Université de Montréal

Genetic Determinants of Clinical Heterogeneity in Sickle Cell Disease

par
Geneviève Galarneau

Département de biochimie
Faculté de médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de doctorat
en bio-informatique

20 Mars 2014

Université de Montréal
Faculté de Médecine

Cette thèse intitulée :

Genetic Determinants of Clinical Heterogeneity in Sickle Cell Disease

Présentée par :
Geneviève Galarneau

a été évaluée par un jury composé des personnes suivantes :

Daniel Sinnett, président-rapporteur
Guillaume Lettre, directeur de recherche
Luis Barreiro, membre du jury
Robert Sladek, examinateur externe
Edward Bradley, représentant du doyen de la FES

Résumé

L'anémie falciforme est une maladie monogénique causée par une mutation dans le locus de la β-globine. Malgré le fait que l'anémie falciforme soit une maladie monogénique, cette maladie présente une grande hétérogénéité clinique. On présume que des facteurs environnementaux et génétiques contribuent à cette hétérogénéité. Il a été observé qu'un haut taux d'hémoglobine fœtale (HbF) diminuait la sévérité et la mortalité des patients atteints de l'anémie falciforme.

Le but de mon projet était d'identifier des variations génétiques modifiant la sévérité clinique de l'anémie falciforme. Dans un premier temps, nous avons effectué la cartographie-fine de trois régions précédemment associées avec le taux d'hémoglobine fœtale. Nous avons ensuite effectué des études d'association pan-génomiques avec deux complications cliniques de l'anémie falciforme ainsi qu'avec le taux d'hémoglobine fœtale. Hormis les régions déjà identifiées comme étant associées au taux d'hémoglobine fœtale, aucun locus n'a atteint le niveau significatif de la puce de génotypage. Pour identifier des groupes de gènes modérément associés au taux d'hémoglobine fœtale qui seraient impliqués dans de mêmes voies biologiques, nous avons effectué une étude des processus biologiques. Finalement, nous avons effectué l'analyse de 19 exomes de patients Jamaïcains ayant des complications cliniques mineures de l'anémie falciforme.

Compte tenu de la taille des cohortes de réplication disponibles, nous n'avons pas les moyens de valider statistiquement les variations identifiées par notre étude. Cependant, nos résultats fournissent de bons gènes candidats pour des études fonctionnelles et pour les réplications futures. Nos résultats suggèrent aussi que le β-hydroxybutyrate en concentration endogène pourraient influencer le taux d'hémoglobine fœtale. De plus, nous montrons que la cartographie-fine des régions associées par des études pan-

génomiques peut identifier des signaux d'association additionnels et augmenter la variation héritable expliquée par cette région.

Mots-clés : Anémie falciforme, hémoglobine fœtale, étude d'association pan-génomique, cartographie-fine, séquençage d'exome, analyse de processus biologiques

Abstract

Sickle cell disease is a monogenic disease caused by a mutation in the β-globin locus. Although it is a monogenic disease, it shows a high clinical heterogeneity. Environmental and genetic factors are thought to play a role in this heterogeneity. It has been observed that a high fetal hemoglobin (HbF) levels correlates with a diminution of the severity and mortality of patients with sickle cell disease.

The goal of my project was to identify genetic modifiers of the clinical severity of sickle cell disease. First, I performed the fine-mapping of three regions previously associated with HbF levels. Second, I performed genome-wide association studies with two clinical complications of sickle cell disease as well as with HbF levels. Since no new loci reached array-wide significance for HbF levels, I performed a pathway analysis to identify additional HbF loci of smaller effect size that might implicate shared biological processes. Finally, I performed the analysis of 19 whole exomes from Jamaican sickle cell disease patients with very mild complications.

In conclusion, given the sample size of the replication cohorts available, we do not currently have the means to statistically validate the association signals. However, these results provide good candidate genes for functional studies and for future replication. Our results also suggest that β-hydroxybutyrate in endogenous levels could influence HbF levels. Furthermore, we show that fine-mapping the loci associated in genome-wide association studies can identify additional signals and increase the explained heritable variation.

Keywords : Sickle cell disease, fetal hemoglobin, genome-wide association study, fine-mapping, whole-exome sequencing, pathway analysis

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Abbreviations and Acronyms

ACS: Acute chest syndrome
bp: Base pairs
BOHB: Beta-hydroxybutyrate
CARe: Candidate-gene Association Resource
CEU: Utah residents with ancestry from northern and western Europe
Chr: Chromosome
CI: Confidence interval
CNV: Copy number variation
CO: Carbon monoxide
Cor: Correlation
CSSCD: Cooperative Study on Sickle Cell Disease
DNA: Deoxyribonucleic acid
eSNP: Single-nucleotide polymorphism associated with gene expression
FDA: Food and Drug Administration
FMS: Faculty of Medical Sciences
GC: Genomic control
gDNA: Genomic DNA
GEO: Gene Expression Omnibus
GHSU: Adult Sickle Cell Clinic of Georgia Health Sciences University
GPx1: Glutathione peroxidase 1 enzyme
GVHD: Graft-versus-host disease
GWAS: Genome-Wide Association Study
h2: Heritability
H_2O_2:Hydrogen peroxide
HbA: Adult hemoglobin
HbC: Hemoglobin variant C
HbE: Hemoglobin variant E
HbF: Fetal hemoglobin
HbFz: Normalized fetal hemoglobin levels
HbS: Sickle-cell hemoglobin
HbSS: Two copies of sickle cell mutation
HDAC: Histone deacetylase
HLA: Human leucocyte antigen
HPFH: Hereditary persistence of fetal haemoglobin
HRM: High-resolution melting
IBC: ITMAT-Broad-CARe
IBD: Identity by descent
ICF-1: Immunodeficiency-centromeric instability-facial anomalies syndrome-1
Indel: Small insertion or deletion
JSCCS: Jamaica Sickle Cell Cohort Study
kb: kilobase
KEGG: Kyoto Encyclopedia of Genes and Genomes

LC-MS: Liquid chromatography –mass spectrometry
LCR: Locus control region
LD: Linkage disequilibrium
MAF: Minor allele frequency
mbq: Minimum base quality
MCHC: Mean cell hemoglobin concentration
mmq: Minimum mapping quality
mRNA: Messenger ribonucleic acid
MSH: Multicenter Study of Hydroxyurea in Sickle Cell Anemia
NCBI: National Center for Biotechnology Information
NEPY: Number of events per year
ng: Nanogram
NHLBI: National Heart, Lung and Blood Institute
nM: Nanomolar
NO: Nitric oxide
OR: odds ratio
P: P-value
PBMC: Peripheral blood mononuclear cells
PCA: Principal component analysis
PCR: Polymerase chain reaction
SCA: Sickle cell anemia
SCD: Sickle cell disease
SITT: Silent cerebral Infarct Transfusion Trial
SCT: Sickle cell trait
SCU: Sickle Cell Unit
S.E.: Standard error
shRNA: Short-hairpin ribonucleic acid
SITT: Silent Cerebral Infarct Transfusion Trial
SNP: Single nucleotide polymorphism
SNV: Single nucleotide variant
Ti/Tv: Transition-to-transversion ratio
Ug: Microgram
uL: Microliter
UHWI: University Hospital of the West Indies
UTR: Untranslated region
UWI: University of the West Indies
WES: Whole exome sequencing
YRI: Yoruba in Ibadan, Nigeria

Acknowledgments

My love for science dates to early childhood. As a child, I collected live snails and caterpillars, planted maple seeds to observe them grow and built an electrical circuit with tools and objects found in the family garage. At school, I always liked science classes. In the last two years of high school, unable to choose between the informatics classes and the science classes (biology, chemistry and physics), I decided to attend all, sacrificing the third of my lunch breaks and an elective sports class. I remember saying at the age of 19, during one of these dreamy conversations that if I could just do anything in life, I would like to learn and investigate on DNA and human genes. Although I totally forgot about this conversation until recently, here I am, nine years later. It seems like my passion for genetics dictated my life choices somewhat unconsciously. I cannot claim to understand everything about human genetics or that I know the function of every gene; many mysteries remain to be solved… and my memory has its limits. Nevertheless, this PhD is a dream come true because it means that I have learned and researched on human genetics. I would like to acknowledge the people who have helped me or encouraged me in the realization of this dream.

First of all, I would like to express my profound gratitude to Dr Guillaume Lettre for entrusting me as his first graduate student. I am honored to have had the chance to be supervised by such a talented, passionate and intelligent researcher. I am grateful for his patience, encouragement and his interest in my ideas. I consider myself very lucky to have worked on projects that truly interested me and kept me enthusiastic throughout the years.

I would like to thank all my colleagues of the Lettre lab over the years: Mélissa Beaudoin, Ken Sin Lo, Amidou N'Diaye, Claire Lavoie-St-Amour, Samuel Lessard, Matthieu Schlögel, Nathalie Chami, Valérie Turcot, Simon Langlois, Cécile Low-Kam, Beatriz Kanzi and

Virginie Bertrand-Lehouillier. Together, we created a dynamic and productive atmosphere where I enjoyed working. I would also like to thank the members of the Rioux lab with whom we interacted frequently.

I wish to thank:

A Superstar: Mélissa Beaudoin (et ses doigts de fée), for her amazing efficiency and rigor in the lab, for her always wise judgment and for her support. Mélissa, during these five years, you have been a contagious carrier of motivation and dynamite energy. Over the years, you have been: my colleague, my training partner, my co-pilot, my friend…

An artist: Ken Sin Lo for his help and patience at my arrival while I was getting acquainted with the various programs and to convince us into things that went from decorating the lab to putting together a lip-sync video!

An accomplice: Claire Lavoie-St-Amour. Claire, your stay at the lab was too short! I am thankful for our frequent, long, and loud conversations and laughter.

A number cruncher: Gabrielle Boucher with whom I spent many hours having statistically flavored discussions.

I also want to thank:

My thesis jury: Daniel Sinnett, Luis Barreiro and Robert Sladek for reviewing my thesis.

Miklós Csürös, John Rioux and Daniel Sinnett who provided insightful comments and suggestions on my thesis committees.

Elaine Meunier: for her assistance with my student file during my many years as a student at Université de Montréal.

I would like to acknowledge the financial support from the following organisms: Fonds de recherche du Québec – Santé (FRQS), Fondation de l'Institut de Cardiologie de Montréal and Fondation Go.

For their advice, help and support, I would like to thank Julie Hussin, Marie-Pier Scott-Boyer, Catherine Labbé and Geneviève David. From both the Lettre lab and the Rioux lab, thanks to Catherine, Frédéric, Geneviève D., Gabrielle, Ken, Claudine, Claire, Marie, Samuel and Matthieu because, like Catherine said: «chaque fois c'est mémorable!»

I would also like to thank my amazing friends who have always been understanding and encouraging, who have been showing interest in my project and who gladly celebrated my successes: Sarah Levesque, Roxane Marcotte, Valéry Mathieu, Julie Beauregard-Racine, Véronique Russell, Véronique Massicotte and Nathaniel Bastien. I would like to thank my parents for encouraging me and helping me in my studies during all these years. I am thankful to Diane Beaupré and Stephen Wyatt who provided me advice and encouragement usually from a 450 km distance but sometimes while being on the other side of the planet. Finally, although it might seem awkward, I would like to take the time to thank myself for all the hard work.

Chapter 1 Introduction

1.1 Introduction

The human genome encodes the necessary information for human development. It contains approximately 3.3 billion deoxyribonucleic acid (DNA) base pairs (bp)[1,2] and 20,500 genes[3]. On average, two non-related human beings share 99.9% of their genetic code. The remaining 0.1% represents inter-individual genetic variations. These variations first occur as *de novo* mutations and are sometimes inherited by descendants. *De novo* mutations can arise from two mechanisms: germline mutations or somatic mutations occurring during early embryo development. Germline mutations are the main source of transmitted *de novo* mutations and occur during meiosis, when a germline cell goes through cellular divisions, chromosomal recombinations (crossovers) and DNA replications to form gametes, either four spermatozoa or four ova. Somatic mutations can occur when the DNA is replicated during mitosis, when a cell replicates its DNA and divides into two daughter cells. Many somatic mutations occur during the lifetime of an individual, but they will only remain specific to a limited number of cells. For example, many cancers are triggered by somatic mutations.

Most *de novo* mutations are single nucleotide variations (SNVs), copy number variations (CNVs) or small insertions/deletions (indels). The human germline mutation rate for SNVs is estimated to be between 0.97×10^{-8} and 3×10^{-8} per position per generation[4-9], meaning that an individual carries on average between 32 and 99 *de novo* SNVs[8]. However, this mutation rate is not homogenous across the genome[5]. Some regions are mutation cold spots and have lower mutation rates, while others are more prone to mutations. Most *de novo* mutations are tolerated because they affect a non-coding region, because they do not induce an amino acid change in the protein sequence or because they are present in only one of the two copies of a gene. Nevertheless, some *de novo* mutations are not viable and lead to a miscarriage. Many aspects of our appearance, health and even personality are influenced by genetic variations. Mendelian traits, like Huntington's disease, cystic fibrosis and oculocutaneous albinism, are encoded by one

gene. Complex traits, like height, hypertension, body mass index and Crohn's disease, are influenced by multiple genetic and environmental factors. Some complex traits are influenced by a multitude of loci, like height, for which over 180 associated regions or loci have been identified[10].

A few generations after they first appeared in an individual, SNVs rarely reach high frequencies; they usually remain at a very low frequency or disappear altogether. Occasionally, a SNV becomes fixed in a population, meaning that all individuals in the population are homozygous for the non-reference allele. Fixation or loss of a SNV in a population is usually a result of genetic drift[11], unless the genetic region is under selection pressure. If a mutation is deleterious or beneficial for the species, it will be governed by selection. Negative selection consists of preventing the spread of a deleterious mutation in a population. Positive selection happens when the new mutation confers an advantage to the organism's survival and reproduction in its habitat; individuals with this mutation are most likely to survive and reproduce. The surviving offspring with the mutation transmits it again and so on, while those without it are less likely to survive and reproduce. When a SNV becomes more frequent in a population, the variant is called a single nucleotide polymorphism (SNP). The frequency threshold to declare a SNV a SNP is not universal, but in this work, we will consider that SNPs are SNVs with minor allele frequencies of at least 1%. On average, there is a SNP every 1,000 to 2,000 nucleotides base pairs in the human genome.

1.2 Evolutionary Pressure and Malaria Infection

Malaria is a classical example of an infectious agent inducing evolutionary pressure. This high evolutionary pressure is due to the severity and mortality of malaria infections. Malaria causes fever, headache, chills, vomiting, jaundice and convulsions and may eventually even cause death. The World Health Organization estimated the number of malaria cases in 2012 at 207 million, resulting in 627,000 deaths[12], meaning that every

minute, a person dies of malaria. Most fatal cases are children because their immunity is not as well-developed as adults. Five protozoans cause malaria in humans: *Plasmodium* (*P.*) *falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi*. Female *Anopheles* mosquitos become carriers of these protozoans after they bite an infected individual and can then transmit these parasites to other individuals. When an infected mosquito bites a human, it injects the infectious form of the malaria parasite, a sporozoite (**Figure 1**), into the bloodstream. The sporozoite will then reach the liver, grow and divide to produce thousand of merozoites per liver cell. Merozoites are the haploid form of malaria parasite that invade erythrocytes, once they leave the liver, and replicate themselves in the bloodstream. Once a critical number of merozoites is present inside an erythrocyte, it bursts, releasing the merozoites that will go on to invade other erythrocytes. In some of the infected erythrocytes, merozoites will develop into sexual forms of the parasite, gametocytes. When a mosquito bites an infected human, the male and female gametocytes enter the mosquito and develop into gametes. The male and female gametes then fuse and form a zygote that will later develop into an ookinete, and finally, an oocyst. Inside the oocyst, the sporozoites grow and multiply themselves, until the oocyst bursts. Once the first sporozoites are released in the mosquito, they invade the salivary glands and are ready to be injected into a host when the mosquito bites. Symptoms of malaria usually appear between 10-15 days after the infectious bite, when the merozoites exit the liver and invade the erythrocytes. The most common varieties of malaria infections are *P. falciparum* and *P. vivax.* Severe cases of malaria can be deadly, especially in the absence of treatment. *P. falciparum* infections are the most fatal[13] and can progress to severe illness within 24 hours after the first symptoms if not treated. Symptoms of severe malaria in children include acute anemia, cerebral malaria and respiratory distress. Artemisinin-based therapies and injectable artesunate have helped to reduce malaria mortality in the past few years. These treatments, combined with increased mosquito control and prevention measures, have helped to reduce deaths due to malaria by 26% since 2000[14].

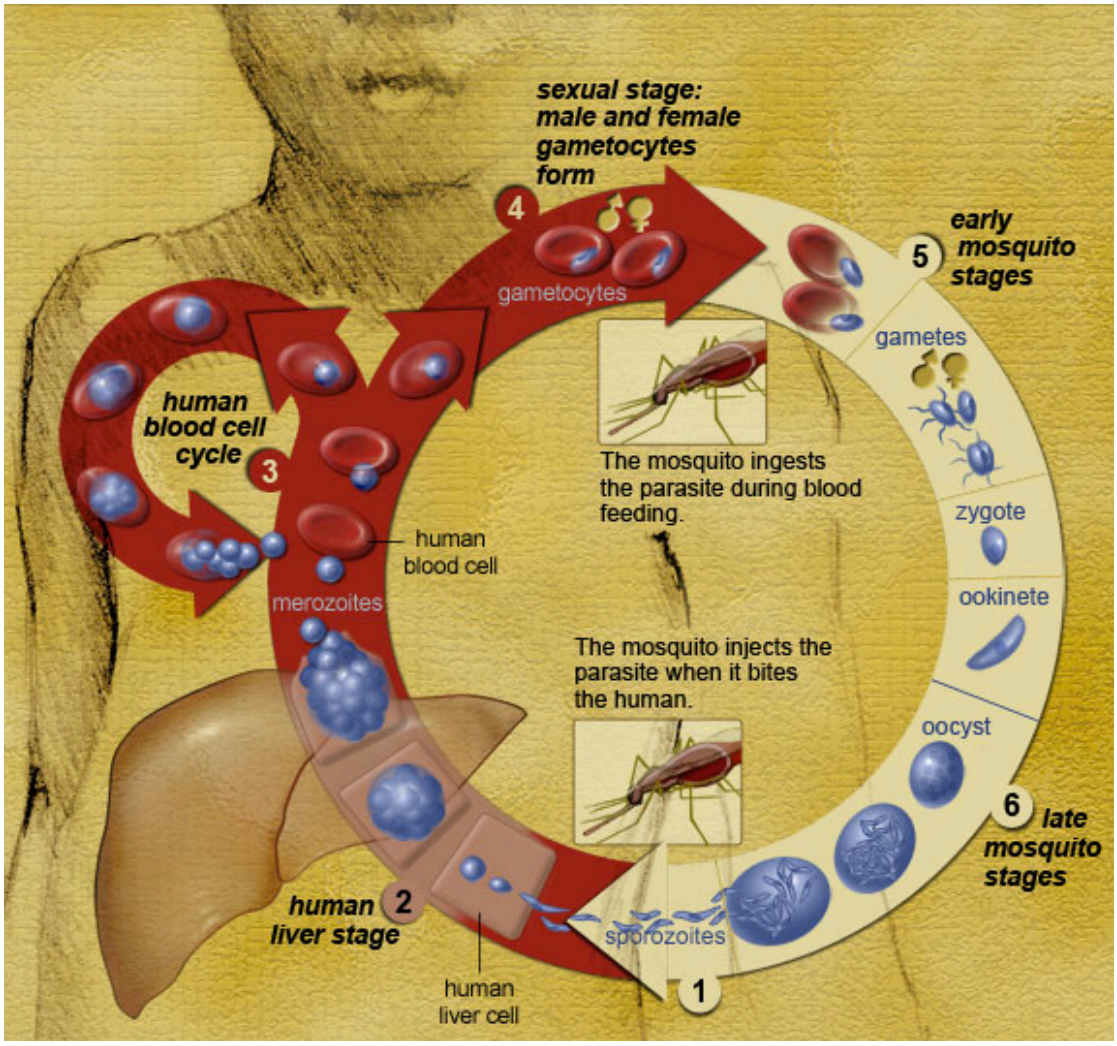


Figure 1 Life cycle of the malaria parasite

(1) (1) When an infected *Anopheles* mosquito bites a human, it injects in the bloodstream the malaria parasite under the form of a sporozoite. (2) The sporozoite will then reach the liver and will divide. (3) The merozoites leave the liver and introduce themselves in the bloodstream to invade the erythrocytes and replicate themselves. Once a critical number of merozoites is present inside the erythrocyte, it bursts, releasing the merozoites that will invade other erythrocytes. (4) In some of the infected erythrocytes, merozoite will develop into sexual forms of the parasite, gametocytes. (5) When a mosquito bites an infected human, the gametocytes are released inside the mosquito and develop into gametes. The male and female gametes then fuse and form a zygote that will later develop in an ookinetes, and finally, an oocysts. (6) Inside the oocyst, the sporozoites grow and multiply themselves, until the oocyst bursts[15].

Malaria infection in humans is believed to have originated approximately 100,000 years ago, creating evolutionary pressure on the human genome. Between 5,000 and 10,000

years ago, this evolutionary pressure became even greater as populations of both humans and malaria mosquito vectors increased rapidly following the introduction of human settlements and agriculture. Three of the most variable genes in the human genome protect against malaria infections: *HBB*, *G6PD* and *DARC*.

1.2.1 Malaria-Protective Effect of *G6PD* Deficiency

The gene *G6PD* is located on the X chromosome and encodes glucose-6-phosphate dehydrogenase, an enzyme implicated in the protection of erythrocytes against oxidative damage. The variants in *GP6D* that cause a deficiency of the enzyme are protective against malaria[16,17]. *G6PD* deficiency is under a very strong balancing selection at an X-linked locus[18,19]. The exact mechanism that protects individuals with a G6PD deficiency against malaria is not fully understood but is thought to be through the effect of the additional oxidative stress resulting from the *G6PD* deficiency.

1.2.2 Malaria-Protective Effect of the Duffy Erythrocyte Silent Alleles

The *DARC* gene encodes the Duffy blood group antigen, a receptor at the human erythrocyte membrane used by *P. vivax* to invade the cell. Four alleles of this gene exists: *FY*A, FY*B* and the Duffy-negative alleles, $FY^{*}B^{ES}$ (erythrocyte silent *FY*B allele*) and $FY^{*}A^{ES}$. The variant leading to the Duffy-negative allele block the expression of the Duffy antigen in the erythrocytes and therefore protects the carriers of this variant against *P. vivax*[20-23]. *The allele* $FY^{*}B^{ES}$ is close to fixation in sub-Saharan Africa. The Duffy gene shows properties of directional selection, a type of selection favoring homozygotes for the new variant, but does not completely fit the theoretical model, suggesting a possible more complex selection signature[24,25].

1.2.3 Malaria-Protective Mutations in the β-Globin Gene

Three different missense mutations in the β-globin gene (*HBB*) protect against malaria. The rs334 mutation, also known as the HbS mutation, changes the seventeenth nucleotide of *HBB* from a thymine to an adenine and the sixth amino acid of the protein from a glutamic acid to a valine (Glu6Val)[26]. When homozygote, this mutation results in sickle cell anemia (SCA). The rs33930165 mutation, or HbC mutation, is located at the same amino acid position as HbS but changes the glutamic acid into a lysine (Glu6Lys)[27]. The third mutation in *HBB* conferring a protection against malaria is rs33946267 (HbE), inducing a change in the amino acid sequence from a glutamic acid to a lysine at position 26 (Glu26Lys)[28]. The HbS mutation, conferring a protection against malaria, is the main mutation causing sickle cell disease (SCD). The definition of SCD includes all the genotypes that cause the clinical conditions of the disease. Individuals with SCD have at least one HbS allele and a deleterious mutation on the other *HBB* allele (**Table 1**). SCA refers to the most common type of SCD, in which an individual carries two copies of the HbS mutation (HbSS). SCA is estimated to represent 70% of SCD cases in populations of African ethnic origin[29].

The HbS mutation in the *HBB* gene is a typical example of a mutation spread by evolutionary pressure. It is mainly found in regions where malaria is, or used to be, endemic (**Figure 2**). Individuals with only one copy of the mutation express the sickle cell trait (SCT). These individuals do not suffer from the most severe SCD-related complications and are protected against severe malaria and deaths from *Plasmodium falciparum*[30-33], the deadliest malaria pathogen[13]. According to two studies, SCT reduces the risk of severe malaria by 90% and hospitalizations due to malaria by 75%-80%[32,34]. From an evolutionary point of view, the protection against malaria of the heterozygotes overcame the SCD suffered by homozygotes, resulting in the transmission and spreading of the HbS mutation. This type of selection, where the heterozygotes are given an advantage compared to homozygotes, is called balancing selection. Based on genetic variation patterns in the same locus (haplotype), it is believed that the HbS mutation

arose and spread five times in human history, between 70,000 and 150,000 years ago[35,36]. Studies have shown that its introduction occurred four times in Africa[37-40] and once in Southern Asia[41]. HbS haplotypes are identified by the region in which they were first discovered: Benin, Senegal, Bantu (also known as Central African Republic) and Arab-Indian (**Figure 3**). The last haplotype is mainly found in Eastern Saudi Arabia and central India. These exceptional events demonstrate the evolutionary pressure caused by malaria and explain why SCD is the most common genetic disorder in the world.

Table 1 Sickle cell disease genotypes

SCD severity	Genotypes	Comments
Severe	HbS/S (β6Glu>Val/β6Glu>Val)	The most common form of sickle cell disease: sickle cell anemia.
Severe	HbS/β^0 thalassaemia	Most prevalent in the eastern Mediterranean region and India[42].
Severe	Severe HbS/β^+ thalassaemia	Most prevalent in the eastern Mediterranean region and India; 1–5% HbA present[42].
Severe	HbS/OArab (β6Glu>Val/β121Glu>Lys)	Reported in north Africa, the Middle East, and the Balkans; relatively rare[42].
Severe	HbS/D Punjab (β6Glu>Val/β121Glu>Gln)	Predominant in northern India but occurs worldwide[42].
Severe	HbS/C Harlem (β6Glu>Val/β6Glu>Val/β, β73Asp>Asn)	Electrophoretically resembles HbSC, but clinically severe; double mutation in β-globin gene; very rare[43].
Severe	HbC/S Antilles (β6Glu>Lys/β6Glu>Val, β23Val–Ile)	Double mutation in β-globin gene results in severe SCD when co-inherited with HbC; very rare[44].
Severe	HbS/Quebec-CHORI (β6Glu>Val/β87Thr>Ile)	Two cases described; resembles SCT with standard analytical techniques[45].
Moderate	HbS/C (β6Glu>Val/β6Glu>Lys)	25–30% cases of SCD in populations of African origin[46].
Moderate	Moderate HbS/β^+ thalassaemia	Most cases in the eastern Mediterranean region; 6–15% HbA present[42].
Moderate	HbA/S Oman (βA /β6Glu>Val, β121Glu>Lys)	Dominant form of SCD caused by double mutation in β-globin gene; very rare[47].
Mild	Mild HbS/β^{++} thalassaemia	Mostly in populations of African origin; 16–30% HbA present[42].
Mild	HbS/E (β6Glu>Val/β26Glu>Lys)	HbE predominates in southeast Asia and so HbSE uncommon, although frequency is increasing with population migration[48].
Mild	HbA/Jamaica Plain (βA /β6Glu>Val, β68Leu/Phe)	Dominant form of SCD; double mutation results in Hb with low oxygen affinity; one case described[49].
Very mild	HbS/HPFH	Group of disorders caused by large deletions of the β-globin gene complex; typically 30% of fetal hemoglobin[42].
Very mild	HbS/other	HbS is co-inherited with many other Hb variants, and symptoms develop only in extreme hypoxia.

Genotypes that have been reported to cause sickle cell disease. SCD: Sickle cell disease. HbS: sickle hemoglobin mutation. HbA: adult hemoglobin, HbE: hemoglobin variant E, HbC: hemoglobin variabt C, Hb: hemoglobin. Modified from [29].

1.2.4 Sickle Cell Trait Protective Mechanism against Malaria Infection

Malaria infections are spread within an individual with the burst of erythrocytes, called hemolysis. Hemolysis releases hemoglobin, and heme quickly releases itself from the oxidized hemoglobin, creating a cytotoxic effect[50]. The survival of the infected individual depends on their ability to prevent the cytotoxic effects of free heme released in the blood due to hemolysis. Both heme oxygenase-1 (HO-1) and carbon monoxide (CO) are protective agents against *P. Plasmodium*[51-54]. The stressed-responsive enzyme HO-1 is expressed with free heme, catabolizes free heme into biliverdin, iron and CO[55] and controls the cytotoxic effect of free heme. CO prevents heme release by the binding the cell-free hemoglobin, preventing it from oxidizing[56]. Sickle hemoglobin induces CO production and HO-1 expression in hematopoietic cells, conferring host tolerance to *Plasmodium* infection[57].

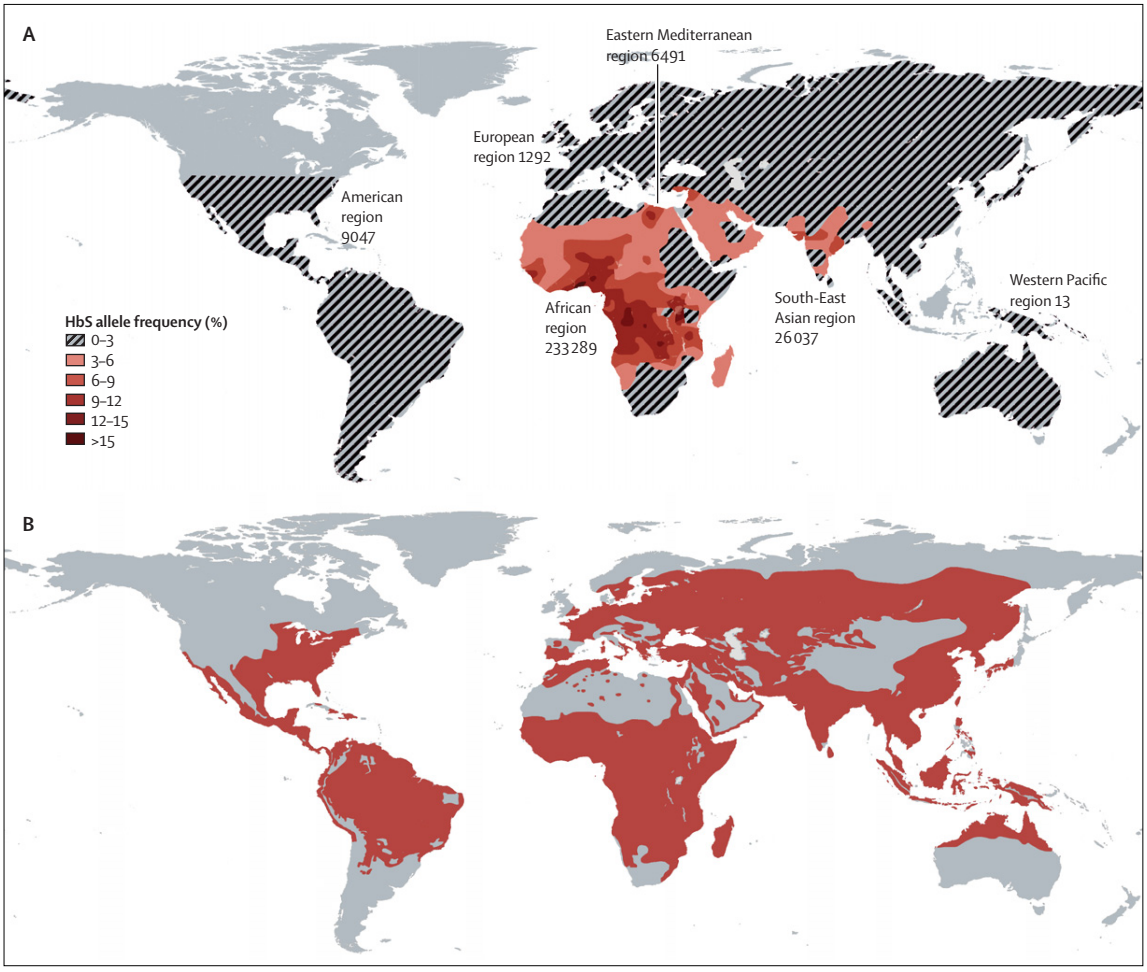


Figure 2 Global distribution of the sickle cell anemia mutation and malaria

(A) Map of the distribution of the sickle cell anemia mutation. (B) Map of the global distribution of malaria (red) before intervention to control malaria. HbS=sickle hemoglobin[29].

Figure 3 Distribution of sickle cell anemia haplotypes among nations with high prevalence of the disease

It is believed that the HbS mutation arose and spread five times in human history, between 70,000 and 150,000 years ago. HbS haplotypes are identified by the region in which they were first discovered: Benin, Senegal, Bantu (also known as Central African Republic) and Arab-Indian[58].

1.2.5 Sickle Cell Disease Epidemiology

Millions of people have SCD around the world. Piel et al. estimated that there were 305,800 newborns with SCD worldwide in 2010[59]. Because of improved survival of SCD patients in high-prevalence low- to middle-income countries, and also because of population migrations to high-income countries, it was suggested that by 2050, the number of newborns with SCD per year would increase to 404,200[59]. An epidemiology study estimated that there are 230,000 SCD newborns in sub-Saharan Africa every year,

representing 0.74% of newborns in this area[60]. In Canada, it is estimated that 1 out of 400 newborns of African descent has SCD[61]. Estimations of the SCD population in the U.S. range from 50,000 to 100,000[62,63].

1.3 Erythrocytes and Hemoglobin

Erythrocytes (red blood cells) are the most common cells in the blood and even in the entire human body. Human erythrocytes are oval biconcave disks and are anucleate, meaning that they do not have a nucleus. Erythrocytes transport the oxygen from the lungs to the tissues via the hemoglobin inside them. Hemoglobin is a tetramer, formed of four globin sub-units each bound to a heme group that contains an iron atom. In humans, there are three different types of hemoglobin: embryonic, fetal and adult. These hemoglobin types differ according to their globin sub-units and the stage of development during which they are produced. The eight genes that encode globin genes are located in two loci: the β-globin gene cluster on chromosome 11 and the α-globin gene cluster on chromosome 16 (**Figure 4**). In both loci, the genes are placed in the same order that they will be expressed throughout development.

Embryonic hemoglobin has three possible conformations: two ζ-subunits and two ε-subunits ($\zeta_2\varepsilon_2$), two α-subunits and two ε-subunits ($\alpha_2\varepsilon_2$) or two ζ-subunits and two γ-subunits ($\zeta_2\gamma_2$). Fetal hemoglobin is composed of two α-subunits and two γ-subunits ($\alpha_2\gamma_2$) (**Figure 4C**). The main form of adult hemoglobin is composed of two α-subunits and two β-subunits ($\alpha_2\beta_2$) (**Figure 4C**), while the alternative form, representing approximately 3% of adult hemoglobin, is composed of two α-subunits and two δ-subunits ($\alpha_2\delta_2$). Different organs throughout the development period produce the hemoglobin (**Figure 5**). Embryonic hemoglobin is produced in the yolk sac for the first six to eight weeks of gestation. Fetal hemoglobin is produced in the liver and the spleen during the gestation. Fetal hemoglobin levels start decreasing shortly before birth and reaches basal levels after the age of one year old. Adult hemoglobin is produced in the

bone marrow. It starts to be produced in low levels after approximately three months of gestation and will slowly increase to become the main type of hemoglobin produced after birth (**Figure 5**).

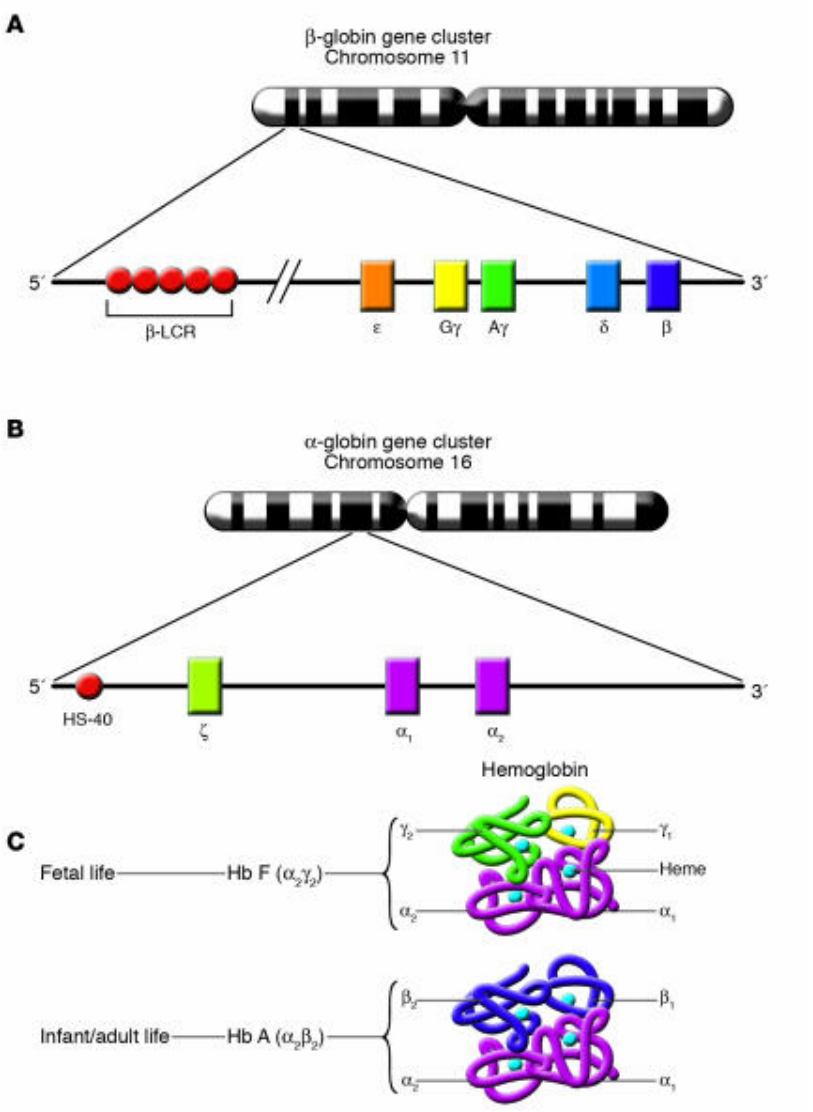


Figure 4 Hemoglobin genes and structure

(A) The β-globin gene cluster on chromosome 11 contains the hemoglobin genes encoding the ε-subunit, the Gγ-subunit, the Aγ-subunit, the δ-subunit and the β-subunit. (B) The α-globin gene cluster on chromosome 16 contains the genes encoding the ζ-subunit the α_1-subunit and the α_2-subunit. (C) Fetal hemoglobin is composed of two α-subunits and two γ-subunits ($\alpha_2\gamma_2$). The main form of adult hemoglobin is composed of two α-subunits and two β-subunits ($\alpha_2\beta_2$)[64].

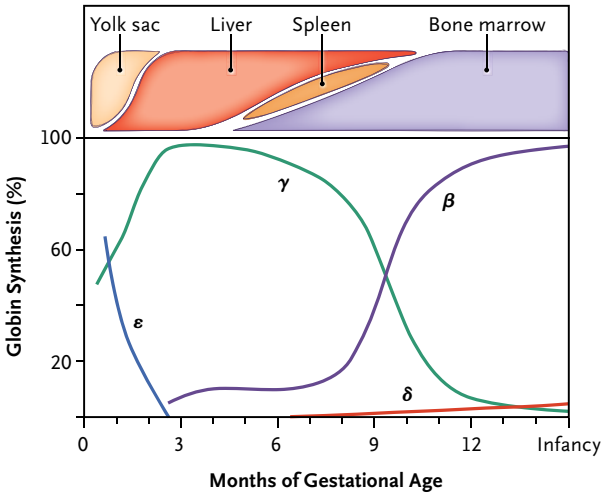


Figure 5 The β-like globin genes expression during development

The ε-globin is expressed in the yolk sac for the first six to eight weeks of gestation. The γ-globin is expressed in the liver and the spleen during the gestation and its expression starts decreasing shortly before birth. The β-globin is expressed in the bone marrow and slowly increases during the end of gestation. The hemoglobin switch occurs when β-globin becomes more expressed than γ-globin[65].

1.4 Sickle Cell Disease

1.4.1 Sickle Cell Disease Pathophysiology

In 1910, Dr James B. Herrick made the first description of SCD from a blood smear of a 20-year-old man from Grenada[66]. Dr Herrick noticed that the young man's erythrocytes had adopted a sickle shape (**Figure 6**). This transformation into the sickle shape is caused by hemoglobin polymerization following deoxygenation inside the red blood cell.

Figure 6 Sickle erythrocyte

The sickle erythrocyte has a different architecture and flexibility than normal erythrocytes[58].

The HbS mutation creates a hydrophobic motif in the deoxygenated hemoglobin structure and creates a binding site between the β1 and β2 chains of two hemoglobin tetramers (**Figure 7**). Interactions between the beta chains favor the growth of hemoglobin polymers within erythrocytes, affecting the cells' flexibility and architecture (**Figure 7**). HbS polymerization inside the erythrocytes is the primary source of pathophysiologic manifestations in SCD patients[67] and leads to erythrocytes remodeling in a sickle shape and promotes cell dehydration. HbS polymerization is proportional to the duration and proportion of deoxygenation. HbS polymerization is also correlated with HbS intracellular concentration and affected by the presence of HbF within the erythrocytes[68,69]. The presence of HbF reduces the polymerization inside the erythrocyte.

Sickle erythrocyte membrane is more rigid and is more likely to break. In a normal individual, an erythrocyte usually circulates for 120 days before it is eliminated by

macrophages, but its survival in SCD patients is reduced to 10-12 days. The burst of the erythrocytes, called hemolysis, releases free heme in the blood. This free heme is toxic[50] and causes the release of adhesion molecules, inducing progressive vasculopathy[70]. The plasma-free heme also scavenges nitric oxide (NO) present in blood. Since NO acts as a vasodilator, this constricts blood vessels. Sickle erythrocytes with increased viscosity and reduced flexibility combined with endothelial activation, vasoconstriction and inflammation block small blood vessels, causing vaso-occlusions (**Figure 8**). These vaso-occlusions can impede the necessary oxygen supply in the affected regions. Hemolysis and vaso-occlusions, both resulting from HbS polymerization, are the main causes of SCD complications.

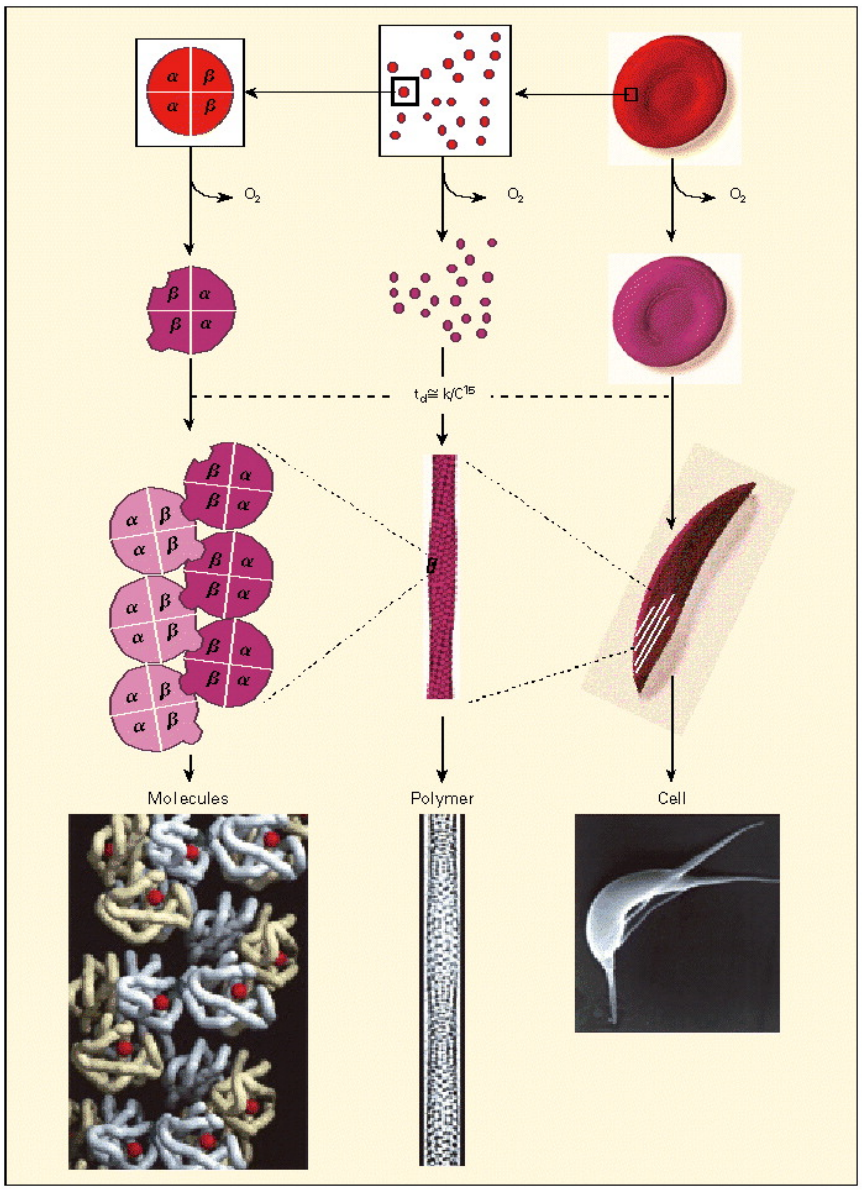


Figure 7 Sickle hemohglobin polymerization

The HbS mutation creates a hydrophobic motif in the deoxygenated hemoglobin structure and creates a binding site between the β1 and β2 chains of two hemoglobin tetramers. Once originated, the polymer grows within the erythrocyte, affecting its flexibility and architecture[68].

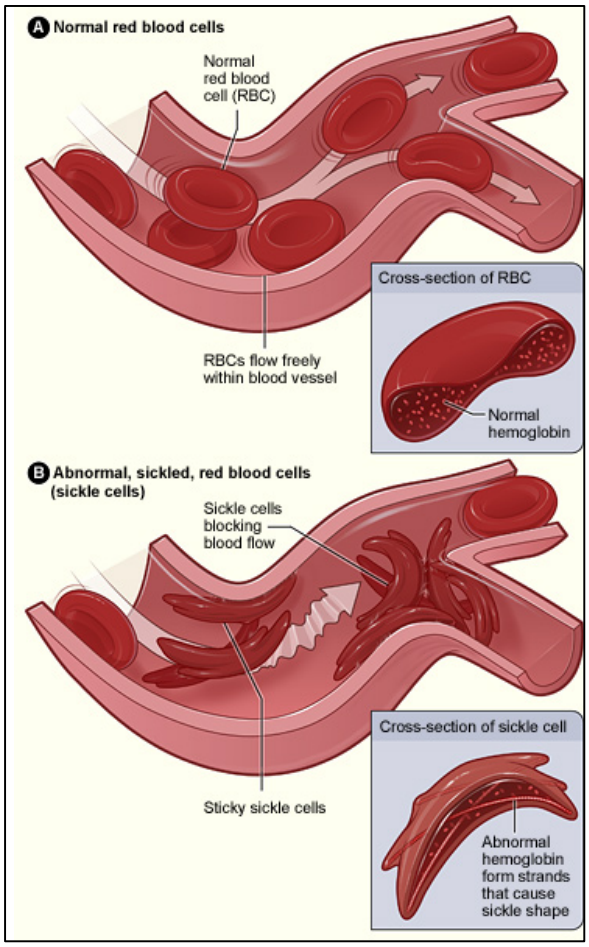


Figure 8 Vaso-occlusion

Sickle erythrocytes are less flexible and more adherent. They can block small blood vessels, causing vaso-occlusion[71].

