Université de Montréal

Genomic architecture of Sickle Cell Disease clinical variation in children from West Africa: a case-control study design

par

Jacklyn Quinlan

École de santé publique

Faculté de médecine

Thèse présentée à la faculté des études supérieures

en vue de l'obtention du grade de Ph.D.

en santé publique

option épidémiologie

Août, 2013

Résumé

Contexte : L'anémie falciforme ou drépanocytose est un problème de santé important, particulièrement pour les patients d'origine africaine. La variation phénotypique de l'anémie falciforme est problématique pour le suivi et le traitement des patients. L'architecture génomique responsable de cette variabilité est peu connue.

Principe : Mieux saisir la contribution génétique de la variation clinique de cette maladie facilitera l'identification des patients à risque de développer des phénotypes sévères, ainsi que l'adaptation des soins.

Objectifs : L'objectif général de cette thèse est de combler les lacunes relatives aux connaissances sur l'épidémiologie génomique de l'anémie falciforme à l'aide d'une cohorte issue au Bénin. Les objectifs spécifiques sont les suivants : 1) caractériser les profils d'expressions génomiques associés à la sévérité de l'anémie falciforme ; 2) identifier des biomarqueurs de la sévérité de l'anémie falciforme ; 3) identifier la régulation génétique des variations transcriptionelles ; 4) identifier des interactions statistiques entre le génotype et le niveau de sévérité associé à l'expression ; 5) identifier des cibles de médicaments pour améliorer l'état des patients atteints d'anémie falciforme.

Méthode : Une étude cas-témoins de 250 patients et 61 frères et soeurs non-atteints a été menée au Centre de Prise en charge Médical Intégré du Nourrisson et de la Femme Enceinte atteints de Drépanocytose, au Bénin entre février et décembre 2010.

Résultats : Notre analyse a montré que des profils d'expressions sont associés avec la sévérité de l'anémie falciforme. Ces profils sont enrichis de génes des voies biologiques qui contribuent à la progression de la maladie : l'activation plaquettaire, les lymphocytes B, le stress, l'inflammation et la prolifération cellulaire. Des biomarqueurs transcriptionnels ont permis de distinguer les patients ayant des niveaux de sévérité clinique différents. La régulation génétique de la variation de l'expression des gènes a été démontrée et des interactions ont été identifiées. Sur la base de ces résultats génétiques, des cibles de médicaments sont proposées.

Conclusion: Ce travail de thèse permet de mieux comprendre l'impact de la génomique sur la sévérité de l'anémie falciforme et ouvre des perspectives de développement de traitements ciblés pour améliorer les soins offerts aux patients.

Mots clés : Drépanocytose, anémie falciforme, génomique, eSNP, expression, interactions, biomarqueurs, pharmacogénetique, Afrique Sub-Saharienne

Abstract

Background: Sickle Cell Disease (SCD) is an important public health issue, particularly in Africa. Phenotypic heterogeneity of SCD is problematic for follow-up and treatment of patients. Little is known about the underlying genomic architecture responsible for this variation.

Rationale: Understanding the genetic contribution to the inter-patient variability will help in identifying patients at risk of developing more severe clinical outcomes, as well as help guide future developments for treatment options.

Objectives: To characterize genome-wide gene expression patterns associated with SCD clinical severities and to identify genetic regulators of this variation. More specifically, our objectives were to associate gene expression profiles with SCD severity, identify transciptional biomarkers, characterise the genetic control of gene expression variation, and propose drug targets.

Methods: A case-control population of 250 SCD patients and 61 unaffected siblings from the National SCD Center in Benin were recruited. Genome-wide gene expression profiles and genotypic data were generated.

Results: Genome-wide gene expression patterns associated with SCD clinical variation were enriched in B-lymphocyte development, platelet activation, stress, inflammation and cell proliferation pathways. Transcriptional biomarkers that can discriminate SCD patients with respect to clinical severities were identified. Hundreds of genetic regulators were significantly associated with gene expression variation and potential drug targets are suggested.

Conclusion: This work improves our understanding of the biological basis of SCD clinical variation and has the potential to guide development of targeted treatments for SCD patients.

Keywords: Sickle cell disease, genomics, eSNP, transcriptomics, interactions, biomarkers, drug repurposing, Sub-Saharan Africa.

Table of content

List of Tables

List of Figures

Abbreviations

A	Acute
AI	Arab-Indian
ANCOVA	Analysis of Covariance
ATYP	atypical haplotypes
AUC	Area under the curve
BEN	Benin
bp	base pairs
C	Control
CAM	Cameroon
CAR	Central African Republic
CDCV	Common disease, common variant
CEPH	Centre d'Étude du Polymorphisme Humain
ClinCat	Clinical Category
ClinStat	Clinical status
CPMI-NFED	Centre de Prise en charge Médicale Intégrée du Nourrisson et de la Femme Enceinte atteints de Drépanocytose
CTD	Comparative Toxicogenomics Database
Ctl	Control
DNA	Deoxyribonicleic Acid
E	Entry
ePC	expression principal component
eQTL	expression quantitative trait loci
eSNP	expression SNP
F	female
FDA	Food drug administration
FDR	False discovery rate
FPR	false positive rate
FU	Follow-up
gPC	genotypic principal component
GSEA	Gene set enrichment analysis
GWA	Genome wide association
GWAS	Genome wide association study
GxE	gene-by-environment
Hb	hemoglobin
HbF	fetal hemoglobin
hg19	human genome version 19

HP	Haptoglobin
HPFP	hereditary persistance of fetal hemoglobin
HPLC	High performance liquid chromatography
HU	hydroxyurea
HWE	Hardy-Wienberg Equilibrium
IBD	Identity-by-Descent
IBS	Identity-by-State
M	male
MAF	Minor allele frequency
MCHC	mean corpuscular hemoglobin concentration
MCV	Mean corpuscular volume
mRNA	messanger RNA
NES	Normalized Enrichment Score
NLP	negative log10 p-value
NSAIDs	non-steroidal anti-inflammatory drugs
NSCDC	National Sickle Cell Disease Center
OR	Odds ratio
PBMC	Peripheral blood mononuclear cells
PCA	Principal component analysis
pval	p-value
QNM	quantile normalization method
RBC	red blood cell
RNA	Ribonucleic Acid
ROC	Receiver Operating Curve
SCD	Sickle cell disease
SEN	Senagal
SNP	Single nucleotide polymorphism
SSS	steady-state satisfactory
SSU	steady-state unsatisfactory
SV	Severity Score
TPR	true positive rate
VCA	Variance component analysis
VOC	vaso-occlusive crisis
WBC	white blood cell
WHO	World Health Organisation
YRI	Yoruban

To Xavier and Dominick, I love you more than words can describe...

“The sequencing of the human genome offers the greatest opportunity for epidemiology since John Snow discovered the Broad Street Pump”

– Shpilberg

“Knowing is not enough; we must apply. Willing is not enough; we must do.”

-- Goethe

Acknowledgements

First and foremost, I am indebted to all the sickle cell patients and their families who participated in this study. Hopefully, this work will be a positive step toward an improved and healthier future for them.

I would like to express my gratitude to my supervisor, Dr. Philip Awadalla and his team at the CHU Ste. Justine. Dr Awadalla's expertise and passion in genetics and genomics have added considerably to my graduate experience. Over the years, he provided me with invaluable direction, support and friendship, for which I am grateful. I am grateful to Dr. Cherif Rahimy and Dr. Ambel Sanni and all other researchers and nurses in Benin for their support in this collaborative research experience.

I am especially indebted to Dr. Youssef Idaghdour for his support, guidance, and mentorship during my PhD. Youssef, you are a great friend and I am so lucky to have worked with you. I am also very thankful for Julie Hussin's input, guidance, and friendship. You too are very special and I will treasure our friendship. Elias Gbeha thank you for your help and support in the field work and in the laboratory genomic experiments. Your attention to detail was very much appreciated!

I am also indebted to all my friends and colleagues who work at the Ste-Justine research center. In particular, I would like to acknowledge Ferran Casals, Thibault de Maliard, Vanessa Bruat, Jean-Philippe Goulet, Jean-Christophe Grenier, Alan Hodginskon, Melanie Capredon, Heloise Gauvin, Elodie Hip-Ki, Claude Bherer, and Armande Ang Houle.

For guidance and support during my academic training, I would like to express my gratitude to Marie-Pierre Dubé, as well as Marie-Hélène Roy-Gagnon, Anita Koushik, and Maria-Victoria Zunzunegui from the Département de Médicine Sociale et Préventive, Université de Montréal.

During the first three years of my doctoral studies, my research was generously supported through a PhD grant from the Fonds de la Recherche en Santé - Quebec (FRSQ). I then received support from the Réseau de la Médecine en Génétique Appliqué (the Louis-Dallaire fellowship) and finally I had a « Bourse de Fins d'Études » from the Département de Médecine Sociale et Préventive from l'Université de Montréal. I gratefully acknowledge their support.

I would like to also express my gratitude to the members of my PhD committee for taking time out from their busy schedules to review this thesis.

Finally, I am indebted to my family and friends for their encouragement and faith in me, even when I wasn't sure of my own strength. Thanks mom and dad for lovingly supporting me and caring for my children while I accomplished this great feat! To my sons, Xavier and Dominick– I love you so much and I am priviledged to have you in my life. Because of you, I found the strength to continue through the tough times over past few years.

A. INTRODUCTION

A.1 Research problem

Sickle Cell Disease (SCD) is an important public health issue [1] affecting millions throughout the world, especially pediatric patients of African ancestry. SCD is the single most important genetic cause of childhood mortality world-wide. An estimated 2.28 per 1000 of all conceptions worldwide are affected by SCD, with higher estimates among populations whose ancestors originate from Sub-Saharan Africa. Among these populations, estimates of SCD prevalence increases to 1 in 500 in African American populations, and 1 in 100 in populations from Nigeria, Benin, and other West African countries [2].

In 2006, the World Health Organisation issued a report that specifically addressed SCD as a prevalent medical condition that contributed to the under-5 death rate on the African continent. An urgent need to provide care and research was emphasized in this report [2]. In 2008, the United Nations recognized SCD as a global health priority and urged all African countries to have a plan to reduce under-5 mortality rates [3]. In order to accomplish this, active North-South and South-South partnerships that prioritize research are required to help sub-Saharan African countries develop robust sickle cell strategies that can provide diagnosis, management, and treatment of SCD.

A major problem in managing SCD is that the underlying phenotypic heterogeneity observed in patients is not understood. Little is known about the underlying genomic architecture responsible for this phenotypic diversity of SCD, which is most likely the result of a combination of host genetic and environmental

factors [1,4]. Understanding the genetic contribution to the inter-patient variability would help identify patients at risk of developing more severe clinical outcomes, as well as help guide future developments in treatment options [5,6].

There is a lack of disease specific treatments available for SCD patients. Drugs used to treat SCD related phenotypes are either non-specific or have a number of side effects and/or contraindications. The FDA has approved only one drug, hydroxyurea (HU), for the specific treatment of SCD related events. However, because of the number of contraindications associated with this medication, the drug's usefulness is limited. Also, HU is not used in developing countries, where comorbidities including malaria and nutritional deficiencies may affect the toxicity profile [7]. New drugs are therefore urgently needed for the treatment of this disease.

In this study, we hypothesized that, in addition to environmental factors, genetic factors impact SCD clinical phenotypes. We aimed to identify genetic factors influencing the disease by integrating the joint analysis of genotyping and gene expression data. The work in this thesis was performed to conduct an in depth study to characterize genome-wide gene expression patterns that are associated with SCD clinical variation in pediatric patients from Benin, West Africa. Using gene expression variation data, we assessed the presence of transcriptional biomarkers that discriminate SCD patients with different clinical severities. Through genome-wide association studies, genetic determinants were tested for association with gene expression levels. This enabled us to better characterize the biological basis of genetic variation as it relates to clinical variation. In a final approach, we evaluated our genetic findings for the presence of known or novel

drug targets, with an emphasis on therapeutic interventions. This work has provided important information which can be used in the development of novel therapeutics, or in repurposing existing ones, and provided guidance for targeted therapies.

A.2 Overview

The research presented in this thesis explored the genetic contribution to gene expression variation in Sickle Cell Disease (SCD) patients with different clinical severities. The results from this work have helped advance the identification of patients at risk of developing more severe clinical outcomes, as well as helped in guiding future developments for treatment options.

The thesis begins with an introduction on Sickle Cell Disease and genomic approaches in public health. The methods and results sections are presented in order to address the four main objectives: to associate gene expression profiles with SCD severity, identify transcriptional biomarkers, characterise the genetic control of gene expression variation, and propose drug targets. In the discussion and conclusion chapters, the implications of the results and examples of how they improve our understanding of the biological basis of SCD clinical variation is explained, as well as how they help guide future research on SCD.

A.3 Sickle Cell Disease (SCD)

Sickle cell disease (SCD) is a congenital, life-long blood disorder characterized by the presence of sickled haemoglobin. It is inherited in an autosomal recessive manner from a single point mutation in the hemoglobin beta gene, either as a homozygote (eg. Hb SS) or as compound heterozygotes (eg.

Hb SC). Sickled hemoglobin affects the globin chain structure, which polymerizes upon deoxygenation, distorting red blood cells (RBCs) into elongated "sickled" cells and consequently leading to their inability to transport oxygen and function effectively.

A.3.1 Epidemiology of SCD

Epidemiologic studies show that the incidence and prevalence of SCD varies significantly depending on geographic location and ethnic background. About 7% of the world's population carries genes responsible for hemoglobinopathies. Each year approximately 300,000-400,000 infants are born with major haemoglobin disorders [1]. More than 200,000 sickle-cell incidents arise yearly in Africa alone [2]. The prevalence of the sickle-cell trait (in which carriers have inherited only one copy of the mutant gene) ranges between 10% and 40% across equatorial Africa and decreases to between 1% and 2% on the North African coast and <1% in South Africa [2]. This distribution is a reflection of the fact that the sickle-cell trait confers a survival advantage against malaria infection among the carriers of a single copy of the SCD mutation, a well-known evolutionary phenomenon referred to as "heterozygote advantage". The selection pressure due to malaria has resulted in an increase in the frequency of the SCD mutant gene in areas of high malarial transmission. In West African countries, the prevalence of SCD is between 15% and 30%, whereas in East Africa, the prevalence of the SCD trait shows marked tribal variations, reaching as much as 45% among the Baamba tribe who live in the west of Uganda. [2].

Table A.1 below shows conservative estimates of prevalence by region as established by the World Health Organisation [8]. At least 5.2% of the world's

population carries a sickle cell variant. Over 70% of all conceptions that have the SCD mutation occur in Africa. Around 2.28 per 1000 conceptions worldwide are affected [8]. Most children born with SCD in high-income countries survive the disease into adulthood but suffer from a chronic disorder, while most children with SCD born in low-income countries die before the age of 5 years. Haemoglobin disorders account for 3.4% of child mortality in children aged under 5 years worldwide and 6.4% of children who die before the age of 5 in Africa [2].

Table A.1. World Health Estimates of SCD prevalence**.**

	Demography 2003				% of pop carrying	Affected SCD	Affected births
Region	Population (millions)	Birth Rate	Annual Births (1000s)	Under-5 mortality rate	sig variant	conceptions (per 1000)	(% of under-5 mortality)
African	586	39	22895	168	18.2	10.68	6.4
American	853	19.5	16609	27	3.0	0.49	2
Mediterranran	573	29.3	16786	108	4.4	0.84	1.4
European	879	11.9	10459	25	1.1	0.07	0.8
South-east Asian	1564	24.4	38139	83	6.6	0.68	1.6
Western Pacific	1761	13.6	23914	38	3.2	0	2
World	6217	20.7	128814	81	5.2	2.28	3.4

* Source: Adapted from Table 1 of World Health Report [8]

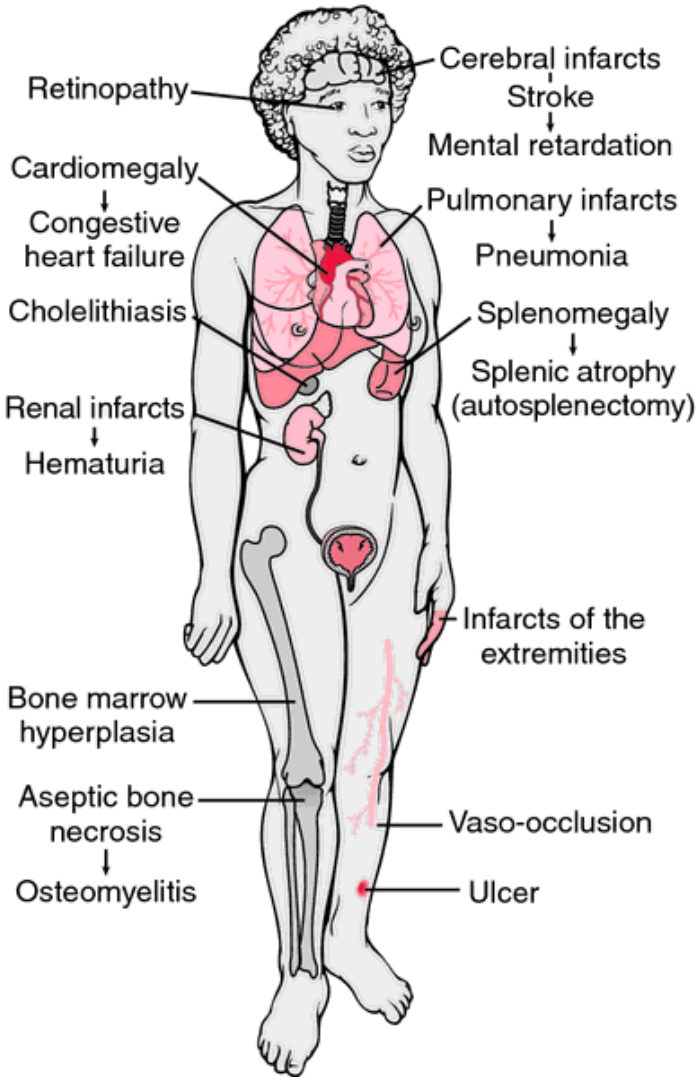
A.3.2 Phenotypic variation and clinical heterogeneity of SCD

As early as 1976, researchers were aware of the phenotypic variability of SCD. At the time, the causal mutation was not known. A quote from Norman Davidson captures the irony of this simple Mendelian disease with a complex phenotype, that even today, we don't understand:

"I remember asking a new graduate student, Harvey Itano, what his research problem was. He said he was going to test your (Linus Pauling's) hunch that there was a difference in hemoglobin molecules between normal people and those with sickle cell anemia. I thought that was a crazy idea; a complicated human disease could not have any such simple cause" [9].

Patients with SCD suffer a wide variety of disease complications [10]. Some individuals have mild manifestations of the disease, which can be clinically unapparent; while others have severe complications [11] that recur frequently, affecting the entire body. Death can occur suddenly and unexpectedly even in apparently stable SCD patients. The underlying cause of this clinical variation remains unknown, but is most likely influenced by a combination of genetic and environmental factors. While some SCD patients have only one crises event every few years, others have several severe and painful crises events every year, often requiring hospitalisation. Figure A.1 below shows the main SCD related clinical phenotypes.

Figure A.1. Main clinical manifestations observed in SCD patients.



* Figure taken from : http://adkteamtalk.wordpress.com/author/adkteamtalk/

Complications of SCD are often from obstructed circulation. Red blood cells often become trapped in the spleen of children with SCD, leading to a serious risk of death from a sudden profound anaemia associated with rapid enlargement of the spleen. Reduced spleen function associated with the condition causes children with SCD to be susceptible to bacterial invasions, bone infections (osteomyelitis), gallbladder infections (cholecystitis), lung infections (pneumonia), and urinary tract infections.

Vaso-occlusive (VOC) events are also common in SCD and occur when tissue damage results from obstructed blood flow. They can be recurrent and are unpredictable, lasting from hours to days. Many affected children present with painful swelling of the hands and/or feet (hand-foot syndrome) when blood flow is blocked to the extremities due to VOC events. Vaso-occlusion can also cause “acute chest syndrome” (pneumonia or pulmonary infarction), bone or joint necrosis (tissues damage caused by insufficient oxygenation of the bone), priapism (painful and prolonged erection), renal failure, and cardiac problems. Poor eyesight or blindness, and ulcers on the lower legs (in affected adolescents and adults) are some other symptoms that may occur as a result of blockage in small blood vessels due to the disease. Cerebrovascular complications from VOC include confusion due to transient ischemic attacks, ischemic strokes, and hemorrhagic strokes, sometimes associated with seizures.

SCD patients experience severe anemia which causes fatigue, paleness, rapid heart rate, shortness of breath, and, or jaundice. Other complications include delayed growth, delayed puberty, and painful joints. SCD patients commonly experience chronic pain in the back bones, the long bones, and in the chest.

Although boys and girls are equally at risk to inherit the mutation that causes SCD, previous studies found that men with sickle cell disease experience more sickle cell crises after puberty than do women. It has also been noted that the median age of death from SCD for men is 42 compared to 48 years for women. Recently, this gender difference for SCD severity was attributed to nitric oxide being significantly higher in woman than in men [12]. Nitric oxide, which is stimulated by estrogen production in women, helps blood vessels to dilate, reducing obstruction.

A.3.3 Genetics of hemoglobin in unaffected and SCD patients

In SCD, the causal mutation is known and is inherited in an autosomal recessive manner [13]. The most common form of SCD is associated with patients homozygous for a single base mutation in codon 6 of the β-globin gene (HbSS) causing a single amino acid substitution in position β6 (glutamic acid to valine) and results in β^s. The second most common abnormal Hb mutation in West Africa, β^c, is also caused by a single mutation/amino acid change at the same position in the β-globin gene, but with lysine replacing glutamic acid. See Table A.2 below.

Table A.2 Adult hemoglobin variants and their corresponding genotypes.

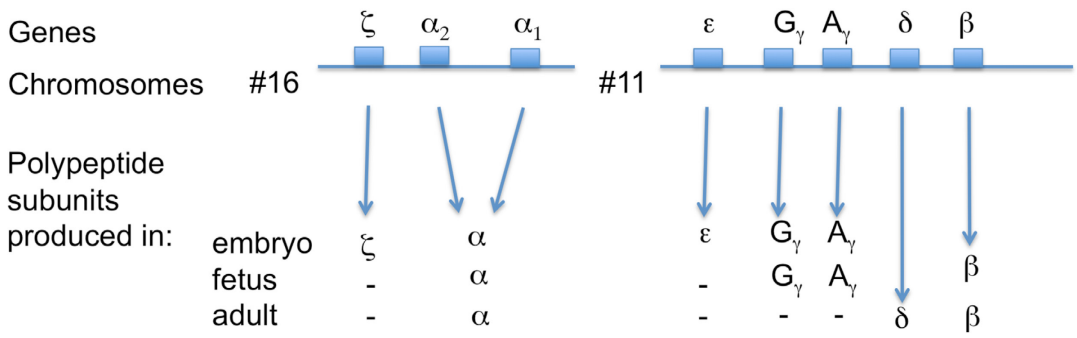
Hemoglobin	Hemoglobin variant	Genotype at codon 6 of β globin gene
HbA (normal)	β6Glu	GAG
HbS	β6Val	GUG
HbC	β6Lys	AAG

In unaffected individuals, β-globin is encoded by a structural gene found in a cluster with four other β-like genes on chromosome 11. The cluster contains five functional genes, ε, $^G\gamma$, $^A\gamma$, δ, β, which are arranged in the order of their developmental expression [14]. ε is an embryonic globin gene expressed primarily in yolk sac-derived cells from 3 to 8 weeks of gestation; $^G\gamma$ and $^A\gamma$ are fetal globin genes expressed primarily in fetal liver-derived cells from 6 weeks of gestation to 6 months after birth; and δ and β are adult globin genes expressed primarily in bone marrow-derived cells starting shortly before birth and persisting throughout adult life. The β gene is responsible for 97-98% of adult β-globin, and the δ gene is responsible for ~2-3% [15]. Upstream of the entire β-globin complex is the locus control region, a regulatory element essential for the expression of all the genes in the complex [14].

β-globin gene products pair up with α-globin gene products to form hemoglobin tetramers that surround a heme core that binds oxygen. α-globin genes are found on chromosome 16, including ζ chain expressed in the embryonic stage, and α chain expressed in the fetal and post-natal stages. β-like chains pair with the α-like globin ζ chain in the embryonic life, and with α-chain in fetal and post-natal life to form the developmentally regulated hemoglobin molecules [16]. The change in globin production from embryonic to fetal and from fetal to adult is called hemoglobin switching [17]. Homozygous HbSS patients have ($\alpha_2\beta^s{}_2$) hemoglobin tetramers, whereas HbSC compound heterozygotes have ($\alpha_2\beta^s\beta^c$) tetramers. See Figure A.2.

Figure A.2 Chromosomal arrangements of globin genes.

The α-hemoglobin genes are located on chromosome 16 and the β-hemoglobin genes are located on chromosome 11. Both α-hemoglobin and β-hemoglobin genes are ordered by their developmental expression along the chromosome. The tetramer proteins that are formed from the hemoglobin genes are shown in the center.



Hemoglobins

embryo	$\zeta_2\varepsilon_2$	Gower I
	$\zeta_2\gamma_2$	Portland I
	$\alpha_2\varepsilon_2$	Gower II
fetus	$\alpha_2\gamma_2$	HbF
	$\alpha_2\beta_2$	HbA
adult	$\alpha_2\beta_2$	HbA
	$\alpha_2\delta_2$	HbA_2

A.3.4 Historical perspective of SCD

The first documented report of SCD was by Dr. James Herrick and Dr. Ernest Irons in a dental student from Grenada in 1904 [18]. They identified irreversibly sickled red blood cells, which they described as “very irregular and with many elongated forms.” The autosomal recessive inheritance of the disease was described by Neel and Beet in 1947. In 1949, Pauling *et al.* demonstrated that the HbS chain in SCD patients had an abnormal electrophoretic mobility which led to the proposal that this was a molecular disease of Hb [13]. This was the first time a genetic disease was linked to a mutation of a specific protein. Ingram et al. demonstrated that the SCD mutation caused a single amino acid change. This was followed by analysing the structure and physical properties of HbS, which formed intracellular polymers upon deoxygenation. These studies put SCD at the leading edge of investigations to elucidate the molecular basis of human diseases. It was at this time that it was noted [19] that infants with SCD rarely have clinical manifestations in the first year of life.

A.3.5 Natural history of SCD

Based on the observations made on SCD infants described above, it was proposed that the high levels of fetal Hb in the red blood cells, which persists during the first year of life, somehow protects SCD infants from clinical manifestations. Childhood clinical manifestations of SCD are seen when the switch from fetal Hb to adult Hb occurs.

It is interesting to note that in a small proportion of individuals (0-0.8% prevalence [20]), hereditary persistence of fetal hemoglobin (HPFH) occurs. HPFH alleviates the severity of SCD since HbF can bind to oxygen with greater affinity than with adult hemoglobin, thus reducing the aggregation of mutated hemoglobin and the sickling of red blood cells. Persistence of fetal hemoglobin provides a hallmark

example of a genetic modifier that results in a decrease in the number of painful episodes in patients with SCD who have a hereditary persistence of fetal Hb.

Once the switch to adult Hb has occurred, the average life expectancy of SCD patients varies, with SCD patients in North America having significantly longer life expectancies than African patients, most likely due to newborn screening, superior health care and follow-up programs, and access to medication. The peak incidence of death among those affected with sickle cell disease is between 1-3 years of age in the USA and in Jamaica; and the median survival of patients with sickle cell disease has been estimated to be 42 years for men and 48 years for women in a Jamaican cohort; and 53 years for men and 58.5 years for women in an American cohort [2].

Survival of patients with SCD on the African continent is dismal [1]. Historically, the life expectancy of sickle cell patients in Africa has been assessed against the yardstick of infant under-five mortality. Using this yardstick, sickle-cell disease contributes the equivalent of 5% of under-five deaths on the African continent, more than 9% in West Africa, and up to 16% of under-five deaths in individual West African countries [2]. In Benin, West Africa, no reliable estimate of mortality rate for SCD children is presently available, however, a rough estimate has been documented that more than 50% of the untreated affected children do not reach their fifth birthday [21]. A recent estimate of between 50-90% of affected children born in Africa with SCD will die before age 5 years, due either to complications of SCD itself or more commonly from complications from pneumococcal disease, malaria, or diarrheal disease [22].

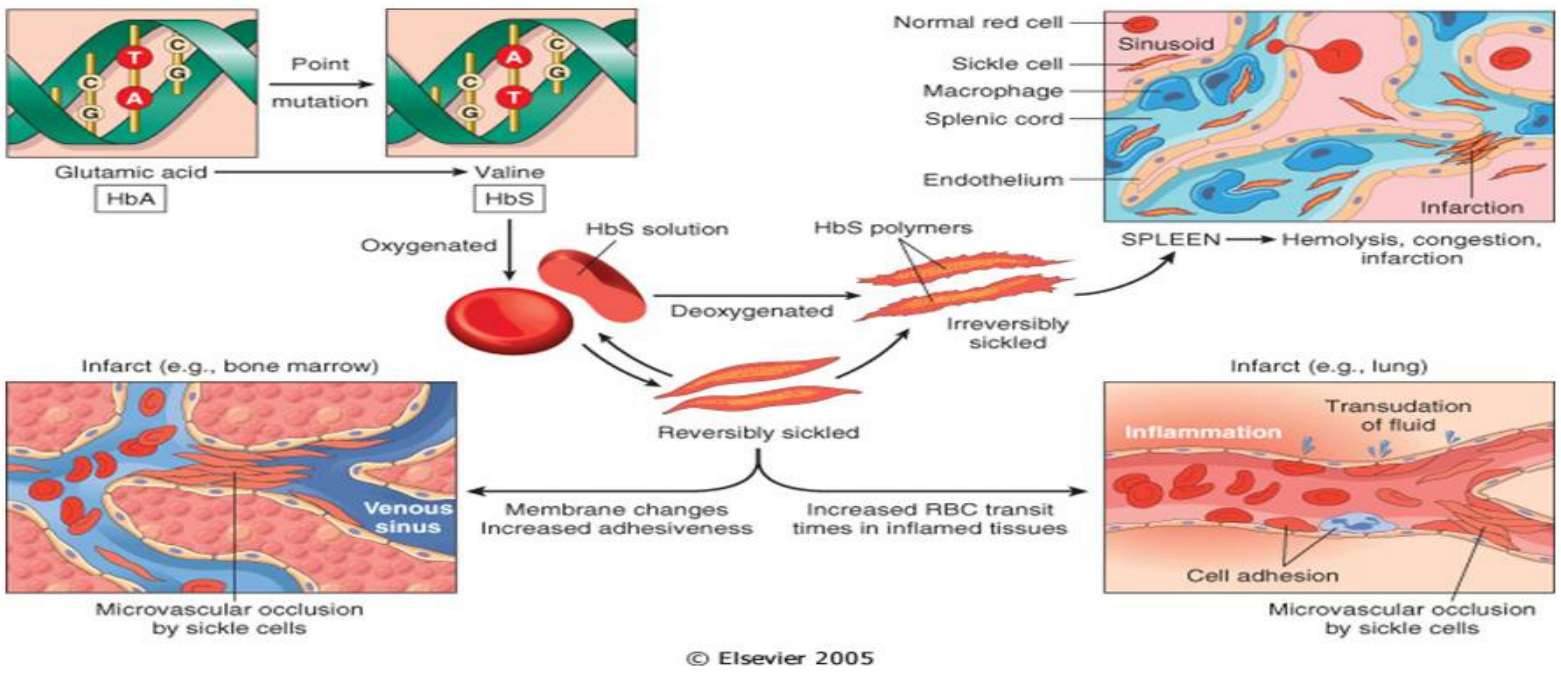
A.3.6 Pathophysiology of SCD

The primary pathophysiological event in the erythrocytes of individuals with SCD is thought to be polymerization of sickle haemoglobin (HbS) [23]. HbS polymerization causes structural damage to the red blood cell membrane, which alters the properties of the erythrocyte, impairs blood flow through the microvasculature, and leads to hemolysis, vaso-occlusive episodes, tissue ischemia and organ dysfunction. The extent of polymerization is determined by the intracellular haemoglobin composition (% HbS, % HbA, and % HbF), total Hb concentration, oxygen saturation, and intracellular pH. Intracellular HbS polymerization leads to a marked decrease in the flexibility of the sickle erythrocytes and obstruction in various microcirculatory beds, as well as chronic anaemia.

Other abnormalities in the properties of the sickled erythrocytes include membrane abnormalities, adhesion between red blood cells, changes in ion fluxes and volume, and endothelial adhesion [24]. These properties result from polymerization events and may in turn increase polymerization. The adhesive interactions between sickled red blood cells and the endothelium is due to the inflammatory reaction of the disease and leads to the microvascular pathogenesis [25]. Decreased flexibility and permeability of the red cell membrane may be equally important pathophysiological events for the onset and frequency of SCD crisis [26]. In Figure A.3, the pathophysiology of SCD is described. Through polymerisation of the hemoglobin molecules, ensuing erythrocyte membrane damage and hemolysis occurs. Intravascular hemolysis of erythrocytes leads to nitric oxide scavenging by plasma hemoglobin and provokes hemolysis-related complications [27].

Figure A.3 Pathophysiology of SCD.

Normal red blood cells that contain homozygous HbAA are quite elastic which enables them to maneuver so that they can pass through small capillaries. The red blood cells of a person with sickle cell disease (homozygous HbSS) lose this elasticity due to the polymerisation of deoxygenated HbS, which form long stiff rods. These stiff rods exert pressure on the walls of the red blood cell membrane causing it to collapse and distort into a variety of shapes including the classic shape of a farmer's sickle, from which SCD derives its name. When re-oxygenated in the lungs, the red blood cells regain their original round shape. However with repeated sickling (deoxygenation) and unsickling (oxygenation) there is increasing crystallization and formation of stiff rods within the red blood cells, causing them to become increasingly hard, brittle and consequently irreversibly sickled. Sickled red blood cells are fragile and break easily. The damaged cells are rapidly destroyed by the reticulo-endothelial cells in the spleen, whereby their lifespan is reduced from the normal 120 days to about 5 to 20 days. This rapid break down of red blood cells in the host leads to the anemia associated with the disease.



* Figure taken from : http://sgugenetics.pbworks.com/w/page/61172304/Pathophysiology%20of%20Sickle%20Cell%20Anemia

A.4 Risk factors of SCD severity

A.4.1 Genetic risk factors of SCD clinical variation

Although SCD is a monogenic disease, the clinical variation manifested in SCD is polygenic, with multiple known genes contributing to its phenotypic variation [28]. Mapping the genetic variants associated with SCD clinical sub-phenotypes have largely been limited to candidate gene approaches and genome-wide association studies.

A.4.1.i SCD causative genotype

The primary genetic determinant of SCD severity is the causative genotype. Individuals who are homozygous for the β^s mutation (HbSS) tend to have the most severe disease followed by those who are compound heterozygous HbSC. HbSS individuals have higher rates of acute chest syndrome and pain crises than individuals with HbSC [29]. However, HbSC individuals have increased risk for thromboembolic complications, retinopathy, and renal papillary necrosis than those with HbSS [29]. While the causative genotype is a key determinant of disease severity, the frequency and severity of complications still vary considerably within each genotypic group.

A.4.1.ii SCD haplotypes

Genetic variation linked to the beta-globin mutation may influence the clinical manifestations of SCD. When these genetic variants that are linked to beta-globin mutations are located on the same genetic background of non-coding DNA, they form a “haplotype” that is specific for each SCD mutation. Five major beta globin haplotypes exist, named according to the country or region of where they were

initially identified: Senegal (SEN), Benin (BEN), Central African Republic (CAR), Arab-Indian (AI) and Cameroon (CAM) (See table A.3), as well as non-canonical atypical haplotypes (ATYP). The four African haplotypes show broad trends in disease severity. The Central African Republic haplotype tends to have the least favourable clinical course, followed by the Benin and Senegal haplotypes [30]. The ranking of the Cameroon haplotype is uncertain. The Arab-Indian haplotype seems to, on average, produce fewer complications than its African counterparts. No clear explanation exists for the differences in severity between haplotypes; however, one hypothesis states that the haplotypes influence baseline HbF levels [31]. The patterns of severity apply only at the population level. Broad overlap in the clinical manifestations prevents the use of haplotypes to predict the clinical course of SCD in a particular individual. Usually, people with SCD outside Africa (e.g., African-Americans) have mixed haplotypes for their sickle cell genes. Analysis of haplotypes in this setting is even less likely to provide clinically useful information.

The HbC mutation is restricted to individuals of West African descent. This mutation is associated with less severe clinical manifestations than the HbS mutation and lies on a different haplotypic background.

Table A.3 Classic HbS haplotypes

SCD haplotypes	HindIII in Gy1	HindIIIin Gy2	HincII	HinfI
Benin	-	-	+	-
CAR/Bantu	+	-	-	-
Cameroon	+	+	+	+
Senegal	+	-	+	+
Saudi Arabia/India	-	-	+	-

*CAR=Central African Republic
Four restriction fragment length polymorphism (RFLP) sites that differentiate the classic HbS haplotypes. The RFLP sites are located in a 70kb region upstream (5') of the HbS mutation. *Ref: Hanchard et al. BMC Genetics 2007 8:52 doi:10.1186/1471-2156-8-52

A.4.1.iii HbF

Fetal hemoglobin (HbF: $\alpha^2\gamma^2$) is a major genetic modulator of the hematologic and clinical features of SCD, an effect mediated by its exclusion from the sickle hemoglobin polymer. Among the majority of individuals with SCD who do not have hereditary persistence of fetal hemoglobin (HPFH), the residual levels of fetal hemoglobin vary considerably (1-25%). Approximately 40% of this variation is accounted for by the X-linked F-cell production locus (Xp22.3-22.2); β-S cluster haplotypes account for an additional 14% [32]; and common SNPs at the BCL11A (2p16), HBS1L-MYB intergenic region (6q23), and beta-globin loci account for an additional 20% of the variation in HbF levels [33]. The mechanism of action of fetal Hb is through disruption of the polymerization of deoxy-HbS and through diluting the intracellular concentration of HbS [34]. Together, these HbF mechanisms effectively prevent many clinical manifestations of SCD.

The distribution of HbF among red blood cells (RBCs) is important. In HPFH, HbF exists at high levels homogenously in all red cells which protects them from sickling. In the absence of HPFH, patients with high levels of HbF have a heterogeneous distribution of fetal hemoglobin among RBCs. For example, a patient in whom half the cells have 30% Hb F and half have 0%. The patient would have 15% HbF overall. However, half the cells would sickle and occlude blood flow through the microcirculation. These deformed cells would block the flow of the normally shaped high HbF cells. Thus, even though this patient has relatively high overall HbF levels, he would still experience SCD clinical manifestations.

Higher HbF levels were associated with reduced rates of acute painful episodes, leg ulcers, and less frequent acute chest syndromes [35]. Genetic factors that explain a large proportion of HbF variation in SCD have been mapped [36].

However, HbF levels have no clear association with other important SCD clinical manifestations such as stroke and silent cerebral infarction, priapism, urine albumin excretion, and systemic blood pressure [35]. Thus, other factors likely contribute to the remaining variation in SCD clinical phenotypes and are most probably controlled by a complex interplay between both genetic and environmental factors [34,37].

A.4.1.iv α-Thalassemia

Thalasemias are inherited hemoglobinopathies that result from quantitative reductions in globin chain synthesis and lead to imbalanced alpha and beta chain tetramers. Thalassemias are different from SCD, since they don't necessarily produce a mutant form of β globin. However, co-inheritance of thalasemias with SCD affects severity of SCD. Those with diminished β-globin chains are termed β-thalassemias, whereas those with decreased α-chain production are called α-thalassemias. More than 30% of most populations with HbSS SCD carry one or more determinants for α thalasemmia. In people of East African descent, this is usually the α-globin gene deletion ($-\alpha^{3.7}$/). The co-existence of the α-thal ($-\alpha^{3.7}$/) deletion with SCD has a quantitative effect on intracellular Hb S concentration, reducing the frequency of Hb S polymerisation and number of irreversibly sickled cells [34]. Thus, co-existing α-thal ($-\alpha^{3.7}$/) has a protective effect against complications related to severe hemolysis in SCD patients. In parts of West Africa, and in Benin in particular, the estimated prevalence of α thalassemia is small [38].

A.4.1.v Genetic factors of SCD phenotypes identified through candidate gene studies

A number of candidate gene studies have associated genetic variants with various SCD phenotypes. Genetic factors that could potentially affect the pathogenesis and modulate the phenotype downstream of the HbS polymorphism event include those involved in pathways mediating inflammation, oxidant injury, blood coagulation and hemolysis, and vascular remodeling. Based on their pathophysiology, candidate genes and SNPs that could plausibly affect the different sickle-related complications have been selected and tested for association in candidate gene studies. However, most candidate gene association studies were characterised by small sample sizes and most of them do not present with a replication of the findings in another study population. Lack of replication is one of the hallmarks of false positive signals in genetics, typically obtained from underpowered studies. This can lead to contradictory results when studies are compared [28]. The most rigorously studied SCD subphenotypes and associated polymorphisms are shown in Table A.2.

It has been proposed that the TGF-β/Smad/BMP pathway play an important role in multiple subphenotypes of SCD based on consistent associations in candidate gene studies [28]. The TGF-β/Smad/BMP pathway regulates diverse cellular processes and plays roles in inflammation, fibrosis, cell proliferation, hematopoiesis, osteogenesis, angiogenesis, nephropathy, wound healing, and the immune response. The many complications of SCD are affected by most of these processes so it is reasonable to suspect that perturbations of this pathway would modulate a SCD patient's development, progression, and resolution.

Table A.4. Candidate gene studies in SCD sub-phenotypes.

Results from the most rigorously studied SCD sub-phenotypes and associated polymorphisms from candidate gene studies.

SCD sub-phenotype	Genes involved
Survival	TGFBR3
Stroke, silent infarction	VCAM1, ILR4, HLA, LDLR
Painful episodes	MBL2
Acute chest syndrome	No gene has been validated
Bacteremia/infection	CCL5, HLA, IGF1R, TGFβ/SMAD/BMP pathway
Osteonecrosis	BMP6
Priaprism	KL, TEK, TGFBR3, AQP1
Leg ulcers	TGFb/SMAD/BMP pathway, KL, HLA
Sickle vasculopathy	BMP6, TGFBR3, ACVR1, BMP2
Cholelithiasis	UGT1A1 promoter
Renal function	DARC FY, TGFb/SMAD/BMP pathway, MYH9, APOL1

* Source adapted from Steinberg *et al* [28].

A.4.1.vi Genetic factors identified by genome-wide association studies (GWAS)

Seven genome-wide association studies have been performed with SCD related phenotypes [39]. The most recent GWAS was performed to identify genetic factors associated with hemolysis in SCD [40]. The authors used principal component analysis of reticulocyte count, lactate dehydrogenase, aspartate aminotransferase and bilirubin levels, measurements of hemolysis, to compute a haemolytic score that was used as a subphenotype in a genome-wide association study. They identified in one cohort (1,117 patients) and replicated in two additional cohorts (n= 549 and 296 patients) the association of a single nucleotide polymorphism in NPRL3. The HBA1/HBA2 regulatory elements are located in introns of NPRL3. Perhaps by independently down-regulating expression of the HBA1/HBA2 genes, NPRL3 reduces haemolysis in SCD.

Another recent GWAS examined total bilirubin and choletithiasis (gallstones) risk in SCD [41]. When hemolysis occurs, circulating heme increases, leading to elevated unconjugated bilirubin levels and increased incidence of cholelithiasis. In a discovery cohort of 1,117 SCD patients, 15 SNPs were significantly associated with total bilirubin levels. SNPs in UGT1A1, UGT1A3, UGT1A6, UGT1A8, and UGT1A10 were identified. All of these associations were validated in four independent sets of more than 3000 SCD patients.

In a GWAS that identified genetic factors associated with mortality in SCD patients [42], the authors used a Bayesian network model to construct disease severity with 24 clinical events and laboratory tests and obtained a score that predicted mortality. The analysis was performed in two independent patient groups. The first patient group consisted of 1,265 patients with either “severe” or “mild” SCD disease based on the network model of disease severity, and discovered 40 SNPs that were strongly associated with SCD severity. Thirty-two of the 40 SNPs were analysed in an independent set of 163 patients. Five of these SNPs were significantly replicated, 8 showed consistent effects, although did not reach statistical significance, and 19 did not show any convincing significance. Among the replicated associations are SNPs in KCNK6, a potassium channel gene. Using an analytical method that examined genetic regions, 27 genes with a strong enrichment of significant SNPs were present and 20 were replicated with varying degrees of confidence. Among the novel genes identified by this analysis as being associated to SCD was the telomere length regulator gene TNKS.

These studies were the first to use GWAS to understand the genetic diversity that accounts for phenotypic heterogeneity of SCD. Nonetheless, a large part of the clinical variation in SCD remains to be explained. The genetic factors associated with

SCD clinical phenotypes listed in Table A.3 and in Appendix I, include 111 genes overall [39]. These genes were identified through candidate genes association studies, GWAS, or gene expression studies.

Table A.5 Genetic factors associated with SCD.

Genetic factors associated with clinical phenotypes observed in SCD that are reported in Human genome epidemiology (HUGE) [39] are shown below. A total of 143 publications on SCD are reported in HUGE, 68 of which are studies that identified genetic factors associated with SCD clinical phenotypes. These 68 publications are grouped by SCD clinical phenotype, study type (candidate gene study, GWAS, or gene expression study), sample size, significant genes, replication cohort size (if done), significant genes replicated, significance (either p-value or Odds Ratio and confidence intervals), first author and year of the publication.

Phenotype	Study type	Sample size (n)	Significant genes	Replication cohort	Significantly replicated genes	pval	OR	Author	Year
Acute chest syndrome (ACS)	candidate gene	186 SCD patients, 86 SCD with ACS	DRB1 HLA haplotype			0.018		Mahdi N	2009
	candidate gene	942 SCD patients	HMOX1				0.25 (0.1-0.81)	Bean CJ	2012
	candidate gene	95 SCD patients with ACS, 62 without	ET-1, ecNOS			significant		Chaar V	2006
	candidate gene	not indicated	eNOS			0.0061		Sharan K	2004
Avascular necrosis (AVN)	candidate gene	SCD with and without AVN	MTHFR			0.006		Kutlar A	2001
Bilirubin, gallstones, cholelithiasis	candidate gene	263 SCD	UGT1A1			0.0001		Vasavda N	2007
	candidate gene	153 SCD	UGT1A1			0.003		Martins R	2008
	candidate gene	171 SCD children 153 SCD adults	UGT1A1			significant		Chaar V	2005
	candidate gene	115 SCD	UGTA1A			0.001-0.002		Passon RG	2001
	GWAS	Discovery n=1117	UGT1A1, UGT1A3, UGT1A6, UGT1A8, UGT1A10	4 Replication cohorts	UGTA	5×10^{-8}		Milton	2012
	candidate gene	324 SCD patients	UGT1A1			0.0001		Carpenter SL	2008
	candidate gene	541 SCD patients and 111 controls	UGT1A1			10^{-5}, 10^{-3}	11.3 ($p-7x10^{-4}$)	Haverfield EV	2005
Kidney function, WBC counts, HU use	candidate gene	227 SCD	Duffy gene (FY)			0.002-0.03		Afenyi-Annan	2008
Chronic transfusions	candidate gene	89 SCD patients transfused	HFE			not significant		Jeng MR	2003
Morphine clearance	candidate gene	24h PK study of morphine clearance in 20 SCD	UGT2B7			0.03		Darbari DS	2008
SCD complications	candidate gene + exp	142 SCD patients and 102 controls	L-selectin and its expression			no association		Ugochukwu CC	2008
Haplotypes	candidate gene	82 Kuwait SCD patients 54 Nigerian SCD patients	Hp-1 Hp-2			0.05		Adekile	2010
Hb F levels	candidate gene	91	Xmn1, HBBP1, intergenic region b/t HBS1L and MYB			0.001, 0.002 and 0.013		Roy P.	2012
	candidate gene	Discovery= 895 SCD patients	BCL11A, HBS1L-MYB, Gamma-globin	Brazil replication cohort= 350		0.04×10^{-42}		Lettre G	2008
	haplotype analysis	47 SCD patients	Benin and Bantu haplotypes					Carvalho-dos Santos	2012
	GWAS	179 individuals from 95th upper and lower %	BCL11A	Replication cohort=90		4x10-16		Menzel S	2007
	candidate gene	4 samples: 177, 631, 87, 75	6q23 region			0.05, 0.002, 0.019, $1.5x10^{-7}$		Creary LE	2009
	candidate gene	57	Xmni gamma			0.002		Nguyen TK	2010
	GWAS	Discovery=848 SCD patients	OR51B5 OR51B6 BCL11A	Replication cohort=305 SCD	OR51B5 OR51B6 BCL11A	$4.7x10^{-8}$		Solovieff N	2010
	candidate gene	131 cases 121 controls	KLF1			enriched		Gallienne AE	2011
	transcription profiling	8 SCD patients used to culture cells	KLF10			$2.54x10^{-6}$		Borg J	2012

Heamoglobin concentration	candidate gene	261	G6PD			0.008		Nouraie M	2010
Hemolytic anemia	GWAS	Discovery=1117 SCD patients	NPRL3	Replication cohorts=449 +296 SCD patients	NPRL3	6.04×10^{-7}		Milton JN et al.	2013
Infection	candidate gene	93 SCD patients, 21 HCV infection	HLA-G				0.41 (0.24-0.71)	Cordero EA	2009
	candidate gene	1473 SCD patients: 145 with bacteria infection, 1248 without	IGF1R, TGF-beta, BMP6, TGFBR3, BMPR1A, SMAD6, SMAD3			0.0031		Adewoye AH	2006
	candidate gene	43 SCD patients with infections, 37 SCD patients without	HLA			0.01		Tamouza R	2002
	candidate gene	115 SSCD patients with/without infections	RANTES			0.01		Dossou-Yovo OP	2009
	candidate gene	72 SCD patients	IFNgamma			0.014		Joannes MO	2010
	candidate gene	73 SCD with bacteria infection 71 SCD without	HLE			$p_{corrected}$ =0.003		Tamouza R	2007
Lipoprotein levels	candidate gene	35 SCD, 15 carriers, 15 SCD/bthal	APOE, factor V Leiden				4.07 (1.01-16.4)	Rahimi Z	2011
Nephropathy	candidate gene	521	MYH9 and APOL1			0.0001		Ashley-Koch	2011
Occlusive vascular complications (OVC)	candidate gene	34 SCD patients with OVC and 63 without	HPA-5			0.0002		Castro V	2004
Osteonecrosis	candidate gene	Not indicated	ANXA2			0.001		Pandey S	2012
	candidate gene	442 SCD with osteonecrosis and 455 controls	BMP-6, Annexin A2, klotho			0.01		Baldwin C	2005
Osteomyelitis	candidate gene	42 SCD with osteomyelitis 140 SCD without	HLA			0.003 (Bonferoni corrected)		Al-Ola K	2008
Pain treatment	candidate gene	72 SCD patients	CYP2D6			0.05		Brousseau	2007
Priaprism	candidate gene	148 SCD with priaprism and 529 controls without	KLOTHO				2.6 (1.4-5.5)	Nolan VG	2005
	candidate gene	199 SCD patients; 83 with priaprism	TGFBR3, AQP1, ITGAV, F13A1			Benjamini-Hochberg sig		Elliott L	2007
Pulmonary hypertension (pTHN)	candidate gene	111 SCD patients : 44 with pHTN, 67 without	TGFbeta, ACVRL1, BMPR2, BMP6			FDR =0.075-0.246		Ashley-Koch	2008
SCD severity	GWAS	Discovery=1265 SCD patients	KCNK6, TNKS	Replication=163 SCD patients	KCNK6, TNKS	$<10^{-6}$		Sebastiani P	2010
SCD vs controls	candidate gene	128 SCD and 542 controls	MBL			no association, 0.002		Dossou-Yovo OP	2007
	candidate gene	106 SCD 156 controls	MTHFR			0.03		Al-Absi IK	2006
	candidate gene	115 SCD and 43 controls	CYP2C19			sign		Babalola CP	2010
Splenic sequestration	candidate gene	210 cases 200 controls	TNF-alpha IL-8			0.001		Cajado C	2011
Stroke	candidate gene	21 SCD stroke patients, 42 SCD non stroke	ATG			0.05	4	Tang DC	2001
	candidate gene	51 cases 51 matched controls	VCAM1			0.04	0.35 (0.15-0.83)	Taylor JG	2002
	candidate gene	49 SCD with stroke 47 without	TNF			0.006	3.27 (1.6-6.9)	Hoppe C	2007

	candidate gene	48 SCD patients 42 controls	G6PD		0.002		Hellani A	2009
	candidate gene	130 cases 103 controls	ANXA2, TGFBR3, TEK, ADCY9		0.05		Flanagan	2011
	GWAS	677 SCD patients: 177 had a stroke	22 variants	GOLGB1/Y1212C, ENPP1/K173Q		0.44 (0.27-0.72)	Flanagen JM	2013
	GWAS	230 cases 400 controls	results not available		results not available		Adams GT	2003
	candidate gene	516	G6PD			2.78 (1.04-7.42)	Thangarajh	2012
	candidate gene	230 SCD patients	TNF and IL4R			5.5 (2.3-13.1)	Hoppe C	2004
	candidate gene	62 TCD cases and 312 stroke free SCD patients	G6PD			3.36 (1.10-10.33)	Bernaudin F	2008
Trachoma (blindness)	candidate gene	836	Hp haplotypes		0.0001	2.0 (1.17-3.44) boys 0.58 (0.32-1.04) girls	Savy M	2010
Vascular complications	candidate gene	53 SCD patients	MTHFR		significant		Moreira Neto F	2006
	candidate gene	177 SCD with complications+100 SCD without complications	MTHFRC667T		0.015		D Hatzhofer BL	2012
Vaso-occlusive crisis (VOC)	candidate gene	104 SCD with VOC, 63 SCD without VOC	HLA		0.05-0.005		Mahdi N	2008
	candidate gene	127 SCD patients with VOC, 130 SCD patients without VOC	HPA 1,2,3,4,5			3.16 (1.4-7.17)	Al-Subaie AM	2009
	candidate gene	39 severe and 48 mild VOC SCD patients	MBL2		0.0188	3.15 (1.19-8.5)	Mendonca TF	2010
	candidate gene	210 VOC patients 114 SCD controls	2VEGF haplotypes			0.68 and 1.89	Al-Habboubi HH	2012

A.4.1.vii Gene expression studies and SCD

Few studies have performed global gene expression profiling on SCD patients. Perhaps the most interesting was by Jison *et al.* [43] who reported that 112 genes are differentially expressed between 27 African-American patients with SCD in steady-state and 13 controls using data generated from peripheral blood mononuclear cells (PBMCs). These genes were involved in heme metabolism, cell-cycle regulation, antioxidant and stress responses, inflammation, and angiogenesis.

Less pertinent studies include one whose goal was to validate a globin reduction protocol and did not characterise differential gene expression as it pertained to SCD [44]. Another study, examined differential miRNA expression between healthy and SCD affected erythrocytes [45]. Recently, miRNAs isolated from platelets from SCD patients and controls were analysed and significant differences were identified in functionally active platelet miRNAs [46].

A.4.2 Environmental risk factors of SCD clinical variation

Observations of clinical variability between identical twins with SCD and also within the same individual affected with SCD at different periods in his/her lifetime, highlight the important contribution of environmental factors to the phenotypic variation in SCD [47]. Environmental factors, such as physical activity, diet, and toxins, can elicit changes in gene expression, altering the epigenome, without changing the DNA code [34]. Such epigenetic changes may be the reason for the discordance of clinical phenotypes in identical twins with SCD [34].

Other environmental factors that affect SCD phenotypes are infections: malarial infections, viral infections, HIV infections, and bacterial infections. The HbSS mutation is a risk factor for death from malaria [48], as well as a potent comorbid

factor for death from bacterial infections, particularly invasive pneumococcal disease [48].

