	1	203652443	1	1.489	C/T	0.002024	missense	*ATP2B4*
1:203693073	-	1	203693073	1	1.838	T/A	0.002024	missense	*ATP2B4*
19:44273140	-	19	44273140	1	-0.044	A/G	0.002024	missense	*KCNN4*
19:44280719	rs78552213	19	44280719	4	0.679	C/T	0.008097	missense	*KCNN4*

Supplemental Figure 1. Concordance of genotypes at *PIEZO1* rs572934641 in 226 sickle cell disease patients from GEN-MOD. On the *x*-axis, we present the number of the common *PIEZO1* functional deletion allele (E756del) as determined by high-coverage whole-exome sequencing (WES). On the *y*-axis, we plot the dose of the same *PIEZO1* allele following imputation using phase 3 haplotypes from the 1000 Genomes Project. The imputation quality metric is high (rsq_hat=0.94) and WES-derived genotypes and imputation doses are highly correlated (Pearson's r=0.75, P<2.2x10^{-16}).



References

1 Herrick, J. B. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *JAMA* **312**, 1063, doi:10.1001/jama.2014.11011 (2014).

2 The Sickle Cell Association of New Jersey, I. *History of sickle cell*, 2013-2016).

3 E.V., H. & E.B, G. Sickle cell anemia. *Arch. Int. Med* **39**, 233 (1927).

4 Pauling, L. Molecular Disease and Evolution. *Bull N Y Acad Med* **40**, 334-342 (1964).

5 Pauling, L., Itano, H. A. & et al. Sickle cell anemia, a molecular disease. *Science* **110**, 543-548 (1949).

6 Neel, J. V. The Inheritance of Sickle Cell Anemia. *Science* **110**, 64-66, doi:10.1126/science.110.2846.64 (1949).

7 Watson, J. The significance of the paucity of sickle cells in newborn Negro infants. *Am J Med Sci* **215**, 419-423 (1948).

8 Allison, A. C. Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J* **1**, 290-294 (1954).

9 Ingram V.M. Abnormal human hemoglobins. The chemical difference between normal and sickle cell hemoglbins. *Biochim Biophys Acta* **36**, 402-411 (1959).

10 Ferrone, F. A. Polymerization and sickle cell disease: a molecular view. *Microcirculation* **11**, 115-128 (2004).

11 Orkin, S. H. MOLECULAR MEDICINE: Found in Translation. *Med (N Y)* **2**, 122-136, doi:10.1016/j.medj.2020.12.011 (2021).

12 Ojodu, J. *et al.* Incidence of sickle cell trait--United States, 2010. *MMWR Morb Mortal Wkly Rep* **63**, 1155-1158 (2014).

13 Nelson, D. A. *et al.* Sickle Cell Trait, Rhabdomyolysis, and Mortality among U.S. Army Soldiers. *N Engl J Med* **375**, 435-442, doi:10.1056/NEJMoa1516257 (2016).

14 Tsaras, G., Owusu-Ansah, A., Boateng, F. O. & Amoateng-Adjepong, Y. Complications associated with sickle cell trait: a brief narrative review. *Am J Med* **122**, 507-512, doi:10.1016/j.amjmed.2008.12.020 (2009).

15 Gergen, P. J., Macri, C. J. & Murrillo, S. The need for sickle cell screening among pediatric latino immigrants. *Arch Pediatr Adolesc Med* **156**, 729 (2002).

16 Italia, Y. *et al.* Feasibility of a newborn screening and follow-up programme for sickle cell disease among South Gujarat (India) tribal populations. *J Med Screen* **22**, 1-7, doi:10.1177/0969141314557372 (2015).

17 Shrikhande, A. V. *et al.* Prevalence of the beta(S) gene among scheduled castes, scheduled tribes and other backward class groups in Central India. *Hemoglobin* **38**, 230-235, doi:10.3109/03630269.2014.931287 (2014).

18 Guler, E., Garipardic, M., Dalkiran, T. & Davutoglu, M. Premarital screening test results for beta-thalassemia and sickle cell anemia trait in east Mediterranean region of Turkey. *Pediatr Hematol Oncol* **27**, 608-613, doi:10.3109/08880018.2010.503772 (2010).

19 Al Hosani, H., Salah, M., Osman, H. M., Farag, H. M. & Anvery, S. M. Incidence of haemoglobinopathies detected through neonatal screening in the United Arab Emirates. *East Mediterr Health J* **11**, 300-307 (2005).

20 Loukopoulos, D. Haemoglobinopathies in Greece: prevention programme over the past 35 years. *Indian J Med Res* **134**, 572-576 (2011).

21 Ladis, V., Karagiorga-Lagana, M., Tsatra, I. & Chouliaras, G. Thirty-year experience in preventing haemoglobinopathies in Greece: achievements and potentials for optimisation. *Eur J Haematol* **90**, 313-322, doi:10.1111/ejh.12076 (2013).

22 Hanchard, N. A., Hambleton, I., Harding, R. M. & McKenzie, C. A. Predicted declines in sickle allele frequency in Jamaica using empirical data. *Am J Hematol* **81**, 817-823, doi:10.1002/ajh.20715 (2006).
23 Saint-Martin, C. *et al.* Universal newborn screening for haemoglobinopathies in Guadeloupe (French West Indies): a 27-year experience. *J Med Screen* **20**, 177-182, doi:10.1177/0969141313507919 (2013).
24 Reeves, S. L. *et al.* Incidence, demographic characteristics, and geographic distribution of sickle cell trait and sickle cell anemia births in Michigan, 1997-2014. *Mol Genet Genomic Med* **7**, e795, doi:10.1002/mgg3.795 (2019).
25 Elguero, E. *et al.* Malaria continues to select for sickle cell trait in Central Africa. *Proc Natl Acad Sci U S A* **112**, 7051-7054, doi:10.1073/pnas.1505665112 (2015).
26 Piel, F. B., Steinberg, M. H. & Rees, D. C. Sickle Cell Disease. *N Engl J Med* **376**, 1561-1573, doi:10.1056/NEJMra1510865 (2017).
27 Piel, F. B., Hay, S. I., Gupta, S., Weatherall, D. J. & Williams, T. N. Global burden of sickle cell anaemia in children under five, 2010-2050: modelling based on demographics, excess mortality, and interventions. *PLoS Med* **10**, e1001484, doi:10.1371/journal.pmed.1001484 (2013).
28 Grosse, S. D. *et al.* Sickle cell disease in Africa: a neglected cause of early childhood mortality. *Am J Prev Med* **41**, S398-405, doi:10.1016/j.amepre.2011.09.013 (2011).
29 Mace, K. E., Lucchi, N. W. & Tan, K. R. Malaria Surveillance - United States, 2017. *MMWR Surveill Summ* **70**, 1-35, doi:10.15585/mmwr.ss7002a1 (2021).
30 Owusu-Ofori, A. K., Betson, M., Parry, C. M., Stothard, J. R. & Bates, I. Transfusion-transmitted malaria in Ghana. *Clin Infect Dis* **56**, 1735-1741, doi:10.1093/cid/cit130 (2013).
31 Cox, F. E. History of the discovery of the malaria parasites and their vectors. *Parasit Vectors* **3**, 5, doi:10.1186/1756-3305-3-5 (2010).
32 GH-DoPDa, M. *The History of Malaria, an Ancient Disease*, <https://www.cdc.gov/malaria/about/history/> (2016).
33 Scully, E. J., Kanjee, U. & Duraisingh, M. T. Molecular interactions governing host-specificity of blood stage malaria parasites. *Curr Opin Microbiol* **40**, 21-31, doi:10.1016/j.mib.2017.10.006 (2017).
34 Martinsen, E. S., Perkins, S. L. & Schall, J. J. A three-genome phylogeny of malaria parasites (Plasmodium and closely related genera): evolution of life-history traits and host switches. *Mol Phylogenet Evol* **47**, 261-273, doi:10.1016/j.ympev.2007.11.012 (2008).
35 Hayakawa, T., Culleton, R., Otani, H., Horii, T. & Tanabe, K. Big bang in the evolution of extant malaria parasites. *Mol Biol Evol* **25**, 2233-2239, doi:10.1093/molbev/msn171 (2008).
36 Escalante, A. A., Freeland, D. E., Collins, W. E. & Lal, A. A. 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* **95**, 8124-8129 (1998).
37 Perkins, S. L. & Schall, J. J. A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *J Parasitol* **88**, 972-978, doi:10.1645/0022-3395(2002)088[0972:AMPOMP]2.0.CO;2 (2002).
38 Vargas-Serrato, E., Corredor, V. & Galinski, M. R. Phylogenetic analysis of CSP and MSP-9 gene sequences demonstrates the close relationship of Plasmodium coatneyi to Plasmodium knowlesi. *Infect Genet Evol* **3**, 67-73 (2003).