1.4.2 Sickle Cell Disease Complications

Sickle cell symptoms do not appear until a few months after birth, when the transition from fetal hemoglobin (HbF) to adult hemoglobin (HbA) is advanced. Because they are phenotypically normal, it is usually impossible to identify newborns with SCD, unless a screening test is performed at birth. Once the switch from fetal to adult hemoglobin is completed, the symptoms may start to appear. Although SCD is a monogenic disease, it shows a high clinical heterogeneity among patients. Some patients will have a very high

morbidity, whereas others will show very mild symptoms. However, the majority of SCD patients suffers from daily pain[72] and progressive organ damage. Many of the SCD complications are recurrent and life-threatening.

Dactylitis

Between 32% and 50% of SCD children suffer from dactylitis[73-75], an inflammation of a digit and temporary modification in its bone architecture. Dactylitis is caused by avascular necrosis of the bone marrow in the carpal and tarsal bones and phalanges. Clinical signs and symptoms include swelling, tenderness, redness and pain in the affected area, as well as fever and a high leukocyte count[76,77]. Dactylitis symptoms usually resolve in 5-31 days[76] and the pain can be treated with analgesia. Most cases of dactylitis are seen in children between the age of six months and three years, and rarely occur in patients older than five years[77] because the bone marrow has by then been replaced by fibrous tissue due to continuous asymptomatic bone marrow infarction[78]. Because it is so frequent in SCD patients and occurs at such a young age, dactylitis is often the revealing manifestation of SCD in the absence of screening[76].

Splenic Complications

One of the first organs affected by SCD in young patients is the spleen. The blood circulating through the spleen is filtered, as worn-out erythrocytes are removed and recycled. The spleen also acts as a blood and platelet reservoir in the case of a hemorrhage. Finally, the spleen plays a vital immunological role by synthesizing antibodies and removing and destroying pathogens recognized by these antibodies. Sickle erythrocytes severely damage the spleen by obstructing its oxygen intake. Because of this spleen damage, SCD patients are exposed to bacterial infections at an early age, so current sickle cell treatment now includes penicillin prophylaxis administration in children[79]. Before the instauration of this treatment, the leading cause of deaths in children with SCD was infection. During adulthood, the spleen of SCD patients undergoes

fibrosis and progressive atrophy because of vaso-occlusions and infarction, increasing susceptibility to infections.

Acute splenic sequestration is a life-threatening complication seen almost exclusively in SCD patients younger than five years-old, and especially in children between six months and two years of age. Acute splenic sequestration prevalence in SCA children varies among studies and was estimated at 12.6% in a French cohort[80], 25% in a Jamaican cohort[81] and 40% in a Brazilian cohort[82]. Symptoms include fatigue, fever, vomiting, diarrhea and abdominal pain. Acute splenic sequestration is caused by sickling red blood cells in the spleen, preventing blood circulation and causing an enlargement of the spleen and a drop in hemoglobin levels. If not treated rapidly, acute splenic sequestration can be deadly, as it negatively affects circulation. Acute splenic sequestration is recurrent in 40%-67% of patients with a first event[80,82-84]. Studies in the 1980s observed a mortality of 12% for the first episodes[83] and of 20% in recurrent episodes[84]. However, studies published between 1995 and 2012 observed lower mortality rates: between 0.53% and 5%[80,85]. This decrease in mortality could be due to genetic screening and SCD awareness programs to help parents to better detect and be more aware of acute splenic sequestration. The treatment for acute splenic sequestration can be blood transfusion or a splenectomy, according to the severity of the crisis.

Another complication seldom seen in SCD children is hypersplenism, consisting of a chronic splenic enlargement. This complication also leads to a new blood cell ratio, a bone marrow expansion and red cell sequestration. Hypersplenism was observed in 5% of the children of a Jamaican SCD cohort[86]. About one third of the children who suffer from an acute splenic sequestration develop hypersplenism[81]. Hypersplenism sometimes requires treatment, which can be a splenectomy or chronic transfusions.

Acute chest syndrome

Acute chest syndrome (ACS) is a pulmonary complication with new pulmonary infiltrates that can be diagnosed with radiography or a lung radioisotope scan. In SCD, it is frequent from early childhood and throughout the patients' life. More than two-thirds of SCD patients suffer from ACS and many have multiple episodes[87]. The pathogenesis of this complication is not fully understood. Components of this complication include infection, infarction, fat embolism and pulmonary sequestration. Fat embolism is caused by a severe vaso-occlusive crisis causing edema and infarction of the bone marrow, leading to its necrosis and the release into the bloodstream of fat marrow cells that can reach the lungs, creating an inflammatory response. Pulmonary sequestration can be created by an imbalance between vasoconstriction and vasodilation, increased expression of adhesion molecules and increased secretion of inflammatory cytokines. These induce prolonged sickle erythrocyte transit time in the lungs, microvascular circulation and sequestration of sickle erythrocytes, creating ischemia. Symptoms of ACS include cough, fever, shortness of breath (dyspnea), chills and severe pain. ACS is one of the leading causes of hospitalization in SCD patients[88], and the mean hospital stay varies between 5.4 days and 10.5 days[87,89]. Since the cause of an ACS event often remains undetermined, there is no optimal treatment; treatment strategies are thus varied and can include bronchodilators, delivery of supplemental oxygen, pain management, transfusions and antibiotics. ACS is fatal in approximately 3% of cases[87]. Adults succumb more to ACS than children (4.3% compared to 1.1%)[89]. ACS accounts for up to 25% of deaths in SCD patients[90-92].

Stroke

A stroke is a sudden loss in brain function due to an impaired blood supply. A stroke can cause severe neurological damage and is a leading cause of death in SCD patients. Two studies reported that 11% of SCA patients under 20 years-old suffered from strokes[93,94], and another reported that 8% of SCA patients have a stroke before the age of 14 years-old[95]. The highest incidence among children appears to occur between the age of two and five years-old[93]. Silent infarcts, asymptomatic strokes that can be detected by magnetic

resonance imaging, were identified in 22% of SCA children between the age of 6 and 19 years-old[96]. Another study reported an incidence of silent infarcts of 37% in SCA children by 14 years-old[97]. Strokes are often recurrent, with 50-70% SCD patients with a second stroke within 3 years[95,98,99]. The pathophysiology of strokes in SCD patients is not fully understood. Vasculopathy, vascular remodeling, the release of adhesion molecules and vaso-occlusion are thought to be contributing factors. Strokes can be classified as ischemic, hemorrhagic and transient ischemic attacks. The youngest and oldest SCD patients are more affected by ischemic strokes, while patients between the age of 20 and 29 years-old are more affected by hemorrhagic strokes[93]. Treatments include transfusion, the use of thrombolytic agents, surgical decompression and control of vasospasm. Periodic transfusions have shown to prevent additional strokes in patients with stroke history[95,99,100].

Aplastic Crisis

Another common complication in children with SCD is aplastic crisis, a decrease in reticulocytes (immature erythrocytes in the bone marrow). In most cases, the aplastic crisis is the result of a Parvovirus B19 infection[101] that temporarily destroys the reticulocytes and blocks erythrocyte production. Since erythrocytes in SCD patients have a much shorter lifetime than in healthy individuals, this infection can lead to life-threatening anemia. Since aplastic crisis also implies reticulocyte destruction, it creates additional hemolysis. The treatment for aplastic crisis is blood transfusion, and the incidence of aplastic crises was estimated to be 28% by 10 years-old[102].

Pain crisis

Pain crisis, also known as sickle crisis, is a self-limited episode of excruciating musculoskeletal pain. They are responsible for 79-94% of emergency room and hospital visits in SCD patients[103] and 91% of hospital admissions[104]. Severe pain crisis can last for days and require on average a nine to eleven days hospital stay[105]. Pain crises are usually

located in the extremities, back, abdomen, chest or head and are caused by the entrapment of erythrocytes and leucocytes in microvessels within the bone marrow. This entrapment leads to vascular obstruction, ischemia and necrosis. Inflammation is often a contributing factor to vaso-occlusions. Pain crises are treated with analgesia and may require morphine. Although pain crisis frequently requires medical attention and significantly contributes to SCD morbidity, it is not a life-threatening complication. In a given year, approximately 60% of SCA patients have at least one episode of pain crisis, and 5.2% of patients have between 3 and 10 episodes[106]. There is an increase in pain crisis frequency between the ages of 15 and 25 years-old[106,107]. This increase in frequency is seen more in males than in females, for whom pain crisis rates do not seem to change throughout life[107]. Platt et al. suggested that the reduction in pain crisis rates following the peak rates during adolescence could be due to early death in more severe patients[106].

Leg ulcers

Leg ulcers are cutaneous manifestations of SCD (**Figure 9**). The incidence of leg ulcers in SCA patients has been estimated to be 10-17%[108,109]. In a Jamaican study, leg ulcers tended to be less frequent past the age of 30 years-old, especially in patients without any history[110], but this decrease was not seen in American patients, in which leg ulcers incidence peaked between the ages of 20 and 30 years-old and remained frequent thereafter[109]. Leg ulcer rates are three times higher in males than in females[109]. The pathogenesis of leg ulcers is still nebulous. They often originate from local traumatic lesions, followed by a delayed healing process that can take up to six months and are often recurrent. They heal faster with bed rest and worsen with prolonged standing. Arteriovenous shunting (a connection between an artery and a vein)[111], ischemia, tissue necrosis and secondary infection might be involved as complicating factors. Leg ulcers can be treated with salves, soaks, whirlpool baths, gel boots and local antibiotics, solely or in combination.

Figure 9 Leg ulcer in a patient with sickle cell disease

Leg ulcers are cutaneous manifestations of sickle cell disease. They often originate from local traumatic lesions, followed by a delayed healing process that can take up to six months[112].

Osteonecrosis

Osteonecrosis is caused by ischemia that leads to bone necrosis (death of the bone tissue). Osteonecrosis occurs mainly in the femoral head, with a few cases in the humeral head[113], and can lead to the collapse of the affected bone. The incidence of femoral head osteonecrosis is 8.9% in SCD patients[114]. In patients with femoral head osteonecrosis, 47.3% did not have pain or limitation of motion at time of diagnosis[114]. Prosthetic surgery can be necessary. The pathophysiology of osteonecrosis is poorly understood. Osteonecrosis is thought to be initiated by the entrapment of sickle erythrocytes inside the bone marrow.

Priapism

Postpubertal males with SCD can suffer from priapism, a painful erection unassociated with sexual desire. Priapism is caused either by an unregulated arterial inflow or the persistent obstruction of venous outflow. There are two types of priapism: stuttering priapism and prolonged priapism. Stuttering priapism is an episode of recurrent and brief episodes that resolve spontaneously after two to four hours. Prolonged priapism can last as long as 12 hours. Severe episodes of priapism need rapid intervention because they can lead to vascular damage and impotence[115]. Studies estimated that between 38-42% of men with SCA suffer from priapism[116,117]. The average age of onset has been estimated between 15[118] and 19[117] years-old.

Renal failure

Kidneys serve as filters of the blood, excreting metabolic waste through urine. They also regulate electrolytes and blood pressure, and maintain the acid-base balance. The renal medulla, the inner part of the kidney, is propitious to HbS polymerization and vaso-occlusions causing renal infarction, papillary necrosis and medullary fibrosis. Renal dysfunction in SCD patients is progressive, irreversible and leads to reduction in the glomerular filtration rate as early as 17 months-old[119]. Many adult patients with SCD develop chronic kidney failure, which contributes to SCD morbidity and mortality[120]. Falk et al[121] estimated that 26% of SCD patients have proteinuria and 7% have renal insufficiency. Once the final stages of renal disease are reached, dialysis or renal transplantation is necessary.

Table 2 Summary of SCD most frequent clinical complications

Complication	Description	Percentage of patients with the complication	Age with highest frequency	Potentially fatal
Dactylitis	Inflammation of a digit and temporary modification in its bone architecture	32-50% of SCD children[73-75]	6 months – 3 years-old	No
Acute splenic sequestration	Enlargement of the spleen and drop in hemoglobin levels caused by sickling red blood cells in the spleen	12.6-40% of SCA children[80-82]	6 months -2 years-old	Yes
Hypersplenism	Chronic splenic enlargement leading to a new blood cell ratio, a bone marrow expansion and red cell sequestration	5%[86]	Children under 5 years-old	No
Acute chest syndrome	Pulmonary complication with new pulmonary infiltrates that can be diagnosed with radiography or a lung radioisotope scan	66%[87]	No clear age with higher frequency	Yes
Stroke	Sudden loss in brain function due to an impaired blood supply	Silent infarcts in 22-37%[96, 97]	2-5 years-old	Yes
Aplastic crisis	Decrease in reticulocytes often resulting of a Parvovirus B19 infection that temporarily destroys the reticulocytes and blocks erythrocyte production	28 % by 10 years-old[102]	Childhood	Yes
Pain crisis	Self-limited episode of excruciating musculoskeletal pain	60% of SCA patients have at least one episode per year[106]	15-25 years-old[106,107]	No
Leg ulcer	Cutaneous manifestations of SCD often originating from local traumatic lesions, followed by a delayed healing process	10-17%[108, 109]	No clear age with higher frequency	No
Osteonecrosis	Bone necrosis (mainly in the femoral head, with a few cases in the humeral head)	8.9%[114]	Adulthood	No
Priapism	Painful erection unassociated with sexual desire	38-42% of males with SCA[116, 117]	Average onset between 15 and 19 years-old[117, 118]	No
Renal failure	Irreversible renal dysfunction in SCD patients due to renal infarction, papillary necrosis and medullary fibrosis	26% have proteinuria and 7% have renal insufficiency[121]	Renal dysfunction is progressive and renal failure is usually reached during adulthood	Yes

1.4.3 Sickle Cell Disease Modifiers

Fetal Hemoglobin

In 1948, Watson described the paucity of complications in SCD infants and observed lower sickling in SCT infants compared to their SCT mothers[122]. She suggested that HbF levels, still elevated in infants, could be responsible for these observations[122]. HbF remains at residual levels in adults and accounts for 1% of total hemoglobin levels in healthy adults. This level is higher in SCD patients and varies from 1-30%, with an average of 8%[29]. High HbF levels have been associated with a better life expectancy (**Figure 10**), and lower morbidity in SCD patients[120,123,124]. HbF does not integrate the hemoglobin polymerization produced by HbS; therefore, increased HbF levels decrease HbS concentration and HbS polymerization inside the erythrocyte, reducing both sickling and hemolysis.

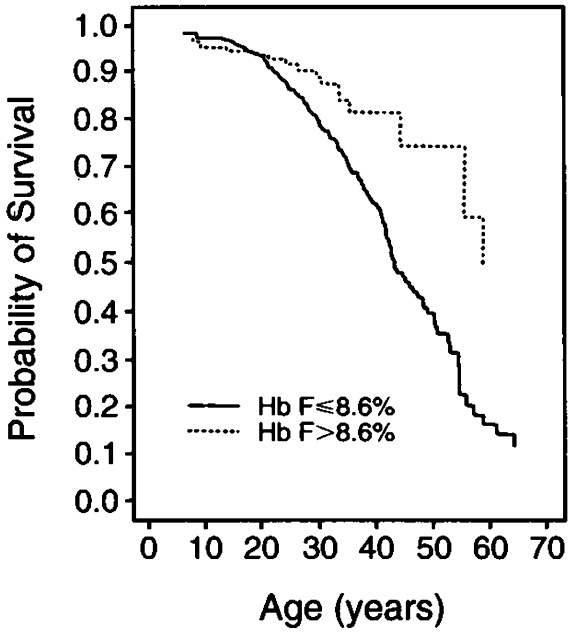


Figure 10 Mortality in patients with sickle cell disease according to fetal hemoglobin levels

High fetal hemoglobin (HbF) levels significantly reduce mortality in patients with sickle cell disease[120].

In addition to being associated with SCD morbidity and mortality, HbF levels have been associated with individual clinical complications. High HbF levels have been associated with fewer events of pain crises[106], acute chest syndrome[88], leg ulcer[109], dactylitis[75], acute splenic sequestration[83] and priapism[116] in SCA patients. Increasing HbF levels in SCD patients has been identified as an interesting potential treatment strategy.

Alpha-thalassemia

In adult hemoglobin, two of the four sub-units are α-globin sub-units. Two adjacent genes on chromosome 16 encode the α-globin sub-unit: *HBA1* and *HBA2*. In individuals with α-thalassemia, there is a deletion of at least one of these genes. It is estimated that 30% of SCD patients of African descent carry the α-thalassemia trait (as heterozygous) and that 4% of patients are homozygous for α-thalassemia[125-127]. The proportion of SCD patients with the α-thalassemia trait is higher than 50% in India[128] and Saudi Arabia[129]. In SCD patients with α-thalassemia, there is a decrease in the mean cell hemoglobin concentration (MCHC)[130]. This decrease in hemoglobin inside the red blood cells reduces hemoglobin polymerization, increases hemoglobin oxygen affinity and decreases hemolysis[106,125,131-133]. SCD patients with concurrent α-thalassemia have demonstrated a reduced occurrence rate of strokes[93,134-137], and leg ulcers[109,125]. Priapism also appears to be slightly less frequent in SCD males with α-thalassemia[138]. The effect of α-thalassemia on ACS is unsettled; some studies claim that it decreases ACS rates[85,125], whereas others observed no effect[88,131]. However, pain crisis rates were significantly higher in SCD patients with α-thalassemia[85,136,139], as was the incidence of osteonecrosis[114,131,140].

Environmental factors

Environmental factors contributing to SCD complications are poorly characterized. Nutrition, hydration, exhaustion, body temperature[86,141] and high altitudes[142] are

believed to affect clinical complications. A handful of studies have observed seasonal variations in certain complications. Two studies in Jamaica observed a seasonal variation in hospital admissions for pain crisis and observed that they were higher during periods of low temperatures[107,143]. Another study conducted in London, UK, observed an association between windy weather and low humidity, and hospital admissions of SCD patients for acute pain[144]. Also, a study on the natural history of dactylitis in SCA reported significantly more episodes during the colder months of the year[75]. In American patients, ACS is more common in winter, especially among children[89].

1.4.4 Sickle Cell Disease Treatment

Potentially fatal complications often occur before SCD symptoms arise in infants with SCD. According to the World Health Organization, many of the babies born in Africa with SCD currently die before the age of 5 years[145,146], and this proportion could be as high as 50% in some regions[86]. In developed countries, there has been tremendous progress in the treatment of SCD children under the age of three years-old, with a 68% decrease in mortality between 1983 and 2002 in the U.S.[147]. This advance can be explained by early diagnosis and prevention measures, such as the implementation of newborn screening programs, vaccination against *Streptococcus pneumonia,* penicillin prophylaxis administration as well as parental counseling programs designed to increase awareness of clinical complications in children with SCD. By 2010, 94% of SCD children in the U.S. reached adulthood[148]. However, the life expectancy of SCD patients still remains well below the general population, with a mean age of death of 39 years old (**Figure 11**)[149]. Currently, bone marrow transplant is the only cure for SCD; other treatments aim to prevent and control SCD complications.

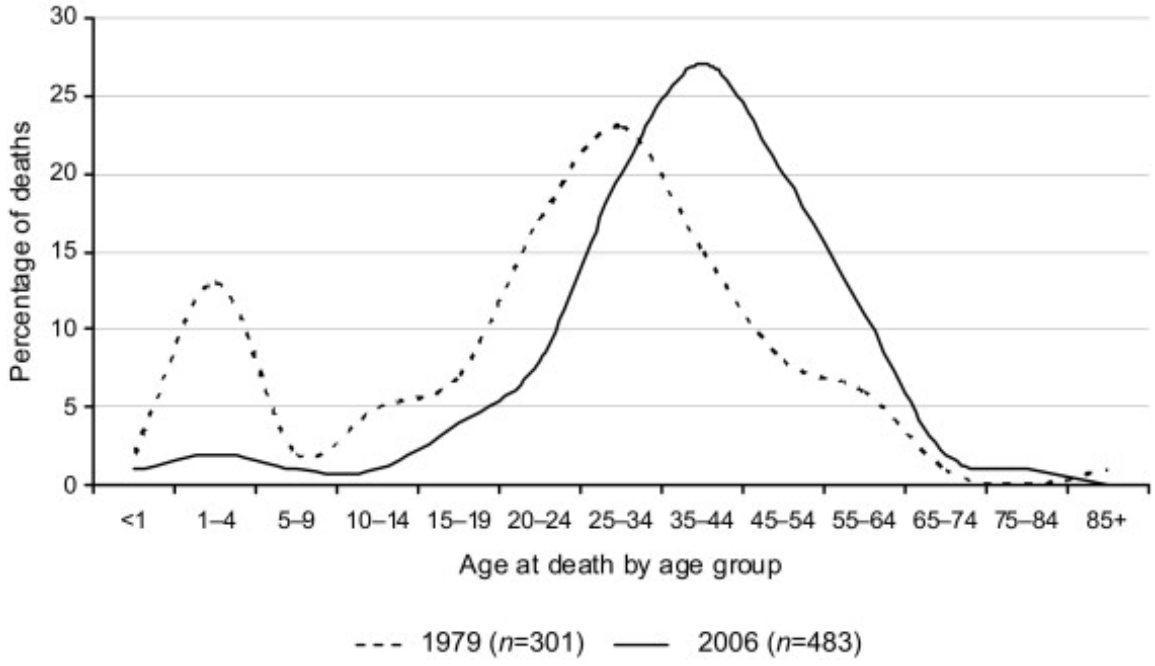


Figure 11 Distribution of age at death of patients with sickle cell disease in 1979 and 2006

The percentage of death in children with sickle cell disease younger than 5 years old has tremendously decrease between 1979 and 2006. However, there was no change in the mean age at death during the same period[149].

Bone Marrow Transplant

Bone marrow transplant is the only cure for SCD. The goal of this procedure is to introduce grafted bone marrow stem cells to produce healthy erythrocytes. First, the recipient is prepared with chemotherapy to destroy their own bone marrow cells. The bone marrow stem cells from the donor are then given to the transplant patient. However, this procedure is very risky, and a compatible donor without SCD is needed. The compatible donor usually has to be a human leucocyte antigen (HLA)-identical sibling. Overall survival in SCD patients who have gone through bone marrow transplantation has now increased above 90%, with disease-free survival rates between 82-100%[150-153]. Nevertheless, 4-14% of these procedures still lead to transplant related mortality[150-153]. Bone marrow transplant is an expensive procedure and therefore

extremely unlikely to be performed in low-income countries where the disease is more prevalent. Because of the difficulty of finding a compatible donor and because of the associated risks, bone marrow transplants are rarely performed in SCD patients, even in high-income countries.

Hydroxyurea

Hydroxyurea, a ribonucleotide reductase inhibitor, was first used in polycythemia vera patients to reduce their abnormally high hematocrit and platelet levels. The effect of hydroxyurea on HbF levels were first observed in baboons[154]. The first study showing that hydroxyurea increased HbF levels in SCD patients was performed in 1984[155]. Hydroxyurea inhibits ribonucleotide reductase activity and selectively inhibits DNA synthesis[156]. It favors the production of F cells, erythrocytes containing high HbF levels and reduces the production of erythrocytes with high levels of sickle hemoglobin. Hydroxyurea's first appeal was that it increased HbF levels in SCD patients[157-159], but we now know that it also reduces platelets and white blood cells, and also stimulates cellular NO production in erythroid progenitors[160]. This treatment also appears to decrease the expression of some adhesion molecules in the vascular endothelium[161].

Hydroxyurea is currently the main treatment for SCD patients. The Food and Drug Administration (FDA) has approved hydroxyurea as a treatment for SCD adult patients with frequent pain crisis. Hydroxyurea shows low cytotoxic effect and was efficient when administered orally[155]. Hydroxyurea treatment has been shown to reduce the frequencies of acute chest syndrome[157,158], pain crisis[157], admissions to hospitals and the need for blood transfusions[157]. In SCD clinical trials, hydroxyurea treatment reduced mortality by 40%[159]. However, de Montalembert et al. estimated that hydroxyurea was not efficient for 20.5% of patients, either because they were not responding or compliant[162].

Blood Transfusions

Blood transfusions in SCD patients help to decrease anemia, reduce the percentage of HbS, reduce hemolysis and reduce the synthesis of HbS. Two types of transfusions can be performed, additive transfusion or exchange transfusion. During an additive transfusion, erythrocytes are simply given to patients, while during an exchange transfusion, the patient's erythrocytes are also removed. Blood transfusions in SCD patients can be used episodically to control complications, or as a chronic transfusion therapy as a preventive strategy in severe cases.

Butyrate

Because it was observed that infants of diabetic mothers had a late switch from fetal to adult hemoglobin[163,164], and that diabetic mothers have higher levels of β-hydroxybutyrate, clinical trials have been performed to attempt to increase HbF levels in SCD patients with butyrate or derivatives. Some clinical trials showed an increase in HbF levels[165], while others were less conclusive and showed a less consistent increase of HbF levels[166]. Because of these mixed results, the possible cytotoxic effect of butyrate and the difficulty of administering the levels of butyrate necessary to increase HbF levels, clinical trials on butyrate have been halted. It is believed that the effect of butyrate would be through the inhibition of histone deacetylases (HDACs)[167].

1.5 From the First Molecular Disease to the Whole-Exome Sequencing of Patients: A Search for Genetic Variants in Sickle Cell Disease Patients

In 1949, Pauling et al. established SCA as the first molecular disease, the first disease caused by an amino acid change[168]. That same year, the autosomal recessive inheritance of SCD was established[169]. However, it was in 1956 that Ingram identified the variant that causes SCA[26]. Between these two events occurred one of the most decisive moment in genetics, the visualization and description of the DNA structure by Watson and Crick[170]. The development of the Sanger sequencing method[171,172] in 1977 also enabled major progress in the field of genetics.

SCD shows a high clinical heterogeneity. Some patients have as only symptoms mild anemia, whereas others have a stroke and multiple pain crises before the age of five years-old. Once the HbS mutation was identified, genetic research on SCD aimed to identify genetic modifiers of the disease. Being the first molecular disease, efforts to identify genetics modifiers of SCD have arisen alongside the development of novel techniques in genetics.

1.5.1 Linkage Studies

Historically, the first genetic association studies were linkage studies, which try to identify segregating chromosomal segments with a Mendelian trait, usually within families. Linkage studies have been successful with very penetrant phenotypes such as Huntington's disease[173] and cystic fibrosis[174]. Linkage studies were also attempted in the context of complex traits. Some led to associations: Factor V^{Leiden} in deep venous thrombosis[175], the *APOEε-4* allele in Alzheimer's disease[176] and PPARγ in type 2

diabetes[177], as well as variants associated with inflammatory bowel disease[178-181], schizophrenia[182] and type 1 diabetes[183]. However, successful associations with linkage studies were much less frequent in complex traits than in Mendelian diseases[184]. This lack of success is due in part to the fact that complex traits are influenced by environmental factors and by multiple common genetic variations of low effect size.

Since the hereditary persistence of fetal hemoglobin (HPFH) is a Mendelian trait that highly reduces SCD severity, it was the first phenotype studied in genetic studies in SCD patients. Linkage studies with HPHF and HbF levels led to the identification of two of the three known loci associated with HbF levels. Many HPFH studies in β-thalassemia and SCD patients identified rare variants in the β-globin cluster inducing HPFH[185-192]. The correlation between the transmission of SCD or β-thalassemia and high HbF levels within families indicated that the β-globin locus was associated with HbF levels and led to the association of the first common variant, *XmnI,* associated with HbF levels[193,194]. Another linkage study in an Asian-Indian healthy kindred with HPFH have also enabled the identification of a novel HbF regulatory region on chromosome 6q[195].

1.5.2 Candidate Gene Studies

Candidate gene studies are performed on genes suspected of being associated with a trait either because of their functions, evidence of associations in previous studies or implication in related diseases. Candidate gene studies target one or a few genes in which a certain number of known variants are genotyped. In the case of HbF levels, a candidate gene study confirmed an association between chromosome 6q and HbF levels, and identified common genetic variants in the *HBS1L-MYB* region associated in healthy individuals[196]. Several candidate gene studies have looked at various SCD complications, but few convincing associations have been reported.

1.5.3 Genome-Wide Association Studies

The completion of the Human Genome Project[1] in 2001 has also been a turning point in the field of human genetics. Data from the first human genome draft and SNP discovery projects such as The International SNP Consortium[197] have led to the first genome-wide map of human genetic variations[198]. By looking at these variations in different individuals, studies have shown the correlation between variants in nearby regions[199-201]. The HapMap project[202], completed in 2005, characterized patterns of variants in three populations. All these events, combined with the crucial development of bioinformatics tools to analyze these colossal datasets, led to the development of whole-genome genotyping arrays and genome-wide association studies (GWAS). These studies interrogate hundreds of thousands common polymorphisms across the genome. Because they are designed based on linkage disequilibrium (correlation patterns among variants in a same region), whole-genome genotyping arrays can cover all the independent signals from common variants in the human genome. In the last decade, GWAS have tremendously helped to identify novel loci associated with complex traits. Between 2005 and June 2012, there were 1,350 GWAS publications[203]. The association between the locus of *BCL11A* on chromosome 2 and HbF levels was identified by a GWAS[204,205].

Because of the number of variants tested in GWAS studies, the significance threshold has to be raised to account for the number of tests performed. The established significance cut-off for GWAS analysis is $\alpha=5x10^{-8}$ [206]. Since in complex traits, effect sizes are small and allele frequencies common, GWAS sample size needs to be very high and may require hundreds of individuals. The replication of the associated locus in an independent dataset has also become a requirement to validate a genetic association. GWAS associations are not sufficient to confirm the causal variant(s) in an associated region. Fine-mapping and functional studies are thus necessary to refine the association signal and validate the causality of the associated variant(s). Even if GWAS have led to the identification of thousands of associated loci, they have so far failed to explain the majority of the expected heritable variation for most complex traits, prompting the

question: “Where is the missing heritability?”[207]. This missing heritability represents the major gap between the expected heritability and the variation explained so far by the variants identified in genetic studies.

1.5.4 High-Throughput DNA Sequencing

The arrival of second generation sequencing technologies resulted in a decrease of the cost of whole-exome and whole-genome sequencing faster than Moore’s law (**Figure 12**). The 1000 Genomes Project was launched in 2008 to identify variants with low frequencies. The first high-throughput sequencing studies mainly focused on the exome because it contains DNA regions encoding the genes, representing one percent of the genome. The first successes of whole-exome sequencing studies were in Mandelian traits: Miller syndrome[208], Schinzel-Giedion syndrome[209], terminal osseous dysplasia[210], nonsyndromic hearing loss DFNB82 [211], ovarian dysgenesis, hearing loss, and ataxia of Perrault syndrome[212], brain malformations[213], Sensenbrenner syndrome[214] and hyperphosphatasia mental retardation syndrome[215]. Most of these studies focused on variants shared between affected individuals and not present in public databases. More recently, whole-exome and whole-genome have shown the burden of de novo mutations in mental illness such as autism or schizophrenia[216,217].

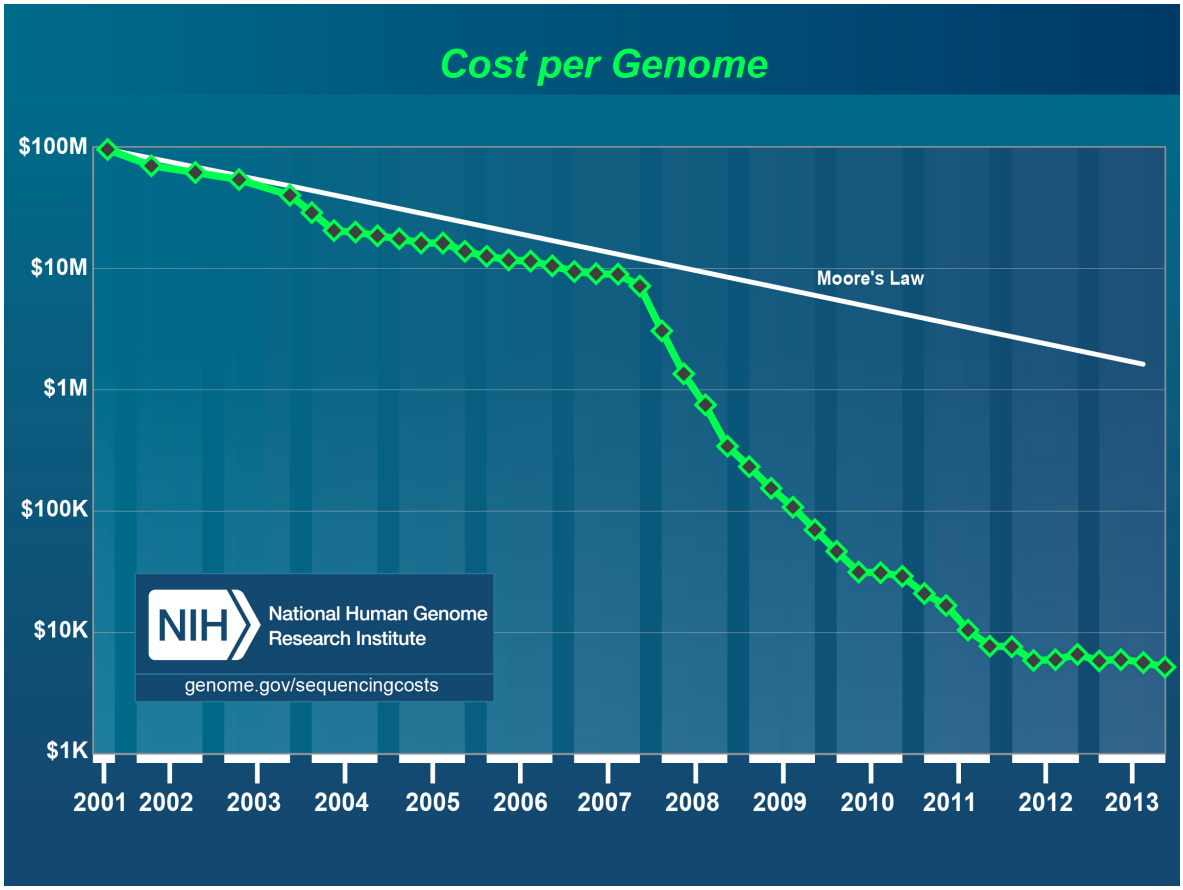


Figure 12 Whole-genome sequencing cost from 2001 to 2013

Cost for whole-genome sequencing decreases faster than Moore's Law, which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years[218].

1.6 The Vital Role of Bioinformatics in GWAS and High-Throughput DNA Sequencing Studies

GWAS datasets contain hundreds of thousands of genotypes for every one of the hundreds, if not thousands, of study participants. A statistical test is performed for each of these variants to verify the association between its genotype and phenotype. PLINK[219] is one of the most widely used bioinformatics toolsets for the analysis of GWAS datasets. The vast majority of

descriptive statistics and classical association tests performed in the context of GWAS are implemented in this software. GWAS were improved by the development of imputation algorithms that enabled the genotype prediction of variants not genotyped on the array by comparing the genotyped variants with reference haplotypes[220-222].

Whole-exome and whole-genome sequencing experiments produce millions of sequences of short DNA fragments, called short reads. All these short reads must go through multiple steps before the variants can be identified. In the first step, each read must be aligned to a reference genome with an alignment tool such as BWA[223]. Once mapped on the genome, a recalibration and realignment step is performed to correct the alignment and the quality score of the reads in regions where small insertions or deletions are present. The last step is the variant calling, in which each difference between the samples sequences and the reference genome is evaluated. To be reported by the variant caller, a variant must meet certain criteria regarding the quality of the reads and the fraction of them that suggest the presence of the variant for a given individual. GATK[224] is one of the most used programs to process high-throughput DNA sequencing data once aligned on the reference genome.

GWAS data and high-throughput sequencing data are gigantic. Without sophisticated bioinformatics tools, the analysis of these datasets would simply be impossible. As whole-exome and whole-genome sequencing prices continue to decrease, the role of bioinformatics will become increasingly crucial for efficiently storing and analyzing these continuously growing datasets.

1.7 Research Questions and Thesis Outline

SCD shows a high clinical heterogeneity from which we ignore the causes. Some patients can suffer from a stroke and multiple pain crises before turning five years old, while others have mild anemia as the only symptom. The best-known modifier of severity and mortality of SCD is HbF levels.

Hypothesis: We expect that both environmental and genetics factors contribute to SCD severity. Genetic association studies could reveal genetic modifiers of SCD severity.

Main goal: This work aimed to identify genetic variants that modify SCD severity. I used different types of genetic association studies to identify such variants.

Specific objectives: In the first project (Chapter 2), I performed the fine-mapping of the three loci identified as associated with HbF levels in SCD patients. I also performed a gene-centric association study with pain crisis rates and acute chest syndrome rates (Chapter 3) and HbF levels (Chapter 4) in SCD patients. Finally, I analyzed the whole-exome sequences of 19 Jamaican patients who exhibited extremely mild SCD (Chapter 5).

Expected impact: Identifying genetic factors modifying SCD severity could help to:

-Identify patients at risk of a more severe spectrum and provide them with more frequent and effective follow-up treatments.

-Better understand the mechanisms underlying SCD pathophysiology and clinical complications and hopefully identify new targets for treatment.

Chapter 2: Fine-Mapping at three Loci Known to Affect Fetal Hemoglobin Levels Explains Additional Genetic Variation

Authors

Geneviève Galarneau, Cameron D. Palmer, Vijay G. Sankaran, Stuart H. Orkin, Joel N. Hirschhorn, Guillaume Lettre

Reference

Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet. 2010 Dec;42(12):1049-51.

Author's Contribution

Conceived and designed the experiment: V.G.S., S.H.O., J.N.H., and G.L. Performed the experiments: G.G., C.D.P., and G.L. Analyzed the data : G.G., C.D.P., and G.L. Contributed reagents/materials/analysis tools: All authors. Wrote the paper: G.G. and G.L., with contributions from all authors.

2.1 Abstract

We used re-sequencing and genotyping in African-American patients with sickle cell anemia (SCA) to characterize association signals to fetal hemoglobin (HbF) levels at the *BCL11A*, *HBS1L-MYB*, and *β-globin* loci. Fine-mapping of HbF association signals at these loci confirmed seven independent SNPs and increased the explained heritable variation in HbF levels from 38.6% to 49.5%. We also identified rare missense variants that causally implicate the *MYB* gene in HbF production.

2.2 Introduction

HbF is a strong and heritable modifier of disease severity for patients with sickle cell disease (SCD; including SCA (HbSS), but also HbSC and HbS-β-thal patients) and β-thalassemia: patients with high HbF levels have less severe complications and a longer life expectancy. Three loci – *BCL11A*, *HBS1L-MYB*, and *β-globin* – carry DNA polymorphisms that modulate HbF levels[196,204,205,225]. To fine-map the HbF association signals, we re-sequenced 175.2 kb from these loci in 190 individuals, including the HapMap CEU and YRI founders, and 70 African-American SCA patients. We discovered 1,489 DNA sequence variants, including 910 previously unreported (**Figure 1**, **Table 1**). Using this information and data from HapMap, we selected and genotyped 95 SNPs in 1,032 African-American SCA patients. We genotyped 17 and 35 SNPs at the *BCL11A* and *HBS1L-MYB* loci, respectively, to characterize previously reported HbF association signals[225]. We also genotyped 43 SNPs at the *β-globin* locus to capture most common genetic variation on the main sickle cell haplotypes.

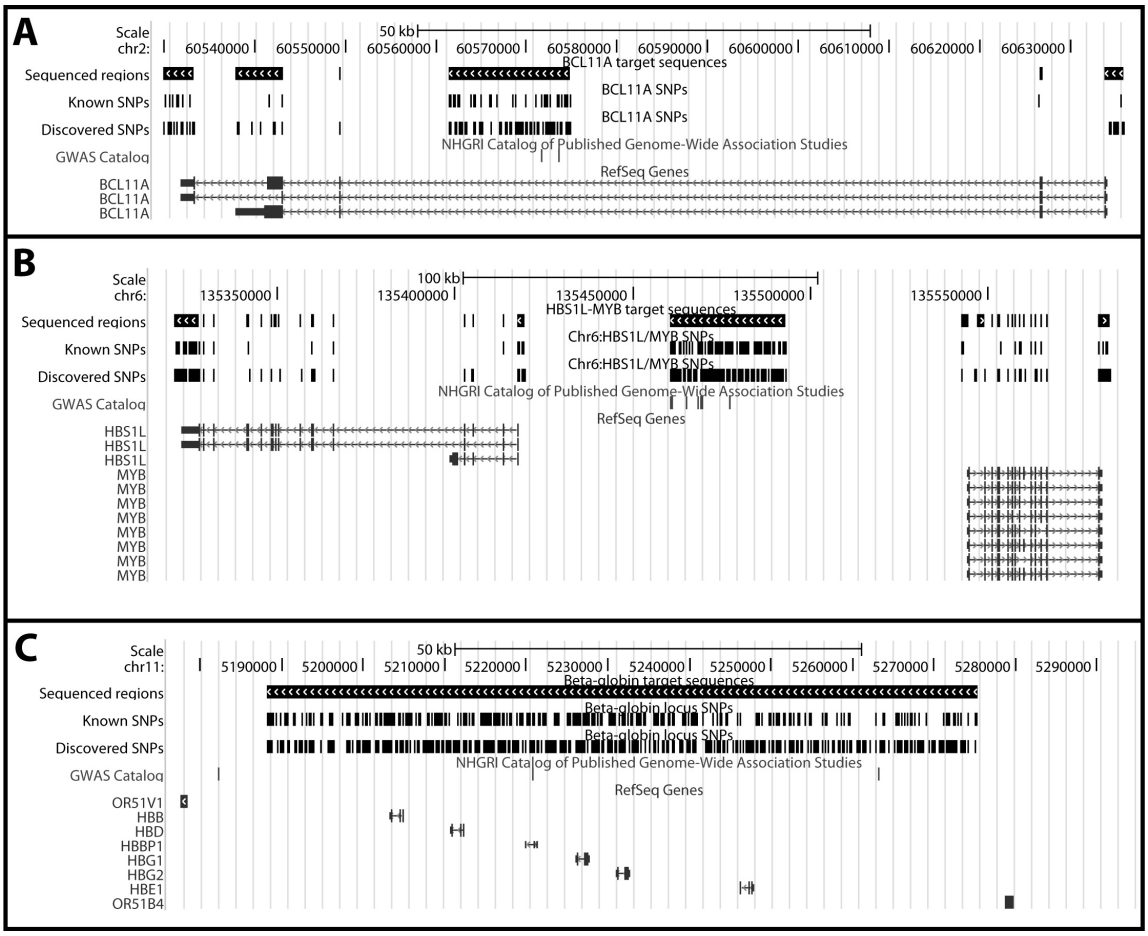


Figure 1 Re-sequencing target regions

The overlay of the UCSC Genome Browser tracks (http://genome.ucsc.edu) with the DNA sequence variants that were previously known ("Known SNPs") and the variants discovered by DNA re-sequencing ("Discovered SNPs") at the *BCL11A* (A), *HBS1L-MYB* (B), and *β-globin* (C) loci. The "Sequenced regions" tract mark regions that we targeted for re-sequencing.

Table 1 Summary of DNA sequence variants identified by re-sequencing the *BCL11A*, *HBS1L-MYB*, and *β-globin* loci

Locus	Targeted region	Number of variants identified
BCL11A	• All *BCL11A* exons (12.1 kb) • Chr2: 60,561,398 - 60,574,851 (13.5 kb)	155 DNA sequence variants (104 novel) 1 non-synonymous (1 novel) 5 synonymous (2 novel) 9 insertion-deletion (9 novel)
HBS1L-MYB	• All *HBS1L* (19.7 kb) and *MYB* (9.8 kb) exons • Chr6: 135,460,328 - 135,493,110 (32.8 kb)	471 DNA sequence variants (271 novel) 18 non-synonymous (12 novel) 10 synonymous (6 novel) 12 insertion-deletion (8 novel)
β-globin	• Chr11: 5,188,125 - 5,275,421 (87.3 kb), including all exons from *HBB* (0.6 kb), *HBD* (0.8 kb), *HBBP1* (0.7 kb), *HBG1* (0.6 kb), *HBG2* (0.6 kb), and *HBE1* (0.8 kb).	863 DNA sequence variants (535 novel) 6 non-synonymous (1 novel) 5 synonymous (2 novel) 29 insertion-deletion (28 novel)
Total	• **175.2 kb**	**1,489 DNA sequence variants (910 novel)**

Genomic positions and annotations are given using NCBI build 36.1. See **Supplementary Table 2** for a complete list of all identified DNA sequence variants, including alleles and functional annotations.**Supplementary Table 2.** 1489 DNA sequence variants identified at the *BCL11A*, *HBS1L-MYB* and *β-globin* loci by DNA re-sequencing in the HapMap Northern European (CEU) and West African (YRI) founders, and in 70 sickle cell anemia (SCA) patients from the Cooperative Study of Sickle Cell Disease (CSSCD).

2.3 Results

BCL11A is a direct repressor of HbF production[226] and a major regulator of developmental globin gene switching[227]. Consistent with previous reports[205,225], rs4671393 in *BCL11A* intron 2 was the genetic marker most strongly associated with HbF levels (P=3.7x10^{-37}) (**Table 2**). Stepwise conditional analyses found two other SNPs, rs7599488 and rs10189857, which independently associate with HbF levels (**Table 2**). These two SNPs, located in *BCL11A* intron 2, are in weak linkage disequilibrium (LD) with rs4671393 (r^2=0.17 and r^2=0.15 for rs7599488 and rs10189857, respectively), but in strong LD with each other (r^2=0.96). When we used principal component analysis to control for admixture, we only observed minor differences (**Supplementary Table 1**).

To further understand the contribution of rs10189857, rs7599488, and rs4671393 to the *BCL11A* HbF association signal, we performed a haplotype analysis. These three SNPs form four haplotypes that represent 99.7% of all haplotypes at this locus. These haplotypes were more strongly associated with HbF levels (P=4.0x10^{-45}) than rs4671393 (P=3.7x10^{-37}), and explained 18.1% of the phenotypic variation in HbF levels (**Table 3**). Thus, these haplotypes explain more phenotypic variance than the cumulative sum of the three *BCL11A* SNPs taken individually (14.7%) (**Table 2**). It is likely that this difference in phenotypic variance explained is due to the presence of HbF-increasing and HbF-decreasing alleles on the same haplotype background, where associated SNPs in LD masked each other's phenotypic effect (**Table 3**). This antagonistic effect could represent an important source of the "hidden" heritability highlighted by genome-wide association study (GWAS) results[207].

Imputation of ungenotyped markers did not reveal other SNPs with stronger association to HbF levels than rs10189857-rs7599488-rs4671393.