Malnutrition and dehydration are also major determinants of sickle cell disease severity. Furthermore, seasonal affects influence SCD severity. Cultural background (some African tribes believe that SCD is a curse, ostracising the individual rather than treating him or her), and lack of medical education and resources contribute to disease severity and clinical variation in parts of Africa, such as in Benin [21].

A.5 Measuring SCD severity

In order to study factors associated with SCD severity, a method to measure disease severity is required. Many methods have been proposed to measure the severity of SCD clinical manifestations. Some methods use categorical classifications, others are quantitative; however, no method has been universally accepted. Below is a description of the most commonly used methods of classifying SCD patients according to category and severity.

A.5.1 Traditional case-control

Most studies recruit SCD patients in steady-state condition. In these studies, the patient is not experiencing a crisis event and has stable hematological values. Steady-state SCD patients are compared to controls and evaluated for differences in the exposure of interest. This method of classification does not measure intra-individual variation observed in SCD patients.

A.5.2 Clinical Categories

In parts of Africa, where only basic clinical lab tests are available, clinical categories are used as a proxy to severity indices. At the National Sickle Cell

Disease Center (NSCDC) in Cotonou, Benin, West Africa, patients are seen on a regular basis and four clinical categories are assigned to them based on their disease course [21]. Upon enrolment into the program and when in steady state, patients are labeled “entry” (E). After at least one year of follow-up, distinction is made between patients with obvious positive changes, including improvement of their general status, increased velocity of linear physical growth and marked reduction in the frequency and severity of SCD-related acute events (group 1) and those with no such improvements (group 2). Both of the groups include patients sampled while in steady state. They are labeled “steady state satisfactory” (SSS) and “steady state unsatisfactory” (SSU), respectively. Finally, all patients followed and who are experiencing a SCD-related crises event are labeled “acute” (A). These clinical categories serve as proxies to a severity gradient of the clinical phenotypes observed in SCD patients.

A.5.3 Severity index

One attempt to establish a severity index of SCD [49,50] was based on the frequency of vaso-occlusive crisis events. This method of ranking severity among individuals with SCD has been proven insufficient because the intra- and inter-patient variability of SCD cannot be measured using this parameter alone.

Miller and colleagues (2000) examined the records of nearly 400 children who were followed at comprehensive sickle cell centers in North America [51]. They performed a multivariate analysis of the clinical courses of these children between infancy and 10 years of age and determined that several factors, including an episode of dactylitis (inflammation of a digit, finger, or toe) prior to one year of age, low hemoglobin levels before 2 years of age, and persistent leucocytosis (elevated

white blood cell counts) in the absence of infection, were associated with severe complications, such as recurrent severe pain episodes, stroke and acute chest syndrome.

A smaller study [52] which tracked the course of adult and pediatric patients over a 7-year period confirmed the factors identified by Miller *et al*. [51] to be associated with SCD severity in adults. However, the smaller study also identified that adults with an elevated white blood cell count experienced more frequent hospital admissions for painful vaso-occlusive crises than did those with lower white blood cell counts [52]. Interestingly, none of the assessed variables were found to be correlated with severity of SCD in children. The smaller size of the study and the greater age range of the children evaluated however most likely account for the difference between the results in this study and the report by Miller and colleagues [51].

Other investigators, such as Hebbel [53] used a scoring system to assess disease severity. One such severity index was proposed by El Hazmi [54,55] which takes into account many clinical manifestations observed in SCD. This severity index allows classification of patients in mild as well as severe forms of sickle cell disease. In Table A.4 below, the parameters that comprise this index are detailed.

Table A.6. SCD Severity Index.

Parameter	Score
Age	
less than 20 years	2
20-40 years	1
more than 40 years	0
No. vaso-occlusive crisis (>48hrs)	
Absent	0
1-2 crisis	1
>2 crisis	2
No. diagnosed infectious episodes	
Absent	0
1-2 times	1
>2 times	2
Antecedents of transfusion	
Absent	0
1-2 times	1
>2 times	2
Chronic complications	
Present	2
Absent	0
Repercussion on activity	
Normal activity	0
Activity disturbed	2

A.5.4 Network model to predict risk of death

Using data from 3380 patients and accounting for all common genotypes of sickle cell disease, a Bayesian network model, which includes 25 SCD clinical events and laboratory test results, was used to predict the risk of death within 5 years for SCD patients [42]. The reliability of the model was supported by its use in the analysis of two independent patient groups. In one group, the severity score was related to disease severity based on the opinion of expert clinicians. In the other group, the severity score was related to the presence and severity of pulmonary hypertension and the risk of death. Along with previously known risk factors for mortality, such as renal insufficiency and leukocytosis, the network identified laboratory markers for severity of hemolytic anemia and its associated clinical events, as contributing risk factors. The authors report that this model can be used to compute a personalized disease severity score for therapeutic decisions to be made according to the prognosis. Also, the severity score could serve as an estimate of overall disease severity in genotype-phenotype association studies, and provides an additional method for studying the complex pathophysiology of sickle cell disease. An online severity calculator based on their model has been made publically available [56].

A.5.5 Fetal hemoglobin levels or F cell distribution as a marker of severity

HbF levels have been used in association studies to identify genetic factors influencing SCD severity. Higher HbF levels were associated with reduced rates of acute painful episodes, leg ulcers, and less frequent acute chest syndromes [35]. However, HbF levels had no clear association with other SCD clinical manifestations such as stroke and silent cerebral infarction, priapism, urine albumin excretion, or systemic blood pressure [35].

A.6 Biomarkers of SCD

The incidence of most clinical complications in SCD varies markedly both with time in the same individual and between different individuals. Meaningful biomarkers, indicators of a biological state that can objectively be measured and evaluated as either an indicator of a normal or a pathogenic biological process, could be useful in the management of SCD. Biomarkers can be used in early diagnosis of complications, detection of chronic organ damage, identification of individuals at risk of a severe clinical course, and monitoring response to treatment.

More than 100 different blood and urine protein biomarkers have been described in SCD [57]. Nearly all of these biomarkers are abnormal in the steady state of SCD, and become more abnormal during complications. Some biomarkers indicate damage to specific organs, whereas others indicate damage to more systemic processes. Unfortunately, however, none of these biomarkers provide specific prognostic or clinical information beyond that which can be provided by the simple measurement of heamoglobin concentration. To date, no prognostically validated biomarker has been identified to predict which SCD patients will develop severe outcomes.

Recently, blood transcriptional profiling has been successfully used to predict disease pathogenesis in tuberculosis, infections, and tumour progression [58,59,60]. Through the identification of aberrant gene expression, it is possible to identify individuals who are susceptible to disease and to predict their outcome. Biomarkers that identify SCD sub-phenotypes exist [57,61,62,63,64], but none have been validated in longitudinal studies. Furthermore, no transcriptional biomarkers of clinical progression have been identified for SCD patients.

A.7 Management of SCD

The vast majority of the therapies offered to patients with SCD are supportive and do not modify or change the underlying pathophysiology of the disease. These therapies include analgesics to relieve acute pain and curb or manage vaso-occlusive crisis events [65,66,67], as well as antipyretics to relieve fever and anti-inflammatories to reduce inflammation associated with the events. Blood transfusions are administered to compensate for RBC death and dehydration. Dehydration is also treated with I-V Saline solution (normal or 5% dextrose in saline). Chronic pain is managed with the use of oral morphine in addition to acetaminophen, NSAIDs (used for deep bone pain), and opiates. Antibiotics are administered when an infection is suspected.

SCD patients are also treated using standard procedures to treat chronic hemolytic anemia, pulmonary hypertension and various organ damage syndromes associated with the disease. Dialysis or kidney transplant for kidney disease, gallbladder removal in patients with gallstone disease, hip replacement for avascular necrosis, wound care for leg ulcers and surgery for patients who have eye problems are all examples of treatments required by some SCD patients.

Prevention of complications is also an important aspect of SCD treatment. To prevent stroke, regular blood transfusions followed by iron chelation for the treatment of hemochromatosis are performed [68]. Blood transfusions are also done in order to reduce HbS percentages to below 30%, thereby reducing the risk of polymerization and blood clot formation. Bypass surgery to restore adequate blood supply may also be required to prevent myocardial infarction.

Neonatal screening for SCD allows for prophylaxis immunization to pediatric patients who are susceptible to infectious diseases. Preventative treatments include

penicillin prophylaxis and immunization against streptococcus pneumonia. Parental teaching about SCD susceptibility to infections and the importance of preventative treatments and vaccinations can also help to increase adherence and thus alleviate the infection morbidity and mortality rate in SCD patients.

Non pharmacological approaches, such as support groups, physical therapy, acupuncture and acupressure can also have an impact in the treatment of SCD by improving quality of life for the SCD patient.

Bone marrow transplantation is a possible cure for SCD patients, however, this difficult and risky procedure is strongly dependent on the availability of a suitable donor. It is only used in cases of severe SCD children who have minimal organ damage to the disease. Bone marrow transplantation is still considered an experimental procedure [26].

A.7.1 Lack of SCD specific drugs

The only FDA approved disease-modifying therapy for SCD is hydroxyurea (HU), but it has variable outcomes and is potentially toxic [69,70].

HU was first approved for use in SCD in February 1998. It was approved for use in reducing the frequency of painful crises and the need for blood transfusions in adult patients with recurrent moderate-to-severe painful crises (generally at least three crises during the preceding 12 months).

HU was shown to promote the production of fetal hemoglobin [71,72] by stimulating development of erythroid cells [73,74], increasing RBC mean corpuscular volume, and reducing the number of dense cells and irreversibly sickled cells in the circulation [75]. HU inhibits ribonucleotide reductase, blocking DNA synthesis and cell division.

Unfortunately, less than half of SCD patients treated with HU benefit from its use [76]. Even when HU is administered, the exact dose of HU needed to prevent painful crises is unknown. Thus, patients receive the maximum tolerated dose, often resulting in the development of macrocytosis (a blood condition of insufficient and unusually large RBCs that leads to oxygen deficiency throughout the body and can result in organ damage). HU administration is not recommended to patients who are pregnant because hydroxyurea has been shown to be teratogenic in mice, although no human studies have been conducted to support this finding for obvious ethical reasons. Administration of HU in children is controversial. A number of concerns have been identified in the treatment of pediatric SCD patients with HU. Among these concerns are the possibility of impaired neurocognitive development and impaired bone maturation. As well, the carcinogenic potential with long-term use of HU is unknown and recent studies of 15 years of follow-up have identified the risk of developing myeloid leukemia. In cultured human cells, it was observed that following HU administration, regions of the cell's genome were duplicated and/or deleted.

The data on hydroxyurea applies only to patients who are homozygous HbSS living in developed counties. Patients with compound heterozygous conditions (e.g., HbSC disease) were excluded from the studies in order to eliminate possible response variability in the data. As a result, compound heterozygous HbSC patients are not eligible to receive HU for their treatment. Also, HU has not been tested on patients living in developing countries where comorbidities, including malaria and nutritional deficiencies, may affect the toxicity profile [7]. Finally, HU remains too expensive for SCD patients in resource poor areas, especially in Africa, where the largest burden of the disease is located.

A.8 Public Health Genomics and SCD

Public health genomics can offer an unbiased, global approach that integrates genome-based knowledge and information to improve the health of SCD populations. It is through recent technological advances that public health genomics has evolved [77] as an extension of genetic epidemiology, which focuses on the role of genetic factors and their interaction with environmental factors in the occurrence of disease in human populations [78]. Public health genomics applies systematic, evidence-based assessments of genomics applications in health practice and works to ensure the delivery of validated, useful genomic tools in medicine. Examples of emerging applications of human genome discoveries for clinical practice and disease prevention are given in Table A.5 below.

Table A.7 Application of Human Genome Discoveries in Public Health.

Type of Application	**Example of Proposed Application**
Therapeutic agents	Herceptin in treatment of breast cancer
Diagnostic tests	BRCA analysis in hereditary breast and ovarian cancer
Pharmacogenomic tests	Genetic testing for warfarin treatment
Prognostic tests	Tumor gene expression profiles in various cancers
Screening tests	Biomarkers for early detection of ovarian cancer
Risk assessment tests	Genome profiles in breast and prostate cancer

***Source adapted from Khoury *et al.* [77]**

A.8.1 Genome-wide association studies (GWAS)

An important development in genetic epidemiology has been the emergence of genome-wide association studies (GWAS). The integration of genome-wide association studies (GWAS) in epidemiology has lead to large collaborative case-control, cross-sectional, and cohort studies that have identified novel genetic variants associated with disease. This prominent genomic approach [79,80,81] uses dense genotyping chips to ascertain the genotype at hundreds of thousands of common single nucleotide polymorphisms (SNPs) in several thousand individuals with and without disease, and significant association between these two variables is assessed without *a priori* hypotheses.

Most GWAS-discovered variants are relatively “weak” risk factors (most with relative risk of 1.05 to 1.50 per allele). Nonetheless, a considerable number of associations have been reported for many disease phenotypes pinpointing to genes and pathways involved in the etiology of the diseases in question.

Although GWAS provide valuable clues to the pathogenesis of complex traits, such as the capacity to provide a relatively unbiased examination of the entire genome for common risk variants, there are certain aspects related to GWAS for which considerable challenges remain. For example, in identifying common risk alleles, GWA studies cannot clearly distinguish between the signal from true risk variants and the statistical noise from the vast numbers of markers that aren't associated with disease. To separate true signals from noise, researchers have to set a high threshold which a marker needs to exceed before it is accepted as a likely disease-causing candidate. This reduces the incidence of false positive results, but it also means that many true disease markers with small effects are lost in the background noise. Furthermore, most efforts at replication have concentrated on the

signals for which the statistical evidence is strongest; whereas, some susceptibility loci with modest effect sizes might also benefit from further exploration in the context of their biological plausibility. By increasing the numbers of samples in disease and control groups, researchers are trying to lower the statistical noise from non-associated markers so that disease genes with even small effects stand out above this noise. As the cost of genotyping has decreased steadily, this approach has become more and more feasible. However, the logistical challenge of collecting large numbers of carefully ascertained patients will always be a serious obstacle. This is because most published GWAS feature case-control designs and thus concerns inherent to this type of design, i.e. selection bias, misclassification bias, and population stratification, remain. Population stratification occurs when there is a systematic difference in allele frequencies in cases and controls due to different ancestries. Past GWAS have been dominated by subjects of Western European ancestry, and only now are we beginning to understand the genetic risk variants in non-European populations [82]. In African populations standard SNP-disease genome-wide mapping has been challenging [82] because of decreased linkage disequilibrium (LD) in these populations [83,84,85]. LD is the non-random association of alleles at two or more loci, usually on the same chromosome. LD is a phenomenon derived from linkage, which is the presence of two or more loci on a chromosome with limited recombination between them.

GWAS relie heavily on the "common disease, common variant" (CDCV) assumption, which states that the genetic risk for common disease is mostly attributable to a relatively small number of common genetic variants [86]. However, the vast majority of the reported associations account for a minor fraction of the disease variance. The missing heritability has been suggested to result from larger

numbers of variants of smaller effects; rarer variants (possibly with larger effects); structural variants poorly captured by existing arrays; and gene–gene and/or gene-environment interactions [87,88,89].

A.8.2 Genome-wide gene expression studies– transcriptomics

Transcriptomics, or genome-wide gene expression profiling, is another important development in public health genomics. Microarray-based transcription profiling is a genomic approach used to quantify gene expression variation for thousands of genes at once [90,91].

The expression of DNA follows the rules of a central dogma which occurs in two stages: (i) Transcription, during which DNA is transcribed into messenger RNA (mRNA) and (ii) Translation, during which mRNA serves as a template for protein synthesis. Regulation of gene expression is the first stage in a multi-step process toward the production of phenotypes and is arguably the most important component in the genetic basis of phenotypic variation [92,93]. Transcript abundance sums the effects of various sources of variation in gene expression including genetic variation, spontaneous inherited epigenetic marks, and environmental factors. These causes of variation in gene expression can be variable at the cell, tissue, organism, or population level [94] and act together at various magnitudes on multiple modulators that include promoters, activators, enhancers, repressors, trans effectors, chromatin, and environment- or genotype-dependent methylation states [93].

Microarrays are used to interrogate total RNA with the relative fluorescence intensity proportional to transcript abundance. This powerful technology shows high repeatability and is cost-effective. Recently, even deeper transcription profiling based

on NextGen sequencing of RNA has become available. However, this technology remains expensive for most population-based studies.

A large number of studies have shown strong correlations between exposures and transcriptomic signatures in blood. The seminal Emilsson *et al.* study [95] on a sample of 1,002 individuals from the Icelandic Family Blood cohort demonstrated the pertinence of using gene expression profiles to explain and predict obesity-related traits even in blood. For example, 2,172 (9.2%) gene expression traits in blood are correlated with Body Mass Index, 1,098 (4.6%) with Percentage Body Fat, and 711 (3.0%) with Waist-to-Hip ratio.

A.8.3 Data Integration in Genetics and Genomics: Functional genomics and systems biology approaches to study disease

As a consequence of the rapid progress and accumulation of new methods and discoveries in genomic technologies, other "omics" related research fields are becoming available, such as transcriptomics (described above), proteomics, and metabolomics. New methods [96] of integrating these "omic" scans has led to the fields of functional genomics and systems biology. Functional genomics approaches study the relationship between an organism's entire genome and its phenotype. Systems biology focuses on complex interactions between genetic and environmental factors within biological systems. Both of these approaches use a more holistic perspective rather than the traditional reductionist approach. As a consequence, integration and analysis of complex data sets from multiple experimental sources, including genomics, transciptomics, and proteomics, are used in order to obtain a global understanding of an organism's health. Using unbiased approaches, that take into account the entirety of the genome and its gene products,

allows for an understanding of the dynamic properties of an organism at the cellular level. This provides a more complete picture of how biological function arises from the information encoded in an organism's genome and in understanding how a particular mutation(s) leads to a given phenotype.

A.8.4 GWAS of gene expression (eQTL/eSNP analysis)

The integration of genome-wide gene expression with genotyping data is a functional genomics and systems biology approach to study disease that jointly explores the environmental and genetic influences on disease phenotypes. Through the joint analysis of gene expression and genotyping data, greater insight is gained than can be provided by either type of data alone [92,97,98,99,100]. This line of research was motivated by the basic idea that transcript abundance is a quantitative trait with a heritable component. Consequently, quantitative linkage mapping (QTL) methods were used to map sources of expression variation and were named expression QTLs (eQTLs). The term expression single nucleotide polymorphism (eSNP) denotes SNPs associated with variation in transcript abundance in a population [101].

The original studies of this nature employed cell lines and clearly established not only that the majority of transcripts are highly heritable, but also that they are often associated with regulatory polymorphisms that account for 15 to 70% of the variance in gene expression in a sampled population [102,103,104]. Sample sizes of less than a hundred individuals are sufficient to obtain genome-wide significance levels [105,106], in stark contrast with sample sizes required to reach this significance level using standard GWA study designs [107], and many of the

associations are consistent across three major human ethnic groups (Africans, Europeans, and Asians) [102,103].

Subsequent studies of peripheral blood isolates confirmed the genetic contribution in gene expression variation somewhat surprisingly since the environmental component of variance should be much higher in leukocyte mixtures isolated from individual people (rather than homogeneous cell lines grown in uniform culture conditions). Large samples of Icelanders [95] have further shown that associations involving so-called *cis*-acting expression SNPs, namely eSNPs located in the same gene as the transcript they regulate, sometimes highlight genes that have been associated with complex disease in other GWAS – even where that disease is manifest in a different tissue. Distal-regulatory eSNPs (where a polymorphism in one gene acts on an unlinked target gene) are less common, in part because the threshold of evidence for genome-wide significance is several orders of magnitude higher owing to the additional multiple comparisons, but also because it seems that cross-correlation is weaker between loci [107].

A.8.5 Transcriptional gene-environment interactions

The meaning of the term “interaction’ can be a cause of confusion. Often, distinction is made between statistical interactions, public health interactions, and biological or causal interactions [108]. Statistical gene-environment (GxE) interactions are described as the differential effect of a given genotype exposed to different environmental conditions [92]. These interactions can be qualitative, where the effect is present in one condition, or going in opposite directions in different strata. However, if the effects go in the same direction, but differ in magnitude, than they are “quantitative interactions” and are scale dependent (e.g. raw or log-transformed for continuous traits; additive or multiplicative for binary traits) [109].

Statistical interactions should be distinguished from public health interactions, which are based on the concept of synergy as a joint effect that is greater than the sum of the excess risks from each factor alone, or biological interaction, which occur when an effect of one factor at the cellular or molecular level is dependent on the presence or absence of the other.

Identifying gene-environment interaction effects has been challenging [92,110]. Using the joint analysis of gene expression and genotyping data, gene-environment interactions are more likely to be significant since they are less likely to suffer from the power and multiple testing issues that conventional GWA studies have had. This is because transcript abundance is a continuous trait, hence potentially more informative than binary outcome data, and thus provides more mapping power. Also transcription is closer to the genetic effect and/or the causal mechanisms of exposure, with genotypes having an effect on transcript abundance on average one order of magnitude stronger than on disease phenotypes, reflecting the tight link between genetic regulatory elements and gene expression traits [92].

Most transcriptional gene-environment interactions have been reported in model organisms [111,112,113,114], and in humans, most of the documented studies report genotype-treatment interactions that are performed *in vitro* [92]. Only a few genome-wide surveys of transcriptional gene-environment interaction *in vivo* have been conducted in humans [115].

In SCD, gene-environment (GxE) interactions may explain part of the phenotypic variation observed in patients. By identifying GxE interaction, an improved understanding of why certain patients do not ameliorate in clinical status even after following a rigorous clinical care program may be possible. Identifying GxE

interactions may also provide a basis for targeting interventions for individuals at high risk of developing a worse outcome.

A.8.6 Pharmacogenomics

Pharmacogenomics is the study of how the entire genetic makeup of an individual affects their response to drugs. A functional genomics or systems biology approach to pharmacogenomics integrates genetic variation on drug response in patients by correlating gene expression with a drug's efficacy or toxicity. There are multiple, highly effective drugs in widespread use whose primary mechanism of action is to affect gene transcription [116]. Understanding gene transcription variation in disease can help in drug discovery. Identifying genes whose expression variation is under genetic control could lead to candidate transcription-modulating drugs which should be investigated as candidates for potential novel treatments.

By applying a genomic approach to study genetic factors that influence SCD clinical variation, it is hoped that an improved understanding of the pathobiology will lead to better care and treatment for this disease.

A.8.7 Drug rescue and repurposing

Genomics discoveries about the molecular basis of disease provide opportunities to translate research into clinically useful products. However, the translation process is long, costly, and has a low success rate. More than 95% of drugs fail the required therapeutic development process [117]. The average time form target selection to approval is ~13 years, and the cost of bringing a new drug to market exceeds US$1 billion [117]. Thus, strategies to reduce the time frame, decrease costs, and improve success rates are urgently needed.

Drug rescue and repurposing offer the advantage of harnessing previous research and development efforts [117]. Drug repurposing takes approved drugs or compounds that have already been tested on humans, and that have detailed information on their pharmacology, formulation, dosing, and potential toxicity, and tests their application for use in alternative diseases. This can enable the rapid testing of new clinical hypotheses, leading to improved health outcomes. By applying drug repurposing to SCD, new therapies might be discovered by identifying regulatory genes involved in the disease that are also drug targets.

B. RATIONALE AND HYPOTHESIS

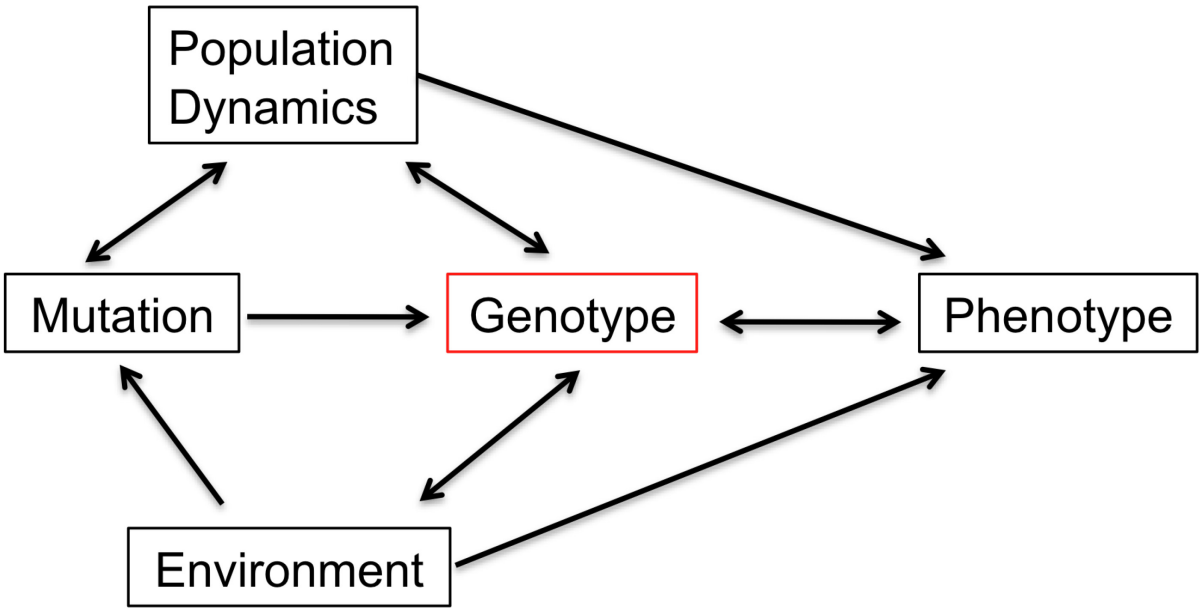
B.1 Rationale

The underlying causes of the clinical heterogeneity in SCD remain unknown. Thus far, almost all attempts at finding genetic factors associated with SCD severity have used classical candidate gene or genome-wide association approaches, but these efforts have had limited success. Important genetic advances in recent years have enabled the investigation of entire human genomes. For African populations, where standard SNP-disease genome-wide mapping has been challenging [82], functional genomics and systems biology approaches offer an attractive analytically-powerful and cost-effective alternative. Since disease in general involves differential expression, a systems genetics approach to map genetic variation associated with gene expression traits that are correlated with SCD clinical phenotypes is likely to reveal regulatory variation modulating SCD and the clinical heterogeneity.

B.2 Conceptual model

Khoury *et al.* [78] used a conceptual model to illustrate the scope of genetic epidemiology as the interface of genetic and environmental interactions in disease. In this model, mutations are the basis for genetic variation in populations. The frequency of specific genotypes, the survival of mutations in subsequent generations, and the frequency of a new mutation in a particular generation are determined by the balance between the occurrence and recurrence of such mutations (possibly influenced by environmental factors) and other dynamic population processes, such as selection, chance fluctuations, and mating patterns. See Figure B.1.

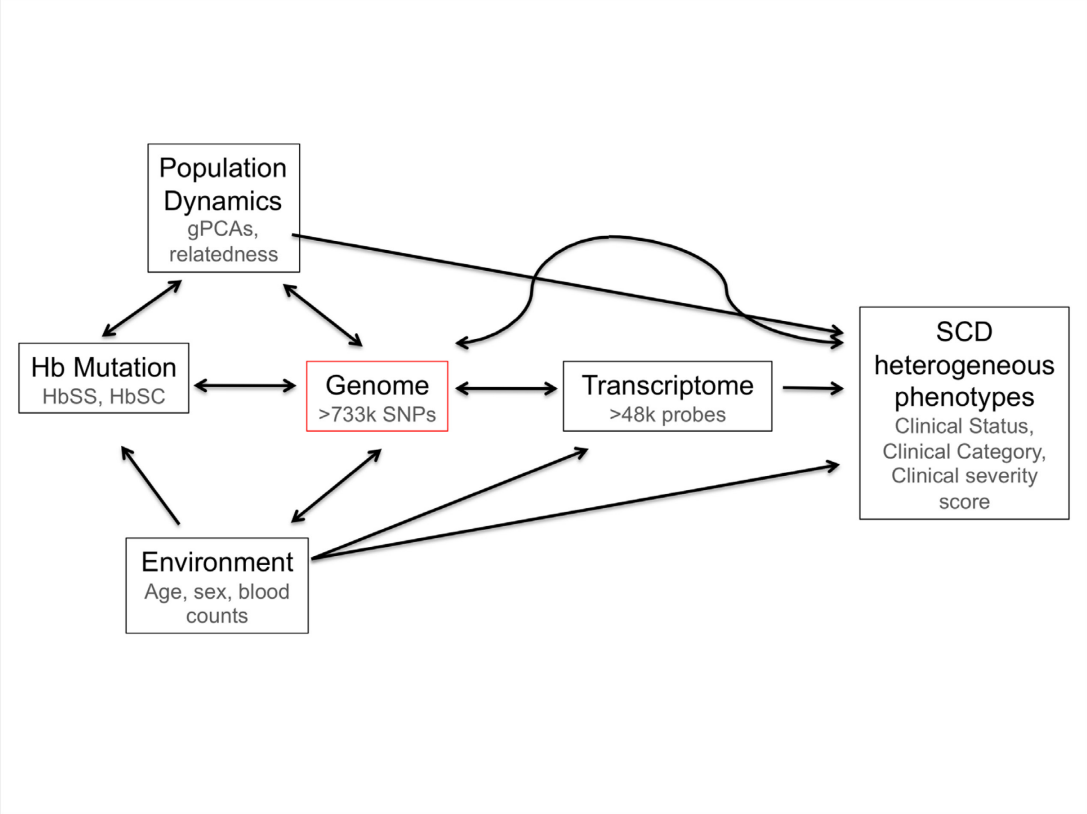
Figure B.1 Scope of genetic epidemiology.



Here, we apply this conceptual model to the case of SCD, a monogenic disease with phenotypic variation that is believed to be caused by the interaction between genetic and environmental factors through gene expression variation.

The genomics revolution has added powerful new potentialities and renewed impetus for understanding how biological and environmental processes act together in health and illness. This revolution has driven a paradigm shift from reductionist approaches that focus on single genetic elements in isolation to systems and functional approaches that focus on the interconnectedness of networks of elements acting as a whole. Using the genetic epidemiology conceptual model proposed by Khoury *et al*. [78], we propose that a SCD patient's clinical outcome is a complex result of their genome, gene products (transcriptome), and environment, all interacting together to modify the phenotypic expression of the Hb mutation. See Figure B.2 below.

Figure B.2 SCD Conceptual Model.



In order to evaluate the genetic determinants that modify SCD clinical outcome, we recruited patients with the two most common SCD mutations, HbS and HbC, as well as unaffected (HbA) participants used as controls. In order to identify genome-wide genetic factors that influence SCD, we genotyped over 733, 000 single nucleotide polymorphisms (SNPs) and measured expression levels of over 48,000 gene-specific probes genome-wide. We had access to environmental factors, including sex, age, and hematological blood counts (which are influenced by infection, diet, treatment, ect.). The two main SCD phenotypes that we used were

discrete categorical variables: clinical status and clinical categories. These variables are used at the National SCD Center in Benin as a proxy to severity.

B.3 Hypothesis

We hypothesize that both genetic and environmental factors contribute to gene expression variation, which in turn accounts for the phenotypic variation observed in SCD patients.

We hypothesize that gene expression differences will be enriched in biological pathways that impact the disease course to account for different clinical severities.

We also hypothesize that transcriptional biomarkers can be identified for the purpose of classifying patients in groups with different SCD severities.

Furthermore, we hypothesize that the genetic factors controlling gene expression can be detected and that interactions between genotype and clinical severity may further explain part of the inter-individual variation of this disease.

Finally, we hypothesize that differentially expressed genes that are under genetic control can be used to identify potential drug targets.

By identifying the underlying genomic architecture of the clinical heterogeneity in SCD, gene expression profiles and the genetic control of the variation associated with SCD severity, we can provide important biologic knowledge which will guide the future development of novel therapeutics and targeted treatments, as well as improved follow-up programs for SCD patients.

B.4 Objectives

The present work proposes the following objectives:

1. Identify gene expression variation associated with SCD clinical severity
    a. Identify biological pathways that impact the course of the disease;
2. Identify transcriptional biomarkers that can classify SCD patients by clinical severity;
3. Identify genetic regulators of gene expression variation in SCD
    a. Test for interaction between genotype and disease severity that may further explain the variation in SCD;
4. Identify known drug targets, using the results of our genomic eSNP analyses, which could be candidates to test in future studies as novel therapies for SCD

C. METHODS

In order to address the main objectives of this study, we implemented a 2-phase approach. In the first phase (Discovery phase, n=157), we explored clinical factors that were associated with gene expression variation and quantified their effects. In the second phase (Replication phase, n=154), we attempted to replicate these findings. Utilising this 2-phase approach allowed us to identify and confirm transcriptional biomarkers of SCD severity. In order to have sufficient power to identify genetic regulators of gene expression variation, interactions effects, and drug targets, we combined the data sets (combined data set I = 263, combined data set II=173).

C.1 Ethics approval

Ethics approval for the study was granted by Sainte-Justine Research Center Ethics Committee and by the Faculté des Sciences de la Santé of the University of Abomey-Calavi in Benin, West Africa. Written informed consent (Appendix II) was obtained by a parent or guardian on behalf of all participants in the study.

C.2 Study design

Case-control-studies, where sampling is conditional to the presence or absence of disease, are widely used in epidemiology for studying associations between disease and potential risk factors. In the current context, the risk factors are represented by genetic data: alleles/genotypes and/or expression profiles. Some of the advantages in using this type of design include that it is relatively quick and inexpensive and that the assessment of the "exposure" variable (i.e. genotypes/expression) is quite straightforward. Yet, in the context of genetic studies,

using a case-control approach raises a few concerns that need to be addressed in order to guarantee the validity of the study. More specifically, study participants, who typically provide data on exposures and other covariates, must also provide biological material – usually blood, for genotyping or gene expression purposes. Such samples are difficult to obtain, inventory and process. In addition, concerns about potential abuses of genetic data as well as the procedure itself (i.e. venopuncture to collect blood) makes using healthy subjects, especially children used as controls, hard to justify and to recruit. The resulting low level of control subject participation can invalidate a study. Furthermore, self-selected controls may not accurately represent the base-population studied.

An important concern regarding the use of unrelated cases and controls in association studies is that it is difficult to distinguish valid association due to linkage from spurious association due to confounding effects. One of the major confounders of importance in genetic association studies is population stratification which occurs if the population from which the cases and controls were sampled consists of latent subpopulations, each with different variant allele frequencies and risks of disease. A spurious association due to this confounding effect will occur for any variant allele that is at an elevated frequency in the subpopulation with the greatest disease prevalence. Several approaches exist to account for population stratification in genomic studies, including methods based on Genomic Control [118] and Principal Component Analysis (PCA) [119]. Genomic Control aims at correcting the Trend test statistic inflated null distribution by estimating an inflation factor, usually called λ, using many markers. The main assumption of this method is that the inflation factor is the same for all markers. PCA-based methods use markers to define continuous axes of variation, called principal components, which reduce the data to few variables

containing most of the information about the genetic variability. These axes often relate the spatial distribution of the ancestries of the samples. Using PCA methods, Price *et al.* propose an association test to account for stratification. It is implemented in the software Eigenstrat [119,120]. In practice, it is common to use principal components to adjust the results of the classical association test and correct for stratification.

C.2.1 Case-control study design

For the present study, we used the case-control study design and recruited SCD patients from the pre-existing National SCD cohort in Benin, West Africa at the Centre de Prise en charge Médicale Intégrée du Nourrisson et de la Femme Enceinte atteints de Drépanocytose (CPMI-NFED). Siblings, unaffected by SCD, of patients seen at the Center were invited to take part as controls. Using this study design both case-control and case-only analyses were performed. A two-phase sampling design was implemented to replicate our gene expression findings. An initial recruitment of patients was performed in the discovery phase (n=126 SCD), followed by recruitment in a replication phase (n=124 SCD). The distribution of SCD clinical severities, Hb genotypes, and sexes were equally proportionate in both phases. The comparison of interest was between SCD patients with different disease severities classified as a quantitative severity index or as discrete categorical classes. Of equal interest was a comparison of SCD patients and controls. Unaffected siblings of SCD were recruited at the CPMI-NFED in 2010 by approaching families who were already at the clinic for a scheduled visit for their affected SCD child. In the discovery phase, 31 controls were recruited. In the replication phase, 30 controls were recruited. The control sample had similar

distributions for age and sex to the SCD patient sample. Genetic ethnicity and relatedness were assessed for the cases and controls.

C.3 Setting

C.3.1 Location

The study was performed at the CPMI-NFED located in Cotonou, the largest city (population of 678 874 inhabitants based on the 2013 census [121]) and the financial capital of the Republic of Benin. Benin (population of 9 983 884 inhabitants based on the 2013 census [121]) is a country in Sub-Saharan West Africa, with one of the highest death rates worldwide for children under the age of five. The CPMI-NFED is the only established SCD Center in the country and follows the majority of the SCD population of Benin. Established SCD centers are rare in West Africa and the CPMI-NFED is a model for SCD clinical care and management for other African countries. See Figure C.1 below for a map showing the location of the recruitment site in Cotonou relative to West Africa.

Figure C.1 Location of recruitment site.

C.3.2 Recruitment dates

Recruitment was performed from February 2010 until April 2010 in the discovery phase, and from April 2010 until December 2010 in the replication phase. The main difference between the two cohorts is time of sampling. An important feature of gene expression variation is that it might vary with time and between seasons. Therefore, a discovery/replication design based on time of sampling when gene expression variation is surveyed seems appropriate to capture the environmental/seasonal component of variation.

C.4 Participants: method of selection and eligibility criteria

C.4.1 SCD patients

All (100%) of the SCD patients recruited for this study were part of a large cohort of SCD children longitudinally followed at the CPMI-NFED. This cohort was initiated as a SCD screening program in 1993 [21]. In 2003, the program was extended to evaluate clinical improvement and to study the disease course of SCD children who were enrolled in this comprehensive follow-up program [21]. Affected infants were and are evaluated at monthly intervals for the first 12 months of life and thereafter every 3 months. At the center, basic management of the disease and its symptoms consists in anti-pneumococcal and anti-malarial prophylactic medication, supplementation with folic and ascorbic acids, and specific vaccinations (Hepatitis B and Hemophilus influenxae B vaccines, and anti-pneumococcal and anti-Salmonella typhi vaccines) in addition to the six vaccines for children recommended by the World Health Organization. During scheduled medical visits, parents of these children are educated about SCD and the importance of nutrition, emphasising the importance of keeping clinic appointments and scheduling regular follow-up visits. In 2003, it was

demonstrated [21] that the children followed by this program showed a marked reduction in frequency and severity of SCD-related acute events, with improvement in general status and physical growth after follow-up. Today, over 2000 infants are enrolled at the CPMI-NFED.

Recruitment of SCD cases was performed by approaching patients who arrived at the Center for their scheduled visit. Recruitment of SCD cases who were referred to the CPMI-NFED as newly diagnosed, first time patients was also performed by approaching patients who arrived at the Center for their scheduled visit.

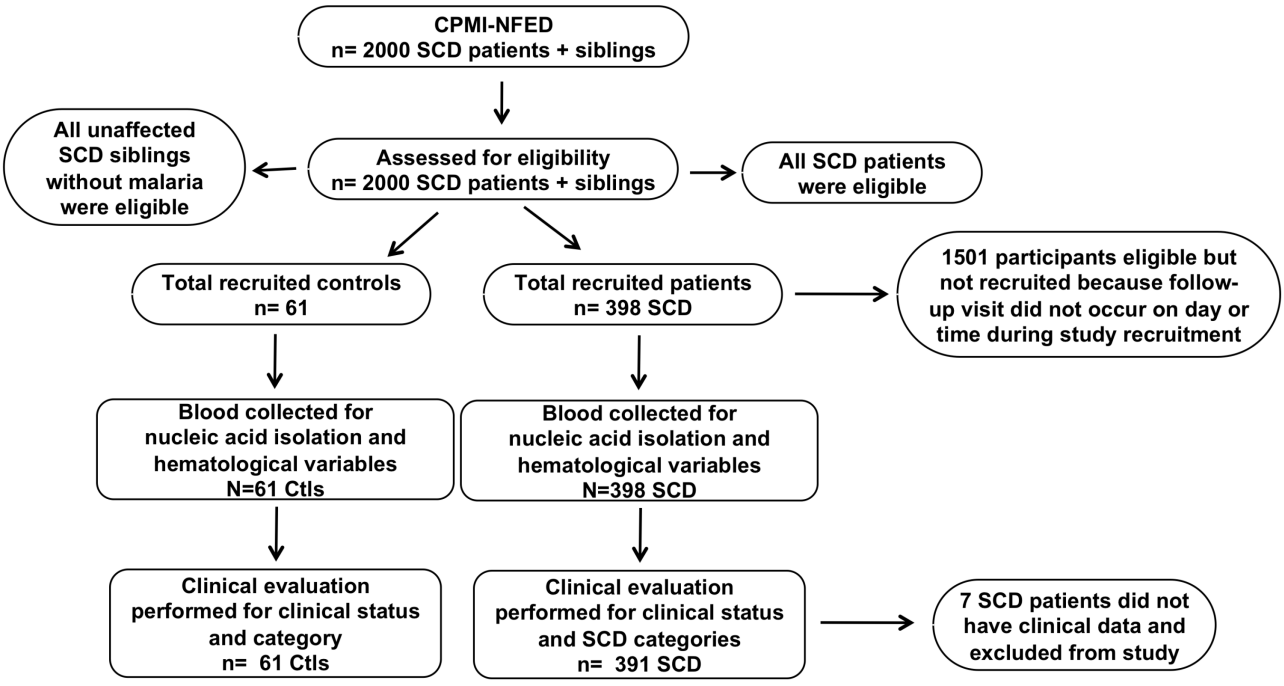
For non-genetic studies, selecting incident cases is ideal in order to avoid biases inherent when using prevalent cases (such as reverse causality, which is a problem when the cause and effect are sampled at the same time and reversed; survival bias, where survivors of a lethal disease are more likely to enter a study than other cases; migration bias, which can occur when the disease risk factors are related to migration from the area, ect.). For genetic studies, however, since the exposure of interest is fixed (genotype) and established prior to the outcome, including both incident and prevalent cases is not a concern. SCD patients who were admitted to the center or already enrolled in the CPMI-NFED programme from February 2010 to December 2010 were invited to participate in the study.

Inclusion criteria included being a patient at the CPMI-NFED and having one of the major forms of SCD (HbSS or HbSC) and providing signed informed consent. All patients seen at the Center were eligible to be included in the study since all had HbSS or HbSC and since sampling the entire spectrum of clinical severity observed in the SCD population was desired. Approximately equal numbers of boys and girls were sampled. No other exclusion or inclusion criteria were imposed.

C.4.2 Controls

Controls (n=61) were also sampled in 2010 from the city of Cotonou at the CPMI-NFED and were children unaffected by SCD and siblings of patients followed at the Center. Only children without clinical signs or symptoms of malaria, who tested negative for the commercially available rapid malaria detection test and the thick smear analysis for parasetemia quantification, and who were confirmed not to have SCD (not HbSS or HbSC) with at least one normal hemoglobin allele (HbA), were eligible to be included as a control. No other exclusion or inclusion criteria were imposed.

Figure C.2 Flow diagram for participant selection.



C.4.3 Similarities in cases and controls

In case-control studies, potential biases can occur if differences exist between cases and controls in the manipulation, timing or processing of samples during the collection procedure. With this in mind, we attempted to minimize the differences by following similar, pre-established protocols for cases and controls. All cases and controls were invited to participate in the study by a trained recruitment officer who approached potential participants at the same site (CPMI-NFED) in the morning (between 9am and noon). No difference in timing of recruitment existed based on case-control status of the participant. After consent, cases and controls had blood drawn in the same manner for all analyses that followed. A trained phlebotomist collected the required samples for cases and controls in a similar manner and stored the samples until the required analysis was performed (complete blood counts and HPLC). Shipment to Montreal was done in a similar manner and not based on case control status. Nucleic acids isolation and genomic experiments were performed in Montreal using identical protocols for cases and controls. All experiments were performed in a similar manner for cases and controls.

C.5 Data sources and measurements

C.5.1 Protection of privacy

All study material (i.e. biological samples and SCD clinical data, ect.) was coded to maintain the anonymity of the participants.

C.5.2 Nucleic acid extractions from whole blood

The same collection procedure was followed for all samples in order to reduce technical heterogeneity (Appendix III). A total of 10 ml of peripheral whole blood was collected from each patient between 9:00 am and 12:00 pm during their visit to the SCD Center and stored at -30^{o}C. Approximately 3 ml of this blood was collected for RNA work in TEMPUS blood RNA Tubes (Life Technologies). TEMPUS tubes contain a stabilizing reagent that immediately lyses whole blood cells and stabilizes RNA by inactivating cellular RNases and selectively precipitates RNA; genomic DNA and proteins remain in solution. Blood drawn in TEMPUS blood RNA Tubes are stable for up to 5 days at room temperature and for several months at -20^{o}C. Approximately 5ml of this blood was collected in EDTA tubes for DNA work. Shipment of the samples to Montreal was done at -20^{o}C. Total RNA was isolated using the TEMPUS RNA extraction kit (Life Technologies) following the manufacturer's recommendations. Since globin mRNA influences the signal quality in gene expression analyses, and since it is heavily expressed in total mRNA isolated from blood, a globin mRNA reduction step was performed using GLOBINclear-Human kit (Life Technologies). Total RNA extractions were quantified and quality was checked using the RNA 6000 Nano LabChip kit and 2100 Bioanalyzer (Agilent Technologies). Only samples of high RNA quality (Agilent's RNA Integrity Number > 7.5) were retained for expression profiling. Whole blood was collected in EDTA tubes

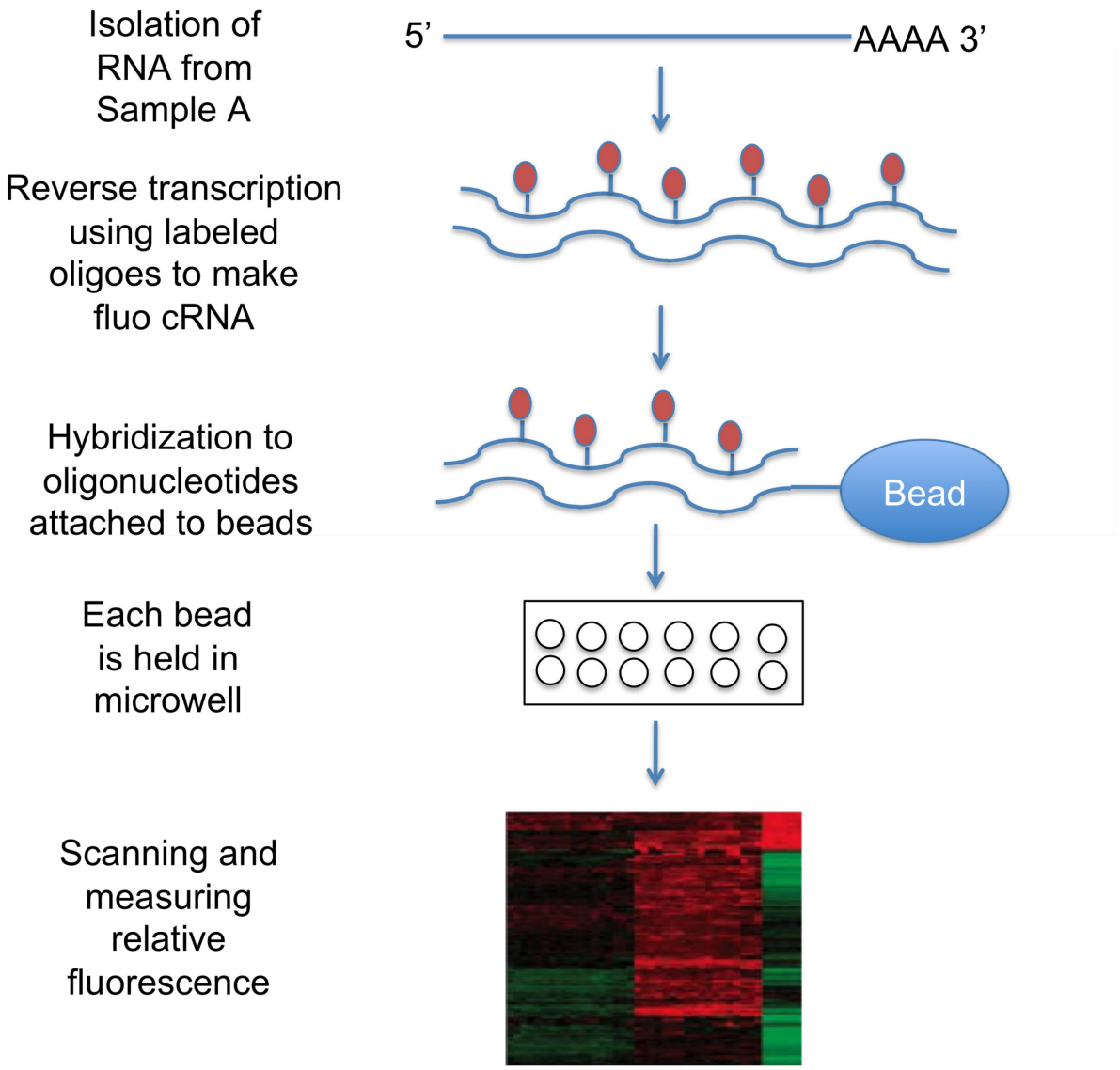
for DNA isolation. DNA samples were extracted using QIAamp DNA Kit (Qiagen). Quantity and quality was checked using Agilent's DNA 6000 Nano LabChip kit and the 2100 Bioanalyzer (Agilent Technologies).

C.5.3 Gene expression profiling

For each participant, total RNA was reverse transcribed, and labeled by incorporating fluorescent oligonucleotides to obtain cRNA. Using 500ng of this labeled cRNA, expression profiles of more than 48,000 probes were generated with Illumina's HumanHT-12 v4 BeadArrays. The manufacturer's recommended protocols were followed. To summarise, for each sample, the labeled cRNAs were hybridised to oligonucleotides which were immobilized to beads held in microwells on the surface of an array. Twelve samples were hybridized to each array. After performing washing and staining steps, the array was scanned and measured for relative fluorescence. The level of fluorescence that is measured for each probe represents the relative expression level for that individual's gene-specific probe. See Figure C.3 below. The raw intensities were extracted using the Gene Expression Module in Illumina's BeadStudio software. Expression intensities were log2 transformed and quantile normalized (QNM) using JMP Genomics v5.0 (SAS) after an outlier filtering procedure was applied. Levene's test of normality of log2 expression data for each probe was performed and probes departing from normality ($p < 0.001$) were removed.

QNM is the most aggressive method of normalization [122,123,124], and remove's technical variability by assigning each measure the mean value across samples for each rank, which creates an identical distribution for all samples. QNM has become the standard method in most gene expression studies [123], and is

Figure C.3 Experimental procedure used for cDNA microarray chips.



appropriate under the assumption that only a small number of measures differ among samples and that variation in the distributions is mostly technical noise that should be removed.

After applying QNM, 28,595 probes with expression at or above background levels averaged across all the arrays were retained for further analysis. These represent probes remaining after removal of 18,404 probe measurements that were considered to lie below background detection levels indicated by the inflection point in a plot of rank-ordered normalized intensities (see Figure C.4 below). Also, 427 probes overlaying known SNPs included in the Illumina's OmniExpress BeadChip were removed from the analysis.

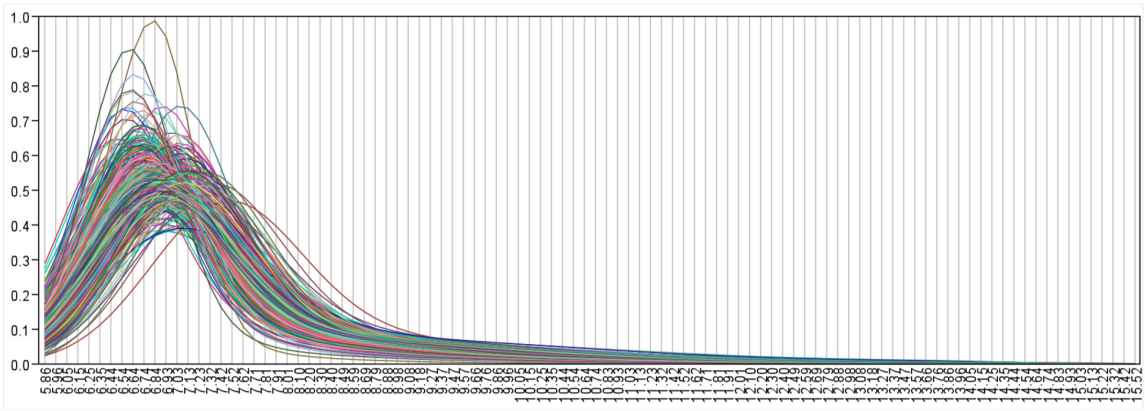
Illumina's HumanHT-12 BeadArrays were the arrays of choice because of their genome-wide coverage that targets more than 25,000 annotated genes using more than 48,000 probes that have a high reproducibility (concordance between replicates shown to be r^2=0.996 [125]).

Gene expression profiling was performed using the high-throughput facilities at the McGill University Genome Quebec Innovation Center in Montreal. (http://www.genomequebecplatforms.com/mcgill/home/index.aspx).

Figure C.4 Transformation and quality control of gene expression data.

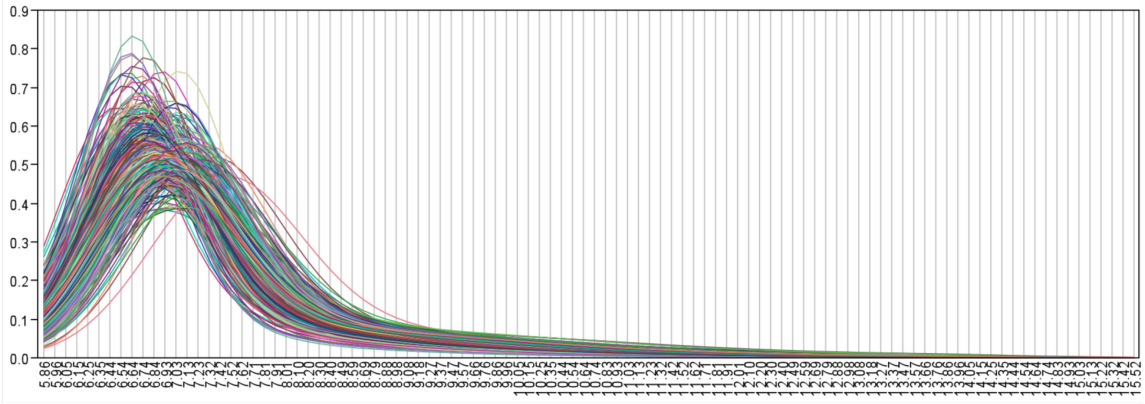
A) Distribution of the log2 transformed data for the 48,000 probes before quantile normalisation. The graph shows overlayed kernel density measures for the 48,000 probes for 324 samples. Two outliers shown in this plot were removed. B) The distribution is shown after removal of the outliers. C) The average expression level of each normalized probe for all samples was calculated. Using the mean values, the 48,000 probes were ordered from least expressed to most expressed. Using the inflection point of the plot of rank-ordered log2 transformed, quantile normalized probe intensities, 28,595 expressed probes were chosen since they were above background detection levels and were retained for further analysis. Below is the plot of the average values of all 48,000 probes, with the 28,595 expressed probes highlighted (dark line).