39 Storm, J. & Craig, A. G. Pathogenesis of cerebral malaria--inflammation and cytoadherence. *Front Cell Infect Microbiol* **4**, 100, doi:10.3389/fcimb.2014.00100 (2014).
40 Wassmer, S. C. *et al.* Investigating the Pathogenesis of Severe Malaria: A Multidisciplinary and Cross-Geographical Approach. *Am J Trop Med Hyg* **93**, 42-56, doi:10.4269/ajtmh.14-0841 (2015).
41 Milner, D. A., Jr. Malaria Pathogenesis. *Cold Spring Harb Perspect Med* **8**, doi:10.1101/cshperspect.a025569 (2018).
42 Marsh, K. *et al.* Indicators of life-threatening malaria in African children. *N Engl J Med* **332**, 1399-1404, doi:10.1056/NEJM199505253322102 (1995).
43 Organization, W. H. World malaria report. (2021).
44 Organization, W. H. World malaria report. *https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2021* (2021).
45 WHO. Global Malaria Control and Elimination: report of a technical review. (2008).
46 LJ., B. C. Essential Malariology. *Wiley Medical* **p.193.** (1985).
47 Hill, A. V. *et al.* Common west African HLA antigens are associated with protection from severe malaria. *Nature* **352**, 595-600, doi:10.1038/352595a0 (1991).
48 Hill, A. V. *et al.* Molecular analysis of the association of HLA-B53 and resistance to severe malaria. *Nature* **360**, 434-439, doi:10.1038/360434a0 (1992).
49 Willcox, M., Bjorkman, A. & Brohult, J. Falciparum malaria and beta-thalassaemia trait in northern Liberia. *Ann Trop Med Parasitol* **77**, 335-347, doi:10.1080/00034983.1983.11811722 (1983).
50 Friedman, M. J. Erythrocytic mechanism of sickle cell resistance to malaria. *Proc Natl Acad Sci U S A* **75**, 1994-1997, doi:10.1073/pnas.75.4.1994 (1978).
51 Archer, N. M. *et al.* Resistance to Plasmodium falciparum in sickle cell trait erythrocytes is driven by oxygen-dependent growth inhibition. *Proc Natl Acad Sci U S A* **115**, 7350-7355, doi:10.1073/pnas.1804388115 (2018).
52 Band, G. *et al.* Malaria protection due to sickle haemoglobin depends on parasite genotype. *Nature* **602**, 106-111, doi:10.1038/s41586-021-04288-3 (2022).
53 Modiano, D. *et al.* Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature* **414**, 305-308, doi:10.1038/35104556 (2001).
54 Lin, M. J., Nagel, R. L. & Hirsch, R. E. Acceleration of hemoglobin C crystallization by hemoglobin S. *Blood* **74**, 1823-1825 (1989).
55 Ha, J., Martinson, R., Iwamoto, S. K. & Nishi, A. Hemoglobin E, malaria and natural selection. *Evol Med Public Health* **2019**, 232-241, doi:10.1093/emph/eoz034 (2019).
56 O'Donnell, A. *et al.* Interaction of malaria with a common form of severe thalassemia in an Asian population. *Proc Natl Acad Sci U S A* **106**, 18716-18721, doi:10.1073/pnas.0910142106 (2009).
57 α +-Thalassemia and Protection from Malaria. *PLoS Medicine* **3**, doi:10.1371/journal.pmed.0030221 (2006).
58 Foo, L. C., Rekhraj, V., Chiang, G. L. & Mak, J. W. Ovalocytosis protects against severe malaria parasitemia in the Malayan aborigines. *Am J Trop Med Hyg* **47**, 271-275, doi:10.4269/ajtmh.1992.47.271 (1992).
59 Boctor, F. N. & Dorion, R. P. Malaria and hereditary elliptocytosis. *Am J Hematol* **83**, 753, doi:10.1002/ajh.21018 (2008).
60 Peters, A. L. & Van Noorden, C. J. Glucose-6-phosphate dehydrogenase deficiency and malaria: cytochemical detection of heterozygous G6PD deficiency in women. *J Histochem Cytochem* **57**, 1003-1011, doi:10.1369/jhc.2009.953828 (2009).

61 Kariuki, S. N. *et al.* Red blood cell tension protects against severe malaria in the Dantu blood group. *Nature* **585**, 579-583, doi:10.1038/s41586-020-2726-6 (2020).
62 Rees, D. C., Williams, T. N. & Gladwin, M. T. Sickle-cell disease. *Lancet* **376**, 2018-2031, doi:10.1016/S0140-6736(10)61029-X (2010).
63 Sundd, P., Gladwin, M. T. & Novelli, E. M. Pathophysiology of Sickle Cell Disease. *Annu Rev Pathol* **14**, 263-292, doi:10.1146/annurev-pathmechdis-012418-012838 (2019).
64 Kato, G. J. *et al.* Sickle cell disease. *Nat Rev Dis Primers* **4**, 18010, doi:10.1038/nrdp.2018.10 (2018).
65 Gladwin, M. T. & Vichinsky, E. Pulmonary complications of sickle cell disease. *N Engl J Med* **359**, 2254-2265, doi:10.1056/NEJMra0804411 (2008).
66 Houwing, M. E. *et al.* Sickle cell disease: Clinical presentation and management of a global health challenge. *Blood Rev* **37**, 100580, doi:10.1016/j.blre.2019.05.004 (2019).
67 Brandow, A. M. *et al.* American Society of Hematology 2020 guidelines for sickle cell disease: management of acute and chronic pain. *Blood Adv* **4**, 2656-2701, doi:10.1182/bloodadvances.2020001851 (2020).
68 Ballas, S. K. *et al.* Definitions of the phenotypic manifestations of sickle cell disease. *Am J Hematol* **85**, 6-13, doi:10.1002/ajh.21550 (2010).
69 Vichinsky, E. P. *et al.* Causes and outcomes of the acute chest syndrome in sickle cell disease. National Acute Chest Syndrome Study Group. *N Engl J Med* **342**, 1855-1865, doi:10.1056/NEJM200006223422502 (2000).
70 Gill, F. M. *et al.* Clinical events in the first decade in a cohort of infants with sickle cell disease. Cooperative Study of Sickle Cell Disease [see comments]. *Blood* **86**, 776-783, doi:10.1182/blood.V86.2.776.bloodjournal862776 (1995).
71 Ohene-Frempong, K. *et al.* Cerebrovascular accidents in sickle cell disease: rates and risk factors. *Blood* **91**, 288-294 (1998).
72 Farooq, F., Mogayzel, P. J., Lanzkron, S., Haywood, C. & Strouse, J. J. Comparison of US Federal and Foundation Funding of Research for Sickle Cell Disease and Cystic Fibrosis and Factors Associated With Research Productivity. *JAMA Netw Open* **3**, e201737, doi:10.1001/jamanetworkopen.2020.1737 (2020).
73 Brousse, V. *et al.* Acute splenic sequestration crisis in sickle cell disease: cohort study of 190 paediatric patients. *Br J Haematol* **156**, 643-648, doi:10.1111/j.1365-2141.2011.08999.x (2012).
74 Goldstein, A. R., Anderson, M. J. & Serjeant, G. R. Parvovirus associated aplastic crisis in homozygous sickle cell disease. *Arch Dis Child* **62**, 585-588, doi:10.1136/adc.62.6.585 (1987).
75 Walker, T. M., Hambleton, I. R. & Serjeant, G. R. Gallstones in sickle cell disease: Observations from The Jamaican Cohort Study. *The Journal of Pediatrics* **136**, 80-85, doi:10.1016/s0022-3476(00)90054-4 (2000).
76 Rogers, Z. R. Priapism in sickle cell disease. *Hematol Oncol Clin North Am* **19**, 917-928, viii, doi:10.1016/j.hoc.2005.08.003 (2005).
77 Adeyoju, A. B. *et al.* Priapism in sickle-cell disease; incidence, risk factors and complications - an international multicentre study. *BJU Int* **90**, 898-902, doi:10.1046/j.1464-410x.2002.03022.x (2002).
78 Tran, H., Gupta, M. & Gupta, K. Targeting novel mechanisms of pain in sickle cell disease. *Hematology Am Soc Hematol Educ Program* **2017**, 546-555, doi:10.1182/asheducation-2017.1.546 (2017).