Table 2. Fetal hemoglobin (HbF) association results in 1,032 sickle cell anemia (SCA) patients from the Cooperative Study of Sickle Cell Disease (CSSCD)

Locus	SNP	Chromosome (Position)	Effect allele / Other allele	Effect allele frequency	BETA (SE)	*P*-value	Variance explained (%)	BETA (SE)	*P*-value	BETA (SE)	*P*-value
BCL11A					**Univariate analysis**			**Conditional on rs4671393**		**Conditional on rs4671393 and rs7599488**	
	rs4671393	2 (60574455)	A / G	0.27	0.604 (0.046)	3.7×10^{-37}	14.7				
	rs7599488	2 (60571851)	T / C	0.31	0.007 (0.046)	0.89	0.002	0.283 (0.046)	1.2×10^{-9}		
	rs10189857	2 (60566739)	G / A	0.31	-0.010 (0.046)	0.83	0.005	0.241 (0.046)	1.6×10^{-7}	-0.794 (0.223)	3.9×10^{-4}
HBS1L-MYB					**Univariate analysis**			**Conditional on rs9402686**		**Conditional on rs9402686 and ss244317976**	
	rs9402686	6 (135469510)	A / G	0.06	0.650 (0.087)	1.9×10^{-13}	5.1				
	ss244317976	6 (135470367)	G / A	0.02	0.567 (0.150)	1.6×10^{-4}	1.4	0.639 (0.146)	1.3×10^{-5}		
	rs28384513	6 (135417902)	G / T	0.21	-0.098 (0.054)	0.070	0.3	-0.162 (0.053)	0.0024	-0.174 (0.054)	0.0013
β-globin					**Univariate analysis**						
	rs10128556	11 (5220259)	T / C	0.09	0.421 (0.069)	1.3×10^{-9}	3.5				

Genomic positions are given using NCBI build 36.1. Effect allele is on the forward strand. Effect size (BETA) and standard error (SE) are given in *Z*-score units. EAF; effect allele frequency.

Table 3 Haplotype analysis of three SNPs genotyped in *BCL11A* intron 2 of 1,032 African-American sickle cell anemia patients from the CSSCD

SNPs	Haplotypes	Frequency	BETA	*P*-value	Variance explained (%)
rs10189857-rs7599488-rs4671393	GCA (↓↓↑)	0.005	Reference	$4.0X10^{-45}$	18.1
	ACA (↑↓↑)	0.269	0.8063		
	GTG (↓↑↓)	0.308	0.3402		
	ACG (↑↓↓)	0.415	0.05869		

Effect size (BETA) is given in *Z*-score units. In this analysis, the null hypothesis is that all haplotypes have the same effect on HbF levels, whereas the alternate hypothesis is that they all have different effects. Arrows indicate if the allele increases (↑) or decreases (↓) HbF levels.

The *HBS1L-MYB* intergenic interval carries DNA polymorphisms that influence HbF levels in healthy Europeans and SCD patients of African ancestry[196,205,225]. We performed single marker regression analysis and identified rs9402686, which was more strongly associated with HbF levels than the previous index HbF SNP at this locus (P=1.9x10^{-13} for rs9402686 vs. P=3.5x10^{-10} for rs9399137 (Ref. [225])) (**Table 2**). Stepwise conditional analysis uncovered two additional SNPs, ss244317976 and rs28384513, which were independently associated with HbF levels (**Table 2**). LD between rs9402686, ss244317976, and rs28384513 is weak (r^2<0.03). As for *BCL11A*, haplotypes defined by these three SNPs explained more variation in HbF levels than the cumulative sum of the phenotypic variance explained by the SNPs individually (7.3% vs. 6.8%) (**Table 4**).

Table 4 Haplotype analysis using three SNPs located in the *HBS1L-MYB* intergenic region that are independently associated with HbF levels in single marker analysis

SNPs	Haplotypes	Frequency	BETA	*P*-value	Variance explained (%)
rs28384513-rs9402686-ss244317976	AGC (↑↓↑)	0.0189	Reference	$6.4x10^{-17}$	7.3
	CAT (↓↑↓)	0.0285	-0.1101		
	AAT (↑↑↓)	0.0333	0.1617		
	CGT (↓↓↓)	0.174	-0.7544		
	AGT (↑↓↓)	0.742	-0.6092		

These SNPs were genotyped in 1,032 African-American sickle cell anemia patients from the CSSCD. In this analysis, the null hypothesis is that all haplotypes have the same effect on HbF levels, whereas the alternate hypothesis is that they all have different effects. Arrows indicate if the allele increases (↑) or decreases (↓) HbF levels.

In contrast to the *BCL11A* locus[226], we do not know the identity of the gene(s) that influence HbF levels in the *HBS1L-MYB* region. MYB is a transcriptional regulator of erythropoiesis, whereas *HBS1L* expression levels correlate with genotypes at HbF-associated SNPs[196]. We can establish causality by identifying rare and penetrant mutations in nearby candidate genes[228]. Re-sequencing 70 SCA patients identified one, six, and four rare missense variants (minor allele frequency <1%) in *BCL11A*, *HBS1L*, and *MYB*, respectively, that were absent from the 120 CEU and YRI samples. We genotyped these 11 rare variants in 1,032 SCA patients to assess their burden at the gene level by comparing normalized HbF levels in carriers and non-carriers (**Table 5**). To minimize ascertainment bias, we removed re-sequenced SCA patients from this analysis. This excluded singletons and left five and three variants to analyze in *HBS1L* and *MYB*, respectively. Results for *HBS1L* were not significant (*P*corrected=1). However, we observed a significant difference for *MYB* (*P*corrected=0.005), with the 25 carriers having on average 1.4% more HbF than the 937 non-carriers (**Table 5**). These data suggest that *MYB* is causally involved in controlling HbF production.

Table 5. Role of rare functional DNA sequence variants in *HBS1L* and *MYB* on fetal hemoglobin (HbF) levels

Gene	SNP	MAF	Annotation	PolyPhen-2 prediction	Mean % HbF in carriers (N)	Mean % HbF in non-carriers (N)	*P*-value
HBS1L	ss212962438	0.0088	Arg44Trp	Probably damaging	5.83 (17)	6.08 (948)	-
	ss212962440	0.0021	Glu55Lys	Probably damaging	7.56 (4)	6.08 (960)	-
	ss212962441	0.0073	Ser65Cys	Benign	5.60 (14)	6.09 (951)	-
	ss212962478	0.0021	Asp13Glu	Probably damaging	7.96 (4)	6.06 (955)	-
	ss212962504	0.0010	Ser672Tyr	Benign	10.00 (2)	6.07 (962)	-
	All five *HBS1L* missense SNPs	**-**	**-**	**-**	**6.09 (40)**	**6.07 (917)**	**1**
MYB	rs73555746	0.0083	Glu626Ala	Probably damaging	7.87 (15)	6.05 (949)	-
	ss212962653	0.0005	Ser661Leu	Probably damaging	6.60(1)	6.08 (964)	-
	ss212962648	0.0062	Gly628Glu	Benign	6.64(10)	6.09 (953)	-
	All three *MYB* missense SNPs	-	-	**-**	**7.47(25)**	**6.06 (937)**	**0.005**

Rare *HBS1L* and *MYB* missense SNPs with minor allele frequency (MAF) <1% were genotyped in 1,032 African-American sickle cell anemia (SCA) patients from the Cooperative Study of Sickle Cell Disease (CSSCD). We excluded 70 SCA patients used in the re-sequencing phase of this project from this analysis. Gene burden was assessed using Wilcoxon's rank test by comparing normalized fetal hemoglobin levels between carriers and non-carriers. *P*-values are corrected for three genes tested.

Recently, it has been suggested that some of the genetic associations identified by GWAS are due to collections of rare variants captured by common variants[229]. We tested whether the HbF association signals with common SNPs in the *HBS1L-MYB* intergenic region are due to the rare variants identified in *MYB*. LD between the three common SNPs and the three rare missense variants, as measured by D', is high (r^2<0.01, D'>0.4; **Table 6**). When we used the three *MYB* missense variants as covariates, association results between HbF levels and the three common *HBS1L-MYB* SNPs were not affected (**Table 7**), indicating that "synthetic associations" with rare markers in *MYB* cannot explain the HbF association signal in the *HBS1L-MYB* intergenic region. These results provide a clear example where both common and rare DNA sequence variants at the same locus are independently associated with the same phenotype.

Table 6 Linkage disequilibrium (LD) between common SNPs in the *HBS1L-MYB* intergenic region and rare missense variants in *MYB*

LD (r^2/D')	**rs73555746**	**ss212962648**	**ss212962653**
rs28384513	0.002/1	0.005/0.47	0.004/1
ss244317976	0/1	0/1	0/1
rs9402686	0.001/1	0/1	0.003/0.43

Table 7 Association results between HbF levels and the three independent common SNPs in the *HBS1L-MYB* intergenic region before and after conditioning on the three rare missense variants in *MYB*

Common SNPs	Univariate association results	Association results with rs73555746-ss212962648-ss212962653 as covariates
rs28384513	BETA (SE) = -0.098 (0.054); P-value = 0.070	BETA (SE) = -0.098 (0.054); P-value = 0.072
ss244317976	BETA (SE) = 0.567 (0.150); P-value = 1.6×10^{-4}	BETA (SE) = 0.579 (0.150); P-value = 1.1×10^{-4}
rs9402686	BETA (SE) = 0.650 (0.087); P-value = 1.9×10^{-13}	BETA (SE) = 0.652 (0.087); P-value = 1.8×10^{-13}

The sickle cell mutation in the *β-globin* locus is associated with five "classic" haplotypes – Benin, Bantu, Cameroon, Senegal, and Arab-Indian – that are characterized by different degree of clinical severity and HbF levels[230]. An *Xmn*I polymorphism (rs7482144) in the proximal promoter of *HBG2* marks the Senegal and Arab-Indian haplotypes, and is associated with HbF levels in African-American SCD patients[225,231]. It remains unclear whether rs7482144-*Xmn*I is a HbF causal marker at the *β-globin* locus. We reproduced the association between rs7482144-*Xmn*I and HbF levels (P=3.7x10^{-7}). However, rs10128556, located downstream of *HBG1*, was two orders-of-magnitude more strongly associated with HbF levels than rs7482144-*Xmn*I (P=1.3x10^{-9}) (**Table 2**). When we conditioned on rs10128556, the HbF association result for rs7482144-*Xmn*I was not significant (P=0.78; P=0.047 for rs10128556 when conditioned on rs7482144-*Xmn*I) (**Figure 2**). This indicates that rs7482144-*Xmn*I is not a causal marker for HbF levels in African-American SCA patients. Similarly, the recently described association between rs5006884 in the olfactory receptor gene cluster upstream of the *β-globin* genes and HbF levels was not significant after conditioning on rs10128556 (P=0.055; P=1.2x10^{-6} for rs10128556 when conditioned on rs5006884) (**Figure 2**)[232]. Finally, when we conducted a haplotype analysis with the 43 SNPs genotyped at the *β-globin* locus and used rs10128556 as a covariate, the result was not significant (P=0.40), indicating that rs10128556 (or a marker in LD) is the principal HbF-influencing variant at the *β-globin* locus in African-American SCA patients (**Table 8**).

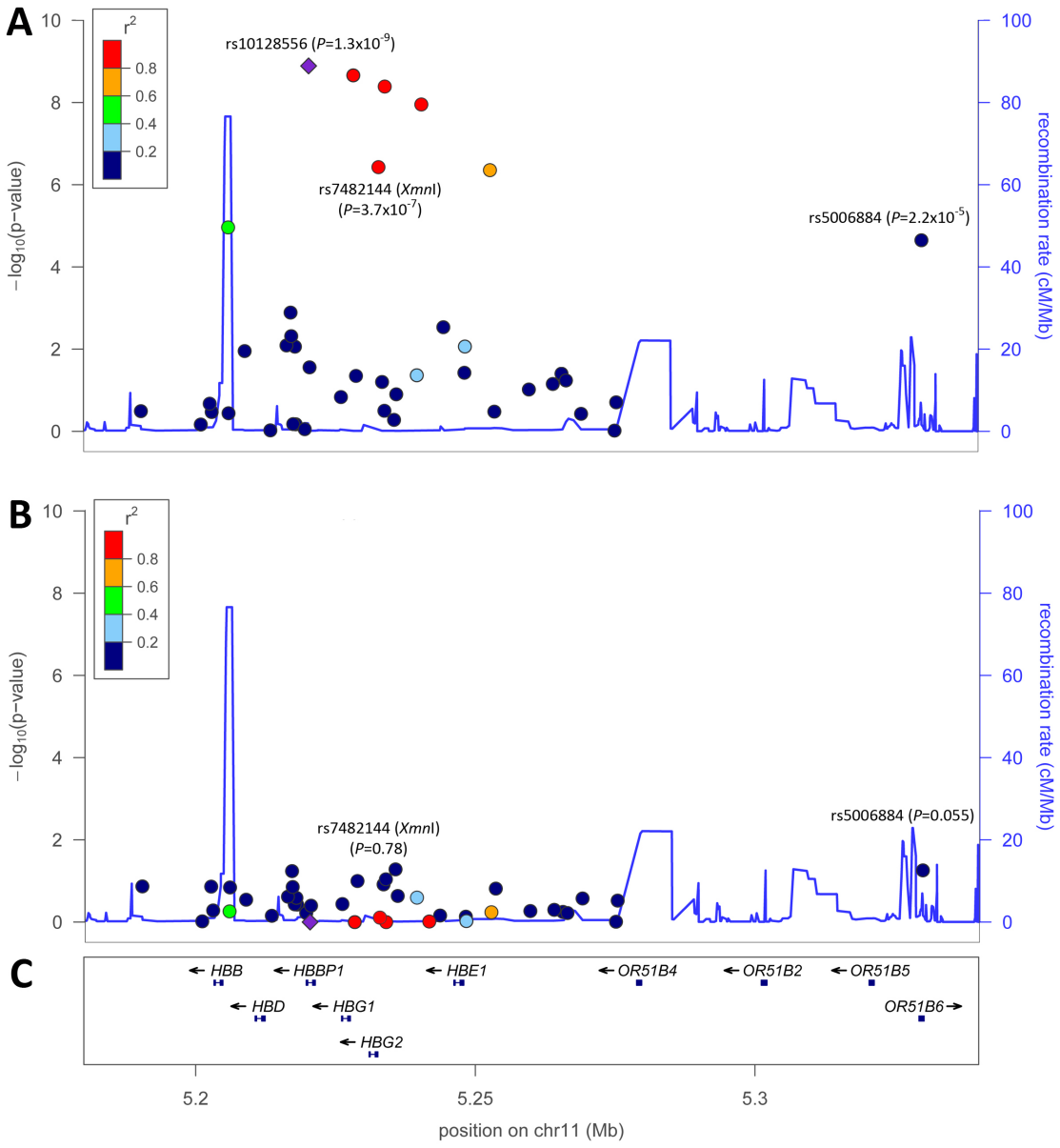


Figure 2 Association to fetal hemoglobin levels for SNPs in the *β-globin* locus that were genotyped in DNA of African-American sickle cell anemia (SCA) patients

The left and right *y*-axes correspond to the strength of association to fetal hemoglobin levels and the recombination rate, respectively. Linkage disequilibrium (r^2) between rs10128556 (diamond) and other genotyped SNPs is based on the Cooperative Study of Sickle Cell Disease (CSSCD) genotype data. The top panel (A) represents single marker association results, whereas the middle panel (B) summarizes marker association results after conditioning on genotype at rs10128556. The bottom panel (C) shows the location of the *β-globin* genes on chromosome 11.

Table 8 Haplotype analysis of the *β-globin* locus using 43 SNPs genotyped in 1,032 African-American sickle cell anemia (SCA) patients from the Cooperative Study of Sickle Cell Disease (CSSCD)

SNPs	Haplotypes	Classic haplotype nomenclature	Haplotype frequency	BETA	P-value	Variance explained (%)
rs6578584-rs10837628-rs11036351-rs10837631-rs16911905-rs1003586-rs6578588-rs9665964-rs2105819-rs60887464-rs968857-rs968856-rs10488679-rs10768687-rs10488678-rs16912203-**rs10128556**-rs16912210-rs2402330-rs2855039-rs2855038-ss212961724-rs7482144 (*Xmn*I)-rs2011051-rs2855122-rs2855121-rs11823688-rs5010981-ss212961284-ss212961261-ss212961256-rs3759071-rs3759070-rs7130110-rs2156918-rs2213169-rs2213167-rs11036635-rs16912979-rs4601817-rs3888708-rs3863292	TACTCTTGGTTTACGG**C**ATCTDGGTGGACCCACGCATACACT	Benin	0.546	Reference	$6.2x10^{-7}$	3.1
	CATTCCCGCTCCAGGG**C**GACTTGGCGCAACGGGGGGCGTACC	Bantu	0.204	-0.091		
	CATTGCTGGTTTACGG**T**AATTTAGCACAATGGGCGGCACACC	Senegal	0.078	0.420		
	CGCAGCTGGTTTACGG**C**AACTDGGCGCCCCCGCGGGCGTACC	Cameroon	0.034	-0.205		
	TACTCTTGGTTTACGG**C**ATCTDGGTGGACCCACGCATACACC	-	0.010	-0.153		
	CGCTCTTGGTTTACGG**C**ATCTDGGTGGACCCACGCATACACT	-	0.009	0.173		
	CATTGCTGGTTTACGG**T**AATTTGGCACAATGGGCGGCACACC	-	0.007	0.228		

Effect size (BETA) is given in *Z*-score units. For each haplotype, the allele for rs10128556 is underlined and in bold, and the allele for rs7482144(*Xmn*I) is highlighted in black with white lettering. We associated the four major haplotypes with their "classic" names using restriction fragment length polymorphism (RFLP)-based genotypes collected by the CSSCD investigators.

2.4 Discussion

The study of the genetic regulation of HbF has provided novel biological insights: BCL11A maintains *γ-globin* silencing and is required for developmental switching within the *β-globin* cluster[226,227]. HbF-associated variants have also shown potential predictive value: they are associated with transfusion-independent β-thalassemia[205,233,234] and reduced pain crisis rate in SCD[225]. In this study, we showed that fine-mapping of known associated loci, through re-sequencing and dense genotyping, can reveal additional independent association signals that could account for a significant fraction of the "hidden" heritability[207]. For HbF levels, we increased the HbF phenotypic variation explained by the same three loci from 23.5% to 30.1%. This translates, assuming a heritability of 60.9%, to an increase from 38.6% to 49.5% of the heritable variation[235]. Thus, characterization of loci identified by GWAS will likely identify untested variants and explain part of the "hidden" heritability.

2.5 Methods

Samples and Phenotypes

The CSSCD is described in details elsewhere[236]. Briefly, the CSSCD was a multicenter, prospective study on the natural history of sickle cell disease and participant enrollment into Phase 1 of the CSSCD began in 1978. Participant entry ended in 1981 for all patients greater than six months of age; however, infants continued to be enrolled until 1988. Both mild and hospital-based sickle cell patients were recruited. A total of 4,085 participants, mostly African Americans and ranging in age from newborns to adults, were enrolled in Phase 1 from 23 centers across the US. Data collection for phase 1 of the CSSCD ended in 1988 (see https://biolincc.nhlbi.nih.gov/studies/csscd/ for more information on the study design).

For re-sequencing, we selected the 60 founders from each the West African (YRI) and Northern European (CEU) HapMap sample collections, as well as 70 sickle cell anemia (SCA) patients from the CSSCD. To maximize the identification of common causal alleles at the HbF loci, we selected these 70 CSSCD participants using the following criteria: (1) 30 patients with low HbF levels (HbF Z-scores ≤ -2 after normalization across the CSSCD) and with the low HbF genotypes at *BCL11A* rs4671393-GG, *HBS1L-MYB* rs9399137-TT and *β-globin XmnI* rs7482144-GG; (2) 30 patients with high HbF levels (HbF Z-scores ≥ 1.2) and with the high HbF genotypes at *BCL11A* rs4671393-AA, *HBS1L-MYB* rs9399137-CT and *β-globin XmnI* rs7482144-AG; and (3) ten patients to over-sample the rarer Cameroun and Senegal sickle cell *β-globin* haplotypes. DNA re-sequencing was performed individually for

each DNA sample selected. For association and conditional analysis, we genotyped CSSCD patients with SCA (N=852) and SCA and α-thalassemia (N=360); 180 DNAs were excluded during the quality control steps (final sample size N=1,032). We excluded from the analysis HbF measurements performed in patients <5 years old, because HbF levels are not yet stable at this early age. To correct for the skewness of the HbF distribution, we log10-transformed and normalized the data to obtain, after correcting for age, gender, and the type of hemoglobinopathy, the quantitative trait used in the association analysis. The protocol for this study was approved by the Institutional Review Board at the Children's Hospital of Boston and the Ethics Committee at the Montreal Heart Institute.

PCR/Sequencing Methods

Resequencing services were provided by the University of Washington, Department of Genome Sciences. Three loci were targeted for DNA re-sequencing (**Figure 1**). First, we re-sequenced all annotated exons of *BCL11A* (12.1 kb), as well as 13.5 kb of *BCL11A* intron 2, flanked by SNPs rs7579014 and rs7557939. This intronic region was re-sequenced to discover all genetic variation in linkage disequilibrium (LD) with rs4671393, a SNP strongly associated with HbF levels in Caucasians and African Americans[205,225]. Second, we re-sequenced all exons from *HBS1L* (19.7 kb) and *MYB* (9.8 kb), as well as 32.8 kb in the *HBS1L-MYB* intergenic region to discover genetic variation in LD with index SNP rs9399137 (flanked by SNPs rs7775698 and rs6934693), a marker strongly associated with HbF levels in African-Americans SCD patients[225]. Third, we re-sequenced the *β-globin* locus (87.3 kb), from 27.5 kb upstream of *HBE1* to 15 kb downstream of *HBB*; this region also includes the *β-globin* locus control region (LCR).

In brief, 5'- M13 tailed-gene specific PCR primers were designed to cover the target region with amplicon sizes ranging from 500-750 bp and with a minimum of 100 bp overlap between adjacent amplicons, where applicable, and resulted in double-stranded coverage of all targeted regions. Overlapping amplicons were used to validate gene-specific primer sequences in independent experiments and rule out the possibility of allele-specific PCR amplifications. All primer sequences were compared to the whole genome assembly to verify uniqueness against pseudogenes and gene families. Following temperature gradient optimization of small-scale reactions to determine optimal thermal cycling conditions, production level PCR amplifications were performed in 96-well plates in a volume of 7 µl comprising 0.2 µl each of 7 µM forward and reverse primers, 2.8 µl DNA (5 ng/µl), and 0.4 µl Elongase Enzyme (Invitrogen) or iProof polymerase (Bio-Rad) per well. Following evaluation by 1% agarose gel electrophoresis, reactions were diluted four to six fold in ddH2O. Dilution of the products eliminated the need for any purification of the PCR products prior to sequencing.

Sequencing reactions were performed in MJ Tetrad PTC 225 thermal cyclers in 384-well format by using 5% BDT v3.1 sequencing chemistry (ABI). Reaction products were precipitated in ethanol with CleanSeq magnetic beads (Agencourt). Perkin Elmer Minitrak, Multiprobe, and Evolution P3 robots were used to automate liquid handling in the setup of PCR, sequencing reactions and precipitation reactions. Reaction products were air dried and diluted to 30 µl with ddH2O. Chromatograms were generated from sequence reaction on an Applied Biosystems ABI 3730XL capillary sequencer. Data flow was tracked by using a custom-designed LIMS system.

SNP Discovery and Analysis

All chromatograms were base-called by using Phred, assembled into contigs by using Phrap, and scanned for SNPs with PolyPhred, version 6.15[237] to identify polymorphic sites. Data quality was monitored and assessed at multiple production checkpoints using numerous methods. For example, each chromatogram was trimmed to remove low-quality sequence (Phred score <25), resulting in analyzed reads averaging >450 bp with an average Phred quality of 40. Following assembly of all chromatograms onto an initial reference sequence, putative polymorphic sites were selectively reviewed by sequence analysts using Consed[238]. Individual polymorphic sites in regions with lower quality data, ambiguous base calls, deviations from Hardy-Weinberg equilibrium or those identified using laboratory quality control tools were reviewed to eliminate potential false positive positions. Outlier genotypes (*i.e.*, deviations from Hardy-Weinberg equilibrium) were scrutinized by data analysts and removed from the dataset if ambiguous. This approach generates sequence-based SNP genotypes with accuracy >99.9%. Variations were deposited into a custom postgreSQL database, formatted and submitted to dbSNP for assignment of ss and rs identification numbers.

To estimate our false negative rate, we used preliminary data from the pilot 1 project (one CEU and one YRI trio at high coverage) of the 1000 Genomes Project (downloaded on August 13 2010). The 1000 Genomes Project identified 154 and 374 novel DNA sequence variants in the genomic regions that we targeted for re-sequencing in CEU and YRI, respectively. Of these, 146 and 341 DNA variants were also identified in our re-sequencing

experiment in CEU and YRI, respectively. Thus, we estimate a false negative rate of 5.2% in CEU and 8.8% in YRI.

DNA Genotyping

All DNA genotyping was performed using the mass-spectrometry based MassArray iPLEX platform from Sequenom[239]. We removed samples with genotyping success rate <90% and SNPs with genotyping success rate <95%. For SNPs passing quality control, the genotyping success rate was >99.5% and the consensus error rate, estimated from replicates, was <0.3%. The Hardy-Weinberg Equilibrium *P*-value was >0.005 for all SNPs. After quality control, genotypes for 95 SNPs and 11 rare sequence variants could be analyzed in the CSSCD. Marker selection for genotyping was based on the following criteria: (1) one, six, and four missense rare mutations in *BCL11A*, *HBS1L*, and *MYB*, respectively, (2) 17 SNPs that tag SNPs in LD ($r^2 \geq 0.1$) in CEU and YRI with HbF-associated *BCL11A* rs4671393, (3) 35 SNPs that tag SNPs in LD ($r^2 \geq 0.1$) in CEU and YRI with HbF-associated *HBS1L-MYB* rs9399137, and (4) 43 SNPs that capture known or discovered genetic variation within the *β-globin* locus. For the rare missense variants, we confirmed that the individuals in which the rare alleles had been discovered by re-sequencing also carry the same alleles using genotyping data.

Imputation

Imputation was performed using MACH 1.0.16 (http://www.sph.umich.edu/csg/abecasis/MaCH/)[222]. MACH requires phased reference haplotypes to perform imputation. For the African Americans, a combined CEU+YRI reference panel was created using publicly available data from HapMap phase 2[240]. This panel includes SNPs segregating in both CEU and YRI, as well as SNPs segregating in one panel and monomorphic and nonmissing in the other. Imputation was performed in two steps. For the first step, a subset of individuals was randomly extracted to generate recombination and error rate estimates. In the second step, these rates were used to impute all individuals across the entire reference panel. Imputation results were filtered at an rsq_hat threshold of 0.3 and a minor allele frequency threshold of 0.01.

Statistical Analysis

All genetic analyses (SNP, imputation dose, and haplotype association testing, conditional analysis, linkage disequilibrium calculations, etc.) were carried out using the software PLINK v1.07[219]. HbF *Z*-scores were analyzed as a quantitative trait using a linear regression framework. Phenotypic variance explained was estimated using the statistical package R 2.10.0 (www.r-project.org/). In conditional analysis, we consider significant association results with P-value less than the pre-defined locus-specific alpha threshold, based on a Bonferroni correction for the number of independent markers tested. The alpha thresholds are: *BCL11A* α=0.0056, *HBS1L-MYB* α=0.0028, *β-globin* α=0.0029. We used likelihood ratio test to show that adding a second and third HbF SNP at the *BCL11A* and *HBS1L-MYB* loci significantly improved data fit with the statistical models; in all cases, adding SNPs

improved the fit of the model (P<0.0025). All analyses tested an additive inheritance model. For the seven SNPs in **Table 2**, when we tested for dominance by comparing additive vs. unconstrained two degrees-of-freedom models using likelihood ratio test, we did not observe deviation from an additive genetic model (P>0.4).

To determine the extent to which the Winner's curse could have estimated upward the effect sizes observed for the SNPs independently associated with HbF levels, we calculated statistical power. For the *BCL11A* and *β-globin* SNPs, when we modeled the allele frequencies observed and an additive inheritance model, we found that we have >80% power even if effect sizes were overestimated twofold (>99% power for the observed effect sizes). For the *HBS1L-MYB* SNPs, power is lower because alleles are rarer: we have 25-65% power to detect these SNPs assuming a twofold overestimation of the effect sizes (55-99% power for the effect sizes observed). It is therefore possible that Winner's curse, mostly at the *HBS1L-MYB* locus, might have slightly inflated the phenotypic variance explained. However, because most of the variation in HbF levels is explained by the index SNP at each of the three loci, the effect of the Winner's curse on the overall variance explained in HbF levels by common genetic variation at the three loci is likely small. Replication in additional larger SCD cohorts would help determine more accurate effect sizes, but such large cohorts currently do not exist.

To evaluate if the phenotypic variance explained by haplotypes is different than the variance explained by the cumulative sum of the individual SNPs, we used bootstrapping procedures (using scripts implemented in the R statistical package – scripts available upon

request). For each model (haplotype or sum of individual SNPs), we performed 1,000 bootstraps. We then compared the distribution of variance explained by both models using a paired Wilcoxon's rank test (similar results were obtained using Student's t-test). For both *BCL11A* and *HBS1L-MYB*, the distributions of variance explained by the two models were very different ($P<2.2x10^{-16}$).

Analysis of Rare Variants

To assess the role of rare putative functional variants on HbF levels variation, we genotyped one *BCL11A*, six *HBS1L*, and four *MYB* missense markers with minor allele frequency (MAF) <1% in the complete CSSCD panel (1,032 African-American SCA patients after QC). One missense variant in each gene was a singleton, that is it was present only once in one of the re-sequenced samples. Because we excluded the 70 SCA patients used for re-sequencing from this rare variants analysis, these three singletons were not considered. We then performed a burden analysis at the gene level in 962 SCA patients (1,032 SCA patients – 70 SCA patients used for re-sequencing): for each gene we divided individuals between non-carriers and carriers of at least one rare functional allele, and used Wilcoxon's rank test to determine if the HbF phenotypic differences observed between non-carriers and carriers were significant. We used normalized phenotypic values for this burden analysis. In support of the result presented in **Table 5** of the main manuscript, we confirmed our results using a recently developed statistical method to analyze rare variants while taking into account missense variant functional effects (as assessed using PolyPhen-2)[241]. Using this variable allele-frequency threshold (VT) method, we obtained corrected P-values of $P_{corrected}=0.023$ and $P_{corrected}=0.49$ for *MYB* and *HBS1L*, respectively.

2.6 Supplementary Information

Supplementary Table 1 Association results with HbF levels after correction for admixture

		Analyses with principal components (N=730)						*Analyses without principal components (N=730)*					
Locus	**SNP**	**BETA**	***P*-value**	**BETA**	***P*-value**	**BETA**	***P*-value**	**BETA**	***P*-value**	**BETA**	***P*-value**	**BETA**	***P*-value**
BCL11A		**Univariate analysis**		**Conditional on rs4671393**		**Conditional on rs4671393 and rs7599488**		**Univariate analysis**		**Conditional on rs4671393**		**Conditional on rs4671393 and rs7599488**	
	rs4671393	0.588	$1.7x10^{-24}$					0.589	$1.8x10^{-24}$				
	rs7599488	0.077	0.176	0.353	$5.6x10^{-10}$			0.076	0.181	0.351	$6.0x10^{-10}$		
	rs10189857	0.055	0.333	0.297	$1.5x10^{-7}$	-0.936	$2.0 x10^{-4}$	0.054	0.341	0.297	$1.4x10^{-7}$	-0.903	$3.2x10^{-4}$
HBS1L-MYB		**Univariate analysis**		**Conditionnal on rs9402686**		**Conditionnal on rs9402686 and ss244317976**		**Univariate analysis**		**Conditionnal on rs9402686**		**Conditionnal on rs9402686 and ss244317976**	
	rs9402686	0.614	$8.7x10^{-9}$					0.590	$2.6x10^{-8}$				
	ss244317976	0.668	$1.8x10^{-4}$	0.725	$3.2x10^{-5}$			0.669	$1.6x10^{-4}$	0.728	$2.7x10^{-5}$		
	rs28384513	-0.159	0.018	-0.211	0.002	-0.220	0.001	-0.155	0.020	-0.209	0.002	-0.218	0.001
β-globin		**Univariate analysis**						**Univariate analysis**					
	rs10128556	0.436	$1.2x10^{-7}$					0.436	$9.5x10^{-8}$				

A subset of the CSSCD participants (N=730) were genotyped on the ITMAT-Broad-CARe (IBC) array by the National Heart, Lung and Blood Institute (NHLBI)-funded Candidate gene Association Resource (CARe) Project[242,243]. The IBC array includes ancestry informative markers (AIMs) that can be used to estimate admixture using principal component analysis (PCA): the first principal component (PC) is almost perfectly correlated with global admixture estimates[244,245]. Correction for global admixture does not change our main conclusions.

2.7 Acknowledgments

We thank all the patients who participated in this study, and Thutrang Nguyen and Mélissa Beaudoin for DNA genotyping support. We thank Soumya Raychaudhuri for critical reading of the manuscript, Gabrielle Boucher for statistical advices, and the CARe Sickle Cell Disease working group for providing the CSSCD principal components. This work was funded by the Fondation de l'Institut de Cardiologie de Montréal (G.L.) and the Doris Duke Charitable Foundation (G.L. and J.N.H.). Resequencing services were provided by the University of Washington, Department of Genome Sciences, under U.S. Federal Government contract number N01-HV-48194 from the National Heart, Lung, and Blood Institute.

Chapter 3: Gene-Centric Association Study of Acute Chest Syndrome and Painful Crisis in Sickle Cell Disease Patients

Authors

Geneviève Galarneau, Sean Coady, Neal Jeffries, Mona Puggal, Dina Paltoo, Antonio Guasch, Abdullah Kutlar, Guillaume Lettre*, George J. Papanicolaou*

*These authors co-directed the study.

Reference

Galarneau G, Coady S, Garrett ME, Jeffries N, Puggal M, Paltoo D, Soldano K, Guasch A, Ashley-Koch AE, Telen MJ, Kutlar A, Lettre G, Papanicolaou GJ. Gene-centric association study of acute chest syndrome and painful crisis in sickle cell disease patients. Blood. 2013 Jul 18;122(3):434-42.

Author's Contribution

G.G., S.C., G.L., and G.J.P. conceived and designed the experiment; G.G., S.C., M.E.G., N.J., K.S., and G.L. performed experiments; G.G., S.C., M.E.G., N.J., M.P., D.P., K.S., A.G., A.E.A.-K., M.J.T., A.K., G.L., and G.J.P. analyzed the results; and G.G. and G.L. wrote the manuscript with contributions from all authors.

3.1 Abstract

Patients with sickle cell disease (SCD) present a wide range of clinical complications. Understanding this clinical heterogeneity offers the prospects to tailor the right treatments to the right patients and also guide the development of novel therapies. Several environmental (e.g. nutrition, body temperature) and non-environmental (e.g. fetal hemoglobin levels, α-thalassemia status) factors are known to modify SCD severity. To find new genetic modifiers of SCD severity, we performed a gene-centric association study in 1,514 African-American participants from the Cooperative Study of Sickle Cell Disease (CSSCD) for acute chest syndrome and painful crisis. From the initial results, we selected 36 SNPs and genotyped them for replication in 387 independent patients from the CSSCD, 318 SCD patients recruited at Georgia Health Sciences University and 472 patients from the Duke SCD cohort. In the combined analysis, an association between ACS and rs6141803 reached array-wide significance (P=4.0x10^{-7}). This SNP is located 8.2 kilobases upstream of *COMMD7*, a gene highly expressed in the lung that interacts with NF-κB signaling. Our results provide new leads to better understand clinical variability in SCD, a "simple" monogenic disease.

3.2 Introduction

Sickle cell disease (SCD) is among the most common Mendelian diseases worldwide and is particularly prevalent in regions where malaria is endemic[246,247]. SCD is caused by mutations in the *β-globin* gene that encodes one of the subunits of the oxygen carrier hemoglobin, and is characterized by a wide spectrum of disease-specific complications. In the deoxygenated state, sickle hemoglobin forms long polymers that alter erythrocyte shape and flexibility, thus increasing hemolysis and the adherence between sickled red blood cells and the endothelium[68,248]. Although hemolysis and cell adherence are the main causes of complications in SCD, little is known about the additional environmental and non-environmental factors that may modify disease severity and therefore explain the remarkable clinical heterogeneity observed in this otherwise simple monogenic disease.

Environmental variables such as nutrition, sufficient hydration and body temperature are linked to clinical heterogeneity in SCD[86,248]. Two non-environmental factors, high fetal hemoglobin (HbF) levels and concomitant α-thalassemia, also correlate with reduced morbidity and mortality in SCD[249]. The identification of additional disease severity modifiers may yield novel insights into SCD pathophysiology. A genetic association study may improve understanding of SCD clinical heterogeneity because it attempts to correlate DNA sequence variants with SCD-specific complications or relevant clinical variables (*e.g.* HbF). Over the last two decades, several genetic associations in SCD have been published, but the results are questionable because of small sample size and lack of replication

(reviewed in ref. [250]). There are, however, three notable exceptions: robust associations between (1) three loci (*BCL11A*, *HBS1L-MYB* and *β-globin*) and HbF levels[204,205], (2) the bilirubin levels-associated *UGT1A1* locus and gallstones[251,252] and (3) the *MYH9-APOL1* locus and SCD nephropathy[253].

To find novel genetic modifiers of SCD, we performed a gene-centric association study in 1,514 participants from the Cooperative Study of Sickle Cell Disease (CSSCD) for acute chest syndrome (ACS) and painful crisis. For genotyping, we used the ITMAT-Broad-CARe (IBC) array, which covers genetic variation at ~2,100 genes important for heart, lung and blood diseases[242]. Although the CSSCD is one of the largest existing SCD cohorts, our discovery power is modest to detect variants of small effect on phenotypic variation. For this reason, we genotyped for replication 36 variants that reached $P<1.0\text{x}10^{-4}$ in the CSSCD discovery sample in the DNA of 387 independent SCD patients from the CSSCD. We also genotyped markers in 318 SCD patients recruited at Georgia Health Sciences University and 472 patients from Duke. Our analysis identified one SNP (rs6141803) near *COMMD7* that reached array-wide significance, defined as $P<2.0\text{x}10^{-6}$ after accounting for the number of independent SNPs present on the IBC array[254]. Overall, our findings prioritize DNA sequence variants and genes for future genetic and functional follow-up experiments in order to better grasp patient-to-patient clinical variability in SCD.

3.3 Material and Methods

Ethics Statement

All participants gave informed written consent. The Candidate-gene Association REsource (CARe) Study is approved by the ethics committees of the participating studies and of the Massachusetts Institute of Technology. This project was also reviewed and approved by the Montreal Heart Institute's ethics committee and the different recruiting centers.

Samples and Genotyping

The CSSCD is described in detail elsewhere[236]. Briefly, the CSSCD was a multicenter, prospective study on the natural history of sickle cell disease and participant enrollment into Phase 1 of the CSSCD began in 1978. Participant entry ended in 1981 for all patients older than six months of age; however, infants continued to be enrolled until 1988. Both mild and hospital-based sickle cell patients were recruited. A total of 4,085 participants, mostly African Americans and ranging in age from newborns to adults, were enrolled in Phase 1 from 23 centers across the US. Data collection for phase 1 of the CSSCD ended in 1988 (see https://biolincc.nhlbi.nih.gov/studies/csscd/ for more information on the study design). In the CSSCD, painful crisis and ACS events were defined as previously described by the CSSCD investigators and as reported in the CSSCD phase 1 clinical database[88,106,225]. Briefly, a painful crisis episode was defined as an occurrence of pain lasting 2 hours or more in the extremities, back, abdomen, chest or head that

could not be explained by a mechanism other than sickle cell. Pain episodes within 14 days were treated as a single episode. An episode of ACS occurred when a participant developed a new infiltrate on chest x-ray and/or had a perfusion defect detected on a lung radioisotope scan. Painful crisis and ACS were analyzed as rates by dividing the number of events by the number of patient-years. 344 patients from the Adult Sickle Cell Clinic of Georgia Health Sciences University Sickle Cell Center were included in this study as a validation cohort, ranging in age from 20-74 and including 185 women and 159 men. Of the 344 patients, 329 had the SS genotype and 15 had SB0 thalassemia. The Duke SCD cohort included 472 adult patients (214 men and 258) and used the two following questions to define ACS (Have you ever experienced acute chest syndrome or pneumonia requiring hospitalization?) and painful crisis (In the past 12 months, have you had painful episodes requiring hospitalization?). Demographics for the three SCD cohorts used in this study are summarized in Table 1.

DNA genotyping on the Illumina ITMAT-Broad-CARe (IBC) array was carried out at the Broad Institute as part of the National Heart, Lung and Blood Institute (NHLBI) CARe Project. The IBC array interrogates genotypes at ~50,000 SNPs and captures genetic variation at ~2,100 genes relevant for heart, lung and blood diseases[242]. Data quality-control and genotype imputation was performed as previously described[254]. We kept in the analysis only markers with an imputation quality rsq_hat >0.6. For imputed markers with strong statistical association, we directly genotyped an overlapping set of CSSCD DNA samples (N=777) and found high

concordance with the imputation results (mean Pearson's correlation coefficient =0.87 (range 0.65-1.0)). The final CSSCD discovery dataset included 1,514 DNA samples with a genotyping success rate >99.8% (47,092 genotyped SNPs and 190,551 imputed SNPs). Genotyping in the CSSCD and Georgia Health Sciences University replication cohorts was performed using the mass-spectrometry-based MassArray iPLEX platform from Sequenom, removing SNPs and DNA samples with genotyping success rate <95% and <90%, respectively. The concordance rate, estimated from replicates, was >99.7%. Genotyping and quality-control filters for the Duke SCD cohort were described elsewhere[255].

Statistical Analysis

In the CSSCD discovery cohort, we tested associations between 237,643 genotyped (0, 1, 2) or imputed (0.0-2.0) common SNPs (minor allele frequency (MAF) ≥ 1%) and phenotypes using Poisson regression (correction for over-dispersion) for painful crisis and ACS rates[225]. We implemented the analysis using custom scripts in the R 2.10.0 statistical package (www.r-project.org/). We used sex, age at baseline and the first ten principal components as covariates. Analyses were stratified based on α-thalassemia status and association results were combined by inverse variance meta-analyses[256]. Analyses of global and local ancestry were performed using, respectively, the EIGENSOFT and HAPMIX software with their default parameters[245,257,258].

Analyses in the CSSCD replication cohort were performed as for the CSSCD discovery cohort, except that we used α-thalassemia status as a covariate because of the small samples size of the cohort and we did not have access to principal components. In the Georgia Health Sciences University SCD cohort, age at baseline and painful crisis information were not available. We analyzed association between genotypes and ACS (dichotomous) using logistic regression in PLINK[219] and sex, year of birth and α-thalassemia as covariates. For the Duke SCD cohort, logistic regression was utilized to determine the effect of genotype on binary definitions of ACS and painful crisis using PLINK[219]. To further examine the impact of number of hospitalizations for painful crisis episodes, ordinal logistic regression was employed using SAS version 9.2 (SAS Systems, Cary, NC). In an attempt to reduce any population substructure that may exist, principal component analysis (PCA) was performed using EIGENSOFT[245]. All models were adjusted for sex, age and the first two principal components.

To analyze the effect of heme on the expression of *COMMD7* in pulmonary endothelial cells, we accessed the relevant gene expression dataset on the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) website (http://www. ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE25014)[259]. We performed the analyses separately for the pulmonary microvascular endothelial cells and the pulmonary artery endothelial cells using the GEO2R analytical module and default parameters (http://www.ncbi.nlm.nih.gov/geo/info/geo2r.html). We corrected P-values using the Benjamini & Hochberg false-discovery rate method.

3.4 Results

Several complications are observed in SCD and are linked to the quality-of-life and life expectancy of patients with this hemoglobinopathy[88,93,106,109,138,260]. The goal of our study was to identify genetic variants associated with these complications, in order to better understand clinical heterogeneity in SCD. We performed our discovery experiment in the CSSCD[236,261], a large longitudinal study with hundreds of clinical variables available. Despite this wealth of phenotypic data, we focus initially on only two quantitative measures, ACS and painful crisis rates, for two reasons. First, statistical power to find genetic associations with quantitative phenotypes was higher than for dichotomous traits (despite remaining relatively modest), even in the large CSSCD. For instance, we estimated that our study design had 50% power to find an association between a quantitative trait and a SNP under the following assumptions: minor allele frequency=25%, variance explained=1% and $\alpha=1x10^{-4}$ (**Table 1**). In comparison, we only had 8% power to find an association with stroke (prevalence=7%) under the same assumptions for a variant with an odds ratio=1.5. Second, because the size of our replication cohorts is small, for low prevalence complications, it is likely that there would be too few affected SCD patients to robustly validate genetic associations observed in the CSSCD discovery cohort.

We modeled and analyzed ACS and painful crisis rates as previously described using Poisson regression[88,106,225]. Whereas for ACS the distribution of the observed test statistics does not show major departure from the expected null distribution

(λ_{GC}=1.032), there was a slight inflation for the painful crisis association results (λ_{GC}=1.069)(**Figure 1**)[262]. For this reason, we corrected both ACS and painful crisis association results using the genomic control (GC) approach. But overall the limited inflation observed indicates that our analysis was appropriate and accounted for the main possible confounders. To declare statistical significance on the IBC array, we selected a threshold of α=2x10^{-6} that is sufficient to account for the number of independent tests performed[254]. Using this criterion, we identified a single locus that met array-wide significance: an association between rs11817401 in the *SORCS1* gene and painful crisis rate (*P*=1.2x10^{-7})(**Figure 1** and **Table 2**).

To confirm this association, and also identify additional loci that appeared promising but did not reach statistical significance in the CSSCD discovery cohort, we selected all SNPs with discovery *P*<1x10^{-4} (before GC correction; 19 SNPs for painful crisis and 17 SNPs for ACS) and genotyped them in 318 independent CSSCD participants. Replication and combined association results are presented in **Table 2** and **Table 3** for painful crisis and ACS, respectively. The association between *SORCS1*-rs11817401 and painful crisis did not replicate (replication *P*=0.52, opposite direction of effect). Overall, we replicated in this small CSSCD cohort two associations at nominal level (*P*<0.05): associations between *FAM193A*-rs11732673 and painful crisis (replication *P*=0.02, combined *P*=9.9x10^{-6}) and between rs6141803 and ACS (replication *P*=0.003, combined *P*=5.2x10^{-7}), the latter reaching array-wide significance when combining the CSSCD discovery and replication results (**Table 3**). *FAM193A* encodes a protein with no clear biological functions.

The rs6141803 SNP is located between the *COMMD7* and *DNMT3B* genes on chromosome 20. *DNMT3B* encodes a DNA methyltransferase important for development and *COMMD7*, a gene highly expressed in the lung, codes for an adaptor protein that interacts with subunits of the NF-κB complex[263]. Importantly, treating human pulmonary endothelial cells with free heme, a model that recapitulates some of the cellular responses observed when ACS is induced in a SCD mouse model[264], significantly modulates the expression of *COMMD7* (differential expression in pulmonary microvascular endothelial cells (P=5x10^{-4}) and in pulmonary artery endothelial cells (P=3x10^{-5}))[259]. This result, together with a role in NF-κB signaling and inflammation, add additional evidences supporting a role for *COMMD7* in ACS.