A)



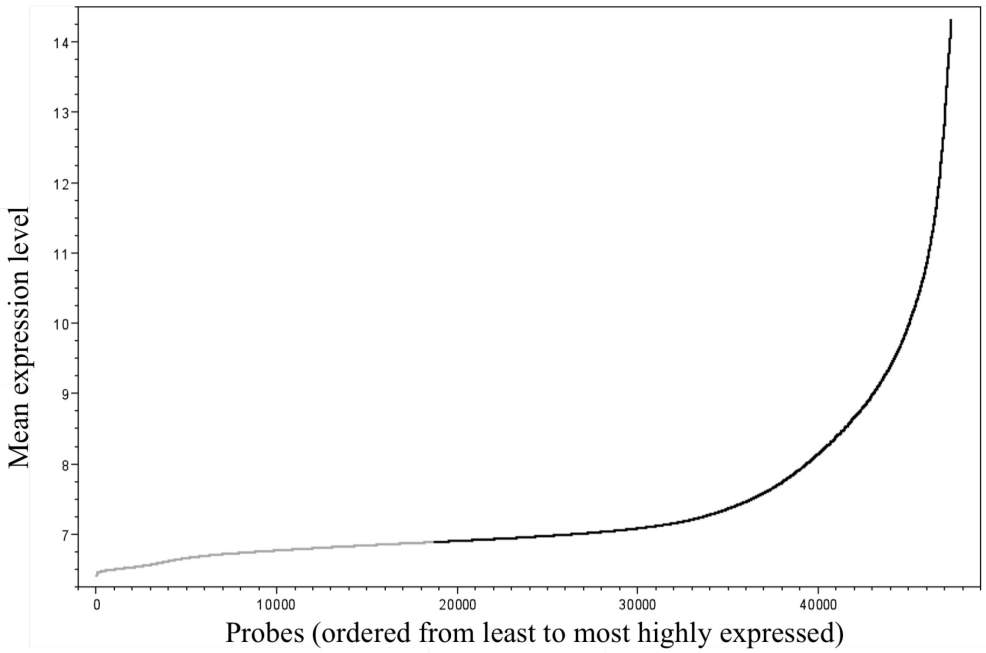
Log2 transformed gene expression data for the 48,000 probes

B)



Log2 transformed gene expression data for the 48,000 probes

C)



C.5.4 Quality control of gene expression data

To minimize chip and batch effects, a randomized design was used during hybridization. Hybridization was performed on two different dates. In order to test for a potential batch effect, four samples from the first hybridization batch were re-hybridized with the second batch as a means of quality control. These four technical replicates clustered adjacent to one another in hierarchical analysis, indicating a negligible batch effect on the data. This was confirmed by testing for batch effect in the probe-by-probe analysis of variance.

C.5.5 Genome-wide genotyping

Genome-wide genotyping data was generated for each participant for over 733,200 single nucleotide polymorphisms (SNPs) using Illumina's HumanOmni Express BeadChip arrays following the manufacturer's protocols. In simplified terms, detection of thousands of polymorphism genotypes for each individual is performed through labeling the isolated DNA with allele-specific oligonucleotides, hybridization of the labeled DNA to an array that contains immobilized nucleic acid sequences of target, washing, and detection using fluorescence. Interpretation of the signal was performed by extracting the fluorescent signal of each allele at each target using the Genotyping Module in Illumina's BeadStudio software.

Illumina's HumanOmni Express BeadChip arrays offer multiple advantages over other genotyping methods including high sample-throughput, comprehensive genomic content with a mean SNP spacing of 4.0 kb, optimized tag SNPs from all three HapMap phases that selects the greatest amount of common SNP variation possible, and high reproducibility (>99.9%). Of particular interest is that the SNP variation in African samples was evaluated by including the Yoruban sample, an

African ethnic group from Nigeria. Although the percent of the genetic variation that is captured in the Yorubans by using the HumanOmni Express is relatively low (0.4), this was the best available technology at the time that this study was performed.

Genotyping was performed using the high-throughput facilities at the McGill University Genome Quebec Innovation Center in Montreal. (http://www.genomequebecplatforms.com/mcgill/home/index.aspx).

C.5.6 Quality control of genotyping and evaluation of genotyping errors

The genotyping process involved a stringent protocol for maintaining the quality of the results acquired. Evaluation of genotyping errors was carried by calculating marker properties for each data set using PLINK [126]. Only SNPs with minor allelic frequency >5%, a call rate >99% and an exact Hardy-Weinberg (HWE) *P* value >0.001 were included. This resulted in sets of SNPs for each data set that were retained for further analyses (see Section C.6.1).

C.5.7 Hematological variables

The remainder of the 10 ml blood sample of whole blood that was collected from each participant (approximately 3 ml) was analysed in Cotonou at the CPMI-NFED and used for complete blood counts using an automated KX-21 blood analyser (Sysmex Corporation, Japan), for identification of the hemoglobin phenotype by high-performance liquid chromatography (HPLC) and Capillary Electrophoresis [127], and for quantitative falciparum malaria parasites determination (thick smear blood analysis). The complete blood count (CBC) hematological variables analysed were red blood cell counts (RBC cells/pL), white blood cell counts (WBC cells/pL), and parasetemia counts. The automatic blood analyser is an accurate and reliable

method of obtaining CBC. HPLC and Capillary Electrophoresis is a standard diagnostic method for SCD and has a high accuracy and reliability [127]. Quantitative falciparum malaria parasites determination was performed by trained laboratory technicians.

C.5.8 Diagnosis of SCD patients

Patients were diagnosed with SCD at the SCD Center in Cotonou after the HbSS or SC proteins were detected by High Performance Liquid Chromatography (HPLC) and Capillary Electrophoresis [127]. This method of SCD diagnosis is a standard, reliable procedure with high accuracy.

C.5.9 Sequenom genotyping of SNPs in β-globin region on chromosome 11

Identification of the rs334 genotype (the SNP that causes the HbS mutation) and characterization of haplotype structure in the Hb locus was performed using Sequenom MassARRAY technology. This method uses a primer extension assay to perform multiplexed genotyping of single nucleotide polymorphisms (SNPs) present in genomic DNA amplified by a multiplex PCR. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry accurately measures the mass of short oligonucleotide primers extended by a single dideoxynucleotide. The multiplexed genotyping assays rely on the natural molecular weight differences of DNA bases. By careful analysis of the genotyping primers, mass spectra of genotyping products can be generated with no ambiguity in allele assignment [128].

Twenty-one SNPs spanning a 61,345 bp region on chromosome 11 were genotyped. Figure C.5 details the chromosomal location of the 21 genotyped SNPs, including those that were used to construct the SCD haplotypes based on SNPs located at the classical RFLP sites used for SCD haplotyping [129]. See Table C.1.

Figure C.5 Chromosomal location of SNPs in globin region.

Twenty-one SNPs spanning a 61,345 bp region on chromosome 11 that includes the β-globin gene. The SNPs located at the four RFLP sites (Xmn1, HindIII, Hinc2, and HbS) were used to construct the SCD haplotypes.

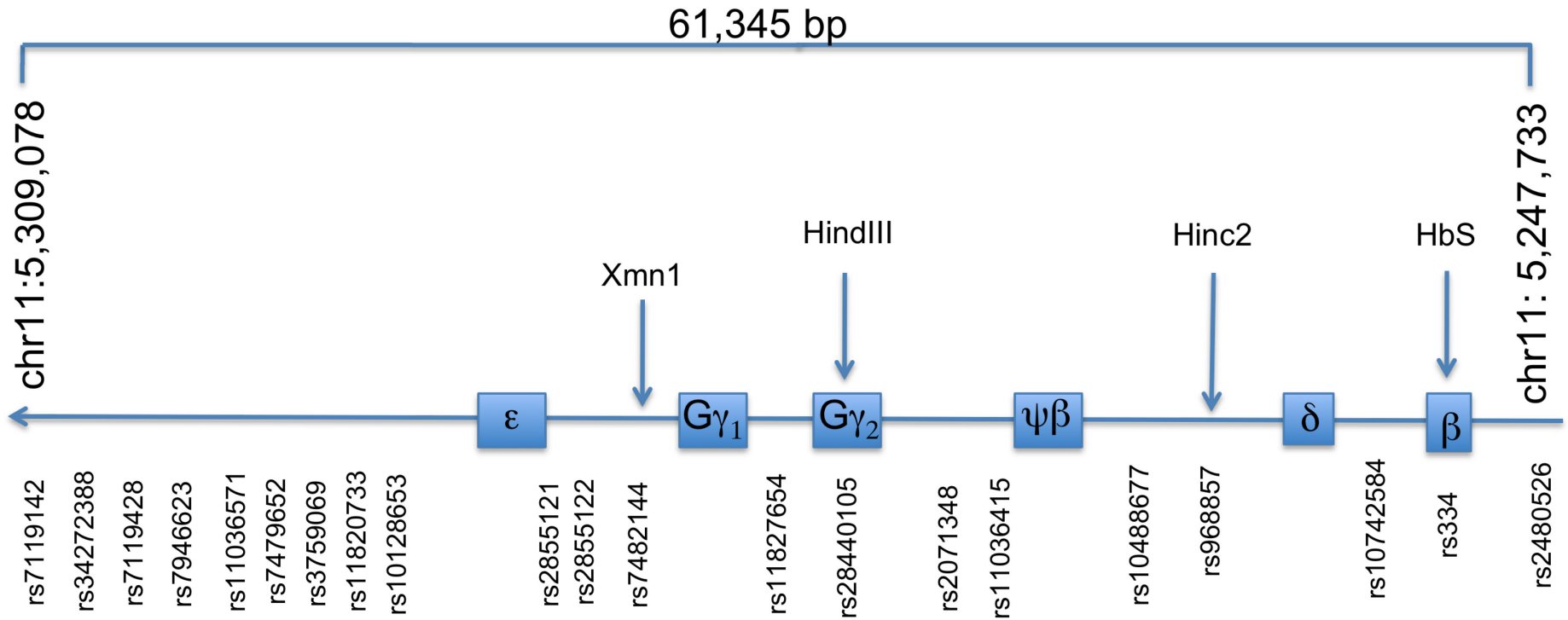


Table C.1 Classic sickle beta-globin haplotypes.

Construction of the classic sickle beta-globin haplotypes was performed based on genotypes of 3 SNPs that mark RFLP sites [129].

	HindIII rs28440105(A/C)	HincII rs968857(A/G)	HbS mutation rs334 (A/T)
Haplotype	A cuts	A cuts	T mutation
Benin	-/-	+/+	+/+
Car	-/-	-/-	+/+
Cameroon	+/+	+/+	+/+
Senegal	-/-	+/+	+/+
Arab	-/-	+/+	+/+

Assays were designed using SpectroDESIGNER software. The Sequenom Genotyping System offers flexible and efficient assay design (i.e. 96% success rate), improved call rates (i.e. 85%) and accuracy (i.e. error rate is less than 0.5%). Six hundred (600) ng of genomic DNA for each participant was used and the manufacturer's recommended protocols were followed at the high-throughput facilities at the McGill University Genome Quebec Innovation Center in Montreal (http://www.genomequebecplatforms.com/mcgill/home/index.aspx). SNPS with greater than 20% missing data were excluded and individuals with less than 75% call rate were excluded.

C.5.10 Measures of SCD clinical severity

In order to test for association between gene expression and clinical severity in SCD, all analyses were performed using two main categorical variables:

SCD clinical status: examines the effect of being followed at the comprehensive clinical care program (2 groups + controls).

SCD clinical categories: captures the clinical heterogeneity in SCD patients and is used as a proxy to SCD clinical severity (4 categories + controls).

C.5.10.1 SCD Clinical status

In order to evaluate the effect of the CPMI-NFED's comprehensive clinical care program, each patient was assigned a discrete categorical variable for clinical status: patients sampled at enrolment into the program and in steady-state were labeled as entry (E), and patients already being followed at the SCD Center were labeled as follow-up (FU). At the Center, most patients that are followed obtain a steady-state condition with general clinical improvement, which involves increased

velocity of linear physical growth and marked reduction in the frequency and severity of SCD-related acute events. However, some followed patients experience no such improvement and remain in an unsatisfactory state. Thus, both environmental and genetic factors influence a patients SCD clinical status.

C.5.10.2 SCD Clinical categories

In order to evaluate the clinical heterogeneity observed in SCD patients, we also assigned each patient a discrete categorical variable for Clinical Category based on their observed evolution in clinical condition that was evaluated at the SCD center over a period of 12-24 months. Patients were labeled as: Steady-State Satisfactory (SSS) if they had obvious positive changes including improvement of their general status, increased velocity of linear physical growth and marked reduction in the frequency and severity of SCD-related crises events. If no such improvement was observed during the followed period in the program, SCD patients were labelled as Steady-State Unsatisfactory (SSU). If patients were experiencing a crisis event related to SCD, they were labeled Acute (A). Finally, all newly enrolled patients into the program and who were in steady state were labelled Entry (E).

C.5.10.3 SCD Severity index

A quantitative SCD severity score (SV) was calculated according to a modified version of the method described by Sabastiani *et al.* [56] using an online sickle cell disease severity calculator (http://www.bu.edu/sicklecell/projects/) [56]. Each patient was assigned a score based on their sex, Hb genotype, mean corpuscular volume (MCV), and white blood cell (WBC) counts. Controls were assigned a score of 0.

C.5.11 Confirmation of control status

The unaffected siblings were confirmed not to have SCD by performing High Performance Liquid Chromatography (HPLC) and Capillary Electrophoresis [128] at the SCD Center in Benin. Individuals with at least one normal (HbA) allele were assigned as controls.

C.6 Variables

C.6.1 Data Sets

Depending on the analysis performed, different subsets of the data were used. See Table C.2 below. For the study that evaluated gene expression profiles implicated in SCD, case-control and case-only analyses were performed using the discovery, replication and combined data sets. For the study that tested for transcriptional biomarkers, analyses were performed to discriminate clinical categories and controls, and to discriminate SSS from SSU, were performed using the discovery and replication data sets. For the study that identified the genetic control of gene expression variation, combined data sets that included cases and controls was performed, or that only included cases for identification of the SNP-by-Clinical category interaction effects. For the study that identified potential SCD drug targets, case-controls analysis was performed using the combined data set, and a case only (SSS and SSU SCD patients) subset of the combined data set.

With combined data set I, we were interested in characterising the entire spectrum of SCD severity (measured by clinical categories) and thus kept all categories and Hb genotypes. In combined data set II, we focused on a sub-set of this sample that included HbSS patients and controls in order to characterize SCD clinical status and follow-up. HbSC individuals were excluded from combined data set II given their small sample size relative to the HbSS group. SCD patients

undergoing an acute event also were excluded from combined data set II to focus on the steady state of the disease.

C.6.2 Variables in the "Gene expression profiles and biological pathways implicated in SCD" project

Our first objective was to identify variables that influence gene expression variation in SCD and to quantify their impact. In order to address this question, the outcome variable of interest was genome-wide gene expression traits using the 28,595 expressed probes. In unsupervised analyses that evaluated the impact of variables on gene expression variation, SCD clinical severity (SCD Clinical Status, SCD Clinical categories, and SCD Severity Index), Hb genotypes (HbSS, HbSC, HbA controls), sex, blood cell counts (WBC and RBC), genetic ethnicity (gPCs), date of sampling, phase, and age were included. Interaction between Hb genotype, Clinical severity, and sex was also evaluated for their impact on gene expression variation.

In order to quantify the effects of SCD clinical severity (Clinical status and Clinical category) and Hb genotypes (exposure variables) on gene expression variation (outcome variable), an ANCOVA was performed that accounted for sex, cell counts, and genetic ethnicity (gPCs) as potential confounders.

Table C.2 Data sets.

Study	Data set	n	Probes	SNPs	Mult.test adj.		Analysis	End point	Tested variable	Covariates
Gene expression										
	Discovery	157	28,595	n/a	FDR 1%		HC, PCA, VCA, ANCOVA	Gene exp	Probes	Sex, RBC, WBC
	Case only	126	28,595	n/a	FDR 1%		VCA	Gene exp	Probes	Sex, RBC, WBC
	Replication	154	28,595	n/a	FDR 1%		HC, PCA, VCA, ANOCVA	Gene exp	Probes	Sex, RBC, WBC
	Case only	124	28,595	n/a	FDR 1%		VCA	Gene exp	Probes	Sex, RBC, WBC
	Combined I	311	28,595	n/a	FDR 1%		HC, PCA, VCA, ANCOVA	Gene exp	Probes	Sex, RBC, WBC
	Combined II	216	28,595	n/a	FDR 1%		ANCOVA, HC, GSEA	Gene exp	Probes	Sex, RBC, WBC
Biomarker										
	Discovery	157	28,595	n/a	Bonferroni		Discriminate analysis, Leave-one-out	ClinCat	Probes	Sex, RBC, WBC
	Replication	154	28,595	n/a	Bonferroni		Discriminate analysis, Leave-one-out	ClinCat	Probes	Sex, RBC, WBC
	SSSvsSSU training	38	28,595	n/a	Bonferroni		Discriminate analysis, Leave-one-out	ClinCat	Probes	Sex, RBC, WBC
	SSSvsSSU test	42	28,595	n/a	Bonferroni		Discriminate analysis, Leave-one-out	ClinCat	Probes	Sex, RBC, WBC
GWAS of gene expression					local	distal				
	Combined I	263	18,890	568,921	Bonferroni	Bonferroni	Multiple linear regression (model 1 ClinCat)	Gene exp	SNP	Sex, RBC, WBC
	Case only	205	4220	399,821	Bonferroni	Bonferroni	Multiple linear regression (SNP-by-ClinCat)	Gene exp	SNP-by-ClinCat	Sex, RBC, WBC
	Combined II	173	19,431	560,675	Bonferroni	Bonferroni	Multiple linear regression (model 1 ClinStatus)	Gene exp	SNP	Sex, RBC, WBC
	Combined II	173	7002	455,750	Bonferroni	Bonferroni	Multiple linear regression (SNP-by-ClinStatus)	Gene exp	SNP-by-ClinStatus	Sex, RBC, WBC
Drug target										
	Cases	205	18,890	568,921	Bonferroni	Bonferroni	Basic linear regression (model 5)	Drugs	Genes	Sex, RBC, WBC
	Controls	58	18,890	568,921	Bonferroni	Bonferroni	Basic linear regression (model 5)	Drugs	Genes	Sex, RBC, WBC
	SSSvsSSU	80	18,890	568,921	Bonferroni	Bonferroni	Linear regression	Drugs	Genes	Sex, RBC, WBC

*For each study indicated, a particular data set was used: Combined I = combined data set I; Combined II = combined data set II; SSS= steady-state satisfactory; SSU= steady-state unsatisfactory. The number of participants in each data set is indicated (n). The tested number of probes and SNPs are listed, as well as the corresponding method of adjusting for multiple testing (Mult.testing adj.). The analyses performed are abreviated: HC= hierarchical clustering; PCA= principal component analysis; VCA= variance component analysis; ANCOVA= analysis of covariance; GSEA= gene set enrichment analysis. The variable of interest in each analysis (end point) is listed, as well as the tested variable and covariates: RBC= red blood cells; WBC= white blood cells.

In order to identify biological pathways that impact the course of SCD through gene expression, the outcome variable of interest was gene expression profiles, and these were produced for patients grouped according to their SCD clinical severity (Clinical Status and Clinical categories). In this analysis, we accounted for Hb genotype, blood cell counts, and sex as potential confounder variables.

C.6.3 Variables in the "Transcriptional biomarkers of SCD clinical severity" project

In order to identify transcriptional biomarkers that discriminate SCD patients by clinical category, the exposure variables were the gene expression levels from the 28,595 expressed probes, which were used to discriminate SCD patients based on their Clinical category (outcome variable).

C.6.4 Variables in the "Genetic control of gene expression variation in SCD patients" project

In order to identify the genetic regulation of gene expression in SCD, linear regression models were run. Gene expression was the outcome variables of interest, and genome-wide genotyped SNPs were exposure variables that were tested for association with gene expression traits. We accounted for Hb genotype, clinical severity (Clinical Status and Clinical Category), cell counts (WBC and RBC), and sex.

In order to determine if there was modification of the association between an individual's genotype (exposure) and their gene expression level (outcome) based on their clinical severity (interaction effect), we used the same variables as explained above, and included an interaction term: SNP-by-Clinical severity (SNP-by-Clinical Status and SNP-by-Clinical Category). For this analysis, we also accounted for relatedness (potential confounder).

C.6.5 Variables in the "Potential SCD drug targets" project

In order to identify known drug targets that could be used in SCD, 2 eSNP analyses were run. In the first eSNP analysis, genes that controlled gene expression in SCD patients and controls were identified. In the second eSNP analysis, genes that controlled gene expression in steady-state satisfactory and unsatisfactory SCD patients were identified. These genes (dependent variable) were then examined to determine if there were drug targets (outcome variable) described in public pharmacogenetic databases.

C.7 Statistical methods

C.7.1 General results methods

C.7.1.i Relatedness

In order for the assumption of independent observations to hold, it is important to account for relatedness in GWAS. We tested for relatedness by estimating genome-wide identity-by-descent (IBD) using Jmp Genomics/SAS and PLINK [126]. IBD is a measure of how many alleles at any marker in two samples come from the same ancestral chromosomes. The probability that zero, one, or two alleles are identical by descent ("shared IBD") is denoted by the notations P(Z=0), P(Z=1), and P(Z=2), respectively. These probabilities may either refer to given markers or be thought of as sample-wide. Pi-hat is a measure of IBD estimated using PLINK, and is equal to P(Z=2) plus one-half of P(Z=1). This is the probable number of shared alleles at any given marker. Although there are other methods to estimate relatedness that may be more accurate [130,131,132], IBD and Pi-hat can achieve reasonable estimates of relatedness among pairs of subjects that is sufficiently appropriate for quality assurance purposes.

In order to avoid biases from groups of correlated markers, SNPs were pruned using PLINK based on marker properties and LD (SNPs with any missing data were removed and only autosomal SNPs were included, all SNPs that were in LD (r^2<0.3) were removed). This left 193,652 SNPs for further analyses. In order to decrease computational time in future analyses that included the relatedness matrix, a random sample of SNPs was chosen, with a per-SNP probability of being kept equal to 0.01. This left a final number of genome-wide SNPs equal to 1,986 SNPs for the relatedness analyses [123]. The relatedness matrix that was generated using 1,986 SNPs was compared to the matrix that was created using 327,554 SNPs. Similar levels of relatedness were obtained (correlation of 0.98; see Appendix IV).

C.7.1.ii Ancestry analysis

Global genotypic variation and ancestry was inferred based on principal component analysis using Eigenstrat proposed by Price *et al.* [119,120]. This method is commonly used to adjust the results of GWAS and correct for stratification. Ancestry analyses of 119, 104 and 235 unrelated individuals from the discovery, replication and combined I data sets were performed using 485,000 genotypes. Significance of genotypic principal components (*gPCs*) was tested using the Tracy-Widom test and the percent of variance in the data that was explained by each gPC was calculated.

C.7.2 Gene expression profiles and biological pathways implicated in SCD analyses

All statistical analyses of the gene expression data were performed using JMP Genomics v5.0 (SAS), and SAS 9.3 (SAS).

C.7.2.i Unsupervised statistical analyses

The unsupervised statistical analyses performed below, are an attempt to find hidden structure in the data without any *a priori* hypothesis.

C.7.2.i.1 Hierarchical clustering

The CLUSTER procedure in SAS was used to perform one-way and two-way hierarchical clustering using Ward's minimum-variance method [133]. This analysis allows an illustration of how the data clusters without any *a priori* hypothesis. One-way clustering (clustering based on each participants' gene expression level or clustering based on samples), identifies co-regulated and functionally related genes or sub-types of related samples. Two-way clustering (clustering of gene expression levels with samples), identifies which genes are the most important for sample clustering. The method is based on an agglomerative hierarchical clustering procedure. Each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of the two closest clusters is repeated until only one cluster remains. One-way clustering has been widely performed in current biological research for discovering and understanding gene functional relationships [134], or used in biomedical research where clustering disease samples to diagnose disease types or disease progress [135,136].

C.7.2.i.2 Principal and Variance component analysis

Principal Component analysis (PCA) finds low dimensional linear combinations of data with maximal variability. Variance Component analysis (VCA) attributes and partitions variability into known sources via a classical random effects model. Both PCA and VCA of the gene expression data were performed such that

the first three expression PCs (ePC) are modeled either simultaneously or individually as a function of various effects in the data: Hemoglobin genotype, SCD severity (Clinical status or Clinical category), Sex, and pair-wise combination of fixed effects.

C.7.2.ii Supervised statistical analyses

The supervised statistical analyses that were performed are described below.

C.7.2.ii.1 ANCOVA

SAS GLM was used to evaluate the magnitude and significance of differentially expressed probes. Probe-level differential expression analyses were performed using analysis of covariance. Variance was partitioned among the hemoglobin genotype (Hb), clinical severity (clinical status effect or the clinical category effect), sex, and total blood cell counts (red (RBC) and white blood cells (WBC)) as covariates. The effects of date of sampling, phase (discovery vs replication), age (in years), and gPCs were tested. Pairwise contrasts (Hb genotype*Sex, Hb genotype*ClinStatus/Clinical category and ClinStatus/Clinical Category*Sex) also were evaluated. The following ANCOVA models were of interest:

Expression = μ + Hb genotype + SCD Clinical Status + Sex + WBC + RBC + ε

Expression = μ + Hb genotype + SCD Clinical Category + Sex + WBC + RBC + ε

The error ε was assumed to be normally distributed with mean equal to zero. The number of probes that were differentially expressed between the 3-way SCD Clinical status effect (EvsFUvsCtls) and the 6-way SCD clinical category effect (AvsSSSvsSSUvsE) were also evaluated. Since multiple comparisons were

performed, the p-value was corrected using the method of False Discovery Rate (FDR=1%) and was applied separately to each term in the analysis of covariance.

C.7.2.ii.2 Enrichment analysis

Enrichment analysis for the differentially expressed genes in the SCD Clinical status contrasts and the SCD Clinical category contrasts were performed using Gene set enrichment analysis (GSEA) [137] on the C2 collections of MsigDB database (http://www.broad.mit.edu/gsea/msigdb). Merged to C2 are 28 gene sets collected from transcriptional analyses of PBMC samples in an immunological context, and are named M x.x (eg. M 1.3). Their descriptions can be found in Chaussabel *et al.* [138]. Also included are 10 gene sets generated from profiling major immune cell subsets found in the human blood [139], prefixed with “Bali”. The resulting p values from the GSEA were adjusted for multiple testing by using Benjamini and Hochberg method [140] to control the false discovery rate. Results were considered to be significant if the adjusted p values are <0.05. A Normalized Enrichment Scores (NES) greater than 0.25 in a SCD patient group relative to the controls was used.

C.7.3 Identification of transcriptional biomarkers analyses

C.7.3.i Discriminant analysis

Using the entire gene expression data (28,595 expressed probes), discriminant analysis was performed to identify biomarkers for each clinical category. Discriminant analysis is a classical statistical method for predicting a classification variable (SCD clinical category) from a set of continuous responses (probes). SAS K-means clustering was used to select one representative predictor from each cluster. The default settings were applied, which included a maximum number of K-means clusters per predictors set to 500. T-tests were performed to filter the predictors

(using FDR 1%). A linear metric was used to compute distances between groups with a forward stepwise variable selection method. Each clinical category's proportion was calculated in the sample and used as a prior. The best model was chosen based on likelihood ratio test.

Using the selected probes identified in the discovery phase to predict clinical category, each individual's clinical category was predicted in the replication sample and accuracy was measured.

For the analysis that predicted SSS from SSU patients, a receiver operating curve (ROC) was generated. A ROC curve is a graphical plot which illustrates the performance of a binary system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR= true positive rate) versus the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate. The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random), points below the line poor results (worse than random). The ROC can be used to generate summary statistics. When using normalized units, the area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming "positive" ranks higher than "negative").

C.7.3.ii Leave-one-out cross validation

In order to assess the accuracy of the results of the discriminant analysis, leave-one-out cross-validation was performed. This technique involves using a single observation from the original sample as the validation data, and the remaining

observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

Leave-one-out cross validation was performed on combined data set I (n=311, K=311), and on the SSS and SSU data set (n=80) using the cv.glm function [141] in the package boot in R [142]. This method refits the generalized linear model with a subset of data (training set: n-1) and calculates the accuracy of the prediction on the remaining data. The SCD severity index was the response variable that was predicted. For the clinical category analysis, a prediction was made for each participant based on the 310 remaining participants' biomarker's gene expression levels. For the SSS and SSU analysis, a prediction was made for each participant based on the 79 remaining participants' biomarker's gene expression levels. R^2 and p-values were calculated for the linear regression analysis between predicted and actual values.

C.7.4 Genetic control of Gene Expression analyses

C.7.4.i GWAS of gene expression

GWAS of gene expression analysis, a method that integrates genome-wide genotying data with genome-wide gene expression data [92,97,98,99,100], was performed using JMP Genomics v5.0 (SAS), SAS 9.3 (SAS), and PLINK [126]. Linear regression analyses were modeled assuming a co-dominance mode of inheritance, where each allele has an additive effect on the level of gene expression that is transcribed.

C.7.4.i.1 Multiple linear regression

Multiple linear regression analyses were performed using PLINK to test for significant associations between gene expression levels and SNP genotype. In these analyses, we tested for association between probe expression levels and SNP genotype while accounting for clinical severity (Clinical Status or Clinical Category), sex and blood cell counts (white and red blood cell counts, WBC and RBC), assuming that the error ε is normally distributed with a mean of zero, where:

Model 1: Expression = μ + SNP+ Clinical Status + WBC + RBC + Sex+ ε

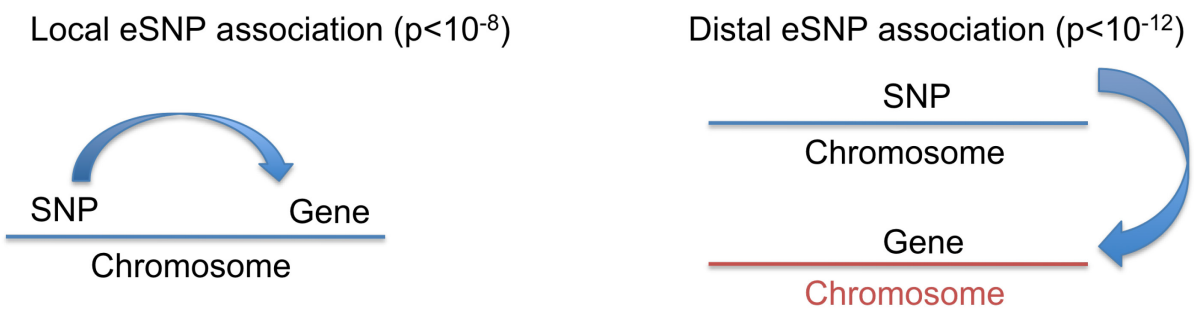
Model 2: Expression = μ + SNP+ Clinical Category + WBC + RBC + Sex+ ε

Only well-annotated, autosomal probes with validated chromosomal location and gene function based on the most recent annotation in NCBI and UCSC as of October 2011 were included for the association tests. In the process, all probes were aligned to the reference genome (hg19), ambiguous probes and all non-RefSeq probes were excluded, and probes overlaying known SNPs (427 of them) were removed from the analysis. SNPs that had a minor allelic frequency <5 %, an exact Hardy-Weinberg (HWE) test *P* value <0.001 and > 1% missing data, calculated seperately for each data set, were excluded. This resulted in a total of 19,431 expressed probes that were tested for association with 560,675 SNPs in model 1, and 18,890 expressed probes that were tested for association with 568,921 SNPs in model 2.

We distinguish between eSNP associations based on the chromosomal location of the probe-SNP pair and the distance between them (see Figure C.6). A local association implicates a probe and a SNP located on the same chromosome that is 1 Megabase pairs (Mb) or less apart from each other (it is assumed that a 1

Mb region has on average 200 SNPs); while a distal cis association is an association between a SNP and probe located on the same chromosome but further than 1 Mb apart. A distal association implicates a probe and a SNP located on different chromosomes. We applied Bonferroni correction for all eSNP associations by accounting for both the number of SNPs and probes tested.

Figure C.6 Representation of local and distal eSNP associations.
Cis associations are eSNP associations between a SNP and a gene located on the same chromosome no more than 1Mb apart between the SNP-probe pair. Distal eSNP associations are on different chromosomes.



Since 560,675 SNPs were tested for association with 19,431 probes in model 1, a genome-wide Bonferroni threshold for distal-associations corresponds to 0.05/(19,431 probes x 560,675 SNP) = 4.59×10^{-12} and for local associations to a Bonferroni threshold of 0.05/(19,431 probes x 200 SNPs) = 1.28×10^{-8} considering an average number of 200 SNPs tested against each probe.

Since 568,921 SNPs were tested for association with 18,890 probes in model 2, a genome-wide Bonferroni threshold for distal-associations corresponds to 0.05/(18,890 probes x 568,921 SNP) = 4.65×10^{-12} and for local associations to a

Bonferroni threshold of 0.05/(18,890 probes x 200 SNPs) = 1.32 x 10^{-8} considering an average number of 200 SNPs tested against each probe.

C.7.4.ii Detection of Interaction effects

Here we test for statistical transcriptional interactions using the method of Idaghdour *et al.* [92], where an interaction is described as the differential effect of a given genotype exposed to different clinical outcomes (Clinical status or clinical category) associated with gene expression. In model 3, we test for SNP-by-Clinical Status interaction effects using 7,002 differentially expressed probes for the 3-way Clinical Status effect and using combined data set II. In model 4, we test for SNP-by-Clinical Category effects using 4,220 differentially expressed probes for the 6-way Clinical Category effect using a subset of combined data set I that contains only SCD patients.

Model 3: Expression = μ + SNP+ Clinical Status + WBC + RBC + Sex + SNPxClinStatus + ε

Model 4: Expression = μ + SNP+ Clinical Category + WBC + RBC + Sex+ SNPxClinCat + ε

where ε is assumed to be normally distributed with a mean of zero.

For these analyses, we reduced the effect of outlier expression values by further filtering the set of genotypes for each sub-group. We calculated marker properties in each of the sub-groups of patients separately and included only SNPs that had a minor allelic frequency <5 %, an exact Hardy-Weinberg (HWE) test *P* value <0.001 and > 1% missing data for each sub-groups of patients. A final number of 455,750 SNPs in combined set II, and 399,821 SNPs in combined set I were used.

This resulted in 7,002 differentially expressed probes for the 3-way Clinical Status effect that were tested for association with 455,750 SNPs in model 3; and 4,220 differentially expressed probes for the 6-way clinical category effect that were tested for association with 399,821 SNPs in model 4.

A genome-wide Bonferroni correction was applied by accounting for both the number of SNPs and probes tested in these analyses. Since 455,750 SNPs were tested for association with 7,002 probes in model 3, a genome-wide Bonferroni threshold for distal-associations corresponds to 0.05/(7,002 probes x 455,750 SNP) = 1.57×10^{-11} and for local associations to a Bonferroni threshold of 0.05/(7,002 probes x 200 SNPs) = 3.57×10^{-08} considering an average number of 200 SNPs tested against each probe.

Since 399,821 SNPs were tested for association with 4,220 probes in model 4, a genome-wide Bonferroni threshold for distal-associations corresponds to 0.05/(4,220 probes x 399,821 SNP) = 2.96×10^{-11} and for local associations to a Bonferroni threshold of 0.05/(4,220 probes x 200 SNPs) = 5.92×10^{-8} considering an average number of 200 SNPs tested against each probe.

C.7.4.ii.1 Q-K mixed model that includes relatedness matrix

Because cases and controls were potentially related, we included the pairwise relatedness estimates (IBD) calculated for each of the SCD individuals and controls and used this matrix to estimate the random effect of shared genetic and environmental components in a Q-K mixed model [143]. Only autosomal SNPs with a MAF>0.1, missingness=0, and that were not in linkage disequilibrium (r^2<0.3) were included in estimating relatedness (final number of SNPs = 1,992 SNPs). Using the GLIMMIX procedure in SAS and including this relatedness matrix in a mixed model,

we reran the analyses to identify interaction effects that remained significant after accounting for the random effects of shared genetic and environmental components.

C.7.4.iii Comparison of significant eQTL results in SCD with published eQTL results

The significant SNP-expression (eSNP) associations identified in SCD in this study were compared to the associations reported in 12 published eQTL studies of peripheral blood or its derivatives at nominal P-values < 10^{-7}. These published associations were accessed using the eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowser/eqtl/) and compared to our study results.

C.7.5 Identification of potential drug targets in SCD analyses

In order to identify drug target genes in SCD, 2 basic eSNP analyses were run for SCD patients (n=205) and controls (n=58), separately (model 5).

Model 5: Expression = μ + SNP + ε

The genes that were significantly associated with SNPs were evaluated to identify if they were drug targets. A drug target is described as a gene that is affected by a chemical or drug in the Comparative Toxicogenomics Database (CTD database: http://ctdbase.org/). The CTD is a public website and research tool that curates scientific data describing relationships between chemicals, genes and human diseases [144]. In Figure C.7, a flow diagram of the approach is shown.

In a second analysis, candidate drug target genes for SCD patients in unsatisfactory conditions (SSU) were identified by examining the overlap between eSNP genes that were differentially expressed between SSS (steady-state

satisfactory) and SSU (steady-state unsatisfactory) patients and the CTD database. See Figure C.8 for the flow diagram for these analyses.

Figure C.7 Flow diagram for SCD drug targets

Steps used in identifying drug targets for SCD patients.

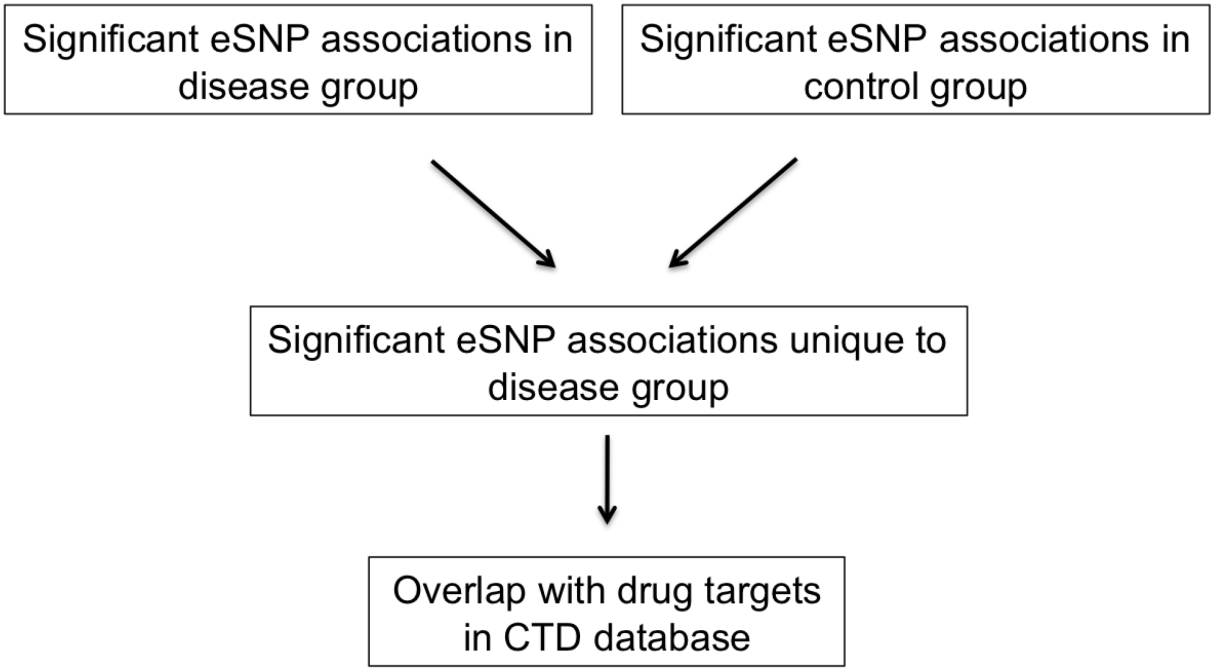
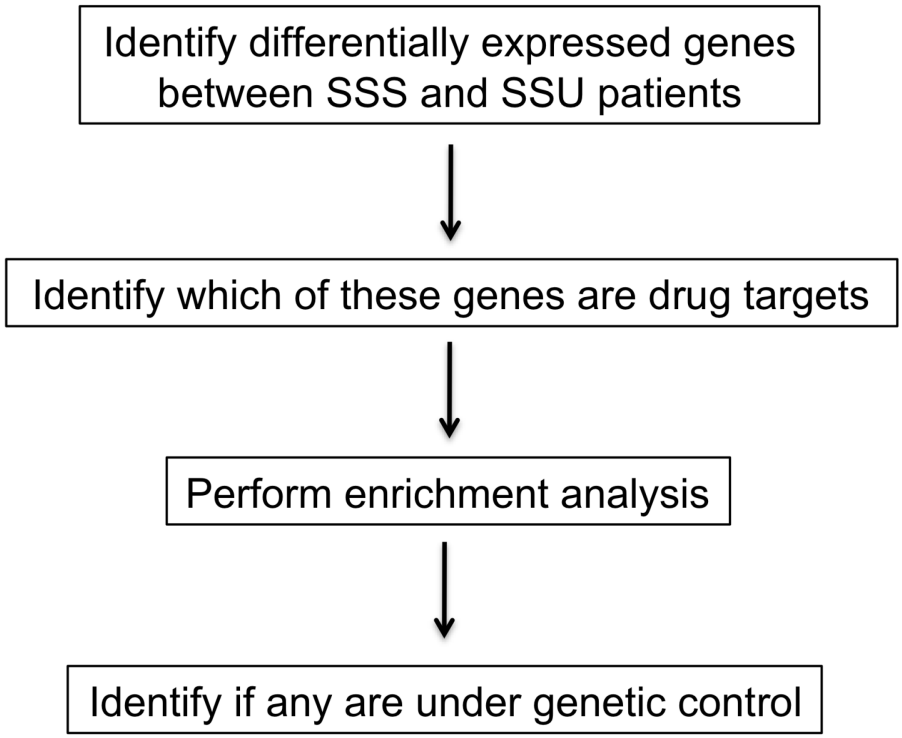


Figure C.8 Flow diagram for drug targets of SCD progression.

Steps used in identifying drug targets for genes differentially expressed between SSS and SSU SCD patients that were under genetic control.



D. RESULTS

D.1 GENERAL RESULTS

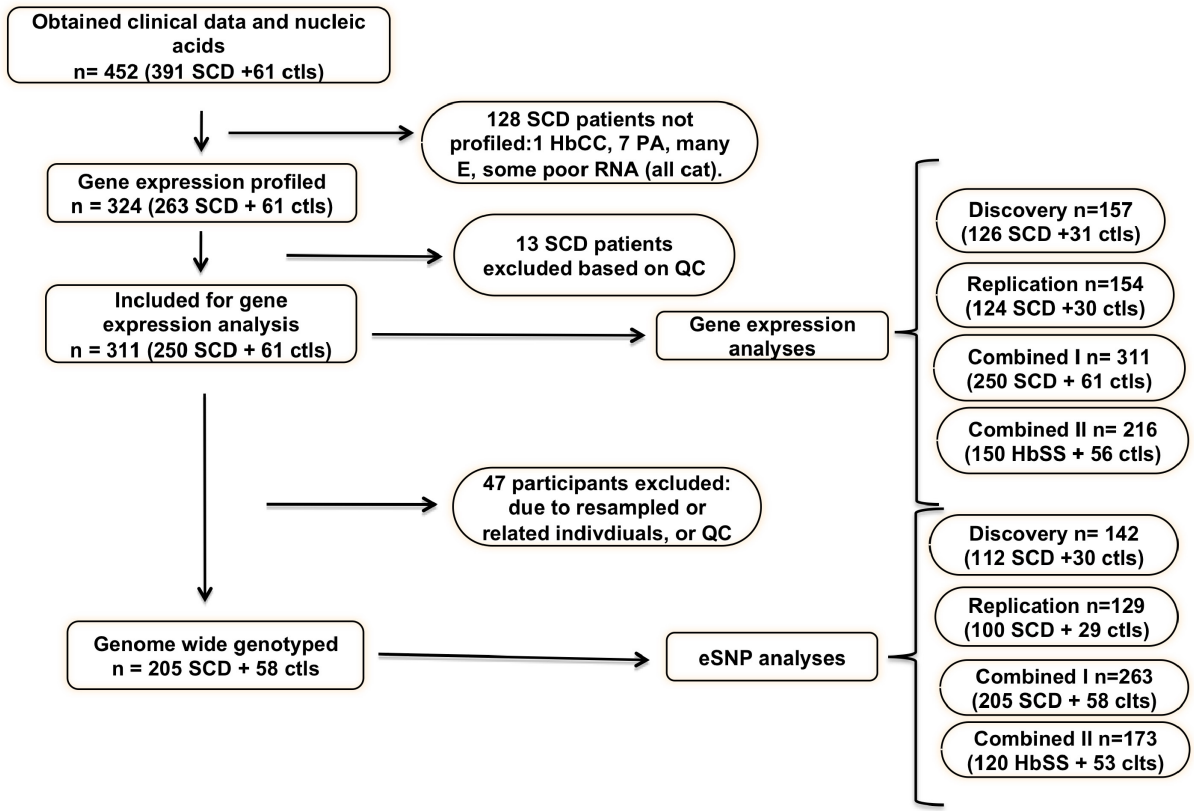
D.1.1 Flow diagram of participant selection

As previously mentioned, participants were recruited from the SCD cohort in Benin during 2010. For the purpose of this study, a total of 452 participants with clinical data and blood samples for nucleic acids isolation were recruited. Three hundred twenty four (324) of these participants were gene expression profiled. We excluded 128 participants due either to small sample sizes in some categories (HbCC and Post-Acute (PA) patients), or because of over-representation of the Entry (E) patients, or because of poor RNA quality in some samples.

After gene expression profiling, three hundred eleven (311) samples were retained for further analysis based on quality control checks during the hybridisation, normalisation and standardisation steps. This made a final sample size of 126 in the Discovery phase, 124 in the Replication phase, and 311 in combined data set I for gene expression analyses.

We genotyped 263 of the 311 patients. Forty seven participants were excluded due to quality control checks, relatedness, or re-sampling. This made a final sample size of 142 in the Discovery phase, 129 in the Replication, and 263 in combined data set I for joint gene expression and genotyping analyses (eSNP analyses).

Figure D.1.1 Flow diagram of participant selection.



D.1.2 Participant characteristics: demographic and clinical data

Demographic and clinical characteristics for the 311 recruited participants are shown in Table D.1.1 below. Approximately three quarters of the SCD sample had the HbSS genotype (n=190; 76%) and one quarter were compound heterozygotes (HbSC; n= 60; 24%). Sixty one participants had at least one HbA allele and were used as controls. The odds ratio of a SCD patient having an HbSS genotype (HbSS vs HbSC) and having a particular SCD clinical category was calculated, but was not significant (Chi Square p value= 0.2783). See Table D.1.2 A and D.1.2.C below. Roughly, equal numbers of boys and girls were recruited. One hundred and thirty five SCD patients were male (54%), and one hundred fifteen were female (46%). More boys were Acute (67%) or SSU (61%) than girls, but this was not significant (OR and

95% confidence intervals, as well as Chi Square p-values are shown in Table D.1.2 B and D.1.2.C below). The distribution of the main variables is similar for each dataset (Figure D.1.2).

Three quarters of our controls were heterozygous HbAS and one quarter were homozygous HbAA. Only 14 probes were differentially expressed between HbAA and HbAS individuals at FDR 1%. Furthermore, none of the variance in the controls was explained by this effect as evidenced by variance component analysis and by the lack of clustering based on Hb genotype in the PCA analysis (Figure D.1.3). For these reasons, we grouped HbAA and HbAS individuals and used them as a control sample.

Table D.1.1 Participant (n=311) characteristics.

Clinical category	Entry			Acute			SSS			SSU			Ctls		
Data set	C	D	R	C	D	R	C	D	R	C	D	R	C	D	R
Caracteristics															
N	134	70	64	36	18	18	42	20	22	38	18	20	61	31	30
Hb genotype															
HbSS	101	53	48	29	15	14	28	15	13	32	16	16	0	0	0
HbSC	33	17	16	7	3	4	14	5	9	6	2	4	0	0	0
Ctls (HbAA +HbAS)	0	0	0	0	0	0	0	0	0	0	0	0	61	31	30
Sex (n)															
M	66	32	34	24	32	10	22	10	10	23	9	14	26	15	12
F	68	38	30	12	4	8	20	10	12	15	9	6	35	16	18
Cell Counts (mean)															
RBC	3.76	3.76	3.8	3.19	3.35	3	3.64	3.5	3.8	3.5	3.33	3.7	4.7**	4.75	4.7
WBC	12.05	12.34	12	16.84*	16.11	18	12.02	14.1	10	13.58	14.78	13	9.29**	9.06	9.5
Parasetemia (# of infec)	2	1	1	1	0	1	4	3	1	4	2	2	0**	0	0
Age (mean years)	3.86	4.05	3.66	4.36	4.57	4.14	4.73	4.4	5.01	5.08	5.19	4.98	2.86*	2.86	2.91
Self-declared Ethnicity															
Fon	64	31	33	16	5	11	16	7	9	18	10	8	18	11	7
Goun	14	8	6	4	4	0	5	3	2	6	3	3	10	2	8
Yoruba	17	9	8	7	3	4	6	4	2	1	0	1	3	2	1
Other	36	21	15	8	6	2	14	5	9	13	5	8	28	15	13
Missing	3	1	2	1	0	1	1	1	0	0	0	0	2	1	1

* Characteristics of participants in Combined dataset I (C), Discovery phase (D) and in Replication phase (R) Hb genotype, Sex (M=males, F=females), Cell counts (RBC=red blood cells, WBC = white blood cell, the number of indivdiuals infected with malaria parasetemia (# of infec), age (years), and self-declared Ethnicity are measured by clinical category: Entry, Acute, steady-state satisfactory (SSS), steady-state- unsatisfactory (SSU) controls. Significant differences are indicated for combined dataset I. ** $p<0.0001$ Nonparametric comparisons for each pair using the Wilcoxon Method. *$p<0.05$ Nonparametric comparisons for each pair using the Wilcoxon Method.

Table D.1.2 SCD clinical categories.

Clinical categories for Combined dataset I (C), Discovery (D), and Replication (R) phases are stratified by A) Hb genotype or by B) sex. C) The odds ratios for a SCD patient having a particular genotype and having a particular clinical category was not significant. Neither was the odds ratios for a SCD patient being a boy or girl and having a particular clinical category.

A)

	Entry			Acute			SSS			SSU			Ctls		
	C	D	R	C	D	R	C	D	R	C	D	R	C	D	R
HbSS	101	53	48	29	15	14	28	15	13	32	16	16	0	0	0
HbSC	33	17	16	7	3	4	14	5	9	6	2	4	0	0	0
Total	134	70	64	36	18	18	42	20	22	38	18	20	0	0	0

*Chi Square=3.849
*p-value=0.2783

B)

	Entry			Acute			SSS			SSU			Ctls		
	C	D	R	C	D	R	C	D	R	C	D	R	C	D	R
M	66	32	34	24	32	10	22	10	10	23	9	14	26	15	12
F	68	38	30	12	4	8	20	10	12	15	9	6	35	16	18

*Chi Square=4.236
*p-value=0.2370

C)

	Hb genotype			Sex		
Comparison	Combined I	Discovery	Replication	Combined I	Discovery	Replication
ClinCat	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
EvsA	0.74 (1.8-0.3)	0.62 (2.4-0.2)	0.86 (2.9-0.25)	0.49 (0.2-1.1)	0.10 (0.3-0.03)	0.91 (2.6-0.3)
EvsSSS	1.53 (3.2-0.7)	1.04 (3.3-0.3)	2.08 (5.8-0.8)	0.88 (0.4-1.8)	0.84 (2.3-0.3)	1.36 (3.6-0.5)
EvsSSU	0.57 (1.5-0.2)	0.39 (1.9-0.08)	0.75 (2.8-0.2)	0.63 (0.3-1.3)	0.84 (2.4-0.3)	0.49 (1.4-0.2)
AvsSSS	2.07 (5.9-0.9)	1.67 (8.3-0.3)	2.42 (9.8-0.6)	1.82 (0.7-4.6)	8 (31.2-2.1)	1.5 (5.2-0.4)
AvsSSU	0.78 (2.6-0.2)	0.63 (4.4-0.09)	0.88 (4.2-0.2)	1.3 (0.5-3.4)	8 (32.1-2.0)	0.54 (2.0-0.2)
SSSvsSSU	0.38 (1.1-0.1)	0.38 (2.2-0.06)	0.36 (1.4-0.09)	0.72 (0.3-1.7)	1 (2.6-0.3)	0.36 (1.2-0.1)

*Calculated using OpenEpi. Version 3, open source calculator-RbyC. Rosner, B. Fundamentals of Biostatistics. 5th ed. Duxbury Thompson Learning. 2000; p.395

Figure D.1.2 Pie charts.

Pie charts of patient characteristics in the discovery, replication, and combined I and II datasets for genome-wide gene expression analysis (Exp) and genome-wide genotyping analysis (Geno). Each main variable (Clinical status: Entry (E), Follow-up (FU), and Controls (C); Hb Genotype: HbSS, HbSC, Controls (Ctls, C); and Sex: Female (F), Male (M)) was sampled in equal proportions in the datasets.

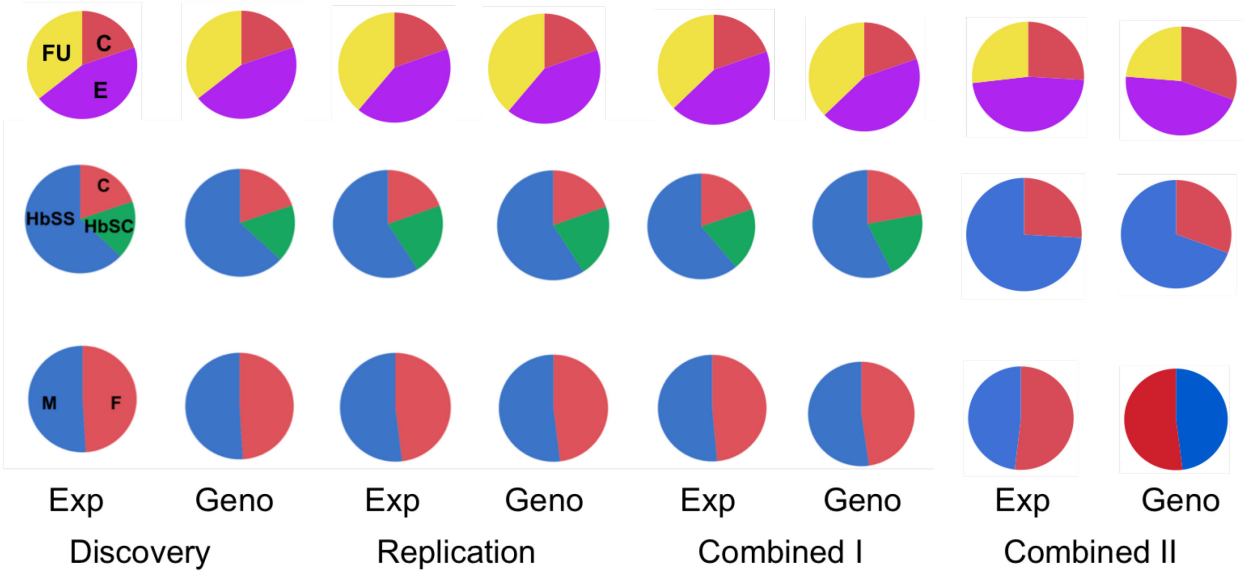
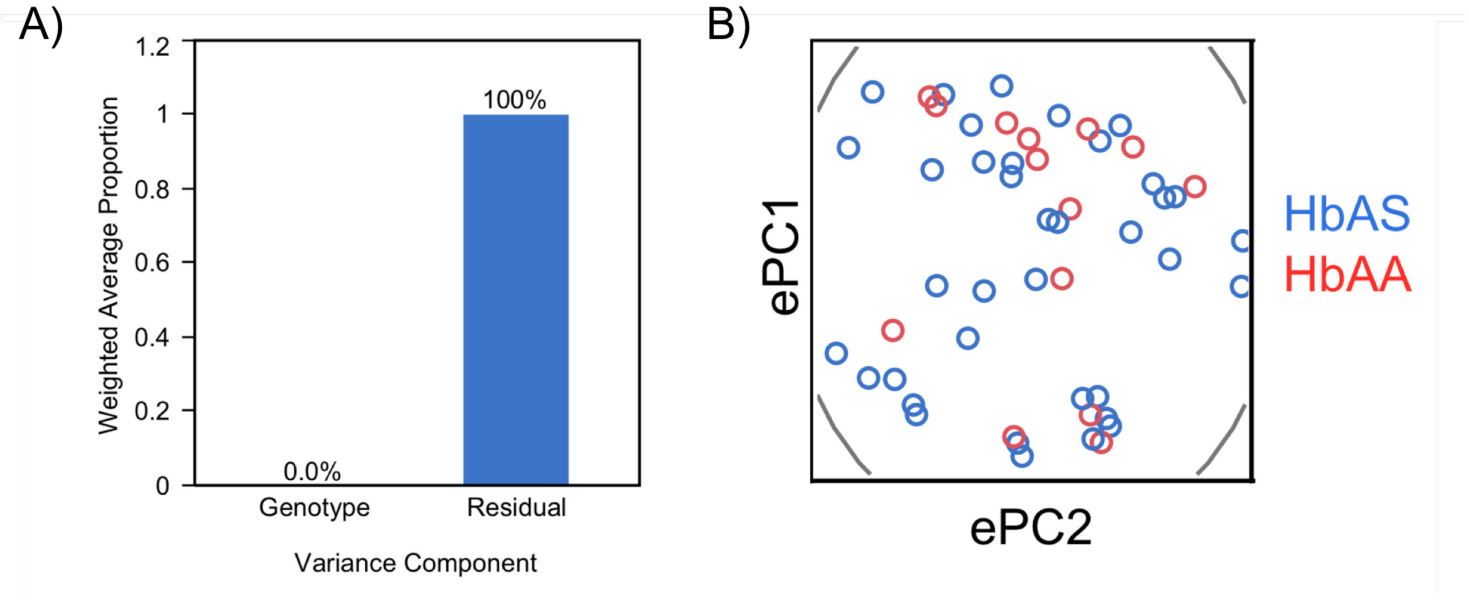


Figure D.1.3 VCA in the controls.