79 Adesina, O., Brunson, A., Keegan, T. H. M. & Wun, T. Osteonecrosis of the femoral head in sickle cell disease: prevalence, comorbidities, and surgical outcomes in California. *Blood Adv* **1**, 1287-1295, doi:10.1182/bloodadvances.2017005256 (2017).
80 Chacko, P., Kraut, E. H., Zweier, J., Hitchcock, C. & Raman, S. V. Myocardial infarction in sickle cell disease: use of translational imaging to diagnose an under-recognized problem. *J Cardiovasc Transl Res* **6**, 752-761, doi:10.1007/s12265-012-9426-z (2013).
81 Jitraruch, S. *et al.* Autoimmune Liver Disease in Children with Sickle Cell Disease. *J Pediatr* **189**, 79-85 e72, doi:10.1016/j.jpeds.2017.06.035 (2017).
82 Pinto, V. M., Gianesin, B., Balocco, M., Bacigalupo, L. & Forni, G. L. Noninvasive monitoring of liver fibrosis in sickle cell disease: Longitudinal observation of a cohort of adult patients. *Am J Hematol* **92**, E666-E668, doi:10.1002/ajh.24918 (2017).
83 Hurtova, M. *et al.* Transplantation for liver failure in patients with sickle cell disease: challenging but feasible. *Liver Transpl* **17**, 381-392, doi:10.1002/lt.22257 (2011).
84 Ataga, K. I., Derebail, V. K. & Archer, D. R. The glomerulopathy of sickle cell disease. *Am J Hematol* **89**, 907-914, doi:10.1002/ajh.23762 (2014).
85 Vichinsky, E. Chronic organ failure in adult sickle cell disease. *Hematology Am Soc Hematol Educ Program* **2017**, 435-439, doi:10.1182/asheducation-2017.1.435 (2017).
86 Sharpe, C. C. & Thein, S. L. How I treat renal complications in sickle cell disease. *Blood* **123**, 3720-3726, doi:10.1182/blood-2014-02-557439 (2014).
87 Aygun, B. *et al.* Hydroxyurea treatment decreases glomerular hyperfiltration in children with sickle cell anemia. *Am J Hematol* **88**, 116-119, doi:10.1002/ajh.23365 (2013).
88 Minniti, C. P. & Kato, G. J. Critical Reviews: How we treat sickle cell patients with leg ulcers. *Am J Hematol* **91**, 22-30, doi:10.1002/ajh.24134 (2016).
89 Klings, E. S. *et al.* An official American Thoracic Society clinical practice guideline: diagnosis, risk stratification, and management of pulmonary hypertension of sickle cell disease. *Am J Respir Crit Care Med* **189**, 727-740, doi:10.1164/rccm.201401-0065ST (2014).
90 Gordeuk, V. R., Castro, O. L. & Machado, R. F. Pathophysiology and treatment of pulmonary hypertension in sickle cell disease. *Blood* **127**, 820-828, doi:10.1182/blood-2015-08-618561 (2016).
91 Klings, E. S. & Steinberg, M. H. Acute chest syndrome of sickle cell disease: genetics, risk factors, prognosis, and management. *Expert Rev Hematol* **15**, 117-125, doi:10.1080/17474086.2022.2041410 (2022).
92 Johnson, S. *et al.* Exercise-induced changes of vital signs in adults with sickle cell disease. *Am J Hematol* **96**, 1630-1638, doi:10.1002/ajh.26369 (2021).
93 Bachmeyer, C. *et al.* Rituximab as an effective treatment of hyperhemolysis syndrome in sickle cell anemia. *Am J Hematol* **85**, 91-92, doi:10.1002/ajh.21578 (2010).
94 Schultz, C. L. *et al.* Reproductive intentions in mothers of young children with sickle cell disease. *Pediatr Blood Cancer* **67**, e28227, doi:10.1002/pbc.28227 (2020).
95 Inusa, B. P., Oyewo, A., Brokke, F., Santhikumaran, G. & Jogeesvaran, K. H. Dilemma in differentiating between acute osteomyelitis and bone infarction in children with sickle cell disease: the role of ultrasound. *PLoS One* **8**, e65001, doi:10.1371/journal.pone.0065001 (2013).
96 Piccin, A. *et al.* Autoimmune disease and sickle cell anaemia: 'Intersecting pathways and differential diagnosis'. *Br J Haematol* **197**, 518-528, doi:10.1111/bjh.18109 (2022).