Our group had access to two additional SCD replication cohorts: 318 patients recruited at Georgia Health Sciences University (GHSU) and 472 SCD patients from Duke University (**Table 1**). Painful crisis information was not available for the GHSU cohort and only available as categories for the Duke cohort. ACS information was available in both cohorts, but only in the form of a binary presence/absence phenotype. In many situations, dichotomizing a quantitative trait can lead to substantial loss in statistical power as individuals with one or several ACS events are all labeled as affected[265]. For replication in the GSHU SCD cohort, we genotyped the top 17 SNPs associated with ACS (**Table 3**). A single variant, rs17728960 in the *NFATC2* gene, was nominally significant (P=0.05), but the combined association result was not significant. The association between ACS and *COMMD7*-rs6141803

was not significant (*P*=0.32) but trended in the right direction (odds ratio (OR)=0.41, **Table 3**). Genome-wide genotype data was available for the Duke SCD cohort. After quality-control steps and genotype imputation (**Materials and Methods**), six painful crisis and 14 ACS SNPs were available for association testing. The association between ACS and rs6141803 near *COMMD7* in the Duke cohort showed a consistent direction of effect (OR=0.16, *P*=0.07)(**Table 3**). When we combine at the P-value level results from the CSSCD discovery, CSSCD replication, GHSU and Duke cohorts using a Z-score method weighted based on sample size, the association between ACS and rs6141803 is array-wide significant (weighted $P=4.0x10^{-7}$).

We noted that *COMMD7*-rs6141803 is an ancestry informative marker: the C-allele has a frequency of 17% and 0% in the HapMap individuals of Northern European (CEU) and African (YRI) ancestry, respectively. This observation raises the possibility that the association between ACS and *COMMD7*-rs6141803 is a false positive result owing to admixture. However, this is unlikely because the ACS rate is not correlated with the first principal component, which captures European vs. African admixture (Spearman's ρ=-0.0188, *P*=0.47) and we used the first ten principal components in our analysis to account for global admixture. Although the association between ACS rate and genotypes at rs6141803 is not spurious because of admixture, we thought of using local ancestry to fine-map the causal variant. We inferred local European vs. African ancestry at the locus and used this estimate as a covariate in our regression model[257]. The strength of the genetic association

between ACS and rs6141803 was reduced when controlling for local ancestry (P=5.4x10^{-5} and P=0.001 without and with local ancestry as covariate), suggesting that rs6141803 is unlikely to be the causal variant.

Table 1 Description of the sickle cell disease cohorts

Phenotypes	CSSCD discovery (N=1514)	CSSCD replication (N=387)	Georgia Health Sciences University (N=318)	Duke (N=472)	Statistical power (0.5% and 1% variance explained)
Males / females	740 / 764	177 / 210	149 / 169	214 / 258	-
Age (year)	14.2 ± 11.9	11.2 ± 12.5	NA	33.6 ± 12.0	-
Follow-up (year)	6.6 ± 1.6	6.1 ± 2.2	NA	NA	-
α-thalassemia (%)	27.1	25.3	NA	NA	-
Acute chest syndrome (events/patient-year or affected / non-affected)	0.12 ± 0.24	0.16 ± 0.48	52 / 262	355 / 117	13%, 50%
Painful crisis (events/patient-year)	0.80 ± 1.45	0.65 ± 1.17	NA	NA	13%, 50%

Analyses in this study were restricted to sickle cell anemia (HbSS) or HbSβ0 patients, with our without α-thalassemia. Means ± standard deviations are provided. NA; not available. Statistical power for the CSSCD discovery cohort was calculated using the following assumptions: minor allele frequency = 25%; effect size = 0.5% or 1% of the phenotypic variance explained and $\alpha=1x10^{-4}$.

Table 2 Painful crisis association results

SNP	Chromosome (position)	EA/OA	EAF	CSSCD Discovery		CSSCD Replication		Combined CSSCD		Duke		Gene (annotation)
				Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value	OR [95% CI]	P-value	
rs12720497	1 (185097043)	T/C	0.10	0.529 (0.125)	$3.9x10^{-5}$	0.336 (0.194)	0.08	0.470 (0.107)	$1.2x10^{-5}$	NA	NA	*PLA2G4A* (intron)
rs540006	2 (70636930)	T/C	0.98	-0.723 (0.167)	$2.9x10^{-5}$	-0.310 (0.338)	0.36	-0.637 (0.154)	$3.5x10^{-5}$	0.91 [0.08, 10.25]	0.94	Intergenic
rs3917296	2 (102151265)	A/G	0.98	-0.773 (0.172)	$1.5x10^{-5}$	1.393 (1.013)	0.17	-0.708 (0.176)	$5.5x10^{-5}$	$1.2x10^{-9}$ [-inf, 0]	1.00	*IL1R1* (intron)
rs324035	3 (115351544)	A/C	0.75	-0.303 (0.076)	$1.2x10^{-4}$	0.061 (0.170)	0.72	-0.239 (0.071)	$8.3x10^{-4}$	NA	NA	*DRD3* (intron)
rs10513478	3 (156723359)	T/C	0.02	0.836 (0.209)	$1.1x10^{-4}$	-1.779 (2.450)	0.47	0.816 (0.216)	$1.5x10^{-4}$	0.57 [0.12, 2.64]	0.48	*PLCH1* (intron)
rs11732673	**4 (2652853)**	**A/G**	**0.77**	**-0.338 (0.089)**	**$2.2 x10^{-4}$**	**-0.374 (0.154)**	**0.02**	**-0.347 (0.079)**	**$9.9x10^{-6}$**	NA	NA	***FAM193A* (intron)**
rs6858735	4 (37592622)	T/C	0.12	0.375 (0.095)	$1.4x10^{-4}$	-0.165 (0.218)	0.45	0.283 (0.090)	$1.6x10^{-3}$	NA	NA	*TBC1D1* (intron)
rs13113915	4 (187702958)	A/G	0.40	0.261 (0.067)	$1.6x10^{-4}$	0.245 (0.131)	0.06	0.258 (0.061)	$2.5x10^{-5}$	NA	NA	*MTNR1A* (intron)
rs10942625	5 (90446448)	A/G	0.26	-0.362 (0.091)	$1.2x10^{-4}$	-0.329 (0.178)	0.07	-0.355 (0.083)	$2.0x10^{-5}$	0.93 [0.68, 1.26]	0.62	*GPR98* (intron)
rs1851426	7 (99220872)	A/G	0.68	-0.290 (0.074)	$1.6x10^{-4}$	0.269 (0.145)	0.07	-0.168 (0.068)	0.01	NA	NA	*CYP3A4* (intergenic)
rs10107231	8 (18098691)	T/C	0.84	-0.375 (0.096)	$1.7x10^{-4}$	-0.105 (0.203)	0.61	-0.322 (0.089)	$3.1x10^{-4}$	NA	NA	*NAT1* (intron)
rs7034457	9 (77807918)	A/G	0.91	-0.418 (0.102)	$7.1x10^{-5}$	0.045 (0.249)	0.86	-0.348 (0.097)	$3.3x10^{-4}$	NA	NA	*PCSK5* (intron)
rs7899453	10 (79469965)	A/C	0.22	0.317 (0.079)	$1.1x10^{-4}$	0.179 (0.149)	0.23	0.285 (0.072)	$7.3x10^{-5}$	NA	NA	*RPS24* (missense)
rs11817401	10 (108823893)	T/C	0.08	0.602 (0.110)	$1.2x10^{-7}$	-0.176 (0.273)	0.52	0.487 (0.105)	$3.4x10^{-6}$	NA	NA	*SORCS1* (intron)
rs17101814	11 (103367886)	T/C	0.09	0.444 (0.100)	$1.8x10^{-5}$	0.128 (0.196)	0.51	0.375 (0.091)	$4.1x10^{-5}$	0.78 [0.48, 1.28]	0.32	*PDGFD* (intron)
rs9933611	16 (52331386)	A/G	0.93	-0.449 (0.116)	$1.8x10^{-4}$	-0.245 (0.272)	0.37	-0.415 (0.110)	$1.5x10^{-4}$	NA	NA	*FTO* (intron)
rs445683	17 (71913117)	T/C	0.67	0.311 (0.079)	$1.5x10^{-4}$	0.089 (0.142)	0.53	0.256 (0.071)	$3.2x10^{-4}$	0.90 [0.68, 1.21]	0.50	*UBE2O* (intron)
rs7507634	19 (10297562)	T/C	0.02	0.899 (0.226)	$1.2x10^{-4}$	-0.821 (0.674)	0.22	0.714 (0.221)	$1.2x10^{-3}$	NA	NA	*RAVER1* (intron)
rs2872817	23 (153211612)	A/G	0.69	-0.249 (0.058)	$3.0x10^{-5}$	-0.158 (0.121)	0.19	-0.231 (0.053)	$1.5x10^{-5}$	NA	NA	*TKTL1* (3'UTR)

(Table 2 continued) We genotyped 19 SNPs with $P<1x10^{-4}$ in the CSSCD Discovery dataset (before genomic control (GC) correction) in an independent subset of the CSSCD. We combined results from the discovery and replication cohorts using an inverse variance meta-analysis approach. Only one SNP (rs11732673, in bold) replicated at nominal significance but no SNPs reached array-wide significance ($P<2x10^{-6}$) in the combined analysis. We also report results for the analysis of painful crisis as a categorical trait in the Duke cohort. P-values for the CSSCD Discovery cohort are GC-corrected. Genomic positions are on NCBI build 37.1. EA, effect allele ; OA, other allele; EAF, Effect allele frequency; NA, not available; SE, standard error; CI, confidence interval.

Table 3 Acute chest syndrome association results

SNP	Chromosome (position)	EA /OA	EAF	CSSCD Discovery		CSSCD Replication		Combined CSSCD		Georgia Health Sciences University		Duke		Gene (annot)
				Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value	OR [95% CI]	P-value	OR [95% CI]	P-value	
rs10399947	1 (149128584)	A/G	0.58	-0.319 (0.082)	$1.2x10^{-4}$	0.339 (0.156)	0.03	-0.174 (0.073)	0.02	0.98 [0.64, 1.50]	0.93	0.91 [0.67, 1.22]	0.51	Intergenic
rs34661029	2 (61002492)	C/G	0.98	-0.822 (0.191)	$2.3x10^{-5}$	-0.103 (0.450)	0.82	-0.709 (0.178)	$7.1x10^{-5}$	0.87 [0.24, 3.19]	0.83	0.76 [0.16, 3.64]	0.73	*REL* (missense)
rs17749316	2 (191548954)	C/G	0.99	-0.891 (0.228)	$1.2x10^{-4}$	0.518 (1.013)	0.61	-0.821 (0.226)	$2.7x10^{-4}$	$3.2x10^{8}$ [0, inf]	1.00	$1.9x10^{7}$ [0, inf]	1.00	*STAT1* (intron)
rs13021001	2 (128125889)	A/G	0.96	-0.631 (0.149)	$3.0x10^{-5}$	0.450 (0.482)	0.35	-0.534 (0.144)	$2.1x10^{-4}$	0.55 [0.14, 2.19]	0.39	0.85 [0.3, 2.38]	0.76	*GPR17* (3'UTR)
rs6778854	3 (563448)	A/G	0.61	0.374 (0.081)	$5.0x10^{-6}$	0.057 (0.161)	0.72	0.309 (0.073)	$2.4x10^{-5}$	0.98 [0.64, 1.50]	0.92	NA	NA	Intergenic
rs13315133	3 (38624908)	A/C	0.78	-0.339 (0.082)	$4.3x10^{-5}$	0.031 (0.197)	0.88	-0.284 (0.076)	$2.0x10^{-4}$	0.90 [0.53, 1.54]	0.71	NA	NA	*SCN5A* (intron)
rs3910551	3 (106659325)	C/G	0.02	0.793 (0.180)	$1.4x10^{-5}$	-0.333 (0.532)	0.53	0.675 (0.172)	$9.1x10^{-5}$	1.13 [0.38, 3.32]	0.83	0.24 [0.07, 0.79]	0.02	*ALCAM* (intron)
rs5030094	3 (187943852)	T/C	0.99	-0.801 (0.221)	$3.5x10^{-4}$	-0.727 (0.424)	0.09	-0.785 (0.198)	$7.5x10^{-5}$	0.54 [0.14, 2,15]	0.38	0.73 [0.11, 4.7]	0.74	*KNG1* (intron)
rs10478813	5 (127859164)	C/G	0.16	0.400 (0.109)	$2.8x10^{-4}$	-0.002 (0.240)	0.99	0.330 (0.100)	$9.9x10^{-4}$	0.81 [0.38, 1.75]	0.60	NA	NA	*FBN2* (intron)
rs5576	7 (24297876)	A/G	0.97	-0.655 (0.152)	$2.1x10^{-5}$	0.078 (0.477)	0.87	-0.586 (0.146)	$6.3x10^{-5}$	1.66 [0.37, 7.55]	0.51	1.07 [0.37, 3.07]	0.90	*NPY* (3'UTR)
rs1176758	11 (113275565)	A/G	0.92	-0.528 (0.117)	$9.7x10^{-6}$	0.110 (0.275)	0.69	-0.427 (0.109)	$9.7x10^{-5}$	1.61 [0.49, 5.33]	0.43	1.50 [0.74, 3.05]	0.26	Intergenic
rs6309	13 (46368601)	A/G	0.04	0.631 (0.159)	$9.0x10^{-5}$	-0.391 (0.477)	0.41	0.526 (0.153)	$5.6x10^{-4}$	1.43 [0.45, 4.56]	0.55	0.74 [0.38, 1.45]	0.38	*HTR2A* (intron)
rs9927848	16 (23740572)	A/C	0.44	-0.354 (0.099)	$4.3x10^{-4}$	0.317 (0.147)	0.03	-0.141 (0.083)	0.09	1.14 [0.74, 1.75]	0.55	1.0 [0.74, 1.35]	0.99	Intergenic
rs12447481	16 (52306470)	T/G	0.98	-0.794 (0.206)	$1.5x10^{-4}$	-0.830 (0.450)	0.07	-0.800 (0.190)	$2.5x10^{-5}$	$3.3x10^{8}$ [0, inf]	1	1.74 [0.31, 9.66]	0.52	*FTO* (intron)
rs6141803	**20 (30804017)**	**T/C**	**0.98**	**-0.701 (0.171)**	**$5.4x10^{-5}$**	**-0.797 (0.266)**	**0.003**	**-0.730 (0.145)**	**$5.2x10^{-7}$**	**0.41 [0.07, 2.34]**	**0.32**	**0.16 [0.02, 1.20]**	**0.07**	**Intergenic**
rs17728960	20 (49563123)	T/C	0.98	-0.789 (0.208)	$1.9x10^{-4}$	-0.053 (0.354)	0.88	-0.595 (0.182)	$1.1x10^{-3}$	0.35 [0.12, 0.99]	0.05	0.26 [0.03, 2.05]	0.20	*NFATC2* (intron)
rs16998437	22 (36418341)	T/G	0.92	-0.483 (0.122)	$9.3x10^{-5}$	-0.061 (0.233)	0.79	0.390 (0.109)	$3.5x10^{-4}$	NA	NA	1.04 [0.51, 2.13]	0.91	*NOL12* (3'UTR)

(Table 3 continued) We genotyped 17 SNPs with $P<1x10^{-4}$ in the CSSCD Discovery dataset (before genomic control (GC) correction) in an independent subset of the CSSCD. We combined results from the discovery and replication CSSCD cohorts using an inverse variance meta-analysis approach. Only one SNP (rs6141803, in bold) replicated at nominal significance and reached array-wide significance ($P<2x10^{-6}$) in the combined analysis. We also report results for the analysis of acute chest syndrome as a dichotomous trait in the Georgia Health Sciences University and Duke cohorts. For rs6141803, the association was not significant but trended in the right direction. P-values for the CSSCD Discovery cohort are GC-corrected. Genomic positions are on NCBI build 37.1. EA, effect allele ; OA, other allele; EAF, Effect allele frequency; SE, standard error; NA, not available; CI, confidence interval; annot: annotation.

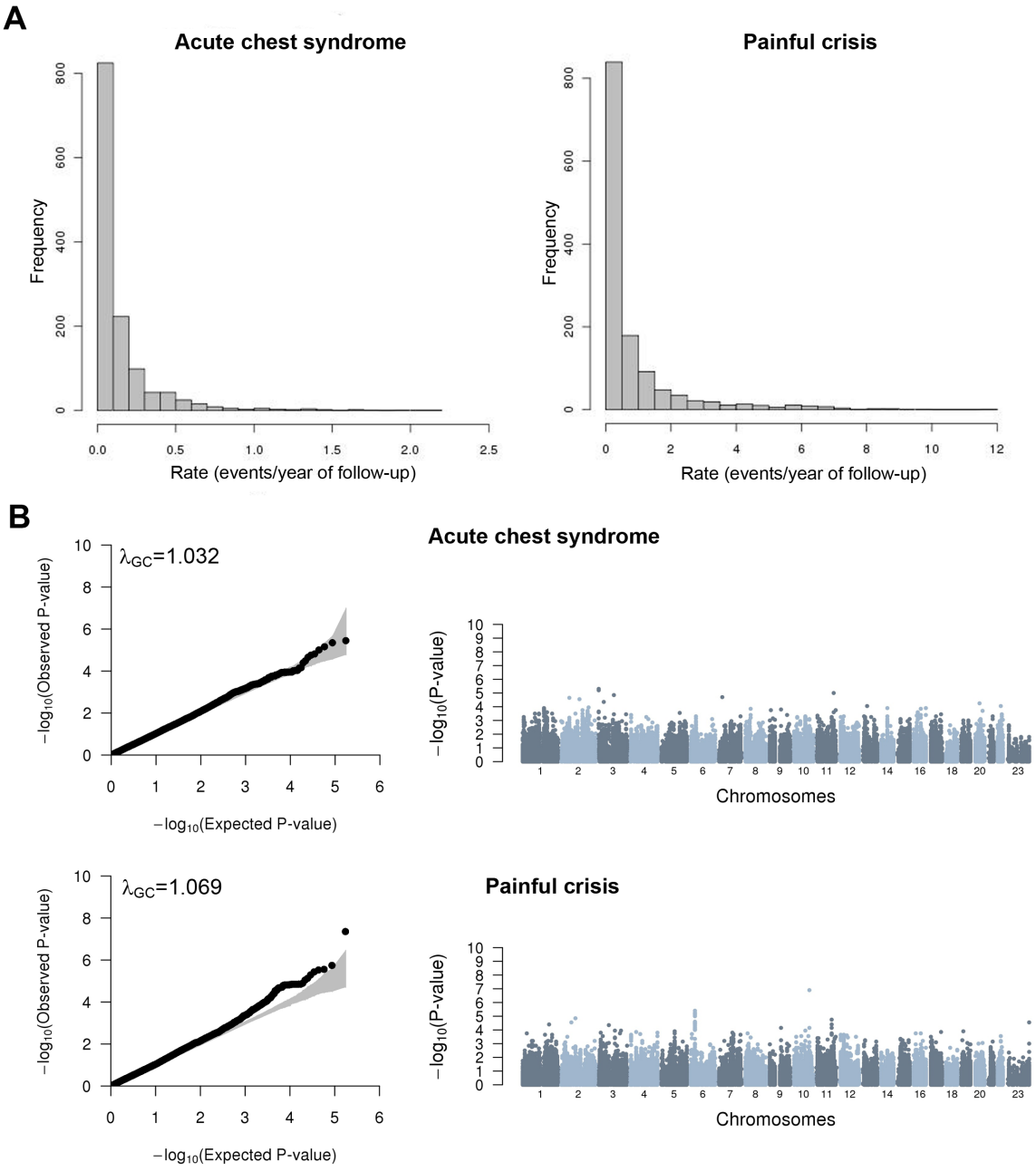


Figure 1. Association results for acute chest syndrome and painful crisis.

(A) Histograms showing the distribution of the acute chest syndrome (left) and painful crisis (right) rates in the Cooperative Study of Sickle Cell Disease (CSSCD) participants analyzed in this study (N=1,514). (B) Quantile-quantile (QQ) (left column) and Manhattan (right column) plots of association results for acute chest syndrome (top) and painful crisis (bottom) analyzed in participants from the CSSCD. In the QQ plots, the grey area corresponds to the 90% confidence interval. The dashed line in the Manhattan plots corresponds to the statistical threshold used ($P<1x10^{-4}$) to select SNPs for replication genotyping. Results are corrected using the genomic control method.

3.5 Discussion

Painful crisis and ACS are respectively the first and second most frequent causes of hospital admissions in patients with SCD[29]. Although they share common causes (*e.g.* vaso-occlusion), it is also clear that some of the triggering factors are different (*e.g.* the role of infections in ACS). We performed here one of the largest genetic association experiments carried out to date to identify DNA sequence variants that modify SCD clinical severity through these two measures of morbidity.

Our experimental design identified a single SNP, rs6141803, that reached array-wide significance (P=5.2x10^{-7} in the CSSCD discovery+replication, P=4.0x10^{-7} if we add the GHSU and Duke SCD samples). Analyzing ACS as a dichotomous phenotype in the GHSU and Duke cohorts, as we did in this study, could account for the loss of statistical power and the non-replication observed. It is also possible that this association is a false positive report. Despite allele frequency differences between the ancestral populations at rs6141803, we performed analyses that suggest that admixture is unlikely to confound this result. Additional replication attempts in large SCD cohorts with quantitative measures of ACS are needed before drawing a final conclusion on the ACS-rs6141803 genetic association.

rs6141803 is an intergenic SNP located between *DNMT3B* and *COMMD7*. *DNMT3B* encodes for a DNA methyltransferase involved in maintenance DNA methylation. Mutations in *DNMT3B* cause immunodeficiency-centromeric instability-facial anomalies syndrome-1 (ICF-1; MIM #242860), a very rare syndrome that has not been linked to

lung-related complications. *COMMD7* encodes a poorly known protein that contains a copper metabolism gene *MURR1* (COMM) domain. The gene is abundantly expressed in the lung[263] and is overexpressed in hepatocellular carcinoma[266]. The knockdown of *COMMD7* using short-hairpin RNA (shRNA) increases apoptosis and cell cycle arrest, in part by interfering with NF-κB signaling[263,266]. NF-κB is a master regulator of acute inflammation; upon stimulation, it transcriptionally activates interleukins, interferon, tumor necrosis factor-α, and adhesion molecules. The expression of *COMMD7* in pulmonary endothelial cells is also affected upon heme treatment[259]; this is promising as free heme can induce ACS in a SCD mouse model[264]. Although more work is needed to clarify the role of *COMMD7* in ACS, these are potentially interesting observations given the importance of inflammation and free radical production in aggravating ACS episodes[267].

A recent large candidate gene study in 942 SCD children identified an association between a microsatellite in the heme oxygenase-1 (*HMOX1*) promoter and ACS: longer alleles were associated with increased rate of hospitalization for ACS[268]. Results from this initial study were not replicated in an independent cohort. Therefore, we queried our own results to test if *HMOX1* SNPs were associated with ACS rate in the CSSCD. Although the *HMOX1* gene was targeted for genotyping on the IBC array (28 genotyped or imputed nearby SNPs), the microsatellite was not directly tested. Of the *HMOX1* SNPs that are accessible on the IBC array, one SNP in the 3' untranslated region of *HMOX1* is associated with ACS at nominal significance (rs12160039, P=0.02). We would need to

directly genotype the promoter microsatellite to determine if this SNP captures the association signal with ACS through linkage disequilibrium.

The second most interesting association identified in our experiment is between painful crisis rate and rs12720497, an intronic SNP in the *PLA2G4A* gene. The association is not array-wide significant but the directions of effect are consistent between the discovery and replication CSSCD panels (replication P=0.08, combined P=1.2x10^{-5}, **Table 2**). *PLA2G4A* encodes a cytosolic phospholipase A2, an enzyme implicated in the production of pro-inflammatory molecules (prostaglandins and leukotrienes) that has previously been implicated in increased sensitivity to pain (hyperalgesia) in humans[269,270]. Enzymes in the phospholipase A2 family can be divided into four groups (cytosolic, secreted, calcium-independent and lipoprotein-associated) and high levels of secreted phospholipase A2 have been suggested to be predictive of future ACS events[271-273]. The link between cytosolic and secreted phospholipase A2 and their role in SCD complications is intriguing, especially because severe ACS often occurs in the course of vaso-occlusive painful crisis. In our data, however, rs12720497 in *PLA2G4A* (coding for cytosolic phospholipase A2) is not associated with ACS rate (P=0.67).

We performed our discovery search in 1,514 participants from the CSSCD whose DNA was genotyped on a gene-centric genotyping arrays[242]. With the caveat that this genotyping platform only captures genetic variation at a subset (~10%) of the predicted human genes, we did not identify loci with moderate-to-strong effect on phenotype, consistent with most reported GWAS results[274]. Our results highlight promising variants

for further replication in independent SCD cohorts and biologically plausible candidate genes (*e.g. COMMD7*, *PLA2G4A*) to test functionally, for instance in SCD mouse models. They are also indicative of the importance of combining GWAS results through meta-analyses between SCD cohorts to gain sufficient statistical power to identify genetic associations of weak phenotypic effect.

3.6 Acknowledgments

The authors acknowledge the contribution of Mélissa Beaudoin for DNA genotyping and Cameron D. Palmer for genotype imputation and thank all the patients who contributed to this study.

CSSCD is supported in part by the National Institutes of Health, National Heart, Lung, and Blood Institute (N01-HB-47110). CARe is supported by the National Heart, Lung, and Blood Institute (HHSN268200625226C). A full listing of the grants and contracts that have supported CARe is provided at http://www.nhlbi.nih.gov/resources/geneticsgenomics/programs/care.htm. The work in the Lettre Laboratory is supported by a Innovation in Clinical Research Award grant from the Doris Duke Charitable Foundation (2009089), the Canada Research Chair Program, the Canadian Institute of Health Research (123382), and the Fonds de Recherche Santé Québec. The work at Duke University was funded in part by the National Heart, Lung, and Blood Institute (RO1 HL079915 and RC2-HL101212).

3.7 Annex

We also performed genetic association studies with logistic regression on four dichotomous traits: osteonecrosis, leg ulcers, stroke and priapism. The association tests were performed with the software PLINK[219] with sex, age at baseline and the first ten principal components as covariates, except for priapism for which only age at baseline and the first ten principal components were used as covariates. Analyses were stratified based on α-thalassemia status and association results were combined by inverse variance meta-analyses[256]. We attempted to replicate in the CSSCD replication cohort signals with $P<1x10^{-5}$ (**Table 4**). None of the SNPs tested replicated.

Table 4 SNPs tested for replication – Dichotomous traits

Phenotype	SNP id	Chr	Bp	Gene	Geno/Impu	EA/OA	EAF	Discovery odds ratio [95% CI]	P-value	CSSCD Repl odds ratio [95% CI]	CSSCD Repl p-value
Osteonecrosis	rs11729	19	19091482	*TMEM161A*	GENO	T / C	0.10	3.52 [2.60, 4.77]	$3.4x10^{-5}$	0.31 [0.12, 0.81]	0.1132
Osteonecrosis	rs6694329	1	103051552	intergenic	IMPU	C / G	0.02	26.66 [13.47, 52.78]	$1.5x10^{-6}$	$1.1 x10^{6}$ [0, inf]	0.9973
Stroke	rs7668014	4	155678178	*PLRG1*	IMPU	T / C	0.10	4.75 [3.05, 7.39]	$4.2x10^{-4}$	$0.04 [7.3x10^{-3}$, 0.27]	0.0834
Stroke	rs8083727	18	59109560	*BCL2*	GENO	A / G	0.04	9.03 [5.22, 15.63]	$6.1x10^{-5}$	0.13 [0.03, 0.56]	0.1574
Priapism	rs10925494	1	235918764	*RYR2*	GENO	T / G	0.16	3.78 [2.72, 5.25]	$5.4x10^{-5}$	1.51 [0.57, 4.03]	0.6745
Priapism	rs13216018	6	155850404	intergenic	IMPU	A / C	0.96	0.04 [0.02, 0.08]	$1.2x10^{-6}$	0.29 [0.04, 1.84]	0.5017
Priapism	rs11203047	10	90991106	*LIPA*	GENO	T / C	0.94	0.14 [0.09, 0.22]	$2.6x10^{-6}$	0.64 [0.17, 2.47]	0.7413
Leg ulcer	rs16961269	15	46711316	*FBN1*	GENO	T / C	0.09	4.41 [3.16, 6.14]	$7.7x10^{-6}$	0.44 [0.17, 1.15]	0.3945

We genotyped eight SNPs with $P<1x10^{-5}$ in the CSSCD Discovery dataset in an independent subset of the CSSCD. We combined results from the discovery and replication CSSCD cohorts using an inverse variance meta-analysis approach. No SNP replicated at nominal significance. Genomic positions are on NCBI build 37.1.

Chr: Chromosome, Bp: base pairs, Geno: Genotyped SNP, Impu: Imputed SNP, EA: Effect allele, OA: Other allele, EAF: Effect allele frequency, CI: Confidence intervals, Repl: Replication.

Chapter 4: Genetic Association Study Based on Gene-Set Enrichment Analysis Identified Biological Pathways Associated with Fetal Hemoglobin Levels in Sickle Cell Disease Patients

Authors

Geneviève Galarneau, Guillaume Lettre.

Reference

Galarneau G, Lettre G. Genetic Association Study Based on Gene-Set Enrichment Analysis Identified Biological Pathways Associated with Fetal Hemoglobin Levels in Sickle Cell Disease Patients. 2014. In preparation

Author's Contribution

Mélissa Beaudoin performed the genotyping, GG and GL conceived and designed the study, GG performed the data analysis and wrote the manuscript.

4.1 Abstract

High levels of fetal hemoglobin (HbF) diminish severity and mortality in sickle cell disease (SCD) patients. In this study, we performed a genetic association study for HbF levels in 1,213 SCD patients from the Cooperative Study of Sickle Cell Disease (CSSCD) genotyped on the ITMAT-Broad-CARe (IBC) array. We first performed a single variant association study. No novel loci reached array-wide significance. We attempted to replicate five single nucleotide polymorphisms (SNPs) in 310 independent CSSCD samples, but none of the variants tested were significant in the replication dataset. Since the variant rs7325795, located in *GPC6*, was near nominal significance (P=0.07), we tried additional replication for this variant in three other cohorts. In the combined analysis including all five cohorts, the p-value for rs7325795 was $P=9.4x10^{-6}$. Since no novel loci reached array-wide significance, we performed a gene-set enrichment analysis in the CSSCD discovery dataset using GenGen software[275,276]. Twelve pathways were nominally significant (P<0.05) and seven of them replicated in an independent set of 276 CSSCD samples. These seven pathways were highly redundant and might represent a single biological process. Our results suggest a role for acetyl-coA in short fatty acid metabolism as well as amino acid degradation in the variation of HbF levels. One of the significant pathways was butyrate metabolism. Butyrate is known to increase HbF levels and has been tested in many clinical trials as a treatment for SCD patients. There was no correlation between HbF and butyrate levels, but we observed an inverse correlation (P=0.01) between HbF levels and β-hydroxybutyrate (BOHB) in 156 SCD patients. This correlation is the opposite of what was expected, and it is unclear if it is a false positive or if endogenous BOHB could reduce HbF levels by increasing histone deacetylase (HDAC) expression.

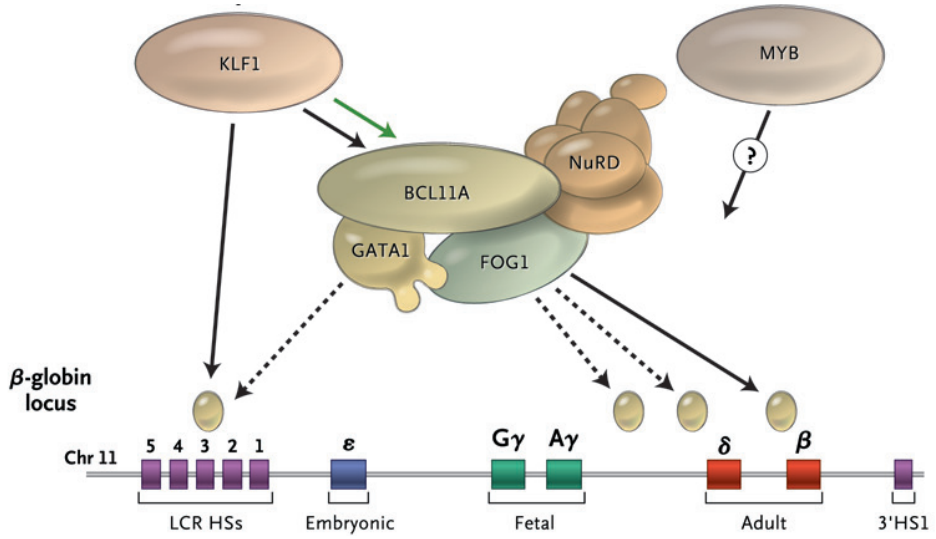
4.2 Introduction

Sickle cell disease (SCD) is a monogenic disease caused by a mutation in the β-globin gene. SCD complications can be life-threatening and include stroke, acute chest syndrome and renal failure. Studies have shown that SCD patients with high fetal hemoglobin (HbF) levels show less morbidity and have a better life expectancy[120,123,124]. HbF is a type of hemoglobin present mainly during fetal development but that remains at basal levels during adulthood. HbF reactivation is the most promising therapeutic strategy for the treatment of SCD patients. Three classes of agents are known to increase HbF levels: chemotherapeutic agents (hydroxyurea, 5-azacytidine and decitabine), short-chain fatty acid derivatives (sodium phenylbutyrate, arginine butyrate and isobutyrate) and erythropoietin. Currently, the only approved treatment for SCD patients is hydroxyurea.

The switch from fetal to adult hemoglobin is a transition from a main form of hemoglobin composed of 2 α-globin sub-units and 2 γ-globin sub-units to a hemoglobin structure composed of 2 α-globin sub-units and 2 β-globin sub-units. Both γ-globin and β-globin genes are located in the β-globin cluster (**Figure 1**). The human β-globin cluster is ~100 kb long and is composed of multiple globin genes placed in the order of their expression during development: -embryonic ε–fetal ^{G}γ-^{A}γ–adult δ-β. The switch from fetal to adult hemoglobin is controlled by transcription factors and chromatin remodeling of the β-globin cluster. (**Figure 1**) HbF level is a highly heritable trait ($h2 \sim 0.6–0.9$)[235,277]. Three loci have previously been associated with HbF levels in GWAS: *BCL11A*, *HBS1L-MYB* and the β-globin locus[196,204,205], and altogether, these three loci explain at least 50% of the heritable variation of HbF levels[196,204,205,278]. The locus of *KLF1* has also been associated with a high persistence

of fetal hemoglobin in linkage studies, though no common variants in this locus have been associated with HbF levels. A better understanding of the genes and mechanisms involved in HbF regulation may lead to novel methods of increasing HbF levels in SCD patients.

A



B

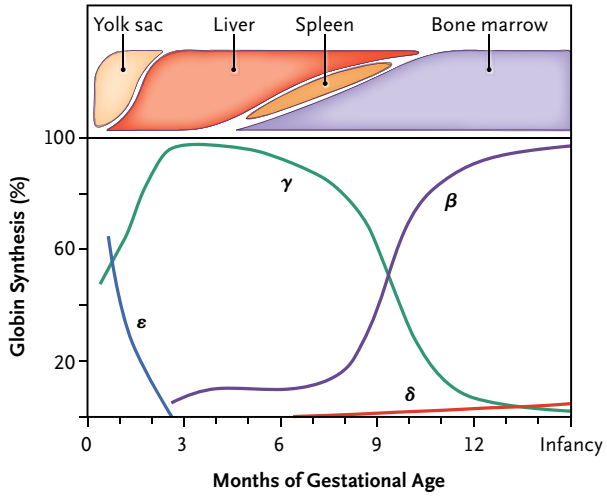


Figure 1 The β-like globin gene locus and expression

(A) The β-globin gene cluster on chromosome 11 contains the hemoglobin genes encoding the ε-subunit, the Gγ-subunit, the Aγ-subunit, the δ-subunit and the β-subunit. BCL11A forms a complex with its protein partners, including GATA1, FOG1 and NuRD, and bind to sequences within the globin locus and repress the expression of the γ-globin genes. KLF1 positively regulates the expression of BCL11A and binds the globin locus to promote the transcription of the adult β-globin gene. (B) Fetal hemoglobin is composed of two α-subunits and two γ-subunits ($\alpha_2\gamma_2$). The main form of adult hemoglobin is composed of two α-subunits and two β-subunits ($\alpha_2\beta_2$). The γ-globin is expressed in the liver and the spleen during the gestation and its expression starts decreasing shortly before birth. The β-globin is expressed in the bone marrow and slowly increases during the end of gestation. The hemoglobin switch occurs when β-globin becomes more expressed than γ-globin. LCR: locus control region; HSs: DNase I hypersensitive sites[279].

Metabolites are small molecules produced at different steps of the metabolic processes to provide energy to the organism. The activity of metabolic processes can be modified by environmental factors such as a diet change or a sudden demand in energy. The regulation of these metabolic reactions to maintain homeostasis is very complex. Metabolites can act as metabolic regulators by affecting gene expression. This is the case for butyrate and β-hydroxybutyrate, two known histone deacetylase (HDAC) inhibitors affecting HbF levels[280,281]. HDACs can induce epigenetic modifications by removing an acetyl group on lysine, which compacts the chromatin. By inhibiting histone deacetylation, these two metabolites reduce the compression of the DNA, thereby facilitating the access to transcription factors and increasing gene expression. Metabolomics studies characterize and quantify small molecule metabolites found in an organism.

Pathway-based approaches in genomic association studies test if multiple genes involved in the same biological processes are implicated in the complex trait under investigation. A pathway is considered to be a set of genes that interact together in a biological process. The definition of the genes included in a pathway can be based on different methods, such as molecular interaction or literature text mining. There are two main types of pathway analysis: over-representation analysis and gene-set enrichment analysis. In both types, the SNPs are first assigned to their closest genes. Over-representation analysis focuses on signals above a selected threshold to test whether genes involved in a pathway are over-represented. Gene-set enrichment analyses aim to identify sets of moderately associated genes involved in a same biological function, usually by assigning a score to each gene. This

gene-score can be based on the overall best p-value in SNPs assigned to that gene, or on the mean of p-values for that gene.

GenGen[275,276] is a gene-set enrichment analysis tool for genomics datasets. Each gene is assigned a score based on the best individual SNP association. To reflect the over representation of genes from a given pathway in higher-ranked genes, it calculates a weighted Kologorov-Smirnov-like running sum statistic. The same calculations are performed in the dataset but with permutated phenotypes. For each pathway, the sum statistic obtained in the real dataset is compared to the ones obtained with permutated phenotypes to calculate a normalized enrichment statistic[275,276].

We previously performed a gene-centric association study on the ITMAT-Broad-CARe (IBC) array[242], an array that covers ~50,000 single nucleotide polymorphisms (SNPs) targeting genes related to heart, lung and blood diseases. In this study, we tested the association with two clinical complications of SCD: pain crisis and acute chest syndrome[282]. These analyses were performed in a prospective cohort of 1,514 African-Americans with SCD from the CSSCD. In the present study, we performed an association study with HbF levels in 1,213 individuals from the same dataset. No SNP was significantly associated in the replication cohort, but the variant rs7325795, located in *GPC6*, was close to significance (P=0.07) with a same direction of effect. In the combined analysis including the discovery cohort and four replication cohorts, the p-value for rs7325795 was $P=9.4x10^{-6}$. Our gene-set enrichment analysis results suggest a role for acetyl-coA and short-chain fatty acid metabolism as well as amino acid degradation in HbF level variation. Butyrate metabolism was one of the

associated pathways that replicated. We observed an inverse correlation between HbF levels and β-hydroxybutyrate (BOHB) levels (P = 0.01) in 156 SCD patients from the CSSCD cohort. These results suggest that endogenous levels of metabolites and genes involved in their metabolism could influence HbF levels.

4.3 Methods

Ethics Statement

Informed consent was obtained for all participants in accordance with the Declaration of Helsinki. The Candidate-gene Association Resource (CARe) Study was approved by the ethics committees of the participating studies, and the Massachusetts Institute of Technology. This project was also reviewed and approved by the Montreal Heart Institute Ethics Committee.

Samples and Genotyping

The CSSCD is a prospective SCD cohort[236] in which detailed phenotypic information on clinical complications such as stroke, pain crisis, osteonecrosis and priapism has been collected. HbF levels were also measured in participating patients. As previously described[282], the samples were genotyped on the Illumina IBC array[242] at Broad Institute and the software MACH 1.0.16[222] was used for genotype imputation. This array covers ~2,100 genes relevant for heart, lung and blood diseases. Both genotyped and imputed SNPs with minor allele frequency (MAF) >0.01 were included in the analysis[282]. There were 1,213 genotyped samples with non-missing HbF measurements in the discovery cohort after quality control, as well as 310 independent SCD samples in the CSSCD replication cohort (**Table 1**). HbF levels were normalized and corrected for age, sex and hemoglobinopathy for samples in both the CSSCD discovery and CSSCD replication datasets. We used the Sequenom MassArray iPLEX platform to genotype the variants selected for replication.

Table 1 SCD cohorts description

	CSSCD discovery	CSSCD replication
Non-missing HbF levels (N)	1,213	310
Males/females	699/725	179/146
Age, y	15.9 ± 11.7	12.6±12.5
Statistical power (0.5% and 1% variance explained)	8%, 34%	24%, 42%

Means ± standard deviations are provided. Statistical power for the CSSCD discovery cohort was calculated using the following assumptions: minor allele frequency = 25%, effect size = 0.5% or 1% of the phenotypic variance explained, and $\alpha = 1 \times 10^{-4}$. Statistical power for the CSSCD replication cohort was calculated using the following assumptions: minor allele frequency = 25%, effect size = 0.5% or 1% of the phenotypic variance explained, and $\alpha = 0.05$.

Single-Variant Association Study

We first attempted to test the genetic association for common variants with HbF levels on the IBC array[242]. In the CSSCD discovery cohort, we performed a linear regression using PLINK v1.07[219] with the normalized HbF levels, and included the first ten principal components as covariates. We performed validation in 310 independent CSSCD samples for SNPs that showed the strongest association ($P<1x10^{-4}$). When building our genotyping pools, we prioritized signals with genotyped variants. We removed variants with genotyping rate <95%. In the CSSCD replication dataset, we performed a linear regression with normalized HbF levels.

Association Tests with rs7325795 in Other Cohorts

We tried to genotype the variant rs7325795 for additional validation in four other cohorts: the Adult Sickle Cell Clinic of Georgia Health Sciences University (GHSU) Sickle Cell Center,

the Jamaica Sickle Cell Cohort Study (JSCCS), the Multicenter Study of Hydroxyurea in Sickle Cell Anemia (MSH) and the Silent cerebral Infarct Transfusion Trial (SITT). The genotyping in these replication cohorts was performed with the Sequenom MassArray iPLEX platform. Genotyping was performed in our lab for the GHSU, the JSCCS and the MSH cohorts and at the John Hopkins School of Medicine for the SITT samples. HbF levels were normalized in all cohorts and were corrected for sex, year of birth and hemoglobinopathy in the GHSU; for sex in the JSCCS; and for age, sex and hemoglobinopathy in the MSH. The HbF levels were also normalized in the JSCCS and in the MSH. We performed a linear regression with PLINK v1.07[219] to test the association with HbF levels in the three additional cohorts.

Gene-Set Enrichment Analysis

With PLINK v1.07[219], we generated 1000 sets of permutated normalized HbF levels within the CSSCD discovery dataset. The discovery phase of the gene-set enrichment analysis included genotyped SNPs with an MAF>0.01. We then ran the 1000 linear regression with the permutated phenotypes using the first 10 principal components as covariates, as in the analysis with the real data. With an in-house annotation script created by Ken Sin Lo, SNPs were assigned to the gene in which they were located or to the nearest gene for intergenic SNPs. We used the KEGG[283] database to define the gene-sets. We used GenGen software[275,276] to calculate gene-set enrichment statistics. In this type of analysis, GenGen assigns a score based on the best individual SNP association to each gene. We used GenGen default maximum distance between a SNP and its nearest gene (500 kb). A total of 171 gene-sets from the KEGG[283] database were tested with GenGen[275,276].

The twelve gene-sets with P<0.05 in the discovery analysis were tested for replication in two cohorts. In the first replication cohort, we used 276 independent CSSCD samples that were genotyped on the Illumina-610 Quad array. Our analyses in the replication cohort were limited to SNPs included in the discovery phase (genotyped on the IBC array with an MAF > 0.01) and genotyped or imputed in the replication dataset. Association tests were performed with Mach2Qtl V1.1.0[222] for genotyped and imputed SNPs with r^2>=0.6.

Association Tests between Metabolites and HbF levels

Butyrate and BOHB levels were measured in 156 serum samples from the CSSCD at the Broad Institute using liquid chromatography-mass spectrometry (LC-MS) analyses. We tested the correlation between HbF levels and these two metabolites with the R 2.10.0 statistical package (www.r-project.org). For each metabolite, we also compared two HbF level models: the null model included the SNPs previously identified as independently associated within the loci associated with HbF levels (*BCL11A*, *HBS1L-MYB* and the β-globin locus): rs4671393, rs7599488, rs10189857, rs9402686, ss244317976, rs28384513 and rs10128556[278]; the second model also included either butyrate or BOHB levels. We performed an ANOVA to compare the two models.

Replication of Most Significant SNPs in Significant Gene-Sets

We attempted to genotype in the CSSCD replication and GHSU cohorts some of the SNPs in the associated pathways. We tested the association with HbF levels and three SNPs (rs9347340, rs11594057 and rs3465) in these two datasets using PLINK v1.07[219]. There

were 88 CSSCD samples genotyped on the IBC array for which we had the metabolite measurements. We tested in these 88 samples the association with BOHB and butyrate, and the 8 SNPs highlighted by the pathway analysis. For each metabolite tested, we performed a linear regression with sex and age as covariates using PLINK v1.07[219].

Power Calculations

Power calculations were performed with the software Quanto[284] under the following assumptions: additive model and $\alpha=1\times10^{-4}$ for the CSSCD discovery dataset and $\alpha=0.05$ for replication.

4.4 Results

HbF level is a known modifier of SCD severity. In the absence of an applicable treatment for the majority of SCD patients, increasing HbF levels can reduce the morbidity of SCD. In this study, we aimed to identify novel loci associated with HbF levels. We first performed a single-variant association study with HbF levels. To verify if there could be an enrichment of SNPs with moderate effects, we attempted a gene-set enrichment. Finally, we measured butyrate and BOHB levels in SCD serum samples and verified the correlation with HbF levels.

Single-Variant Association Study

We tested the single-variant genetic association with HbF levels on the IBC array[242] in 1,213 individuals. A total of 175,020 genotyped or imputed SNPs were tested. The test statistics distribution does not show major inflation (**Figure 2**). Three loci have already been associated with HbF levels in GWAS: *BCL11A*, *HBS1L-MYB* and the β-globin locus[196,204,205,225]. SNPs in both *BCL11A* and β-globin reached significance (**Table 2**, **Figure 3**). The genotyping array did not cover the *HBS1L-MYB* intergenic region. No SNPs in novel regions reached the array-wide significance level of $\alpha=2x10^{-6}$ [254] (**Table 2**). We attempted to validate variants with $P<1x10^{-4}$ in an independent set of 310 CSSCD samples, prioritizing signals with genotyped markers. We tried to genotype eight SNPs, but three of them failed genotyping or were filtered after quality control. None of the SNPs were significant ($P<0.05$) in the replication set (**Table 3**). The association between the variant rs7325795 and HbF levels

was close to significance (P=0.07), with an effect consistent with the direction observed in the discovery set.