Variance component analysis for the Hb genotype effect (HbAS vs HbAA) in the controls on the first three expression principal components (ePC1-3) explains zero percent of the total variance in the combined dataset (A), and PC analysis identified a lack of clustering based on Hb genotype (B).



D.1.3 Hematological variables

Complete blood counts (CBC) were obtained for each participant. The CBC hematological variables that were analysed were red blood cell counts (RBC cells/pL), white blood cell counts (WBC cells/pL), and malaria parasetemia levels. The hematological variables were evaluated for correlation with each other. RBC were negatively correlated with WBC (r^2= -0.4892). Nonparametric tests were evaluated using the Wilcoxon method for significant differences between the hematological variables and SCD clinical categories (see Table D.1.1). Acute SCD patients had a significant difference in their white blood cell counts ($p<0.05$) when compared to Entry and SSS SCD patients. Controls were significantly different from all SCD clinical category patient groups for all three hematological variables ($p<0.001$).

D.1.4 β-SCD Haplotypes

It is possible that SCD severity would be influenced by β-SCD haplotypes. In order to ensure that SCD clinical severity was not driven by a β-SCD haplotype in the study, clinical category and β-SCD haplotypes were tested for association. A 61,345 base pair (bp) region on chromosome 11 that spanned the β-globin locus was genotyped using 21 additional SNPs that were not included on the Illumina array and β-SCD haplotypes were constructed. In the Appendix V, marker properties of the 21 genotyped SNPs are shown. One hundred nineteen (119) out of 169 SCD patients that were successfully genotyped were homozygous for the causal rs334 SNP genotype, confirming their HbSS genotype (Table D.1.3). Out of the 119 homozygous HbSS SCD patients, 109 (92%) had the Benin haplotype and 10 had other minor, atypical haplotypes. Fifty SCD patients were heterozygous for the rs334 genotype

and were presumed to have the HbSC genotype. Out of the 50 controls, 34 were HbAS carriers based on their heterozygous status for the rs334 mutation, and 16 were HbAA homozygous for the normal rs334 genotype. No β-SCD haplotype was significantly associated with a SCD clinical category.

Table D.1.3 β-SCD haplotypes.

Characterisation was performed in 237 genotyped participants were not significantly associated with clinical categories.

HbS mutation	Haplotype (Xmn1, Hind3, Hinc2)	Hb geno	Total	Clinical Category
"+/+"	"-/-,+/+,+/+" Benin	HbSS	109	20A,19SSU,16SSS,49E,5C
"+/+"	"+/-,+/-, +/+" Minor	HbSS	1	1A
"+/+"	"+/-, +/+, +/+" Minor	HbSS	5	1A, 4E
"+/+"	"-/-, +/-, +/+" Minor	HbSS	4	1A, 2SSS, 1E
"+/-"	HbS "+/-" suspected HbC	HbSC	50	7A, 7SSU, 13SSS, 23E
"+/-"	HbS "+/-" carrier	HbAS	34	34C
"-/-"	HbS "-/-" controls	HbAA	16	16C
Failed	Failed	-	18	n/a

D.1.5 Relatedness of SCD participants, genetic ethnicity and population structure

Identity by descent (IBD) and Pi-hat, estimates of relatedness, were calculated for SCD patients using Jmp/SAS and PLINK, respectively. The distribution of IBD and pi-hat in the genotyped individuals are plotted below in Figure D.1.4 and D.1.5. The Relationship matrix for the SCD patients was constructed using the IBD values across all marker variables (Figure D.1.6). Overall, there was marginal relatedness among the samples. Ten (10) pairs of participants had a pi-hat greater than 0.04 and were excluded from the Principal Component Analysis (PCA) that was performed using Eigenstrat in order to identify SCD participants' genetic ethnicity. In this analysis, individuals from HapMap were included and 518,000 shared genome-wide genotyped SNPs were used. After filtering by minor allele frequency and removal of outliers, the first 2 genotypic principal components (gPCs) were significant. In Figure D.1.7, these two gPCs are shown. The first gPC separates European CEPH (green) from Africans, and the second gPC separates Benin SCD participants (red) from Nigerian Yorubans (YRI: black). Additional PCAs were run in order to evaluate the population structure within the SCD population (Figure D.1.8). No significant gPCs were identified, indicating that population structure is not substantial in this sample. Also, no correlation was observed between gPCs and clinical status, clinical category, sex, or Hb genotype (Figure D.1.8).

Figure D.1.4 IBD distribution

IBD distribution of pairs of genotyped SCD participants identified few related participants.

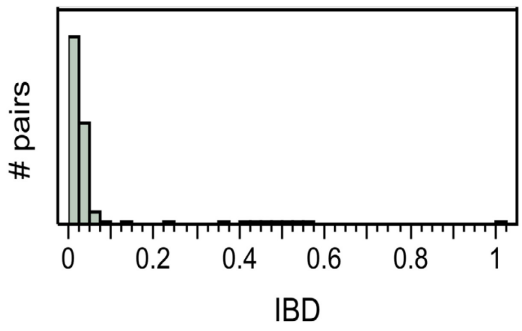


Figure D.1.5 Pi-hat distribution

Pi-hat distribution of pairs of genotyped SCD participants identifies related participants.

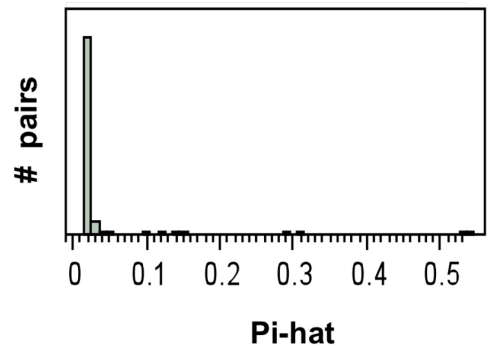


Figure D.1.6 Relationship matrix

Relationship matrix for 263 participants was calculated using IBD estimates computed across genome-wide marker variables.

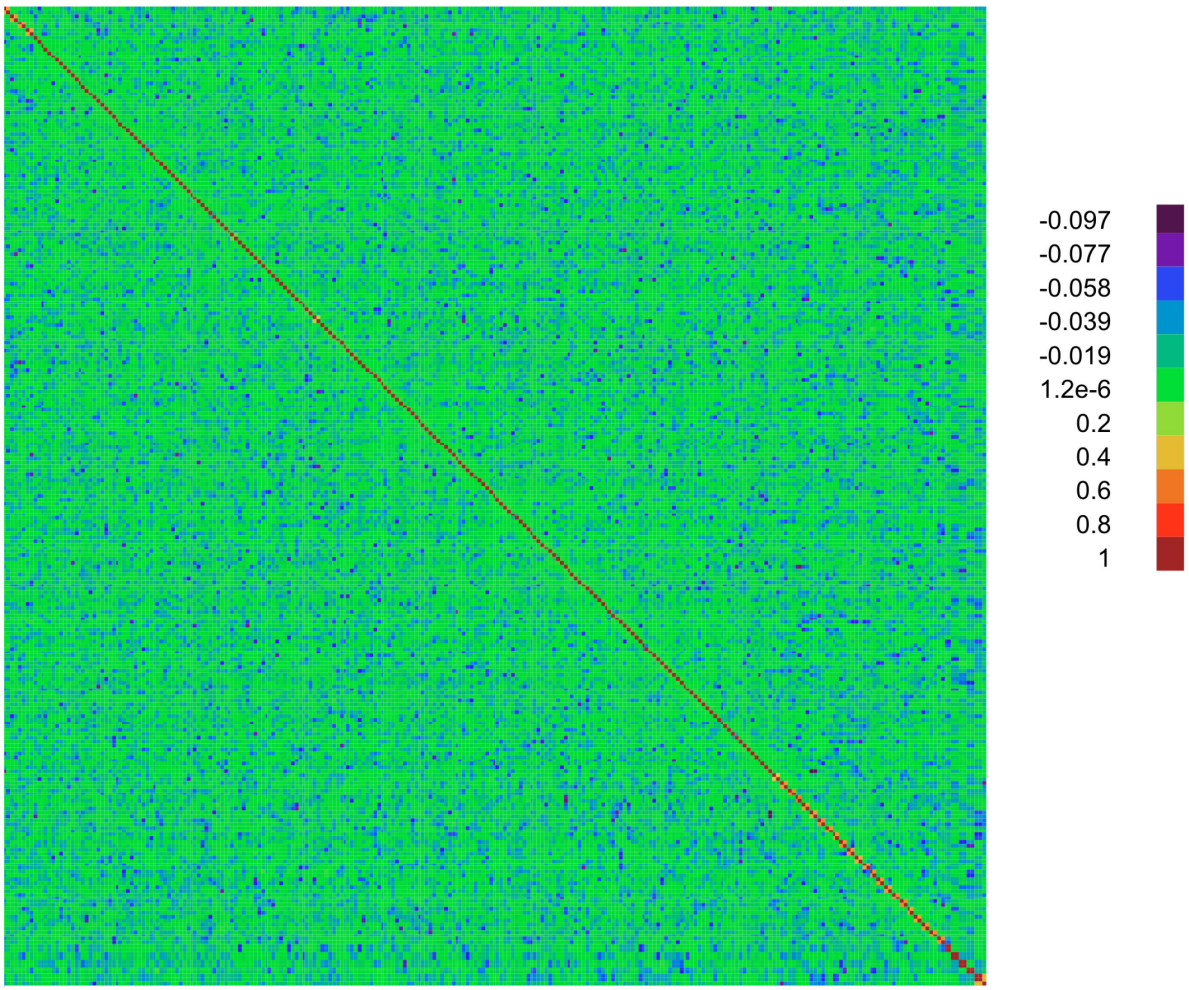


Figure D.1.7 gPCA.

Principal Component Analysis using 518,000 shared genome-wide genotyped SNPs from SCD patients, YRI Yoruban Africans and CEPH European HapMap individuals identifies SCD participants to be ethnically similar to YRI individuals.

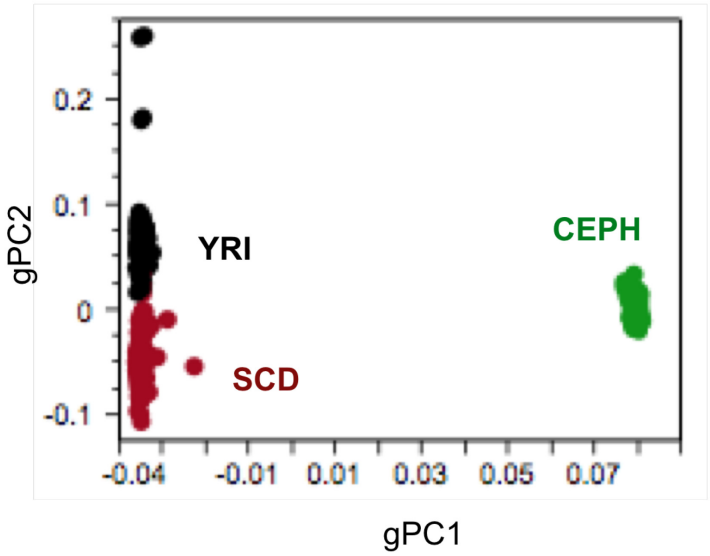
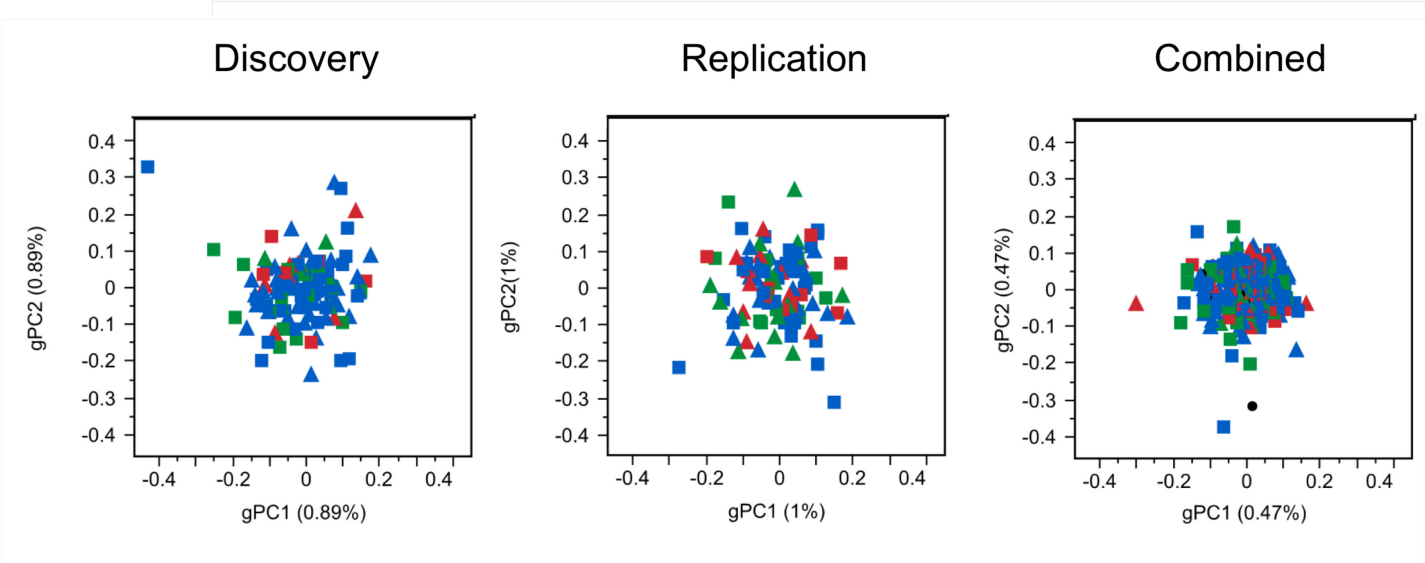


Figure D.1.8 Plot of gPC1-2 of SCD patients.

Ancestry analyses of 119, 104 and 235 unrelated individuals from the discovery phase, replication phase and the combined dataset I using 485,000 genotypes. No obvious population structure was observed with all genotypic principal components (gPCs) explaining 1% or less of the total variance. The plots show the first two gPCs for the discovery, replication, and the combined dataset I. No correlation was observed between gPCs and clinical status, sex (males, squares; females, triangles), or Hb genotype (blue, HbSS; green, HbSC; red, controls).



D.2 GENE EXPRESSION PROFILES AND BIOLOGICAL PATHWAYS IMPLICATED IN SCD

D.2.1 Unsupervised gene expression analysis – discovery phase

Analysis of gene expression shows that SCD has substantial influence on the transcriptome. One-way unsupervised hierarchical clustering analysis of the genome-wide gene expression correlation matrix revealed that individual gene expression profiles cluster largely according to Hb genotype, SCD severity score (SV), and clinical status (ClinStatus; Figure D.2.1). Principal Component Analysis (PCA) revealed the presence of strong correlation structure in the data (Figure D.2.2) such that the first three expression principal components (ePC1-3) explain over a third of the total variance (Table D.2.1 and D.2.2). Variance component analysis of the first three ePCs further confirms the substantial effect of Hb genotype (explaining 45.6% of the variance) followed by clinical status and clinical category (explaining 7% of the variance) (Figure D.2.3 and D.2.4). Variance of ePC1 was explained primarily by Hb genotype (>70%) while ePC2 and 3 were dominated by the effect of clinical status, explaining 20% of the variance of each ePC, or clinical category, explaining 15% and 20% of the variance, respectively. Sex and interaction effects had negligible effects on the variance (Figure D.2.3 and D.2.4). Repeating this analysis with only SCD patients (n=126) revealed that a third of the variance (31%) was captured by the first three ePCs, with Hb genotype and the follow-up effect explaining 19.5% and 8.6% of the variance, respectively (Figure D.2.3, Table D.2.1), or Hb genotype and clinical category explaining 20% and 6% of the variance, respectively (Figure D.2.4, Table D.2.2).

Figure D.2.1 Hierarchical clustering of gene expression data.

One-way hierarchical clustering of the genome-wide gene expression correlation matrix for the combined dataset I (n=311). The heat map shows the clustering of individual expression profiles based on similarity. The highest level of clustering is observed for the Hb genotype (HB) effect followed by SCD severity score (SV) and clinical status (ClinStatus). The color coded heat-map displays the largest expression values in red (hot) and the smallest values in blue (cool). Intermediate values are displayed in different shades of red and blue.

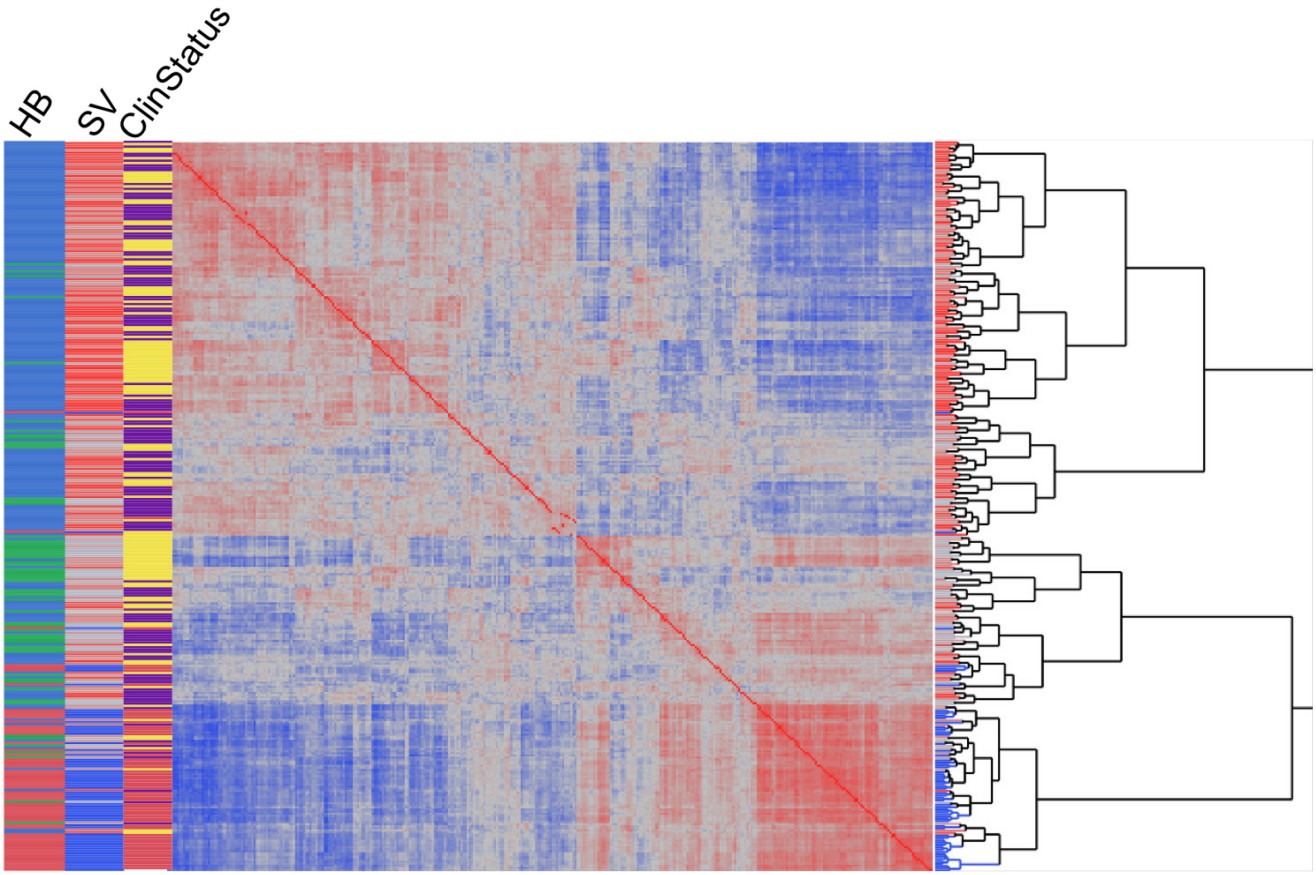


Figure D.2.2 Principal component analysis using gene expression data.

Principal componant analysis using genome-wide gene expression data. The first two expression principal components (ePC) from PC analysis of the discovery and replication phase samples, and in the combined dataset are plotted. Individuals are coloured according to Hb genotype (HbSS, blue; HbSC, green; and Controls, red), SCD severity score (SV, red to blue indicates high to low severity) and Clinical Status effect (FU, yellow; E, purple, Ctls, red).

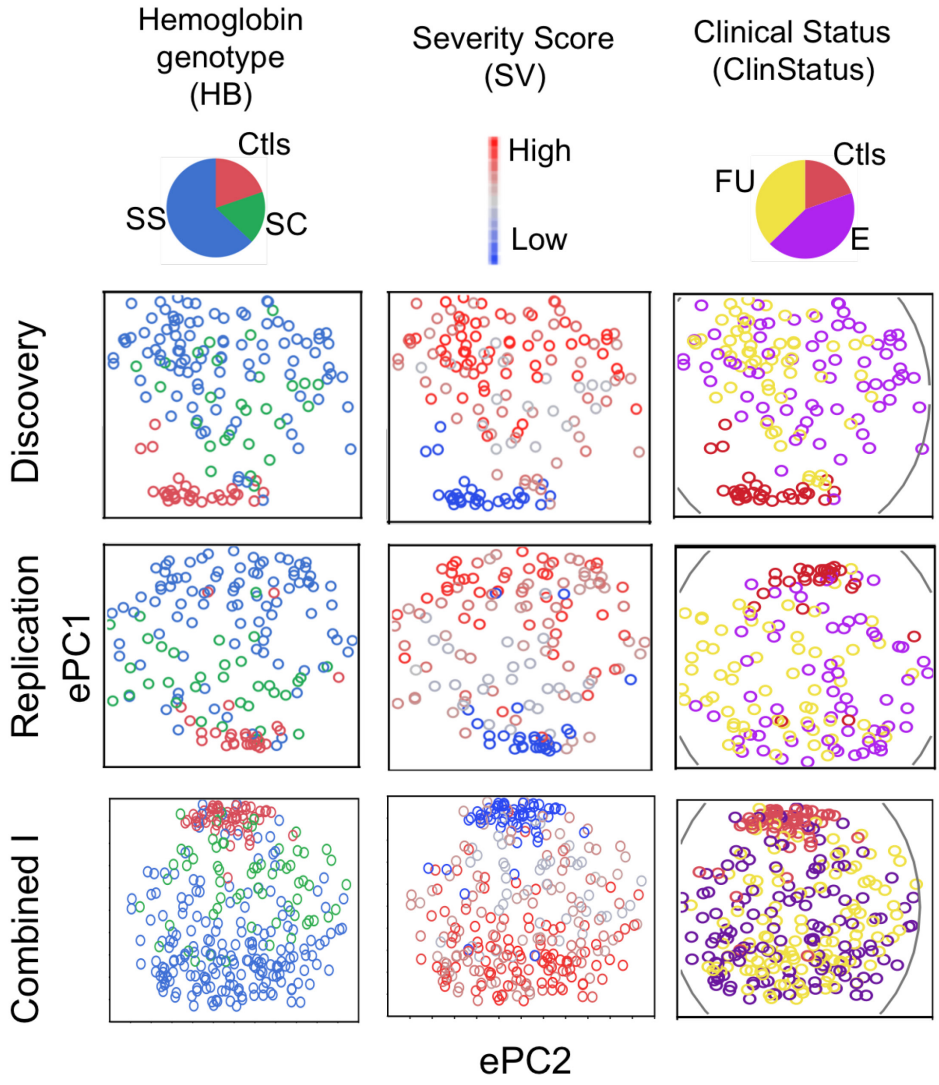


Figure D.2.3 Histogram of the proportion of variance explained by expression principle componenets (ePC) 1 to 3 for each modeled variable (ClinStatus).

Variance component analysis of ePC1-3. The proportion of variance explained (y-axis) by each of the main variables (x-axis) is shown for the variance component analysis (VCA) of the first three expression principal components (ePC1-3). These first 3 ePCs explain 36%, 37% and 37% of the total variance in the discovery, replication, and in the combined dataset (See Appendix VI). The two main variables that explain this variance are Hb genotype and follow-up effect. The proportion of the variance explained by each variable is similar in the discovery, replication and combined datasets. VCA of SCD patients alone shows that the proportion of the variance explained by clinical category was similar to that when the controls were included but the proportion of the variance explained by Hb genotype dropped by 25-50%.

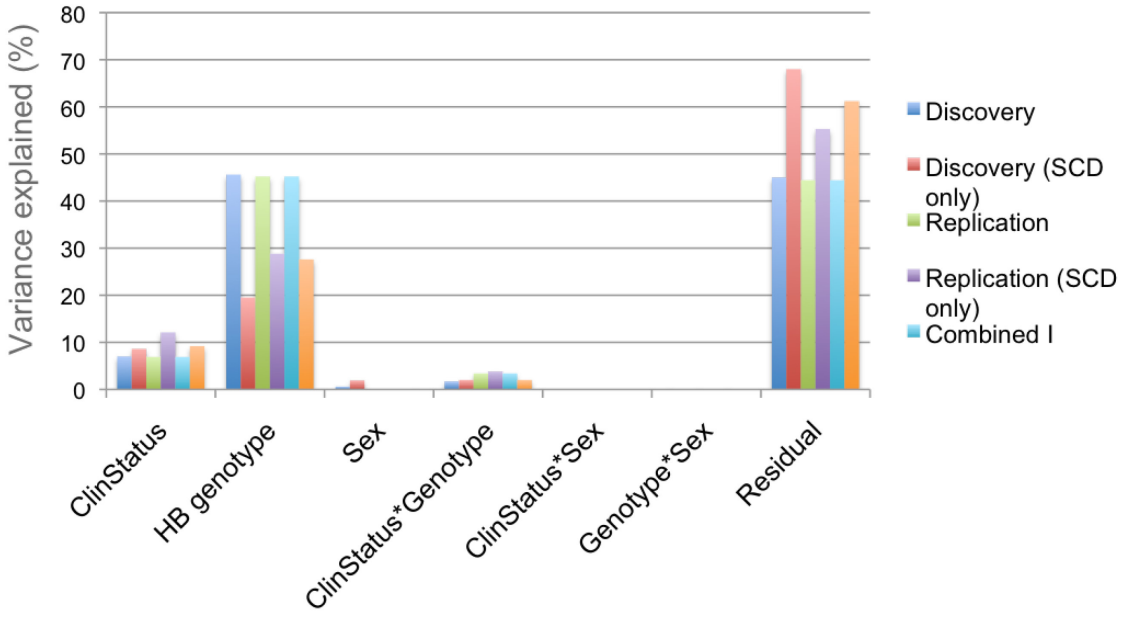


Figure D.2.4 Histogram of the proportion of variance explained by expression principle componenets (ePC) 1 to 3 for each modeled variable (ClinCategory).

Variance component analysis of ePC1-3. The proportion of variance explained (y-axis) by each of the main variables (x-axis) is shown for the variance component analysis (VCA) of the first three expression principal components (ePC1-3). These ePCs explain 36%, 37% and 35% of the total variance in the discovery, replication, and in the combined dataset (see Appendix VII). The two main variables that explain this variance are Hb genotype and clinical category. The proportion of the variance explained by each variable is similar in the discovery, replication and combined datasets. VCA of SCD patients alone shows that the proportion of the variance explained by clinical category was similar to that when the controls were included but the proportion of the variance explained by Hb genotype dropped by 25-50%.

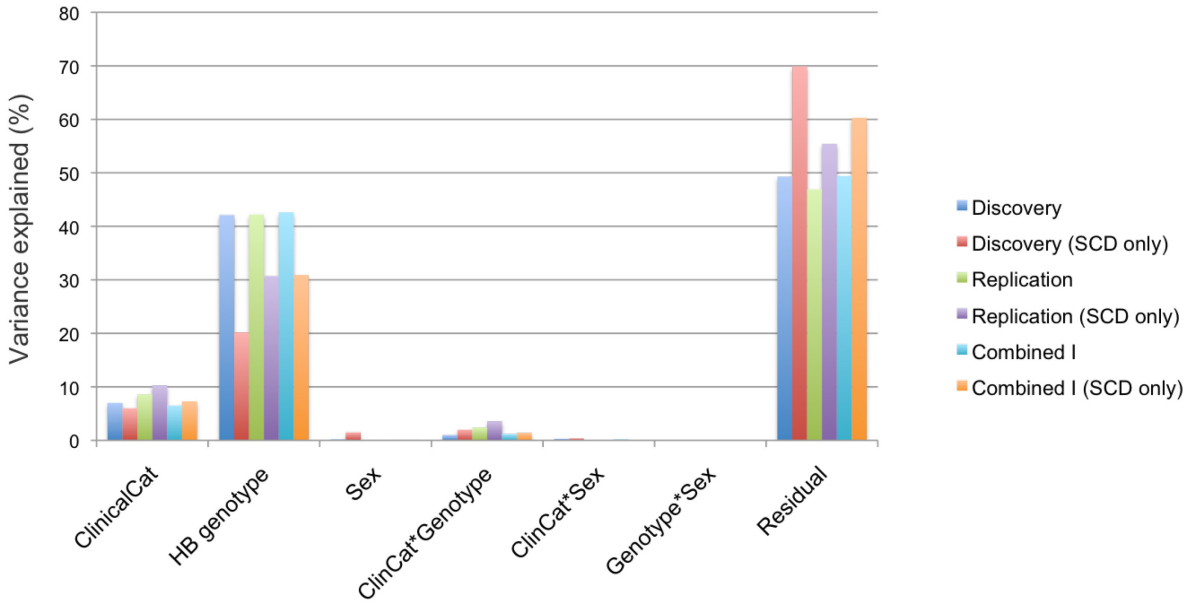


Table D.2.1 Table of the proportion of variance explained by expression principle components (ePC) 1 to 3 for each modelled variable (ClinStatus).

Variance component analysis on the first three expression principal components (ePC1-3) explains over a third of the total variance in the discovery phase, the replication phase, and in the combined data set. This total variance that is explained by the first ePCs is reduced slightly when only SCD patients were included in the analysis. The variance explained by each variable in the model are indicated (ClinStatus: SCD clinical status, Hb genotype, Sex, and (*) interaction effects). The proportion of variance explained by each variable is similar in the discovery, replication and combined data set. Here, the SCD patients are categorised by Clinical Status (ClinStatus).

N	Phase	Variance Explained (ePC1-3)	ClinStatus	HB genotype	Sex	ClinStatus*Genotype	ClinStatus*Sex	Genotype*Sex	Residual
157	Discovery	36	7	45.6	0.6	1.7	0	0.1	45
126	Discovery (SCD only)	31	8.6	19.5	1.9	2	0	0	68
154	Replication	37	6.9	45.2	0	3.4	0	0.1	44.4
124	Replication (SCD only)	32	12.1	28.8	0	3.8	0	0	55.3
311	Combined I	37	6.9	45.2	0	3.4	0	0.1	44.4
250	Combined I (SCD only)	35	9.2	27.6	0.1	2	0	0	61.2

Table D.2.2 Table of the proportion of variance explained by expression principle components (ePC) 1 to 3 for each modelled variable (ClinCat).

Variance component analysis on the first three expression principal components (ePC1-3) explains over a third of the total variance in the discovery phase, the replication phase, and in the combined data set. This total variance that is explained by the first ePCs is reduced slightly when only SCD patients were included in the analysis. The variance explained by each variable in the model are indicated (ClinicalCat: SCD clinical category, Hb genotype, Sex, and (*) interaction effects). The proportion of variance explained by each variable is similar in the discovery, replication and combined data set. Here, the SCD patients are categorised by Clinical Category (ClinicalCat).

N	Phase	Variance Explained (ePC1-3)	ClinicalCat	HB genotype	Sex	ClinCat*Genotype	ClinCat*Sex	Genotype*Sex	Residual
157	Discovery	0.36	7	42.1	0.2	1	0.3	0	49.3
126	Discovery (SCD only)	0.32	6	20.2	1.5	2	0.4	0	69.9
154	Replication	0.37	8.6	42.2	0	2.4	0	0	46.9
124	Replication (SCD only)	0.3	10.3	30.7	0	3.6	0	0	55.4
311	Combined I	0.35	6.5	42.6	0.1	1.2	0.2	0	49.4
250	Combined I (SCD only)	0.31	7.3	30.9	0.1	1.4	0	0	60.3

D.2.2 Supervised gene expression analysis – discovery phase

The magnitude and significance of differentially expressed genes between SCD clinical status, SCD clinical categories and controls were evaluated. Given that a fraction of the variation in expression PCs is likely due to differences in the proportion of cell types between SCD patients, we performed a probe-by-probe analysis of covariance (ANCOVA) of the discovery sample that accounts for total blood cell counts (red and white blood cell counts), in addition to sex, and genetic ethnicity using individuals' scores at significant genotypic PC axes. This analysis revealed significant differences between SCD patients (E and FU) and controls (Ctls) and between clinical categories and controls. A quarter or more of the transcriptome was differentially expressed for the follow-up effect and for the clinical category effect at 1% False Discovery Rate (FDR) (Table D.2.3). Thousands of genes were also significantly differentially expressed between Hb genotypes (HbSS, HbSC, controls) while minor differences were observed between sexes and no effect of the genome-wide genotypic ethnicity effect (gPCs) was detected. Since meaningful population structure in the sample was not observed (Figure D.1.10) and since no probes were significant for the gPC effect (FDR 1%), genetic ancestry is unlikely to significantly contribute to the observed gene expression differences in our sample.

Table D.2.3 ANCOVA

Number of differentially expressed probes for the following effects: SCD Clinical Status (ClinStatus: E=Entry, FU=Follow-up and Ctls=Controls, Acute=A), SCD Clinical Category (ClinCategory: Acute=A. E=Entry, SSS=Steady-state satisfactory, SSU=Steady-state unsatisfactory, Ctls=Controls), Hb genotypes (HbSS, HbSC, Ctls), sexes (M=males, F= females), gPCs (genotypic principal components), and cell counts (RBCs=red blood cell counts, WBCs=white blood cell counts). The 3way-FU effect is between EvsFUvsCtls, and the 6-wasy ClinCategory comparison is between SCD patients: AvsEvsSSSvsSSU. These results were obtained from an analysis of covariance (ANCOVA, FDR 1%) of the discovery, replication and combined datasets I and II and accounts for sex and total blood cell counts (RBC and WBC).

	ANCOVA (FDR1%)			
Effect	**Discovery**	**Replication**	**Combined I**	**Combined II**
ClinStatus				
E vs FU	3189	2821	4733	1453
E vs Ctls	6733	2855	6677	5398
FU vs Ctls	5254	3189	6570	4268
3-way ClinStatus comp	9577	6279	10924	7002
A vs Ctls	2471	2380	4274	
A vs E	1667	2727	4276	
A vs FU	439	659	1647	
E vs Ctls	6168	2827	6279	
FU vs Ctls	4842	2881	6089	
E vs FU	2115	1742	3190	
ClinCategory				
A vs Ctls	2248	2025	4012	
E vs Ctls	5736	2416	5937	
SSS vs Ctls	2870	1329	3733	
SSU vs Ctls	4443	1919	5582	
6-way ClinCat comp	1195	1638	4220	
Hb genotype				
HbSS vs Ctls	4934	2729	6403	
HbSC vs Ctls	3838	1509	4325	
HbSS vs HbSC	690	926	2641	
Sex				
Males vs Females	31	59	113	
gPCs				
gPC1	0			
gPC2	0			
Cell counts				
RBCs	908			
WBCs	0			

D.2.3 Replication of differential gene expression among SCD participants

To evaluate if we could replicate the patterns of differential gene expression observed in the discovery phase, we performed the analyses described above on the replication group (n=154) and the combined datasets (combined dataset I and II). We observed similar results, with high correlation between discovery and replication phase principal component scores of all genes (Figure D.2.5). Unsupervised analysis identified similar clustering by Hb genotype and SCD severity score (SV). Forty five percent (45.2%) and 6.9% of the variance in the first three ePCs was explained by Hb genotype and clinical status (Table D.2.1), respectively. Thirty one percent (31%) and 10% of the variance in the first three ePCs was explained by Hb genotype and clinical category (Table D.2.2), respectively. When only SCD patients were included, Hb genotype and the follow-up effect explained 28.8% and 12.1% of the variance in the first three ePCs, respectively; while Hb genotype and clinical category explained 31% and 10% of the variance in the first three ePCs, respectively (Table D.2.1 and D.2.2). The magnitude and significance of differentially expressed probes for the clinical status, clinical category and Hb genotype effects were highly consistent in both replication and discovery phases (Figure D.2.6 and Figure D.2.7).

Figure D.2.5 Correlation PCA between phases.

Pearson correlation of Principal Components (PC) scores for each gene in the discovery (set 1), and replication phases (set 2). PCA of the gene expression data was performed for the discovery and replication. Principal component scores of all genes for each phase were contrasted. This analysis shows high correlation between discovery and replication phases. The correlations are shown for the first three ePCs; in G) ePC1, H) ePC2, and I) ePC3. Set 1= discovery phase, while set2 = replication phase.

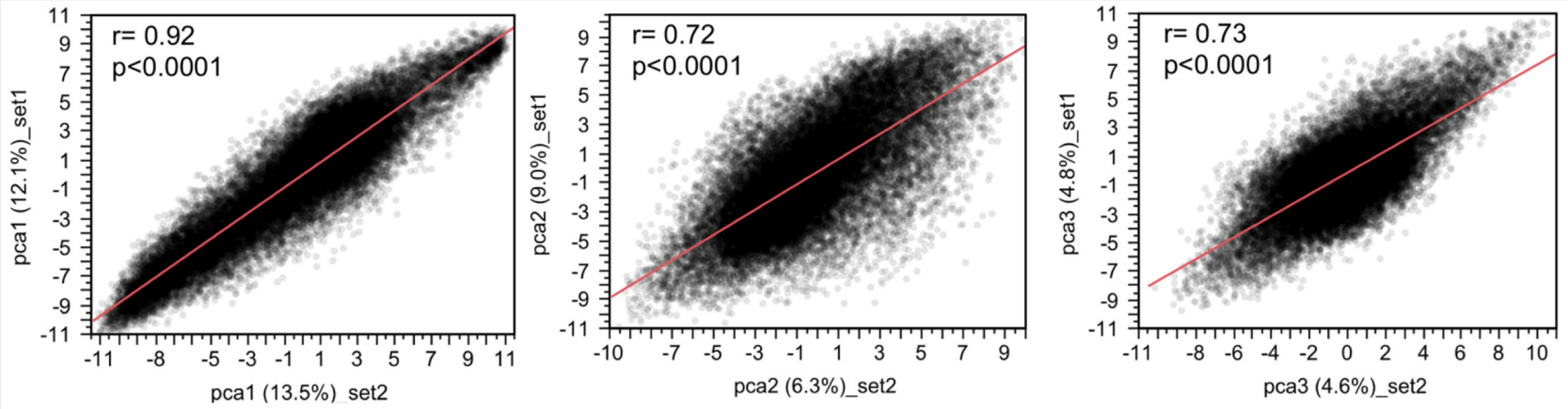


Figure D.2.6. Volcano plots for ClinStauts effect.

Volcano plots of differentially expressed probes from ANCOVA (FDR 1%) in discovery, replication, and combined data sets for the clinical status effect (E vs FU vs Ctls). Volcano plots showing significance (y-axis, -log10 p-value) vs log2 fold change (x-axis) of all probes analyzed in the contrasts between SCD follow-up and controls using the discovery (A-C), replication (D-F), and combined I dataset (G-I). Probes that are differentially expressed at FDR 1% are above the dotted red line. Correlation of gene expression fold change differences in the discovery and replication phases for all probes that are differentially expressed in the contrasts CvsE (J), CvsFU (K), and EvsFU (L) are shown. Correlations between the discovery and replication phases are significant ($p<0.001$).

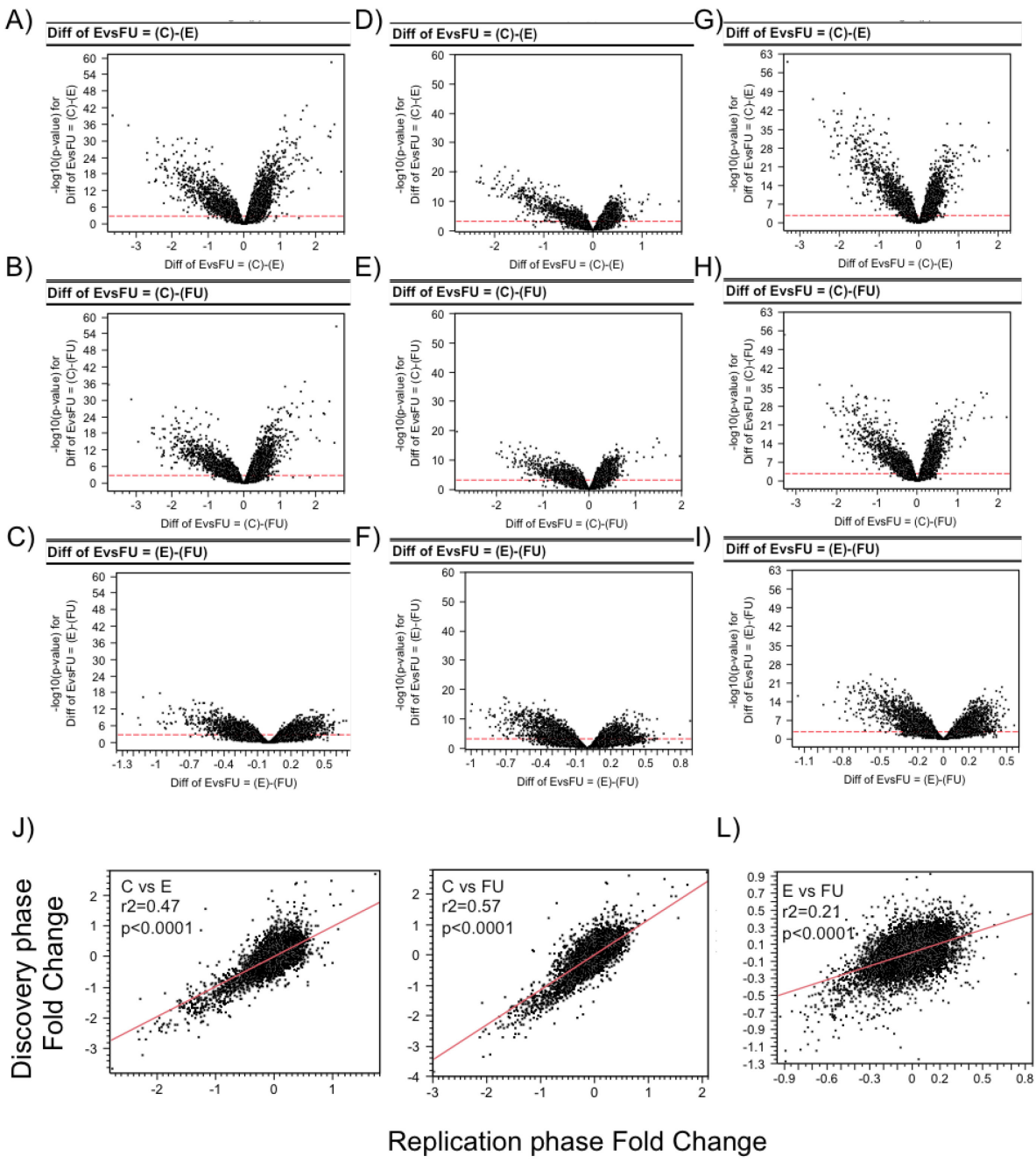
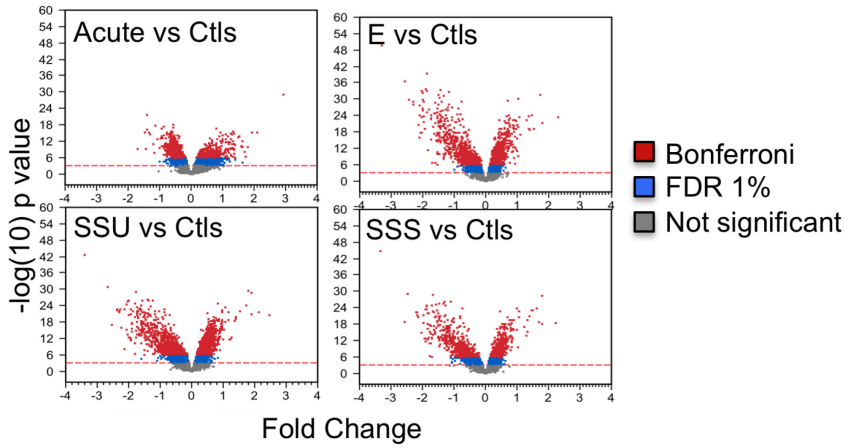

A)
Diff of EvsFU = (C)-(E)
-log10(p-value) for Diff of EvsFU = (C)-(E)
Diff of EvsFU = (C)-(E)
D)
Diff of EvsFU = (C)-(E)
-log10(p-value) for Diff of EvsFU = (C)-(E)
Diff of EvsFU = (C)-(E)
G)
Diff of EvsFU = (C)-(E)
-log10(p-value) for Diff of EvsFU = (C)-(E)
Diff of EvsFU = (C)-(E)
B)
Diff of EvsFU = (C)-(FU)
-log10(p-value) for Diff of EvsFU = (C)-(FU)
Diff of EvsFU = (C)-(FU)
E)
Diff of EvsFU = (C)-(FU)
-log10(p-value) for Diff of EvsFU = (C)-(FU)
Diff of EvsFU = (C)-(FU)
H)
Diff of EvsFU = (C)-(FU)
-log10(p-value) for Diff of EvsFU = (C)-(FU)
Diff of EvsFU = (C)-(FU)
C)
Diff of EvsFU = (E)-(FU)
-log10(p-value) for Diff of EvsFU = (E)-(FU)
Diff of EvsFU = (E)-(FU)
F)
Diff of EvsFU = (E)-(FU)
-log10(p-value) for Diff of EvsFU = (E)-(FU)
Diff of EvsFU = (E)-(FU)
I)
Diff of EvsFU = (E)-(FU)
-log10(p-value) for Diff of EvsFU = (E)-(FU)
Diff of EvsFU = (E)-(FU)
J)
L)
Discovery phase Fold Change
C vs E
r2=0.47
p<0.0001
C vs FU
r2=0.57
p<0.0001
E vs FU
r2=0.21
p<0.0001
Replication phase Fold Change

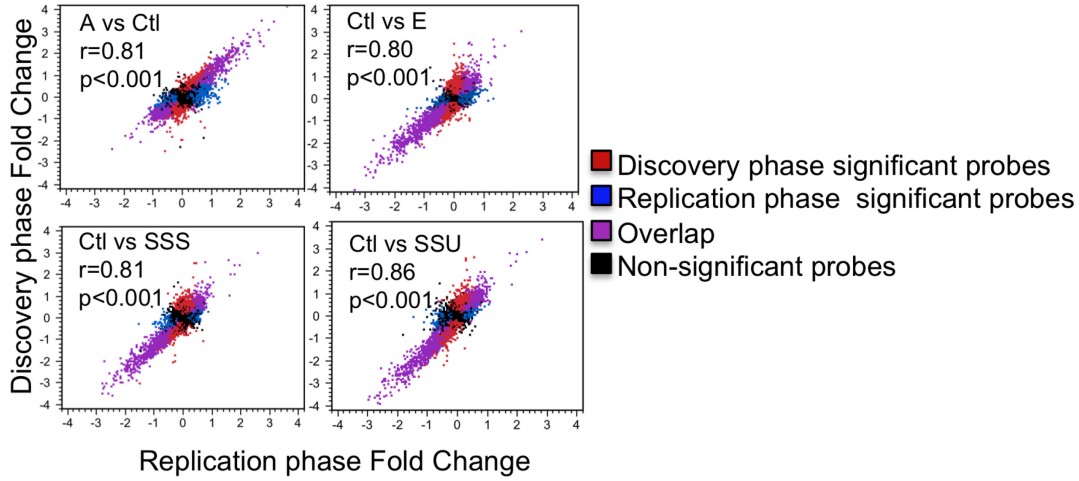
Figure D.2.7. Volcano plots for ClinCat effect.

A) Volcano plots showing significance (y-axis, -log10 p-value) vs log2 fold change (x-axis) of all probes analyzed in the contrasts between SCD clinical categories and controls using the combined dataset. Probes that are differentially expressed at Bonferroni significance ($p < 1.75 \times 10^{-6}$) and at FDR 1% are colored in red and blue, respectively. B) Correlation between fold change differences in the discovery and replication phases for the set of differentially expressed probes between each of the SCD clinical categories (A, E, SSS, SSU) and the controls (Ctls). In red are probes that are differentially expressed only in the discovery phase; in blue are probes that are differentially expressed only in the replication phase; in purple are probes that are differentially expressed for both phases; in black are probes that are not differentially expressed in either phase. All correlations between the discovery and replication phases are significant ($p<0.001$).

A)


Acute vs Ctls
E vs Ctls
SSU vs Ctls
SSS vs Ctls
-log(10) p value
Fold Change
Bonferroni
FDR 1%
Not significant

B)

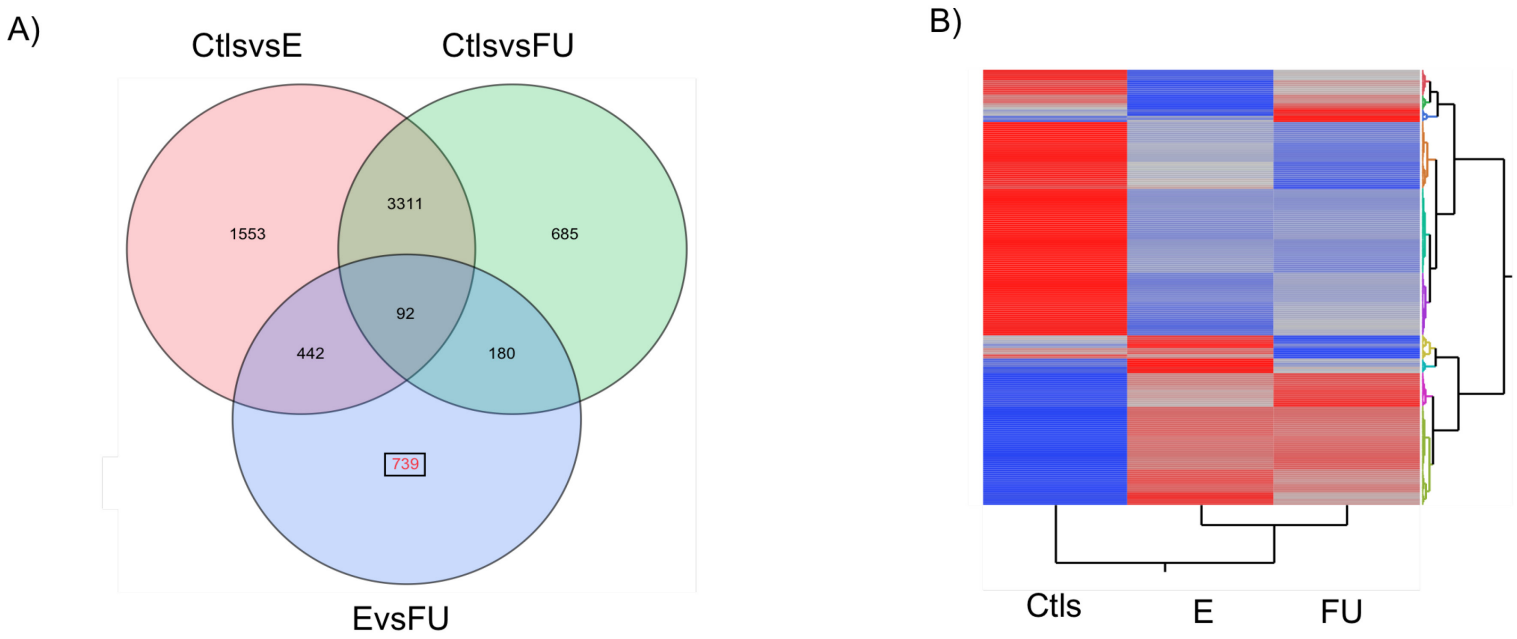

A vs Ctl
r=0.81
p<0.001
Ctl vs E
r=0.80
p<0.001
Ctl vs SSS
r=0.81
p<0.001
Ctl vs SSU
r=0.86
p<0.001
Discovery phase Fold Change
Replication phase Fold Change
Discovery phase significant probes
Replication phase significant probes
Overlap
Non-significant probes

D.2.3.1 Gene expression signatures associated with SCD clinical status

Next, we focused on 160 SCD HbSS patients and 56 controls (combined data set II, n=216) to characterize the transcriptional signatures associated with SCD clinical status and follow-up considering only the HbSS group and the controls. HbSC individuals were excluded from this analysis given their small sample size relative to the HbSS group. SCD patients undergoing an acute event were also excluded to focus on the steady state of the disease. An ANCOVA (FDR 1%) of this dataset accounting for sex and cell counts revealed that over seven thousand probes were significantly differentially expressed (1% FDR) for the clinical status effect (EvsFUvsCtls) and 739 probes were differentially expressed between SCD patients for the follow-up effect (EvsFU; Table D.2.3). The effect of clinical status is visually shown in a heat map generated using a 2-way hierarchical clustering of per-group mean expression levels of differentially expressed probes (Figure D.2.8).

Figure D.2.8 Gene expression signatures associated with ClinStatus.

A) Venn diagram of the 7002 differentially expressed probes for the 3-way clinical status effect in the combined data set II. 739 probes are uniquely differentially expressed for the follow-up effect (EvsFU) for SCD patients. B) Two-way hierarchical clustering of the mean expression levels for the 7002 differentially expressed probes in the combined data set II for each group of patients (E, FU, Ctls) is shown. Mean expression from this class of genes cluster controls from SCD entry and follow-up patients.