97 Nagant, C. *et al.* Alteration of humoral, cellular and cytokine immune response to inactivated influenza vaccine in patients with Sickle Cell Disease. *PLoS One* **14**, e0223991, doi:10.1371/journal.pone.0223991 (2019).
98 Rogers, Z. R. *et al.* Biomarkers of splenic function in infants with sickle cell anemia: baseline data from the BABY HUG Trial. *Blood* **117**, 2614-2617, doi:10.1182/blood-2010-04-278747 (2011).
99 Kavanagh, P. L., Fasipe, T. A. & Wun, T. Sickle Cell Disease: A Review. *JAMA* **328**, 57-68, doi:10.1001/jama.2022.10233 (2022).
100 Tisdale, J. F., Thein, S. L. & Eaton, W. A. Treating sickle cell anemia. *Science* **367**, 1198-1199, doi:10.1126/science.aba3827 (2020).
101 McGann, P. T. & Ware, R. E. Hydroxyurea therapy for sickle cell anemia. *Expert Opin Drug Saf* **14**, 1749-1758, doi:10.1517/14740338.2015.1088827 (2015).
102 Charache, S. *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* **332**, 1317-1322, doi:10.1056/NEJM199505183322001 (1995).
103 Wong, T. E., Brandow, A. M., Lim, W. & Lottenberg, R. Update on the use of hydroxyurea therapy in sickle cell disease. *Blood* **124**, 3850-3857; quiz 4004, doi:10.1182/blood-2014-08-435768 (2014).
104 Ataga, K. I. *et al.* Crizanlizumab for the Prevention of Pain Crises in Sickle Cell Disease. *N Engl J Med* **376**, 429-439, doi:10.1056/NEJMoa1611770 (2017).
105 McGann, P. T. *et al.* Hydroxyurea Therapy for Children With Sickle Cell Anemia in Sub-Saharan Africa: Rationale and Design of the REACH Trial. *Pediatr Blood Cancer* **63**, 98-104, doi:10.1002/pbc.25705 (2016).
106 Gluckman, E. *et al.* Sickle cell disease: an international survey of results of HLA-identical sibling hematopoietic stem cell transplantation. *Blood* **129**, 1548-1556, doi:10.1182/blood-2016-10-745711 (2017).
107 Walters, M. C. *et al.* Indications and Results of HLA-Identical Sibling Hematopoietic Cell Transplantation for Sickle Cell Disease. *Biol Blood Marrow Transplant* **22**, 207-211, doi:10.1016/j.bbmt.2015.10.017 (2016).
108 Gluckman, E. Allogeneic transplantation strategies including haploidentical transplantation in sickle cell disease. *Hematology Am Soc Hematol Educ Program* **2013**, 370-376, doi:10.1182/asheducation-2013.1.370 (2013).
109 Niihara, Y. *et al.* A Phase 3 Trial of l-Glutamine in Sickle Cell Disease. *N Engl J Med* **379**, 226-235, doi:10.1056/NEJMoa1715971 (2018).
110 Vichinsky, E. *et al.* A Phase 3 Randomized Trial of Voxelotor in Sickle Cell Disease. *N Engl J Med* **381**, 509-519, doi:10.1056/NEJMoa1903212 (2019).
111 Kanter, J. *et al.* Biologic and Clinical Efficacy of LentiGlobin for Sickle Cell Disease. *N Engl J Med* **386**, 617-628, doi:10.1056/NEJMoa2117175 (2022).
112 Esrick, E. B. *et al.* Post-Transcriptional Genetic Silencing of BCL11A to Treat Sickle Cell Disease. *N Engl J Med* **384**, 205-215, doi:10.1056/NEJMoa2029392 (2021).
113 Frangoul, H. *et al.* CRISPR-Cas9 Gene Editing for Sickle Cell Disease and beta-Thalassemia. *N Engl J Med* **384**, 252-260, doi:10.1056/NEJMoa2031054 (2021).
114 van Dijk, M. J. *et al.* Safety and efficacy of mitapivat, an oral pyruvate kinase activator, in sickle cell disease: A phase 2, open-label study. *Am J Hematol* **97**, E226-E229, doi:10.1002/ajh.26554 (2022).
115 Rees, D. C. *et al.* A randomized, placebo-controlled, double-blind trial of canakinumab in children and young adults with sickle cell anemia. *Blood* **139**, 2642-2652, doi:10.1182/blood.2021013674 (2022).

116 Nagel, R. L. *et al.* Structural bases of the inhibitory effects of hemoglobin F and hemoglobin A2 on the polymerization of hemoglobin S. *Proc Natl Acad Sci U S A* **76**, 670-672, doi:10.1073/pnas.76.2.670 (1979).
117 Yawn, B. P. *et al.* Management of sickle cell disease: summary of the 2014 evidence-based report by expert panel members. *JAMA* **312**, 1033-1048, doi:10.1001/jama.2014.10517 (2014).
118 Platt, O. S. *et al.* Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med* **330**, 1639-1644, doi:10.1056/NEJM199406093302303 (1994).
119 Lettre, G. *et al.* DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc Natl Acad Sci U S A* **105**, 11869-11874, doi:10.1073/pnas.0804799105 (2008).
120 Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat Genet* **42**, 1049-1051, doi:10.1038/ng.707 (2010).
121 Cato, L. D. *et al.* Genetic regulation of fetal hemoglobin across global populations. *medRxiv*, doi:10.1101/2023.03.24.23287659 (2023).
122 Brugnara, C. Sickle cell dehydration: Pathophysiology and therapeutic applications. *Clin Hemorheol Microcirc* **68**, 187-204, doi:10.3233/CH-189007 (2018).
123 Bartolucci, P. *et al.* Erythrocyte density in sickle cell syndromes is associated with specific clinical manifestations and hemolysis. *Blood* **120**, 3136-3141, doi:10.1182/blood-2012-04-424184 (2012).
124 Ma, S. *et al.* Common PIEZO1 Allele in African Populations Causes RBC Dehydration and Attenuates Plasmodium Infection. *Cell* **173**, 443-455 e412, doi:10.1016/j.cell.2018.02.047 (2018).
125 Tiffert, T. *et al.* The hydration state of human red blood cells and their susceptibility to invasion by Plasmodium falciparum. *Blood* **105**, 4853-4860, doi:10.1182/blood-2004-12-4948 (2005).
126 Ilboudo, Y. *et al.* Genome-wide association study of erythrocyte density in sickle cell disease patients. *Blood Cells Mol Dis* **65**, 60-65, doi:10.1016/j.bcmd.2017.05.005 (2017).
127 Lessard, S. *et al.* An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility. *J Clin Invest* **127**, 3065-3074, doi:10.1172/JCI94378 (2017).
128 Joiner, C. H. Cation transport and volume regulation in sickle red blood cells. *The American journal of physiology* **264**, C251-270 (1993).
129 Anthony J. McGoron, C. H. J., 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* **95**, 2164-2168 (2000).
130 Robert S. Franco, H. T., 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* **90**, 3195-3203 (1997).
131 Boettger, T. *et al.* Loss of K-Cl co-transporter KCC3 causes deafness, neurodegeneration and reduced seizure threshold. *EMBO J* **22**, 5422-5434, doi:10.1093/emboj/cdg519 (2003).
132 Boettger, T. *et al.* Deafness and renal tubular acidosis in mice lacking the K-Cl co-transporter Kcc4. *Nature* **416**, 874-878, doi:10.1038/416874a (2002).