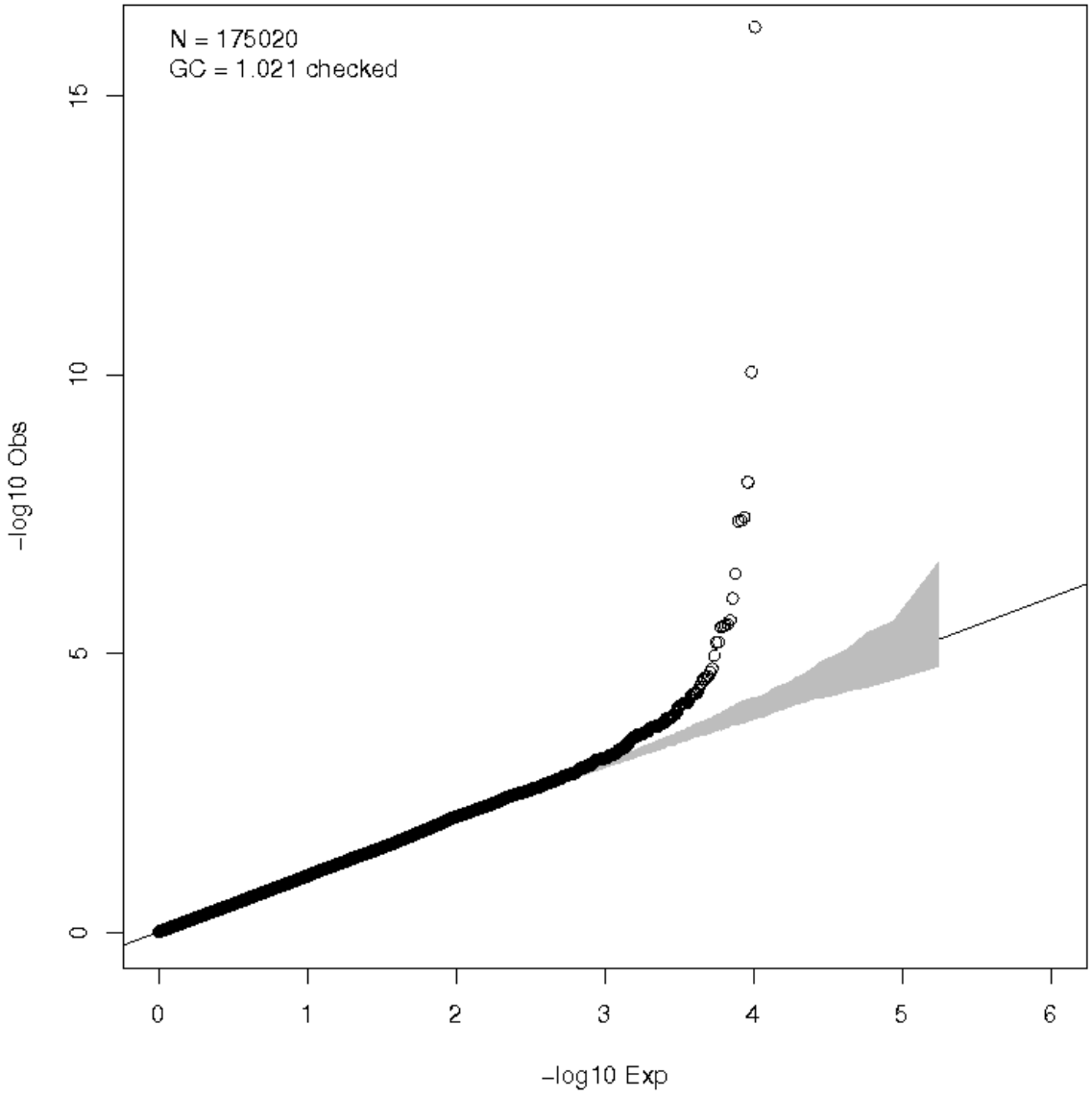


Figure 2 HbF single-marker association results QQ-plot

Quantile-quantile plot for the linear regression in the single-marker discovery dataset. The gray area corresponds to the 90% CI.

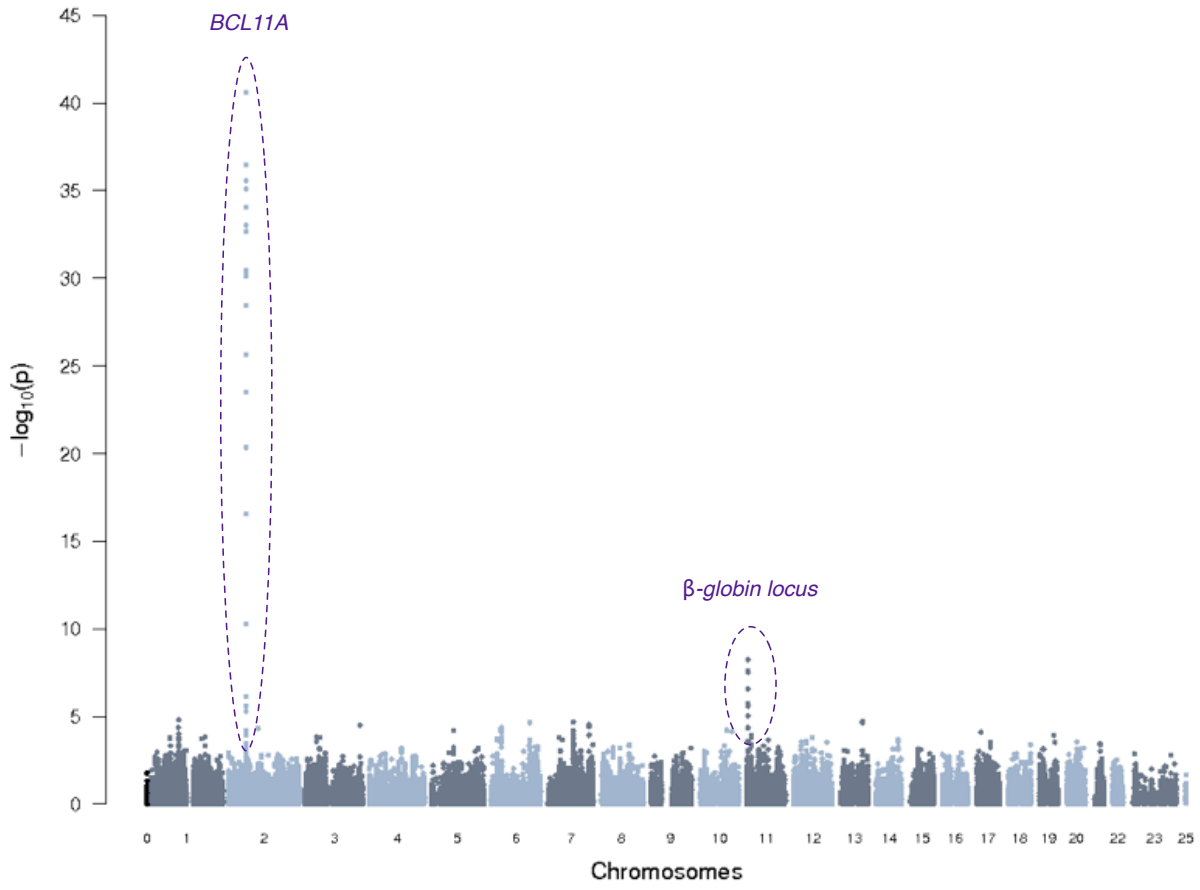


Figure 3 HbF single-marker association results Manhattan plot

Manhattan plot of results for HbF single-marker association test in the CSSCD discovery dataset.

Table 2 HbF single variant association results ($P<1x10^{-4}$)

SNP	Geno/ Impu	Rsq	Chr	BP	EA	OA	EA Frq	b	SE	P-value
rs1427407	GENO	-	2	60571547	T	G	0.24	0.602	0.043	$2.7x10^{-41}$
rs766432	IMPU	0.97	2	60573474	A	C	0.73	-0.567	0.043	$3.5 x10^{-37}$
rs7606173	IMPU	0.93	2	60578955	C	G	0.42	-0.503	0.039	$2.7 x10^{-36}$
rs4671393	IMPU	0.97	2	60574455	A	G	0.28	0.554	0.043	$8.2 x10^{-36}$
rs6706648	GENO	-	2	60575544	T	C	0.40	-0.477	0.038	$9.0 x10^{-35}$
rs1896294	IMPU	0.95	2	60572578	C	T	0.30	0.526	0.042	$9.1 x10^{-34}$
rs11886868	IMPU	0.97	2	60573750	C	T	0.30	0.519	0.042	$2.2 x10^{-33}$
rs6738440	IMPU	0.83	2	60575745	A	G	0.69	0.529	0.044	$3.3 x10^{-31}$
rs10172646	IMPU	1.00	2	60574261	A	G	0.69	-0.488	0.041	$6.2 x10^{-31}$
rs10195871	GENO	-	2	60574093	A	G	0.30	0.487	0.041	$6.8 x10^{-31}$
rs7584113	IMPU	0.99	2	60574815	A	G	0.31	0.489	0.041	$7.4 x10^{-31}$
rs7557939	IMPU	0.99	2	60574851	A	G	0.69	-0.489	0.041	$7.5 x10^{-31}$
rs6709302	IMPU	0.72	2	60581133	A	G	0.29	-0.555	0.048	$3.6 x10^{-29}$
rs766431	IMPU	0.71	2	60573422	A	G	0.83	-0.656	0.060	$2.3 x10^{-26}$
rs11692396	IMPU	0.85	2	60564242	A	G	0.75	-0.493	0.048	$3.2 x10^{-24}$
rs13019832	IMPU	0.86	2	60564075	A	G	0.44	-0.390	0.041	$4.4 x10^{-21}$
rs6732518	IMPU	0.78	2	60562101	C	T	0.31	0.410	0.048	$2.8 x10^{-17}$
rs13024177	IMPU	0.68	2	60569812	C	G	0.13	-0.469	0.071	$5.8 x10^{-11}$
rs3759074	GENO	-	11	5214354	A	G	0.10	0.378	0.065	$6.0 x10^{-9}$
rs2855039	GENO	-	11	5228247	T	C	0.10	0.361	0.065	$2.6 x10^{-8}$
rs10128556	IMPU	0.99	11	5220259	C	T	0.91	-0.362	0.065	$3.0 x10^{-8}$
rs2071348	IMPU	0.99	11	5220722	G	T	0.09	0.361	0.065	$3.1 x10^{-8}$
rs2499959	GENO	-	11	4969610	G	A	0.15	0.282	0.055	$2.8 x10^{-7}$
rs13031396	IMPU	0.65	2	60570903	G	T	0.91	0.412	0.083	$8.0 x10^{-7}$
rs16907838	IMPU	0.83	11	4953149	C	G	0.90	-0.329	0.069	$2.0 x10^{-6}$
rs11820492	IMPU	0.99	11	4960668	C	G	0.88	-0.281	0.059	$2.4 x10^{-6}$
rs16907831	IMPU	0.95	11	4952024	C	T	0.88	-0.285	0.060	$2.5 x10^{-6}$
rs12477097	IMPU	0.99	2	60551901	A	C	0.64	0.192	0.041	$2.6 x10^{-6}$
rs7935991	IMPU	0.70	11	4945445	A	G	0.91	-0.373	0.079	$2.7 x10^{-6}$
rs4672393	IMPU	0.99	2	60551158	A	C	0.36	-0.187	0.041	$5.0 x10^{-6}$
rs4672394	GENO	-	2	60551965	T	C	0.36	-0.186	0.041	$5.2 x10^{-6}$
rs11821225	GENO	-	11	4961984	T	C	0.13	0.262	0.059	$9.0 x10^{-6}$
rs12070573	IMPU	0.90	1	94272320	C	T	0.50	0.176	0.040	$1.5 x10^{-5}$
rs7325795	GENO	-	13	93853857	A	G	0.07	-0.317	0.074	$1.8 x10^{-5}$
rs1044498	GENO	-	6	132214061	A	C	0.22	0.213	0.050	$2.1 x10^{-5}$
rs10263111	IMPU	0.61	7	87765328	C	G	0.40	0.221	0.052	$2.2 x10^{-5}$
rs9669943	IMPU	0.62	13	93864798	C	T	0.91	0.376	0.089	$2.4x10^{-5}$
rs6569761	IMPU	0.99	6	132209220	C	G	0.79	-0.211	0.050	$2.4x10^{-5}$
rs2734221	IMPU	0.91	7	142187348	A	G	0.21	-0.212	0.050	$2.8 x10^{-5}$
rs5030115	GENO	-	3	187940026	T	C	0.05	0.371	0.089	$3.3 x10^{-5}$
rs1800907	GENO	-	7	142208236	T	C	0.24	-0.191	0.046	$3.8 x10^{-5}$
rs3789393	IMPU	0.90	1	94271721	C	T	0.53	-0.166	0.041	$4.3 x10^{-5}$
rs9470224	IMPU	0.79	6	36248614	C	T	0.76	0.211	0.052	$4.4 x10^{-5}$
rs16911905	GENO	-	11	5205866	G	C	0.18	0.211	0.051	$4.4 x10^{-5}$
rs11465635	GENO	-	2	102364515	A	G	0.01	-0.870	0.213	$4.5 x10^{-5}$
rs7130110	GENO	-	11	5252680	C	G	0.12	0.241	0.059	$4.8 x10^{-5}$
rs17106864	IMPU	0.98	10	93374053	C	T	0.05	0.381	0.094	$5.8 x10^{-5}$
rs43111	IMPU	0.80	7	87694451	A	G	0.56	-0.180	0.045	$6.4 x10^{-5}$
rs16872536	GENO	-	5	74714241	T	G	0.06	-0.336	0.084	$6.4 x10^{-5}$
rs17028290	IMPU	0.83	2	60536877	A	G	0.67	0.182	0.046	$6.5 x10^{-5}$
rs1801823	IMPU	0.65	6	33367560	A	G	0.10	-0.325	0.081	$6.6 x10^{-5}$
rs17106868	GENO	-	10	93376651	C	T	0.05	0.369	0.092	$6.8 x10^{-5}$
rs10884984	GENO	-	10	112245884	T	C	0.09	0.292	0.073	$7.0 x10^{-5}$
rs9470228	IMPU	0.60	6	36271919	C	T	0.24	-0.234	0.059	$7.3 x10^{-5}$
rs6906612	IMPU	0.62	6	36259155	C	T	0.76	0.232	0.058	$7.4 x10^{-5}$
rs8073291	GENO	-	17	12588133	G	A	0.11	-0.239	0.060	$7.8 x10^{-5}$
rs4147846	IMPU	0.94	1	94267995	C	T	0.55	0.156	0.040	$9.7 x10^{-5}$
rs3945204	IMPU	0.92	1	94265361	C	T	0.58	-0.160	0.041	$9.7 x10^{-5}$

(Table 2 continued)

SNPs with p-values < $1x10^{-4}$ in the single-marker association test with linear regression in the CSSCD discovery cohort. Legend: blue: known locus, purple: genotyped for replication, red: failed genotyping or filtered after qc, yellow: did not fit in genotyping pool, grey: SNP in same locus as another SNP attempted for replication (purple, red, or yellow). Geno: Genotyped SNP, Impu: Imputed SNP, Rsq: Imputation r^2, Chr: Chromosome, BP: Base pair, EA: Effect allele, OA: Other allele, EA frq: Effect allele frequency, SE: Standard error.

Association Tests with rs7325795 in Other Cohorts

Given that the SNP rs7325795 was close to significance in the CSSCD replication dataset and that the effect was similar to the one observed in the discovery set, we attempted to validate this variant in three additional cohorts. The genotyping of rs7325795 failed in the SITT cohort. The association in the GHSU (n=266) was, again, close to significance (p=0.07) with a consistent direction of effect (**Table 4**). In the MSH (n=71) and JSCCS (n=88), rs7325795 was not significant (**Table 4**). Association results for rs7325795 were combined by inverse variance meta-analyses and the combined p-value for the CSSCD discovery set, CSSCD replication set, GHSU, JSCCS and MSH was $P=9.4x10^{-6}$ (**Table 4**).

Table 3 HbF single-marker replication results

SNP	Chr	BP	EA/OA	CSSCD Discovery				CSSCD Replication			
				EA Frq	β	SE	P-value	N	β	SE	P-value
rs5030115	3	187940026	T/C	0.05	0.371	0.089	$3.3x10^{-5}$	310	-0.235	0.204	0.25
rs10263111	7	87765328	C/G	0.40	0.221	0.052	$2.2x10^{-5}$	309	0.048	0.078	0.54
rs17106864	10	93374053	C/T	0.05	0.381	0.094	$5.8x10^{-5}$	309	0.149	0.187	0.43
rs10884984	10	112245884	T/C	0.09	0.292	0.073	$7.0 x10^{-5}$	309	-0.063	0.145	0.66
rs7325795	13	93853857	A/G	0.07	-0.317	0.074	$1.8x10^{-5}$	309	-0.298	0.163	0.07

SNPs tested in the CSSCD replication cohort in the single-marker association analysis with linear regression.

Chr: Chromosome, BP: Base pair, EA: Effect allele, OA: Other allele, EA Frq: Effect allele frequency, SE: Standard error.

Table 4 Replication of rs7325795 in additional cohorts

Cohort	N	Effect allele	Other allele	EAF	b	SE	P-value
CSSCD	1212	A	G	0.07	-0.317	0.074	$1.8x10^{-5}$
CSSCD rep	309	A	G	0.07	-0.298	0.163	0.07
GHSU	266	A	G	0.06	-0.323	0.177	0.07
SITT	Failed genotyping						
MSH	71	A	G	0.20	0.409	0.368	0.27
JSCCS	88	A	G	0.14	0.362	0.306	0.24
Combined		A	G		-0.269	0.061	$9.4x10^{-6}$

Association results for rs7325795 with linear regression in additional replication cohorts

EAF: Effect allele frequency, SE: Standard error.

Gene-Set Enrichment Analysis

To verify if there were signals with moderate effects implicated in shared biological processes, we performed a gene-set enrichment analysis using phenotype permutations. We tested the association between HbF levels and 171 gene sets, 12 of which showed nominal association (P<0.05) (**Table 5**). We attempted to replicate the association of the 12 pathways showing nominal association in an independent subset of 276 CSSCD samples genotyped on the Illumina-610-Quad array. Seven highly redundant pathways replicated (P<0.05) (**Table 6, Table 7**), and almost all of them shared genes such as *ACAT2*, *ACAT1*, *ALDH2* and *ECHS1* (**Table 7**).

Table 5 Gene-set enrichment analysis results (P < 0.05)

Gene-set name	Markers included (N)	Enrichment score	Normalized enrichment score	Nominal P-value	Main genes
Butyrate metabolism	14	0.723	3.018	0.001	*ACAT2*, *GAD2*, *ACSM1*, *ALDH2*, *ECHS1*
Pyruvate metabolism	12	0.735	2.746	0.004	*ACAT2*, *PCK1*, *ALDH2*, *ACSS2*, *ACOT12*
Valine, leucine and isoleucine degradation	17	0.668	2.662	0.010	*ACAA1*, *ACAT2*, *ALDH2*, *BCAT1*, *PCCA*
Pathogenic Escherichia coli infection	16	0.680	2.446	0.006	*TLR4*, *ROCK2*, *RHOA*, *CTNNB1*, *TUBA8*
Shigellosis	16	0.680	2.446	0.006	*TLR4*, *ROCK2*, *RHOA*, *CTNNB1*, *TUBA8*
Fatty acid metabolism	24	0.582	2.293	0.013	*ACAA1*, *ACAT2*, *ALDH2*, *ADH1C*, *ECHS1*
Benzoate degradation via CoA ligation	8	0.753	2.261	0.011	*ACAT2*, *ACOT11*, *ECHS1*, *GCDH*, *ACAT1*
Lysine degradation	14	0.710	2.260	0.002	*ACAT2*, *ALDH2*, *ECHS1*, *GCDH*, *PLOD2*
Glutamate metabolism	8	0.822	2.217	0.009	*GCLM*, *GAD2*, *GSS*, *CPS1*, *GSR*
Propanoate metabolism	12	0.629	1.608	0.023	*ACAT2*, *ALDH2*, *ACSS2 PCCA*, *ECHS1*
Basal transcription factors	6	0.714	0.996	0.046	*GTF2H3*, *GTF2A2*, *TAF1*
Nucleotide excision repair	9	0.654	1.140	0.049	*RFC1*, *CUL4B*, *ERCC4*, *GTF2H3*, *POLD1*

Gene-set enrichment analysis results with P < 0.05.

Table 6 Gene-set enrichment analysis replication results (P < 0.05)

Gene-set name	P-value discovery	P-value replication	Main genes
Butyrate metabolism	0.001	0.001	*ACAT2*, *GAD2*, *ACSM1*, *ALDH2*, *ECHS1*
Lysine degradation	0.002	0.001	*ACAT2*, *ALDH2*, *ECHS1*, *GCDH*, *PLOD2*
Pyruvate metabolism	0.004	0.001	*ACAT2*, *PCK1*, *ALDH2*, *ACSS2*, *ACOT12*
Valine, leucine and isoleucine degradation	0.010	0.001	*ACAA1*, *ACAT2*, *ALDH2*, *BCAT1*, *PCCA*
Benzoate degradation via CoA ligation	0.011	0.001	*ACAT2*, *ACOT11*, *ECHS1*, *GCDH*, *ACAT1*
Fatty acid metabolism	0.013	0.001	*ACAA1*, *ACAT2*, *ALDH2*, *ADH1C*
Propanoate metabolism	0.023	0.001	*ACAT2*, *ALDH2, ACSS2 PCCA*, *ECHS1*

Gene-set enrichment analysis results with P < 0.05 in the CSSCD samples genotyped on the Illumina 610-quad array.

Table 7 Main genes in associated pathways

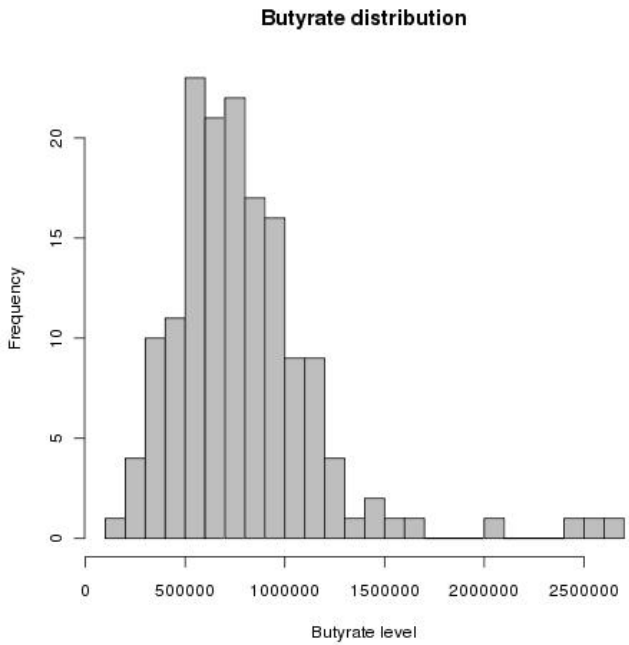
Pathway	ACAA1	ACAT2	PCK1	GAD2	ACOT11	ACSM1	ALDH2	BCAT1	ACSS2	PCCA	ADH1C	ACOT12	ME3	ECHS1	ACADS	GCDH	ACAT1	HSD17B4
Butyrate Metabolism		✓		✓		✓	✓							✓	✓		✓	✓
Lysine Degradation		✓					✓							✓		✓	✓	✓
Pyruvate Metabolism		✓	✓				✓		✓			✓	✓				✓	
Valine, Leucine, Isoleucine Degradation	✓	✓					✓	✓		✓				✓	✓		✓	✓
Benzoate Degradation via CoA ligation		✓			✓									✓		✓	✓	
Fatty acid metabolism	✓	✓					✓				✓			✓	✓	✓	✓	✓
Propanoate metabolism		✓					✓		✓	✓				✓			✓	

Main genes showing moderate association (in the CSSCD discovery cohort) in the pathways that replicated in the CSSCD samples genotyped on the Illumina 610-quad array.

Association Tests between Metabolites and HbF levels

We measured butyrate and BOHB serum levels in 156 CSSCD serum samples (**Figure 4**) and then tested the association between HbF levels and these two metabolites. There was no correlation between butyrate and HbF levels (P=0.26), while there appeared to be an inverse correlation (P=0.01, correlation=-0.20) between HbF levels and BOHB (**Figure 5**). Finally, we compared two linear models of HbFz levels. The null model included the seven SNPs identified in the fine-mapping of the three HbF-associated loci: *BCL11A*, *HBS1L-MYB* and β-globin[278]. The second model also included BOHB levels. The inclusion of BOHB level significantly improved (P=0.02) the HbF modelization and increased the explained variance from 51% to 54%. The same test was not significant for butyrate levels.

A



B

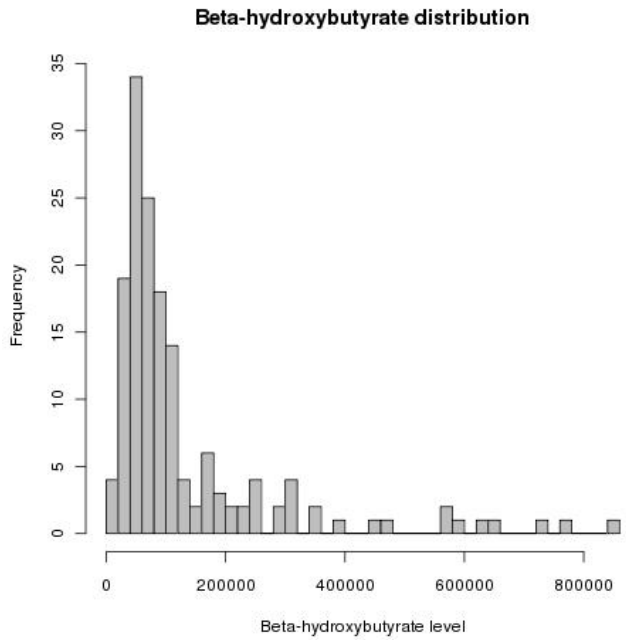


Figure 4 Metabolite levels distribution in 156 CSSCD serum samples

(A) Butyrate levels distribution. (B) BOHB levels distribution.

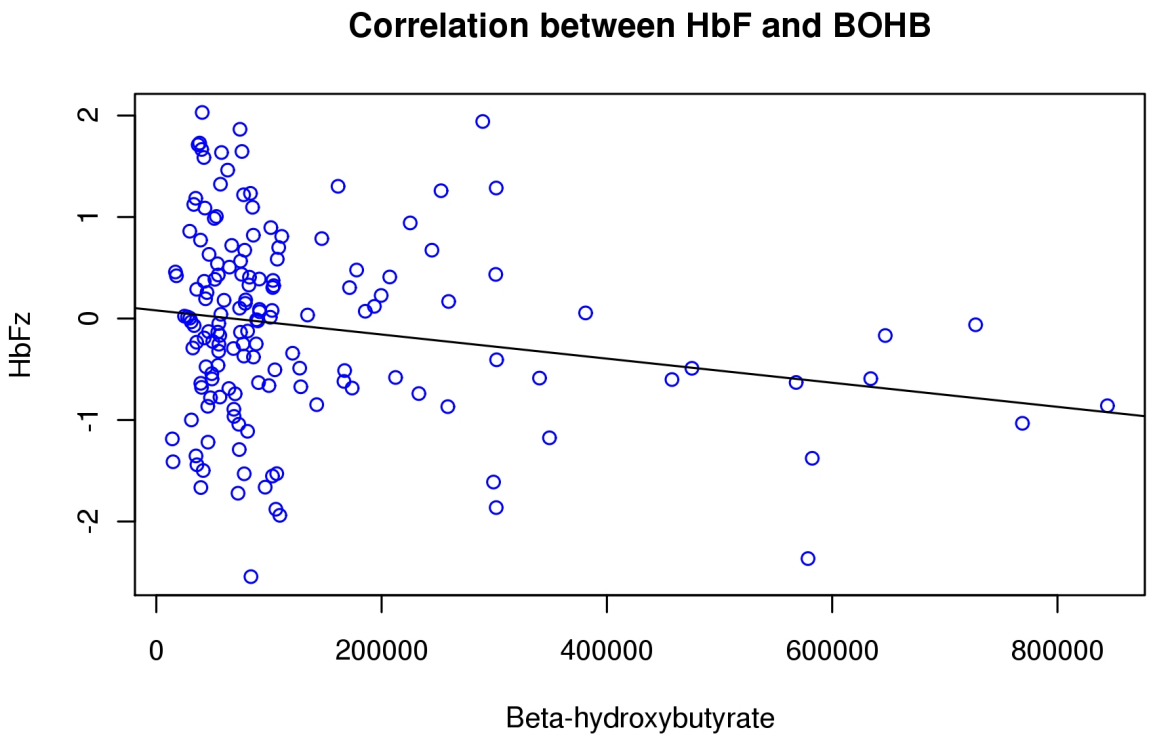


Figure 5 Correlation between HbF and BOBH levels

Correlation between HbF and BOHB integrated LC-MS peak areas (N=156, P=0.01, correlation=-0.20). The data are integrated LC-MS peak areas, which are proportional to concentration but are unitless values

Replication of Most Significant SNPs in Significant Gene-Sets

In the CSSCD replication, we tested the association of SNPs highlighted in the pathway analysis. The three SNPs that we were able to fit in our genotyping pools (rs9347340, rs11594057 and rs3465) were located in the genes *ACAT2* and *GAD2*. Overall, *ACAT2* is the second most associated gene in all replicated pathways and is shared amongst all of these pathways (**Table 7**). No SNP replicated (P<0.05) (**Table 8**). When testing the genetic association with BOHB or butyrate levels, the only SNP that showed association was

rs7311852 in *ALDH2* ($P=2.0x10^{-3}$), with BOHB (**Table 9**). The effect allele was associated with an increase in both HbF levels and BOHB.

Table 8 HbF replication results for associated SNPs in significant gene-sets

SNP	Chr	BP	EA	OA	EA Frq	CSSCD discovery			CSSCD replication		Annotation description	Gene name
						β	SE	P-value	β	P-value		
rs9347340	6	160108881	T	C	0.45	0.137	0.039	$5.2x10^{-4}$	-0.117	0.14	Intron	*ACAT2*
rs11594057	10	26613082	A	G	0.07	-0.243	0.075	$1.2x10^{-3}$	-0.138	0.37	Intron	GAD2
rs3465	6	160118385	A	G	0.17	-0.108	0.053	0.04	0.018	0.85	Synonymous	ACAT2

Single-variant association results in additional replication cohorts with linear regression for SNPs highlighted by the gene-set enrichment analysis.

Chr: chromosome, BP: base pair, EA: Effect allele, OA: Other allele, EAF: Effect allele frequency, SE: Standard error.

Table 9 Metabolite association results for associated SNPs in significant gene-sets

Chr	SNP	BP	Gene	EA	HbF (N=1,213)			β-hydroxybutyrate (N=88)			Butyrate (N=88)		
					β	SE	P-value	β	SE	P-value	β	SE	P-value
5	rs11539471	118888837	*HSD17B4*	C	0.093	0.043	0.03	-21580	26100	0.41	25880	52830	0.63
6	rs9347340	160108881	*ACAT2*	T	0.137	0.039	$5.2x10^{-4}$	-16200	25420	0.53	1444	51450	0.98
10	rs11594057	26613082	*GAD2*	A	-0.243	0.075	$1.2x10^{-3}$	29200	37470	0.44	-44580	75770	0.56
10	rs4604	135026081	*ECHS1*	A	-0.098	0.041	0.02	-1266	24260	0.95	42740	48760	0.38
11	rs11212525	107521224	*ACAT1*	T	0.125	0.056	0.03	-8144	38440	0.83	-42990	77500	0.58
12	rs7311852	110709687	*ALDH2*	G	0.235	0.084	$5.1x10^{-3}$	171300	53610	$2.0x10^{-3}$	-65020	114400	0.57
12	rs639667	119662134	*ACADS*	A	0.212	0.094	0.025	-50190	41960	0.24	-18870	85420	0.83
16	rs992381	20603065	*ACSM1*	A	-0.182	0.065	$5.0x10^{-3}$	14410	45700	0.75	-169700	90450	0.06

Linear regression association results with HbF, BOHB and butyrate for associated SNPs in the pathways highlighted by the gene-set enrichment analysis.

Chr: chromosome, BP: base pair, EA: Effect allele, SE: Standard error.

4.5 Discussion

We were able to replicate in our dataset the two previously associated regions covered by the array—*BCL11A* and the β-globin locus—indicating the reliability of our dataset. No SNPs in novel regions reached array-wide association. We attempted the validation of SNPs in novel regions with $P<1x10^{-4}$ in 310 independent CSSCD samples. None of the SNPs tested for replication showed significant association in the replication samples. Given the allele frequencies and effects observed in the discovery cohort, our statistical power in the CSSCD replication dataset for the five variants tested was between 52% and 55%, except for rs10263111, which had a statistical power of 77%. Therefore, the lack of replication could be due to insufficient statistical power in the replication dataset. To reach an 80% statistical power in a replication cohort (α=0.05), between 558 and 596 samples would be necessary for the following SNPs: rs5030115, rs17106864, rs10884984 and rs7325795.

In an effort to validate the most promising candidate, we attempted to genotype the variant rs7325795 in additional SCD cohorts. There was no significant association between rs7325795 and HbF levels in any of the three other replication cohorts. As in the CSSCD replication cohort, the rs7325795 p-value in the GHSU (N=266) was close to significance (P=0.07), and the effect was in the same direction as in the discovery cohort. When combining the association results of all cohorts in which rs7325795 was genotyped, $P=9.4x10^{-6}$ (**Table 4**). According to the HapMap data, the variant rs7325795 is monomorphic in the individuals of European descent. The variant could be a false

positive, but it is also possible that it did not replicate due to a lack of power in our replication cohorts. Given its allele frequency and effect in the discovery cohort, we estimated that our statistical power to replicate the rs7325795 association with HbF levels, if real, was 16%, 19%, 46% and 52% for MSH, JSCCS, GHSU and CSSCD replication dataset, respectively. The variant rs7325795 is located in the 3'-UTR of *GPC6* which encodes a glypican protein that is not well characterized but is thought to be a putative cell-surface coreceptor. This gene is mostly expressed in the ovaries, liver and kidneys[285]. Further analysis will be needed to confirm the association between the variant rs7325795 and HbF levels. Globally, our single-variant association results will help prioritize replication for future association studies.

We estimated that our statistical power in the discovery cohort was 8% for a variant with an effect size of 0.5% of the phenotypic variance explained, and 34% for an effect size of 1% (MAF=25%, $\alpha=1 \times 10^{-4}$). Since our statistical power in the discovery cohort was limited, and no SNPs reached the array-wide significance threshold except for the previously associated regions, we opted to perform a gene-set enrichment analysis to verify if SNPs with moderate effects were involved in shared biological processes. We used the GenGen software[275,276] to perform this analysis.

Since the individuals in the discovery and replication cohorts were not genotyped on the same array, we chose only to include in the replication steps the SNPs tested in the discovery set. Seven of the twelve pathways tested for replication were significant ($P<0.05$) in the 276 independent CSSCD samples. These pathways were highly

redundant. The genes *ACAT2*, *ACAT1*, *ALDH2* and *ECHS1* were shared amongst the majority of the replicated pathways. These gene sets most likely represent only one association signal in a biological process involving some of the overlapping genes.

The first evidence of butyrate implication in HbF regulation was when it was observed that infants born from diabetic mothers have a delayed hemoglobin switch[163,164]. Diabetic mothers are known to have elevated hydroxybutyrate levels. It was later observed that butyrate and some derivatives could delay the fetal to adult hemoglobin switch in sheep[286] and increase HbF levels by stimulating selectively the γ-globin gene expression in cultured human erythroid cells and baboons[280,281]. Many clinical trials to increase HbF levels with butyrate or derivatives have been initiated in SCD patients[165,166,287-291]. Some appeared to be conclusive in small numbers of patients, while others showed no significant increase.

The mechanism by which butyrate increases HbF levels is not fully understood but is thought to act through epigenetics and HDAC inhibition. Epigenetics consist of non-coding DNA modifications not affecting the DNA nucleotide sequence but that can affect gene expression with mechanisms like methylation and chromatin remodeling. HDACs can induce epigenetic modifications by removing an acetyl group on lysine, which compacts the chromatin. Erythroblasts treated with butyrate from patients with β-globin disorders show alterations in the protein-DNA interactions in the γ-globin gene promoter[292]. Fathallah et al. observed that exposure to butyrate increases histone

acetylation and decreases DNA methylation in the γ-globin gene and observed the opposite effect in the β-globin gene[167]. The inverse correlation between BOHB and HbF levels observed in 156 SCD patients is intriguing. Based on previous literature, we would expect to see an increase in HbF levels with BOHB. It is possible that BOHB at endogenous levels can decrease HbF levels and that higher levels could increase it through a mechanism like a feedback loop. Amajian et al. observed that *HDAC1*, *HDAC3*, *HDAC5* and *HDAC6* expression is increased in cultured mouse neural cells after butyrate treatment, suggesting that HDAC inhibitors like butyrate would also regulate these HDACs[293]. In this model, endogenous BOHB levels would increase HDAC expression and activity, and therefore repress HbF (**Figure 6**). A supplementary intake of BOHB or a derivative would enable individuals to reach a level where HDACs would inhibit themselves and thereafter increase HbF levels. If this model was confirmed, genetic variants in genes involved in butyrate metabolism might influence the dose response in SCD patients treated with butyrate. This inverse correlation between BOHB and HbF levels could also be a false positive because of the limited number of samples or sample heterogeneity due to the fact the samples were collected at different centers.

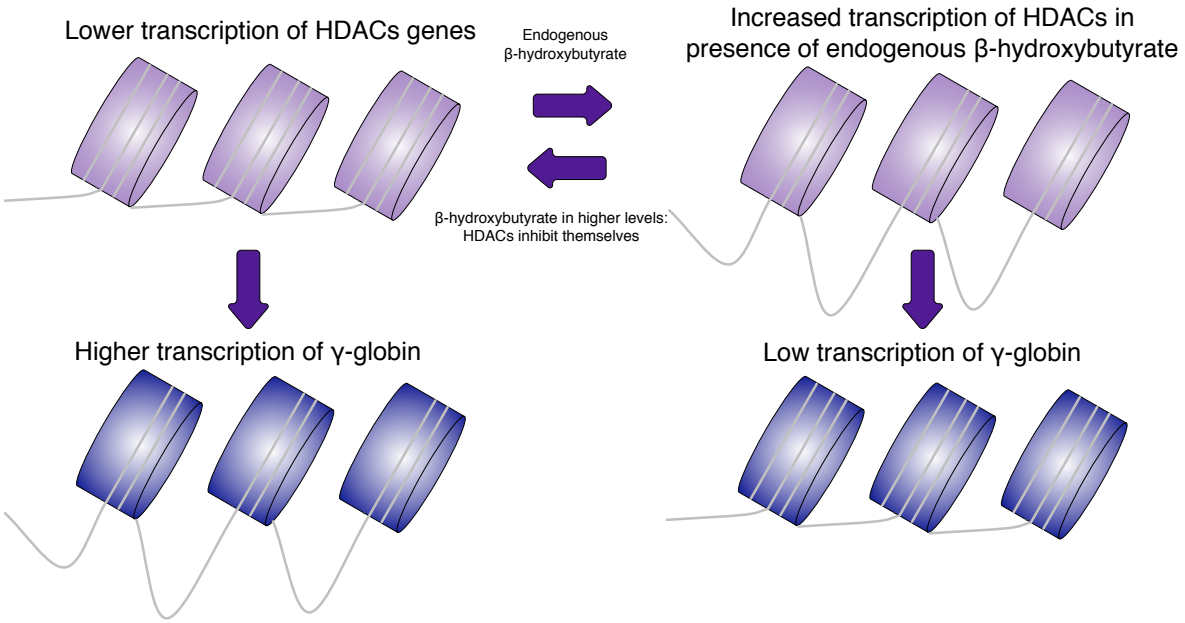


Figure 6 Possible model to explain inverse correlation between HbF levels and endogenous β-hydroxybutyrate levels

Histone deacetylase (HDAC) inhibitors like β-hydroxybutyrate (BOHB) could regulate HDAC through a feedback loop. BOHB at endogenous levels could increase HDAC transcription and therefore decrease HbF levels. Higher levels of BOHB could increase HDAC activity until they inhibit themselves, at which point HbF levels would increase.

The inverse correlation observed between β-hydroxybutyrate and HbF levels will need to be confirmed in a larger dataset, ideally where the samples would have all been collected at the same center and for which the two samples were collected at the same moment. To test the model of HbF regulation by BOHB with a HDAC feedback loop, we could differentiate cell lines such as CD34+ peripheral blood mononuclear cells (PBMCs). We could then measure γ-globin, β-globin and HDAC expression at baseline, after treatment with endogenous levels of BOHB and with levels similar to those reached during butyrate treatment in SCD patients.

In conclusion, we conducted a gene-centric association study with HbF levels in SCD patients. We were not able to validate novel single-variant associations. We identified and replicated seven biological pathways that are enriched for SNPs that influence HbF levels in SCD patients. One of these pathways was butyrate metabolism. We observed an inverse correlation between HbF levels and BOHB in SCD patients. Further analysis will be necessary to determine if endogenous BOHB levels could reduce HbF levels by increasing HDAC levels.

4.6 Acknowledgements

We would like to thank all the individuals who participated in this study. The authors also acknowledge the contribution of Mélissa Beaudoin for DNA genotyping, and Cameron D. Palmer for genotype imputation.

Chapter 5: Whole-Exome Sequencing of Nineteen Extremely Mild Sickle Cell Disease Patients

Authors

Geneviève Galarneau, Mélissa Beaudoin, Ken Sin Lo, Bamidele O. Tayo, Richard S. Cooper, Marvin Reid, Ian R. Hambleton, Joel N. Hirschhorn, Colin A. McKenzie, Guillaume Lettre.

Reference

Galarneau G, Beaudoin M, Lo KS, Tayo BO, Cooper RS, Reid M, Hambleton IR, Hirschhorn JN, McKenzie CA, Lettre G. Chapter 5: Whole-Exome Sequencing of Nineteen Extremely Mild Sickle Cell Disease Patients**.** 2014. In preparation

Author's Contribution

MB performed the whole-exome sequencing of the Jamaican samples and the genotyping, GG and KSL performed the data analysis, GG, MB, KSL, BOT, RSC, MR, IRH, JNH, CAM and GL contributed reagents, GG, JNH and GL conceived and designed the study, GG wrote the manuscript.

5.1 Abstract

Sickle cell disease (SCD) is caused by mutations in the β-globin locus. Although it is a monogenic disease, patients show a high clinical heterogeneity. The Jamaica Sickle Cell Cohort Study is a prospective study in which each participant was assessed a severity score based on the average number of complications per year during the first 18 years of life. To identify variants that potentially reduce disease severity, we sequenced the whole-exomes of 19 SCD patients with extremely few complications. Our first analysis strategy was to calculate a gene score based on the novel functional mutations and their Polyphen-2 score. Amongst the genes with highest scores was the gene *GPX1*, which protects the erythrocytes against oxidative stress. The signal in *GPX1* was due to one common variant not present in the database used for filtering, rs1050450. We genotyped rs1050450 in the rest of the individuals for which DNA was available. The association between rs1050450 and the severity score did not reach significance (P=0.09, N=76) but trended in the right direction. In a combined analysis including sequenced individuals (N=95), there appeared to be a nominal association between the severity score and rs1050450 (P=0.02). We attempted to replicate this result in 1,313 patients from the Cooperative Study of Sickle Cell Disease (CSSCD) by testing the association between a perfect linkage disequilibrium (LD) proxy for rs1050450 and individual clinical complications, but none showed significant association. We later gained access to the whole-exome sequence data of a set of 50 Nigerians that we used as a control group in a case-control analysis. No variant reached genome-wide significance. We were unable to identify novel variants that caused the extremely mild status of the 19 patients

sequenced. Further analysis will be necessary to determine whether the known variant rs1050450 in *GPX1* is a modifier of SCD severity.

5.2 Introduction

Sickle cell disease (SCD) is caused by a single missense polymorphism located in the β-globin gene. Although it is caused by a single amino acid change, the range of complications varies widely from one patient to another. Severe sickle cell patients suffer from complications such as pain crisis, acute chest syndrome, osteonecrosis and stroke and usually require frequent hospitalization. Patients with extremely mild cases may not suffer from any complications for decades. This clinical heterogeneity of SCD is thought to be due to both environmental and genetic factors. Amongst the environmental factors, nutrition, hydration and body temperature[86,141] are thought to affect clinical complications, and physicians advise SCD patients to avoid dehydration, excessive stress, exhaustion, extreme body temperatures and high altitudes[142]. Current treatments available for SCD patients include hydroxyurea, blood transfusions and pain management.

Systematic genetic screening of newborns for SCD is now performed in many high-income countries and a few middle-income countries[294], and can lead to the identification of SCD cases that might not be otherwise identified for years because of their extremely mild status. Some of these rare mild SCD patients do not experience any complications for many years. We thought that some of these exceptionally mild SCD patients might be carriers of novel and functional variants causing their extreme phenotype. Moreover, the identification of such variants that decrease SCD severity could

help understand the biological mechanisms involved in SCD mildness and aid in the treatment of SCD patients.

The Jamaica Sickle Cell Cohort Study (JSCCS) recruited 311 newborns with SCD between 1973 and 1981 in Jamaica[295]. The clinical events developed by these patients while they were followed were compiled. For each patient, a severity score was generated by calculating the average number of events per year (NEPY) during the first 18 years of their lives (**Figure 1**). Amongst the patients classified as potentially mild, the ones with low NEPY did not show any statistical difference in regards to total hemoglobin, fetal hemoglobin, packed cell volume, red blood cell count, mean cell volume, mean cell hemoglobin, mean cell hemoglobin concentration, nucleated blood count and reticulocyte count (**Table 1**). Individuals with SCD with extremely rare complications could be carriers of novel penetrant and protective functional mutations.

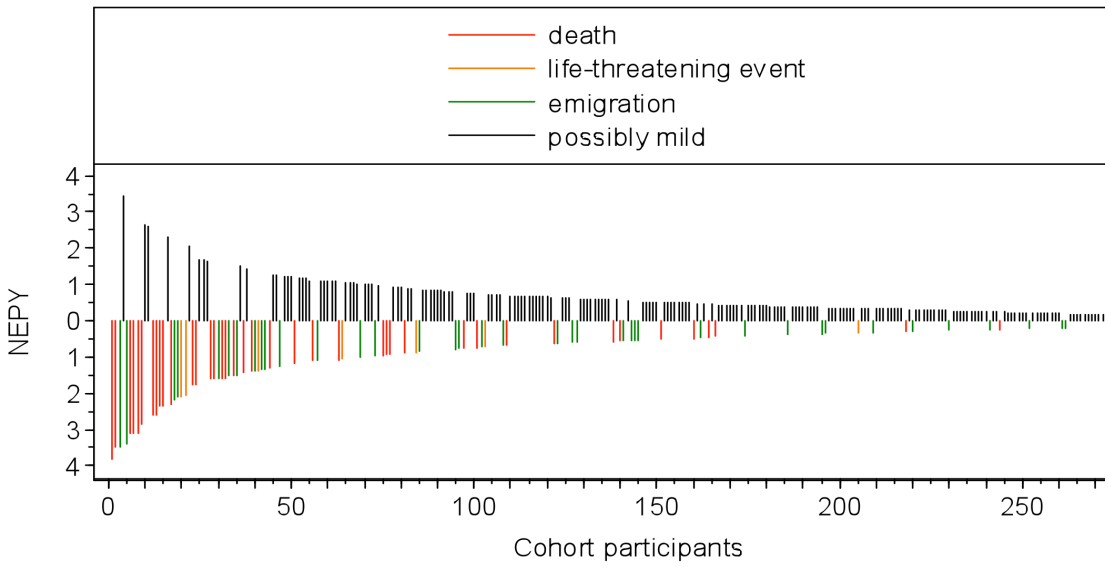


Figure 1 NEPY distribution in the Jamaica Sickle Cell Cohort Study

(Adapted from: Colin McKenzie and Ian Hambleton, unpublished)

Cohort participants rank ordered by NEPY. Patients with severe disease are shown in the lower half, patients with "possibly mild" disease are shown in the upper half of the figure. The NEPY corresponds to the number of clinical events during the first 18 years of the patient's life.