D.2.4 Identification of biologically relevant pathways in SCD

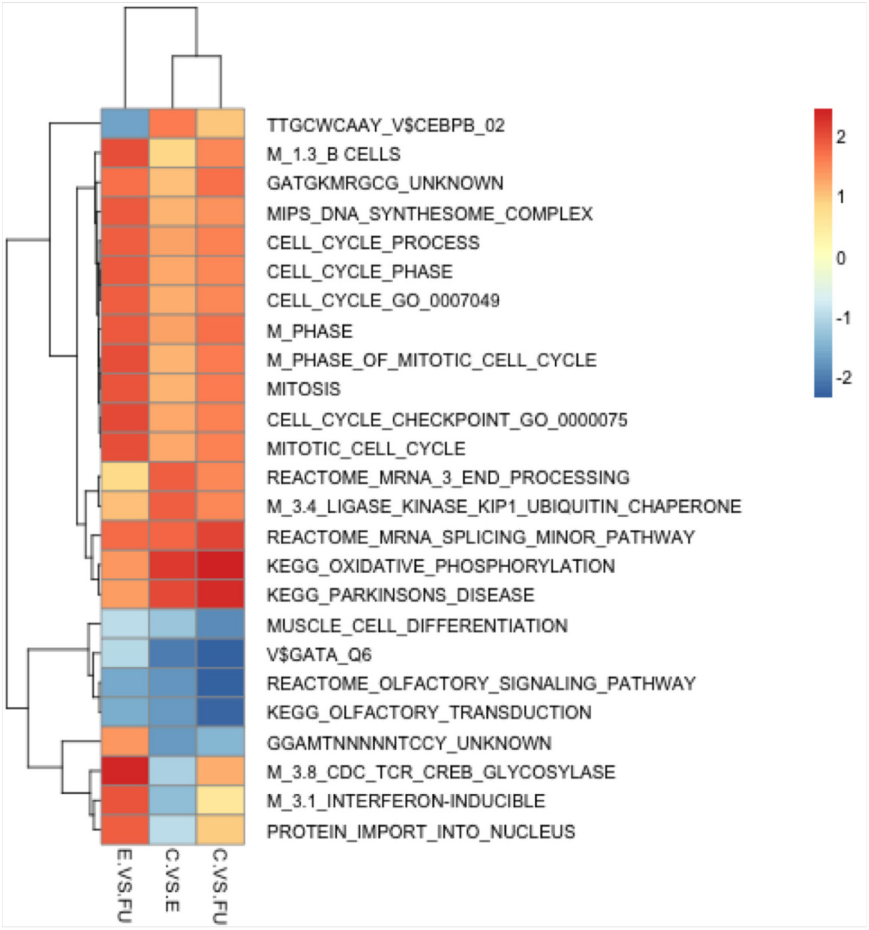
In order to identify biological pathways subject to the effects of differential expression associated with SCD clinical status and clinical categories, Gene Set Enrichment Analysis (GSEA) [137] was performed using the differentially expressed genes in the discovery, replication and combined datasets (I, II).

In order to identify sets of genes that differentiated SCD patients who took part in the follow-up program from SCD patients at entry and from controls, GSEA for the clinical status effect was performed using combined data set II, where we focused on the follow-up effect in HbSS patients. The results are shown in Figure D.2.9. A strong activation of biological pathways previously reported to be associated with SCD [43], pathways associated with lymphocyte development, stress (glucocorticoid and hypoxia related therapies), and uncontrolled cell growth is observed for the CvsE and CvsFU contrasts. Many of these pathways are also activated in the clinical course of SCD follow-up (EvsFU contrast).

The 739 probes that uniquely differentiate the E from FU patients (the follow-up effect; Figure D.2.8) were tested for significant overlap with genes belonging to pathways in the GSEA. A significant overlap was found in genes that were up-regulated in B-lymphocytes expressing phosphotylated CD5 (Gary_CD5_Targets_UP). This pathway most likely also plays a role in the clinical course of SCD follow-up.

Figure D.2.9 GSEA for ClinStatus effect.

Gene Set Enrichment Analysis (GSEA) was performed for each contrast clinical status effect using the combined dataset II. This analysis identified biological pathways and sets of individual genes that are significantly enriched (Benjamini-Hochberg-adjusted P < 0.05) in each contrast. Selection of the most distinctive significantly enriched pathways between entry and follow-up groups is shown. Cells are coloured by their respective Normalized Enrichment Scores for a given contrast.



The GSEA results for the clinical category effect (Figure D.2.10) show a strong inflammatory response signature in the Acute patients driven by an enrichment in myeloid lineage, complement and coagulation cascade, interferon-inducible and inflammation pathways. The signature observed in SSU patients highlights a strong activation of platelets. Pathways associated with B- and T-cell stimulation, as well as metabolism-related pathways are upregulated in E and SSS patients.

GSEA results for the discovery and replication phases are shown in Appendix VIII. The GSEA results in the discovery and replication phases are correlated (Figure D.2.11) and are similar to those in the combined data sets.

Figure D.2.10 GSEA for ClinCat effect.

Gene Set Enrichment Analysis (GSEA) was performed for each clinical category versus controls contrast using the combined dataset I. This analysis identified biological pathways and sets of individual genes that are significantly enriched in each contrast. Only pathways and modules significantly enriched (Benjamini-Hochberg-adjusted P < 0.05) from at least one contrast are shown. Colours in the heat map indicate the enrichment score relative to the controls.

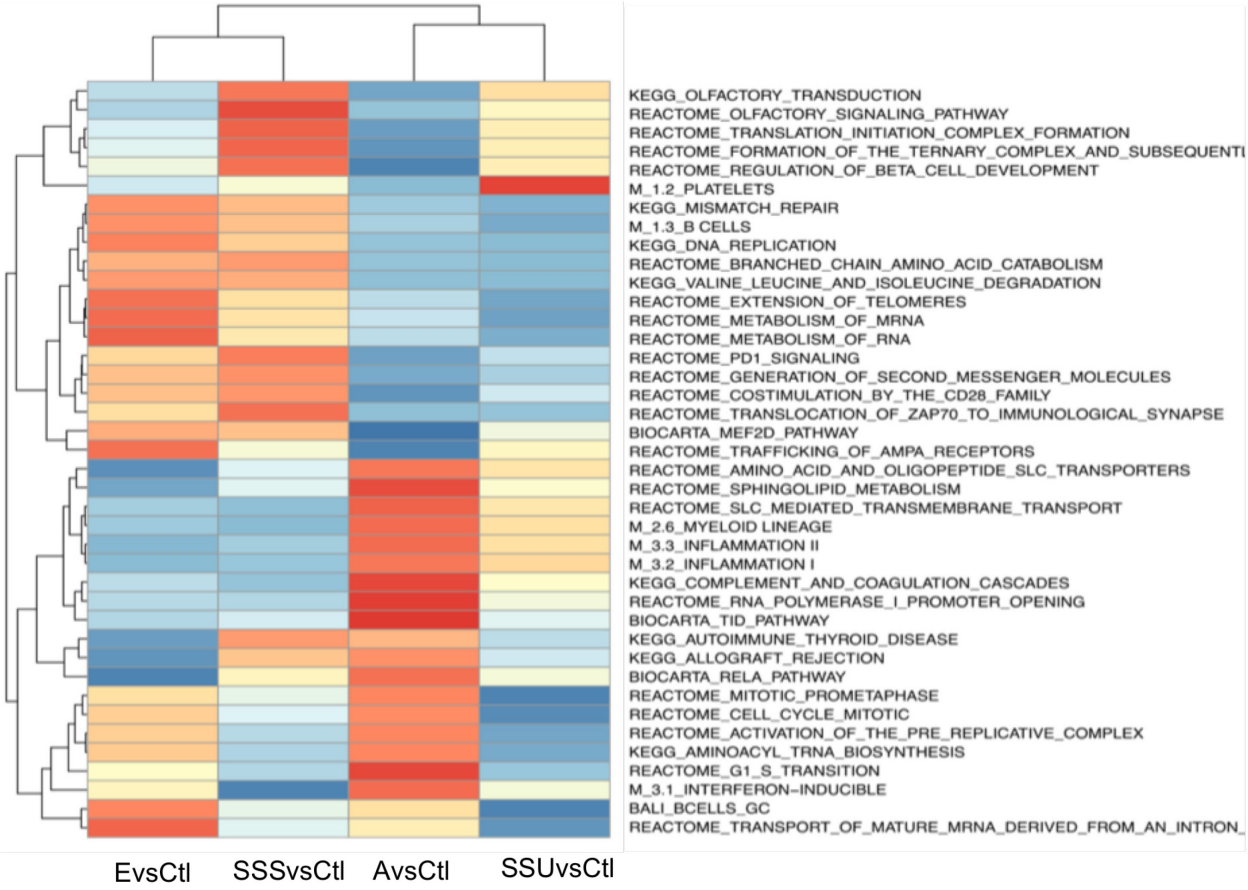
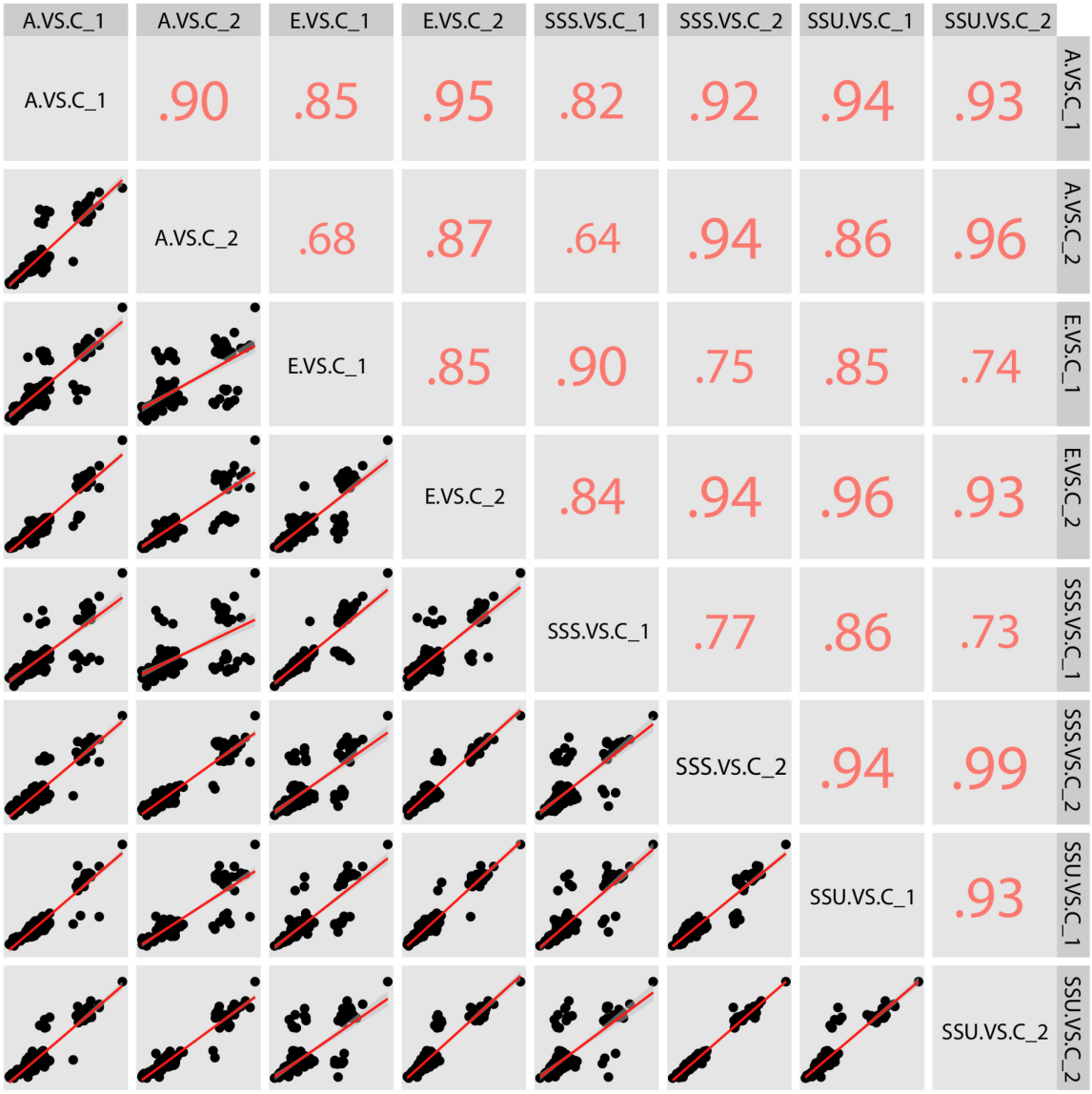


Figure D.2.11. Correlation of GSEA between phases.

Spearman correlations of NES for significant GSEA pathways (FDR =0.05) for the discovery and replication phases in at least one clinical category contrast. Gene set enrichment analysis (GSEA) was performed for contrasts of differentially expressed probes between SCD clinical categories (A, E, SSS, SSU) versus controls (C) for the discovery (phase 1) and replication (phase 2) phases. All pairwise clinical category comparisons from both phases are contrasted to each other in the following way: for each contrast, correlation of normalized enrichment scores (NES) of the genes from the pathways that are enriched in at least one comparison is tested. These plots show a high degree of correlation observed for these contrasts with the majority of the genes showing the trend of direction in gene expression differentiation.


A.VS.C_1
A.VS.C_2
E.VS.C_1
E.VS.C_2
SSS.VS.C_1
SSS.VS.C_2
SSU.VS.C_1
SSU.VS.C_2
A.VS.C_1
.90
.85
.95
.82
.92
.94
.93
A.VS.C_2
.68
.87
.64
.94
.86
.96
E.VS.C_1
.85
.90
.75
.85
.74
E.VS.C_2
.84
.94
.96
.93
SSS.VS.C_1
.77
.86
.73
SSS.VS.C_2
.94
.99
SSU.VS.C_1
.93
SSU.VS.C_2

D.3. TRANSCRIPTIONAL BIOMARKERS OF SCD CLINICAL SEVERITY

D.3.1 Identification of transcriptional biomarkers of SCD clinical categories

Patients with SCD suffer a wide variety of disease complications [10]. The incidence of most clinical complications in SCD varies markedly both with time in the same individual and between different individuals. Meaningful biomarkers for SCD would be useful in the management of SCD complications. Recently, blood transcriptional profiling has been successfully used to predict disease pathogenesis in tuberculosis, infections, and tumour progression [58,59,60]. Through the identification of aberrant gene expression, it is possible to identify individuals who are susceptible to disease and to predict their outcome.

To identify transcriptional biomarkers for clinical categories, discriminant analysis was performed using the full gene expression dataset (28,595 probes). Individuals in the discovery phase were used as a "training set". This analysis revealed 14 probes that differentiate clinical category with 97% accuracy ($p<0.0001$, Figure D.3.1). Since these 14 genes discriminated SSU patients with only 89% accuracy and since the canonical values for each SSS and SSU patients overlapped (Appendix IX), an additional discriminant analysis was performed on a sub-set of the discovery phase that included only SSS and SSU patients (n=37). Five additional probes that differentiated these two clinical categories with 100% accuracy were included (Figure D.3.1). Using the 5 additional probes and the original 14 probes made a final set of 19 probes to be tested in the replication phase individuals ("test set") for accuracy in predicting clinical category. Overall, these 19 biomarkers predicted clinical category with 80.1% accuracy ($p<0.0001$, Figure D.3.1) in the replication phase. "Leave-one-out" cross validation of the 19 biomarkers was performed using combined dataset I (n=311). Since a quantitative outcome variable is required, we used the severity score of each participant. In Figure D.3.2, the linear regression of this prediction is

shown to be significant ($p<0.0001$) and explained 70% of the variance. The mean and standard deviation of each clinical category's severity score is also shown in Figure D.3.2.

Figure D.3.1 Transcriptional biomarkers for SCD clinical categories.

A) Discriminant analysis performed on individuals in the discovery phase using 28,595 expressed probes identified 14 probes that distinguish clinical categories with 97% accuracy ($p<0.0001$). An additional discriminant analysis including only SSS and SSU SCD patients (n=37) from the discovery phase identified 5 extra probes that distinguished between SSS and SSU clinical categories with 100% accuracy. Using these 19 probes, we were able to predict clinical category in the replication phase individuals with 80% accuracy ($p<0.0001$). B) Two way hierarchical clustering analysis of the 311 SCD patients and controls using the nineteen biomarkers clustered individuals by clinical category.

a)

Data set	Discovery Phase	Sub-analysis (SSS +SSU)	Replication Phase
Category	Accuracy	Accuracy	Accuracy
Acute	0.944	n/a	0.722
SSU	0.889	1.00	0.967
SSS	1.000	1.00	0.875
E	1.000	n/a	0.591
Ctls	1.000	n/a	0.850
Total	0.967	1.00	0.801
Likelihood Ratio	p<0.0001	p<0.0001	p<0.0001

b)

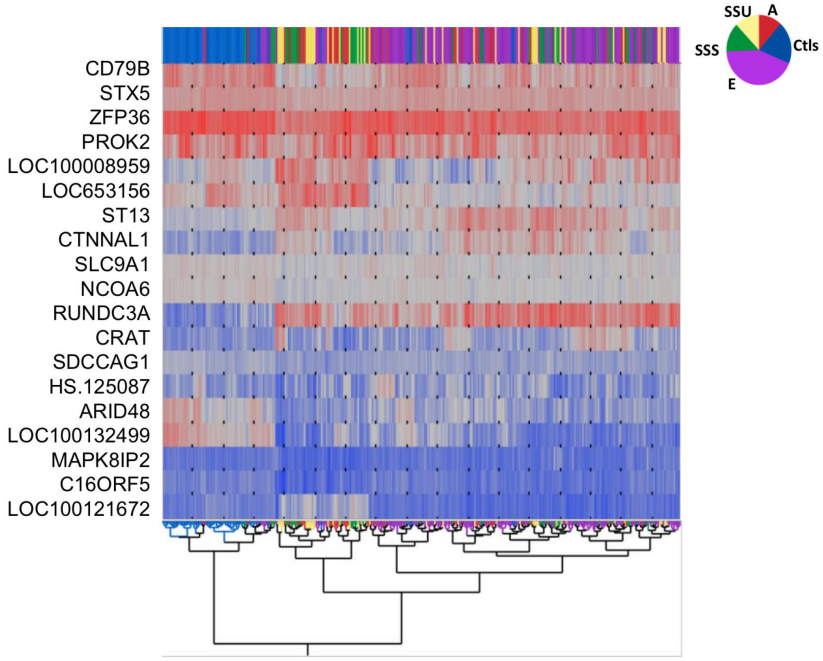
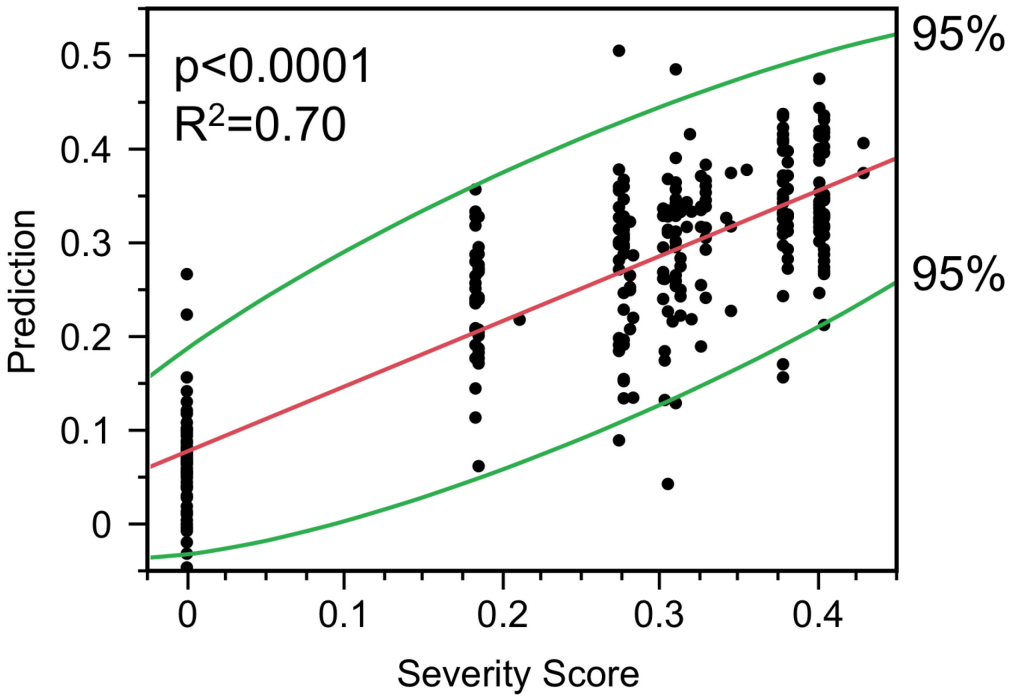


Figure D.3.2 Leave-one-out cross validation for 19 biomarkers

A) Leave-one-out cross validation was performed in the combined dataset I using gene expression levels for the 19 transcriptional biomarkers to predict the severity score for each patient. A plot of the 311 individuals' predicted (y-axis) vs true (x-axis) severity score values is shown. The linear regression was estimated (red line) and shown to be significant ($p<0.0001$) and explains 70% of the variation. The 95% intervals are shown in green. B) The mean and standard deviation values for the severity score for each clinical category show a progressively higher (worse) severity score for Entry, SSS, SSU, and A SCD patients.

A)



B)

Category	N	Mean Severity Score	Std. Dev
C	61	0.00	0.00
E	134	0.31	0.07
SSS	42	0.30	0.09
SSU	38	0.33	0.08
A	36	0.35	0.06

D.3.2 Identification of transcriptional biomarkers for SCD patient progression

Since we are ultimately interested in identifying patients who are unsatisfactory even after follow-up, an additional discriminant anaysis was performed including only SSS and SSU SCD patients. This resulted in 38 SCD patients in the training set (20 SSS patients and 18 SSU patients), and 42 SCD patients in the test set (22 SSS patients and 20 SSU patients). A discriminant analysis on the training set, using the entire gene expression data (28,628 expressed probes), identified 3 probes for genes GCAT, CD79A, NUCKS1 that discriminated SSS from SSU patients with 100% accuracy (Figure D.3.3 below). These probes were 74% accurate in discriminating patients in the test set (Figure D.3.3). Leave-one-out cross validation was performed to predict each patients severity score based on the 3 gene expression levels. The linear regression was significant (p=0.0001), but explained a small proportion of the variance (R^2=0.17) (Figure D.3.4).

Figure D.3.3 Discriminant analysis for SSS and SSU patients.

Discriminant analysis performed on training set (n=38) identified 3 probes that discriminate between SCD patients that were in SSS or SSU condition with 100% accuracy. These 3 probes discriminated SSS from SSU patients with 74% accuracy in the test set (n=42). The receiver operating curve (ROC), illustrates that the results are perfect in the training set and above the diagonal in the test set, which represents good classification (better than random).

	Training Set			Test Set		
Category	n	Accuracy (%)	ROC	n	Accuracy (%)	ROC
SSS	20	100		22	86	
SSU	18	100		20	60	
Overall	38	100	1	42	74	0.78

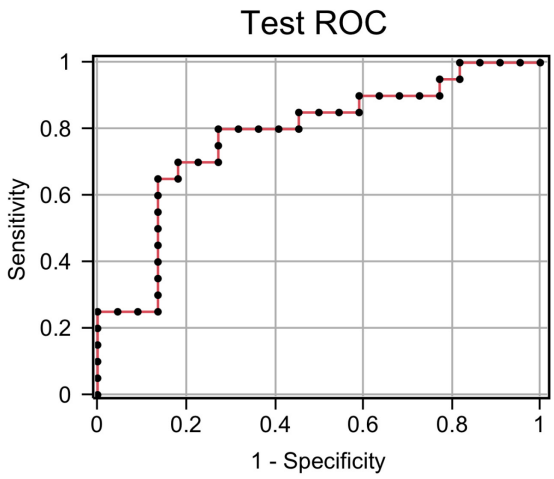
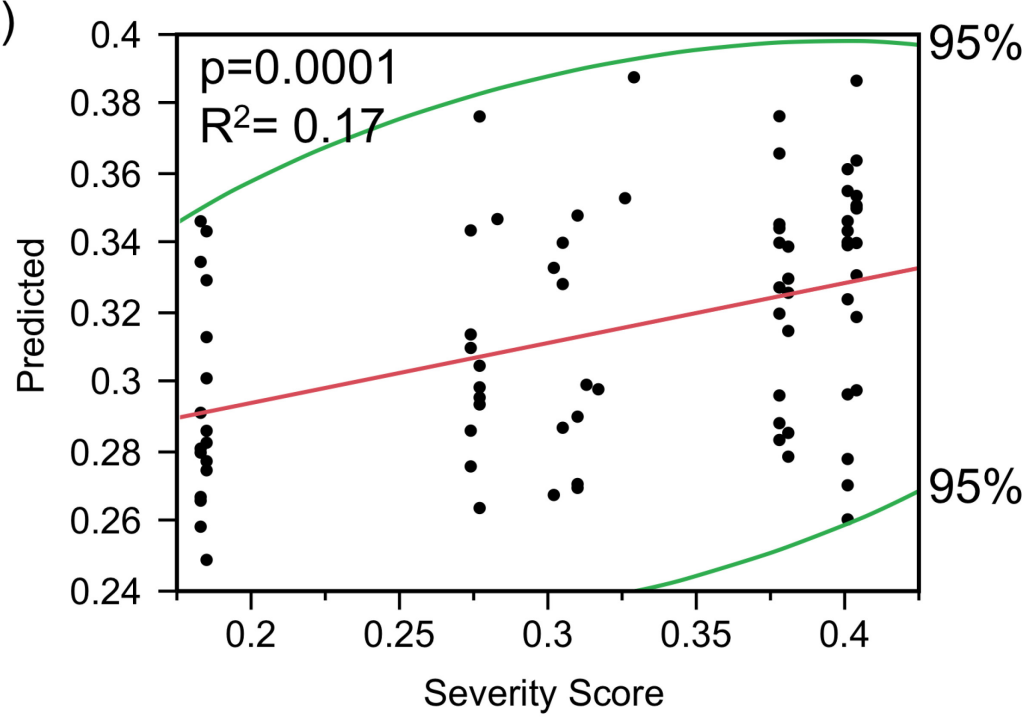


Figure D.3.4 Leave-one-out cross validation using 3 biomarkers.

Leave-one-out cross validation using the 3 probe gene expression level's predicted the severity score for the 42 SCD patients. The actual severity score (x axis) is plotted against the predicted severity score (y axis) and a regression line was fit (p=0.0001), although the variance explained is negligible (r^2= 0.17). The mean severity score for the SSS and SSU patients was calculated.

A)



B)

Category	N	Mean Severity Score	Std Dev
SSS	42	0.30	0.09
SSU	38	0.33	0.08

D.4. THE GENETIC REGULATION OF GENE EXPRESSION VARIATION IN SCD PATIENTS

D.4.1 The genetic architecture of transcript abundance in SCD

The genetic architecture of transcript abundance in SCD was investigated through genome-wide association analysis of gene expression traits in SCD patients and controls. Two models tested for association between SNP genotype and gene expression traits: model 1 accounted for SCD Clinical Status and model 2 accounted for SCD Clinical Category. Both models also accounted for blood cell counts (RBC and WBC), and sex (see Table D.4.1 and the Methods).

Given the high degrees of correlation in the results of gene expression analyses for the discovery and replication phases, and to increase mapping power, we performed the analysis by combining the datasets. Subsets of the combined datasets, for which both gene expression and genotypic data were available, were used. Model 1 (that accounted for Clinical status) used a subset of combined data set II and had a sample size of n=173 (120 not acute HbSS SCD patients and 53 controls). Model 2 (that accounted for Clinical category) used a subset of combined data set I and had a sample size of n=263 (205 SCD patients and 58 controls).

The expression data for the combined datasets were re-normalized in order to minimize potential batch effects. This resulted in a final set of 19,431 probes tested in model 1 and 18,890 probes tested in model 2. Marker properties were also re-calculated for the combined data sets and quality control checks were applied. After filtering by marker properties, 560,675 SNP genotypes were used in model 1 and 568,921 SNP genotypes in model 2. Multiple regression analyses were run and Bonferroni correction for multiple testing was applied.

Since 560,675 SNPs were tested for association with 19,431 probes in model 1, a genome-wide Bonferroni threshold for distal-associations corresponds to

0.05/(19,431 probes x 560,675 SNP) = 4.59 × 10^{-12}. For local associations, a Bonferroni threshold corresponds to 0.05/(19,431 probes x 200 SNPs) = 1.28 x 10^{-08} considering an average number of 200 SNPs tested against each probe.

Since 568,921 SNPs were tested for association with 18,890 probes in model 2, a genome-wide Bonferroni threshold for distal-associations corresponds to 0.05/(18,890 probes x 568,921 SNP) = 4.65 × 10^{-12}. For local associations, a Bonferroni threshold corresponds to 0.05/(18,890 probes x 200 SNPs) = 1.32 x 10^{-08} considering an average number of 200 SNPs tested against each probe.

Table D.4.1 eSNP models and results.

Description of eSNP models and their results. For each model, different data sets were included, as well as different variables: (SNP, ClinStatus = SCD Clinical Status, Blood counts (RBC, WBC), Sex, ClinCat=SCD Clinical Category), and interaction effects). The thresholds needed to reach a significant p value for local and distal eSNP associations were calculated using a Bonferonni threshold. The final number of significant peak associations (Assoc) are indicated for each model, with the corresponding number of local and distal eSNP associations.

Model	Data set	Variables	n	SNPs	Probes	pval local	pval distal	Assoc	local	distal
Model 1	Combined dataset II	SNP, ClinStatus, Blood counts, Sex	173	560,675	19,431	1.28×10^{-08}	4.59×10^{-12}	390	371	19
Model 2	Combined dataset I	SNP, ClinCat, Blood counts, Sex	263	568,921	18,890	1.32×10^{-08}	4.65×10^{-12}	581	579	9
Model 3	Combined dataset II	SNP, ClinStatus, Blood counts, Sex, SNP-by-ClinStatus	173	455,750	7002	3.57×10^{-08}	1.27×10^{-11}	8	8	0
Model 4	Combined dataset I	SNP, ClinCat, Blood counts, Sex, SNP-by-ClinCat	263	399,821	4220	5.92×10^{-08}	2.96×10^{-11}	11	9	2

***** For each eSNP model (Model), a specific data set (Data set), number of individuals (n), and variables (Variables) were included. The number of SNPs and probes that were tested in each of the corresponding models are indicated, with the corresponding Bonferroni corrected p-value thresholds for local (pval local) and distal (pval distal) associations. Based on these thresholds, the final number of significant associations are indicated (Assoc), broken down by the number of local and distal associations.

Model 1 identified three hundred and ninety genome-wide significant peak SNP-probe (eSNP) associations. This corresponded to 371 local and 19 distal effects. See Figure D.4.1 for an illustration of Model 1 eSNP associations.

Model 2 identified five hundred and eighty-eight genome-wide significant peak SNP-probe associations, corresponding to 579 local and 9 distal effects. See Figure D.4.2 for an illustration of Model 2 eSNP associations.

Seventy five percent of the local associations implicate SNPs located within 1 Mb from the probe coordinates. These associations explain on average 20% of the variance of transcript abundance (Figure D.4.3).

Of the genes that are differentially regulated among SCD clinical severities, many are also under genetic control. One hundred of the 390 significant eSNP associations in model 1 implicated genes that are differentially expressed for the 3-way SCD Clinical status effect. Eighty-nine out of the 588 significant eSNP associations in model 2 implicated genes that are differentially expressed for the 6-way SCD-clinical category effect.

Figure D.4.1 Model 1 eSNP results.

a) Manhattan plot of 390 peak eSNP associations identified in Model 1. Each association is plotted by SNP chromosomal location on x-axis and by significance (negative –log10 p-values (NLP)) on the y-axis. The colour code refers to the SNP chromosomal location. Local associations are circled and were significant at a NLP value of 7.89 or more. Distal associations are in triangles and were significant at a NLP of 11.34 or more.

b) Plot of eSNP associations based on their SNP (x-axis) and probe (y-axis) chromosomal location. Associations along the diagonal are local (the SNP- probe pair are located on same chromosome) and those off the diagonal are distal (the SNP-probe pair are located on different chromosomes). Out of the 390 peak associations, 371 are local, and 19 are distal.

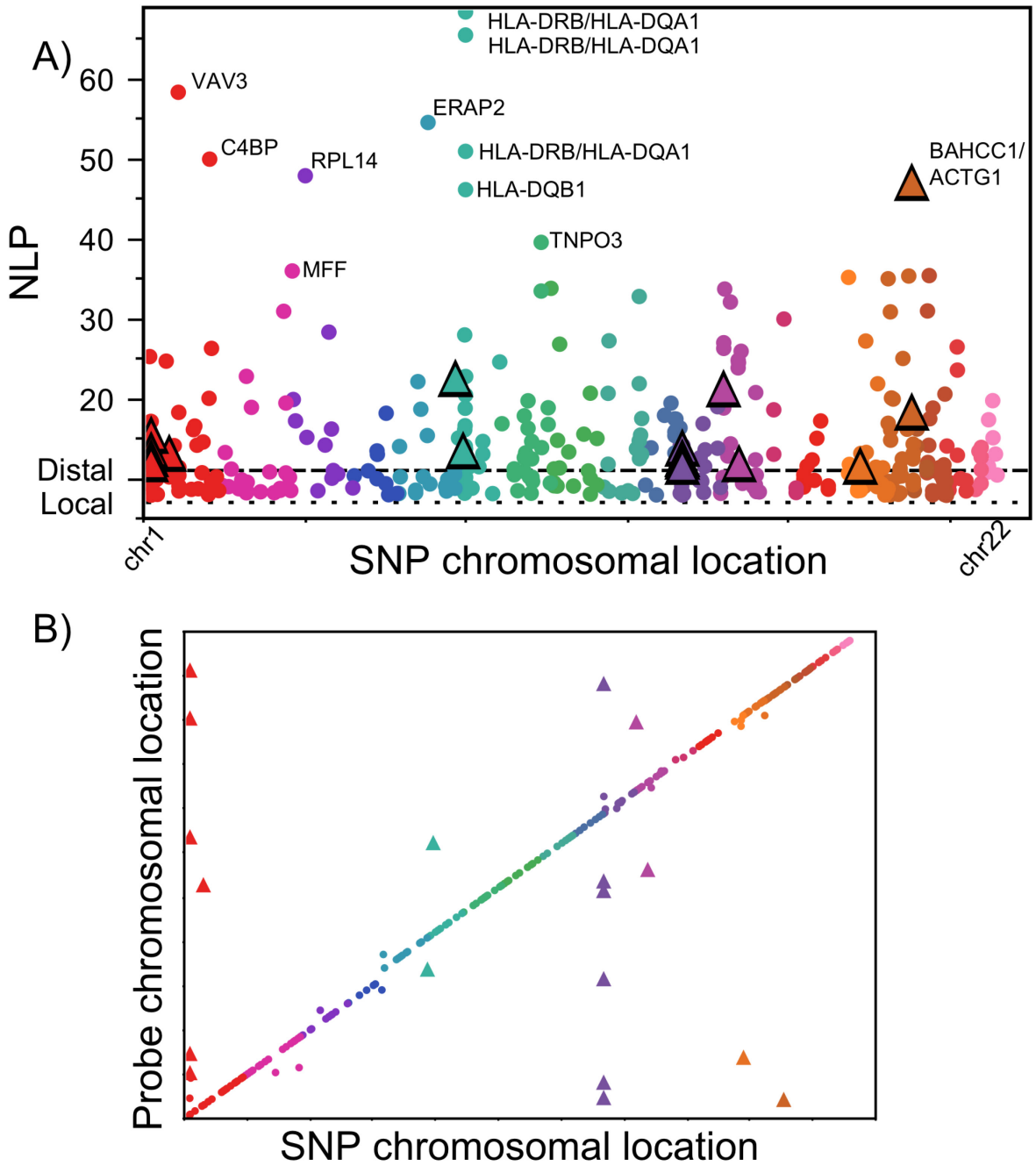

A)
NLP
60
50
40
30
20
Distal
Local
VAV3
C4BP
RPL14
ERAP2
HLA-DRB/HLA-DQA1
HLA-DRB/HLA-DQA1
HLA-DRB/HLA-DQA1
HLA-DQB1
TNPO3
MFF
BAHCC1/
ACTG1
chr1
chr22
SNP chromosomal location
B)
Probe chromosomal location
SNP chromosomal location

Figure D.4.2 Model 2 eSNP results.

Manhattan plot of peak 588 associations identified in Model 2. a) Each association is plotted by SNP chromosomal location on the x-axis and by significance (negative – log10 p-values (NLP)) on the y-axis. The colour code refers to the SNP chromosomal location. Local associations are circled and were significant at a NLP value of 7.88 or more. Distal associations are in triangles and were significant at a NLP of 11.34 or more.

b) Plot of eSNP associations based on their SNP (x-axis) and Probe (y-axis) chromosomal location. Those associations along the diagonal are local (the SNP-probe pair are located on same chromosome) and those off the diagonal are distal (the SNP-probe pair are located on different chromosomes). Out of the 588 peak associations, 579 are local and 9 are distal.

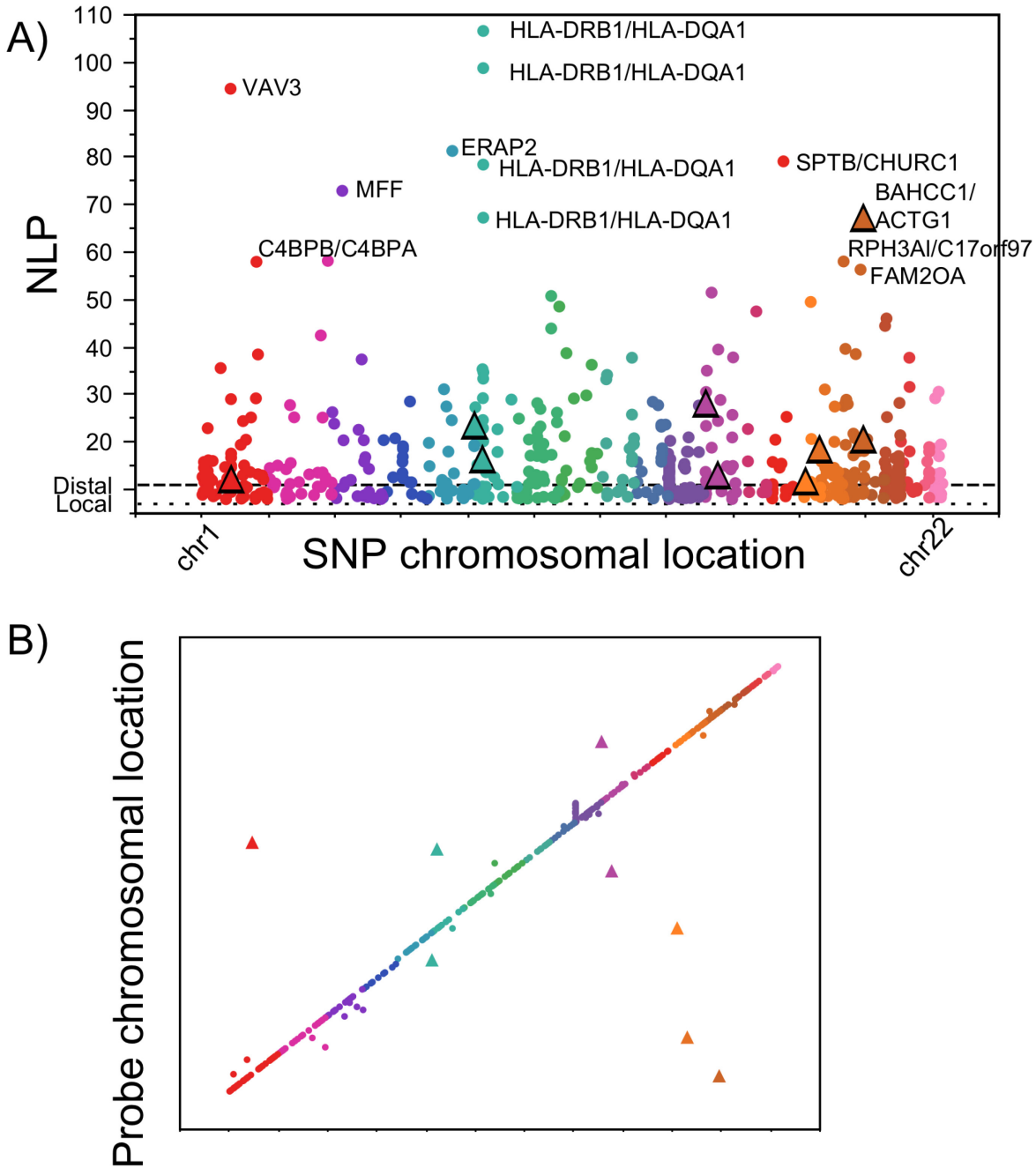
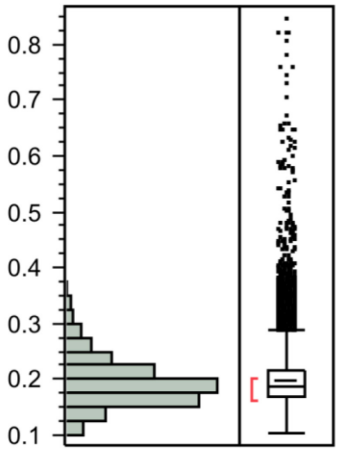

A)
NLP
110
100
90
80
70
60
50
40
30
20
Distal
Local
VAV3
HLA-DRB1/HLA-DQA1
HLA-DRB1/HLA-DQA1
ERAP2
HLA-DRB1/HLA-DQA1
SPTB/CHURC1
MFF
BAHCC1/
ACTG1
HLA-DRB1/HLA-DQA1
C4BPB/C4BPA
RPH3AI/C17orf97
FAM2OA
chr1
SNP chromosomal location
chr22
B)
Probe chromosomal location

Figure D.4.3 Variance explained by eSNP associations.

Histogram of variance explained by each significant eSNP association in the basic model of the combined data set. The mean proportion of the variation (R^2) that was explained by each eSNP association in the basic model for the combined data set was 0.1967 +/- 0.0494 s.d. The range was from 0.104–0.847.



In model 1, five eSNP genes, GSTM1, GSTT1, HLA.DQB1, HLA.DRB1, and SLC14.A1, were associated with SCD phenotypes in previously reported association studies (Table D.4.2). Model 2 identified eight genes, FCGR3B, GSTM1, GSTT1, HLA-DQB1, HLA-DRB1, HLA-G, HP, and SLC4A1 that are associated with an eSNP and which were previously associated with SCD phenotypes (Table D.4.3).

Almost half of the eSNP genes (150/390 eSNP genes in model 1; 229/527 eSNP genes in model 2) overlapped with previously reported significant eQTL associations (Table D.4.2 and D.4.3). Out of these, 58 in model 1 and 21 in model 2 were exact SNP-gene eSNP pairs. (Table D.4.2 and D.4.2.3).

Table D.4.2. Comparison of Model 1 eSNP results with literature.

A) In model 1, 6 eSNP genes were also previously associated with SCD in candidate gene studies, with one of them being associated with an interaction. The list of candidate genes previously associated with SCD was obtained from the PhenoPedia database [39]. The overlap between genes that had previously been associated with SCD phenotypes and with genes under eSNP control identified in our study was obtained. The six genes' and their respective references are shown in the table below, along with the probe-SNP pair that we identified. B) Overlap between genes associated with eSNPs in the present study and genes in previous eQTL studies that was obtained through the eQTL database (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/). The table shows the number of genes that were identified in our study to be associated with a SNP (detected at genome-wide significance) that overlap with genes reported in the twelve published studies. The comparison includes all genes reported in these studies that are significant at negaltive log10 p-value (NLP) greater than 7 (the genome wide threshold for local eSNP associations). The total number of genes that overlapped is two hundred eighteen, one hundred and fifty of which are significant at NLP > 7. Sixty eight are exact gene-SNP pairs, with fifty eight being significant at NLP > 7.

A

Gene	Probe	Effect	No.Publications	Reference
GSTM1	ILMN_1668134	SNP	1	Silva et al. 2011
GSTM1	ILMN_1762255	SNP	1	Silva et al. 2011
GSTM1	ILMN_2391861	SNP	1	Silva et al. 2011
GSTT1	ILMN_1730054	SNP	1	Silva et al. 2011
HLA-DQB1	ILMN_1661266	SNP	4	Mahdi et al. 2009; Mahdi et al. 2008; Al-Ola et al. 2008; Tamouza et al. 2002
HLA-DRB1	ILMN_1715169	SNP	4	Mahdi et al. 2009; Mahdi et al. 2008; Al-Ola et al. 2008; Tamouza et al. 2002
SLC14A1	ILMN_1805561	SNP	1	Ware et al. 2011
SLC14A1	ILMN_2197659	SNP	1	Ware et al. 2011
CCR2	ILMN_1777461	Interaction	1	Vargas et al. 2005

B

Associations	Number
total SCD eSNP associations	390
overlap with 12 published eQTL studies*	218
overlap with 12 published eQTL studies (NLP > 7)	150
overlap with exact SNP-gene associations	68
overlap with exact SNP-gene association (NLP > 7)	58

* Ref: Dimas09_Tcell, Dimas09_Fibro, Dimas09_Lympho, Montgomery10_exo, Montgomery10_trans, Myers, Pickrell10_eQTL, Pickrell10_sQTL, Schadt, Stranger, Vayrieras, Zellers10

Table D.4.3 Comparison of Model 2 eSNP results with literature.

A) In model 2, 8 eSNP genes were previously associated with SCD in candidate gene studies. The list of candidate genes associated with SCD was obtained from the PhenoPedia database [39]. The eight genes' and their respective references are shown in the table below, along with the probe-SNP pair that we identiifed. B) Overlap between genes associated with eSNPs in the present study and previous eQTL studies available through the eQTL database (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/). The table shows the number of genes in our study (detected at genome-wide significance) that overlap with genes reported in the twelve published studies indicated. The comparison was limited to genes reported in these studies at negative 10log p-value (NLP) greater than 7. The total number of genes that overlapped is two hundred twenty nine, twenty one of which are exact gene-SNP pairs.

A

eSNP Genes	# Probes	# Publications	Reference
FCGR3B	1	2	Taylor et al. 2002; Kuwano et al. 2000
GSTM1	3	1	Silva et al. 2011
GSTT1	1	1	Silva et al. 2011
HLA-DQB1	1	4	Mahdi et al. 2009; Mahdi et al. 2008; Al-Ola et al. 2008; Tamouza et al. 2002
HLA-DRB1	1	4	Mahdi et al. 2009; Mahdi et al. 2008; Al-Ola et al. 2008; Tamouza et al. 2002
HLA-G	1	1	Cordero et al. 2009
HP	1	3	Fowkes et al. 2006; Savy et al. 2010; Adekile et al. 2010
SLC4A1	1	1	Ware et al. 2011

B

Study	Genes (eSNP association with NLP>7)
Dimas09_Tcell	29
Dimas09_Fibro	20
Dimas09_Lympho	26
Montgomery10_exon	27
Montgomery10_transcript	18
Myers	9
Pickrell10_eQTL	31
Pickrell10_sQTL	12
Schadt	29
Stranger	91
Veyrieras_Pvalue	78
Zeller10	283
Total # unique overlapping eSNP genes	229
Exact # of overlapping gene-SNP association pairs	21

D.4.2 SNP-by-clinical severity interactions explain gene expression variability

Differential expression analysis revealed thousands of genes differentially expressed between clinical severities: 7,002 probes were differentially expressed for the 3-way clinical status effect (EvsFuvsCtls); and 4,220 probes were differentially expressed for the 6-way clinical category effect (EvsSSSvsSSUvsA). In order to identify regulatory effects that are dependent on clinical severity for these genes, we tested for SNP-by-Clinical Status interaction effects and SNP-by-Clinical category interaction effects in Models 3 and 4 (See Table D.4.1 and the Methods for details).

D.4.3 SNP-by-Clinical Status interactions

Because testing for interaction effects between the control and SCD follow-up groups might be sensitive to differences in the representation of each group within each genotypic class, we applied an additional filter to the list of SNPs in these models and excluded all SNPs not in HWE in the clinical groups (FU, Entry, Controls). This resulted in a final number of 455,750 SNPs that were retained for eSNP analyses. Bonferroni correction for multiple testing in this analysis resulted in a genome-wide significance threshold of p = 0.05/(7,002 probes x 200 SNPs) = 3.57 x 10^{-8} for local associations and p = 0.05/(7,002 probes x 455,750 SNP) = 1.27 × 10^{-11} for distal-associations.

Model 3 identified 13 significant interaction effects, eight of which remained genome-wide significant after accounting for relatedness in the entire sample using a Q-K mixed model ([143], see Methods for details): ZSCAN12L1 (p-value = 4.26 x 10^{-10}), C9ORF173 (p-value = 8.94 x 10^{-9}), CAPZA1 (p-value = 1.33 x 10^{-8}), SVIL (p-value = 2.41 x 10^{-8}), MEF2A (p-value = 1.69 x 10^{-8}), and C1ORF88 (p-value = 5.42 x 10^{-9}). These interactions are illustrated in Figures D.4.4 and in Appendix X. The significance of the associations before and after the QK-mixed model is shown in Table D.4. Figure D.4.3a-c shows three local eSNP interaction effects where higher expression levels of the corresponding gene in the

follow-up group relative to both the Entry group and the controls is driven by the minor allele of the eSNP in question. Figure D.4.3d-e shows two associations where the higher expression levels in the controls relative to SCD patients is observed only in the presence of the minor allele for the corresponding eSNP. A comprehensive circos plot illustrating the genome-wide significant SNP-probe associations from model 1 and the SNP-by-Clinical Status interaction effects from model 3 is shown in Figure D.4.5.

Figure D.4.4 Examples of significant SNP-by-clinical status interaction effects.

Five SNP-by-clinical status interaction effects are shown. All are local eSNP interactions. Expression levels are shown on the y-axis, and SNP genotype on the x-axis. The eSNP interaction involving gene zinc finger and SCAN domain containing 12 pseudogene 1 (ZSCAN12L1) is shown in (A); chromosome 9 open reading frame 173 (C9ORF173) is shown in (B); capping protein (actin filament) muscle Z-line, alpha 1 (CAPZA1) is shown in (C); supervillin (SVIL) is shown in (D); and myocyte enhancer factor 2A (MEF2A) is shown in (E). Linear regression for each group is plotted and colored: yellow for follow-up (FU), purple for entry (E), and red for controls (Ctls).

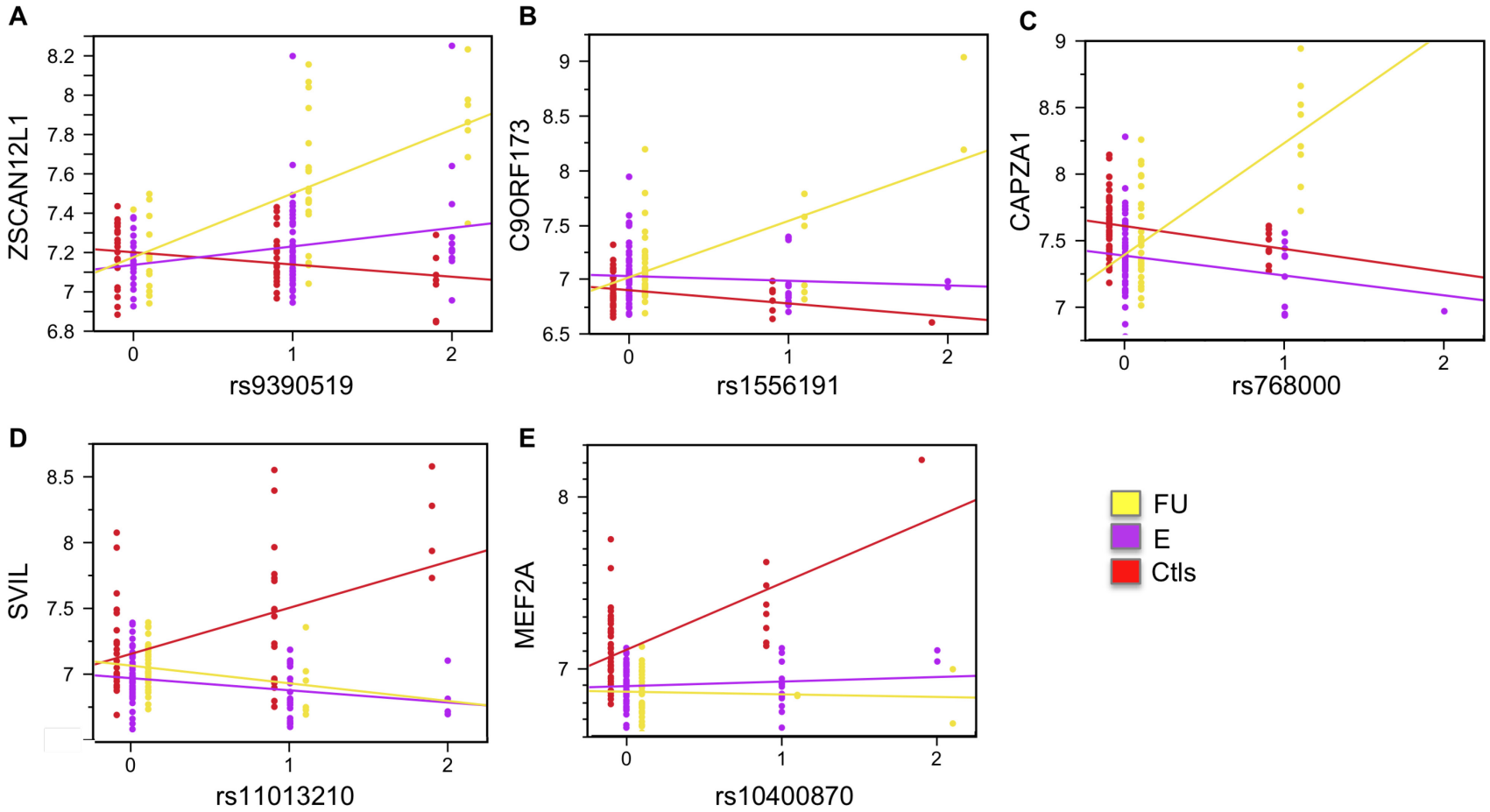

A
ZSCAN12L1
8.2
8
7.8
7.6
7.4
7.2
7
6.8
0
1
2
rs9390519
B
C9ORF173
9
8.5
8
7.5
7
6.5
0
1
2
rs1556191
C
CAPZA1
9
8.5
8
7.5
7
0
1
2
rs768000
D
SVIL
8.5
8
7.5
7
0
1
2
rs11013210
E
MEF2A
8
7
0
1
2
rs10400870
FU
E
Ctls

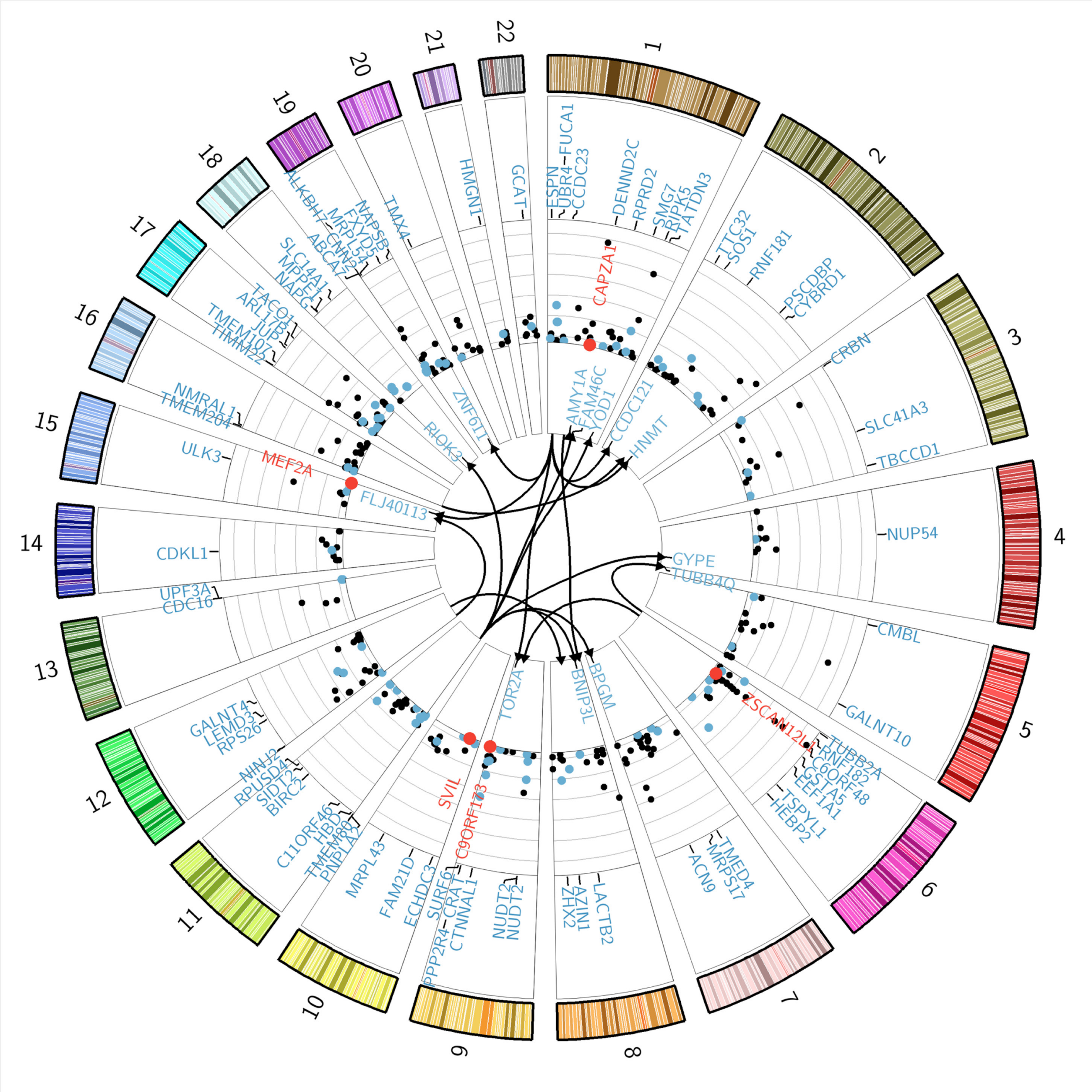
Table D.4.4 ClinStat interactions before and after accounting for relatedness.