133 Hubner, C. A. *et al.* Disruption of KCC2 reveals an essential role of K-Cl cotransport already in early synaptic inhibition. *Neuron* **30**, 515-524 (2001).
134 Lew VL, B. R. Ion transport pathology in the mechanism of sickle cell dehydration. *Physiol Rev* **85** (2005;).
135 Ataga, K. I. *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* **153**, 92-104, doi:10.1111/j.1365-2141.2010.08520.x (2011).
136 Zhang, Y. *et al.* Elevated sphingosine-1-phosphate promotes sickling and sickle cell disease progression. *J Clin Invest* **124**, 2750-2761, doi:10.1172/JCI74604 (2014).
137 Zhang, Y. *et al.* Detrimental effects of adenosine signaling in sickle cell disease. *Nat Med* **17**, 79-86, doi:10.1038/nm.2280 (2011).
138 Sun, K. *et al.* Structural and Functional Insight of Sphingosine 1-Phosphate-Mediated Pathogenic Metabolic Reprogramming in Sickle Cell Disease. *Sci Rep* **7**, 15281, doi:10.1038/s41598-017-13667-8 (2017).
139 Darghouth, D. *et al.* Pathophysiology of sickle cell disease is mirrored by the red blood cell metabolome. *Blood* **117**, e57-66, doi:10.1182/blood-2010-07-299636 (2011).
140 Zerez, C. R., Lachant, N. A., Lee, S. J. & Tanaka, K. R. Decreased erythrocyte nicotinamide adenine dinucleotide redox potential and abnormal pyridine nucleotide content in sickle cell disease. *Blood* **71**, 512-515 (1988).
141 Morris, C. R. *et al.* Erythrocyte glutamine depletion, altered redox environment, and pulmonary hypertension in sickle cell disease. *Blood* **111**, 402-410, doi:10.1182/blood-2007-04-081703 (2008).
142 Adebiyi, M. G., Manalo, J. M. & Xia, Y. Metabolomic and molecular insights into sickle cell disease and innovative therapies. *Blood Adv* **3**, 1347-1355, doi:10.1182/bloodadvances.2018030619 (2019).
143 Vinuela, A. *et al.* Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nat Commun* **11**, 4912, doi:10.1038/s41467-020-18581-8 (2020).
144 Steinberg, J. *et al.* A molecular quantitative trait locus map for osteoarthritis. *Nat Commun* **12**, 1309, doi:10.1038/s41467-021-21593-7 (2021).
145 Tobias Österlund, Marija Cvijovic & Kristiansson, E. in *Systems Biology, VI* (2017).
146 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
147 Salzberg, B. L. a. S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
148 Li, X. & Heyer, W. D. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res* **18**, 99-113, doi:10.1038/cr.2008.1 (2008).
149 Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501, doi:10.1093/bioinformatics/btw018 (2016).
150 Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908 (2014).
151 Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722, doi:10.1371/journal.pgen.1004722 (2014).

152 Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1273-1300, doi:10.1111/rssb.12388 (2020).
153 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).
154 Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, doi:10.1038/s43586-021-00056-9 (2021).
155 Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet* **39**, 1197-1199, doi:10.1038/ng2108 (2007).
156 Uda, M. *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* **105**, 1620-1625, doi:10.1073/pnas.0711566105 (2008).
157 Milton, J. N. *et al.* A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. *PLoS One* **7**, e34741, doi:10.1371/journal.pone.0034741 (2012).
158 Batista, J. *et al.* Influence of UGT1A1 promoter polymorphism, alpha-thalassemia and beta(s) haplotype in bilirubin levels and cholelithiasis in a large sickle cell anemia cohort. *Ann Hematol* **100**, 903-911, doi:10.1007/s00277-021-04422-1 (2021).
159 Flanagan, J. M. *et al.* Genetic predictors for stroke in children with sickle cell anemia. *Blood* **117**, 6681-6684, doi:10.1182/blood-2011-01-332205 (2011).
160 Saraf, S. L. *et al.* Genetic variants and cell-free hemoglobin processing in sickle cell nephropathy. *Haematologica* **100**, 1275-1284, doi:10.3324/haematol.2015.124875 (2015).
161 Adebayo, O. C. *et al.* Clinical and genetic factors are associated with kidney complications in African children with sickle cell anaemia. *Br J Haematol* **196**, 204-214, doi:10.1111/bjh.17832 (2022).
162 Pincez, T., Ashley-Koch, A. E., Lettre, G. & Telen, M. J. Genetic Modifiers of Sickle Cell Disease. *Hematol Oncol Clin North Am* **36**, 1097-1124, doi:10.1016/j.hoc.2022.06.006 (2022).
163 Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299, doi:10.1038/s41586-021-03205-y (2021).
164 Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-290, doi:10.1038/ng.3190 (2015).
165 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-354, doi:10.1038/ng.548 (2010).
166 Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423-1426, doi:10.1093/bioinformatics/btw079 (2016).
167 Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am J Hum Genet* **110**, 179-194, doi:10.1016/j.ajhg.2022.12.011 (2023).
168 Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am J Hum Genet* **109**, 767-782, doi:10.1016/j.ajhg.2022.04.001 (2022).

169 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).
170 Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572-580, doi:10.1016/S0140-6736(12)60312-2 (2012).
171 Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, doi:10.7554/eLife.34408 (2018).
172 Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* **46**, 1734-1739, doi:10.1093/ije/dyx034 (2017).
173 Sanderson, E. *et al.* Mendelian randomization. *Nature Reviews Methods Primers* **2**, doi:10.1038/s43586-021-00092-5 (2022).
174 Holmes, M. V., Ala-Korpela, M. & Smith, G. D. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat Rev Cardiol* **14**, 577-590, doi:10.1038/nrcardio.2017.78 (2017).
175 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).
176 Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17, doi:10.2202/1544-6115.1128 (2005).
177 Song, W. M. & Zhang, B. Multiscale Embedded Gene Co-expression Network Analysis. *PLoS Comput Biol* **11**, e1004574, doi:10.1371/journal.pcbi.1004574 (2015).
178 Long, Q. *et al.* Inter-tissue coexpression network analysis reveals DPP4 as an important gene in heart to blood communication. *Genome Med* **8**, 15, doi:10.1186/s13073-016-0268-1 (2016).
179 Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P. & Zitzler, E. BicAT: a biclustering analysis toolbox. *Bioinformatics* **22**, 1282-1283, doi:10.1093/bioinformatics/btl099 (2006).
180 Ragoussis, J. Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet* **10**, 117-133, doi:10.1146/annurev-genom-082908-150116 (2009).
181 Verlouw, J. A. M. *et al.* A comparison of genotyping arrays. *Eur J Hum Genet* **29**, 1611-1624, doi:10.1038/s41431-021-00917-7 (2021).
182 Purcell, S. M. in *Charney & Nestler's Neurobiology of Mental Illness (5 edn)* Ch. Genetic Methodologies and Applications, (2017).
183 LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* **37**, 4181-4193, doi:10.1093/nar/gkp552 (2009).
184 Zhao, S. *et al.* Strategies for processing and quality control of Illumina genotyping arrays. *Brief Bioinform* **19**, 765-775, doi:10.1093/bib/bbx012 (2018).
185 Marchini, J. UKBiobank Phasing and Imputation Documentation. (2015). <https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/impute_ukb_v1.pdf>.
186 Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147**, 1462-1465, doi:10.1126/science.147.3664.1462 (1965).
187 Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-1656, doi:10.1126/science.2047873 (1991).
188 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-351, doi:10.1038/nrg.2016.49 (2016).

189 Liu, L. *et al.* Comparison of next-generation sequencing systems. *J Biomed Biotechnol* **2012**, 251364, doi:10.1155/2012/251364 (2012).
190 Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13, doi:10.1186/s13059-016-0881-8 (2016).
191 Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* **2015**, 951-969, doi:10.1101/pdb.top084970 (2015).
192 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).
193 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).
194 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).
195 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:10.1038/nature11233 (2012).
196 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
197 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
198 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).
199 Lu, W. *et al.* Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *Annu Rev Biochem* **86**, 277-304, doi:10.1146/annurev-biochem-061516-044952 (2017).
200 Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* **18**, 747-756, doi:10.1038/s41592-021-01197-1 (2021).
201 Cambiaghi, A., Ferrario, M. & Masseroli, M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Brief Bioinform* **18**, 498-510, doi:10.1093/bib/bbw031 (2017).
202 Misra, B. B. New software tools, databases, and resources in metabolomics: updates from 2020. *Metabolomics* **17**, 49, doi:10.1007/s11306-021-01796-1 (2021).
203 Danjou, F. *et al.* Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat Genet* **47**, 1264-1271, doi:10.1038/ng.3307 (2015).
204 Magi, R. *et al.* Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum Mol Genet* **26**, 3639-3650, doi:10.1093/hmg/ddx280 (2017).
205 Thein, S. L. The molecular basis of beta-thalassemia. *Cold Spring Harb Perspect Med* **3**, a011700, doi:10.1101/cshperspect.a011700 (2013).
206 Weatherall, D., Akinyanju, O., Fucharoen, S., Olivieri, N. & Musgrove, P. in *Disease Control Priorities in Developing Countries* (eds nd *et al.*) (2006).
207 Qin, K. *et al.* Publisher Correction: Dual function NFI factors control fetal hemoglobin silencing in adult erythroid cells. *Nat Genet* **54**, 906, doi:10.1038/s41588-022-01112-0 (2022).