Table 1 Comparison of steady-state hematology between low NEPY and the rest of the 'possibly mild' participants

	Low NEPY	Other NEPY	P-value
Group summary			
Number in group	32	163	
Years of follow-up (mean, SD)	16.81 (2.07)	17.20 (1.43)	0.35
NEPY (mean, 95% CI)	0.073 (0.057- 0.088)	0.571 (0.497 – 0.645)	<0.001
Steady-state hematology			
Total hemoglobin (g/dL)	7.91 (1.08)	7.76 (0.96)	0.43
Fetal hemoglobin (%)	7.05 (5.50)	5.72 (4.28)	0.13
Packed cell volume (%)	23.21 (3.22)	22.78 (2.84)	0.44
Red blood cell count (10^{12}/L)	2.75 (0.43)	2.71 (0.42)	0.66
Mean cell volume (fl)	85.50 (7.37)	85.18 (7.52)	0.83
Mean cell hemoglobin (pg)	29.16 (2.84)	28.98 (2.88)	0.74
Mean cell hemoglobin concentration	34.24 (1.10)	34.23 (1.14)	0.94
Nucleated blood count (10^9/L)	14.22 (3.64)	13.78 (2.67)	0.43
Reticulocyte count (%)	11.32 (2.81)	11.70 (2.79)	0.48

Comparison of hematological characteristics between potentially mild patients subgroups. There were no significant differences between potentially mild patients with low NEPY from the rest of the potentially mild patients.

In this study, we aimed to identify novel functional variants modifying the severity of SCD by performing whole-exome sequencing of very mild cases of SCD from the JSCCS. The selected patients had very few or no complications during their first 18 years of life (very low NEPY) and did not suffer from any life-threatening complications during the time they were followed in the study. Our first approach consisted of ranking genes by calculating a gene score based on novel variants identified in the 19 sequenced individuals. Amongst the genes with highest scores was the gene *GPX1*, encoding a glutathione peroxidase 1, which protects the erythrocytes against oxidative stress[296]. The signal in *GPX1* was due to one variant, rs1050450. We genotyped rs1050450 in the rest of the individuals of the JSCCS for which deoxyribonucleic acid (DNA) was available. The association between rs1050450 and the severity score did not reach significance

(P=0.09, N=76) but trended in the right direction. In a combined analysis also including individuals selected for the sequencing (N=95), there appeared to be an association between the severity score and rs1050450 (P=0.02). We attempted to replicate these results by testing the association between a proxy for rs1050450 and complications in 1,313 individuals from the Cooperative Study on Sickle Cell Disease (CSSCD)[236], but none of the clinical complications showed a significant association with the variant. It is unclear if this lack of replication is due to the difference between the phenotypes tested or if the association between the variant rs1050450 and the NEPY is a false positive. We later gained access to the whole-exome sequence data of a group of 50 Nigerians recruited for a hypertension study and performed a case-control analysis as a second approach. No variant reached genome-wide significance. Nevertheless, our results suggest that a known variant in *GPX1* could potentially be a genetic modifier of SCD, but further validation will be needed.

5.3 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 Loyola University Chicago Ethics Committee, the University of West Indies Ethics Committee and the Montreal Heart Institute Ethics Committee.

Jamaica Sickle Cell Cohort Study

The JSCCS identified 315 newborns with SCD from 100,000 consecutive live births between 1973 and 1981 in Jamaica[295]. Of these, 311 were recruited in a prospective study. Cohort participants were followed at the clinic of the Sickle Cell Unit (SCU), University of the West Indies (UWI), Kingston, Jamaica, at least annually when they were well, to collect information on steady-state parameters, and were encouraged to visit the clinic whenever they were sick.

NEPY Calculation and Mild Status Assessment

The NEPY aimed to reflect patient severity and consisted of the number of clinical events divided by the number of follow-up years during the first 18 years of life. The complications considered as events were: painful crisis events (dactylitis, bone and abdominal painful crisis), avascular necrosis, acute chest syndrome, acute splenic sequestration, septicemia, and parvovirus B19 infection. In addition to the NEPY, the

patients of the cohort were classified in two groups (severe or potentially mild) depending on whether or not they suffered from life-threatening complications. Death and complications such as stroke excluded patients from a mild disease classification. Thirty-two individuals were identified as mild disease patients with a very low NEPY. Of these, 19 unrelated patients with NEPY values ranging from 0 to 0.08 were selected for whole-exome sequencing.

Hemoglobinopathy Validation

We validated the hemoglobinopathy of the individuals for which we had DNA samples by genotyping the rs334 variant using restriction fragment length polymorphism. The single nucleotide polymorphism (SNP) rs334 creates a desulfovibrio desulfuricans I (DdeI) restriction consensus site (CTNAG) on the wild type allele (allele A) but not on the mutant allele (allele T). We used 100 ng of genomic DNA (gDNA) to perform a polymerase chain reaction (PCR) amplification of a 457-base-pair (bp) fragment surrounding the SNP followed by an enzymatic digestion with DdeI. The wild type allele produces two fragments, one of 268 bp and one of 189 bp. The mutant allele is not digested and remains a 457 bp fragment.

Whole-Exome Re-Sequencing of the Nineteen Jamaican Samples

We followed the SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library protocol (Sure-Select Human A11 exon and human a11-exon + version 2.0.1, May 2010), except when specified. The DNA quantification was performed

using PicoGreen (Invitrogen). We used 3 ug of gDNA per library. DNA shearing was performed with Covaris S2.

Instead of Agencourt AMPure XP beads, we used Qiagen Qiaquick columns for the DNA purification following the manufacturer's instructions after the shearing step. DNA was quantified with the Agilent 2100 Bioanalyzer using a DNA 1000 chip; a peak was observed at 150 bp. After the repair end step, the purification was performed with Qiagen Qiaquick column. We used a MinElute column to purify the DNA after the addition of an Adenosine base. For the ligation of the paired-end adapters, we used 10 uL of the library and 4 uL of water instead of 14 uL of the library. After the ligation step, we observed a peak of 200 bp on a HS Bioanalyzer chip. We performed 4 PCR cycles for the amplification of the adapter-ligated library.

The hybridization was performed with the Agilent SureSelect All-Exon 50 Mb capture library. We used 500 ng of each DNA library. To generate the clusters and to perform the sequencing reaction, we used the Paired-end Cluster Generation Kit v4 (catalog # PE-203-4001), the 36 bp sequencing kit v5 (catalog # FC-104-5001) and Flow-cell v4. The samples were sequenced on the Illumina GAIIx next-generation DNA sequencer. We performed paired-end sequencing generating reads of 100 bp. Each individual was sequenced on two lanes. The library concentration was 10 pMol/lane.

Whole-Exome Re-Sequencing of the Fifty Nigerian Samples

We obtained from collaborators the whole-exome sequence data for 50 Nigerians sequenced as part of a hypertension study. These individuals were unrelated and represented the two extremes of a blood-pressure distribution. A sequencing service platform performed the sequencing of the Nigerian samples. Briefly, the hybridization was performed with the Agilent SureSelect Human All Exon Kit v2. The captured DNA was amplified with barcoded primers. Two libraries with different barcodes were pooled for the sequencing of the Nigerian samples. The samples were sequenced on a Hi-Seq 2000 with two libraries per lane.

Variant Calling

Three years have passed between the first analysis of the Jamaican samples and the reception of the Nigerian whole-exome sequence data. New versions of the tools used in the sequencing data analysis and variant calling process were released during that period. Therefore, we redid the variant calling with both Jamaican and Nigerian samples upon reception of the Nigerian sequencing data with the software's most recent version. **Table 2** shows the versions used for both rounds of variant calling. The read alignment, the removal of duplicated reads, the depth of coverage calculations, the recalibration, the realignment and the variant calling were performed with GATK[224] according to the best practices guidelines[297,298] latest version at the time that each round was performed.

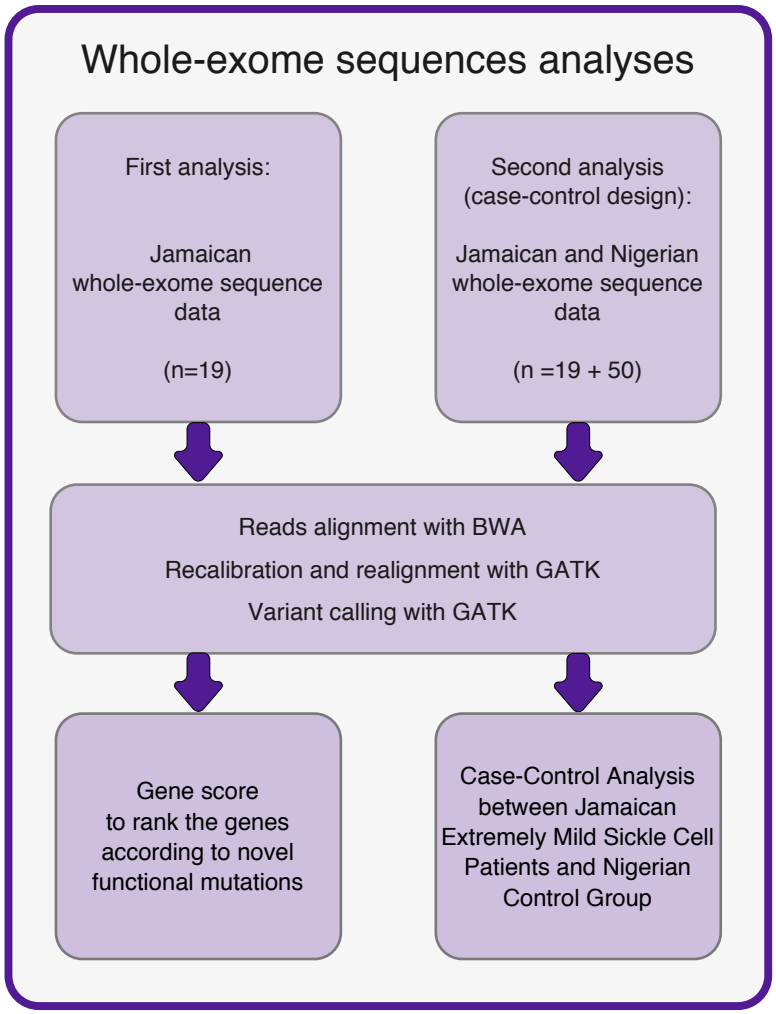


Figure 2 Whole-exome sequences analyses

A first analysis was performed on the 19 whole-exome sequences of the 19 SCD Jamaican patients with low complications. In the first analysis, a gene score was calculated based on the novel functional mutations. A second analysis with a case-control design has been done. In the second analysis, the variant calling has been redone to include both Jamaican and Nigerian samples.

Table 2 Variant calling bioinformatics pipeline

	Jamaicans (Gene score)	Jamaicans (Case-control)	Nigerians
Reference genome	hg19		
dbSNP	dbSNP131	dbSNP135	dbSNP135
BWA version	0.5.9	0.6.1	0.6.2
SAMtools version	0.1.11	0.1.18	0.1.18
Picard version	1.4.1	1.68	1.78
GATK version	1.0.5467	1.6.5	2.1.13
GATK unified genotyper options	-stand_call_conf 50.0 -stand_emit_conf 10.0 -dcov 200 -glm BOTH	-stand_call_conf 0.0 -stand_emit_conf 0.0 -dcov 200 -glm BOTH	

Bioinformatics tools and versions used for each round of read alignment and variant calling

The reads were aligned to human reference genome hg19 with BWA[223]. Alignments in SAM format were generated with BWA[223] sampe command. SAM files were merged with Picard[299], with the validation stringency set to lenient. Reads were indexed with Samtools[300]. Duplicated reads were marked and removed with Picard[299], with the validation stringency set to lenient.

The read realignment and the score recalibration were performed with GATK[224]. In the first alignment of the Jamaican samples, the realignment of the reads was performed after the recalibration of the read scores, whereas the realignment was done before the recalibration in the second variant calling pipeline.

Depth of coverage was calculated with GATK[224] with the following options: omitBaseOutput was set to true, minimum mapping quality (mmq) was set to 9 and minimum base quality (mbq) was set to 9. We calculated the percentage of bases covered at 1x, 5x, 10x, 20x and 30x. The variant calling was performed with GATK[224] unified genotyper with options -stand_call_conf 50.0 -stand_emit_conf 10.0 -dcov 200 -glm BOTH for the Jamaicans samples variant calling and with options -stand_call_conf 0.0 -stand_emit_conf 0.0 -dcov 200 -glm BOTH for the variant calling round with both Jamaicans and Nigerians samples.

Gene Score

We established a gene score calculation in order to rank the genes according to the novel functional mutations they contain. The gene score took into account the impact of novel functional mutations on the protein it encodes as predicted by the PolyPhen-2[301] score. PolyPhen-2 is an algorithm that predicts the impact of an amino acid substitution on the structure and function of a protein. This algorithm uses both physical and comparative properties. The gene score also includes a correction for the gene length. Since the PolyPhen-2 score is limited to missense variants, we assigned to non-sense and splice-site mutations the equivalent of the highest possible PolyPhen-2 score: 1.0. The gene score was the sum of the two highest PolyPhen-2 scores of novel functional variants in this gene per individual divided by the square root of the gene length.

$$Score = \frac{\sum_{i=1}^{n}(2\ highest\ polyphen\ scores)_i}{\sqrt{gene\ length}}$$

GPX1 rs1050450 High-Resolution Melting Genotyping with Nested PCR

Because the SNP rs1050450 is located in a region highly homologous to regions on chromosome 21 and chromosome X (**Figure 3**), we performed a high-resolution melting genotyping (HRM) with nested PCR. We genotyped the variant rs1050450 in the 96 JSCCS samples for which DNA was available. The genotyping failed for one sample. The first PCR amplification (**Table 3**) was performed on 25 ng of gDNA with a final concentration of 400 nM for each sens and antisens primer **(Figure 3)**, using the Gotaq Hostart PCR kit following protocol specifications. For the second PCR reaction **(Table 3)**, 1 uL of a 1/100 dilution of the first PCR reaction was used. The second PCR was performed on Illumina Eco with 5x MBI EVOlution Evagreen PCRmix and a final concentration of 250 nM for each primer **(Figure 3)**. Sanger sequencing was also used to validate the HRM genotyping protocol. The purification of the second PCR reaction was made with Qiagen Qiaquick columns. Three samples of the Jamaican cohort and one heterozygous sample of northern and western Europe ancestry (CEU) were sequenced.

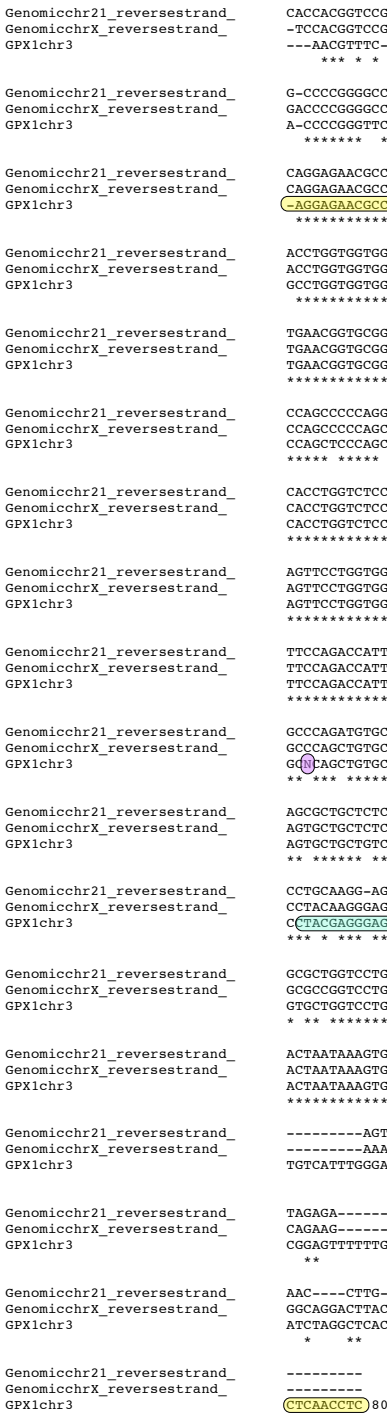


Figure 3 *GPX1* rs1050450 nested PCR primers

First set of primers (yellow), second set of primers (green), and targeted SNP (purple).

Table 3 PCR specifications for rs1050450 genotyping with nested PCR

	1st PCR	2nd PCR
Initial	95°C 2 minutes	95°C 5 minutes
Denaturation	94°C 30 seconds	95°C 10 seconds
Annealing	65°C 30 seconds	60°C 30 seconds
Elongation	72°C 1 minute	72°C 15 seconds
Cycles numbers	30	40
Elongation (Final)	72°C 5 minutes	-
High Resolution Melting curve	No	Yes

Association Test between rs1050450 and NEPY in the JSCCS

To validate the rs1050450, we performed a linear regression to test the association between rs1050450 and the NEPY in 76 samples genotyped by HRM and that were not selected for sequencing. We also performed a second association test, this time including all individuals genotyped (N=95).

Association Test between rs1050450 Proxy rs9858280 and Clinical Complications in the CSSCD Cohort

The CSSCD is a multicentered prospective study of the natural history of SCD in African Americans[236]. Complications studied in the CSSCD included acute chest syndrome, pain crisis, osteonecrosis, stroke, leg ulcers and priapism. Of the 4,085 individuals in the CSSCD cohort, 1,313 were genotyped on the ITMAT-Broad-CARe (IBC) array[242,282]. The

SNP rs1050450 was neither genotyped nor imputed on the IBC array. To identify a proxy for rs1050450 on the IBC array, we performed a high-resolution melting rs1050450 genotyping with nested PCR, as described above, in 94 CSSCD samples. We identified an imputed proxy for rs1050450 (r^2=1.0) on the IBC array and validated it in the 1000 Genomes samples[302]. We tested the association between this proxy, rs9858280, and the different clinical complications available in the CSSCD cohort. For acute chest syndrome and pain crisis rates, we performed a Poisson regression with custom scripts in the R 2.10.0 statistical package (www.r-project.org) with sex and age at baseline and the first 10 principal components as covariates[282]. For osteonecrosis, priapism, stroke and leg ulcer, we performed a logistic regression with PLINK v1.07[219] with sex and age at baseline and the first 10 principal components as covariates. Analyses for all complications were stratified based on α-thalassemia status and association results were combined by inverse variance meta-analyses with the software Metal[256]. We also tested the association between the imputed SNP and normalized fetal hemoglobin levels (HbFz) corrected for age at baseline, sex and hemoglobinopathy using the software PLINK v1.07[219]. The first 10 principal components were included as covariates. Finally, we performed a reverse regression, a regression of the imputed SNP rs9858280 on clinical complications under a quasi-binomial model with R 2.10.0. In this last analysis, the dependent variable was the genotype, and the clinical complications were the predictors. We performed an ANOVA to compare the general linear null model comprising age, sex and the 10 principal components with a model also including the following severe complications: pain crisis, acute chest syndrome, osteonecrosis and stroke.

Case-Control Analysis between Jamaican Extremely Mild Sickle Cell Patients and Nigerian Control Group

The file containing the variant calls (.vcf) of the second variant calling round including both Jamaican extremely mild sickle cell patients and the Nigerian control group was transformed in the tped/tfam format file using PLINK/SEQ v0.08 (https://atgu.mgh.harvard.edu/plinkseq). We performed an estimation of pairwise identity by descent (IBD) with PLINK v1.07[219]. One of the Nigerian samples was removed because 20% of its variants were identical-by-descent with half of the Nigerians samples, suggesting contamination of this sample. We performed a logistic regression with PLINK v1.07[219] on all variants with a minor allele frequency (MAF) >0.05 in the combined dataset (Jamaican and Nigerian individuals) and included sex as a covariate. We performed two rare variants analyses, including rare missense, non-sense and splice site variants with MAF<=0.05 with SKAT-O v0.81[303] and a burden test. The burden test was a custom script implemented with the R 2.10.0 statistical package (www.r-project.org/). This burden test consisted of a fisher's exact test for each gene to compare the number of carriers and non-carriers of rare functional mutations in the cases and in the controls. SKAT[303] is a statistical program to test the association between sets of rare variants and either continuous or dichotomous phenotypes. This tool uses kernel machine methods, which provides the possibility to include covariates in the analysis. As opposed to the burden test, SKAT can detect situations in which variants in a same set affect the phenotype in opposite directions. For the analysis with SKAT, we included sex as a covariate. For the burden analysis, we divided the individuals of each population into

two groups depending on whether or not they were carriers of at least one functional variant for the gene tested and performed a Fisher test.

Validation of the Variant at chr22: 30184798

We attempted to validate the variant at chr22: 30184798. This variant showed high association in the case-control study and seemed to be enriched in the Jamaican sequenced samples. For the PCR amplification, we used the following primers: "AGGGGAAGTGGTGTCTTGTG" and "CATGATCCCATCCTGAGACC" and then performed Sanger sequencing.

5.4 Results

Sickle cell disease patients show a high clinical heterogeneity. Few of these patients show a very mild status and suffer from very few complications or none at all. This mild SCD status could be of genetic origin. In this study, we aimed to identify such variants by performing whole-exome sequencing of 19 SCD patients from the JSCCS with extremely few complications. Discovering variants reducing SCD severity would help to understand the disease's pathophysiology and provide novel targets for the treatment of SCD clinical complications.

Whole-Exome Re-Sequencing Quality of the Nineteen Jamaicans Samples

We performed whole-exome re-sequencing of 19 unrelated very mild SCD patients. We obtained a mean coverage of 129X in our Jamaican samples and 84% of the targeted genomic regions were re-sequenced at ≥20X (**Table 4**). Fifty-eight percent of the reads were on target. We identified 41,829 high-quality non-synonymous (nonsense, missense) or splice site variants, 23% of which have not yet been reported in dbSNP132. (**Table 5**). We obtained a transition-to-transversion (Ti/Tv) ratio of 3.15 for the variants identified in our samples. The theoretical Ti/Tv ratio in the human exome is ~3.2. These numbers are consistent with other large-scale next-generation DNA re-sequencing projects, suggesting that the quality of our data is high.

Table 4 Mean coverage and percentage of reads on target per sample for the re-sequencing of the 19 Jamaicans patients

Sample id	NEPY	Mean coverage	% 20x coverage	% on target
55	0.07	143	84.5	57.3
150	0.07	145	85.3	59.1
524	0.07	134	82.7	56.9
850	0.04	129	84.5	56.7
1134	0.07	150	85.1	61.1
1338	0.07	139	84.6	57.7
1398	0.08	107	84.9	58.7
1614	0	114	85.1	59.3
1743	0.04	104	81.7	64.5
1893	0	129	84.0	55.8
2023	0.04	103	81.4	61.8
2263	0.08	142	85.0	56.4
2288	0	122	84.3	56.1
2312	0.04	156	84.9	60.5
2320	0.08	130	85.8	57.3
2819	0.08	142	84.8	58.9
2820	0.07	142	84.9	58.2
2845	0.04	102	81.7	57.2
2857	0.08	122	86.0	54.6
Mean	**0.054**	**129.2**	**84.3**	**58.3**

Table 5 Variants identified by whole-exome re-sequencing of 19 Jamaican with mild sickle cell disease

Annotation	Total	Known	New
Nonsense	398	244	154
Missense	41,128	31,620	9,508
Splice site	303	139	164
Synonymous	43,273	36,077	7,196
5'-UTR	3,966	2,985	981
3'-UTR	5,455	3,994	1,461
Intron	57,822	43,606	14,216
Near gene 5' (2000 bp)	853	614	239
Near gene 3' (500 bp)	279	213	66
Intragenic	153,477	119,492	33,985
Intergenic	8,928	6,065	2,863
Total	**162,405**	**125,557**	**36,848**

Variants annotation was performed with dbSNP132.

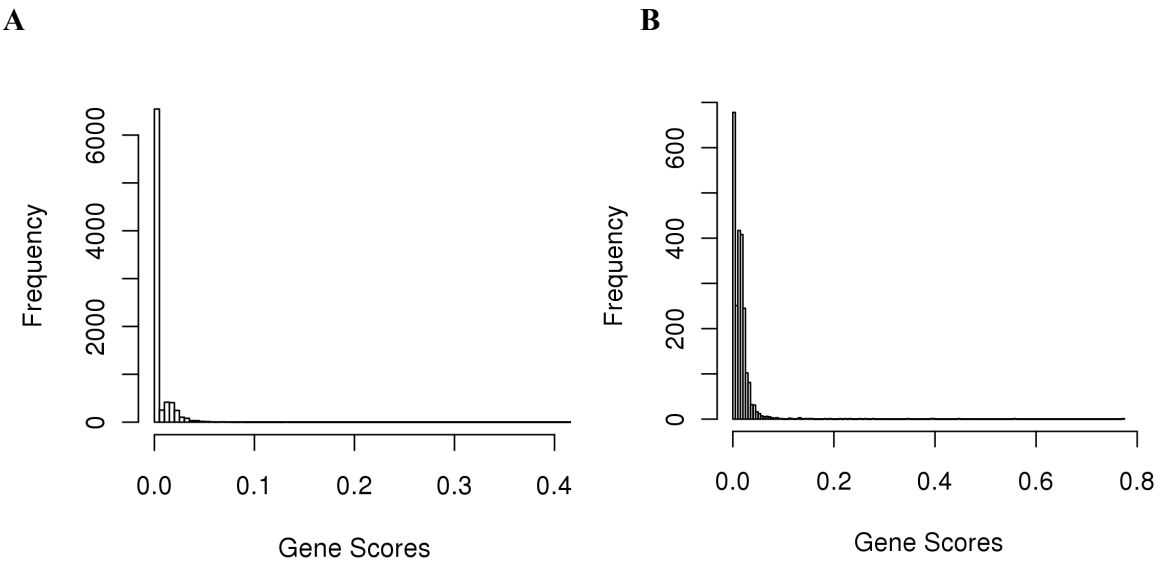


Figure 4 Gene scores distribution

We calculated a gene score based on the novel functional mutations they carried and corrected for the length of the gene. A. Genes scores distribution (all genes). B. Genes scores distribution of genes with scores >0.

Table 6 Genes with the highest scores

Gene	Length	Carriers	Novel variants	Score	Mutations type	Gene function
TSPAN10	1613	19	1	0.7719	Splice site	Unknown
CTAGE9	2577	18	4	0.5558	Missense	Unknown
FAM86B2	1267	11	2	0.4477	Missense	Unknown
RGPD8	7299	19	5	0.3979	Missense	Unknown
ANKRD20A4	3517	19	11	0.3928	Missense Nonsense	Unknown
SHKBP1	2365	12	1	0.3496	Splice site	Unknown
CD24	2180	8	2	0.2784	Nonsense	Modulates B-cell activation responses. Promotes AG-dependent proliferation of B-cells, and prevents their terminal differentiation into antibody-forming cells.
UBTFL1	1182	9	2	0.2618	Missense	Essential for proliferation of the inner cell mass and trophectodermal cells in peri-Implantation development.
CDC27	5610	19	1	0.2537	Splice site	Component of the anaphase promoting complex/cyclosome (APC/C), a cell cycle-regulated E3 ubiquitin ligase that controls progression through mitosis and the G1 phase of the cell cycle. The APC/C complex acts by mediating ubiquitination and subsequent degradation of target proteins.
MUC4	17108	19	23	0.2303	Missense	Has anti-adhesive properties, Seems to alter cellular behavior through both anti-adhesive effects on cell-cell and cell-extracellular matrix interactions and in its ability to act as an intramembrane ligand for ERBB2, Plays an important role in cell proliferation and differentiation of epithelial cells.
GOLGA6B	3178	9	2	0.2227	Missense	Unknown
C12orf57	543	4	1	0.2146	Splice site	Unknown
USP17L2	1593	8	12	0.1931	Missense	Deubiquitinating enzyme that specifically removes conjugated ubiquitin from CDC25A, thereby playing a key role in cell cycle regulation. Has transforming capabilities on cell lines. May also function in cell apoptosis.
RGPD4	7172	16	8	0.1811	Missense Nonsense	Unknown
LOC100132396	1207	9	3	0.1590	Missense	Unknown
MLL3	16862	17	3	0.1540	Splice site	Histone methyltransferase. Central component of the MLL2/MLL3 complex, a coactivator complex of nuclear receptors, involved in transcriptional coactivation. May be involved in leukemogenesis and developmental disorder.
FABP12	423	3	1	0.1459	Missense	May play a role in lipid transport.
GPX1	1183	15	1	0.1450	Missense	Protects the hemoglobin in erythrocytes from oxidative breakdown
FRG1	1043	19	2	0.1335	Missense	May have a role in processing of pre-rRNA or in the assembly of rRNA into ribosomal subunits. May be involved in pre-mRNA splicing
ANKRD36B	5986	19	10	0.1334	Missense	Unknown
OR4C45	919	4	1	0.1319	Missense	Odorant receptor
RGPD2	5502	5	3	0.1298	Missense	Unknown

Genes with scores >0.12 and their function.

Length: Gene length, Carriers: Number of individuals carrying a novel functional variant in the given gene, Novel variants: Number of novel functional variant identified in the given gene, Mutations type: Type of novel functional mutations found in the given gene.

Gene Score

We developed a gene score to rank genes according to the novel functional mutations found in the 19 Jamaican patients with very low disease severity. A total of 18,743 genes were assessed a score. Most genes did not carry damaging mutations and had a gene score of 0 (**Figure 4a**). Nearly half of the genes with the highest scores had no clear known biological function (**Table 6**). Amongst the highest scoring genes with known functions was *GPX1,* which encodes a glutathione peroxidase 1. Since this enzyme protects the erythrocytes against oxidative stress and vascular damage by free radicals is implicated in SCD severity, it was an obvious candidate. The signal in *GPX1* was caused by one single mutation, rs1050450, a missense mutation inducing a proline to leucine modification at residue 200. This variant is highly conserved across species, with a GERP++ score[304] of 4.85. Although this variant was previously known, it was not filtered out from a gene score analysis because it was not present in dbSNP134, used for variant filtering. Although our analysis targeted novel variants, we decided to pursue investigation on this variant in *GPX1* because of its role in erythrocyte protection against oxidative stress.

Association Test between rs1050450 and NEPY in the JSCCS

We genotyped the variant rs1050450 in all of the Jamaican samples for which we had DNA to validate the variant and verify if there was a correlation between rs1050450 genotype and the NEPY. The correlation between rs1050450 genotype and the NEPY in the 76 independent samples did not reach significance (P=0.09) but trended in that

direction i.e. the Leucine allele is associated with reduced disease severity (**Figure 5**). When the 19 sequenced individuals were also included in the analysis (N=95), the p-value was P=0.02.

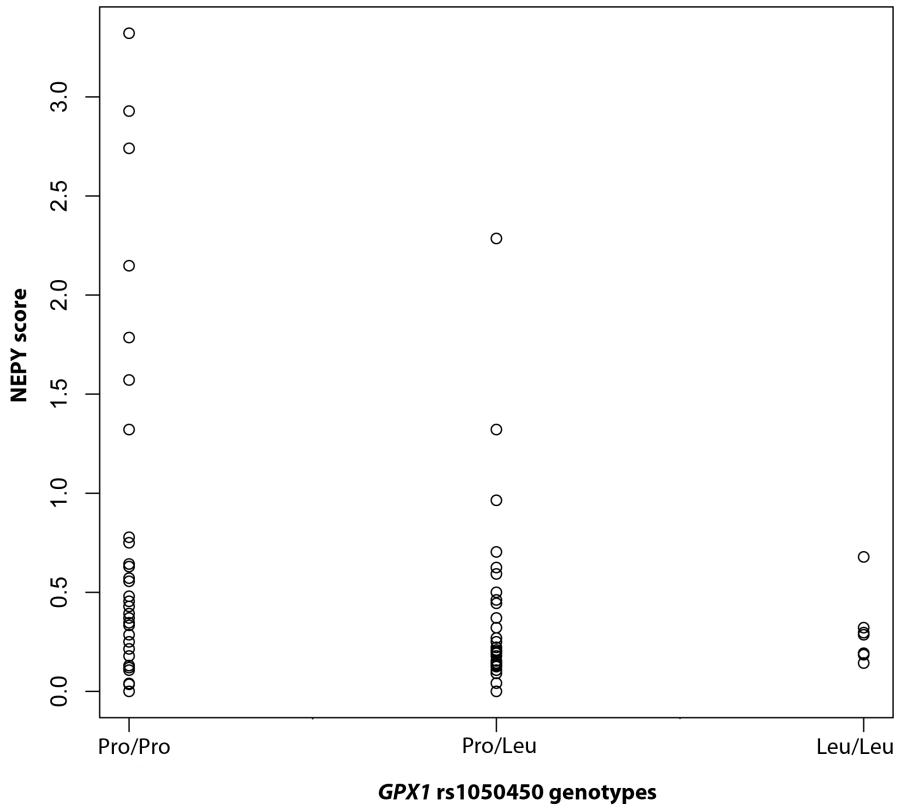


Figure 5 Correlation between NEPY and rs1050450 genotype (n=76)

Association test between the genotype of rs1050450 and disease severity as assessed by the NEPY in 76 independent SCD patients from the JSCCS (P=0.09).

Association Test between rs1050450 Proxy rs9858280 and Clinical Complications in the CSSCD Cohort

To further validate the association between the variant and disease severity, we attempted replication in an additional cohort, the CSSCD. The CSSCD cohort is a

prospective SCD study[236]. The variant rs1050450 was neither genotyped nor imputed on the IBC array, but we found an imputed proxy in perfect LD (r^2=1.0): rs9858280. We tested in 1,313 CSSCD samples the association between rs9858280 and clinical complications such as pain crisis, acute chest syndrome, osteonecrosis, leg ulcer, priapism and stroke. There was no significant association between the SNP rs9858280 and clinical complications in the CSSCD (**Table 7**). There was a slight association (P=0.01) between rs9858280 and HbFz (**Table 7**, HbFz).

Table 7 Association between rs9858280 and clinical complications in the CSSCD

Phenotype	Sample size	Effect (S.E.)	OR [95% CI]	P-value
HbFz	1,213	0.116 (0.045)		0.01
Pain crisis	1,313	-0.057 (0.079)		0.47
Acute chest syndrome	1,313	-0.024 (0.088)		0.79
Leg ulcer	206/1,072	-0.060 (0.140)	0.90 [0.70, 1.16]	0.67
Osteonecrosis	224/1,089	0.127 (0.132)	1.26 [0.99, 1.60]	0.34
Priapism	117/517	-0.199 (0.178)	0.70 [0.51, 0.96]	0.26
Stroke	98/1,215	-0.253 (0.171)	0.63 [0.46, 0.86]	0.14

Association test between the dosage of rs9858280 (T allele) and each clinical complication. The association with fetal hemoglobin z-scores corrected for age and sex was tested by linear regression with the first ten principal components as covariates. Pain crisis and acute chest syndrome were tested using Poisson regression. Effect sizes for pain crisis and acute chest syndrome represent the effects on rates per year. Leg ulcer, osteonecrosis, priapism and stroke were tested with a logistic regression. Sex, age and the ten first principal components were included as covariates for all complications, except for priapism, for which sex was not included. Analyses for all complications were stratified based on α-thalassemia status and association results were combined by inverse variance meta-analyses

S.E. Standard error, OR odds ratio, CI confidence intervals.

We also tested the association between rs9858280 and clinical complications in the CSSCD by performing a regression where the dependent variable was the rs9858280

genotype and the clinical complications were the explanatory parameters. The association was not significant (P=0.36).

Whole-Exome Re-Sequencing Quality of the Fifty Nigerian Samples

We later gained access to the whole-exome sequencing data of 50 Nigerians recruited for a hypertension study. Since the variant calling accuracy increases with the number of samples included, we performed a new variant calling round with the 50 Nigerian and the 19 sickle cell Jamaican samples. We then performed a case-control study with a logistic regression. The mean coverage for the Nigerian samples was 69.9X and 71% of the targeted bases were covered at least at 20X and 68% of the reads were on target (**Table 8**). The reads statistics were very similar for the Jamaican samples between the first and the second rounds of reads alignment. The mean coverage for the Jamaican samples following the second round of reads alignment was 134.2X and 85% of the regions targeted were covered at least at 20X (**Table 9**). Fifty-seven percent of the reads aligned to the targeted regions (**Table 9**). We identified 63,382 non-synonymous variants in the variant calling including both the Nigerian and Jamaican samples with approximately 16% of these not present in dbSNP135.

Table 8 Mean coverage and percentage of reads on target per sample (Nigerians)

Sample id	NEPY	Mean coverage	% 20x coverage	% on target
55	0.07	143	85.4	56.7
150	0.07	148	86.4	58.7
524	0.07	143	83.8	56.9
850	0.04	136	85.4	56.7
1134	0.07	152	86.2	60.6
1338	0.07	137	85.4	57.2
1398	0.08	108	85.8	58.5
1614	0	119	85.7	59.3
1743	0.04	117	81.4	58.5
1893	0	135	84.7	55.7
2023	0.04	117	80.7	55.2
2263	0.08	148	85.7	56.3
2288	0	126	84.7	56.1
2312	0.04	156	85.8	60.0
2320	0.08	135	86.2	57.3
2819	0.08	143	85.7	58.5
2820	0.07	143	86.0	57.9
2845	0.04	116	81.2	51.4
2857	0.08	127	86.5	54.6
Mean	0.05	134.2	84.9	57.2

Table 9 Mean coverage and percentage of reads on target per sample (Jamaicans)

Nigerian Sample id	Mean coverage	% 20x coverage	% on target
1	66	70.7	79.1
2	48	45.0	81.3
3	78	73.2	80.4
4	79	73.8	77.2
5	71	72.0	82.7
6	84	74.3	81.2
7	82	74.8	84.2
8	45	67.5	80.8
9	74	73.3	79.3
10	80	73.9	79.2
11	43	66.4	79.6
12	66	71.4	76.5
13	74	73.5	79.8
14	74	74.0	79.4
15	78	72.6	79.8
16	79	73.7	80.7
17	72	73.0	79.8
18	92	75.4	79.9
19	79	74.8	79.4
20	82	74.5	80.4
21	71	73.4	81.5
22	75	74.5	80.0
23	75	74.6	79.2
24	73	73.9	80.3
25	87	76.7	84.0
26	98	77.2	81.3
27	98	77.1	78.1
28	105	77.4	78.2
29	97	77.7	82.4
30	101	77.6	78.4
31	91	76.3	75.9
32	94	77.9	81.1
33	93	76.7	77.8
34	104	77.6	81.6
35	96	77.6	81.0
36	92	77.6	79.8
37	37	65.2	30.4
38	37	65.2	30.6
39	35	62.5	30.7
40	36	63.9	29.8
41	38	65.4	30.5
42	38	64.9	30.3
43	38	64.9	30.3
44	36	64.2	28.5
45	35	62.6	31.1
46	38	65.3	30.8
47	37	65.0	30.2
48	39	65.7	30.4
49	78	73.3	79.5
50	77	73.0	79.7
Mean	69.9	71.5	68.1

Table 10 Variants identified for each variant calling step (JAM+NIG)

Annotation	Total	Known	New
Nonsense	681	458	223
Missense	62,459	52,843	9,616
Splice site	242	127	115
Synonymous	60,624	55,260	5,364
5' UTR	28,479	19,119	9,360
3' UTR	92,450	66,312	26,138
Intron	862	641	221
Near gene 5' (2000 bp)	182	128	54
Near gene 3' (500 bp)	65	53	12
Intragenic	246,044	194,941	51,103
Intergenic	844	638	206
Total	**246,888**	**195,579**	**51,309**

Variants annotation was performed with dbSNP135.

Case-Control Analysis between Jamaican Extremely Mild Sickle Cell Patients and Nigerian Controls

We performed a logistic regression on variants with MAF>5% including sex as a covariate. Logistic regression association results showed deflation compared to what would be expected (**Figure 6**). This deflation is probably due to the limited number of samples in our analysis. No SNP reached genome-wide significance (**Table 11**). Many SNPs in the β-globin locus were found within the most associated SNPs (**Table 11**), which is expected since the main difference between the two groups of individuals is that the one is composed of SCD patients but not the other. The variant that showed the highest association, chr1: 26801762, is a very rare but known variant (rs199694531).

This variant, located in the gene *HMGN2,* was enriched in the Nigerian samples. Although it is a known variant, this variant was not observed in the 1000 Genomes Project samples[302]. Since we are mostly interested in variants enriched in the Jamaican samples, we discarded this extremely rare variant that seemed to be enriched in the Nigerian samples. The second most associated variant, located at chr22: 30184798 is also a known variant, rs1048001, not observed in the 1000 Genomes Project samples[302]. This variant seemed to be enriched in the Jamaican individuals with very low NEPY (MAF=39% in the Jamaican samples compared to MAF=2% in the Nigerian samples). We were not able to validate this variant by Sanger sequencing, and it appears that the variant is a false positive that could be due to the low coverage (all samples ≤ 10X, mean of 5X) of this base.

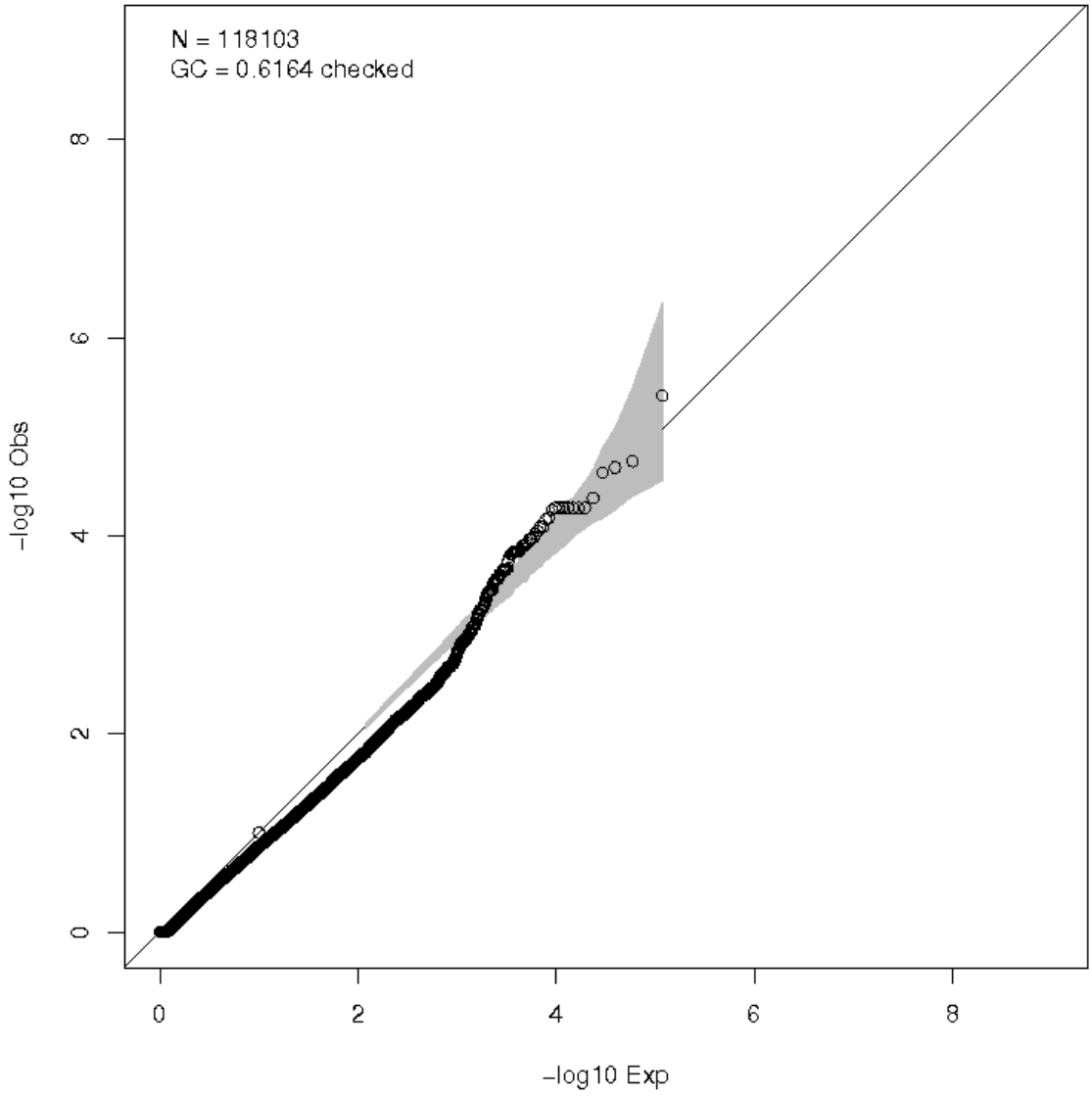


Figure 6 Association results with logistic regression in common variants QQ-plot

Quantile-quantile plot for the logistic regression on variants with MAF>0.05 in the combined dataset. The gray area corresponds to the 90% CI.

Table 11 Association results with logistic regression

CHR	BP	Ref	Alt	N	OR	SE	P	Annotation description	Gene	Gene function
1	26801762	G	T	68	0.002	1.304	$3.9x10^{-6}$	3'-UTR	HMGN2	Binds to the inner side of the nucleosomal DNA thus altering the interaction between the DNA and the histone octamer.
22	30184798	C	T	66	140.8	1.153	$1.8x10^{-5}$	3'-UTR	ASCC2	Enhances NF-kappa-B, SRF and AP1 transactivation.
11	5221233	G	A	68	22.2	0.728	$2.1x10^{-5}$	Missense	OR51V1	Odorant receptor.
11	4869792	A	G	68	21.1	0.720	$2.3x10^{-5}$	Missense	OR51S1	Odorant receptor.
11	4869649	G	A	68	0.055	0.708	$4.2x10^{-5}$	Missense	OR51S1	Odorant receptor.
11	5344552	C	T	68	11.4	0.601	$5.2x10^{-5}$	3'-UTR	OR51B2	Odorant receptor.
11	5344561	G	A	68	11.4	0.601	$5.2x10^{-5}$	3'-UTR	OR51B2	Odorant receptor.
11	5344772	T	C	68	11.4	0.601	$5.2x10^{-5}$	Synonymous	OR51B2	Odorant receptor.
11	5344847	A	G	68	11.4	0.601	$5.2x10^{-5}$	Synonymous	OR51B2	Odorant receptor.
11	5344902	C	G	68	11.4	0.601	$5.2x10^{-5}$	Missense	OR51B2	Odorant receptor.
11	5345170	A	G	68	11.4	0.601	$5.2x10^{-5}$	Missense	OR51B2	Odorant receptor.
11	5345486	G	A	68	11.4	0.601	$5.2x10^{-5}$	Synonymous	OR51B2	Odorant receptor.
17	17395057	G	C	54	37.8	0.901	$5.5x10^{-5}$	3'-UTR	MED9	Component of the Mediator complex, a coactivator involved in the regulated transcription of nearly all RNA polymerase II-dependent genes. Mediator functions as a bridge to convey information from gene-specific regulatory proteins to the basal RNA polymerase II transcription machinery.
8	144993377	A	G	68	21.3	0.767	$6.6x10^{-5}$	Synonymous	PLEC	Interlinks intermediate filaments with microtubules and microfilaments and anchors intermediate filaments to desmosomes or hemidesmosomes. Could also bind muscle proteins such as actin to membrane complexes in muscle. May be involved not only in the filaments network, but also in the regulation of their dynamics. Structural component of muscle.
17	17395055	T	A	55	36.0	0.900	$6.9x10^{-5}$	3'-UTR	MED9	Component of the Mediator complex, a coactivator involved in the regulated transcription of nearly all RNA polymerase II-dependent genes. Mediator functions as a bridge to convey information from gene-specific regulatory proteins to the basal RNA polymerase II transcription machinery.
11	4944661	C	G	68	18.2	0.737	$8.2x10^{-5}$	Missense	OR51G1	Odorant receptor.
11	4944790	C	T	68	18.2	0.737	$8.2x10^{-5}$	Synonymous	OR51G1	Odorant receptor.
11	4936401	G	A	68	8.5	0.548	$9.0x10^{-5}$	Missense	OR51G2	Odorant receptor.
18	70209321	C	A	61	16.6	0.720	$9.5x10^{-5}$	Synonymous	CBLN2	Unknown
11	5373646	T	C	68	0.109	0.572	$1.0x10^{-4}$	Synonymous	OR51B6	Odorant receptor.