Significance (–log(10) p-values (NLP)) for eSNP interactions effects between SNPxclinical status associated with gene expression levels. Using combined data set II, a multiple linear regression analysis was performed that accounted for Hb genotype, clinical status, sex and cell counts and tested for significant GenotypexClinical status interaction effects. 13 peak genome-wide significant SNPxClinical category interaction effects were identified all of which were local (Local_Distal column: local=0, distal =1). The negative log p values (NLP) for the interaction effects are shown before and after accounting for relatedness (NLP_before_QK; NLP_after_QK).

Gene	SNP	Local_Distal	NLP_before QK	NLP_after QK
SVIL	rs110113210	0	7.62	7.87
SVIL	rs7907625	0	7.62	7.87
SVIL	rs17103281	0	7.49	9.81
RPN2	rs2284277	0	7.64	7.15
CCR2	rs1531871	0	7.48	7.27
HIP1R	rs10862789	0	7.48	6.9
C9ORF173	rs1556191	0	8.05	9.01
CAPZA1	rs768000	0	7.88	13.17
C1ORF88	rs12026788	0	8.27	8.33
NBPF20	rs1935562	0	7.58	7.13
NBPF8	rs1935562	0	7.71	6.48
ZSCAN12L1	rs9390519	0	9.37	8.18
MEF2A	rs10400870	0	7.77	10.2

Figure D.4.5 Genetic regulation of gene expression in SCD patients.

Circularised Manhattan plot showing genome-wide significant SNP-probe associations from the analysis of the combined II dataset. Bonferroni correction for multiple testing was applied to all of our analyses with a genome-wide significance threshold of $p < 0.05/(19{,}431 \text{ probes} \times 200 \text{ SNPs}) = 1.28 \times 10^{-08}$ (NLP=7.89) for local associations in model 1 and $p < 0.05/(19{,}431 \text{ probes} \times 560{,}675 \text{ SNP}) = 4.59 \times 10^{-12}$ (NLP=11.34) for distal-associations in model 1; while model 3 thresholds were $p < 0.05/(7{,}002 \text{ probes} \times 200 \text{ SNPs}) = 3.57. \times 10^{-08}$ (NLP = 7.45) for local associations and $p < 0.05/(7{,}002 \text{ probes} \times 455{,}750 \text{ SNP}) = 1.57 \times 10^{-11}$ (NLP=10.80) for *distal*-associations. Distal associations are shown in the center of the plot. All genes involved in an interaction effect are differentially expressed and shown in red. eSNP genes from model 1 that are differentially expressed for the clinical status effect are shown in blue. The y-axis of the Manhattan plot indicates significance values (-log10 p-values) for the local-associations. Genes under eSNP control that are not differentially expressed for the clinical status effect (in the ANCOVA analysis at FDR 1%) are shown in black.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
CAPZA1
MEF2A
ZSCAN12L1
SVIL
C9ORF173
GYPE
TUBB4Q
FLJ40113
CDKL1
ULK3
NUP54
CMBL
GALNT10
CRBN
SLC41A3
TBCCD1
RNF181
GCAT
HMGN1
TMX4
TOR2A
BNIP3L
BPGM
LACTB2
MRPL43
FAM21D
NUDT2

D.4.4 SNP-by-Clinical Category interactions

Because testing for interaction effects between the SCD clinical categories might also be sensitive to differences in the representation of each group within each genotype class, we applied the additional filter to the list of SNPs in this model and excluded all SNPs not in HWE in each clinical category group (Entry, SSS, SSU, Acute). A final number of 399,821 SNP were retained for eSNP analyses. Bonferroni correction for multiple testing in this analysis resulted in a genome-wide significance threshold of p = 0.05/(4,220 probes x 200 SNPs) = 5.92 x 10^{-08} for local associations and p = 0.05/(4,220 probes x 399,821 SNP) = 2.96 × 10^{-11} for distal-associations. This analysis revealed 11 SNP-by-clinical category interaction effects (shown in Figure D.4.6 and Appendix XI), 2 of which remained genome-wide significant after accounting for relatedness in the entire sample using a Q-K mixed model ([143], see Methods for details, Table D.4.5): HP and ZNF16). A comprehensive circos plot illustrates the genome-wide significant SNP-probe associations from model 2 and the SNP-by-Clinical Category interaction effects from model 4 are shown in Figure D.4.7.

Table D.4.5 ClinCat interactions before and after accounting for relatedness.

Significance (–log(10) p-values (NLP)) for eSNP interactions effects between SNPxclinical category associated with gene expression levels. Using combined data set I, a multiple linear regression analysis was performed that accounted for Hb genotype, clinical category, sex and cell counts and tested for significant GenotypexClinical category interaction effects. 10 peak genome-wide significant SNPxClinical category interaction effects were identified: 8 local and 2 distal (Local_Distal column: local=0, distal =1). The negative log p values (NLP) for the interaction effects are shown before and after accounting for relatedness (NLP_before_QK; NLP_after_QK).

Gene	SNP	Local_Distal	NLP_before_QK	NLP_after_QK
ATP6V0A2	rs2711746	0	7.23	3.78
AZIN1	rs2460221	0	7.28	4.49
NOP56	rs6023423	0	7.68	3.88
OBFC1	rs7080456	0	7.29	5.83
PBX1	rs1294387	0	7.33	3.98
PIK3CG	rs10224307	0	7.21	3.42
PREPL	rs1919430	0	7.42	5.01
PTPRA	rs2328014	0	7.81	4.68
TMEM180	rs6573755	1	10.73	6.70
ZNF716	rs758676	0	9.99	11.60
HP	rs742898	1	14.03	12.2

Figure D.4.6. Examples of significant SNP-by-clinical category interactions.

Four eSNP interaction effects involve the genes PIK3CG, associated with SNP rs10224307 (NLP = 7.2, Figure D.4.6a), HP, associated with SNP rs742898 (NLP = 13.6, Figure D.4.6b), AZIN1, associated with the SNP rs2460221 (NLP = 7.3, Figure D.4.6c), and OBFC1, associated with the SNP rs7080456 (NLP = 7.3, Figure D.4.6d). These examples show how the eSNP effect is dependent on SCD clinical category being most significant mainly in the the Acutes and in SSU patients relative to other categories.

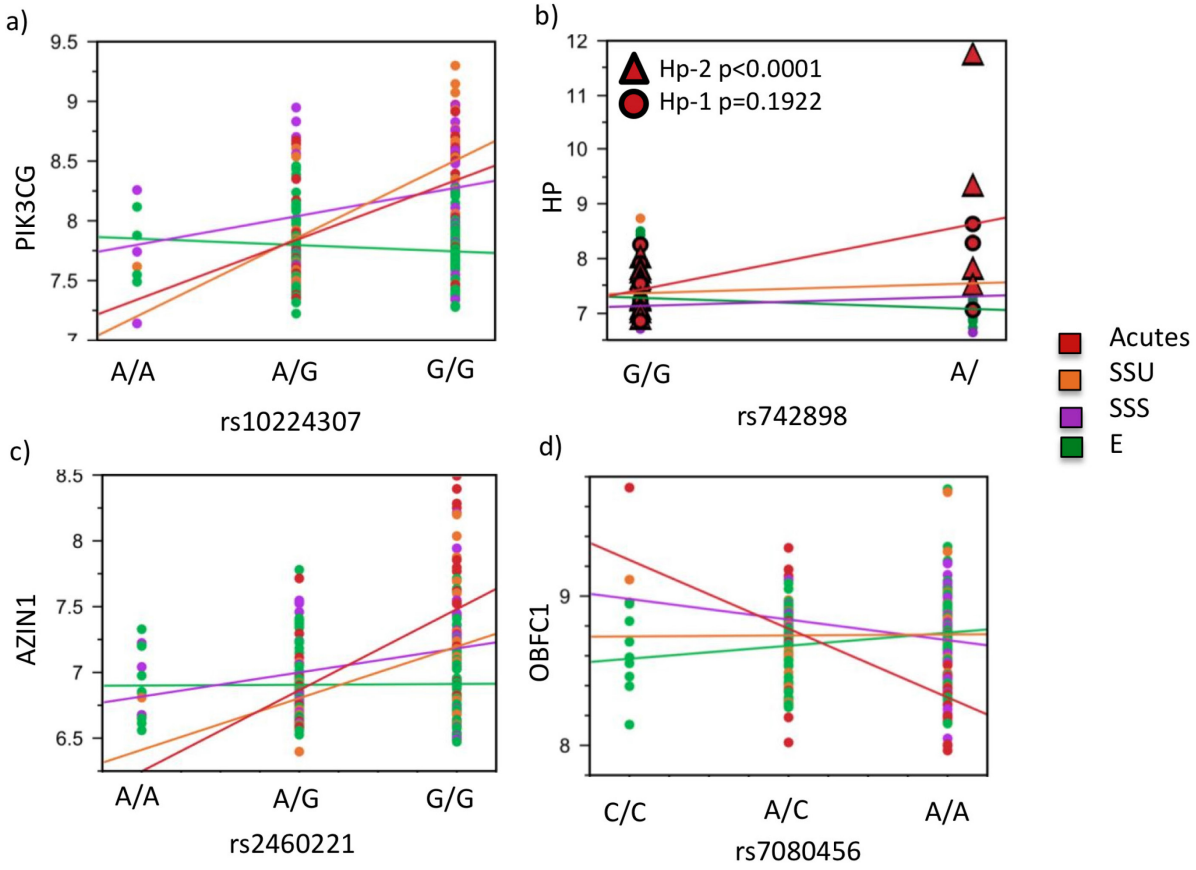
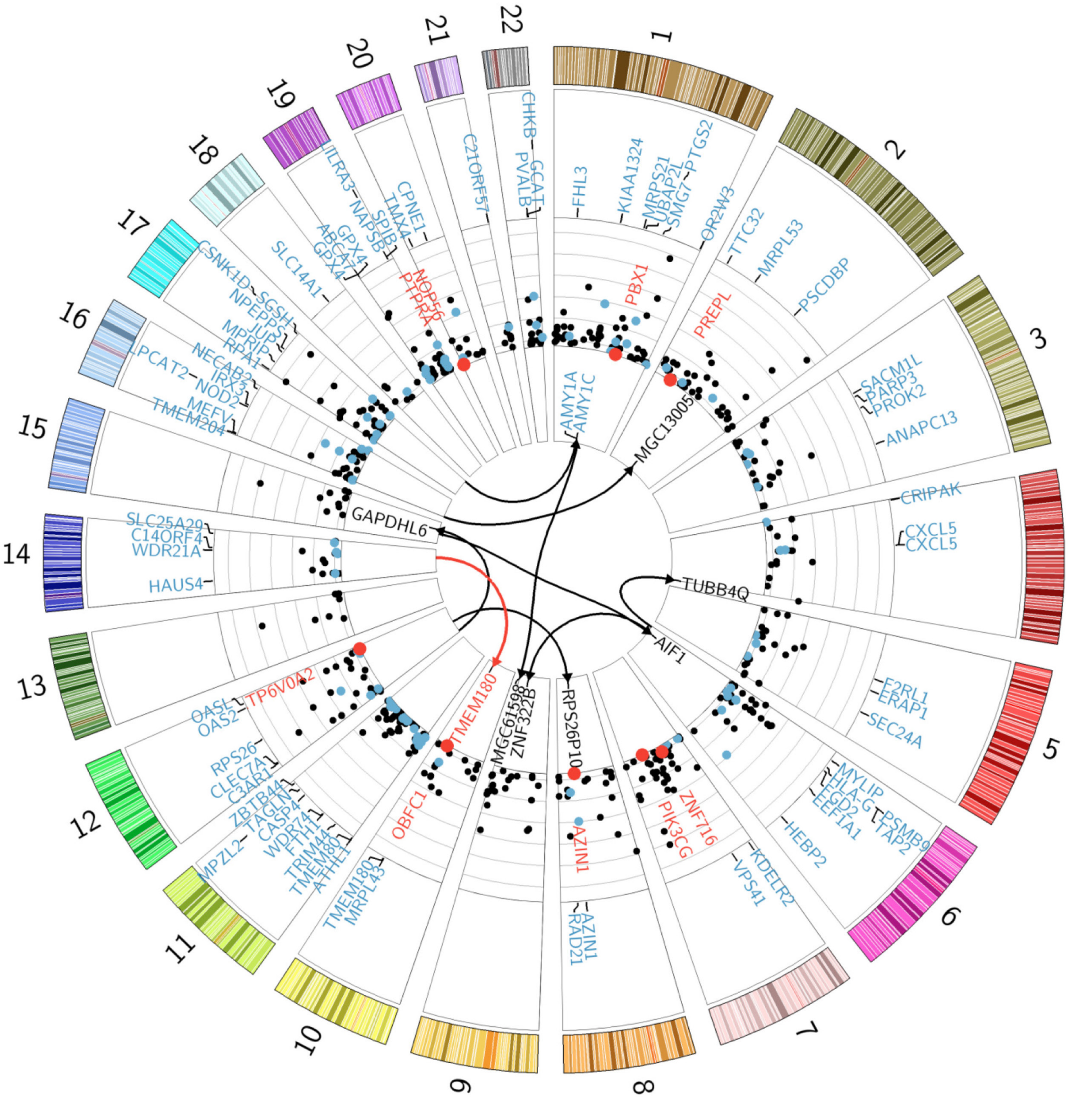

a)
PIK3CG
9.5
9
8.5
8
7.5
A/A
A/G
G/G
rs10224307
b)
HP
12
11
10
9
8
7
Hp-2 p<0.0001
Hp-1 p=0.1922
G/G
A/
rs742898
Acutes
SSU
SSS
E
c)
AZIN1
8.5
8
7.5
7
6.5
A/A
A/G
G/G
rs2460221
d)
OBFC1
9
8
C/C
A/C
A/A
rs7080456

Figure D.4.6 Genetic regulation of gene expression in SCD patients.

Circularised Manhattan plot showing genome-wide significant SNP-probe associations from the analysis of the combined dataset. Bonferroni correction for multiple testing was applied to all of our analyses with a genome-wide significance threshold of $p < 0.05/(18,890 \text{ probes} \times 200 \text{ SNPs}) = 1.32 \times 10^{-08}$ (NLP=7.89) for local associations in model 2 and $p < 0.05/(18,890 \text{ probes} \times 568,921 \text{ SNP}) = 4.65 \times 10^{-12}$ (NLP=11.33) for distal-associations in model 2; while model 4 thresholds were $p < 0.05/(4,220 \text{ probes} \times 200 \text{ SNPs}) = 5.92. \times 10^{-08}$ (NLP = 7.2) for local associations and $p < 0.05/(4,220 \text{ probes} \times 399,821 \text{ SNP}) = 2.96 \times 10^{-11}$ (NLP=10.5) for *distal*-associations. Distal associations are shown in the center of the plot. All genes involved in an interaction effect are differentially expressed and shown in red. eSNP genes from model 2 that are differentially expressed for the clinical category effect are shown in blue. The y-axis of the Manhattan plot indicates significance values (-log10 p-values) for the local-associations. Genes under eSNP control that are not differentially expressed for the clinical category effect (in the ANCOVA analysis at FDR 1%) are shown in black.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
FHL3
KIAA1324
MRPS21
UBAP2L
SMG7
PTGS2
OR2W3
TTC32
MRPL53
PSCDBP
PBX1
PREPL
SACM1L
PARP3
PROK2
ANAPC13
CRIPAK
CXCL5
CXCL5
F2RL1
ERAP1
SEC24A
MYLIP
PSMB9
TAP2
EEF1A1
HEBP2
KDELR2
VPS41
ZNF716
PIK3CG
AZIN1
AZIN1
RAD21
RPS26P10
MGC61598
ZNF322B
TMEM180
OBFC1
TMEM180
MRPL43
ATHL1
TRIM44
FTH1
WDR74
CASP4
TAGLN
ZBTB44
C3AR1
CLEC7A
RPS26
MPZL2
OASL
OAS2
TP6V0A2
HAUS4
WDR21A
C14ORF4
SLC25A29
TMEM204
MEFV
NOD2
IRX3
NECAB2
LPCAT2
RPA1
MPRIP
JUP
NPEPPS
SGSH
CSNK1D
SLC14A1
GPX4
ABCA7
NAPSB
SPIB
TMX4
CPNE1
NOP56
PTPRA
C21ORF57
PVALB
GCAT
CHKB
AMY1A
AMY1C
MGC13005
GAPDHL6
TUBB4Q
AIF1

D.5 POTENTIAL SCD DRUG TARGETS

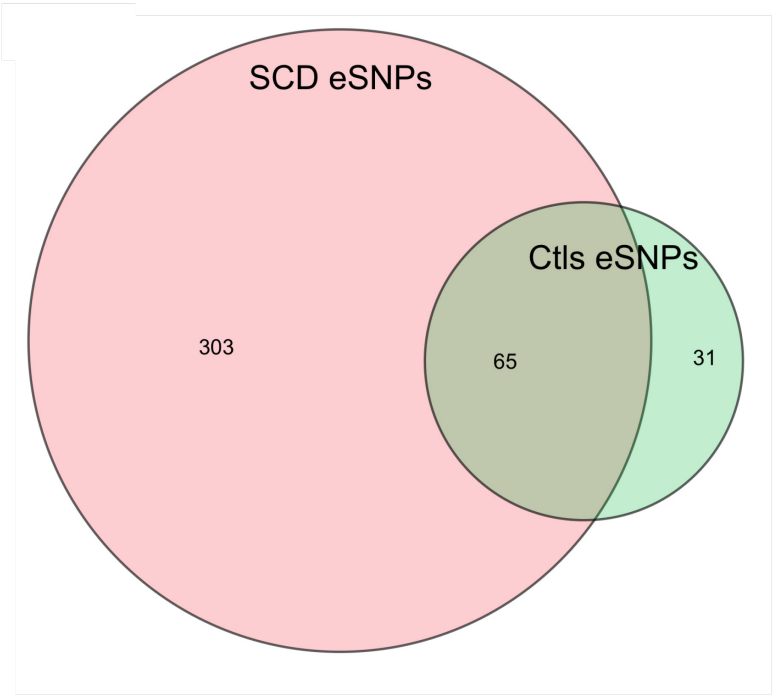
D.5.1 Drug target genes identified for SCD patients

Methods of translating our genomic results into improved clinical care can be accomplished by identifying overlaps between genes that are important in the pathobiology of SCD and genes that are known drug targets. These lists provide candidate drugs that can be evaluated for their effectiveness in improving SCD outcome.

In order to identify drug targets that are specific to the disease, we identified genes that control differential expression between SCD patients and controls. Basic eSNP analyses were run separately for 205 SCD patients and 58 controls which identified significant associations between SNP genotype and probe expression levels for each group. After stringent filtering based on quality control and marker properties (see Methods) and using Bonferroni thresholds, 368 eSNP associations remained for the SCD patients and 96 eSNP associations for the control group. This gave a final number of 303 eSNP associations unique to the SCD group (Figure D.5.1).

Figure D.5.1 Venn diagram for eSNP associations.

The overlap of eSNP associations between SCD patients and controls.



Thirty four out of the 368 eSNP genes in the SCD group were drug targets identified in the CTD database, whereas eleven out of the 96 eSNP genes in the control group were drug targets identified in the CTD database. Since we were interested in identifying genes that were potential drug targets for SCD patients, we restricted further investigations to the genes that were unique to the SCD group. Of the 34 drug targets in the SCD group, twenty five (25) were unique to the SCD group. Table D.5.1 gives a list of these 25 genes that are drug targets along with the corresponding drug. In Appendix XII, illustrations of the 25 eSNP associations are shown.

Table D.5.1 Twenty-five drug targets for SCD.

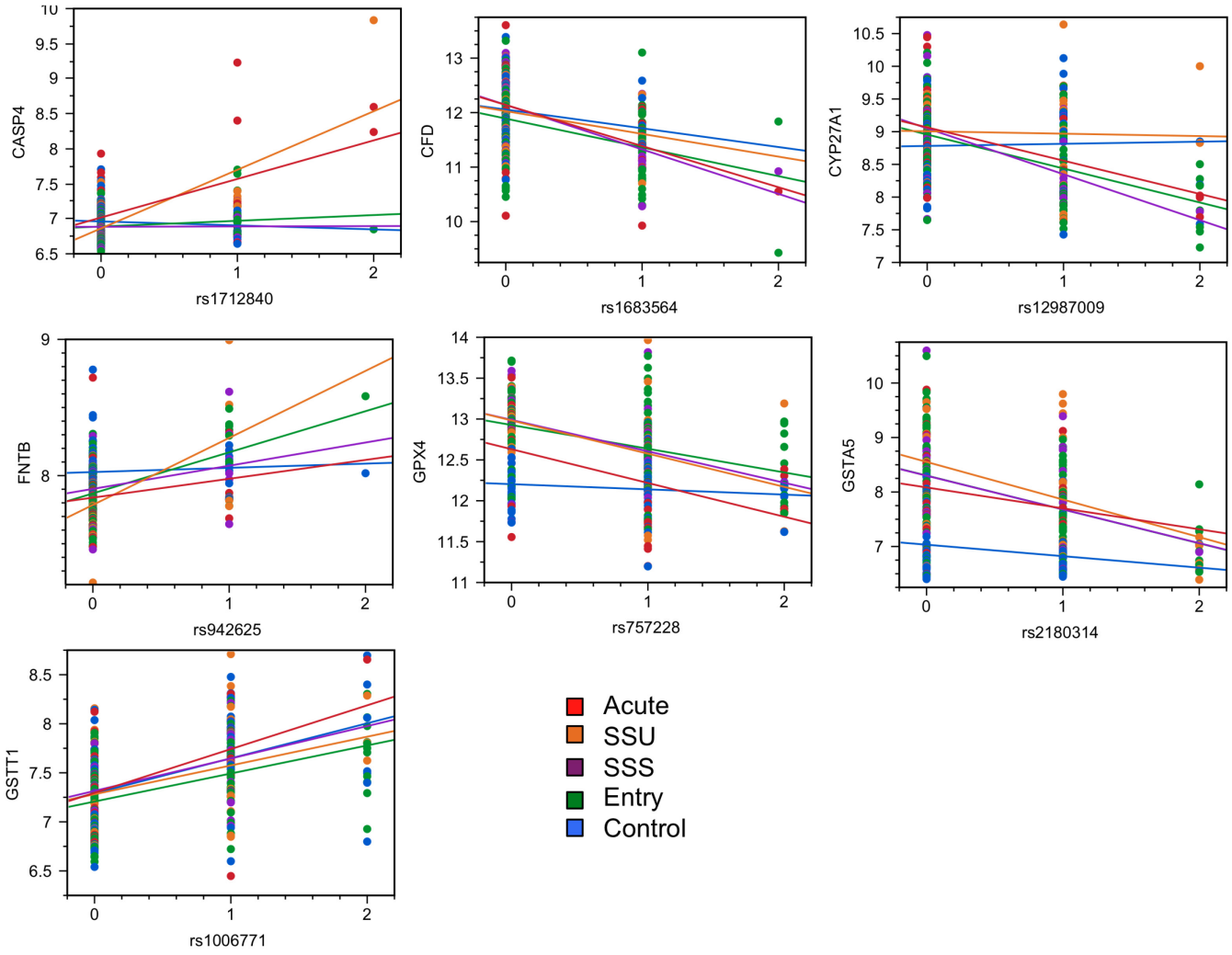
The twenty five drug targets were significant eSNPs uniquely in the SCD patient group and not significant in the control group. The eSNP genes and corresponding drugs are listed in the table.

eSNP gene	SCD drug
BIN3	Paclitaxel
CASP4	Folic Acid|Paclitaxel
CD36	alpha-Tocopherol|Paclitaxel|Zinc
CFD	Paclitaxel
CXCL5	Paclitaxel|Zinc
CYP19A1	Paclitaxel
CYP27A1	Morphine
EEF1A1	Zinc
EPHX2	Morphine
FDFT1	Paclitaxel
FNTB	Zinc
GPX4	alpha-Tocopherol
GSTA5	Oxycodone
GSTM1	Paclitaxel|Prednisone
GSTT1	Paclitaxel
HBG2	alpha-Tocopherol
HNRNPL	Morphine
MPZL2	Paclitaxel
MTRR	Folic Acid
PAM	Zinc
POMZP3	Paclitaxel
PTGS2	Morphine
SLC39A8	Zinc
TAGLN	Folic Acid
ZFAND2A	Zinc

Of the 25 eSNP drug targets, seven eSNP associations were dependent on clinical category (Figure D.5.2). This would suggest that SCD patients may not respond equally to medications that target these genes.

Figure D.5.2 Drug targets by ClinCat.

Seven examples of eSNP associations plotted by clinical category. The linear regression for the associations are coloured by clinical category.



D.5.2 Drug target genes identified for SCD patients after follow-up

We were particularly interested in identifying treatments to improve patients who do not improve in their condition even after follow-up. Gene expression analysis identified 824 probes (775 genes) that were differentially expressed between SSS and SSU patients after accounting for Hb genotype, clinical category, and sex (FDR=5%; see Appendix XIII). Of these differentially expressed genes, 86 were drug targets identified in the CTD database. Enrichment analysis for these 86 gene targets that are differentially expressed between SSS and SSU patients was performed using ToppFun (http://toppgene.cchmc.org/enrichment.jsp). Significant enrichment in the vitamin binding category for the GO Molecular Function category was identified (Appendix XIV). Enrichment in this pathway may suggest potential therapies that should be investigated for clinical use in improving SCD patients who are unsatisfactory.

Examining which of the differentially expressed genes were under genetic control identified six of the 84 differentially expressed drug targets to be under genetic control (Table D.5.3). Of particular interest, 2 were involved in vitamin binding: peptidylglycine alpha-amidating mono-oxygenase (PAM) and glycine acetyltransferase (GCAT). In Figure D.5.4, the eSNP association for the 6 genes are illustrated. As can be seen, the eSNP association is dependent on the SCD clinical category.

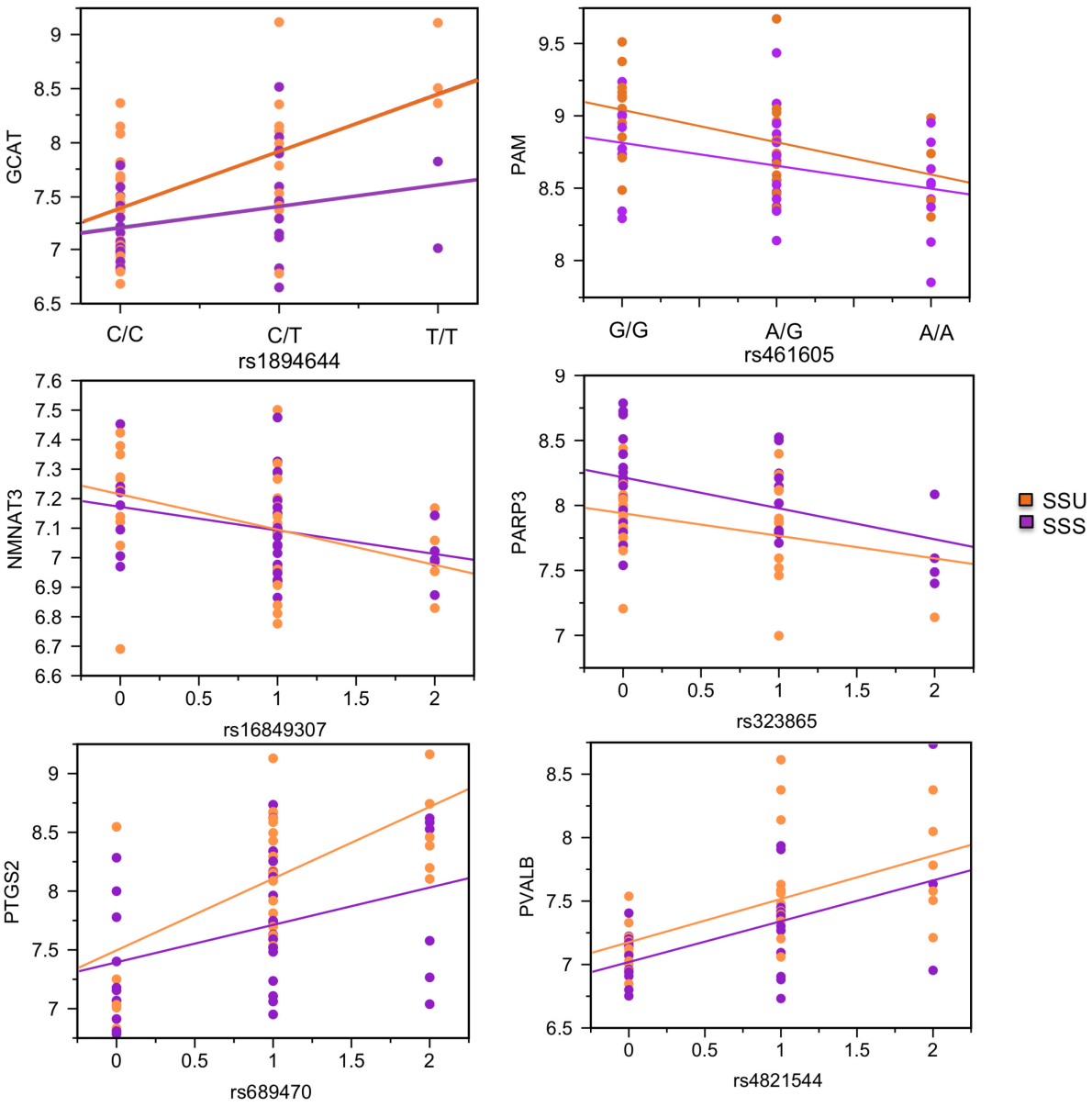
Table D.5.2 Six potential drug targets identified for SSS and SSU patients.

The six potential drug targets are also eSNP genes that are differentially expressed between SSS and SSU patients. All associations are local eSNP associations (the SNP-probe pair are on the same chromosome). Significance for all associations were genome-wide after Bonferonni correction (negative log10 pvalue (NLP)).

DrugTarget	SNP	SNP Chr	Probe	Probe Chr	Local Distal	NLP
GCAT	rs1894644	22	ILMN_1724437	22	Local	17.07
NMNAT3	rs16849307	3	ILMN_1665123	3	Local	9.55
PAM	rs461605	5	ILMN_2313901	5	Local	13.44
PARP3	rs323865	3	ILMN_2397954	3	Local	11.55
PTGS2	rs689470	1	ILMN_2054297	1	Local	25.21
PVALB	rs4821544	22	ILMN_2069224	22	Local	30.62

Figure D.5.3. Six drug targets.

The 6 eSNP associations that are drug targets are shown below and colored by clinical category (SSS in orange and SSU in purple). On the y-axis, gene expression levels are plotted, and SNP genotype is plotted on the x-axis.



E. DISCUSSION

E.1 Discussion of key results

In 2008, SCD was recognised as a global health priority by the United Nations who urged African countries to adopt a national strategy to reduce under-5 mortality by improving care and treatment of this disorder. In order to improve treatment, basic research is required to increase our knowledge of the factors that contribute to the clinical heterogeneity of SCD. Characterising the sources of the clinical variation in SCD patients is important to improve therapies and follow-up programs, lessening the burden of the disease on the public health system. Despite many candidate gene and genome-wide association studies having been performed, the majority of the underlying genetic factors that contribute to the phenotypic variation in SCD remain unknown. In this study, we used the joint power of genotyping and gene expression analysis [96] to characterise the genomic architecture of SCD. Utilising this approach increased our chance of success. A summary of our key findings and their relavence is explained below.

E.1.1 Discussion of General Results

A total of 311 children from Cotonou, Benin, West Africa, were recruited for this study, including 250 pediatric SCD patients with either the HbSS or HbSC genotype sampled in two phases. A two-phase approach allowed us to replicate our findings. The distribution of SCD severity, clinical status, clinical categories, Hb genotypes, and sex were proportionate in both phases. Sixty one healthy siblings of SCD patients at the CPMI-NFED, with at least one normal hemoglobin allele, and of roughly equal age and proportions of sex were also recruited. Participants were distantly or not related with each other. Since cases and controls were not significantly related, the assumption of independent observations that is needed in

most association test statistics was not violated. Ethnicity analyses revealed that the sample was of similar ethnic background (based on gPCA), thus eliminating the possibility of population stratification.

Three quarters of our controls were heterozygous HbAS and one quarter were homozygous HbAA. Few probes were differentially expressed between HbAA and HbAS individuals and none of the variance in the controls was explained by the genotype effect. For these reasons, we grouped HbAA and HbAS individuals and used them as a control sample.

In addition to nucleic acids and clinical phenotypes, we obtained hematological samples from each participant in this study. In order to maximise relevancy and minimize redundancy, we only retained red blood cells (RBCs) and white blood cells (WBCs) in our analyses. Low RBC count can indicate anemia, an important determinant of SCD severity, and elevated WBC count can be an indication of infection and an acute event in SCD patients. Since cell count differences might confound differential expression, variability in RBCs and WBCs were included as covariates in the analyses.

Although the Hb genotype is the primary determinant of SCD severity, we did not detect evidence for association between Hb genotype and clinical severity as measured by SCD clinical categories. Nor did we detect significant associations between β-SCD haplotypes and SCD clinical categories. Thus, other genetic factors, located in other genomic regions, or environmental factors must influence the clinical heterogeneity observed in our SCD sample.

Previous studies identified that men with sickle cell disease experience more sickle cell crises than do women, but this was documented after puberty [12]. We did

not detect evidence for association between sex and clinical category, which is most likely explained by fact that the SCD patients in our sample are young and pre-pubescent.

E.1.2 The influence of SCD on the human transcriptome

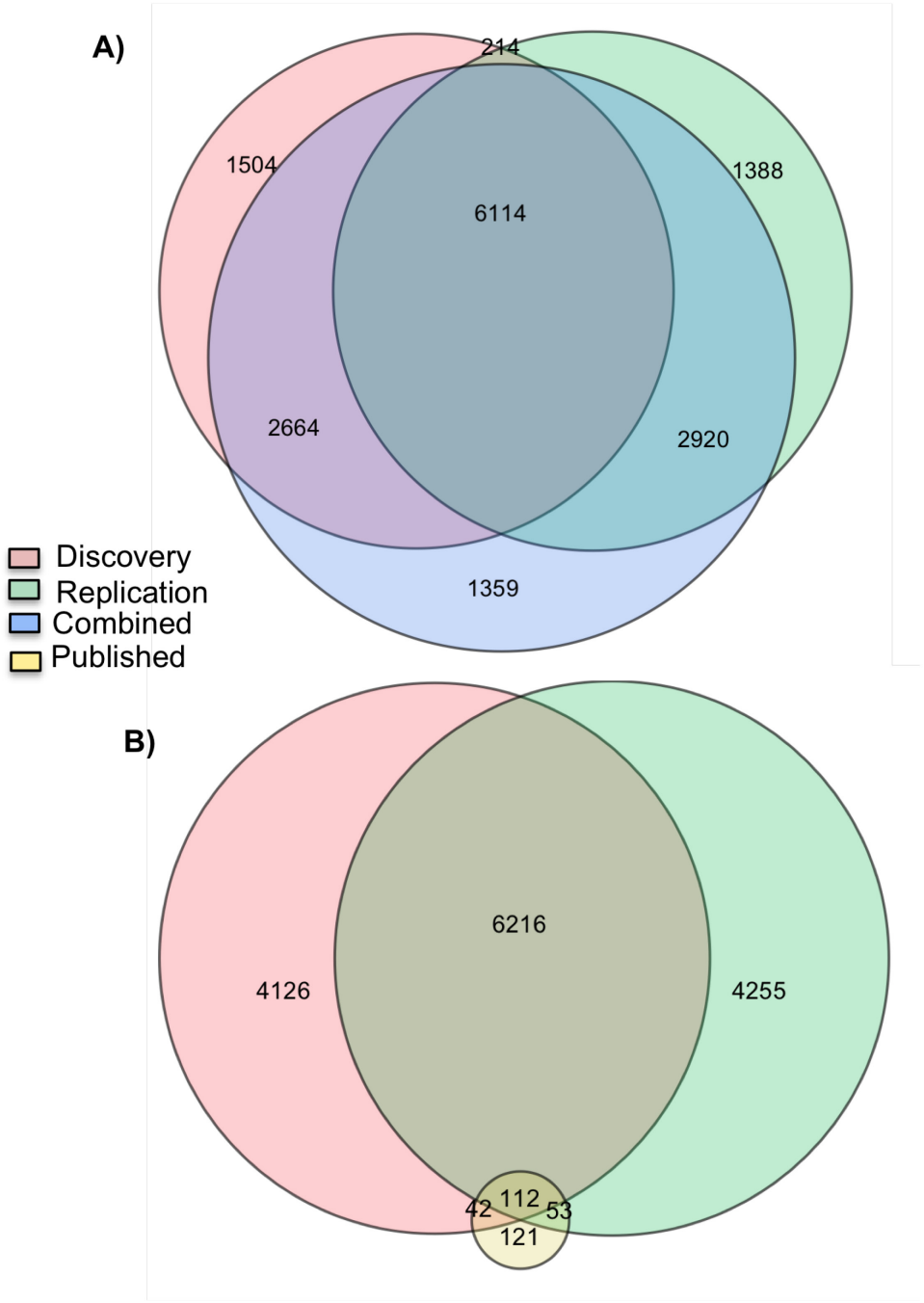
We first identified the extent of gene expression variation in SCD patients that is explained by clinical phenotypes and measured the magnitude and significance of gene expression differences between SCD patients in an initial discovery phase. The unsupervised analysis of gene expression profiles shows that SCD has substantial influence on the human transcriptome, explaining over a third of the total variance. Hb genotype, SCD clinical status and SCD clinical category are the main variables that influence this gene expression variation. The Hb genotype had the strongest effect on gene expression. However, significant differences in gene expression profiles between SCD patients (grouped according to clinical status or clinical category) and controls were also observed, with over a quarter of the transcriptome being differentially expressed. We replicated these findings in a replication cohort. The only other documented work on differential expression in SCD patients examined differences between HbSS patients and controls [43]. When we compare our results with the previously published results from this study by Jison *et al.* [43], we found an overlap of 207 genes that were differentially expressed. One-hundred and twelve (112) of these genes were differentially expressed in the discovery, the replication, and in Jison's published data set (see Figure E.1.1). Differences in the number of differentially expressed genes identified in our study and in the one documented by Jison *et al*. may be due to the differences in sample size (17 untreated SCD patients and 13 controls in the Jison study, as compared to our sample of 250 SCD patients and 61 controls). Furthermore, the analysis was

performed using different blood types (we used whole blood, while Jison restricted their analysis to peripheral blood mononuclear cells without platelets and neutrophils). Most likely some of the gene expression variation observed in our study is driven by the differences in the subsets of cell types that are present in whole blood. By using whole blood, we capture more variation; however, we haven't measured cell specific effects. It is possible to infer cell specific effects by using the genomic signature of flow-cytometry-sorted immune cell types reported by Nakaya *et al.* [139] where cell type-specific modules are constructed based on the level of expression levels of each gene relative to each other cell type in the PBMC mixture[139]. We identified significant associations between SCD clinical category contrasts and the six cell type-specific expression profiles investigated (Appendix XV). Finally, the differences in technologies that were used in our study and Jison's study (we used Illumina HT12v4 chips that captured over 47,000 probes while Jison used the HU95Av2 gene chip from Affymetrix containing only 12,626 sequenced genes) probably also played a role in differences in the results. However, in both our study and in Jison's study, metabolism, cell-cycle regulation, angiogenesis, inflammation, antioxidant and stress response pathways were enriched from the genes that were differentially expressed.

Other studies have examined gene expression differences in SCD, but these are not comparable to ours for various reasons. One study used gene expression to validate a protocol and did not analyse their data beyond the validation step [44]. The second study performed their analysis on microarray chips that captured miRNAs and not genes [45].

Figure E.1.1 Comparison of gene expression results with literature.

A) Venn diagram of the gene that were differentially expressed between HbSS SCD patients and controls in the discovery, replication and combined data set I. B) Venn diagram of the genes that were differentially expressed between HbSS SCD patients and controls in the discovery and replication data sets and in the published results from Jison et al [43].



E.1.3 Biological pathways implicated in SCD

Gene Set Enrichment Analysis (GSEA) [137] procedures allowed us to identify and replicate biological pathways involved in the clinical course of SCD. Strong modulation of the transcriptome implicates pathways affecting core hematological cell functions.

Enrichment of genes that were uniquely differentiated between the entry and follow-up patients identified a significant up-regulation in B-lymphocytes expressing phosphorylated CD5, B-cell Receptor Signalling and upstream regulation of B-cells by PAX5. PAX5 expression has been shown to increase the quantity and the commitment of B cells [145]. These observations reflect perturbed cellular profiles in the entry groups and more stable profiles after clinical follow-up. Furthermore, markers of mitosis, cell cycle and DNA synthesis were identified in the analysis on combined data set II and likely suggest a more stable state of blood cells in the follow-up group in general. The strong interferon related signature also suggests a more perturbed and potentially more pathogenic state of blood cells prior to clinical follow-up. The over expression of activated B lymphocyte markers in the entry group tends to point in that same direction. Previous studies have shown that changes in B cell function occurs during vaso-occlusive crisis (VOC) in patients with SCD [146]. Thus, follow-up of SCD patients may act on these pathways.

The results of the GSEA that was performed on differentially expressed genes between clinical categories identified a strong inflammatory response signature in acute patients, which is consistent with the processes induced during SCD crises events such as vaso-occlusive crisis (VOC) [147]. Activation of platelets in SSU patients suggests their implication in SCD complications that contribute to the

unsatisfactory clinical state, including SCD vasculopathy [148] and hemolysis-associated pulmonary hypertension [149]. Pathways associated with B- and T-cell stimulation, as well as metabolism-related pathways that are up-regulated in entry patients and SSS patients suggest an induction of these processes and a role in driving a clinically satisfactory state. Several of these pathways were previously shown to be differentially regulated between SCD steady-state patients and controls [43].

E.1.4 Genetic control of gene expression in SCD

To the best of my knowledge, this is the first study to characterize the genetic architecture of transcript abundance in SCD patients and controls. Hundreds of peak eSNP associations were identified. Three hundred and ninety genome-wide significant peak SNP-probe associations were identified in model 1 (which includes the Clinical Status effect) and five hundred and eighty-eight in model 2 (which includes the clinical category effect). This corresponds to 371 local and 19 distal effects in model 1, and 579 local and 9 distal effects in model 2. These associations explain on average 20% of the variance in transcript abundance in either model. Seventy five percent of the local associations implicate SNPs located within 1 Mb from the probe coordinates.

Of the genes that are differentially regulated among SCD clinical severities, many are also under genetic control. One hundred (100) of the 390 significant eSNP associations in model 1 implicated genes that are differentially expressed for the 3-way SCD clinical status effect. Eighty-nine out of the 588 significant eSNP associations implicated genes that are differentially expressed for the 6-way SCD-clinical category effect. We identified an overlap between eSNP genes identified in this study and with genes previously identified to be associated with SCD severity.

Five genes that are associated with an eSNP in model 1 were previously associated with SCD phenotypes in reported association studies, and eight genes which are associated with an eSNP in model 2 were previously associated with SCD phenotypes. We observed significant overlap between the eSNP associations for SCD and those reported in 12 published eQTL studies of various tissues including peripheral blood and its derivatives. Almost half of the eSNP genes in model 1 (150/390 eSNP genes) are replicated. Of these, 58 were exact SNP-gene eSNP pairs. Almost half of the eSNP genes in model 2 (229/527 eSNP genes) were also replicated. Of these, 21 were exact SNP-gene eSNP pairs (21/229). The other associations in our datasets are either novel, of weaker strength in the 12 eQTL studies or might be reported in other published studies.

E.1.5 Identification of interaction effects

Using the joint analysis of gene expression and genotyping, we identified statistical SNP-by-clinical severity interaction effects that were associated with log-transformed gene expression levels. The regression coefficients for the interaction terms represent the change in the genotypic differences in mean log-transformed expression levels.

Differential gene expression analyses identified thousands of genes differentially expressed between SCD patients with different clinical severities (Clinical Status or Clinical Category). We restricted the eSNP interaction analyses to this class of genes: for the SNP-by-Clinical status effect (model 3) we tested for significant interactions with 7002 genes; for the SNP-by-Clinical category effect (model 4) we tested for significant interactions with 4220 genes. In doing so, we identified regulatory eSNP effects that are dependent on clinical status or clinical category for genes that are differentially expressed.

Model 3, which included the SNP-by-Clinical status effect, identified 13 significant statistical interaction effects, eight of which remained genome-wide significant after accounting for relatedness in the entire sample using a Q-K mixed model [143]. The genes that are associated with these 8 interactions are ZSCAN12L1 (p-value = 4.26 x 10^{-10}), C9ORF173 (p-value = 8.94 x 10^{-9}), CAPZA1 (p-value = 1.33 x 10^{-8}), SVIL (p-value = 2.41 x 10^{-8}), MEF2A (p-value = 1.69 x 10^{-8}), and C1ORF88 (p-value = 5.42 x 10^{-9}). Three of these eSNP interaction effects have higher expression levels of the corresponding gene (ZSCAN12L1, C9ORF173, and CAPZA1) in the follow-up group relative to both the Entry group and the controls, and this is driven by the minor allele of the eSNP in question. These interactions may help explain why certain SCD patients do not improve even after following a rigorous clinical care program. Two of the remaining eSNP interactions have higher expression levels in the controls relative to SCD patients being observed only in the presence of the minor allele for the corresponding eSNP (SVIL, and MEF2A). The genes associated with the clinical status interactions have not been documented to have a link with SCD, and thus the biological mechanism of how they would affect the clinical progression of the disease is not known. Nonetheless, it remains of interest to identify if these interactions can provide a basis for targeted interventions for SCD patients with a worse clinical phenotype.

In model 4, 11 significant SNP-by-Clinical category effects were identified that are dependent on clinical category, 2 of which remained genome-wide significant after accounting for relatedness in the entire sample using a Q-K mixed model [143]. The genes that are associated with the 11 interactions are ATP6VOA2, AZIN1, NOP56, OBFC1, PBX1, PIK3CG, PREPL, PTPRA, TMEM180, ZNF716 (NLP = 11.60) , HP (NLP =12.2). Seven of these interaction effects involve genes that have

been associated with disease, with three local and one distal effect implicating genes known to be involved in biological processes implicated in the pathobiology of SCD. A description of these effects and how they may impact SCD is given below.

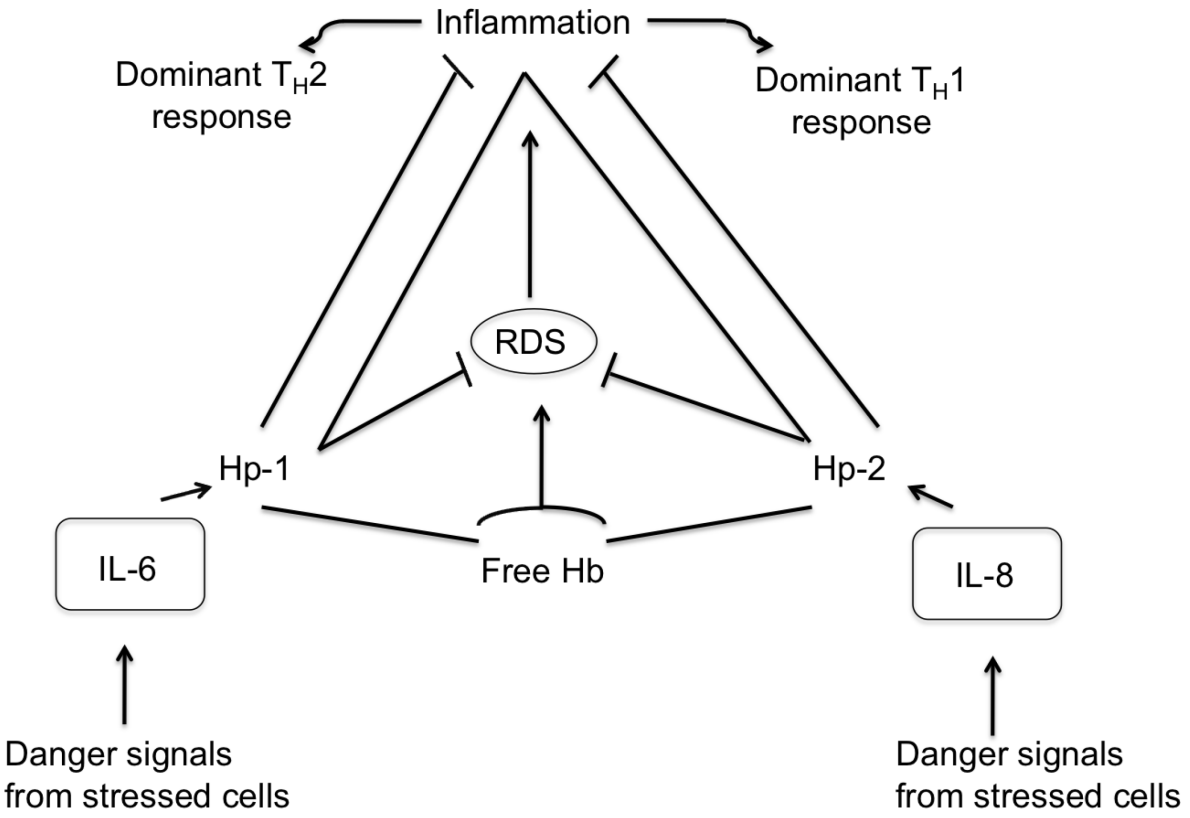
A local eQTL interaction involved the PIK3CG gene. This gene has recently been associated with platelet aggregation [150], and in another GWAS, a SNP associated with PIK3CG transcript levels and protein function was associated with mean platelet volume and platelet count [151]. Furthermore, mice lacking the PIK3CG gene have profound defects in platelet aggregation [152]. Our finding suggests that SSU and acute patients that are homozygote for the major allele at SNP rs10224307 would have a higher risk of platelet aggregation given the higher level of expression.

A distal eSNP interaction effect implicates expression of the haptoglobin (*HP*) gene, which is responsible for binding free hemoglobin during intravascular hemolysis [153]. Transcription of this gene is elevated in SCD acute patients, and this pattern is dependent on the genotype of the rs742898 SNP located in a predicted enhancer region of the *SERPINA3* gene. Clustering analysis of gene expression profiles have previously shown that HP and SERPINA3 cluster together with other acute phase response genes, suggesting shared regulatory mechanisms [154]. We do not detect a local eSNP association between rs742898 SNP and SERPINA3 gene expression. This is perhaps not surprising, however, since SERPINA3 is expressed predominantly in hepatocytes [155], and not in the tissue (blood) that was sampled. Another explanation for not detecting a significant local eSNP association is that the SERPINA3 gene is alternatively spliced, and the probe used in this study captured the effects from the non-differentially expressed transcript [156].

Interestingly, a common polymorphism in the *HP* gene, characterized by alleles Hp-1 and Hp-2, gives rise to structurally and functionally distinct haptoglobin protein phenotypes, with Hp-2 previously associated with the prevalence of infections, autoimmune diseases, and other disorders [153]. A model described by Quaye [153] proposed a role of HP in the inflammatory response (Figure E.1.2) with the different Hp alleles activating different immune responses (T_H1 and T_H2). We characterized the HP haplotypic structure in the Acute patients (see the Appendix XVI for more details on the method) and demonstrated that the interaction effect is driven by the association of the Hp-2 heterozygote patients with higher levels of HP expression resulting in the eSNP association being significant in the Acute patients who have the Hp-2 allele, but not in those with the Hp-1 allele.

Figure E.1.2 Model of the role of HP in the inflammatory response.

Signals from stressed cells induce expression of IL-6/8 which in turn induce expression of Hp. The strong haemoglobin (Hb) binding, antioxidant, and anti-inflammatory activity of Hp-1 lead to a T_H2 dominant cytokine expression. The corollary holds for the Hp-2 phenotype. ROS: reactive oxygen species.



*Adapted from Quaye, 2008 [153].

The AZIN1 gene was previously associated with risk of liver fibrosis progression in patients with chronic hepatitis C virus (HCV) infection at the polymorphic site rs62522600 [157,158,159]. This AZIN1 SNP leads to enhanced generation of a novel alternative splice form of AZIN1, causing differences in gene expression and progression of liver fibrosis [160]. Interestingly, SCD patients have a higher risk of HCV morbidity [161,162]. Here we show that Acute and SSU SCD

patients with the major allele for SNP rs2460221 have higher gene expression levels for the AZIN1 gene.

In a recent GWAS [163], SNP rs4387287, located close to the OBFC1 gene, was associated with leukocyte telomere length (LTL). In another GWAS, a key gene involved in telomere length (TNKS) was associated with sickle cell anemia severity [164]. Here we give supportive evidence of regulation of OBFC1 gene expression, in Acute SCD patients, as being dependent on genotypic class for the SNP rs7080456.

Three additional genes associated with an interaction effect have been associated with other diseases: NOP56 has been associated with spinocerebellar ataxia [165], PBX1, pre-b cell leukemia homeobox 1, is associated with cancer [166], type 2 diabetes [167,168,169], and congenital heart defects [170], and PTPRA has been associated with Alzheimer disease [171] and schizophrenia [172].

Using eQTL approaches, transcriptional genotype-by-environment interactions have previously been reported in humans [173,174,175,176] but mostly using *in vitro* systems. Here we report a demonstration of the existence of these effects *in vivo* in SCD. The genes implicated in these interactions show differential eSNP effects depending on SCD follow-up status or clinical category. These interactions show how the genetic control of gene expression through allelic variation is likely to impact processes modulating SCD severity, as well as in clinical follow-up programs.

E.2 Implications for public health

The genomic results generated for this project can be applied to potentially improve public health programs and care for SCD patients. Using the gene expression data, we provide candidate biomarkers for clinical progression. Examining

genes that control differential gene expression led to identification of potential drug targets.

E.2.1 Identification of transcriptional biomarkers of SCD severity

Up till now, efforts to identify biomarkers of clinical progression in SCD has had limited success [57,61,62,63,64,177]. Although more than 100 different blood and urine biomarkers have been described in SCD, the clinical value of these biomarkers has not been assessed for clinical validity and utility. Ultimately, the goal is to identify biomarkers that are specific, independent indicators of future risk. New technologies, such as transcriptomics and proteomics, has lead to the discovery of more useful molecules [178] in other diseases. While mRNAs do not play as important a role in cellular functions as proteins, there are a number of reasons why one might prefer doing mRNA expression profiling. The principal reason is that nucleic acids (such as mRNAs) are much easier to separate, purify, detect and quantify than proteins for the purpose of biomarkers. Also since protein concentrations can be considered to be integrals of mRNA concentrations, the variability at the mRNA level is usually larger than the variability at the protein level. Nonetheless, mRNA and protein expression measurements complement each other and thus both types of biomarkers are useful.