208 Kawabata, E. *et al.* Identification of Novel Variants Associated with Fetal Hemoglobin Levels in Healthy Donors (the INTERVAL study). *Blood* **134**, 2243-2243, doi:10.1182/blood-2019-123977 (2019).
209 Olave, I. A., Doneanu, C., Fang, X., Stamatoyannopoulos, G. & Li, Q. Purification and identification of proteins that bind to the hereditary persistence of fetal hemoglobin -198 mutation in the gamma-globin gene promoter. *J Biol Chem* **282**, 853-862, doi:10.1074/jbc.M610404200 (2007).
210 Xu, J. *et al.* Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc Natl Acad Sci U S A* **110**, 6518-6523, doi:10.1073/pnas.1303976110 (2013).
211 Gong, Y. *et al.* A natural DNMT1 mutation elevates the fetal hemoglobin level via epigenetic derepression of the gamma-globin gene in beta-thalassemia. *Blood* **137**, 1652-1657, doi:10.1182/blood.2020006425 (2021).
212 Basak, A. *et al.* Control of human hemoglobin switching by LIN28B-mediated regulation of BCL11A translation. *Nat Genet* **52**, 138-145, doi:10.1038/s41588-019-0568-7 (2020).
213 Masuda, T. *et al.* Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science* **351**, 285-289, doi:10.1126/science.aad3312 (2016).
214 Satta, S. *et al.* Compound heterozygosity for KLF1 mutations associated with remarkable increase of fetal hemoglobin and red cell protoporphyrin. *Haematologica* **96**, 767-770, doi:10.3324/haematol.2010.037333 (2011).
215 Grevet, J. D. *et al.* Domain-focused CRISPR screen identifies HRI as a fetal hemoglobin regulator in human erythroid cells. *Science* **361**, 285-290, doi:10.1126/science.aao0932 (2018).
216 Vinjamur, D. S. *et al.* ZNF410 represses fetal globin by singular control of CHD4. *Nat Genet* **53**, 719-728, doi:10.1038/s41588-021-00843-w (2021).
217 Solovieff, N. *et al.* Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5 ' olfactory receptor gene cluster. *Blood* **115**, 1815-1822, doi:10.1182/blood-2009-08-239517 (2010).
218 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).
219 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).
220 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).
221 Chen, M. H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213 e1114, doi:10.1016/j.cell.2020.06.045 (2020).
222 Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat Genet* **46**, 200-204, doi:10.1038/ng.2852 (2014).
223 Liu, D. *RareMETALS*, <https://genome.sph.umich.edu/wiki/RareMETALS> (2017).
224 Mtatiro, S. N. *et al.* Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One* **9**, e111464, doi:10.1371/journal.pone.0111464 (2014).
225 Xu, J. *et al.* Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* **23**, 796-811, doi:10.1016/j.devcel.2012.09.003 (2012).
226 Gautier, E. F. *et al.* Comprehensive Proteomic Analysis of Human Erythropoiesis. *Cell Rep* **16**, 1470-1484, doi:10.1016/j.celrep.2016.06.085 (2016).

227 Ludwig, L. S. *et al.* Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis. *Cell Rep* **27**, 3228-3240 e3227, doi:10.1016/j.celrep.2019.05.046 (2019).
228 Wood, W. G., Weatherall, D. J. & Clegg, J. B. Interaction of heterocellular hereditary persistence of foetal haemoglobin with beta thalassaemia and sickle cell anaemia. *Nature* **264**, 247-249, doi:10.1038/264247a0 (1976).
229 Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat Genet* **42**, 801-805, doi:10.1038/ng.630 (2010).
230 Blobel, G. A. *et al.* An international effort to cure a global health problem: A report on the 19th Hemoglobin Switching Conference. *Exp Hematol* **43**, 821-837, doi:10.1016/j.exphem.2015.06.008 (2015).
231 Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom* **2**, 100168, doi:10.1016/j.xgen.2022.100168 (2022).
232 Harding, S. D. *et al.* The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Res* **50**, D1282-D1294, doi:10.1093/nar/gkab1010 (2022).
233 Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664-1669, doi:10.1038/s41588-019-0538-0 (2019).
234 Kauf, T. L., Coates, T. D., Huazhi, L., Mody-Patel, N. & Hartzema, A. G. The cost of health care for children and adults with sickle cell disease. *Am J Hematol* **84**, 323-327, doi:10.1002/ajh.21408 (2009).
235 Telen, M. J. Beyond hydroxyurea: new and old drugs in the pipeline for sickle cell disease. *Blood* **127**, 810-819, doi:10.1182/blood-2015-09-618553 (2016).
236 Orkin, S. H. & Bauer, D. E. Emerging Genetic Therapy for Sickle Cell Disease. *Annu Rev Med* **70**, 257-271, doi:10.1146/annurev-med-041817-125507 (2019).
237 Chirico, E. N. & Pialoux, V. Role of oxidative stress in the pathogenesis of sickle cell disease. *IUBMB Life* **64**, 72-80, doi:10.1002/iub.584 (2012).
238 Niihara, Y. *et al.* L-glutamine therapy reduces endothelial adhesion of sickle red blood cells to human umbilical vein endothelial cells. *BMC Blood Disord* **5**, 4, doi:10.1186/1471-2326-5-4 (2005).
239 Suhre, K. & Gieger, C. Genetic variation in metabolic phenotypes: study designs and applications. *Nat Rev Genet* **13**, 759-769, doi:10.1038/nrg3314 (2012).
240 Zampieri, M., Sekar, K., Zamboni, N. & Sauer, U. Frontiers of high-throughput metabolomics. *Curr Opin Chem Biol* **36**, 15-23, doi:10.1016/j.cbpa.2016.12.006 (2017).
241 Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet* **27**, R195-R208, doi:10.1093/hmg/ddy163 (2018).
242 Smith, G. D. *et al.* Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med* **4**, e352, doi:10.1371/journal.pmed.0040352 (2007).
243 Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22, doi:10.1093/ije/dyg070 (2003).
244 Hung, R. J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633-637, doi:10.1038/nature06885 (2008).