SNPs with p-values < $1x10^{-4}$ in the logistic regression (variants with MAF>0.05 in the combined dataset).

CHR: chromosome, BP: base pair, Ref: reference allele, Alt: alternate allele, N: number of individuals with non-missing genotype, OR: odd ratio, SE: standard error, P: p-value

In addition to the logistic regression on common variants, we also performed two rare variant analyses: a SKAT optimal analysis and a burden test. For both analyses, no gene reached the significance threshold of $\alpha=4x10^{-6}$ when applying a Bonferroni correction for the number of genes tested (**Table 13** and **Table 14**). Five genes had p-values <0.001 in both analyses: *AKR7A2*, *ZNF696*, *ZNF575*, *RELT* and *TMEM187*.

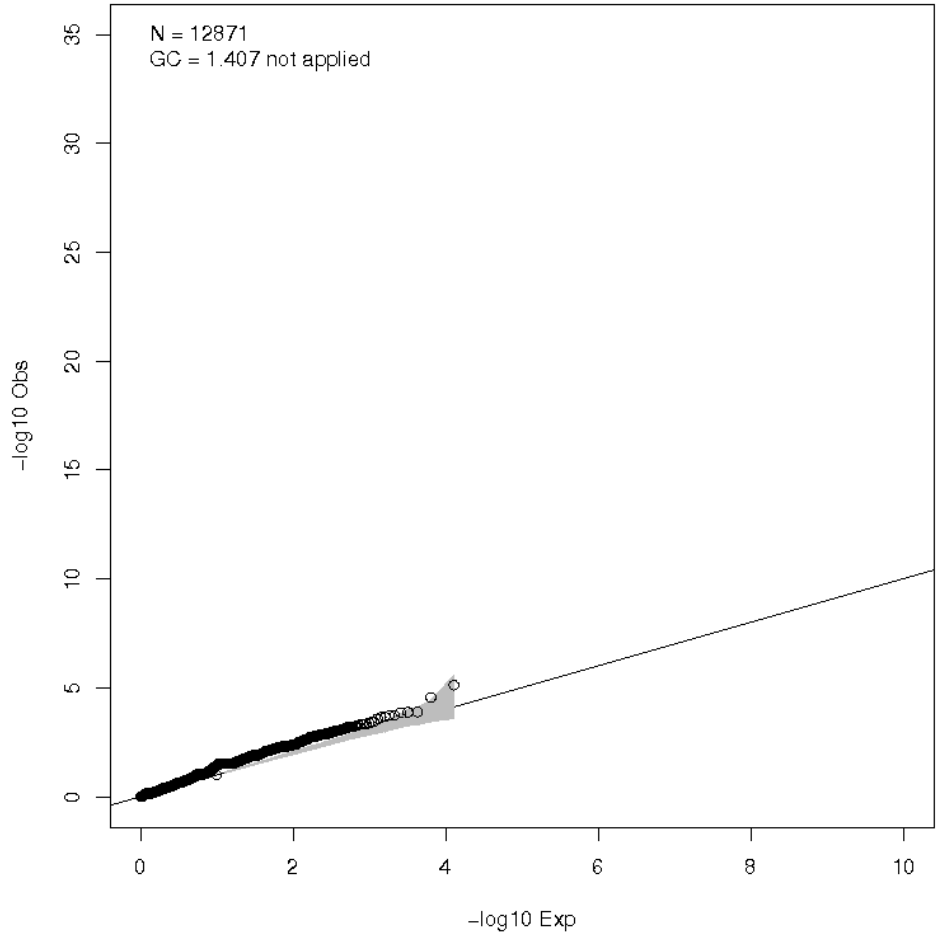


Figure 7 Rare variants analysis with SKAT-O QQ-plot

Quantile-quantile plot for the SKAT-O analysis on rare functional variants with MAF<0.05. The gray area corresponds to the 90% CI.

Table 12 Rare variants analysis with SKAT-O ($p<1x10^{-3}$)

Gene	Variants	P-value	Gene function
ZNF696	2	$7.6x10^{-6}$	May be involved in transcriptional regulation.
AKR7A2	3	$2.8x10^{-5}$	Catalyzes the NADPH-dependent reduction of succinic
AATK	4	$1.3x10^{-4}$	May be involved in neuronal differentiation.
ISLR2	3	$1.3x10^{-4}$	Required for axon extension during neural development
ZNF575	2	$1.4x10^{-4}$	May be involved in transcriptional regulation.
TEP1	27	$1.9x10^{-4}$	Tumor suppressor. Acts as a dual-specificity protein phosphatase,
ZNF497	9	$1.9x10^{-4}$	May be involved in transcriptional regulation.
MKL1	6	$2.1x10^{-4}$	Transcriptional coactivator of serum response factor (SRF) with
SIX6	1	$2.3x10^{-4}$	May be involved in eye development.
TLR6	6	$2.7x10^{-4}$	Participates in the innate immune response to Gram-positive
RELT	4	$3.7x10^{-4}$	Mediates activation of NF-kappa-B. May play a role in T-cell
ASPDH	2	$3.8x10^{-4}$	Specifically catalyzes the NAD or NADP-dependent
SLC39A4	2	$4.2x10^{-4}$	Plays an important role in cellular zinc homeostasis as a zinc
TMEM187	3	$4.6x10^{-4}$	Encodes a multi-pass membrane protein.
ZNF697	2	$4.7x10^{-4}$	May be involved in transcriptional regulation.
SEPT14	3	$4.7x10^{-4}$	Filament-forming cytoskeletal GTPase (By similarity). May play a
SVEP1	26	$4.9x10^{-4}$	May play a role in the cell attachment process.
PTPRH	13	$5.3x10^{-4}$	May contribute to contact inhibition of cell growth and motility by
SAFB2	3	$5.4x10^{-4}$	Binds to scaffold/matrix attachment region (S/MAR) DNA. Can
STOML3	2	$5.9x10^{-4}$	NA
SLC5A5	3	$6.1x10^{-4}$	Mediates iodide uptake in the thyroid gland.
FNDC9	2	$6.2x10^{-4}$	Unknown
SP110	5	$6.3x10^{-4}$	Transcription factor. May be a nuclear hormone receptor
PRRC2A	9	$6.5x10^{-4}$	May play a role in the regulation of pre-mRNA splicing.
C1QTNF8	1	$6.7x10^{-4}$	Unknown
SH2B3	6	$7.3x10^{-4}$	Links T-cell receptor activation signal to phospholipase C-gamma-
HSPBAP1	5	$7.3x10^{-4}$	Enhances STAT6-dependent transcription (By similarity). Has
ABCB1	8	$7.8x10^{-4}$	May mediate critical mitochondrial transport functions related to
WFDC10B	1	$8.0x10^{-4}$	This gene encodes a member of the WAP-type four-disulfide core
ZNF462	7	$8.6x10^{-4}$	May be involved in transcriptional regulation.
MYOM2	22	$8.9x10^{-4}$	Major component of the vertebrate myofibrillar M band. Binds
HLA-F	3	$9.0x10^{-4}$	Involved in the presentation of foreign antigens to the immune
SLC22A4	3	$9.1x10^{-4}$	Sodium-ion dependent, low affinity carnitine transporter. Probably
PPIG	3	$9.1x10^{-4}$	PPIases accelerate the folding of proteins. It catalyzes the cis-trans
IBSP	2	$9.8x10^{-4}$	Binds tightly to hydroxyapatite. Appears to form an integral part of
KCNJ15	1	$9.8x10^{-4}$	Inward rectifier potassium channels are characterized by a greater

Functional rare variants analysis with SKAT-O including variants inducing a missense, nonsense or splice site mutation with MAF <0.05.

Table 13 Rare variants analysis with burden test ($p<1x10^{-3}$)

Gene	Variants tested (N)	Non-carriers in controls (N)	Carriers in controls (N)	Non-carriers in cases (N)	Carriers in cases (N)	Missing genotype (N)	P-value	Gene function
AKR7A2	3	47	2	10	9	0	$7.4x10^{-5}$	Proton-linked monocarboxylate transporter. Catalyzes the rapid transport across the plasma membrane of many monocarboxylates such as lactate, pyruvate, branched-chain oxo acids derived from leucine, valine and isoleucine, and the ketone bodies acetoacetate, beta-hydroxybutyrate and acetate.
ZNF696	2	46	0	11	6	5	$1.8x10^{-4}$	May be involved in transcriptional regulation.
ZNF575	2	44	0	11	6	7	$2.2x10^{-4}$	May be involved in transcriptional regulation.
PTPRH	13	44	5	9	10	0	$4.4x10^{-4}$	May contribute to contact inhibition of cell growth and motility by mediating the dephosphorylation of focal adhesion-associated substrates and thus negatively regulating integrin-promoted signaling processes. Induces apoptotic cell death by at least two distinct mechanisms: inhibition of cell survival signaling mediated by PI 3-kinase, Akt, and ILK and activation of a caspase-dependent proapoptotic pathway.
NOTCH1	11	35	9	6	13	5	$4.5x10^{-4}$	Functions as a receptor for membrane-bound ligands Jagged1, Jagged2 and Delta1 to regulate cell-fate determination. Upon ligand activation through the released notch intracellular domain (NICD) it forms a transcriptional activator complex with RBPJ/RBPSUH and activates genes of the enhancer of split locus. Affects the implementation of differentiation, proliferation and apoptotic programs. May be important for normal lymphocyte function.
RELT	4	45	4	10	9	0	$6.8x10^{-4}$	Mediates activation of NF-kappa-B. May play a role in T-cell activation.
TMEM187	3	48	0	13	5	2	$9.6x10^{-4}$	Encodes a multi-pass membrane protein.

Functional rare variants analysis including variants inducing a missense, nonsense or splice site mutation with MAF <0.05. Individuals of each population were separated in two groups: carriers and non-carriers of rare functional variants to perform a Fisher test.

5.5 Discussion

SCD patients tend to suffer from complications such as pain crisis, acute chest syndrome, osteonecrosis and stroke and require frequent hospitalizations. Few SCD patients have rare or no complications and have a mild disease status. The exceptional phenotype of these extremely mild SCD patients could be due to genetic variants. In this study, we intended to identify novel genetic variants modifying SCD severity by performing whole-exome sequencing on 19 SCD patients from the JSCCS with extremely few complications.

Our first approach consisted of building a gene score to rank the genes according to the number of carriers of novel and potentially functional mutations for each gene. To predict the protein alteration induced by missense variants, we used the PolyPhen-2 score, which is based on both physical-chemical properties and conservation[301]. We focused our analysis on the novel variants and used the database dbSNP134 to perform the filtering of known variants. Since the variants remaining in the analysis would be novel variants, we hypothesized that they would be rare and that the chance of finding two distinct novel variants on a same allele of a gene was low. For each individual, we included the two highest PolyPhen-2 scores for novel mutations in this gene and assumed they are on different alleles of the gene. As longer genes are more likely to carry novel mutations, the gene score included a correction for the gene length. One of the limitations of this correction in the gene score calculation is that it does not take into account that the mutation rate is not homogenous across the genome[5].

We built the gene score with the aim of ranking genes according to the novel functional mutations they contained. Half of the 24 genes with the highest scores, including the six top genes, had unknown functions. Additional follow-up for many high-scoring genes will be necessary. One of the genes with a high score with a known function was *GPX1*, which encodes for a glutathione peroxidase 1. This enzyme protects the erythrocytes against oxidative stress. Since vascular damage caused by free radicals contributes to SCD severity, *GPX1* appeared like a promising candidate as a modifier of disease severity. The signal in *GPX1* was caused by one variant, rs1050450, inducing a proline-to-leucine change at position 200 in the protein sequence. Although this variant was not present in dbSNP134, it turned out to be common in individuals of African descent in the 1000 Genomes Project[302]. We observed an enrichment of this variant in the 19 sequenced Jamaican samples, in which the frequency was 47%, compared to minor allele frequencies between 25% and 34% in the populations of African ancestry of the 1000 Genomes Project[302]. Even though the mutation was known, we attempted replication by genotyping the variant in the Jamaican individuals for which we had DNA samples to test if there was a correlation between the variant rs1050450 and the NEPY. Since rs1050450 is located in a region sharing high homology with two other regions located on chromosome 21 and chromosome X, we performed a nested PCR to genotype the variant. The p-value of the association between the genotypes at rs1050450 and the NEPY was P=0.09 in the 76 additional Jamaican SCD patients but trended in the right direction (**Figure 5**). When we included the 19 individuals selected for whole-exome sequencing along with the 76 independent samples (N=95), the p-value was P=0.02. We attempted to replicate the association between the variant rs1050450 and SCD severity

in the CSSCD cohort with a proxy, rs9858280, that was imputed on the IBC array. The alternate allele of rs1050450 seems to be slightly associated with an increase of fetal hemoglobin (P=0.01), which could reduce disease severity. There was no significant association in the CSSCD between the proxy rs9858280 and the clinical complications when tested individually.

The gene *GPX1* remains somewhat interesting and further analyses will be needed before being able to draw final conclusions. This gene has been identified as an interesting candidate based on previous literature. Oxidative stress is caused by reactive oxygen species. Due to their role as oxygen carriers, erythrocytes are constantly under oxidative stress. Hydrogen peroxide (H_2O_2) oxides hemoglobin to methemoglobin. *GPx1* protects erythrocytes against oxidative damage and oxidative hemolysis caused by H_2O_2 by catalyzing the oxidation of reduced glutathione by hydrogen peroxide molecules into water and dioxygen, involving in the reaction two GSH molecules as electron donors[305]. Individuals with at least one alternate allele at rs1050450 are expected to show a higher relative increase in the enzyme activity[306]. It has been suggested that variants in *GPX1* interact with the sickle cell mutation (HbS) through epistasis[307]. Cho et al[308] showed that there was a 33% decrease in *GPx1* activity in red blood cells of patients with SCD and that *GPx1* loss of activity was correlated with hemolysis in SCD patients. Also, red blood cells from sickle patients treated with hydroxyurea treatment showed 90% higher *GPx1* activity compared to red blood cells from untreated SCD patients[308]. It is unclear whether rs1050450 is a false positive and further analysis will be needed.

The phenotypes tested to validate the association with rs1050450 or its proxy in the JSCCS and CSSCD were different. In the JSCCS, we tested the association between rs1050450 and the NEPY, a measure of severity of the disease based on the average number of occurrences of severe complications per year during the first 18 years of life. In the CSSCD, we separately tested the association between the proxy rs9858280 and each clinical complication. For pain crisis and acute chest syndrome, we used the average complication rate per year during the time of follow-up. For osteonecrosis, stroke, priapism and leg ulcer, we performed a case-control analysis. To test an association closer to disease severity in the CSSCD, we performed a reverse regression, where the dependent variable was rs9858280 dosage, and the clinical complications were the predictors. This association was not significant either. Although both JSCCS and CSSCD cohorts have detailed phenotypic information, the lack of association of rs9858280 with clinical complications in the CSSCD could be due to the difference between the phenotypes tested in the two cohorts. Another possible explanation for this lack of replication could be due to a gene-environment interaction or to the ethnic differences between the two populations. It is also possible that the variant is a false positive.

The occurrence of some of the SCD complications is age-specific. Since the Jamaican cohort is a birth cohort, this was taken into account when building the severity score. When testing the association with an individual complication, age-specific effects are considered by correcting for age. However, it would not be as feasible to try to build a severity score comparable to the NEPY in the CSSCD since the age of recruitment is variable among patients. Also, the average follow-up time for participants was much

shorter in the CSSCD (6.6± 1.6 years) than in the JSCCS, where the NEPY was established based on the complications developed during the first 18 years of the participants' life. Therefore, a severity score in the CSSCD would most likely not be as robust as the NEPY in the JSCCS.

Jamaican Genetic Structure

In our case-control analysis, we compared our Jamaican SCD patients with the Nigerians recruited for a hypertension study. We did not have access to Jamaican controls that had already been sequenced. Historically, 90.4% of slaves arriving in Jamaica were brought from West Africa[309]. However, over the years, the Jamaican population became an admixed population, with genetic input from both European and East Asian populations[310]. Simms et al. estimated that the Jamaican admixture was between 17.9% and 23.5%[310]. In the absence of Jamaican controls, the Nigerians seemed to be a good control group for a case-control study with Jamaicans. It is unlikely, but possible, that this admixture in the Jamaican population would create stratification in the case-control study between the Jamaicans and Nigerians. Genotyping the rest of the Jamaican cohort would enable us to identify such a situation.

Case-Control Analysis

We performed a logistic regression on variants with MAF above 5% in the combined dataset. Many variants located in proximity to the β–globin locus showed association (**Table 11**). Since the main difference between the two groups of individuals is that the

Jamaicans have SCD, these variants can be considered as positive controls. The reason why the main mutation causing SCD, rs334, is not the most highly associated variant in the β-globin locus is that 12 individuals of the Nigerian cohort carry the sickle cell trait (that is, they are heterozygous for the mutation). The variant showing the highest association in the case-control study of common variants at chr1: 26801762 ($P=3.9x10^{-6}$) is a known variant, rs199694531. This variant is homozygous for the reference allele in the samples of the 1000 Genomes Project[302] and hence, very rare. Since this rare mutation is only observed in the Nigerian samples, we do not consider it a variant capable of explaining the mild disease status of the 19 Jamaicans sequenced. The second most associated variant, at chr22:30184798, is also a known very rare variant, rs1048001. The alternate allele is not present in the 1000 Genomes Project samples[302]. The sequencing data suggested an enrichment in the Jamaican samples with a MAF=39%, compared to a MAF=2% in the Nigerian samples. This variant is located in the gene *ASCC2*. This gene is a transcription factor that is thought to play an essential role in *AP-1*, *SRF* and NF-kappaB transactivation[311]. *ASSC2* is also thought to play a role in the transrepression between the nuclear receptors and either *AP-1* or NF-kappaB[311]. We attempted to validate this variant with Sanger sequencing, but it did not validate and appears to be a sequencing false positive.

This study was designed at the beginning of the second generation sequencing era when it was expected that many of the variants causing Mendelian syndromes could be identified with whole-exome sequencing. Since 2009, whole-exome sequencing has identified causal genes/mutations for numerous Mendelian syndromes, including Miller

syndrome[208], Schinzel-Giedion syndrome[209], terminal osseous dysplasia[210], nonsyndromic hearing loss DFNB82[211], ovarian dysgenesis, hearing loss, and ataxia of Perrault syndrome[212], brain malformations[213], Sensenbrenner syndrome[214] and hyperphosphatasia mental retardation syndrome[215]. Most of these studies focused on novel functional variants shared between affected individuals. When designing this study, we expected that extremely mild cases of SCD could be due to few variants with high penetrance. Since environmental factors tend to increase disease severity, we expected to see more heterogeneity in individuals with extremely high severity. We also expected to have a better power by sequencing more individuals with extremely few complications than by sequencing half this number at both ends of the distribution. Although whole-exome sequencing has led to successful identification of causal variants for Mendelian diseases, whole-exome sequencing success stories for complex traits are not as frequent. Retrospectively, our number of sequenced samples might have been sufficient if our trait had been completely Mendelian, but insufficient for a complex trait. Overall, given the small number of samples, our study had limited power. Indeed, even for a variant with MAF=25% with an odds ratio of 3 the power is still below 1% (under a log-additive model and when a bonferroni correction is applied for the number of variants tested $\alpha=2.0x10^{-7}$). No SNP reached genome-wide significance in the case-control analysis. Nor did any gene reach the significance threshold when applying a Bonferroni correction for the number of genes tested in the two types of rare variants analyses that we performed (SKAT optimal analysis and a burden test). As for the gene score, it is hard to establish a score cut-off. Our results do offer prioritization for potential candidate genes modifying SCD severity.

In conclusion, we performed the whole exome sequencing of 19 extremely mild SCD patients to identify potential variants modifying SCD severity. However, due to the limited size of our cohort, no SNP or gene reached significance. The variant rs1050450 in the gene *GPX1* appeared to be a promising candidate but did not replicate in the CSSCD. This lack of replication could be due to the difference between the phenotypes tested in the CSSCD and in the JSCCS. Additional analysis will be needed to validate this variant.

5.6 Acknowledgments

We would like to thank all the individuals who participated in this study, as well as Gabrielle Boucher for statistical advice. This work was graciously funded by The Doris Duke Charitable Fundation. G. Galarneau is the recipient of a scholarship from Fonds de recherche du Québec–Santé.

Chapter 6: Discussion

Many topics could be discussed following the results in the four projects presented in this thesis. In this section, I will focus on four: the implication of our results, the future of genetic studies in complex traits, the future of genetic studies on sickle cell disease (SCD) clinical complications and the prospective of an applicable cure for SCD in the next decade.

6.1 Implication of our Results

6.1.1 Study Limitations

The projects presented in this thesis have some limitations. For example, for the gene-centric association studies on fetal hemoglobin (HbF) levels (Chapter 4), pain crisis and acute chest syndrome (Chapter 3), the samples were genotyped on the ITMAT-Broad-CARe (IBC) array, focusing on 2,100 genes related to heart, lung and blood diseases. Our analyses did not cover the entire genome. Therefore, there might be genetic associations not found in our studies due to the coverage of the IBC array that could have been found using a genome-wide array. Also, using the genotypes from a genome-wide array instead of the IBC array in the gene-set enrichment analysis, and hence testing all the genes in the human genome, might have helped to better discern the causal pathway from the seven that replicated in our analysis. Nevertheless, given that ischemia, vaso-constriction and adhesion molecules in the blood all contribute to SCD complications, the IBC array constitutes a good genotyping array to test genetic association with SCD complications and HbF levels, especially when considering that our statistical power to detect new signals associated with these traits was limited. Using the IBC array instead of a genome-

wide array, we increased our statistical power by reducing our significance threshold due to multiple testing (from $\alpha=5x10^{-8}$ [206] for a genome-wide array to $\alpha=2x10^{-6}$ [254] for the IBC array).

Our analysis of the whole-exome sequences of patients with extremely mild SCD was limited by the small number of sequenced patients and by the lack of Jamaican controls. If we had found a genetic association between the Jamaican patients and the Nigerian controls, this association could have been due to population stratification and further analysis of the variant would have been required. Finally, our hypothesis assumed that the patients with extremely mild SCD would be carriers of a rare coding variant. However, it is possible that the extremely mild SCD phenotype is due to non-coding variants or to common variants.

6.1.2 Comprehension on Sickle Cell Disease Severity

The fine-mapping of *BCL11A*, *HBS1L-MYB* and the β-globin locus (Chapter 2) helps to better understand the regulation of HbF levels at these loci. First, we now know that multiple SNPs are independently associated in both *BCL11A* and *HBS1L-MYB*. Additionally, the fine-mapping of the *HBS1L-MYB* region implicates the gene *MYB* in the regulation of HbF levels. We also show that the variant *XmnI* is not the causal variant in the β-globin locus as previously thought.

Results from both projects on HbF levels (Chapter 2) and (Chapter 4) indicate that additional genetic factors could be regulating HbF levels through epigenetics mechanisms. *MYB* is a transcriptional regulator of erythropoiesis. Both *BCL11A* and *MYB* are transcription factors and epigenetics modifications could affect their regulation of HbF levels. Also, our gene-set enrichment analysis (Chapter 4) suggests an association between HbF and genes implicated in acetyl-coA and short fatty acids metabolism and we observe a negative correlation between β-hydroxybutyrate and HbF levels. Since β-hydroxybutyrate acts as a histone deacetylase (HDAC) inhibitor, the genes showing moderate associations in these pathways could influence HbF levels through an epigenetic effect. Our results also provide the first evidence that metabolites in endogenous levels could influence HbF levels and show that metabolomics studies with HbF levels might uncover additional regulators of HbF levels.

Regarding the genetic association study with clinical complications (Chapter 3), the variant rs6141803 is the only variant that reached array-wide significance when combining the association results. This variant is located in the intergenic region of *DNMT3B* and *COMMD7* and shows association with acute chest syndrome. Interestingly, *COMMD7* expression in human pulmonary endothelial cells is modulated by the presence of free heme, a contributing factor to the SCD clinical complications. Our association studies with pain crisis and acute chest syndrome (Chapter 3) and our analysis of the whole-exome of mild SCD patients (Chapter 4) provide good candidate genes that might modify SCD severity: *FAM193A*, *PLA2G4A*, *COMMD7* and *GPX1*. Additional analyses will be necessary to validate these candidates.

6.1.3 Missing Heritability

One of the main questions following the first genome-wide association studies (GWAS) regarded "the missing heritability". The missing heritability represents the proportion of the estimated heritability that GWAS results did not explain. Our fine-mapping study of the three loci previously associated with HbF levels (Chapter 2) show that part of this missing heritability can be found by fine-mapping the loci identified in GWAS.

The remaining missing heritability is probably due to many factors. It could be due to an overestimate of the heritability. As currently calculated, the gene-environment interactions and the gene-gene interactions are not properly considered in the heritability calculations and inflate the heritability estimates. Additional missing heritability could possibly be found in variants not covered by genome-wide arrays such as rare variants and structural variants like deletions, duplications and inversions. Finally, mechanisms such as allelic expression, non-coding RNA and methylation might also explain a portion of the missing heritability.

6.2 Future Genetic Studies in Complex Traits

The majority of current genetic studies strictly rely on single-variant association tests. As seen in the results of this thesis, one big limitation for GWAS in complex traits is the genome-wide significance threshold and the sample size required to reach this significance threshold. The combination of results from multiple studies by meta-analysis has led to the identification of novel genome-wide significant loci for many phenotypes. Except when performing these meta-analyses, little follow-up is currently done on variants showing moderate association (not reaching genome-wide significance threshold).

Height is probably the phenotype for which the largest genetic studies have been performed, because it is almost always systemically measured in genetic studies on other traits or diseases. A recent meta-analysis study on height included over 260,000 individuals[312]. Such sample sizes might be reachable for anthropometric traits and traits routinely measured such as blood pressure, heart rate and hematological traits. However for many other traits, and disease phenotypes in particular, it would simply be impossible to reach sample sizes as big as the most recent meta-analysis on height. Each newer and bigger meta-analysis on height still uncovers new associated loci with smaller effect sizes[312,313]. Since such sample sizes are impossible to reach for some traits, we can conclude that for many complex traits, it will not be possible to rely on single-variant association studies alone to identify at genome-wide significance all the associated genetic variants. As shown in our gene-set enrichment analysis (Chapter 4), moderately associated variants can still provide insightful information on potentially implicated loci or pathways. There is a need to develop strategies to integrate known biological information in the analysis of these moderately associated signals. These strategies may consist of gene sets analysis to implicate biological processes or Bayesian approaches and machine-learning methods to prioritize moderately associated loci.

6.2.1 Gene Set Analysis in Genome-Wide Association Studies

The first genome-wide analyses of moderately associated signals integrating biological knowledge in GWAS were gene-set enrichment analyses based on biological pathways[276,314-317]. These analyses test if there is an enrichment of moderately

associated genes involved in a same biological process based on known biological pathways. Over the years, over a dozen of gene-sets or pathway-based methods for GWAS have been developed. Some rely on text-mining[318,319], others perform gene-set analysis[219,320-323], gene-set enrichment based analysis for meta-analysis[324] or gene-set enrichment analysis[275,317,325,326] like GenGen, used in our pathway analysis (Chapter 4). These algorithms differ in the calculation of the gene statistic, in the input data, in the statistical test for pathway association, in whether or not the method takes into account linkage disequilibrium, gene size and pathway size. Gene-sets analysis results will vary depending on the tool chosen for the analysis and the pathway definitions provided.

In addition to gene-set approaches based on biological processes, gene-sets based on eQTLs, co-expression networks or epigenetic modifications in relevant cell types could also be tested when tissues or data are available. Zhong et al[327] integrated information from gene expression studies. They derived SNPs associated with gene expression (eSNPs) sets based on liver and adipose tissue expression studies. These eSNPs sets were integrated in gene-sets defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database and calculated the enrichment of eSNPs in these pathways. Using this method, they identified and replicated novel pathways associated with type 2 diabetes. A study on Parkinson's disease, compared pathway analysis reports between both GWAS and gene expression data and found that four of the seven most significant pathways were shared between the two approaches[328].

6.2.2 Bayesian Approaches in Genome-Wide Association Studies

Bayesian approaches could help to prioritize genes with moderate association based on other types of studies such as gene expression studies to establish priors. Bayesian approaches are already used in genetic association studies to account for the model of inheritance or to account for the imputation quality. The convergent functional genomics approach is so far the main Bayesian approach utilized to identify and prioritize candidate genes in GWAS study. Le-Niculescu et al[329] used a convergent functional genomics approach as a way of mining moderately associated signals in existing bipolar disorder GWAS datasets. They integrated to the GWAS results gene expression data from blood and postmortem brain tissues for both human and model animals. They observed a fourfold enrichment for the top 41 genes in an independent GWAS study. Convergent functional studies have been applied to mood disorders[330], bipolar disorder[331], anxiety disorders[332], chronic fatigue syndrome[333] schizophrenia[334] and autism[335]. The software SNPTEST now enables the user to perform simple genome-wide Bayesian analyses with t-distribution priors[336].

6.2.3 Machine-Learning to Prioritize Candidate Genes

Machine-learning programs consist of algorithms that are first trained to identify the characteristics differentiating categorical data to later be able to classify similar datasets. The first studies using machine-learning algorithms on GWAS datasets aimed to predict disease-status. Type 1 diabetes status can be well predicted with GWAS data using support-vector machine[337,338]. Recent studies suggest that using support-vector machine

to predict disease-status could also help to prioritize candidate genes. Negi et al[338] provided to a support vector machine algorithm the SNPs from their GWAS on rheumatoid arthritis with association p-values ≤ 0.001. This analysis assigns a weight to each SNP based on its predictive value in the model and identified five potential new loci. Most of these new loci are known to be associated with other autoimmune diseases and interact with genes implicated in rheumatoid arthritis. A stand-alone user-friendly Java-based software for machine learning analysis for disease-status prediction has been developed by Mittag et al[339].

6.2.4 Integrative GWAS

The genome, transcriptome, proteome, metabolome are inter-related, and together with environmental elements, can explain observed complex diseases and traits. Many efforts have been made to identify quantitative traits loci associated with gene expression or metabolic traits. GWAS on gene expression have been initiated in many tissues including lymphocytes[340-343], monocytes[344], blood[345-347], liver[348-350], subcutaneous tissue[350], adipose tissue[346,350] and brain[347,351,352]. The first GWAS on metabolic traits[353-361] have been published. Datasets with metabolomic, transcriptomic and genomic data are now available[362]. Initiatives such as the National Institute of Health (NIH) Genotype-Tissue Expression Project (GTEx Project), aiming to identify eQTLs by analyzing gene expression in a multitude of tissues of densely genotyped individuals are also underway. All these initiatives will help to identify a multitude of SNPs that are associated with gene expression and metabolite concentrations and provide insightful information to assess

the function of variants identified by GWAS. Given our association results with HbF levels in the pathway analysis (Chapter 4), GWAS on metabolic traits could be help to prioritize moderate association signals with HbF levels.

We also start to see association studies integrating genomic, transcriptomic and metabolomics in the same dataset for phenotypes such as high-density lipoprotein cholesterol[363] and we should expect to soon see many more of these studies within the next few years. In these studies, the association of moderately associated variants could be corroborated by other evidences such as modifications in gene expression or changes in the metabolism of a metabolite. These studies will require the development of new bioinformatics and statistical tools. Such analyses will, however, raise further questions about multiple testing, as the number of tests would once again increase.

6.2.5 The Importance of Fine-Mapping Studies and Functional Studies

The fine-mapping of the three loci previously associated with fetal hemoglobin (HbF) levels described in Chapter 2 provided one of the answers to a central question that arose from the first GWAS: "Where is the missing heritability?" This project showed that fine-mapping can reveal independent signals in associated loci and can increase the number of heritable variations explained by the loci identified in GWAS. Although fine-mapping studies might be considered less scientifically exciting than conducting new GWAS, they are necessary, especially if newer and bigger GWAS are justified by the fact that the previous ones did not explain all the heritable variations for a given trait.

Functional studies will often remain necessary to identify and/or validate causal variants. Only a minority of GWAS or fine-mapping studies will lead to a single candidate variant with no other variant in strong linkage disequilibrium (LD). In the fine-mapping of known loci associated with HbF levels (Chapter 2), the LD between the variants showing the highest association was still too strong to determine the causal ones. Bauer et al identified the causal variants in the *BCL11A* locus and showed that they are located in a powerful stage-specific, lineage restricted enhancer[364]. The strongest causal variant disrupts a transcription factor binding motif in this enhancer[364].

Given the number of variants currently identified by GWAS, it is clear that we will need large-scale functional studies to validate these variants and loci. One of the promising strategies to perform these large-scale functional studies would be the recently developed genome engineering using CRISPR/Cas9 system[365,366]. This method is a precision multiplex genome-engineering tool for mammalian genomes. CAS9 is used to induce DNA double-strand breaks at specific loci surrounding the targeted DNA sequence through a single-guide RNA. Once the target DNA sequence is cleaved, this method can leave the deletion created as is, insert a mutated version of the sequence cleaved or insert another sequence. By introducing frame shifts insertions or deletions, the CRISPR/Cas9 system can create a loss-of-function allele. As opposed to RNAi, the CRISPR/Cas9 system can also target regions across the entire genome, including intergenic regions, introns, enhancers and promoters, permitting a genome-scale knockout screening[367].

As seen in the gene score analysis of the whole-exome sequences of mild SCD patients (Chapter 5), it is hard to prioritize genes showing moderate association if half of them have unknown function. Approximately 41% of genes in the human genome still have unknown functions[2], so functional studies will remain fundamental to understand how associated genes disturb or modify the biological processes leading to observed phenotypes. These functional studies could also uncover interacting proteins that are only moderately associated with the phenotype. Also, as gene-set enrichment analyses results are dependent on the functional knowledge of the genes evaluated, it will be important to characterize the genes with unknown function and their biological processes.

6.3 Future Genetic Studies on Sickle Cell Disease Severity

6.3.1 The Feasibility of Creating Larger Sickle Cell Disease Cohorts

Although recruitment in the Cooperative Study of Sickle Cell Disease (CSSCD) ended over 20 years ago, this is still the largest sickle cell disease (SCD) cohort. In addition, the data was collected in individuals before hydroxyurea treatment was applied, providing unbiased statistics on natural incidence and the prevalence of clinical complications. Despite its age, the CSSCD is still relevant because of its size and detailed phenotypic information. In this work, the association study with clinical complications was performed in 1,514 individuals for the discovery phase. Because of the limited power shown by this study, it could be argued that we should create larger SCD cohorts. I do not think that it is feasible at the moment to create such a large and detailed cohort.

Assuming this new cohort would be created in a developed country, the U.S. would be the best location, being the country with the highest number of cases (estimated at between 50,000 and 100,000). A sample size of 7,800 individuals would be necessary to provide an 80% statistical power (under the assumption that the variant explains 0.05% of the phenotype variation with a minor allele frequency of 25% under an additive model for a normally distributed quantitative phenotype such as fetal hemoglobin levels). Such a cohort would require the recruitment in a multi-center study of the equivalent of 7.8%-15.6% of the United States SCD population. To insure an 80% power to validate the results at α=0.05 obtained in the discovery phase, a replication cohort with at least 1,566 individuals would be necessary. The fact that many SCD patients in developed countries are now treated with hydroxyurea could impair association results on clinical complications and HbF levels leading to the identification of loci influencing hydroxyurea response instead. This would be undesirable if the goal of the study is to find novel targets for the treatment of these complications. Hydroxyurea treatment could also create additional variation in the phenotypes studied, reducing the variation explained by the genetic variants when compared to a study of untreated patients like the CSSCD. Complication rates and incidence would probably be decreased in this cohort, meaning a decrease in statistical power.

Many countries where SCD is highly prevalent are low- or middle-income countries. Given the financial and technical resources necessary to create a detailed genetic cohort on SCD clinical complications, creating a high quality cohort in such countries would be difficult. Also, nutrition and infections could play major roles in SCD mortality in these

countries, especially in children. Unfortunately, there are no exact statistics on the mortality rate in children under five years old in Africa. According to the World Health Organization, many of the babies born in Africa with SCD currently die before the age of 5 years-old[145,146], and one study even suggests that this proportion could be as high as 50% in some regions[86]. Data in the U.S. show a dramatic decrease in deaths of young children since the establishment of penicillin prophylaxis administration and *Streptococcus pneumonia* vaccination. For the time being, it would not be advisable to do research to predict the disease severity in countries where SCD screening in newborns is not yet performed. I think that the first step before creating SCD cohorts in countries with high SCD prevalence would be to establish global newborn screening and preventive care programs in these regions. Since 1993, a newborn screening program for SCD have been initiated in Kumasi, Ghana[368] and newborn screening is about to be implemented on a national scale in Ghana[369]. A newborn screening program was also carried out in a hospital in Benin City, Nigeria[370].

6.4 An Applicable Cure in the Next Decade?

The deployment of penicillin therapy and *Streptococcus pneumonia* vaccination in developed countries has greatly reduced the mortality rate in SCD children under 5 years-old, whereas hydroxyurea treatment has helped to reduce the incidence of most clinical complications in SCD patients. However, the average age at death of SCD patients in 2006 was still 39 years-old[149], demonstrating the crucial need for an applicable cure for SCD. I think (and very much hope) that in the next decade, major improvements in

stem cell transplantation or successful clinical trials of gene therapy will provide a realistic cure for the majority of SCD patients.

6.4.1 Stem Cell Transplantation

Major improvement in stem cell transplantation could make this procedure a cure for many SCD patients. Currently, the main stem cell transplantation performed in SCD patients is a bone marrow transplantation from a human leucocyte antigen (HLA)-identical sibling. In France, over 220 SCD patients have had bone marrow transplantations[371]. However, one of the major current obstacles to bone marrow transplantation as a treatment for SCD is the requirement of an HLA-identical sibling as a donor. For now, the use of partially HLA-matched donors is associated with an increased risk of graft failure, acute and chronic graft-versus-host disease (GVHD) and slow immune system recovery. If partially HLA-matched donors became adequate donors for SCD recipients, as it is already the case for some cancer patients, bone-marrow transplantation would be possible in a higher proportion of SCD patients. Research is already ongoing into stem cell transplantations in SCD patients with partially-matched donors[372].

The risks associated with stem cell transplantation are a major source of concern. Because of their immunocompetence, highly proliferative bone marrow and immunization acquired through many blood transfusions, SCD patients are more at risk of complications after stem cell transplantation compared to patients with hematologic

malignancies. Encouragingly, a significant decrease in mortality has been observed among SCD patients who underwent bone marrow transplantation between 2000-2004, compared to those who did in 1988-2000[150]. The overall survival rate among SCD patients who undergo bone marrow transplantation has now increased above 90%, with disease-free survival reaching between 82-100%[150-153]. One study suggests that event-free survival might now be as high as 95.3%[150]. If the complications and mortality rates following bone marrow transplantation continue to decrease, and if partially matched donors are deemed to be adequate donors, this procedure could become a highly viable solution for severe patients with healthy siblings in the near future. Another way to increase potential candidacy for stem cell transplantation in SCD patients would be with cord blood transplantations. However, the event-free survival rate in cord blood transplantations, which is still experimental for SCD patients, is estimated at 50%[373]. If an acceptable event-free survival rate was reached for cord blood transplantations, cord blood biobanks could provide donors for many SCD patients with no compatible sibling.

The medical care, loss of productivity and premature mortality among SCD patients result in a financial burden both for patients and society. On average, allogeneic bone marrow transplantation costs $203,026 in the United States[374]. A study estimated that the average monthly cost per SCD patient in the U.S. for medical treatment was $1,389[375]. Thus, although bone-marrow transplantation is expensive, it is less expensive than long-term medical care for SCD.

6.4.2 Gene Therapy

Genetic therapy consists in the delivery of therapeutic DNA in an organism to treat a disease. In the case of SCD, three approaches could be adopted in gene therapy. The first approach would consist of correcting the HbS mutation by homologous recombination. The second approach would be to add a normal β-globin gene in the genome. Now that the causal variants modifying HbF levels in the *BCL11A* locus have been identified[364], a third approach would be the editing of the variant rs1427407, disrupting the binding motif in the enhancer in the *BCL11A* locus to increase HbF levels.

In 2001, Pawliuk et al. corrected SCD in transgenic mice by introducing a lentiviral vector containing a human β-globin gene with a variant at position 87 that prevented HbS polymerization[376]. In 2006, an attempt was made using a vector combining both a γ-globin gene and a small hairpin RNA targeting the sickle β-globin messenger ribonucleic acid (mRNA), which corrected human sickle cells in vitro[377]. This vector decreased the production of HbS and increased HbF expression.

The first clinical trials for gene therapy in patients with hemoglobinopathies have begun, and the first successful gene therapy in a β-thalassemia patient occurred in 2010[378]. A phase 1 clinical trial has been initiated in France for both β-thalassemia and SCD patients using a globin lentiviral vector with the same mutation as studied by Pawliuk et al[379]. Another trial in β-thalassemia patients has started in the U.S.[380]. If these first trials are successful, severe SCD patients with no potential bone marrow donors could be the first

candidates for gene therapy in phase 2 and phase 3 clinical trials. If these are conclusive, SCD patients in developed countries could eventually be treated with gene therapy.

6.5 Conclusion

In this work, I used different methods to identify genetic modifiers of SCD. The fine-mapping of previously associated loci uncovered additional and independently associated variants in two of the three loci and increased the heritable variation explained by these three loci. We have identified good candidate modifier genes for acute chest syndrome and pain crisis symptoms in a gene-centric association study. Our results in the gene-set enrichment analysis with HbF levels suggest an association with genes involved in butyrate metabolism and β-hydroxybutyrate concentration in SCD patients' serum. Finally, the whole-exome sequencing of extremely mild SCD patients identified the gene *GPX1* as a potential modifier of disease severity.

Overall, these results provided novel insight for genetic studies of complex traits and showed the importance of fine-mapping studies to assess the complete variation explained by the loci identified in the GWAS. The outcome of this project, as in many others, also stresses the fact that many genetic cohorts are underpowered for GWAS and that novel methods integrating biological knowledge in the analysis of moderately associated signals, like gene-set enrichment analysis, need to be developed. I believe that Bayesian methods or machine-learning algorithms could benefit GWAS by identifying the most promising moderately associated variants.

This project has led to good candidates for genetic modifiers of SCD but additional validation will be needed. If confirmed, these results could be utilized in the prediction of

SCD severity or in the development of treatment for SCD patients. As bone marrow transplantation becomes more frequent in severe patients and as the first clinical trials for gene therapy in SCD patients are underway, these results might be more useful in the research on drug development for SCD patients in Africa, where these cures will not be accessible in a near future.

References

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-51 (2001).
3. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* **104**, 19428-33 (2007).
4. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-9 (2010).
5. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
6. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).
7. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21**, 12-27 (2003).
8. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712-4 (2011).
9. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).
10. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).
11. Kimura, M. *The Neutral Theory of Molecular Evolution.*, (Cambridge University Press, 1983).
12. World Malaria Report 2013. (World Health Organization, Geneva, 2013).
13. Snow, R.W., Guerra, C.A., Noor, A.M., Myint, H.Y. & Hay, S.I. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature* **434**, 214-7 (2005).
14. World malaria report 2012. (World Health Organization. , Geneva, 2012).
15. Malaria. Vol. 2014 (National Institute of Allergy and Infectious Diseases – National Institutes of Health).
16. Ruwende, C. *et al.* Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* **376**, 246-9 (1995).
17. Guindo, A., Fairhurst, R.M., Doumbo, O.K., Wellems, T.E. & Diallo, D.A. X-linked G6PD deficiency protects hemizygous males but not heterozygous females against severe malaria. *PLoS Med* **4**, e66 (2007).
18. Bienzle, U., Ayeni, O., Lucas, A.O. & Luzzatto, L. Glucose-6-phosphate dehydrogenase and malaria. Greater resistance of females heterozygous for enzyme deficiency and of males with non-deficient variant. *Lancet* **1**, 107-10 (1972).
19. Clark, T.G. *et al.* Allelic heterogeneity of G6PD deficiency in West Africa and severe malaria susceptibility. *Eur J Hum Genet* **17**, 1080-5 (2009).

20. Miller, L.H., Mason, S.J., Clyde, D.F. & McGinniss, M.H. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med* **295**, 302-4 (1976).
21. Kasehagen, L.J. *et al.* Reduced Plasmodium vivax erythrocyte infection in PNG Duffy-negative heterozygotes. *PLoS One* **2**, e336 (2007).
22. Rowe, J.A. *et al.* Blood group O protects against severe Plasmodium falciparum malaria through the mechanism. *PNAS* **104**, 17471-6 (2007).
23. Fry, A.E. *et al.* Common variation in the ABO glycosyltransferase is associated with susceptibility to severe Plasmodium falciparum malaria. *Hum Mol Genet* **17**, 567-76 (2008).
24. Hamblin, M.T., Thompson, E.E. & A., D.R. Complex signatures of natural selection at the Duffy blood group locus. . *American Journal Human Genetics* **70**, 369-383 (2002).
25. Hamblin, M.T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* **66**, 1669-79 (2000).
26. Ingram, V.M. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature* **178**, 792-4 (1956).
27. Modiano, D. *et al.* Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature* **414**, 305-8 (2001).
28. Hutagalung, R. *et al.* Influence of hemoglobin E trait on the severity of Falciparum malaria. *J Infect Dis* **179**, 283-6 (1999).
29. Rees, D.C., Williams, T.N. & Gladwin, M.T. Sickle-cell disease. *Lancet* **376**, 2018-31 (2010).
30. Allison, A.C. Polymorphism and Natural Selection in Human Populations. *Cold Spring Harb Symp Quant Biol* **29**, 137-49 (1964).
31. Aidoo, M. *et al.* Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* **359**, 1311-2 (2002).
32. Williams, T.N. *et al.* Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases. *J Infect Dis* **192**, 178-86 (2005).
33. Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet* **41**, 657-65 (2009).
34. Williams, T.N. *et al.* Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. *Nat Genet* **37**, 1253-7 (2005).
35. Kurnit, D.M. Evolution of sickle variant gene. *Lancet* **1**, 104 (1979).
36. Solomon, E. & Bodmer, W.F. Evolution of sickle variant gene. *Lancet* **1**, 923 (1979).
37. Nagel, R.L. *et al.* Hematologically and genetically distinct forms of sickle cell anemia in Africa. The Senegal type and the Benin type. *N Engl J Med* **312**, 880-4 (1985).
38. Pagnier, J. *et al.* Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc Natl Acad Sci U S A* **81**, 1771-3 (1984).
39. Cheboune, Y. *et al.* Structural analysis of the 5' flanking region of the j3-globin gene in African sickle cell anemia patients: Further evidence for three origins of the sickle cell mutation in Africa. *PNAS* **85**, 4431-4435 (1988).
40. Lapouméroulie C *et al.* A novel sickle cell mutation of yet another origin in Africa: the Cameroon type. . *Hum Genet.* **89**, 333-7. (1992).