Here, using discriminant analysis, we identify transcriptional biomarkers of SCD clinical categories. We identified 19 transcriptional biomarkers of SCD clinical categories in a discovery phase. Using these 19 biomarkers, we were able to classify SCD patients with 80.1% accuracy in the replication phase.

Identifying patients who do not improve in their clinical state even after intensive clinical follow-up is of clinical importance. We performed an additional

discriminant analysis to identify transcriptional biomarkers for SCD patients who were steady-state satisfactory (SSS) or unsatisfactory (SSU). These patients are followed for at least one year before a diagnosis is made on their condition. Predicting which SCD patients will follow an unsatisfactory clinical course could help in order to tailor follow-up programs by providing a pro-active treatment plan for these patients who would be expected to otherwise be unsatisfactory. We identified 3 additional transcriptional biomarkers for this group of patients. These biomarkers warrant further investigation for their diagnostic and prognostic utility in SCD.

E.2.2 Potential drug targets for SCD

Finding a widely available cure for SCD remains a challenge. Bone marrow transplant offers the only potential cure for SCD. But finding a donor is difficult and the procedure has serious risks associated with it, including death. As a result, treatment for SCD is usually aimed at avoiding crises, relieving symptoms and preventing complications. Even under the best of conditions, an individual afflicted with SCD can only expect to live to their mid-forties or early fifties. SCD patients need to visit their physician regularly to check their red blood cell count and monitor their health. Treatments may include medications to reduce pain and prevent complications, blood transfusions and supplemental oxygen.

The only disease-modifying drug that is presently available and that is approved by the FDA is hydroxyurea (HU). When taken daily, HU reduces the frequency of painful crises and may reduce the need for blood transfusions. HU stimulates production of fetal hemoglobin, which helps prevent the formation of sickle cells. However, HU reduces white blood cell counts, red blood cell counts, and platelets, thus increasing the risk of infections, anemia, and hemorhage. Also, there is some concern that long-term use of this drug may cause tumours or leukemia.

Other side effects include nausea, gastro-intestinal upset, skin, nail and hair modifications, infertility, and teratogenic effects. The drug is not beneficial for everyone and is not FDA approved for young children or pregnant woman. Hence, there is an urgent need for more specific and effective treatments for SCD.

The standard approach to developing therapeutics involves testing many thousands of compounds against a known target in order to identify a lead compound. The lead compound can then be further refined in silico and in vitro before heading into the lengthy and costly clinical trials pipeline. This process, which consists of phases I, II, III and IV before final drug approval, involves 10-17 years of drug development from target identification until FDA approval, with only a 10% probability of success [179]. As a result, the pharmaceutical industry spends an average of about 1.2 billion US dollars to bring each new drug to market [180]. There is also a high risk associated with de novo drugs due to unforeseen adverse side effects, as seen in the case of Thalidomide, a drug used to treat morning sickness which resulted in devastating birth defects [181].

A novel approach to therapeutic development is to identify new applications for drugs that have already been approved, or have successfully completed phase I clinical trials [182]. This process of “drug repositioning” aims not to develop drugs de novo, but associate existing therapeutics with new phenotypes. Using our genomics results, we attempted to identify candidate drug targets that could be investigated for repurposing in SCD. Of course, due diligence must be used in all instances of drugs repositioning hypotheses. As pointed out by Wang and Zhang [183], the lack of knowledge on directionality could lead to potential side effects rather than therapeutic benefits. Our eSNP data allowed us to identify genes that are modulated by known drugs, and identify if they are modulated in the correct direction with respect to the

pathology of SCD. In this study, we identify twenty-five drug target genes that are genetically controlled uniquely in the SCD patient group. Among these drug target genes, some have differential gene expression levels between SCD clinical categories that are dependent on SNP genotype. This suggests that not all SCD patients would react in the same way to these medications. Nonetheless, these genes are interesting candidates for drug repurposing in SCD.

For example, CASP4, which we identified as a drug target, is an interesting candidate for SCD because of its link with folic acid. Folic acid is a form of the water-soluble vitamin B9. Its biologically active forms (tetrahydrofolate and others) are essential for nucleotide biosynthesis and homocysteine remethylation. Studies performed by Novakovic *et al.* [184] showed that when folic acid was restricted, CASP4 had increased gene expression levels. We showed that SCD patients in an acute or steady-state unsatisfactory condition and who have the minor allele for SNP rs1712840 have higher gene expression levels of CASP4 mRNA. Our results suggest that acute and SSU SCD patients with the minor allele for SNP rs1712840 would potentially benefit from folic acid supplementation. By supplementing them with folic acid, these patients might have their CASP4 gene expression levels reduced to similar, “normal”, values as non-acute SCD patients and controls. Of course, before exposing any patient to a drug, further analysis *in vitro* and in animal models would be required to test the effects of folic acid on SCD clinical improvement. Also, protein levels should be examined since post-translational modification might influence the results.

Identifying improved therapies for SCD patients who are not experiencing an acute event yet remain unsatisfactory over the course of a follow-up program (steady-state unsatisfactory patients) is a major concern. By comparing the genes

that are differentially expressed between steady-state satisfactory and unsatisfactory patients with genes in a public pharmacogenetic database, and evaluating which ones are under genetic control, we suggest candidate drug target genes that control differential expression between steady-state satisfactory and unsatisfactory SCD patients. One of the candidate drug target genes is prostaglandin-endoperoxide synthase (PTGS), also known as cyclooxygenase. PTGS is the key enzyme in prostaglandin biosynthesis, and acts both as a dioxygenase and as a peroxidase [185]. There are two isozymes of PTGS: a constitutive PTGS1 and an inducible PTGS2, which differ in their regulation of expression and tissue distribution. This gene encodes the inducible isozyme. It is regulated by specific stimulatory events, suggesting that it is responsible for the prostanoid biosynthesis involved in inflammation and mitogenesis. PTGS2 is targeted by many drugs, including nonsteroidal anti-inflammatory drugs such as acetaminophen and ibuprofen, and morphine. We see in our analysis that SSU patients with the minor allele for rs689470 have increased expression of PTGS2. If any of the potential drugs that target PTGS2 could decrease its expression in patients with the minor allele for rs689470 and who are in SSU condition, it may improve their clinical outcome.

Interestingly, genes that are differentially expressed between SSS and SSU patients and that are drug targets are enriched in vitamin binding properties. Since the late 1980's, under-nutrition has been considered as a serious complication of SCD that should be treated as part of standard clinical care of SCD patients. Growing interest in the nutritional problems of SCD has created a body of literature from researchers seeking nutritional alternatives as a means of decreasing morbidity and improving quality of life among SCD patients. In general, SCD patients are lacking in nutrients for adequate growth and development, despite apparently sufficient dietary

intakes. Although there is still a paucity of data supporting the efficacy of macronutrient supplementation, it is becoming clearer that recommended dietary allowances (RDAs) for the general population are insufficient for sickle cell patients. A similar shortage is likely to be true for micronutrient deficiencies, including recent findings of vitamin D deficiency that may be associated with incomplete ossification and bone disease, which are known complications of SCD. By identifying drug target genes enriched in vitamin binding properties, our results suggest that nutrient supplementation should be investigated for use in SCD clinical care prgrams.

There are many pharmacogenetic databases, each with different characteristics. CTD promotes understanding about the effects of chemicals on human health by integrating data from curated scientific literature to describe chemical interactions with genes and proteins, and associations between diseases and chemicals, and diseases and genes. We compared the drug target genes that we identified in CTD database with drug target genes in another open source database: DrugBank (http://www.drugbank.ca/extractor). The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. The database contains 6825 drug entries including 1541 FDA-approved small molecule drugs, 150 FDA-approved biotech (protein/peptide) drugs, 86 nutraceuticals and 5082 experimental drugs. In this study, we identified that 13 of the 30 eSNP genes that were drug targets identified in CTD, were also drug targets in DrugBank. Interestingly, the drug target genes we identified in DrugBank have multiple drugs that are associated with them (Appendix XIX). These additional drugs need to also be investigated in their repurposing use in SCD.

E.3 Strengths and Limitations

A two-phase genome-wide study was performed employing a case-control design to examine the association of gene expression profiles with SCD clinical severity and its genetic control. Using the joint power of genotyping and gene expression analysis to characterise the genomic architecture of SCD, we increased our chance of successfully capturing associated variants, since genotypes have effects on transcript abundance, on average, one order of magnitude stronger than disease phenotypes [92]. Genome-wide thresholds were used for all analyses to control for false positives. By using gPCA, the cases and controls were confirmed to be of similar ethnic backgrounds, making the problem of population stratification inherent in the case-control design irrelevant. Using a two phase design in our study enabled us to replicate our findings. In the replication phase, all analyses were with (i) the same phenotype, (ii) and the same markers/variables to assess the associations in a second sample. Replication studies generally require investigations in separate populations that are independent of the original study. We recruited individuals that were part of the same large cohort for which the discovery sample was obtained, but stratified by date of sampling. Since selection factors that might have influenced any disparities between the two sample phases were similar and not under the influence of the exposure in this study (i.e. genotypes), it is believed that it is unlikely that any selection factor would significantly influence the validity of these findings. Other biases, if any, are likely to be minimal since (i) the clinical demographic features and characteristics of the patients in either sample were similar; and (ii) the allele frequencies of the SNPs were similar for the control samples in the discovery and in the replication phase. This suggests that both samples were generated from the same source population.

We performed our experiments *in vivo*. Using *in vitro* experiments, environmental factors can be controlled; however, *in vivo* experiments capture the complexity of intercellular signaling and interactions that are present in real life. By sampling diseased and healthy individuals *in vivo*, we captured the complexity of SCD and were able to characterise biological pathways involved in the development of the disease.

Non-genetic exposures that are known risk factors for the disease may confound genetic associations, if: (i) the genotype influences the exposure, or (ii) if the gene being studied is in LD (or in association) with another gene that influences the exposure. Although environmental factors are unlikely to confound genetic associations, it is possible for this to happen if they are associated with disease and genotype. If any of the environmental factors that are associated with SCD severity are also associated with allele frequencies for the SNPs associated with gene expression, then confounding would be a problem. In order to examine the possibility of confounding due to malaria (a possible environmental source of confounding), we compared the list of genes that were under genetic control in both SCD and in malaria [115]. More than 50% of the genes were significant in only the SCD analyses, and not significant in the malaria analyses (Appendix XVII). Importantly, when we examined the parasitemia levels in our SCD sample, only 11 patients had detectable parasitemia levels. Thus, exposure to malaria is not likely to confound our association results. Other non-genetic factors may possibly confound the SNP associations we detected in SCD; however, this is unlikely.

Gene expression data can be influenced by many *in vivo* and *in vitro* factors and counfound results. Multiple environmental and biological factors, such as age, sex, cell count differences, life style, geography, and place of residence, nutrition,

time of sampling, treatment type, and disease status are known to influence differential expression between individuals. In SCD, all of these factors likely impact gene expression since they are risk factors for clinical severity. Gene expression heterogeneity is also caused by systematic bias from sources such as technical variation in microarray manufacturing [186]. We attempted to reduce the differences due to technical variation to the best of our capabilities. Using TEMPUS tubes, blood samples are lysed and stabilised instantly, reducing heterogeneity due to RNA degradation. Sample collection was done in a standardised manner following pre-established protocols. The time of sampling was performed within a 3 hours window (between 9am and noon). Shipment of samples was done on dry ice to ensure limited degradation and a reliable company that allowed us to track the samples was employed. RNA was extracted from samples in a systematic manner and an attempt to randomize the order that this was done was made so as not to create batch effects. We also randomised our cDNA synthesis step, the order the samples were put on the microarray chips, and the scanning procedure. All of these steps were performed to limit the technical variation that can cause confounding in gene expression studies.

We measured the impact of age, sex, genetic principal components (a proxy to ethnicity), cell count differences (RBCs and WBCs), and SCD clinical severity in the gene expression analysis. Nonetheless, many other non-measured variables could influence our results. It is possible to identify these hidden expression artifacts that arise from technical, demographic, genetic and environmental factors through surrogate variable analysis [187]. We used surrogate variable (SV) analysis to identify and estimate the amount of expression heterogeneity in our data and identified 11 significant SVs. However, these SVs did not substantially alter the level

of gene expression differentiation for the clinical status effect after including them in the ANCOVA model. Based on the high correlation (Appendix XX) between the significance levels in ANCOVA models before and after accounting for SVs, the results from the gene expression analysis are believed to be robust and are not significantly influenced by non-accounted for variables.

Although we did not match cases with controls, we did sample cases and controls with similar sex and age distributions. Furthermore, we accounted for age and sex in our models. Cell count differences can also confound gene expression studies if they are associated with both the disease outcome and with gene expression differences. Much debate exists on whether cell count differences should be accounted for in gene expression studies. By including RBC and WBC counts in our regression models, we accounted for cell count differences in this study. Thus, gene expression differences and genetic associations that are independent of the differences in cell counts were identified.

The possibility of selection bias influencing the results was considered. In particular, prevalence-incidence bias, which can occur in studies where asymptomatic, mild, clinically resolved, or fatal cases are inadvertently excluded from the case group because the selected cases were examined some time after the disease process has already begun (i.e. looking at prevalent versus incident cases) [188], was considered. It is possible that some patients who had SCD were missed, either because they died from the disease before being sampled, or because they were mild cases and not followed at the Center. If the distribution of the alleles in these cases were different from those that took part in the study, selection bias would influence the results. Since there is still no newborn screening for SCD in Benin, it is

possible that very severe cases were missed. However, since individuals suspected of having SCD are referred to the Center due to it being the only SCD Center in the country, and since clinical symptoms do not generally appear until about 2 years of age because infants are protected by the production of HbF, severe SCD are not believed to be missing from our sample. Since all SCD patients eventually undergo crises, even those with mild clinical symptoms will eventually require treatment or be seen in clinic. Thus, these mild patients are most likely included in our sample as well.

Information bias, i.e. genotyping and transcription profiling accuracy, needs to be considered for all genetic and genomic studies, including this one. In using Illumina's chip assays, only high quality data was retained. The accuracy for Illumina's OmniExpress and HT-12 Bead Chips are very high (>99.8% [189], >99.6 [125], respectively). As genotyping was carried out blinded to the case/control status of the subject, misclassification bias if any, was also likely to be minimal.

There is a possibility that misclassification occurred when SCD patients were assigned a clinical category. Patients were assigned a clinical category by an experienced hematologist who based his decision on a patient's clinical course, complete blood counts and hematological analysis, weight gain, and overall well-being after several months of follow-up. It is possible that misclassification occurred if any of the methods used to generate the data used for the clinician's decision were inaccurate. If this were to happen, bias would be introduced in our results. If the extent of the misclassification were different between the cases and controls, or if it occurred more often in one clinical category as compared to another, than differential misclassification would occur, and could lead to under- or over-estimation of the true magnitude of the measure of association. If the degree of misclassification was

uniform between groups, then non-differential misclassification would occur. The non-differential misclassification would result in a dilution of the measure of association and bias the results toward the null value of no association. Since all patients followed the same protocol, if there was misclassification, it was probably non-differential. Also, if other variables that were not taken into account by the clinician affected a patient's clinical category assignment, than error would be introduced into our results.

Genetic factors that are known to influence hemoglobin, such as HbF, might influence our results. However, this is not believed to be an issue since we performed a globin reduction step, removing expressed globin genes, thereby ensuring that signal from other genetic factors that might influence severity beyond the known effects of globin genes could be detected. Also, insignificant population structure and limited genetic differentiation was observed when 541 genotypes from a subset of genes known to influence hemoglobin levels (alpha-globin, G6PD, BCL11A, MYB and HBS1L) were used in PCA and gene-wise Fst analysis performed to estimate the magnitude of genetic differentiation among the SCD clinical categories, or between SCD patients and controls. Thus, confounding due to these genes is unlikely or limited.

Other genetic factors or processes not captured by the Illumina chip might account for variability in gene expression, including, but not limited to SNPs not captured by the chip, gene methylation, alternative splicing, or protein modification. If any of these factors are involved in the clinical heterogeneity observed in SCD patients, our study would not have captured these effects.

Gene-environment interactions are a fundamental component of the genetic architecture of complex traits and disease susceptibility. Interaction effects have been recognized for many years to play an important part in disease etiology. Gene-environment interactions add sources of complexity in the mapping relationship between genotype and phenotype. As such, a need for research strategies that embrace, rather than ignore, this complexity are needed. In the current study, we addressed this necessity by implementing SNP-by-SCD clinical severity terms in our linear regression models to identify statistical transcriptional interaction effects. One of the major limitations for traditional GWAS in identifying interaction effects has been a lack of power. Sample size requirements for GWA GxE studies can be enormous. A useful rule-of-thumb is that detection of an interaction requires at least four times the sample size than for detecting a main effect of comparable magnitude [190]. By using an integrative approach that capitalizes on the advantage of using of gene expression as an endophenotype, we gain sufficient power to identify significant statistical transcriptional interaction effects.

It has been argued that formal statistical tests for GxE interactions are less useful than biological or public health interactions [191]. Statistical tests depend on the trait measurement scale, and may yield little insight even after formal rejection or retention of the statistical model, since multiple (potentially contradictory) biological models can be consistent with the same statistical model for interaction [191]. This may be true, however, as some of the interactions we identified involve genes previously reported to be linked to SCD, it is conceivable that they contribute to the clinical heterogeneity observed in SCD. This needs to be confirmed. These interaction effects should be tested in other SCD populations to validate their contribution to the clinical heterogeneity. For example, the PIK3CG gene should be

evaluated for its contribution in platelet aggregation in unsatisfactory SCD patients who have the minor allele at SNP rs10224307. Acute inflammation should be demonstrated to occur in patients who have elevated HP gene expression, who have the Hp-2 allele, and who have the minor allele of rs742898 SNP in the SERPINA3 gene. Furthermore, the mechanism of HP gene expression regulation by rs742898 SNP in SERPINA3 gene should also be evaluated. SCD patients with liver cirhosis and chronic Hepatitis C infection should be tested for association with AZIN1 gene expression conditional on SNP rs2460221. And finally, telomere lengths should be tested for association with OBFC1 gene expression conditional on SNP rs7080456 alleles in SCD patients with different clinical severities. All of the interactions we have identified should be further investigated as potential markers for identifying targeted programs or treatments for SCD patients.

We suggest potential transcriptional biomarkers of SCD clinical categories. One of the limitations of our analysis is that we do not test these biomarkers at predicting clinical category in patients after follow-up. Ultimately, these biomarkers need to be tested for their accuracy in predicting a patient's clinical category in longitudinal studies. The potential misclassification of a patient's clinical category may limit our results. A more precise method of measuring clinical severity might improve the accuracy in identifying clinically useful biomarkers.

Since we have not included information on patient treatment and nutritional status in the drug target analysis, it is possible that our results are biased. In order to better evaluate the usefulness of these drugs, additional medical information, including patients' medications and their nutritional intake would be needed. Furthermore, and as explained above, *in vitro* and *in vivo* testing on animal models would be needed to validate the proper modulation and evaluate the clinical

usefulness, as well as potential side effects, of a candidate drug before exposing SCD patients to it.

E.4 Public Health Relevance: importance, recommendations, and generalisability.

Our study has provided valuable genomic information to the scientific community on SCD. In order to meet the United Nations goal of reducing under five mortality caused by SCD, newborn screening programs are being integrated into many countries, followed by the need to improve clinical care. By studying the underlying genomic factors associated with SCD severity, we contribute to this effort. Additional studies will need to confirm our results and test the applications of these findings for SCD treatment and care.

Here, we have identified gene expression profiles associated with SCD severity. Utilizing transcriptomic information has improved our understanding of the molecular basis of this disease and has provided candidate transcriptional biomarkers of SCD progression. We have also identified many novel genetic loci that control gene expression traits in SCD, as well as interaction effects. Gene-by-environment (GxE) interactions are worth studying in public health for several reasons [109]. They can shed light on fundamental biological mechanisms. They can also be important for risk prediction and for evaluating the benefit of changes in modifiable environmental exposures or environmental regulations. From a public health perspective, the idea of personalized recommendations and targeted interventions has been questioned, as the overall benefit of small changes at a population level may be larger than that of large changes in high-risk individual [192]. Personalised recommendations, however, may be considered reasonable for cases

when an exposure has a null or negative effect in one genotype group and a protective effect in another genotype group [193]. Personalised recommendations can also be applied in pharmacogenetics, where different patients will react differently to drugs depending on their genetic “make-up”. Thus interactions can help in choosing the best treatment for an individual to optimize therapy based on genetic predisposition [109]. The interactions we have identified should be considered when prescribing medications for SCD patients. Further investigation in novel treatments that take into consideration the drug targets we identified should be performed. Before being put into general practice, there is a need for additional studies to validate our findings, including clinical trials to test the suggested drugs on SCD clinical improvement, as previously stated.

The SCD patients recruited for this study were sampled in order to obtain a representative sample of the clinical heterogeneity of SCD observed at the National SCD Center. Since there is only one SCD Center in Benin, and it is believed that all Benin SCD patients are referred to it (personal communications with MC Rahimy), our findings can be generalised to the pediatric SCD population of Benin. In order to generalise our findings to all SCD patients, additional studies are needed to validate our results in other populations (including adults SCD patients), environmental settings and contexts.

F. CONCLUSIONS AND FUTURE DIRECTIONS

Using a two-stage case-control sampling design, we identified and replicated a strong transcriptional signature for clinical state in SCD patients that implicates core biological pathways involved in the pathobiology of the disease. Identifying specific biological pathways involved in SCD will help make more personalized strategies for diagnosing, treating and preventing SCD possible. By improving our understanding of these pathways, we can open the door to new improvements. We have provided a genome-wide picture of regulatory variation *in vivo* in SCD patients and highlighted genotype-by-clinical severity interaction effects that likely contribute to the clinical heterogeneity observed in SCD patients. Using gene expression data, we suggest potential transcriptional biomarkers for SCD patients. These results further our understanding of the transcriptional events occurring in SCD patients and their genetic regulatory control and may guide future treatment strategies and biomarker development. The genetic and transcriptional markers reported in this study can potentially guide follow-up programs. Furthermore, the markers detected in whole blood, a readily and ethically accessible source of biological material in children, will particularly be useful in populations where the disease is most prevalent.

F.1 Validation of results in different environments and cohorts

It remains important to validate our results in different environments using other SCD cohorts. However, since there is no universal method of measuring SCD severity, a major challenge will be in comparing our results based on SCD clinical categories with other methods of measuring clinical severity.

Recently, we have obtained ethics approval to examine the genomics of SCD in pediatric patients in Montreal. Of particular interest will be cell specific gene expression analyses, which will allow the dissection of specific gene expression profiles according to the isolated cell type. We also propose to use even deeper transcription profiling based on NextGen sequencing of RNA to complement the microarray-based transcription profiling. This Montreal cohort will allow us to test for replication of the results we identified in the African study, but it will also allow us to investigate environmental and geographical factors that impact SCD severity.

F.2 Application of SCD genomics to other disease: malaria

Sickle cell disease carriers (HbAS) are protected against clinical malaria caused by infection with the most virulent species, *Plasmodium falciparum* [194]. Our current knowledge of the molecular mechanisms for this protection is not completely understood. However, a recent *in vitro* study [195] showed that human microRNAs were found to be more abundant in SCD patients and carriers than in individuals with normal HbAA. It was demonstrated that these microRNAs translocate and incorporate into the *P.falciparum* parasite mRNAs, thereby inhibiting translation of mRNA and reducing the parasite's growth [195]. These results warrant further investigation *in vivo* and offer exciting possibilities for a novel malaria treatment. Applying genomic information identified in SCD to other diseases, such as malaria, may offer novel treatments beyond SCD patients.

F.3 Challenges of integrating genomics into African SCD public health programs

The effective development and application of genomics-based interventions to improve public health in developing countries has become a priority that is recognised by the World Health Organization (WHO) [196]. The current trend is for countries in the developing world to collaborate with more developed nations (north-south collaborations). Nonetheless, there is also an increase in the trend toward south-south collaborations, in which developing countries pool their limited resources, help each other and learn from each other's experience. Integrating results from genomic research in SCD public health practice remains challenging. In 2010, the WHO issued a report on SCD that provided specific targets and goals for sub-Saharan African countries to adopt a national strategy for comprehensive care and treatment of this disorder [3]. By 2015, 25% of countries should have a plan, and by 2020, 50% of countries should have a plan to reduce under-5 mortality by 30%. Such bold targets will not be met easily, however. Active north-south and south-south partnerships that prioritize research will be required to help sub-Saharan African countries develop robust sickle cell strategies that can provide diagnosis, management, and treatment of SCD [197]. Active research partnerships can begin with networking, as supported by the new Global Sickle Cell Disease Network [198]. Recent initiatives by the National Institutes of Health, the UK's Medical Research Council, and other academic and non-profit organizations in drug repurposing has great possibilities of translating genomic information in order to find new indications for old drugs and/or failed candidates [199]. Applying drug repurposing to SCD could identify novel treatments at a fraction of the cost than traditional drug discovery studies can offer. Through similar novel initiatives, research projects that investigate and translate genomics studies will help improve the survival and quality of life of disadvantaged children with SCD wherever they live.

G. REFERENCES

1. Weatherall DJ, Clegg JB (2001) Inherited haemoglobin disorders: an increasing global health problem. Bulletin of the World Health Organization 79: 704-712.
2. World Health Organization tWHA (2006) Sickle Cell Anaemia, Report by the Secretariat. Geneva: World Health Organization.
3. Report of the Regional Director WHOROfA (2010) Sickle-cell disease: a strategy for the WHO African Regions. Geneva, Switzerland.
4. Allison AC (2009) Genetic control of resistance to human malaria. Current Opinion in Immunology 21: 499-505.
5. Makis AC, Hatzimichael EC, Stebbing J (2006) The genomics of new drugs in sickle cell disease. Pharmacogenomics 7: 909-917.
6. Holly SP, Parise LV (2011) Big science for small cells: systems approaches for platelets. Curr Drug Targets 12: 1859-1870.
7. Ware RE (2013) Is sickle cell anemia a neglected tropical disease? PLoS Negl Trop Dis 7: e2120.
8. Bernadette Modell MD (2008) Global epidemiology of haemoglobin disorders and derived service indicators. 480-487 p.
9. Davidson N (1976) Letter to Linus Pauling, January 26, 1976.
10. Driss A, Asare KO, Hibbert JM, Gee BE, Adamkiewicz TV, et al. (2009) Sickle Cell Disease in the Post Genomic Era: A Monogenic Disease with a Polygenic Phenotype. Genomics Insights 2009: 23-48.
11. Ballas SK, Lieff S, Benjamin LJ, Dampier CD, Heeney MM, et al. (2010) Definitions of the phenotypic manifestations of sickle cell disease. American Journal of Hematology 85: 6-13.
12. Gladwin MT, Schechter AN, Ognibene FP, Coles WA, Reiter CD, et al. (2003) Divergent nitric oxide bioavailability in men and women with sickle cell disease. Circulation 107: 271-278.
13. Pauling L, Itano HA, et al. (1949) Sickle cell anemia, a molecular disease. Science 109: 443.
14. Bulger M, Bender MA, van Doorninck JH, Wertman B, Farrell CM, et al. (2000) Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters. Proc Natl Acad Sci U S A 97: 14560-14565.
15. Whitelaw E, Tsai SF, Hogben P, Orkin SH (1990) Regulated expression of globin chains and the erythroid transcription factor GATA-1 during erythropoiesis in the developing mouse. Mol Cell Biol 10: 6596-6606.
16. Harmening DM, Zeringer, H., Brugnara, C. (2002) Hemolytic Anemias Intracorpuscular Defects III: The Hemoglobinopathies. ; DM H, editor. Philadelphia: Davis Company. 165-185 p.
17. Peterson KR (2003) Hemoglobin switching: new insights. Curr Opin Hematol 10: 123-129.
18. Herrick JB (2001) Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. 1910. Yale J Biol Med 74: 179-184.
19. Watson J (1948) A study of sickling of young erythrocytes in sickle cell anemia. Blood 3: 465-469.
20. Ahern EJ, Swan AV, Ahern VN (1973) The prevalence of the rarer inherited haemoglobin defects in adult Jamaicans. Br J Haematol 25: 437-444.
21. Rahimy MC, Gangbo A, Ahouignan G, Adjou R, Deguenon C, et al. (2003) Effect of a comprehensive clinical care program on disease course in severely ill children with sickle cell anemia in a sub-Saharan African setting. Blood 102: 834-838.
22. Grosse SD, Odame I, Atrash HK, Amendah DD, Piel FB, et al. (2011) Sickle cell disease in Africa: a neglected cause of early childhood mortality. Am J Prev Med 41: S398-405.
23. Noguchi CT, Schechter AN, Rodgers GP (1993) Sickle cell disease pathophysiology. Baillieres Clin Haematol 6: 57-91.

24. Manodori AB, Matsui NM, Chen JY, Embury SH (1998) Enhanced adherence of sickle erythrocytes to thrombin-treated endothelial cells involves interendothelial cell gap formation. Blood 92: 3445-3454.
25. Gladwin MT (2008) Current and future therapies of sickle cell anemia: an historical perspective. Hematology Am Soc Hematol Educ Program: 176.
26. Vekilov PG (2007) Sickle-cell haemoglobin polymerization: is it the primary pathogenic event of sickle-cell anaemia? Br J Haematol 139: 173-184.
27. Jeffers A, Gladwin MT, Kim-Shapiro DB (2006) Computation of plasma hemoglobin nitric oxide scavenging in hemolytic anemias. Free Radic Biol Med 41: 1557-1565.
28. Steinberg MH, Sebastiani P (2012) Genetic modifiers of sickle cell disease. Am J Hematol 87: 795-803.
29. Ashley-Koch A, Yang Q, Olney RS (2000) Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. Am J Epidemiol 151: 839-845.
30. Nagel RL, Fabry ME, Pagnier J, Zohoun I, Wajcman H, et al. (1985) Hematologically and genetically distinct forms of sickle cell anemia in Africa. The Senegal type and the Benin type. N Engl J Med 312: 880-884.
31. Steinberg MH, Hsu H, Nagel RL, Milner PF, Adams JG, et al. (1995) Gender and haplotype effects upon hematological manifestations of adult sickle cell anemia. Am J Hematol 48: 175-181.
32. Atweh GF, DeSimone J, Sauntharajah Y, Fathallah H, Weinberg RS, et al. (2003) Hemoglobinopathies. Hematology Am Soc Hematol Educ Program: 14-39.
33. Lettre G, Sankaran VG, Bezerra MA, Araujo AS, Uda M, et al. (2008) DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proc Natl Acad Sci U S A 105: 11869-11874.
34. Thein SL (2011) Genetic modifiers of sickle cell disease. Hemoglobin 35: 589-606.
35. Akinsheye I, Alsultan A, Solovieff N, Ngo D, Baldwin CT, et al. (2011) Fetal hemoglobin in sickle cell anemia. Blood 118: 19-27.
36. Sankaran VG, Lettre G, Orkin SH, Hirschhorn JN (2010) Modifier genes in Mendelian disorders: the example of hemoglobin disorders. Ann N Y Acad Sci 1214: 47-56.
37. Steinberg MH, Sebastiani P (2012) Genetic modifiers of sickle cell disease. Am J Hematol.
38. Rahimy MC, Gangbo A, Ahouignan G, Alihonou E (2009) Newborn screening for sickle cell disease in the Republic of Benin. Journal of Clinical Pathology 62: 46-48.
39. Phenopedia H (2013) www.hugenavigator.net/HuGENavigator/startPagePhenoPedia.do.
40. Milton JN, Rooks H, Drasar E, McCabe EL, Baldwin CT, et al. (2013) Genetic determinants of haemolysis in sickle cell anaemia. Br J Haematol 161: 270-278.
41. Milton JN, Sebastiani P, Solovieff N, Hartley SW, Bhatnagar P, et al. (2012) A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. PLoS One 7: e34741.
42. Sebastiani P, Nolan VG, Baldwin CT, Abad-Grau MM, Wang L, et al. (2007) A network model to predict the risk of death in sickle cell disease. Blood 110: 2727-2735.
43. Jison ML, Munson PJ, Barb JJ, Suffredini AF, Talwar S, et al. (2004) Blood mononuclear cell gene expression profiles characterize the oxidant, hemolytic, and inflammatory stress of sickle cell disease. Blood 104: 270-280.
44. Raghavachari N, Xu X, Munson PJ, Gladwin MT (2009) Characterization of whole blood gene expression profiles as a sequel to globin mRNA reduction in patients with sickle cell disease. PLoS ONE 4: e6484.
45. Chen SY, Wang Y, Telen MJ, Chi JT (2008) The genomic analysis of erythrocyte microRNA expression in sickle cell diseases. PLoS ONE 3: e2360.
46. Jain S, Kapetanaki MG, Raghavachari N, Woodhouse K, Yu G, et al. (2013) Expression of regulatory platelet microRNAs in patients with sickle cell disease. PLoS One 8: e60932.
47. Amin BR, Bauersachs RM, Meiselman HJ, Mohandas N, Hebbel RP, et al. (1991) Monozygotic twins with sickle cell anemia and discordant clinical courses: clinical and laboratory studies. Hemoglobin 15: 247-256.

48. Williams TN, Uyoga S, Macharia A, Ndila C, McAuley CF, et al. (2009) Bacteraemia in Kenyan children with sickle-cell anaemia: a retrospective cohort and case-control study. Lancet 374: 1364-1370.
49. Baum KF, Dunn DT, Maude GH, Serjeant GR (1987) The painful crisis of homozygous sickle cell disease. A study of the risk factors. Arch Intern Med 147: 1231-1234.
50. Powars DR, Chan LS (1987) Is sickle cell crisis a valid measure of clinical severity in sickle cell anemia? Prog Clin Biol Res 240: 393-402.
51. Miller ST, Sleeper LA, Pegelow CH, Enos LE, Wang WC, et al. (2000) Prediction of adverse outcomes in children with sickle cell disease. N Engl J Med 342: 83-89.
52. Olatunji PO, Davies SC (2000) The predictive value of white cell count in assessing clinical severity of sickle cell anaemia in Afro-Caribbeans patients. Afr J Med Med Sci 29: 27-30.
53. Hebbel RP, Moldow CF, Steinberg MH (1981) Modulation of erythrocyte-endothelial interactions and the vasocclusive severity of sickling disorders. Blood 58: 947-952.
54. el-Hazmi MA, Bahakim HM, Warsy AS (1992) DNA polymorphism in the beta-globin gene cluster in Saudi Arabs: relation to severity of sickle cell anaemia. Acta Haematol 88: 61-66.
55. el-Hazmi MA, Bahakim HM, Warsy AS, al-Momen A, al-Wazzan A, et al. (1993) Does G gamma/A gamma ratio and Hb F level influence the severity of sickle cell anaemia. Mol Cell Biochem 124: 17-22.
56. Sebastiani P Sickle cell disease severity calculator.
57. Rees DC, Gibson JS (2012) Biomarkers in sickle cell disease. Br J Haematol 156: 433-445.
58. Berry MP, Graham CM, McNab FW, Xu Z, Bloch SA, et al. (2010) An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 466: 973-977.
59. Ardura MI, Banchereau R, Mejias A, Di Pucchio T, Glaser C, et al. (2009) Enhanced monocyte response and decreased central memory T cells in children with invasive Staphylococcus aureus infections. PLoS One 4: e5446.
60. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, et al. (2007) Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood 109: 2066-2077.
61. van den Tweel XW, van der Lee JH, Heijboer H, Peters M, Fijnvandraat K (2010) Development and validation of a pediatric severity index for sickle cell patients. Am J Hematol 85: 746-751.
62. Setty BN, Key NS, Rao AK, Gayen-Betal S, Krishnan S, et al. (2012) Tissue factor-positive monocytes in children with sickle cell disease: correlation with biomarkers of haemolysis. Br J Haematol 157: 370-380.
63. Qari MH, Dier U, Mousa SA (2012) Biomarkers of inflammation, growth factor, and coagulation activation in patients with sickle cell disease. Clin Appl Thromb Hemost 18: 195-200.
64. Tumblin A, Tailor A, Hoehn GT, Mack AK, Mendelsohn L, et al. (2010) Apolipoprotein A-I and serum amyloid A plasma levels are biomarkers of acute painful episodes in patients with sickle cell disease. Haematologica 95: 1467-1472.
65. Steinberg MH (1999) Management of sickle cell disease. N Engl J Med 340: 1021-1030.
66. Adams RJ, McKie VC, Hsu L, Files B, Vichinsky E, et al. (1998) Prevention of a first stroke by transfusions in children with sickle cell anemia and abnormal results on transcranial Doppler ultrasonography. N Engl J Med 339: 5-11.
67. Wang WC, Kovnar EH, Tonkin IL, Mulhern RK, Langston JW, et al. (1991) High risk of recurrent stroke after discontinuance of five to twelve years of transfusion therapy in patients with sickle cell disease. J Pediatr 118: 377-382.
68. Adams RJ, McKie VC, Brambilla D, Carl E, Gallagher D, et al. (1998) Stroke prevention trial in sickle cell anemia. Control Clin Trials 19: 110-129.
69. Steinberg MH, Barton F, Castro O, Pegelow CH, Ballas SK, et al. (2003) Effect of hydroxyurea on mortality and morbidity in adult sickle cell anemia: risks and benefits up to 9 years of treatment. JAMA 289: 1645-1651.
70. Saleh AW, Jr., Velvis HJ, Gu LH, Hillen HF, Huisman TH (1997) Hydroxyurea therapy in sickle cell anemia patients in Curacao, The Netherlands Antilles. Acta Haematol 98: 125-129.

71. Letvin NL, Linch DC, Beardsley GP, McIntyre KW, Nathan DG (1984) Augmentation of fetal-hemoglobin production in anemic monkeys by hydroxyurea. N Engl J Med 310: 869-873.
72. Platt OS (2008) Hydroxyurea for the treatment of sickle cell anemia. N Engl J Med 358: 1362-1369.
73. Platt OS, Orkin SH, Dover G, Beardsley GP, Miller B, et al. (1984) Hydroxyurea enhances fetal hemoglobin production in sickle cell anemia. J Clin Invest 74: 652-656.
74. Platt OS, Orkin SH, Dover G, Beardsley GP, Miller B, et al. (1984) Hydroxyurea increases fetal hemoglobin production in sickle cell anemia. Trans Assoc Am Physicians 97: 268-274.
75. Goldberg MA, Brugnara C, Dover GJ, Schapira L, Lacroix L, et al. (1992) Hydroxyurea and erythropoietin therapy in sickle cell anemia. Semin Oncol 19: 74-81.
76. Bridges KR, Barabino GD, Brugnara C, Cho MR, Christoph GW, et al. (1996) A multiparameter analysis of sickle erythrocytes in patients undergoing hydroxyurea therapy. Blood 88: 4701-4710.
77. Khoury M. BS, Gwinn M., Higgins J., Ioannidis J., and Little J. (2009) Human Genome Epidemiology; Building the evidence for using genetic information to inprove health and prevent disease.
78. Khoury M. LJ, and Burke W. (2004) Human Genome Epidemiology. New York: Oxford University Press, Inc.
79. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108.
80. Wang TH, Wang HS (2009) A genome-wide association study primer for clinicians. Taiwan J Obstet Gynecol 48: 89-95.
81. Hust M, Jostock T, Menzel C, Voedisch B, Mohr A, et al. (2007) Single chain Fab (scFab) fragment. BMC Biotechnology 7: 14.
82. Goldstein DB, Weale ME (2001) Population genomics: linkage disequilibrium holds the key. Curr Biol 11: R576-579.
83. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A (2003) Linkage disequilibrium patterns of the human genome across populations. Hum Mol Genet 12: 771-776.
84. Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet 3: 611-621.
85. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204.
86. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17: 502-510.
87. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8: e1000294.
88. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-838.
89. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
90. Nica AC, Dermitzakis ET (2008) Using gene expression to investigate the genetic basis of complex disorders. Hum Mol Genet 17: R129-134.
91. Gibson G (2008) The environmental contribution to gene expression profiles. Nat Rev Genet 9: 575-581.
92. Idaghdour Y, Awadalla P (2012) Exploiting gene expression variation to capture gene-environment interactions for disease. Front Genet 3: 228.
93. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20: 1377-1419.
94. Raser JM, O'Shea EK (2005) Noise in gene expression: origins, consequences, and control. Science 309: 2010-2013.
95. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423-428.

96. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, et al. (2009) Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics 2009.
97. Gibson G, Weir B (2005) The quantitative genetics of transcription. Trends Genet 21: 616-623.
98. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17: 388-391.
99. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743-747.
100. Stamatoyannopoulos JA (2004) The genomics of gene expression. Genomics 84: 449-457.
101. Kim J, Gibson G (2010) Insights from GWAS into the quantitative genetics of transcription in humans. Genet Res (Camb) 92: 361-369.
102. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nat Genet 39: 1202-1207.
103. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217-1224.
104. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365-1369.
105. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, et al. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33: 422-425.
106. Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet 32: 261-266.
107. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10: 184-194.
108. Clayton D, McKeigue PM (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358: 1356-1360.
109. Thomas D (2010) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health 31: 21-36.
110. Thomas D (2010) Gene--environment-wide association studies: emerging approaches. Nat Rev Genet 11: 259-272.
111. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL (2007) Genetic properties influencing the evolvability of gene expression. Science 317: 118-121.
112. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, et al. (2006) Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet 2: e222.
113. Sambandan D, Carbone MA, Anholt RR, Mackay TF (2008) Phenotypic plasticity and genotype by environment interaction for olfactory behavior in Drosophila melanogaster. Genetics 179: 1079-1088.
114. Smith EN, Kruglyak L (2008) Gene-environment interaction in yeast gene expression. PLoS Biol 6: e83.
115. Idaghdour Y, Quinlan J, Goulet JP, Berghout J, Gbeha E, et al. (2012) Evidence for additive and interaction effects of host genotype and infection in malaria. Proc Natl Acad Sci U S A 109: 16786-16793.
116. Cai W, Hu L, Foulkes JG (1996) Transcription-modulating drugs: mechanism and selectivity. Curr Opin Biotechnol 7: 608-615.
117. Collins FS (2011) Mining for therapeutic gold. Nat Rev Drug Discov 10: 397.
118. Yan T, Hou B, Yang Y (2009) Correcting for cryptic relatedness by a regression-based genomic control method. BMC Genet 10: 78.
119. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909.
120. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

121. http://www.geohive.com/cntry/benin.aspx.
122. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.
123. Qin S, Jinhee Kim, Dalia Arafat, and Greg Gibson (2012) Effect of normalization on statistical and biological interpretation of gene expression profiles. Fronters in Genetics 3: 1-11.
124. Hansen KD, Irizarry RA, Wu Z (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13: 204-216.
125. Illumina (2013) HumanHT-12 v3 Expression BeadChip Data Sheet: RNA analysis.
126. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
127. Hanash SM, Shapiro DN (1981) Separation of human hemoglobins by ion exchange high performance liquid chromatography. Hemoglobin 5: 165-175.
128. Ross P, Hall L, Smirnov I, Haff L (1998) High level multiplex genotyping by MALDI-TOF mass spectrometry. Nat Biotechnol 16: 1347-1351.
129. Hanchard N, Elzein A, Trafford C, Rockett K, Pinder M, et al. (2007) Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. BMC Genet 8: 52.
130. Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. Am J Hum Genet 88: 173-182.
131. Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86: 526-539.
132. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res 19: 318-326.
133. Ward JH (1963) Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58: 263-244.
134. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.
135. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1: 24-45.
136. Prior M, Eisenmajer R, Leekam S, Wing L, Gould J, et al. (1998) Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. J Child Psychol Psychiatry 39: 893-902.
137. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545-15550.
138. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, et al. (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity 29: 150-164.
139. Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, et al. (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12: 786-795.
140. Benjamini YaH, Y. (1995) Controlling for the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. JR Statis Soc B57: 289-300.
141. .
142. Ihaka RGaR R statistical package.
143. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208.
144. Mattingly CJ, Rosenstein MC, Colby GT, Forrest JN, Jr., Boyer JL (2006) The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. J Exp Zool A Comp Exp Biol 305: 689-692.

145. Horcher M, Souabni A, Busslinger M (2001) Pax5/BSAP maintains the identity of B cells in late B lymphopoiesis. Immunity 14: 779-790.
146. Venkataraman M, Westerman MP (1985) B-cell changes occur in patients with sickle cell anemia. Am J Clin Pathol 84: 153-158.
147. Musa BO, Onyemelukwe GC, Hambolu JO, Mamman AI, Isa AH (2010) Pattern of serum cytokine expression and T-cell subsets in sickle cell disease patients in vaso-occlusive crisis. Clin Vaccine Immunol 17: 602-608.
148. Raghavachari N, Xu X, Harris A, Villagra J, Logun C, et al. (2007) Amplified expression profiling of platelet transcriptome reveals changes in arginine metabolic pathways in patients with sickle cell disease. Circulation 115: 1551-1562.
149. Villagra J, Shiva S, Hunter LA, Machado RF, Gladwin MT, et al. (2007) Platelet activation in patients with sickle disease, hemolysis-associated pulmonary hypertension, and nitric oxide scavenging by cell-free hemoglobin. Blood 110: 2166-2172.
150. Johnson AD, Yanek LR, Chen MH, Faraday N, Larson MG, et al. (2010) Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. Nat Genet 42: 608-613.
151. Soranzo N, Rendon A, Gieger C, Jones CI, Watkins NA, et al. (2009) A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. Blood 113: 3831-3837.
152. Schoenwaelder SM, Ono A, Sturgeon S, Chan SM, Mangin P, et al. (2007) Identification of a unique co-operative phosphoinositide 3-kinase signaling mechanism regulating integrin alpha IIb beta 3 adhesive function in platelets. J Biol Chem 282: 28648-28658.
153. Quaye IK (2008) Haptoglobin, inflammation and disease. Trans R Soc Trop Med Hyg 102: 735-742.
154. Langlais D, Couture C, Balsalobre A, Drouin J (2008) Regulatory network analyses reveal genome-wide potentiation of LIF signaling by glucocorticoids and define an innate cell defense response. PLoS Genet 4: e1000224.
155. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, et al. (2011) Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet 7: e1002078.
156. Huntington JA (2011) Serpin structure, function and dysfunction. J Thromb Haemost 9 Suppl 1: 26-34.
157. Curto TM, Lagier RJ, Lok AS, Everhart JE, Rowland CM, et al. (2011) Predicting cirrhosis and clinical outcomes in patients with advanced chronic hepatitis C with a panel of genetic markers (CRS7). Pharmacogenet Genomics 21: 851-860.
158. Trepo E, Potthoff A, Pradat P, Bakshi R, Young B, et al. (2011) Role of a cirrhosis risk score for the early prediction of fibrosis progression in hepatitis C patients with minimal liver disease. J Hepatol 55: 38-44.
159. do ON, Eurich D, Schmitz P, Schmeding M, Heidenhain C, et al. (2012) A 7-gene signature of the recipient predicts the progression of fibrosis after liver transplantation for hepatitis C virus infection. Liver Transpl 18: 298-304.
160. Paris AJ, Snapir Z, Christopherson CD, Kwok SY, Lee UE, et al. (2011) A polymorphism that delays fibrosis in hepatitis C promotes alternative splicing of AZIN1, reducing fibrogenesis. Hepatology 54: 2198-2207.
161. Neto JP, Lyra IM, Reis MG, Goncalves MS (2011) The association of infection and clinical severity in sickle cell anaemia patients. Trans R Soc Trop Med Hyg 105: 121-126.
162. Nouraie M, Nekhai S, Gordeuk VR (2012) Sickle cell disease is associated with decreased HIV but higher HBV and HCV comorbidities in US hospital discharge records: a cross-sectional study. Sex Transm Infect.
163. Levy D, Neuhausen SL, Hunt SC, Kimura M, Hwang SJ, et al. (2010) Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. Proc Natl Acad Sci U S A 107: 9293-9298.

164. Sebastiani P, Solovieff N, Hartley SW, Milton JN, Riva A, et al. (2010) Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study. American Journal of Hematology 85: 29-35.
165. Kobayashi H, Abe K, Matsuura T, Ikeda Y, Hitomi T, et al. (2011) Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. Am J Hum Genet 89: 121-130.
166. Familiades J, Bousquet M, Lafage-Pochitaloff M, Bene MC, Beldjord K, et al. (2009) PAX5 mutations occur frequently in adult B-cell progenitor acute lymphoblastic leukemia and PAX5 haploinsufficiency is associated with BCR-ABL1 and TCF3-PBX1 fusion genes: a GRAALL study. Leukemia 23: 1989-1998.
167. Thameem F, Wolford JK, Bogardus C, Prochazka M (2001) Analysis of PBX1 as a candidate gene for type 2 diabetes mellitus in Pima Indians. Biochim Biophys Acta 1518: 215-220.
168. Wang H, Chu W, Wang X, Zhang Z, Elbein SC (2005) Evaluation of sequence variants in the pre-B cell leukemia transcription factor 1 gene: a positional and functional candidate for type 2 diabetes and impaired insulin secretion. Mol Genet Metab 86: 384-391.
169. Duesing K, Charpentier G, Marre M, Tichet J, Hercberg S, et al. (2008) Evaluating the association of common PBX1 variants with type 2 diabetes. BMC Med Genet 9: 14.
170. Arrington CB, Dowse BR, Bleyl SB, Bowles NE (2012) Non-synonymous variants in pre-B cell leukemia homeobox (PBX) genes are associated with congenital heart defects. Eur J Med Genet 55: 235-237.
171. Hu CJ, Sung SM, Liu HC, Hsu WC, Lee LS, et al. (2000) Genetic risk factors of sporadic Alzheimer's disease among Chinese in Taiwan. J Neurol Sci 181: 127-131.
172. Takahashi N, Nielsen KS, Aleksic B, Petersen S, Ikeda M, et al. (2011) Loss of function studies in mice and genetic association link receptor protein tyrosine phosphatase alpha to schizophrenia. Biol Psychiatry 70: 626-635.
173. Smirnov DA, Morley M, Shin E, Spielman RS, Cheung VG (2009) Genetic analysis of radiation-induced changes in human gene expression. Nature 459: 587-591.
174. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, et al. (2012) Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A 109: 1204-1209.
175. Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, et al. (2010) Systems genetics analysis of gene-by-environment interactions in human cells. Am J Hum Genet 86: 399-410.
176. Idaghdour Y, Quinlan J, Goulet JP, Berghout J, Gbeha E, et al. (2012) Evidence for additive and interaction effects of host genotype and infection in malaria. Proc Natl Acad Sci U S A.
177. Naprawa JT, Bonsu BK, Goodman DG, Ranalli MA (2005) Serum biomarkers for identifying acute chest syndrome among patients who have sickle cell disease and present to the emergency department. Pediatrics 116: e420-425.
178. Yuditskaya S, Suffredini AF, Kato GJ (2010) The proteome of sickle cell disease: insights from exploratory proteomic profiling. Expert Rev Proteomics 7: 833-848.
179. Ashburn TT, Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3: 673-683.
180. America PRaMo (2013) PhRMA annual membership survey. Washington, DC.
181. Lary J DK, Erickson J, Roberts H, Moore C (1999) The return of thalidomide: can birth defects be prevented? Drug Safety 21: 161-169.
182. Chong CR, Sullivan DJ, Jr. (2007) New uses for old drugs. Nature 448: 645-646.
183. Wang ZY, Zhang HY (2013) Rational drug repositioning by medical genetics. Nat Biotechnol 31: 1080-1082.
184. Novakovic P, Stempak JM, Sohn KJ, Kim YI (2006) Effects of folate deficiency on gene expression in the apoptosis and cancer pathways in colon cancer cells. Carcinogenesis 27: 916-924.
185. Picot D, Loll PJ, Garavito RM (1994) The X-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. Nature 367: 243-249.

186. Akey JM, Biswas S, Leek JT, Storey JD (2007) On the design and analysis of gene expression studies in human populations. Nat Genet 39: 807-808; author reply 808-809.
187. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics 3: 1724-1735.
188. Oleckno WA (2002) Essential Epidemiology, principles and applications. Long Grove, IL: Waveland Press, Inc.
189. Jiang L, Willner D, Danoy P, Xu H, Brown MA (2013) Comparison of the performance of two commercial genome-wide association study genotyping platforms in Han Chinese samples. G3 (Bethesda) 3: 23-29.
190. Smith PG, Day NE (1984) The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 13: 356-365.
191. Kraft P, Hunter D (2005) Integrating epidemiology and genetic association: the challenge of gene-environment interaction. Philos Trans R Soc Lond B Biol Sci 360: 1609-1616.
192. Rose G (2001) Sick individuals and sick populations. Int J Epidemiol 30: 427-432; discussion 433-424.
193. Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, et al. (2008) Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet 16: 1164-1172.
194. Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, et al. (1991) Common west African HLA antigens are associated with protection from severe malaria. Nature 352: 595-600.
195. LaMonte G, Philip N, Reardon J, Lacsina JR, Majoros W, et al. (2012) Translocation of sickle cell erythrocyte microRNAs into Plasmodium falciparum inhibits parasite translation and contributes to malaria resistance. Cell Host Microbe 12: 187-199.
196. WHO WHO (2012) Grand Challenges in Genomics for Public Health in Developing Countries: Top 10 policy and research priorities to harness genomics for the greatest public health problems. Geneva, Switzerland.
197. Weatherall D, Hofman K, Rodgers G, Ruffin J, Hrynkow S (2005) A case for developing North-South partnerships for research in sickle cell disease. Blood 105: 921-923.
198. Global Sickle Cell Disease Network
199. Allarakhia M (2013) Open-source approaches for the repurposing of existing or failed candidate drugs: learning from and applying the lessons across diseases. Drug Des Devel Ther 7: 753-766.
200. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, et al. (2010) Cell type-specific gene expression differences in complex tissues. Nat Methods 7: 287-289.
201. Koch W, Latz W, Eichinger M, Roguin A, Levy AP, et al. (2002) Genotyping of the common haptoglobin Hp 1/2 polymorphism based on PCR. Clin Chem 48: 1377-1382.

H. APPENDICES

Appendix I: Genes associated with SCD

Genes associated with Sickle Cell Anemia that are reported in Human genome epidemiology (HUGE). Genes are ranked by the evidence strength that was calculated based on the volume of different types of published literature in human genome epidemiology (data source: HUGE Literature Finder). The number of publications, genetic associations, and GWAS that were identified for each gene is indicated.