245 Ference, B. A. *et al.* Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. *J Am Coll Cardiol* **60**, 2631-2639, doi:10.1016/j.jacc.2012.09.017 (2012).
246 Collaboration, C. R. P. C. H. D. G. *et al.* Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ* **342**, d548, doi:10.1136/bmj.d548 (2011).
247 Xu, J. Z. *et al.* Clinical and metabolomic risk factors associated with rapid renal function decline in sickle cell disease. *Am J Hematol* **93**, 1451-1460, doi:10.1002/ajh.25263 (2018).
248 Solovieff, N. *et al.* Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood* **115**, 1815-1822, doi:10.1182/blood-2009-08-239517 (2010).
249 Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525, doi:10.1093/bioinformatics/17.6.520 (2001).
250 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883, doi:10.1093/bioinformatics/bts034 (2012).
251 Asnani, M., Serjeant, G., Royal-Thomas, T. & Reid, M. Predictors of renal function progression in adults with homozygous sickle cell disease. *Br J Haematol* **173**, 461-468, doi:10.1111/bjh.13967 (2016).
252 Arlet, J. B. *et al.* Determination of the best method to estimate glomerular filtration rate from serum creatinine in adult patients with sickle cell disease: a prospective observational cohort study. *BMC Nephrol* **13**, 83, doi:10.1186/1471-2369-13-83 (2012).
253 Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550, doi:10.1038/ng.2982 (2014).
254 Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet* **49**, 568-578, doi:10.1038/ng.3809 (2017).
255 Staley, J. R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207-3209, doi:10.1093/bioinformatics/btw373 (2016).
256 Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304-314, doi:10.1002/gepi.21965 (2016).
257 Bowden, J. *et al.* Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol* **45**, 1961-1974, doi:10.1093/ije/dyw220 (2016).
258 Verbanck, M., Chen, C. Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* **50**, 693-698, doi:10.1038/s41588-018-0099-7 (2018).
259 Niihara, Y., Zerez, C. R., Akiyama, D. S. & Tanaka, K. R. Oral L-glutamine therapy for sickle cell anemia: I. subjective clinical improvement and favorable change in red cell NAD redox potential. *American Journal of Hematology* **58**, 117-121, doi:10.1002/(sici)1096-8652(199806)58:2<117::Aid-ajh5>3.0.Co;2-v (1998).
260 Passon, R. G., Howard, T. A., Zimmerman, S. A., Schultz, W. H. & Ware, R. E. Influence of bilirubin uridine diphosphate-glucuronosyltransferase 1A promoter polymorphisms on serum bilirubin levels and cholelithiasis in children with sickle cell

anemia. *J Pediatr Hematol Oncol* **23**, 448-451, doi:10.1097/00043426-200110000-00011 (2001).
261 Stender, S., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjaerg-Hansen, A. Extreme bilirubin levels as a causal risk factor for symptomatic gallstone disease. *JAMA Intern Med* **173**, 1222-1228, doi:10.1001/jamainternmed.2013.6465 (2013).
262 Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin levels. *Hum Mol Genet* **18**, 2700-2710, doi:10.1093/hmg/ddp202 (2009).
263 Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3, doi:10.1093/nar/gky837 (2019).
264 Hara, N. *et al.* Molecular identification of human glutamine- and ammonia-dependent NAD synthetases. Carbon-nitrogen hydrolase domain confers glutamine dependency. *J Biol Chem* **278**, 10914-10921, doi:10.1074/jbc.M209203200 (2003).
265 Wojcik, M., Seidle, H. F., Bieganowski, P. & Brenner, C. Glutamine-dependent NAD+ synthetase. How a two-domain, three-substrate enzyme avoids waste. *J Biol Chem* **281**, 33395-33402, doi:10.1074/jbc.M607111200 (2006).
266 Hagglund, M. G. A. *et al.* Transport of L-glutamine, L-alanine, L-arginine and L-histidine by the neuron-specific Slc38a8 (SNAT8) in CNS. *J Mol Biol* **427**, 1495-1512, doi:10.1016/j.jmb.2014.10.016 (2015).
267 Toral, M. A. *et al.* Structural modeling of a novel SLC38A8 mutation that causes foveal hypoplasia. *Mol Genet Genomic Med* **5**, 202-209, doi:10.1002/mgg3.266 (2017).
268 Reid, M. A. *et al.* The B55alpha subunit of PP2A drives a p53-dependent metabolic adaptation to glutamine deprivation. *Mol Cell* **50**, 200-211, doi:10.1016/j.molcel.2013.02.008 (2013).
269 Sookoian, S. & Pirola, C. J. Alanine and aspartate aminotransferase and glutamine-cycling pathway: their roles in pathogenesis of metabolic syndrome. *World J Gastroenterol* **18**, 3775-3781, doi:10.3748/wjg.v18.i29.3775 (2012).
270 Osman, W. M. *et al.* Clinical and genetic associations of renal function and diabetic kidney disease in the United Arab Emirates: a cross-sectional study. *BMJ Open* **8**, e020759, doi:10.1136/bmjopen-2017-020759 (2018).
271 Nunes, D. *et al.* LPIN1 deficiency: A novel mutation associated with different phenotypes in the same family. *Mol Genet Metab Rep* **9**, 29-30, doi:10.1016/j.ymgmr.2016.09.004 (2016).
272 Schweitzer, G. G. *et al.* Loss of lipin 1-mediated phosphatidic acid phosphohydrolase activity in muscle leads to skeletal myopathy in mice. *FASEB J* **33**, 652-667, doi:10.1096/fj.201800361R (2019).
273 Stepien, K. M. *et al.* Long-term outcomes in a 25-year-old female affected with lipin-1 deficiency. *JIMD Rep* **46**, 4-10, doi:10.1002/jmd2.12016 (2019).
274 Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov* **15**, 473-484, doi:10.1038/nrd.2016.32 (2016).
275 Adams, C. D. A Mendelian randomization study of circulating glutamine and red blood cell traits. *Pediatr Blood Cancer*, e28333, doi:10.1002/pbc.28333 (2020).
276 Calvani, M. *et al.* Carnitine replacement in end-stage renal disease and hemodialysis. *Ann N Y Acad Sci* **1033**, 52-66, doi:10.1196/annals.1320.005 (2004).
277 Rhee, E. P. *et al.* A combined epidemiologic and metabolomic approach improves CKD prediction. *J Am Soc Nephrol* **24**, 1330-1338, doi:10.1681/ASN.2012101006 (2013).

278 Gerdle, B. *et al.* Decreased muscle concentrations of ATP and PCR in the quadriceps muscle of fibromyalgia patients--a 31P-MRS study. *Eur J Pain* **17**, 1205-1215, doi:10.1002/j.1532-2149.2013.00284.x (2013).
279 Alhourani, H. M., Kumar, A., George, L. K., Sarwar, T. & Wall, B. M. Recurrent Pyroglutamic Acidosis Related to Therapeutic Acetaminophen. *Am J Med Sci* **355**, 387-389, doi:10.1016/j.amjms.2017.08.001 (2018).
280 Moolenaar, S. H. *et al.* beta-Ureidopropionase deficiency: a novel inborn error of metabolism discovered using NMR spectroscopy on urine. *Magn Reson Med* **46**, 1014-1017, doi:10.1002/mrm.1289 (2001).
281 Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635-649, doi:10.1016/j.ajhg.2017.03.004 (2017).
282 DeBaun, M. *et al.* Sickle Cell Disease and Gene Therapy - Patient and Physician Perspectives. *N Engl J Med* **387**, e28, doi:10.1056/NEJMp2212269 (2022).
283 Metaferia, B. *et al.*, doi:10.1101/2022.06.23.497377 (2022).
284 Cannon, M. *et al.* Large-Scale Drug Screen Identifies FDA-Approved Drugs for Repurposing in Sickle-Cell Disease. *J Clin Med* **9**, doi:10.3390/jcm9072276 (2020).
285 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
286 Shen, Y. *et al.* A unified model of human hemoglobin switching through single-cell genome editing. *Nat Commun* **12**, 4991, doi:10.1038/s41467-021-25298-9 (2021).
287 Cheng, L. *et al.* Single-nucleotide-level mapping of DNA regulatory elements that control fetal hemoglobin expression. *Nat Genet* **53**, 869-880, doi:10.1038/s41588-021-00861-8 (2021).
288 Pusztai, L., Hatzis, C. & Andre, F. Reproducibility of research and preclinical validation: problems and solutions. *Nat Rev Clin Oncol* **10**, 720-724, doi:10.1038/nrclinonc.2013.171 (2013).
289 Sever, R. *et al.*, doi:10.1101/833400 (2019).
290 Empty rhetoric over data sharing slows science. *Nature* **546**, 327, doi:10.1038/546327a (2017).
291 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985, doi:10.1093/nar/gkac1010 (2023).
292 Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330, doi:10.1126/science.aaz1776 (2020).
293 Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* **53**, 420-425, doi:10.1038/s41588-021-00783-5 (2021).
294 Saez-Rodriguez, J. *et al.* Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet* **17**, 470-486, doi:10.1038/nrg.2016.69 (2016).
295 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53, doi:10.1126/science.abj6987 (2022).
296 Liao, W. W. *et al.* A draft human pangenome reference. *Nature* **617**, 312-324, doi:10.1038/s41586-023-05896-x (2023).
297 Mukamel, R. E. *et al.* Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499-1505, doi:10.1126/science.abg8289 (2021).
298 Tabet, D., Parikh, V., Mali, P., Roth, F. P. & Claussnitzer, M. Scalable Functional Assays for the Interpretation of Human Genetic Variation. *Annu Rev Genet* **56**, 441-465, doi:10.1146/annurev-genet-072920-032107 (2022).