41. Kulozik, A.E. *et al.* Geographical survey of beta S-globin gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. *Am J Hum Genet* **39**, 239-44 (1986).
42. Serjeant, G.R. & Serjeant, B.E. *Sickle cell disease.* , (Oxford University Press, Oxford, UK., 2001).
43. Moo-Penn, W. *et al.* The presence of hemoglobin S and C Harlem in an individual in the United States. *Blood* **46**, 363-7 (1975).
44. Monplaisir, N. *et al.* Hemoglobin S Antilles: a variant with lower solubility than hemoglobin S and producing sickle cell disease in heterozygotes. *Proc Natl Acad Sci U S A* **83**, 9363-7 (1986).
45. Witkowska, H.E. *et al.* Sickle cell disease in a patient with sickle cell trait and compound heterozygosity for hemoglobin S and hemoglobin Quebec-Chori. *N Engl J Med* **325**, 1150-4 (1991).
46. Nagel, R.L., Fabry, M.E. & Steinberg, M.H. The paradox of hemoglobin SC disease. *Blood Rev* **17**, 167-78 (2003).
47. Nagel, R.L. *et al.* HbS-oman heterozygote: a new dominant sickle syndrome. *Blood* **92**, 4375-82 (1998).
48. Masiello, D. *et al.* Hemoglobin SE disease: a concise review. *Am J Hematol* **82**, 643-9 (2007).
49. Geva, A. *et al.* Hemoglobin Jamaica plain--a sickling hemoglobin with reduced oxygen affinity. *N Engl J Med* **351**, 1532-8 (2004).
50. Sadrzadeh, S.M. & Eaton, J.W. Hemoglobin-mediated oxidant damage to the central nervous system requires endogenous ascorbate. *J Clin Invest* **82**, 1510-5 (1988).
51. Balla, G. *et al.* Ferritin: a cytoprotective antioxidant strategem of endothelium. *J Biol Chem* **267**, 18148-53 (1992).
52. Seixas, E. *et al.* Heme oxygenase-1 affords protection against noncerebral forms of severe malaria. *Proc Natl Acad Sci U S A* **106**, 15837-42 (2009).
53. Gozzelino, R., Jeney, V. & Soares, M.P. Mechanisms of cell protection by heme oxygenase-1. *Annu Rev Pharmacol Toxicol* **50**, 323-54 (2010).
54. Ferreira, A., Balla, J., Jeney, V., Balla, G. & Soares, M.P. A central role for free heme in the pathogenesis of severe malaria: the missing link? *J Mol Med (Berl)* **86**, 1097-111 (2008).
55. Tenhunen, R., Marver, H.S. & Schmid, R. The enzymatic conversion of heme to bilirubin by microsomal heme oxygenase. *Proc Natl Acad Sci U S A* **61**, 748-55 (1968).
56. Hebbel, R.P., Morgan, W.T., Eaton, J.W. & Hedlund, B.E. Accelerated autoxidation and heme loss due to instability of sickle hemoglobin. *Proc Natl Acad Sci U S A.* **85**, 237-241 (1988).
57. Ferreira, A. *et al.* Sickle hemoglobin confers tolerance to Plasmodium infection. *Cell* **145**, 398-409 (2011).
58. Gabriel, A.P., J. Sickle-cell anemia: A Look at Global Haplotype Distribution. . *Nature Education* **3**, 2 (2010).
59. 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 (2013).

60. Modell, B. & Darlison, M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull World Health Organ* **86**, 480-7 (2008).
61. Health Promotion and Population Outreach in the Black and Caribbean Canadian Community. (Health Canada, Ottawa, 2000).
62. Orkin, S.H. & Higgs, D.R. Sickle Cell Disease at 100 Years. *Science* **329**, 291-292 (2010).
63. Delaney, K.M. *et al.* Leg ulcers in sickle cell disease: current patterns and practices. *Hemoglobin* **37**, 325-32 (2013).
64. Frenette, P.S. & Atweh, G.F. Sickle cell disease: old discoveries, new concepts, and future promise. *J Clin Invest* **117**, 850-8 (2007).
65. Forget, B.G. Progress in understanding the hemoglobin switch. *N Engl J Med* **365**, 852-4 (2011).
66. Herrick, J. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *Arch Intern Med* **6**, 517–21. (1910).
67. Brittenham, G.M., Schechter, A.N. & Noguchi, C.T. Hemoglobin S polymerization: primary determinant of the hemolytic and clinical severity of the sickling syndromes. *Blood* **65**, 183-9 (1985).
68. Bunn, H.F. Pathogenesis and treatment of sickle cell disease. *N Engl J Med* **337**, 762-9 (1997).
69. Noguchi, C.T., Rodgers, G.P., Serjeant, G. & Schechter, A.N. Levels of fetal hemoglobin necessary for treatment of sickle cell disease. *N Engl J Med* **318**, 96-9 (1988).
70. Rother, R.P., Bell, L., Hillmen, P. & Gladwin, M.T. The clinical sequelae of intravascular hemolysis and extracellular plasma hemoglobin: a novel mechanism of human disease. *JAMA* **293**, 1653-62 (2005).
71. What is Sickle Cell Anemia? Vol. 2014 (National Heart, Lung, and Blood Institute).
72. Smith, W.R. *et al.* Daily assessment of pain in adults with sickle cell disease. *Ann Intern Med* **148**, 94-101 (2008).
73. Bainbridge, R., Higgs, D.R., Maude, G.H. & Serjeant, G.R. Clinical presentation of homozygous sickle cell disease. *J Pediatr* **106**, 881-5 (1985).
74. el Mouzan, M.I., al Awamy, B.H. & al Torki, M.T. Clinical features of sickle cell disease in eastern Saudi Arab children. *Am J Pediatr Hematol Oncol* **12**, 51-5 (1990).
75. Stevens, M.C., Padwick, M. & Serjeant, G.R. Observations on the natural history of dactylitis in homozygous sickle cell disease. *Clin Pediatr (Phila)* **20**, 311-7 (1981).
76. Worrall, V.T. & Butera, V. Sickle-cell dactylitis. *J Bone Joint Surg Am* **58**, 1161-3 (1976).
77. Ballas, S.K. Defining the phenotypes of sickle cell disease. *Hemoglobin* **35**, 511-9 (2011).
78. Olivieri, I., Scarano, E., Padula, A., Giasi, V. & Priolo, F. Dactylitis, a term for different digit diseases. *Scand J Rheumatol* **35**, 333-40 (2006).
79. Gaston, M.H. *et al.* Prophylaxis with oral penicillin in children with sickle cell anemia. A randomized trial. *N Engl J Med* **314**, 1593-9 (1986).
80. Brousse, V. *et al.* Acute splenic sequestration crisis in sickle cell disease: cohort study of 190 paediatric patients. *Br J Haematol* **156**, 643-8 (2012).

81. Topley, J.M., Rogers, D.W., Stevens, M.C. & Serjeant, G.R. Acute splenic sequestration and hypersplenism in the first five years in homozygous sickle cell disease. *Arch Dis Child* **56**, 765-9 (1981).
82. Rezende, P.V., Viana, M.B., Murao, M., Chaves, A.C. & Ribeiro, A.C. Acute splenic sequestration in a cohort of children with sickle cell anemia. *J Pediatr (Rio J)* **85**, 163-9 (2009).
83. Emond, A.M. *et al.* Acute splenic sequestration in homozygous sickle cell disease: natural history and management. *J Pediatr* **107**, 201-6 (1985).
84. Mills, M.L. Life-threatening complications of sickle cell disease in children. *JAMA* **254**, 1487-91 (1985).
85. 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. *Blood* **86**, 776-83 (1995).
86. Serjeant, G.R. Natural history and determinants of clinical severity of sickle cell disease. *Curr Opin Hematol* **2**, 103-8 (1995).
87. 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**, 1855-65 (2000).
88. Castro, O. *et al.* The acute chest syndrome in sickle cell disease: incidence and risk factors. The Cooperative Study of Sickle Cell Disease. *Blood* **84**, 643-9 (1994).
89. Vichinsky, E.P. *et al.* Acute chest syndrome in sickle cell disease: clinical presentation and course. Cooperative Study of Sickle Cell Disease. *Blood* **89**, 1787-92 (1997).
90. Serjeant, G.R. *Sickle Cell Disease.* , (Oxford, UK, Oxford,, 1986).
91. Gray, A., Anionwu, E.N., Davies, S.C. & Brozovic, M. Patterns of mortality in sickle cell disease in the United Kingdom. *J Clin Pathol* **44**, 459-63 (1991).
92. Thomas, A.N., Pattison, C. & Serjeant, G.R. Causes of death in sickle-cell disease in Jamaica. *Br Med J (Clin Res Ed)* **285**, 633-5 (1982).
93. Ohene-Frempong, K. *et al.* Cerebrovascular accidents in sickle cell disease: rates and risk factors. *Blood* **91**, 288-94 (1998).
94. Quinn, C.T., Rogers, Z.R. & Buchanan, G.R. Survival of children with sickle cell disease. *Blood* **103**, 4023-7 (2004).
95. Balkaran, B. *et al.* Stroke in a cohort of patients with homozygous sickle cell disease. *J Pediatr* **120**, 360-6 (1992).
96. Pegelow, C.H. *et al.* Longitudinal changes in brain magnetic resonance imaging findings in children with sickle cell disease. *Blood* **99**, 3014-8 (2002).
97. Bernaudin, F. *et al.* Impact of early transcranial Doppler screening and intensive therapy on cerebral vasculopathy outcome in a newborn sickle cell anemia cohort. *Blood* **117**, 1130-40; quiz 1436 (2011).
98. Powars, D., Wilson, B., Imbus, C., Pegelow, C. & Allen, J. The natural history of stroke in sickle cell disease. *Am J Med* **65**, 461-71 (1978).
99. Sarnaik, S., Soorya, D., Kim, J., Ravindranath, Y. & Lusher, J. Periodic transfusions for sickle cell anemia and CNS infarction. *Am J Dis Child* **133**, 1254-7 (1979).
100. Pegelow, C.H. *et al.* Risk of recurrent stroke in patients with sickle cell disease treated with erythrocyte transfusions. *J Pediatr* **126**, 896-9 (1995).
101. Serjeant, G.R. *et al.* Outbreak of aplastic crises in sickle cell anaemia associated with parvovirus-like agent. *Lancet* **2**, 595-7 (1981).

102. Goldstein, A.R., Anderson, M.J. & Serjeant, G.R. Parvovirus associated aplastic crisis in homozygous sickle cell disease. *Arch Dis Child* **62**, 585-8 (1987).
103. Yang, Y.M., Shah, A.K., Watson, M. & Mankad, V.N. Comparison of costs to the health sector of comprehensive and episodic health care for sickle cell disease patients. *Public Health Rep* **110**, 80-6 (1995).
104. Brozovic, M., Davies, S.C. & Brownell, A.I. Acute admissions of patients with sickle cell disease who live in Britain. *Br Med J (Clin Res Ed)* **294**, 1206-8 (1987).
105. Ballas, S.K. The sickle cell painful crisis in adults: phases and objective signs. *Hemoglobin* **19**, 323-33 (1995).
106. Platt, O.S. *et al.* Pain in sickle cell disease. Rates and risk factors. *N Engl J Med* **325**, 11-6 (1991).
107. Baum, K.F., Dunn, D.T., Maude, G.H. & Serjeant, G.R. The painful crisis of homozygous sickle cell disease. A study of the risk factors. *Arch Intern Med* **147**, 1231-4 (1987).
108. Cumming, V., King, L., Fraser, R., Serjeant, G. & Reid, M. Venous incompetence, poverty and lactate dehydrogenase in Jamaica are important predictors of leg ulceration in sickle cell anaemia. *Br J Haematol* **142**, 119-25 (2008).
109. Koshy, M. *et al.* Leg ulcers in patients with sickle cell disease. *Blood* **74**, 1403-8 (1989).
110. Serjeant, G.R. Leg ulceration in sickle cell anemia. *Arch Intern Med* **133**, 690-4 (1974).
111. Minniti, C.P., Eckman, J., Sebastiani, P., Steinberg, M.H. & Ballas, S.K. Leg ulcers in sickle cell disease. *Am J Hematol* **85**, 831-3 (2010).
112. Trent, J.T. & Kirsner, R.S. Leg ulcers in sickle cell disease. *Adv Skin Wound Care* **17**, 410-6 (2004).
113. Hernigou, P., Bachir, D. & Galacteros, F. The natural history of symptomatic osteonecrosis in adults with sickle-cell disease. *J Bone Joint Surg Am* **85-A**, 500-4 (2003).
114. Milner, P.F. *et al.* Sickle cell disease as a cause of osteonecrosis of the femoral head. *N Engl J Med.* **325**, 1476-81 (1991).
115. Rogers, Z.R. Priapism in sickle cell disease. *Hematol Oncol Clin North Am* **19**, 917-28, viii (2005).
116. Emond, A.M., Holman, R., Hayes, R.J. & Serjeant, G.R. Priapism and impotence in homozygous sickle cell disease. *Arch Intern Med* **140**, 1434-7 (1980).
117. Fowler, J.E., Jr., Koshy, M., Strub, M. & Chinn, S.K. Priapism associated with the sickle cell hemoglobinopathies: prevalence, natural history and sequelae. *J Urol* **145**, 65-8 (1991).
118. Adeyoju, A.B. *et al.* Priapism in sickle-cell disease; incidence, risk factors and complications - an international multicentre study. *BJU Int* **90**, 898-902 (2002).
119. Ware, R.E. *et al.* Renal function in infants with sickle cell anemia: baseline data from the BABY HUG trial. *J Pediatr* **156**, 66-70 e1 (2010).
120. Platt, O.S. *et al.* Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med* **330**, 1639-44 (1994).

121. Falk, R.J. *et al.* Prevalence and pathologic features of sickle cell nephropathy and response to inhibition of angiotensin-converting enzyme. *N Engl J Med* **326**, 910-5 (1992).
122. Watson, J. The significance of the paucity of sickle cells in newborn Negro infants. *Am J Med Sci* **215**, 419-23 (1948).
123. Morrison, J.C., Whybrew, W.D., Bucovaz, E.T. & Wiser, W.L. Fluctuation of fetal hemoglobin in sickle-cell anemia. *Am J Obstet Gynecol* **125**, 1085-8 (1976).
124. Haghshenass, M., Ismail-Beigi, F., Clegg, J.B. & Weatherall, D.J. Mild sickle-cell anaemia in Iran associated with high levels of fetal haemoglobin. *J Med Genet* **14**, 168-71 (1977).
125. Higgs, D.R. *et al.* The interaction of alpha-thalassemia and homozygous sickle-cell disease. *N Engl J Med* **306**, 1441-6 (1982).
126. Higgs, D.R. *et al.* Detection of alpha thalassaemia in Negro infants. *Br J Haematol* **46**, 39-46 (1980).
127. Steinberg, M.H. & Embury, S.H. Alpha-thalassemia in blacks: genetic and clinical aspects and interactions with the sickle hemo- globin gene. *Blood* **68**, 985-990 (1986).
128. Kar, B.C. *et al.* Sickle cell disease in Orissa State, India. *Lancet* **2**, 1198-201 (1986).
129. Padmos, M.A. *et al.* Two different forms of homozygous sickle cell disease occur in Saudi Arabia. *Br J Haematol* **79**, 93-8 (1991).
130. Embury, S.H. *et al.* Concurrent sickle-cell anemia and alpha-thalassemia: effect on severity of anemia. *N Engl J Med* **306**, 270-4 (1982).
131. Steinberg, M.H. *et al.* Effects of thalassemia and microcytosis on the hematologic and vasoocclusive severity of sickle cell anemia. *Blood* **63**, 1353-60 (1984).
132. Seakins, M., Gibbs, W.N., Milner, P.F. & Bertles, J.F. Erythrocyte Hb-S concentration. An important factor in the low oxygen affinity of blood in sickle cell anemia. *J Clin Invest* **52**, 422-32 (1973).
133. de Ceulaer, K. *et al.* alpha-Thalassemia reduces the hemolytic rate in homozygous sickle-cell disease. *N Engl J Med* **309**, 189-90 (1983).
134. Bernaudin, F. *et al.* G6PD deficiency, absence of alpha-thalassemia, and hemolytic rate at baseline are significant independent risk factors for abnormally high cerebral velocities in patients with sickle cell anemia. *Blood* **112**, 4314-7 (2008).
135. Adams, R.J. *et al.* Alpha thalassemia and stroke risk in sickle cell anemia. *Am J Hematol* **45**, 279-82 (1994).
136. Neonato, M.G. *et al.* Acute clinical events in 299 homozygous sickle cell patients living in France. French Study Group on Sickle Cell Disease. *Eur J Haematol* **65**, 155-64 (2000).
137. Hsu, L.L. *et al.* Alpha Thalassemia is associated with decreased risk of abnormal transcranial Doppler ultrasonography in children with sickle cell anemia. . *J Pediatr Hematol Oncol.* **25**, 622-8 (2003).
138. Nolan, V.G., Wyszynski, D.F., Farrer, L.A. & Steinberg, M.H. Hemolysis-associated priapism in sickle cell disease. *Blood* **106**, 3264-7 (2005).
139. Billett, H.H., Kim, K., Fabry, M.E. & Nagel, R.L. The percentage of dense red cells does not predict incidence of sickle cell painful crisis. *Blood* **68**, 301-3 (1986).

140. Ballas, S.K., Talacki, C.A., Rao, V.M. & Steiner, R.M. The prevalence of avascular necrosis in sickle cell anemia: correlation with alpha-thalassemia. *Hemoglobin* **13**, 649-55 (1989).
141. Steinberg MH, Forget BG, Higgs DR & D, W. *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management*, (Cambridge, United Kingdom: Cambridge University Press, 2009).
142. U.S. Department of Health and Human Services, P.H.S., National Institutes of Health, National Heart, Lung, and Blood Institute. The Management of Sickle Cell Disease (NIH Publication No. 02-2117 4th edition). (2002).
143. Redwood, A.M., Williams, E.M., Desal, P. & Serjeant, G.R. Climate and painful crisis of sickle-cell disease in Jamaica. *Br Med J* **1**, 66-8 (1976).
144. Jones, S. *et al.* Windy weather and low humidity are associated with an increased number of hospital admissions for acute pain and sickle cell disease in an urban environment with a maritime temperate climate. *Br J Haematol* **131**, 530-3 (2005).
145. Makani, J. *et al.* Malaria in patients with sickle cell anemia: burden, risk factors, and outcome at the outpatient clinic and during hospitalization. *Blood* **115**, 215-20 (2010).
146. Weatherall, D.J. The inherited diseases of hemoglobin are an emerging global health burden. *Blood* **115**, 4331-6 (2010).
147. Yanni, E., Grosse, S.D., Yang, Q. & Olney, R.S. Trends in pediatric sickle cell disease-related mortality in the United States, 1983-2002. *J Pediatr* **154**, 541-5 (2009).
148. Sheth, S., Licursi, M. & Bhatia, M. Sickle cell disease: time for a closer look at treatment options? *Br J Haematol* **162**, 455-64 (2013).
149. Hassell, K.L. Population estimates of sickle cell disease in the U.S. *Am J Prev Med* **38**, S512-21 (2010).
150. Bernaudin, F. *et al.* Long-term results of related myeloablative stem-cell transplantation to cure sickle cell disease. *Blood* **110**, 2749-56 (2007).
151. Vermylen, C. *et al.* Haematopoietic stem cell transplantation for sickle cell anaemia: the first 50 patients transplanted in Belgium. *Bone Marrow Transplant* **22**, 1-6 (1998).
152. Walters, M.C. *et al.* Stable mixed hematopoietic chimerism after bone marrow transplantation for sickle cell anemia. *Biol Blood Marrow Transplant* **7**, 665-73 (2001).
153. Lucarelli, G. *et al.* Allogeneic cellular gene therapy in hemoglobinopathies--evaluation of hematopoietic SCT in sickle cell anemia. *Bone Marrow Transplant* **47**, 227-30 (2012).
154. Letvin, N.L., Linch, D.C., Beardsley, G.P., McIntyre, K.W. & Nathan, D.G. Augmentation of fetal-hemoglobin production in anemic monkeys by hydroxyurea. *N Engl J Med* **310**, 869-73 (1984).
155. Platt, O.S. *et al.* Hydroxyurea enhances fetal hemoglobin production in sickle cell anemia. *J Clin Invest* **74**, 652-6 (1984).
156. Yarbro, J.W. Mechanism of action of hydroxyurea. *Semin Oncol* **19**, 1-10 (1992).
157. 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**, 1317-22 (1995).
158. Hankins, J.S. *et al.* Long-term hydroxyurea therapy for infants with sickle cell anemia: the HUSOFT extension study. *Blood* **106**, 2269-75 (2005).

159. Steinberg, M.H. *et al.* Effect of hydroxyurea on mortality and morbidity in adult sickle cell anemia: risks and benefits up to 9 years of treatment. *JAMA* **289**, 1645-51 (2003).
160. Lou, T.F., Singh, M., Mackie, A., Li, W. & Pace, B.S. Hydroxyurea generates nitric oxide in human erythroid cells: mechanisms for gamma-globin gene activation. *Exp Biol Med (Maywood)* **234**, 1374-82 (2009).
161. Laurance, S. *et al.* Differential modulation of adhesion molecule expression by hydroxycarbamide in human endothelial cells from the micro- and macrocirculation: potential implications in sickle cell disease vasoocclusive events. *Haematologica* **96**, 534-42 (2011).
162. de Montalembert, M. *et al.* Long-term hydroxyurea treatment in children with sickle cell disease: tolerance and clinical outcomes. *Haematologica* **91**, 125-8 (2006).
163. Bard, H. & Prosmanne, J. Relative rates of fetal hemoglobin and adult hemoglobin synthesis in cord blood of infants of insulin-dependent diabetic mothers. *Pediatrics* **75**, 1143-7 (1985).
164. Perrine, S.P., Greene, M.F. & Faller, D.V. Delay in the fetal globin switch in infants of diabetic mothers. *N Engl J Med* **312**, 334-8 (1985).
165. Perrine, S.P. *et al.* A short-term trial of butyrate to stimulate fetal-globin-gene expression in the beta-globin disorders. *N Engl J Med* **328**, 81-6 (1993).
166. Sher, G.D. *et al.* Extended therapy with intravenous arginine butyrate in patients with beta-hemoglobinopathies. *N Engl J Med* **332**, 1606-10 (1995).
167. Fathallah, H., Weinberg, R.S., Galperin, Y., Sutton, M. & Atweh, G.F. Role of epigenetic modifications in normal globin gene regulation and butyrate-mediated induction of fetal hemoglobin. *Blood* **110**, 3391-7 (2007).
168. Pauling L, Itano HA, Singer SJ & IC, W. Sickle cell anemia, a molecular disease. *Science*, 543-48 (1949).
169. Neel, J.V. The Inheritance of Sickle Cell Anemia. *Science* **110**, 64-6 (1949).
170. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-8 (1953).
171. Sanger, F. & Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-8 (1975).
172. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-7 (1977).
173. Gusella, J.F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-8 (1983).
174. Riordan, J.R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066-73 (1989).
175. Bertina, R.M. *et al.* Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64-7 (1994).
176. Corder, E.H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-3 (1993).
177. Altshuler, D. *et al.* The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. . *Nat Genet.* **26**, 76-80 (2000).
178. Hugot, J.P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603 (2001).

179. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603-6 (2001).
180. Rioux, J.D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* **29**, 223-8 (2001).
181. Stoll, M. *et al.* Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat Genet* **36**, 476-80 (2004).
182. Stefansson, H. *et al.* Neuregulin 1 and susceptibility to schizophrenia. *Am J Hum Genet* **71**, 877-92 (2002).
183. Nisticò, L. *et al.* The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. . *Hum Mol Genet.* **5**, 1075-80 (1996).
184. Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet.* **69**, 936-50 (2001).
185. Marinucci M *et al.* Beta-Thalassemia associated with increased HbF production. Evidence for the existence of a heterocellular hereditary persistence of fetal hemoglobin (HPFH) determinant linked to beta-thalassemia in a southern Italian population. *Hemoglobin.* **5**, 1-17 (1981).
186. Fritsch, E.F., Lawn, R.M. & Maniatis, T. Characterisation of deletions which affect the expression of fetal globin genes in man. . *Nature* **279**, 598-603 (1979).
187. Bethlenfalvay, N. *et al.* Hereditary persistence of fetal hemoglobin, beta thalassemia, and the hemoglobin delta-beta locus: further family data and genetic interpretations. *American Journal of human genetics* **27**, 140-154 (1975).
188. Cappellini MD, Fiorelli G & LF., B. Interaction between homozygous beta (0) thalassaemia and the Swiss type of hereditary persistence of fetal haemoglobin. *Br J Haematol.* **48**, 561-72 (1981).
189. Giampaolo, A. *et al.* Heterocellular hereditary persistence of fetal hemoglobin (HPFH). Molecular mechanisms of abnormal gamma-gene expression in association with beta thalassemia and linkage relationship with the beta-globin gene cluster. *Hum Genet* **66**, 151-6 (1984).
190. Milner, P.F. *et al.* Increased HbF in sickle cell anemia is determined by a factor linked to the beta S gene from one parent. *Blood* **63**, 64-72 (1984).
191. Thein, S.L. & Weatherall, D.J. A non-deletion hereditary persistence of fetal hemoglobin (HPFH) determinant not linked to the beta-globin gene complex. *Prog Clin Biol Res.* **316B**, 97-111 (1989).
192. Pistidda, P. *et al.* Fetal hemoglobin expression in compound heterozygotes for -117 (G-->A)A gamma HPFH and beta zero 39 nonsense thalassemia. *Am J Hematol* **49**, 267-70 (1995).
193. Dover, G.J., Boyer, S.H. & Pembrey, M.E. F-cell production in sickle cell anemia: regulation by genes linked to beta-hemoglobin locus. *Science* **211**, 1441-4 (1981).
194. Gilman, J.G. & Huisman, T.H. DNA sequence variation associated with elevated fetal G gamma globin production. *Blood* **66**, 783-7 (1985).
195. Craig, J.E. *et al.* Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nat Genet* **12**, 58-64 (1996).

196. Thein, S.L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc Natl Acad Sci U S A* **104**, 11346-51 (2007).
197. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513-6 (2000).
198. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-33 (2001).
199. Nickerson, D.A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet* **19**, 233-40 (1998).
200. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229-32 (2001).
201. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
202. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-9 (2005).
203. Hindorff LA, M.J.E.B.I., Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: http://www.genome.gov/gwastudies. Accessed February 13th 2014.
204. 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**, 1197-9 (2007).
205. 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**, 1620-5 (2008).
206. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol.* **32**, 381-5 (2008).
207. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
208. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-5 (2010).
209. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* **42**, 483-5 (2010).
210. Sun, Y. *et al.* Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. *Am J Hum Genet* **87**, 146-53 (2010).
211. Walsh, T. *et al.* Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am J Hum Genet* **87**, 90-4 (2010).
212. Pierce, S.B. *et al.* Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. *Am J Hum Genet* **87**, 282-8 (2010).
213. Bilguvar, K. *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**, 207-10 (2010).
214. Gilissen, C. *et al.* Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am J Hum Genet* **87**, 418-23 (2010).

215. Krawitz, P.M. *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* **42**, 827-9 (2010).
216. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-42 (2012).
217. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* **43**, 864-8 (2011).
218. Wetterstrand, K.A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Vol. 2014.
219. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
220. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-23 (2009).
221. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
222. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
223. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
224. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
225. 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**, 11869-74 (2008).
226. Sankaran, V.G. *et al.* Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* **322**, 1839-42 (2008).
227. Sankaran, V.G. *et al.* Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460**, 1093-7 (2009).
228. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387-9 (2009).
229. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8**, e1000294 (2010).
230. Embury, S.H. & Hebbel, R.P. Sickle cell disease: Basic principles and clinical practice. (1994).
231. Labie, D. *et al.* Common haplotype dependency of high G gamma-globin gene expression and high Hb F levels in beta-thalassemia and sickle cell anemia patients. *Proc Natl Acad Sci U S A* **82**, 2111-4 (1985).
232. 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**, 1815-22 (2010).
233. Galanello, R. *et al.* Amelioration of Sardinian beta0 thalassemia by genetic modifiers. *Blood* **114**, 3935-7 (2009).

234. Nuinoon, M. *et al.* A genome-wide association identified the common genetic variants influence disease severity in beta(0)-thalassemia/hemoglobin E. *Hum Genet* (2009).
235. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
236. Farber, M.D., Koshy, M. & Kinney, T.R. Cooperative Study of Sickle Cell Disease: Demographic and socioeconomic characteristics of patients and families with sickle cell disease. *J Chronic Dis* **38**, 495-505 (1985).
237. Stephens, M., Sloan, J.S., Robertson, P.D., Scheet, P. & Nickerson, D.A. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat Genet* **38**, 375-81 (2006).
238. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195-202 (1998).
239. Campbell, C.D. *et al.* Association studies of BMI and type 2 diabetes in the neuropeptide Y pathway: a possible role for NPY2R as a candidate gene for type 2 diabetes in men. *Diabetes* **56**, 1460-7 (2007).
240. Kang, S.J. *et al.* Genome-wide association of anthropometric traits in African- and African-derived populations. *Hum Mol Genet* **19**, 2725-38 (2010).
241. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* **86**, 832-8 (2010).
242. Keating, B.J. *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS One* **3**, e3583 (2008).
243. Musunuru, K. *et al.* Candidate Gene Association Resource (CARe): Design, Methods, and Proof of Concept. *Circ Cardiovasc Genet* **3**, 267-75 (2010).
244. Price, A.L. *et al.* Discerning the Ancestry of European Americans in Genetic Association Studies. *PLoS Genet* **4**, e236 (2008).
245. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
246. Weatherall, D.J. The inherited diseases of hemoglobin are an emerging global health burden. *Blood* (2010).
247. Weatherall, D.J. & Clegg, J.B. Inherited haemoglobin disorders: an increasing global health problem. *Bull World Health Organ* **79**, 704-12 (2001).
248. Steinberg, M.H., Forget, B.G., Higgs, D.R. & Weatherall, D. *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management* 846 (Cambridge University Press, Cambridge, U.K., 2009).
249. Sankaran, V.G., Lettre, G., Orkin, S.H. & Hirschhorn, J.N. Modifier genes in Mendelian disorders: the example of hemoglobin disorders. *Ann N Y Acad Sci* **1214**, 47-56 (2010).
250. Lettre, G. The Search for Genetic Modifiers of Disease Severity in the beta-Hemoglobinopathies. *Cold Spring Harb Perspect Med* **2**(2012).
251. 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**, 448-51 (2001).

252. Milton, J.N. *et al.* A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. *PLoS ONE* **7**, e34741 (2012).
253. Ashley-Koch, A.E. *et al.* MYH9 and APOL1 are both associated with sickle cell disease nephropathy. *Br J Haematol* **155**, 386-94 (2011).
254. Lo, K.S. *et al.* Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Hum Genet* **129**, 307-17 (2011).
255. Sebastiani, P. *et al.* Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study. *Am J Hematol* **85**, 29-35 (2010).
256. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
257. Price, A.L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**, e1000519 (2009).
258. Lettre, G. *et al.* Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. *PLoS Genet* **7**, e1001300 (2011).
259. Ghosh, S. *et al.* Global gene expression profiling of endothelium exposed to heme reveals an organ-specific induction of cytoprotective enzymes in sickle cell disease. *PLoS One* **6**, e18399 (2011).
260. Milner, P.F. *et al.* Sickle cell disease as a cause of osteonecrosis of the femoral head. *N Engl J Med* **325**, 1476-81 (1991).
261. Gaston, M. *et al.* Recruitment in the Cooperative Study of Sickle Cell Disease (CSSCD). *Control Clin Trials* **8**, 131S-140S (1987).
262. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
263. Burstein, E. *et al.* COMMD proteins, a novel family of structural and functional homologs of MURR1. *J Biol Chem* **280**, 22222-32 (2005).
264. Ofori-Acquah, S., Ghosh, S. & Adisa, O. Acute Elevation Of Protein-Free Plasma Heme Triggers Acute Chest Syndrome In Mouse Models Of Sickle Cell Anemia. *American Journal of Respiratory and Critical Care Medicine* **183**, A3757 (2011).
265. Fardo, D., Celedon, J.C., Raby, B.A., Weiss, S.T. & Lange, C. On dichotomizing phenotypes in family-based association tests: quantitative phenotypes are not always the optimal choice. *Genet Epidemiol* **31**, 376-82 (2007).
266. Zheng, L. *et al.* ShRNA-Targeted COMMD7 Suppresses Hepatocellular Carcinoma Growth. *PLoS ONE* **7**, e45412 (2012).
267. Miller, A.C. & Gladwin, M.T. Pulmonary complications of sickle cell disease. *Am J Respir Crit Care Med* **185**, 1154-65 (2012).
268. Bean, C.J. *et al.* Heme oxygenase-1 gene promoter polymorphism is associated with reduced incidence of acute chest syndrome among children with sickle cell disease. *Blood* **120**, 3822-8 (2012).
269. Svensson, C.I. *et al.* Spinal phospholipase A2 in inflammatory hyperalgesia: role of the small, secretory phospholipase A2. *Neuroscience* **133**, 543-53 (2005).
270. Lucas, K.K., Svensson, C.I., Hua, X.Y., Yaksh, T.L. & Dennis, E.A. Spinal phospholipase A2 in inflammatory hyperalgesia: role of group IVA cPLA2. *Br J Pharmacol* **144**, 940-52 (2005).

271. Styles, L.A. *et al.* Phospholipase A2 levels in acute chest syndrome of sickle cell disease. *Blood* **87**, 2573-8 (1996).
272. Styles, L.A., Aarsman, A.J., Vichinsky, E.P. & Kuypers, F.A. Secretory phospholipase A(2) predicts impending acute chest syndrome in sickle cell disease. *Blood* **96**, 3276-8 (2000).
273. Styles, L. *et al.* Refining the value of secretory phospholipase A2 as a predictor of acute chest syndrome in sickle cell disease: results of a feasibility study (PROACTIVE). *Br J Haematol* **157**, 627-36 (2012).
274. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
275. Wang, K. *et al.* Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *Am J Hum Genet* **84**, 399-405 (2009).
276. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81**, 1278-83 (2007).
277. Thein, S.L. & Craig, J.E. Genetics of Hb F/F cell variance in adults and heterocellular hereditary persistence of fetal hemoglobin. *Hemoglobin* **22**, 401-14 (1998).
278. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat Genet* **42**, 1049-51 (2010).
279. Sankaran, V.G. & Nathan, D.G. Reversing the hemoglobin switch. *N Engl J Med* **363**, 2258-60 (2010).
280. Perrine, S.P. *et al.* Sodium butyrate enhances fetal globin gene expression in erythroid progenitors of patients with Hb SS and beta thalassemia. *Blood* **74**, 454-9 (1989).
281. Constantoulakis, P., Knitter, G. & Stamatoyannopoulos, G. On the induction of fetal hemoglobin by butyrates: in vivo and in vitro studies with sodium butyrate and comparison of combination treatments with 5-AzaC and AraC. *Blood* **74**, 1963-71 (1989).
282. Galarneau, G. *et al.* Gene-centric association study of acute chest syndrome and painful crisis in sickle cell disease patients. *Blood* **122**, 434-42 (2013).
283. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
284. Gauderman, W.J. & Morrison, J.M. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe. (2006).
285. Paine-Saunders, S., Viviano, B.L. & Saunders, S. GPC6, a novel member of the glypican gene family, encodes a product structurally related to GPC4 and is colocalized with GPC5 on human chromosome 13. *Genomics* **57**, 455-8 (1999).
286. Perrine, S.P. *et al.* Butyrate infusions in the ovine fetus delay the biologic clock for globin gene switching. *Proc Natl Acad Sci U S A* **85**, 8540-2 (1988).
287. Dover, G.J., Brusilow, S. & Charache, S. Induction of fetal hemoglobin production in subjects with sickle cell anemia by oral sodium phenylbutyrate. *Blood* **84**, 339-43 (1994).
288. Atweh, G.F. *et al.* Sustained induction of fetal hemoglobin by pulse butyrate therapy in sickle cell disease. *Blood* **93**, 1790-7 (1999).

289. Resar, L.M. *et al.* Induction of fetal hemoglobin synthesis in children with sickle cell anemia on low-dose oral sodium phenylbutyrate therapy. *J Pediatr Hematol Oncol* **24**, 737-41 (2002).
290. Kutlar, A. *et al.* A phase 1/2 trial of HQK-1001, an oral fetal globin inducer, in sickle cell disease. *Am J Hematol* **87**, 1017-21 (2012).
291. Kutlar, A. *et al.* A dose-escalation phase IIa study of 2,2-dimethylbutyrate (HQK-1001), an oral fetal globin inducer, in sickle cell disease. *Am J Hematol* **88**, E255-60 (2013).
292. Ikuta, T., Kan, Y.W., Swerdlow, P.S., Faller, D.V. & Perrine, S.P. Alterations in protein-DNA interactions in the gamma-globin gene promoter in response to butyrate therapy. *Blood* **92**, 2924-33 (1998).
293. Ajamian, F., Salminen, A. & Reeben, M. Selective regulation of class I and class II histone deacetylases expression by inhibitors of histone deacetylases in cultured mouse neural cells. *Neurosci Lett* **365**, 64-8 (2004).
294. Piel, F.B. *et al.* Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. *Lancet* **381**, 142-51 (2013).
295. Serjeant, G.R. *et al.* Haemoglobin gene frequencies in the Jamaican population: a study in 100,000 newborns. *Br J Haematol* **64**, 253-62 (1986).
296. Cohen, G. & Hochstein, P. Glutathione Peroxidase: The Primary Agent for the Elimination of Hydrogen Peroxide in Erythrocytes. *Biochemistry* **2**, 1420-8 (1963).
297. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
298. Van der Auwera, G.A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. in *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2012).
299. Picard.
300. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
301. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
302. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
303. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
304. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. . *PLoS Comput Biol.* **6**, e1001025 (2010).
305. Mills, G.C. Hemoglobin catabolism. I. Glutathione peroxidase, an erythrocyte enzyme which protects hemoglobin from oxidative breakdown. *J Biol Chem.* **229**, 189-97. (1957).
306. Jablonska, E. *et al.* Association between GPx1 Pro198Leu polymorphism, GPx1 activity and plasma selenium concentration in humans. *Eur J Nutr* **48**, 383-6 (2009).
307. Destro-Bisol, G. & Spedini, G. Anthropological survey on red cell glutathione peroxidase (GPX1) polymorphism in central western Africa: a tentative hypothesis on

the interaction between GPX1*2 and Hb beta *S allelic products. *American Journal of Physical Anthropology* **79**, 217-24 (1989).
308. Cho, C.S. *et al.* Hydroxyurea-induced expression of glutathione peroxidase 1 in red blood cells of individuals with sickle cell anemia. *Antioxid Redox Signal* **13**, 1-11 (2010).
309. Pepin, J. From the Old World to the New World: an ecologic study of population susceptibility to HIV infection. *Trop Med Int Health* **10**, 627-39 (2005).
310. Simms, T.M., Rodriguez, C.E., Rodriguez, R. & Herrera, R.J. The genetic structure of populations from Haiti and Jamaica reflect divergent demographic histories. *Am J Phys Anthropol* **142**, 49-66 (2010).
311. Jung, D.J. *et al.* Novel transcription coactivator complex containing activating signal cointegrator 1. *Mol Cell Biol* **22**, 5203-11 (2002).
312. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet.* **45**, 501-12. (2013).
313. Lanktree, M.B. *et al.* Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. *Am J Hum Genet* **88**, 6-18 (2011).
314. Jia, P. *et al.* A bias-reducing pathway enrichment analysis of genome-wide association data confirmed association of the MHC region with schizophrenia. *J Med Genet* **49**, 96-103 (2012).
315. Zhang, L. *et al.* Pathway-based genome-wide association analysis identified the importance of regulation-of-autophagy pathway for ultradistal radius BMD. *J Bone Miner Res* **25**, 1572-80 (2010).
316. Lambert, J.C. *et al.* Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *J Alzheimers Dis* **20**, 1107-18 (2010).
317. Holmans, P. *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. . *Am J Hum Genet.* **85**, 13-24. (2009).
318. Chen, H. & Sharp, B.M. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5**, 147 (2004).
319. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. . *PLoS Genet.* **5**, e1000534. (2009).
320. Medina, I. *et al.* Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res* **37**, W340-4 (2009).
321. Nam, D., Kim, J., Kim, S.Y. & Kim, S. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res* **38**, W749-54 (2010).
322. Yaspan, B.L. *et al.* Genetic analysis of biological pathway data through genomic randomization. . *Hum Genet.* **129**, 563-71 (2011).
323. Chen, L.S. *et al.* Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. . *Am J Hum Genet.* **86**, 860-71. (2010).
324. Segrè, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. . *PLoS Genet.* **6**, e1001058. (2010).

325. Holden, M., Deng, S., Wojnowski, L. & Kulle, B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24**, 2784-5 (2008).
326. O'Dushlaine, C. *et al.* The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics* **25**, 2762-3 (2009).
327. Zhong, H., Yang, X., Kaplan, L.M., Molony, C. & Schadt, E.E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* **86**, 581-91 (2010).
328. Edwards, Y.J. *et al.* Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. *PLoS One.* **6**, e16917 (2011).
329. Le-Niculescu, H. *et al.* Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms. *Am J Med Genet B Neuropsychiatr Genet* **150B**, 155-81 (2009).
330. Le-Niculescu, H. *et al.* Identifying blood biomarkers for mood disorders using convergent functional genomics. . *Mol Psychiatry.* **14**, 156-74 (2009).
331. Patel, S.D. *et al.* Coming to grips with complex disorders: genetic risk prediction in bipolar disorder using panels of genes identified through convergent functional genomics. . *Am J Med Genet B Neuropsychiatr Genet.* **153B**, 850-77. (2010).
332. Le-Niculescu, H. *et al.* Convergent functional genomics of anxiety disorders: translational identification of genes, biomarkers, pathways and mechanisms. *Transl Psychiatry* **1**, e9 (2011).
333. Smith, A.K., Fang, H., Whistler, T., Unger, E.R. & Rajeevan, M.S. Convergent genomic studies identify association of GRIK2 and NPAS2 with chronic fatigue syndrome. *Neuropsychobiology* **64**, 183-94 (2011).
334. Ayalew, M. *et al.* Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry* **17**, 887-905 (2012).
335. Carayol, J. *et al.* A scoring strategy combining statistics and functional genomics supports a possible role for common polygenic variation in autism. *Front Genet* **5**, 33 (2014).
336. 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**, 906-13 (2007).
337. Wei, Z. *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet* **5**, e1000678 (2009).
338. Negi, S. *et al.* A genome-wide association study reveals ARL15, a novel non-HLA susceptibility gene for rheumatoid arthritis in North Indians. *Arthritis Rheum* **65**, 3026-35 (2013).
339. Mittag, F. *et al.* Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. . *Hum Mutat.* **33**, 1708-18. (2012).
340. Göring, H.H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. . *Nat Genet.* **39**, 1208-16. (2007).
341. Murphy, A. *et al.* Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum Mol Genet* **19**, 4745-57 (2010).

342. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).
343. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-7 (2007).
344. Rotival, M. *et al.* Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genet* **7**, e1002367 (2011).
345. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur J Hum Genet* **21**, 48-54 (2013).
346. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
347. Heinzen, E.L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol* **6**, e1 (2008).
348. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107 (2008).
349. Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet* **7**, e1002078 (2011).
350. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* **8**, e1002431 (2012).
351. Myers, A.J. *et al.* A survey of genetic human cortical gene expression. *Nat Genet* **39**, 1494-9 (2007).
352. Liu, C. *et al.* Whole-genome association mapping of gene expression in the human prefrontal cortex. *Mol Psychiatry* **15**, 779-84 (2010).
353. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* **42**, 137-41 (2010).
354. Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine. *Nat Genet* **43**, 565-9 (2011).
355. Rueedi, R. *et al.* Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS Genet* **10**, e1004132 (2014).
356. Rhee, E.P. *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab* **18**, 130-43 (2013).
357. Luykx, J.J. *et al.* Genome-wide association study of monoamine metabolite levels in human cerebrospinal fluid. *Mol Psychiatry* **19**, 228-34 (2014).
358. Hong, M.G. *et al.* A genome-wide assessment of variability in human serum metabolism. *Hum Mutat* **34**, 515-24 (2013).
359. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* **44**, 269-76 (2012).
360. Nicholson, G. *et al.* A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet* **7**, e1002270 (2011).
361. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* **4**, e1000282 (2008).
362. Inouye, M. *et al.* Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol* **6**, 441 (2010).

363. Laurila, P.P. *et al.* Genomic, transcriptomic, and lipidomic profiling highlights the role of inflammation in individuals with low high-density lipoprotein cholesterol. *Arterioscler Thromb Vasc Biol* **33**, 847-57 (2013).
364. Bauer, D.E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253-7 (2013).
365. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-23. (2013).
366. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-6 (2013).
367. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-7 (2014).
368. Bain, B.J. Neonatal/newborn haemoglobinopathy screening in Europe and Africa. *J Clin Pathol* **62**, 53-6 (2009).
369. Ansong, D., Akoto, A.O., Ocloo, D. & Ohene-Frempong, K. Sickle cell disease: management options and challenges in developing countries. *Mediterr J Hematol Infect Dis* **5**, e2013062 (2013).
370. Odunvbun, M.E., Okolo, A.A. & Rahimy, C.M. Newborn screening for sickle cell disease in a Nigerian hospital. *Public Health* **122**, 1111-6 (2008).
371. Dalle, J.H. Hematopoietic stem cell transplantation in SCD. *C R Biol* **336**, 148-51 (2013).
372. Kharbanda, S. *et al.* Unrelated Donor Allogeneic Hematopoietic Stem Cell Transplantation for Patients with Hemoglobinopathies Using a Reduced-Intensity Conditioning Regimen and Third-Party Mesenchymal Stromal Cells. . *Biol Blood Marrow Transplant.* **S1083-8791**(2013).
373. Ruggeri, A. *et al.* Umbilical cord blood transplantation for children with thalassemia and sickle cell disease. *Biol Blood Marrow Transplant* **17**, 1375-82 (2011).
374. Majhail, N.S., Mau, L.W., Denzen, E.M. & Arneson, T.J. Costs of autologous and allogeneic hematopoietic cell transplantation in the United States: a study using a large national private claims database. . *Bone Marrow Transplant.* **48**, 294-300 (2013).
375. 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**, 323-7 (2009).
376. Pawliuk, R. *et al.* Correction of sickle cell disease in transgenic mouse models by gene therapy. *Science* **294**, 2368-71 (2001).
377. Samakoglu, S. *et al.* A genetic strategy to treat sickle cell anemia by coregulating globin trans-gene expression and RNA interference. . *Nat Biotechnol.* **24**, 89-94 (2006).
378. Cavazzana-Calvo, M. *et al.* Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. *Nature* **467**, 318-22 (2010).
379. Bank, A., Dorazio, R. & Leboulch, P. A phase I/II clinical trial of beta-globin gene therapy for beta-thalassemia. *Ann N Y Acad Sci* **1054**, 308-16 (2005).
380. Sadelain, M. *et al.* Strategy for a multicenter phase I clinical trial to evaluate globin gene transfer in beta-thalassemia. *Ann N Y Acad Sci* **1202**, 52-8 (2010).