299 Martin-Rufino, J. D. *et al.* Massively parallel base editing to map variant effects in human hematopoiesis. *Cell*, doi:10.1016/j.cell.2023.03.035 (2023).
300 Picard, M., Scott-Boyer, M. P., Bodein, A., Perin, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* **19**, 3735-3746, doi:10.1016/j.csbj.2021.06.030 (2021).
301 Metaferia, B. *et al.* Phenotypic screening of the ReFRAME drug repurposing library to discover new drugs for treating sickle cell disease. *Proc Natl Acad Sci U S A* **119**, e2210779119, doi:10.1073/pnas.2210779119 (2022).
302 Cai, S. *et al.* Deep Learning Detection of Sea Fan Neovascularization From Ultra-Widefield Color Fundus Photographs of Patients With Sickle Cell Hemoglobinopathy. *JAMA Ophthalmol* **139**, 206-213, doi:10.1001/jamaophthalmol.2020.5900 (2021).
303 Charache, S. *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)* **75**, 300-326 (1996).
304 Serjeant, G. R. *et al.* Causes of death and early life determinants of survival in homozygous sickle cell disease: The Jamaican cohort study from birth. *PLoS One* **13**, e0192710, doi:10.1371/journal.pone.0192710 (2018).
305 Pincez, T. *et al.* Clonal hematopoiesis in sickle cell disease. *Blood* **138**, 2148-2152, doi:10.1182/blood.2021011121 (2021).
306 Zhang, Y. *et al.* Metformin induces FOXO3-dependent fetal hemoglobin production in human primary erythroid cells. *Blood* **132**, 321-333, doi:10.1182/blood-2017-11-814335 (2018).
307 Sun, Q. *et al.* From GWAS variant to function: A study of approximately 148,000 variants for blood cell traits. *HGG Adv* **3**, 100063, doi:10.1016/j.xhgg.2021.100063 (2022).
308 Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* **49**, 139-145, doi:10.1038/ng.3737 (2017).
309 Võsa, U. *et al.*, doi:10.1101/447367 (2018).
310 Jansen, R. *et al.* Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum Mol Genet* **26**, 1444-1451, doi:10.1093/hmg/ddx043 (2017).
311 Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* **24**, 14-24, doi:10.1101/gr.155192.113 (2014).
312 Westra, H. J. & Franke, L. From genome to function by studying eQTLs. *Biochim Biophys Acta* **1842**, 1896-1902, doi:10.1016/j.bbadis.2014.04.024 (2014).
313 Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**, 1339-1348, doi:10.1038/s41588-019-0481-0 (2019).
314 Iolascon, A., Andolfo, I. & Russo, R. Advances in understanding the pathogenesis of red cell membrane disorders. *British Journal of Haematology* **187**, 13-24, doi:10.1111/bjh.16126 (2019).
315 Zarkowsky, H. S., Oski, F. A., Sha'afi, R., Shohet, S. B. & Nathan, D. G. Congenital hemolytic anemia with high sodium, low potassium red cells. I. Studies of membrane permeability. *N Engl J Med* **278**, 573-581, doi:10.1056/NEJM196803142781101 (1968).
316 Andolfo, I., Russo, R., Gambale, A. & Iolascon, A. Hereditary stomatocytosis: An underdiagnosed condition. *Am J Hematol* **93**, 107-121, doi:10.1002/ajh.24929 (2018).
317 De Franceschi, L. *et al.* Evidence for a protective role of the Gardos channel against hemolysis in murine spherocytosis. *Blood* **106**, 1454-1459, doi:10.1182/blood-2005-01-0368 (2005).

318 Hertz, L. *et al.* Is Increased Intracellular Calcium in Red Blood Cells a Common Component in the Molecular Mechanism Causing Anemia? *Front Physiol* **8**, 673, doi:10.3389/fphys.2017.00673 (2017).
319 Petkova-Kirova, P. *et al.* Red Blood Cell Membrane Conductance in Hereditary Haemolytic Anaemias. *Front Physiol* **10**, 386, doi:10.3389/fphys.2019.00386 (2019).
320 Vandorpe, D. H. *et al.* Hypoxia activates a Ca2+-permeable cation conductance sensitive to carbon monoxide and to GsMTx-4 in human and mouse sickle erythrocytes. *PLoS One* **5**, e8732, doi:10.1371/journal.pone.0008732 (2010).
321 An, X. *et al.* Conformational stabilities of the structural repeats of erythroid spectrin and their functional implications. *J Biol Chem* **281**, 10527-10532, doi:10.1074/jbc.M513725200 (2006).
322 Cox, C. D. *et al.* Removal of the mechanoprotective influence of the cytoskeleton reveals PIEZO1 is gated by bilayer tension. *Nat Commun* **7**, 10366, doi:10.1038/ncomms10366 (2016).
323 Flatt, J. F. & Bruce, L. J. The Molecular Basis for Altered Cation Permeability in Hereditary Stomatocytic Human Red Blood Cells. *Front Physiol* **9**, 367, doi:10.3389/fphys.2018.00367 (2018).
324 Burton, N. M. & Bruce, L. J. Modelling the structure of the red cell membrane. *Biochem Cell Biol* **89**, 200-215, doi:10.1139/o10-154 (2011).
325 Evans, E. L. *et al.* RBCs prevent rapid PIEZO1 inactivation and expose slow deactivation as a mechanism of dehydrated hereditary stomatocytosis. *Blood* **136**, 140-144, doi:10.1182/blood.2019004174 (2020).
326 Shi, J. *et al.* Sphingomyelinase Disables Inactivation in Endogenous PIEZO1 Channels. *Cell Rep* **33**, 108225, doi:10.1016/j.celrep.2020.108225 (2020).
327 Ilboudo, Y. *et al.* A common functional PIEZO1 deletion allele associates with red blood cell density in sickle cell disease patients. *Am J Hematol* **93**, E362-E365, doi:10.1002/ajh.25245 (2018).
328 Lettre, G. & Bauer, D. Fetal Hemoglobon in sickle-cell disease: from genetic epidemiology to new therapeutic strategies. *Lancet* **387** (2016).
329 Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-913, doi:10.1038/ng2088 (2007).
330 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).