Rank	Gene	Publications	Genetic Associations	GWAS
1	HBB	55	33	0
2	G6PD	11	9	0
3	BCL11A	9	9	3
4	UGT1A1	12	12	1
5	MTHFR	12	9	1
6	F5	9	8	1
7	F2	9	7	1
8	HBA1	9	9	0
9	TNF	5	5	1
10	ITGB3	4	4	1
11	HBS1L	5	5	0
12	HLA-DRB1	5	5	0
13	MYB	5	5	0
14	NOS3	5	5	0
15	GP1BA	3	3	1
16	APOE	3	3	1
17	ITGA2	3	3	1
18	VCAM1	3	3	1
19	HLA-DQB1	4	4	0
20	ACE	2	2	1
21	FGB	2	2	1
22	ITGA2B	2	2	1
23	SERPINE1	2	2	1
24	MBL2	4	3	0
25	ANXA2	3	3	0
26	HFE	3	3	0
27	HP	3	3	0
28	TGFBR3	3	3	0
29	F7	1	1	1
30	F8	1	1	1
31	FGA	1	1	1
32	GOLGB1	1	1	1
33	OR51B5	1	1	1
34	APOA1	1	1	1

35	ICAM1	1	1	1
36	OR51B6	1	1	1
37	ENPP1	1	1	1
38	UGT1A10	1	1	1
39	UGT1A8	1	1	1
40	UGT1A6	1	1	1
41	UGT1A3	1	1	1
42	SELE	1	1	1
43	SELL	1	1	1
44	SELP	1	1	1
45	THBD	1	1	1
46	NPRL3	1	1	1
47	KCNK6	1	1	1
48	CCR5	2	2	0
49	ADRB2	2	2	0
50	CYP2D6	2	2	0
51	F13A1	2	2	0
52	ABO	2	2	0
53	GSTM1	2	2	0
54	GSTT1	2	2	0
55	HBG2	2	2	0
56	HLA-G	2	2	0
57	HMOX1	2	2	0
58	IL4R	2	2	0
59	MYH9	2	2	0
60	NOS2A	2	2	0
61	BMP6	2	2	0
62	APOL1	2	2	0
63	FCGR2A	2	1	0
64	FCGR3B	2	1	0
65	KLF1	1	1	0
66	ADCY9	1	1	0
67	CCR2	1	1	0
68	CYBA	1	1	0
69	CYP2C19	1	1	0
70	AGT	1	1	0
71	EDN1	1	1	0
72	FCGR3A	1	1	0
73	DARC	1	1	0
74	GPM6B	1	1	0
75	GSTP1	1	1	0
76	HBBP1	1	1	0
77	HLA-A	1	1	0
78	HLA-E	1	1	0
79	APOB	1	1	0
80	IFNG	1	1	0
81	IGF1R	1	1	0
82	IL8	1	1	0
83	AQP1	1	1	0
84	ITGAV	1	1	0
85	ARG1	1	1	0
86	ARG2	1	1	0
87	LDLR	1	1	0

88	SMAD3	1	1	0
89	SMAD6	1	1	0
90	NOS1	1	1	0
91	SAR1A	1	1	0
92	RHCE	1	1	0
93	CCL5	1	1	0
94	BMP4	1	1	0
95	SLC4A1	1	1	0
96	SLC14A1	1	1	0
97	BMPR1A	1	1	0
98	BMPR2	1	1	0
99	SOD2	1	1	0
100	TRIM21	1	1	0
101	TEK	1	1	0
102	KLF10	1	1	0
103	UGT2B7	1	1	0
104	VEGFA	1	1	0
105	SLC14A2	1	1	0
106	SELI	1	1	0
107	KL	1	1	0
108	ACVRL1	1	1	0
109	CD40LG	1	1	0
110	TOX	1	1	0
111	TNFRSF1A	1	0	0

Appendix II : Copy of the Consent Form

CHU Sainte-Justine
Le centre hospitalier universitaire mère-enfant

Pour l'amour des enfants

Université de Montréal

FORMULAIRE DE CONSENTEMENT

Titre de l'étude: Intercations Génomiques et Environmentales de l'Anémie Falciforme et du Paludisme en Afrique de l'Ouest.

Personnes responsables: Philip Awadalla, PhD, U. of Montreal, CHU Sainte-Justine

Greg Gibson, PhD, U. of Queensland

Chérif Rahimy, MD, PhD, l'Université d'Abomey-Calavi (UAC) au Bénin.

Ambaliou Sanni, MD, l'Université d'Abomey-Calavi (UAC) au Bénin

Michel Duval, MD, U. of Montreal, CHU Sainte-Justine

Nancy Robitaille, MD,U. of Montreal, CHU Sainte-Justine

Source de financement: Cette recherche est financée par le Centre de recherche du CHU Sainte-Justine, Fonds de la recherche en sante de Quebec (FRSQ), The National Academies of Science and the Keck Foundation, The Human Frontiers in Science Program.

Nom du patient: _____

Votre enfant ou vous-même est atteint d'anémie falciforme (drépanocytose) et /ou malaria. Nous vous proposons de participer à une étude de recherche.

L'éthique médicale et la loi exigent d'obtenir un consentement écrit avant d'entreprendre un procédé de recherche.

L'étude ci-haut mentionnée est proposée par le CHU Sainte-Justine et l'Université d'Abomey-Calavi (UAC) au Bénin. Elle vous a été expliquée par le Dr _____.

Nous vous offrons de participer à un projet de recherche. La participation à l'étude implique un prélèvement de sang réalisé lors d'une visite au centre d'anémie falciforme. Ces projets de recherche sont menés seulement auprès de personnes qui acceptent d'y participer. Prenez le temps nécessaire pour évaluer ce document. Si vous le souhaitez, parlez-en avec votre famille et vos amis avant de prendre votre décision.

Par souci de simplicité, dans le reste du document, le terme "**vous**" doit être compris comme « vous-même ou votre enfant », et le terme "**je**" doit être compris comme « moi-même ou mon enfant ».

1. QUELLE EST LA NATURE DE L'ÉTUDE ?

Pertinence de la recherche

La drépanocytose, ou anémie falciforme, est due à la présence de différentes formes du gène de l'hémoglobine, notamment la forme « S ».

Toutes les personnes atteintes d'anémie falciforme ou drépanocytose, partageant la même forme S, ne sont pas malades de la même façon. Certaines sont très malades, d'autres beaucoup moins. Ces différences sont dues ;

- Soit à d'autres gènes hérités des parents

- Soit à des facteurs d'environnement : lieu de vie, habitudes de vie, etc.

Si on pouvait connaitre la cause de ces différences entres les personnes atteintes d'anémie falciforme ou drépanocytose, on pourrait adapter le traitement à chaque personne et même prévenir certaines complications.

Il est bien connu que le gène de l'anémie falciforme ou drépanocytose confère une résistance au paludisme. Le paludisme est une maladie sévère, très fréquente en Afrique de l'Ouest. Il est causé par un parasite appelé Plasmodium falciparum, transmis par une piqûre de moustique.

Objectifs de la recherche

Notre premier objectif est de comprendre pourquoi il y a des différences de gravité entre deux personnes atteintes de drépanocytose ou anémie falciforme. Nous voulons identifier les déterminants génétiques et les facteurs environnementaux expliquant ces différences.

Notre second objectif est de comprendre les liens entre les symptômes du paludisme, la quantité de parasites dans le sang et les facteurs génétiques et d'environnement.

2. COMMENT SE DÉROULERA L'ÉTUDE ?

Lors de votre visite au Centre d'anémie falciforme à Cotonou, Bénin, il vous sera prélevé 5 ml de sang. Ce prélèvement servira aux études génétiques. Les parents devront répondre à des questions sur l`histoire médicale de l`enfant. Cela nécessitera environ 15 minutes et se fera lors de la même visite au Centre.

Identification des échantillons: La confidentialité des échantillons sera assurée en leur assignant un code spécifique. Ce code permettra de vous lier à

l'échantillon mais le décodage ne pourra se faire que par Chérif Rahimy, l'investigateur principal au Bénin, ou par une personne déléguée par ce dernier. Votre échantillon ne portera pas de nom et ne permettra pas de vous identifier directement.

Autres recherches: En signant ce formulaire, vous nous autorisez à conserver les échantillons non utilisés dans cette étude pour d'autres recherches sur la drépanocytose ou anémie falciforme ainsi que toutes autres maladies associées. Ces recherches seront approuvés par un comité d'éthique et pourraient impliquer l'envoi des échantillons à d'autres chercheurs, même à l'extérieur de notre institution. Les échantillons demeureront codés. Les échantillons seront conservés au Service d'hématologie-oncologie du CHU Sainte-Justine et gardés tout le temps que le service pourra en assurer la bonne gestion.

Confidentialité et communication des résultats: Les résultats personnels des études ne vous seront pas divulgués directement, ni à votre médecin traitant à moins que ceux-ci aient une importance significative pour votre santé. Dans ce cas, les résultats seront communiqués à votre médecin qui vous informera et seront ensuite enregistrés dans votre dossier médical afin d'assurer un meilleur suivi médical. Vous pourrez communiquer avec l'équipe de recherche afin d'obtenir de l'information sur l'avancement des travaux ou les résultats généraux du projet de recherche. Par ailleurs, les résultats de cette étude pourront être publiés ou communiqués dans des congrès ou dans des articles scientifiques mais aucune information pouvant vous identifier ne sera dévoilée. L'équipe de recherche du CHU Sainte-Justine consultera le dossier médical de votre enfant pour obtenir les informations pertinentes à ce projet de recherche. Tous les renseignements obtenus dans le cadre de ce projet de recherche seront confidentiels, à moins d'une autorisation de votre part ou d'une exception de la loi. Les dossiers sous étude seront conservés au CHU Sainte-Justine sous la responsabilité du Dr. Philip Awadalla. La confidentialité de l'ordinateur est soigneusement gardée. Cependant, aux fins de vérifier la saine gestion de la recherche, il est possible qu'un délégué du Comité d'éthique de la recherche du CHU Sainte-Justine ou un représentant des organismes subventionnaires consulte vos données de recherche et votre dossier médical.

3. QUELS SONT LES AVANTAGES ET BÉNÉFICES ?

Il est difficile de prédire si des bénéfices directs résulteront de ce programme de recherche. Nous espérons que les connaissances acquises grâce à cette étude seront utiles dans l'avenir à l'amélioration du diagnostic et du traitement de la drépanocytose ou anémie falciforme et de la malaria.

4. QUELS SONT LES INCONVÉNIENTS ET RISQUES ?

Toute personne donnant de son sang encoure un risque de douleur, de saignement et d'ecchymose (bleu) au site d'introduction de l'aiguille ou un épisode d'étourdissement et d`évanouissement. Les soins appropriés seront pris pour éviter ces complications.

5. QUELS SONT MES DROITS EN TANT QUE PARTICIPANT ?

En signant ce formulaire de consentement, vous ne renoncez à aucun de vos droits prévus par la loi. De plus, vous ne libérez pas les investigateurs de leur responsabilité légale et professionnelle advenant une situation qui vous causerait préjudice.

Pour plus d'information concernant cette recherche, vous pouvez contacter M Chérif Rahimy, l'investigateur responsable au Bénin, au +229 21 30 72 42.

Pour des informations regardant les droits des patients sous programme de recherche, vous pouvez contacter le commissaire local aux plaintes et à la qualité des services du CHU Sainte-Justine au 514-345-4749.

6. Y A-T-IL UNE COMPENSATION PRÉVUE ?

Aucune compensation n'est accordée pour votre participation à cette étude.

Cette étude pourrait contribuer à la création de produits commerciaux, ou à la commercialisation plus large de produits existants, dont vous ne pourrez retirer aucun avantage financier.

7. SUIS-JE LIBRE DE PARTICIPER À CETTE ÉTUDE ?

Vous êtes totalement libre de participer ou non à cette étude. Toute nouvelle connaissance susceptible de remettre en question votre participation vous sera communiquée. Vous êtes libre de vous retirer de ce programme de recherche, à tout moment, sans qu'aucun préjudice ne soit porté aux traitements subséquents. Si vous vous retirez du programme de recherche, vos échantillons sanguins seront détruits.

Un refus de participation n'implique aucune pénalité ou perte de certains bénéfices. Vous êtes libre de recevoir les soins du médecin de votre choix à tout moment. Si vous ne participez pas à l'étude ou si vous vous retirez, cela n'affectera pas la qualité des soins qui vous seront offerts.

8. RENSEIGNEMENTS COMPLÉMENTAIRES

Pour des renseignements plus complets sur la drépanocytose ou anémie falciforme, le Service d'hématologie-oncologie du CHU Sainte-Justine ou de l'Université d'Abomey-Calavi (UAC) au Bénin vous a remis un document d'information et reste à votre disposition pour répondre à toutes vos questions.

FORMULAIRE D'ADHÉSION ET SIGNATURES

Protocole:

On m'a expliqué la nature et le déroulement du projet de recherche. J'ai pris connaissance du formulaire de consentement et on m'en a remis un exemplaire. J'ai eu l'occasion de poser des questions auxquelles on a répondu. Après réflexion, j'accepte que mon enfant participe à ce projet de recherche. J'autorise l'équipe de recherche à consulter le dossier médical de mon enfant pour obtenir les informations pertinentes à ce projet.

Nom du participant Date	Assentiment du participant capable de	
(Lettres moulées)	comprendre la nature du projet (signature)	

Nom du parent, tuteur pour participant de moins de 18 ans Date	Consentement (signature)	
(Lettres moulées)		

Assentiment verbal d'un participant incapable de signer mais capable de comprendre la nature de ce

projet : oui___ non___

J'ai expliqué au participant et/ou à son parent/tuteur tous les aspects pertinents de la recherche et j'ai répondu aux questions qu'ils m'ont posées. Je leur ai indiqué que la participation au projet de recherche est libre et volontaire et que la participation peut être cessée en tout temps.

Nom de la personne qui a obtenu (signature) Date

le consentement (Lettres moulées)

Le projet de recherche doit être décrit au participant et/ou à son parent/tuteur ainsi que les modalités de la participation. Un membre de l'équipe de recherche doit répondre à leurs questions et doit leur expliquer que la participation au projet de recherche est libre et volontaire. L'équipe de recherche s'engage à respecter ce qui a été convenu dans le formulaire de consentement.

Nom du chercheur responsable (signature)
Date

(Lettres moulées)

Personnes ressources:

- Dr Philip Awadalla
- Dr Chérif Rahimy
- Dr Nancy Robitaille

Appendix III : Blood collection procedure

Protocol de Collecte d'Échantillons de Sang pour le Centre de Drépanocytose

Matériel :

Fiche de l'échantillon

Formulaire de consentement

Fiche de données cliniques

Matériel de collecte de sang

Marqueur indélébile

Tube Tempus (Bleu)

Tube EDTA (Violet)

Porte-tubes

Congélateur

Procédure :

1- Remplir le Formulaire de Consentement

2- Remplir la Fiche de l'Échantillon

3- Marquer le **Numéro de l'Échantillon** ainsi que la date et l'heure sur les tubes Tempus (Bleu) et EDTA (Violet).

4- Procédez à la collecte d'échantillon de sang

4- Collectez environ **3 ml** de sang dans le tube Tempus (**Bleu**). Le tiret noir sur le tube indique le niveau 3 ml **Mélangez bien le tube en l'inversant 25 fois** Mettez le tube dans le porte-tubes.	

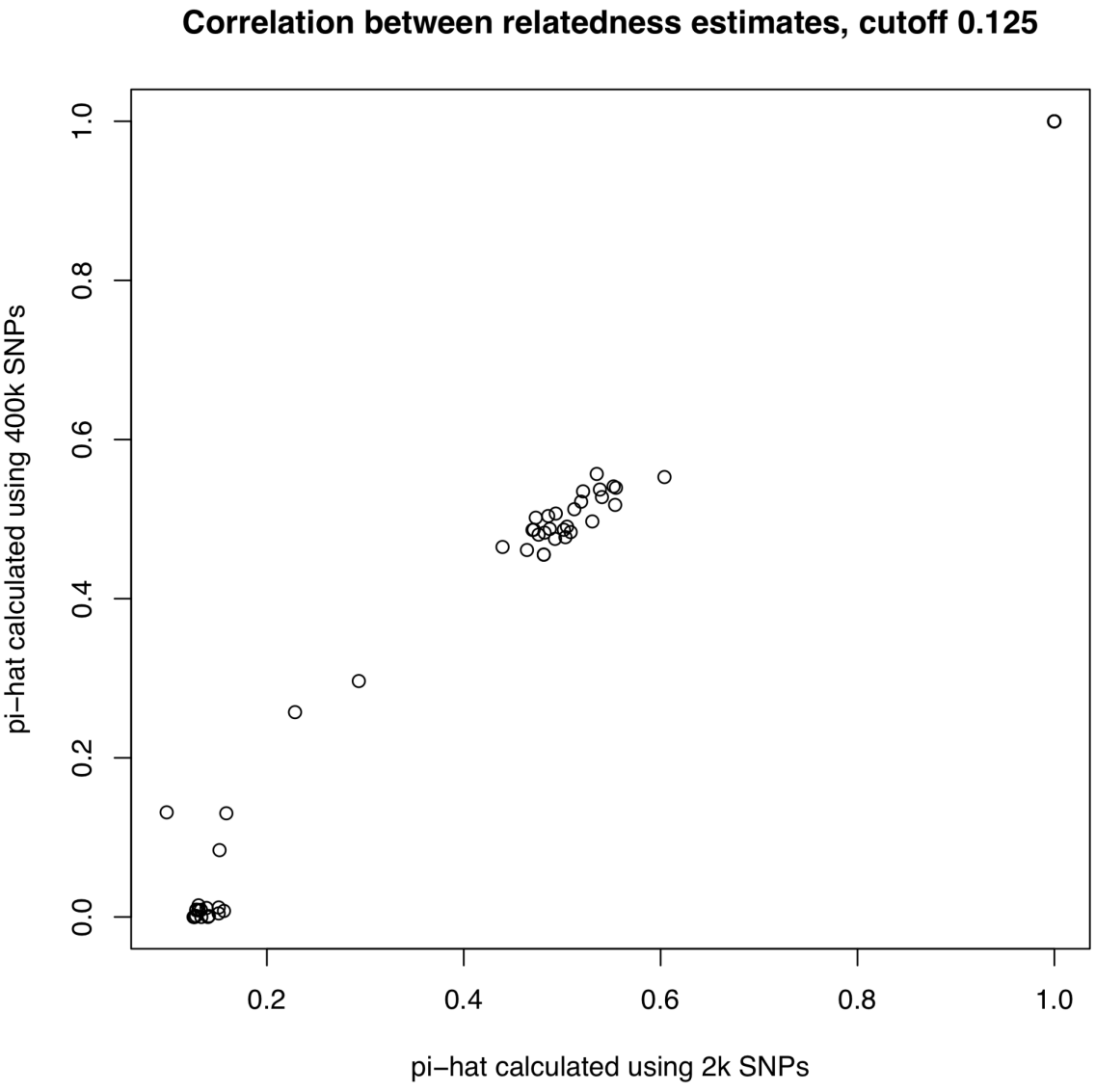
5- Collectez **1 ml** de sang additionnel dans le tube EDTA (**Violet**). **Mélangez bien le tube en l'inversant 5-10 fois** Mettez le tube dans le porte-tubes EDTA.	

6- **Conserver les 2 tubes dans un congélateur à température -20C ou inferieure**

7- Préparer 3 lames pour faire le test de parasitémie (goutte épaisse et frotti pour chaque lame) à partir du sang d'un tube EDTA avec un capillaire

Appendix IV: Correlation of relatedness estimates using different numbers of SNPs

Relatedness estimates generated using 1,986 SNPs (2k) were compared to estiamtes generated using 327,554 SNPs (400k). Removing unrelated individuals with pi hat values of 0.125 or less (which corresponds to first cousins), gave a correlation of 0.98 between the two estimates. Thus, to infer closely individuals (with a pi-hat of at least 0.125), 1986 genome-wide SNPs is sufficient and can be used in the analyses we performed.

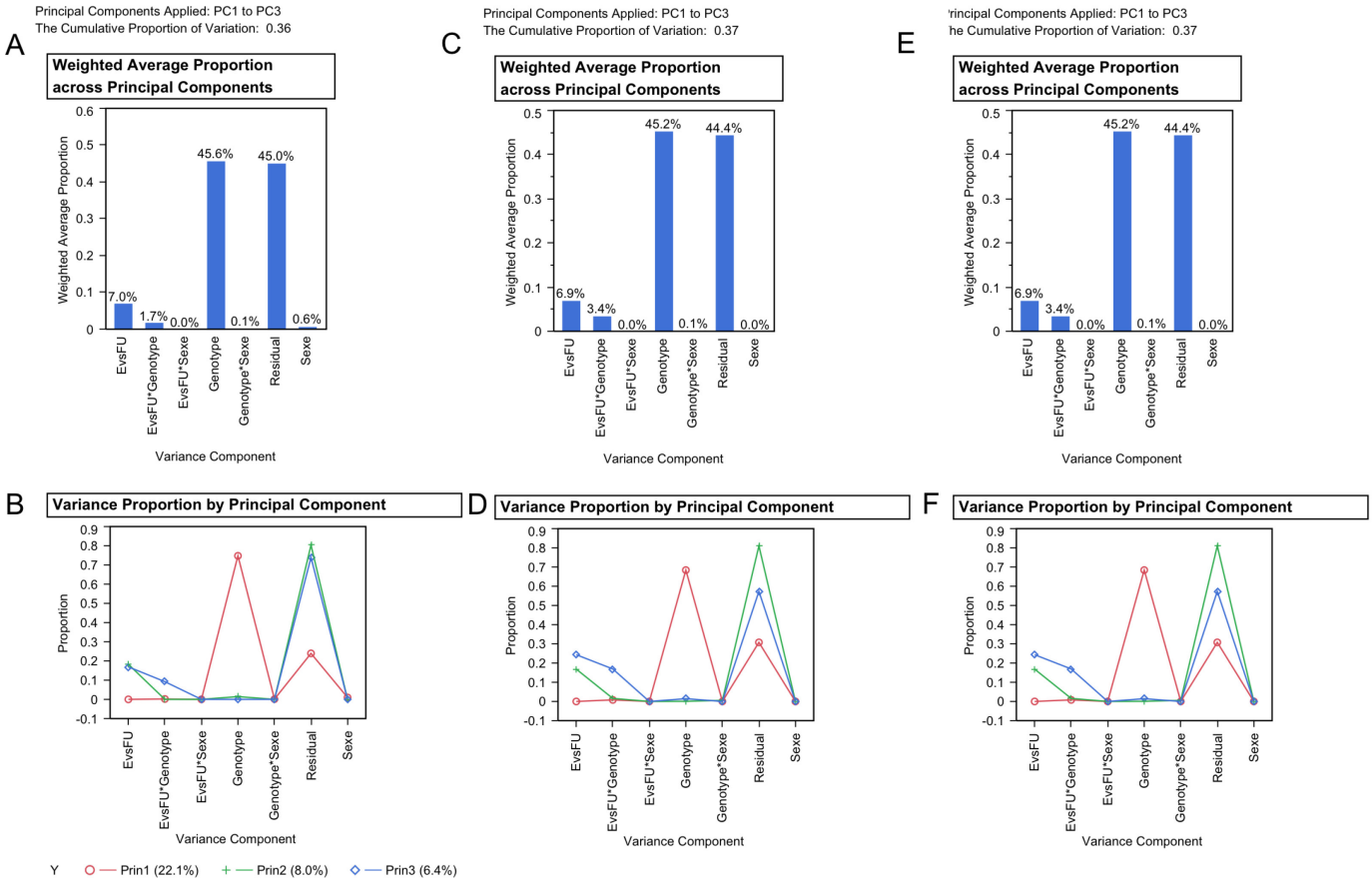


Appendix V: Marker properties of SNPs genotyped using Sequenom

Marker	Position	Map	MAF	Heterozygosity	Call rate
rs7480526	chr11:5247733	37_1	6.22%	9.68%	99.09%
rs334	chr11:5248232	37_1	26.48%	38.36%	100%
rs10742584	chr11:5248770	37_1	0.23%	0.46%	100%
rs968857	chr11:5260458	37_1	4.57%	9.13%	100%
rs10488677	chr11:5261470	37_1	6.42%	10.09%	99.54%
rs11036415	chr11:5262782	37_1	26.50%	41.94%	99.09%
rs2071348	chr11:5264146	37_1	1.14%	2.28%	100%
rs28440105	chr11:5269799	37_1	17.35%	33.79%	100%
rs11827654	chr11:5272521	37_1	22.02%	36.70%	99.54%
rs7482144	chr11:5276169	37_1	1.15%	2.29%	99.54%
rs2855122	chr11:5277236	37_1	16.91%	29.90%	93.15%
rs2855121	chr11:5277291	37_1	1.14%	2.28%	100%
rs10128653	chr11:5277461	37_1	3.00%	5.07%	99.09%
rs11820733	chr11:5281264	37_1	21.92%	36.53%	100%
rs3759069	chr11:5291830	37_1	24.77%	38.53%	99.54%
rs7479652	chr11:5293113	37_1	10.73%	19.63%	100%
rs11036571	chr11:5295644	37_1	0.00%	0.00%	100%
rs7946623	chr11:5298322	37_1	26.03%	39.27%	100%
rs7119428	chr11:5302080	37_1	22.54%	28.17%	97.26%
rs34272388	chr11:5303086	37_1	24.09%	29.53%	88.13%
rs7119142	chr11:5309078	37_1	11.93%	18.35%	99.54%

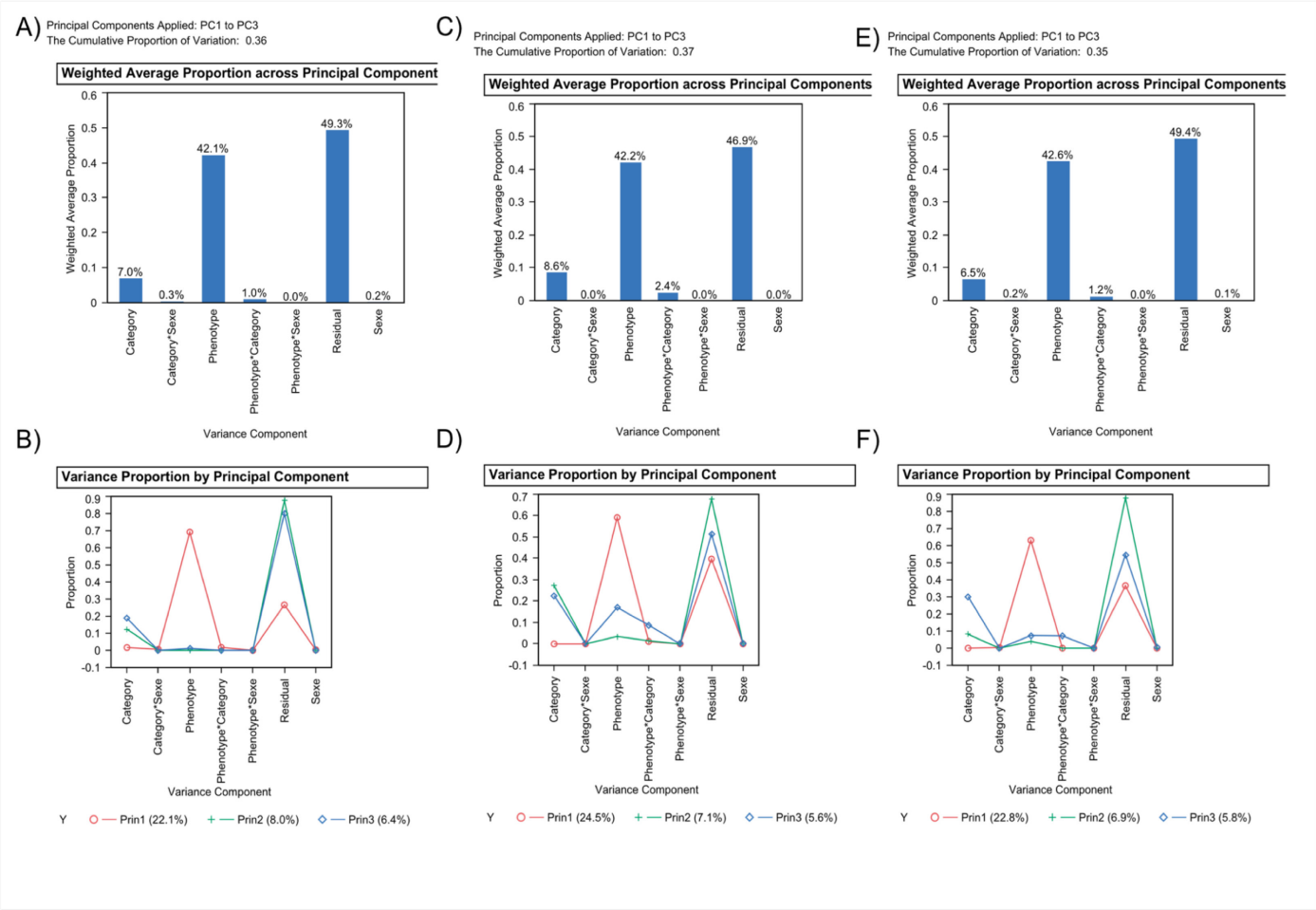
Appendix VI: Additonal VCA results (ClinStatus)

Variance component analysis on the first three expression principal components (ePC1-3) explains over a third of the total variance in the discovery phase (A and B), the replication phase (C and D), and in the combined data set (E and F). ePC1 is primarily explained by Hb genotype (phenotype), while ePC2 and ePC3 are driven by clinical status (EvsFUvsC) and follow-up (EvsFU) in all three datasets.



Appendix VII: Additional VCA results (ClinCat)

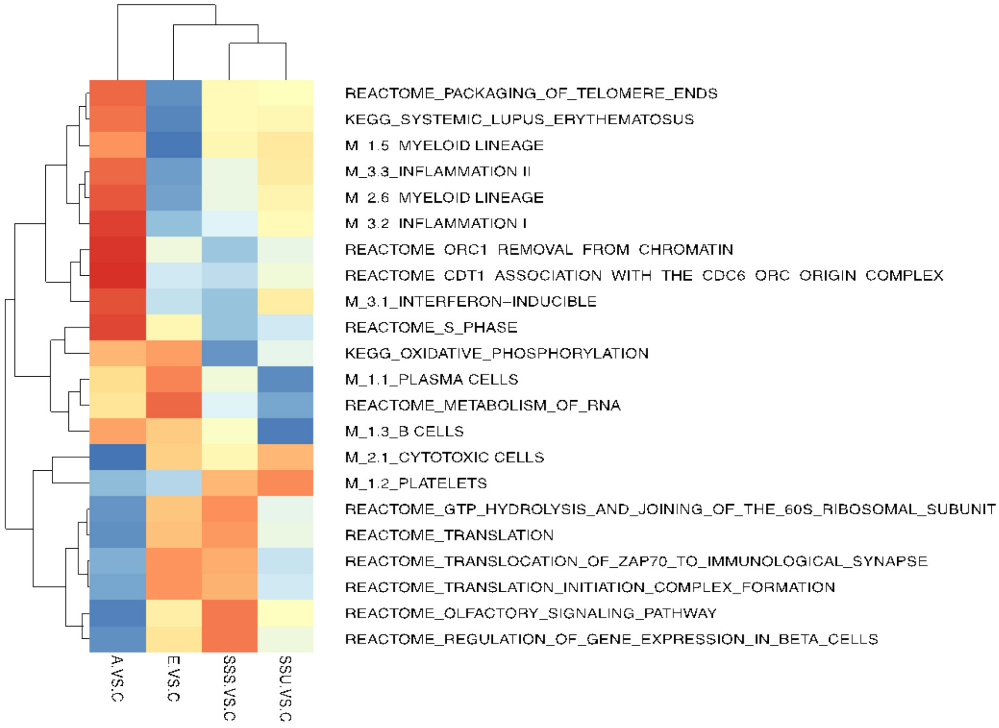
Variance component analysis on the first three expression principal components (ePC1-3) explains over a third of the total variance in the discovery phase (A and B), the replication phase (C and D), and in the combined data set (E and F). ePC1 is primarily explained by Hb genotype (phenotype), while ePC2 and ePC3 are driven by the clinical category effect in all three datasets.



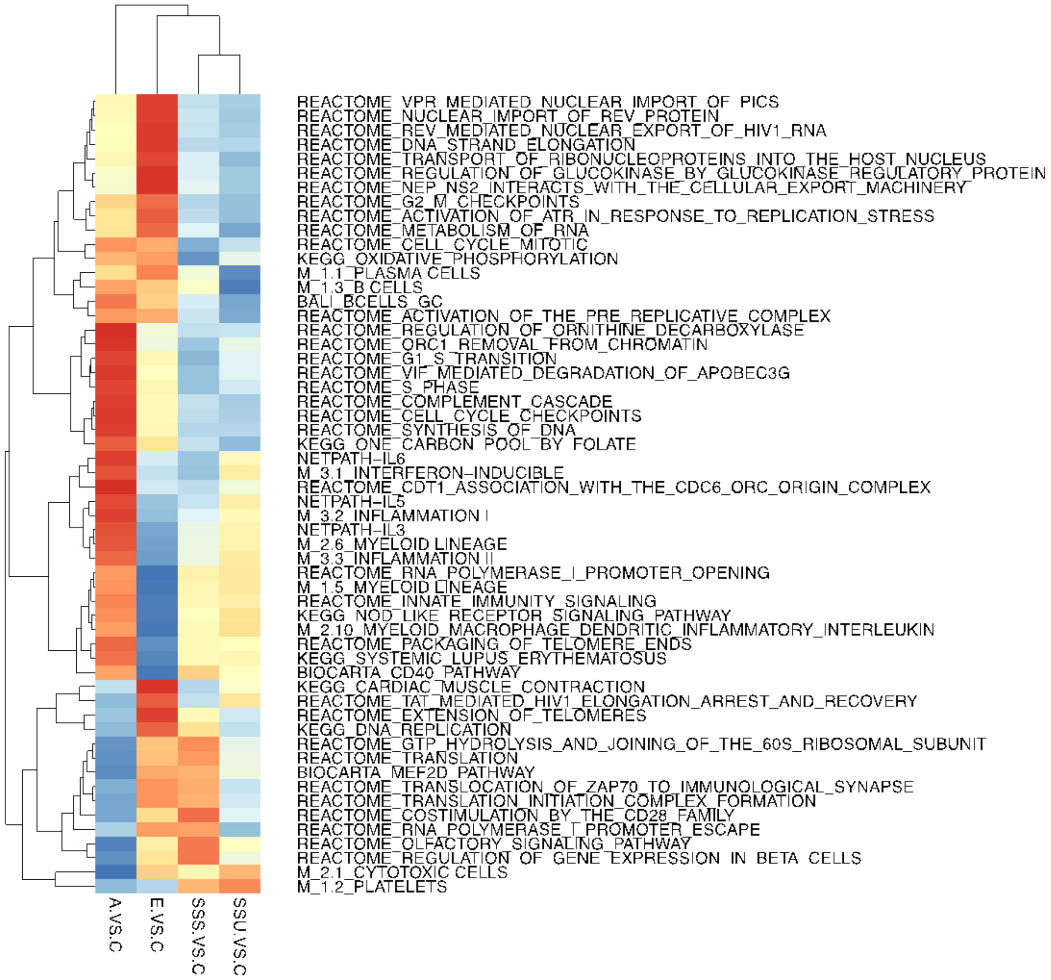
Appendix VIII: GSEA Discovery and Replication phases

Gene Set Enrichment Analysis for the discovery and replication phases for SCD patients grouped accordng to Clinical Categories. Clustered scaled normalized enrichment scores from the GSEA for the discovery and replication phases are shown. Only pathways and modules significantly enriched (Benjamini Hochberg P < 0.05) from at least one contrast are shown. Colors in the heat map indicate the enrichment score relative to the controls.

Discovery Phase

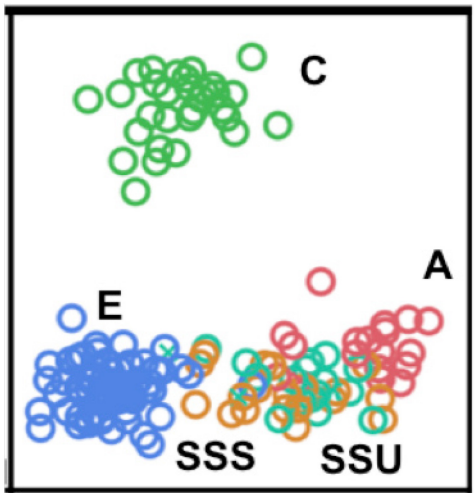


Replication phase



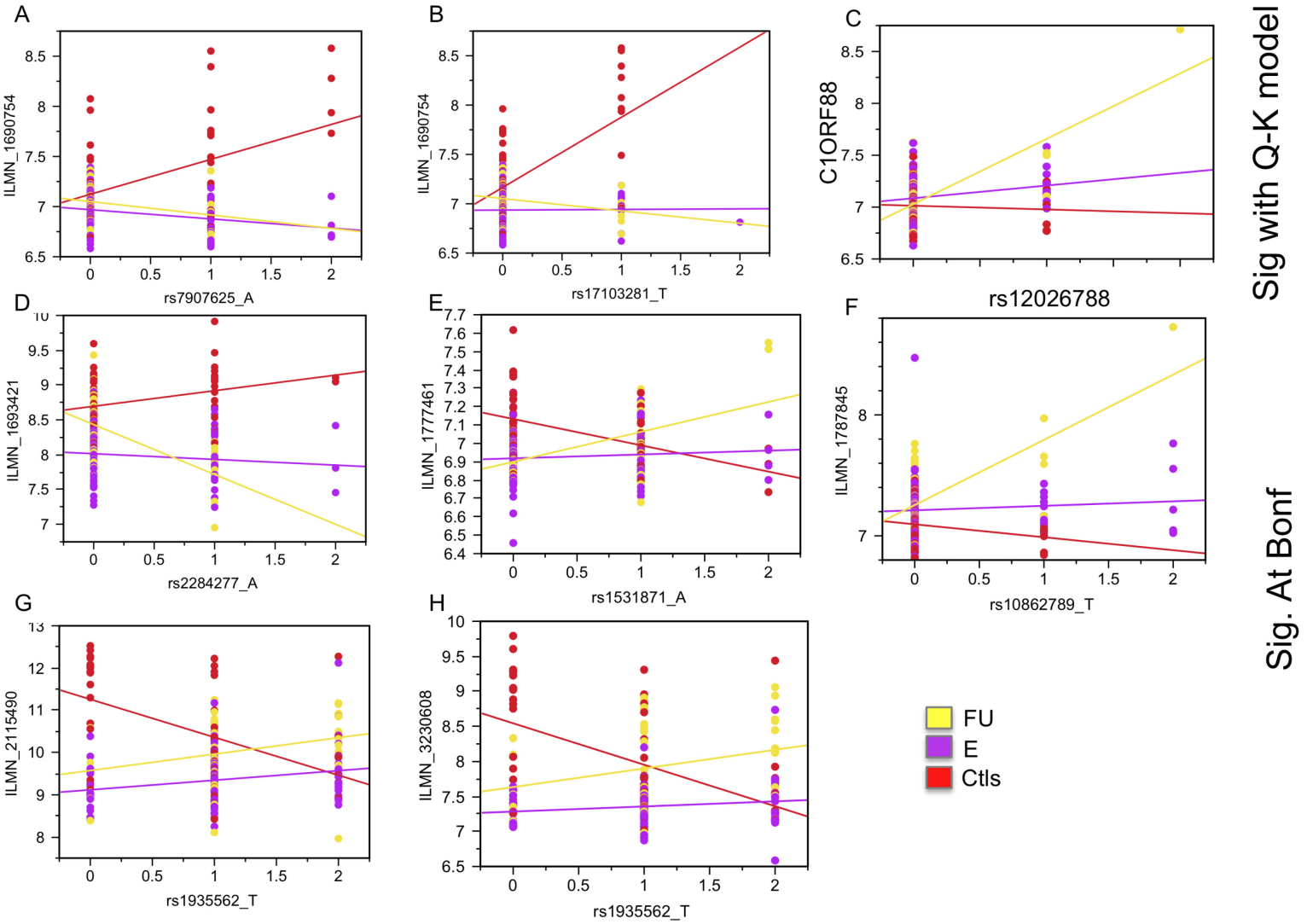
Appendix IX: Canonical values from discriminant analysis

Canonical values for each individual in discovery phase was calculated based on the discriminant analysis.The first 2 canonical values are plotted against each other. Controls (C) are colored in green, entry (E) are blue, acute (A) are red, steady-state satisfactory (SSS) are orange, and steady-state unsatisfactory (SSU) are turquoise. As can be seen, the canonical values for the SSS and SSU patients overlap.



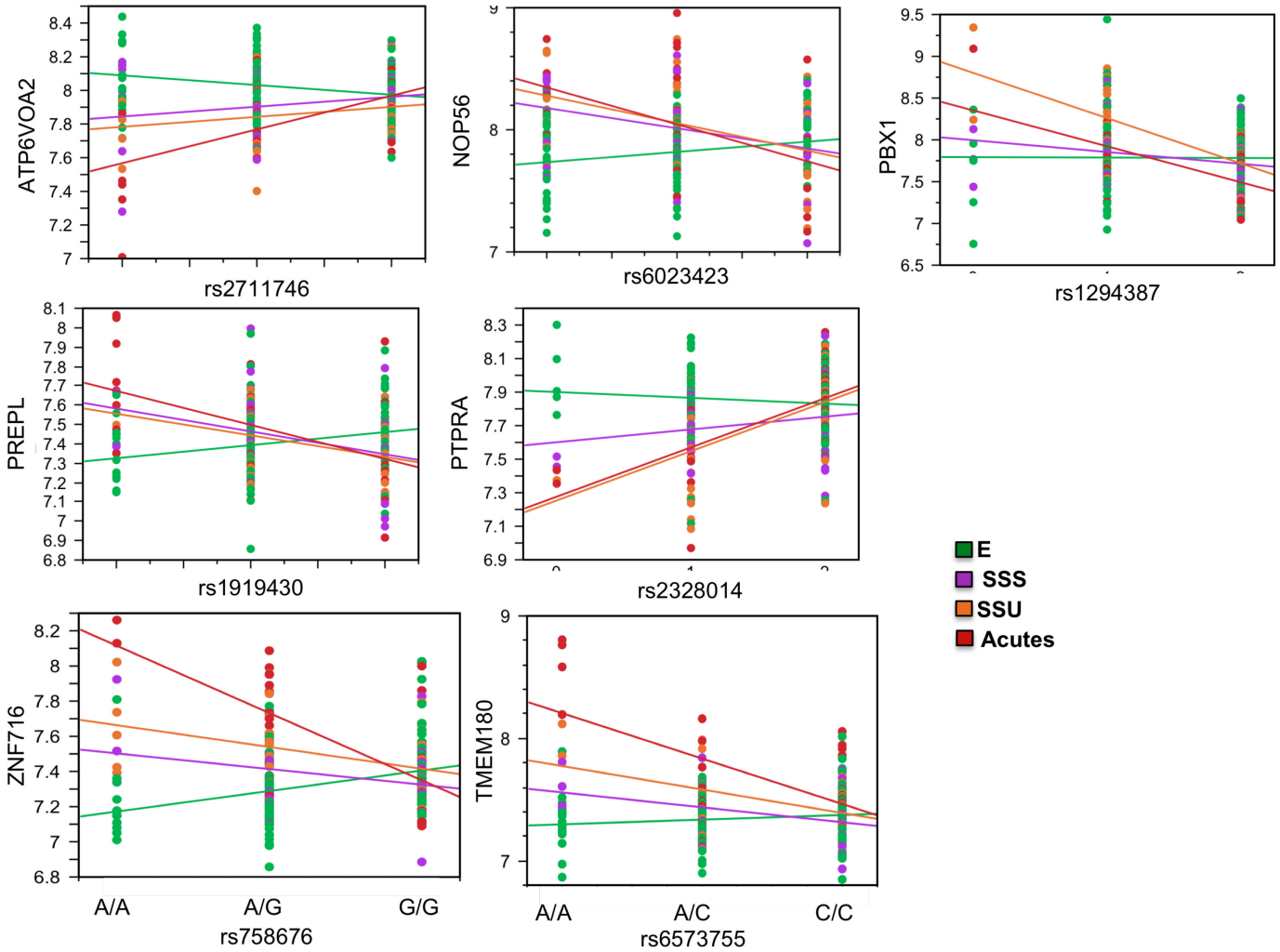
Appendix X : Remaining SNP-by-ClinStatus interactions

Eight remaining SNP-by-ClinStatus interaction effects of genes differentially expressed between SCD clinical status. Using the combined dataset II, a multiple linear regression analysis was performed that accounted for clinical status, sex and cell counts and tested for significant interaction effects for 7002 genes that are differentially expressed between the Clinical Status effect (EvsFUvsCtls). Thirteen peak genome-wide significant interaction effects were identified; eight of which remained significant after running a Q-K mixed model that accounts for relatedness. Five of the 8 interactions are plotted in Fig. D.4.5, and the remainder of interactions are plotted below (A-C are those that remain significant for the Q-K mixed model; D-H the remainder of the 13 interactions). These interaction effects show how the eSNP effect is modulated by clinical status. All interactions are local. The color code for the clinical status is indicated on the right hand side. Expression levels are shown on the y-axis and SNP genotypic class on the x-axis.


A
ILMN_1690754
rs7907625_A
B
ILMN_1690754
rs17103281_T
C
C10RF88
rs12026788
D
ILMN_1693421
rs2284277_A
E
ILMN_1777461
rs1531871_A
F
ILMN_1787845
rs10862789_T
G
ILMN_2115490
rs1935562_T
H
ILMN_3230608
rs1935562_T
FU
E
Ctls
Sig with Q-K model
Sig. At Bonf

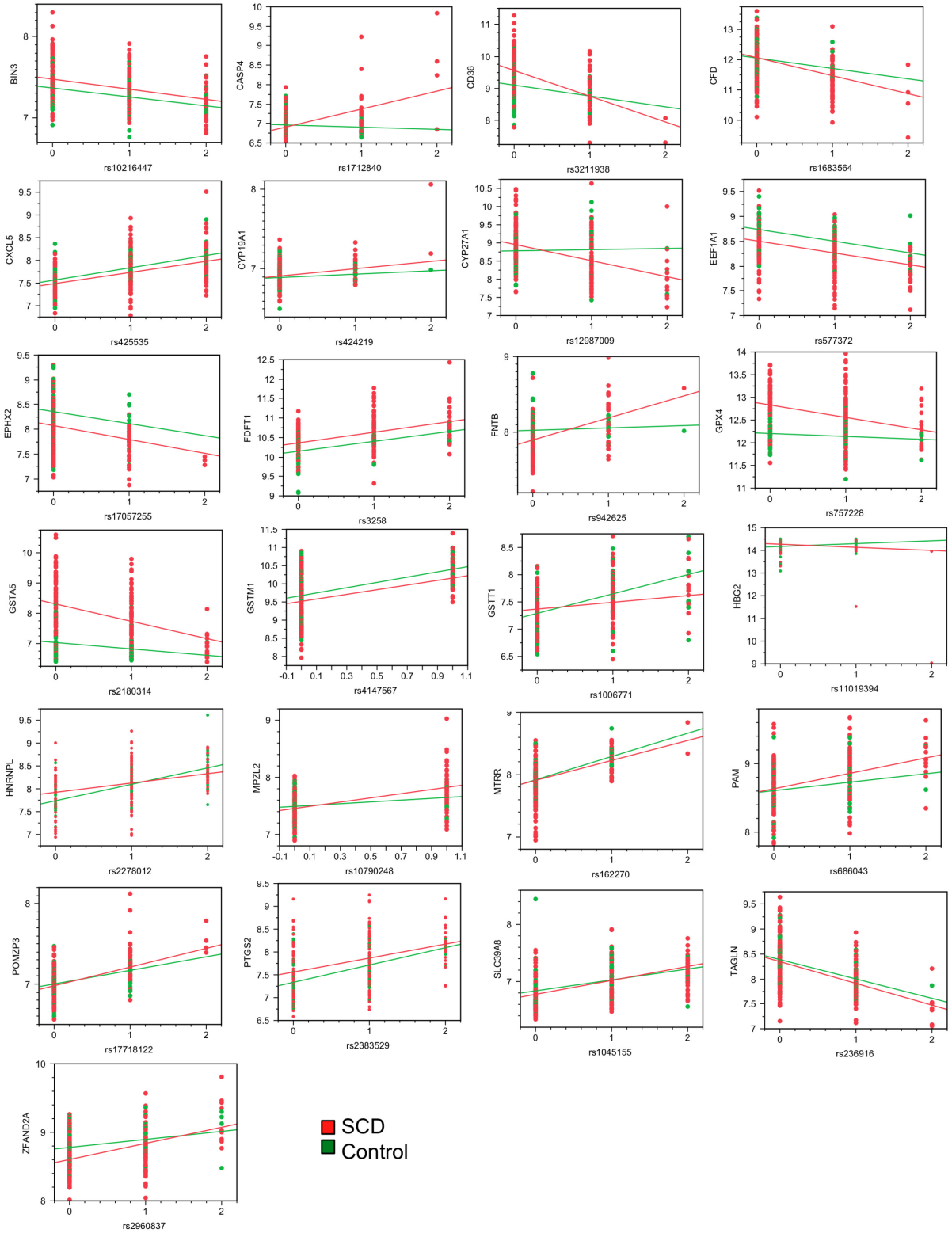
Appendix XI: Remaining SNP-by-ClinCat interactions

Remaining SNP-by-ClinCategory interaction effects of genes differentially expressed between SCD clinical categories. Using the combined dataset, a multiple linear regression analysis was performed that accounted for Hb genotype, clinical category, sex and cell counts and tested for significant SNPxClinical category interaction effects for 4220 genes that are differentially expressed between clinical categories. Eleven peak genome-wide significant SNPxClinical category interaction effects were identified: nine of which are local and two are distal. Four of the 11 interactions are plotted in Fig. D.4.7, and the other seven interactions are plotted below. These seven interaction effects show how the eSNP effect is modulated by SCD clinical state. Six out of the seven interactions plotted here are local, and one (TMEM180 and rs6573755) is distal. The color code for the clinical category is indicated on the left hand side. Expression levels are shown on the y-axis and SNP genotypic class on the x-axis.


ATP6VOA2
rs2711746
NOP56
rs6023423
PBX1
rs1294387
PREPL
rs1919430
PTPRA
rs2328014
ZNF716
A/A
A/G
G/G
rs758676
TMEM180
A/A
A/C
C/C
rs6573755
E
SSS
SSU
Acutes

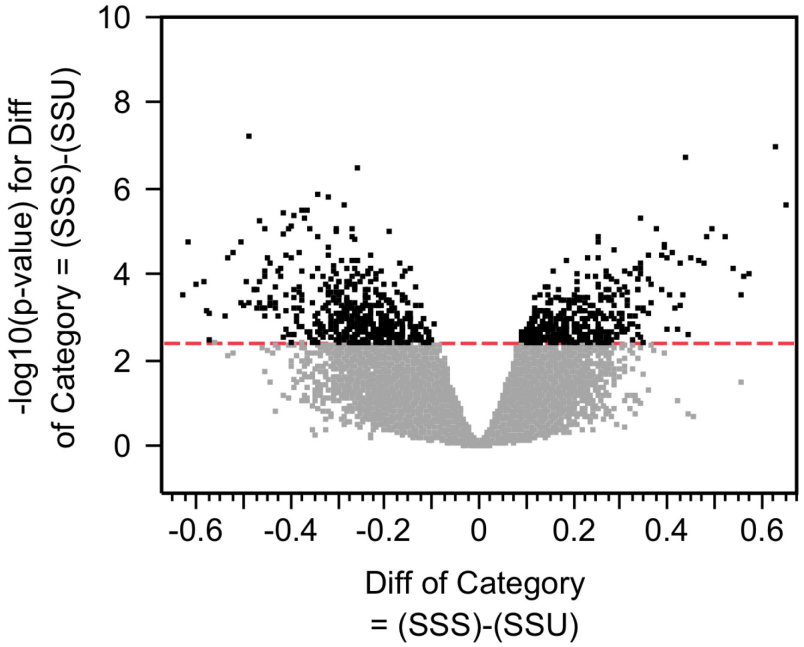
Appendix XII: Plots of 25 eSNP associations that are drug targets

Twenty-five eSNP associations significant in SCD patients and not in controls. The genes for these associations are drug target.



Appendix XIII : Volcano plot of genes differentially expressed b/t SSS and SSU

Volcano plot of the differentially expressed genes between SSS and SSU SCD patients. 824 probes (which corresponds to 775 genes) were differentially expressed (FDR= 5%) between SSS and SSU SCD patients when an ANCOVA was performed that accounted for Hb genotype, clinical category, and sex. The dotted red line indicates the threshold of significance, with all dark dots above this line indicating a probe that was significant. NLP (-log(pvalue)) is plotted on the y-axis, and the fold change value between SSS and SSU patients is plotted on the x-axis.



Appendix XIV : Toppfun results

ToppFun results for enriched GO categories**.** The 86 drug targets that were differentially expressed between SSS and SSU patients were included in the ToppFun enrichment analysis, with all 28,000 expressed probes being the reference. The significant GO categories, pathway name, and p-values are shown below**.**

Category	Name	P-value
GO: Molecular Function	vitamin binding	0.0001
GO: Molecular Function	oxidoreductase activity	0.000361
GO: Molecular Function	ligase activity, forming carbon-sulfur bonds	0.004258
GO: Molecular Function	IgG receptor activity	0.008838
GO: Molecular Function	protein-methionine-R-oxide reductase activity	0.026435
GO: Molecular Function	acetate-CoA ligase activity	0.026435
GO: Molecular Function	transferase activity, transferring pentosyl groups	0.028416
GO: Molecular Function	acid-thiol ligase activity	0.042045
GO: Molecular Function	L-ascorbic acid binding	0.048893
GO: Biological Process	coenzyme biosynthetic process	0
GO: Biological Process	carboxylic acid metabolic process	0
GO: Biological Process	oxoacid metabolic process	0
GO: Biological Process	coenzyme metabolic process	0
GO: Biological Process	organic acid metabolic process	0.000001
GO: Biological Process	cellular ketone metabolic process	0.000001
GO: Biological Process	response to bacterium	0.000001
GO: Biological Process	cofactor biosynthetic process	0.000002
GO: Biological Process	oxidation-reduction process	0.000002
GO: Biological Process	cofactor metabolic process	0.000006
GO: Biological Process	response to organic substance	0.000006
GO: Biological Process	positive regulation of inflammatory response	0.000026
GO: Biological Process	response to lipopolysaccharide	0.000043

GO: Biological Process	response to endogenous stimulus	0.000044
GO: Biological Process	small molecule biosynthetic process	0.000047
GO: Biological Process	response to molecule of bacterial origin	0.000107
GO: Biological Process	cellular nitrogen compound biosynthetic process	0.000108
GO: Biological Process	positive regulation of response to external stimulus	0.000112
GO: Biological Process	acetyl-CoA metabolic process	0.0002
GO: Biological Process	cellular response to chemical stimulus	0.000321
GO: Biological Process	response to other organism	0.000385
GO: Biological Process	response to hormone stimulus	0.000447
GO: Biological Process	regulation of inflammatory response	0.000466
GO: Biological Process	positive regulation of nitric-oxide synthase biosynthetic process	0.000801
GO: Biological Process	inflammatory response	0.000854
GO: Biological Process	response to peptide hormone stimulus	0.001334
GO: Biological Process	defense response	0.00135
GO: Biological Process	response to wounding	0.001562
GO: Biological Process	monocarboxylic acid metabolic process	0.001833
GO: Biological Process	multi-organism process	0.001967
GO: Biological Process	nitric-oxide synthase biosynthetic process	0.001996
GO: Biological Process	regulation of nitric-oxide synthase biosynthetic process	0.001996
GO: Biological Process	acetyl-CoA biosynthetic process	0.002021
GO: Biological Process	response to hypoxia	0.002208
GO: Biological Process	response to oxygen levels	0.003271
GO: Biological Process	homeostatic process	0.003534
GO: Biological Process	response to biotic stimulus	0.003857
GO: Biological Process	positive regulation of defense response	0.003868
GO: Biological Process	response to external stimulus	0.005684
GO: Biological Process	regulation of response to external stimulus	0.007488
GO: Biological Process	xenobiotic metabolic process	0.007853
GO: Biological Process	cellular response to xenobiotic stimulus	0.007853

GO: Biological Process	positive regulation of immune effector process	0.008185
GO: Biological Process	response to xenobiotic stimulus	0.008628
GO: Biological Process	positive regulation of acute inflammatory response	0.012303
GO: Biological Process	lipid biosynthetic process	0.015471
GO: Biological Process	positive regulation of cytokine biosynthetic process	0.016006
GO: Biological Process	positive regulation of tumor necrosis factor biosynthetic process	0.016549
GO: Biological Process	regulation of lipid metabolic process	0.019728
GO: Biological Process	response to steroid hormone stimulus	0.022107
GO: Biological Process	regulation of defense response	0.024549
GO: Biological Process	positive regulation of response to stimulus	0.024586
GO: Biological Process	regulation of chemokine production	0.025446
GO: Biological Process	alcohol metabolic process	0.027733
GO: Biological Process	chemokine production	0.02898
GO: Biological Process	positive regulation of nitric oxide biosynthetic process	0.037114
GO: Biological Process	phosphorylation	0.040278
GO: Biological Process	response to insulin stimulus	0.040292
GO: Biological Process	response to interleukin-15	0.044903
GO: Biological Process	acetate biosynthetic process	0.044903
GO: Biological Process	propionate biosynthetic process	0.044903
GO: Biological Process	cell activation	0.049536

Appendix XV: Cell type specific expression profiles

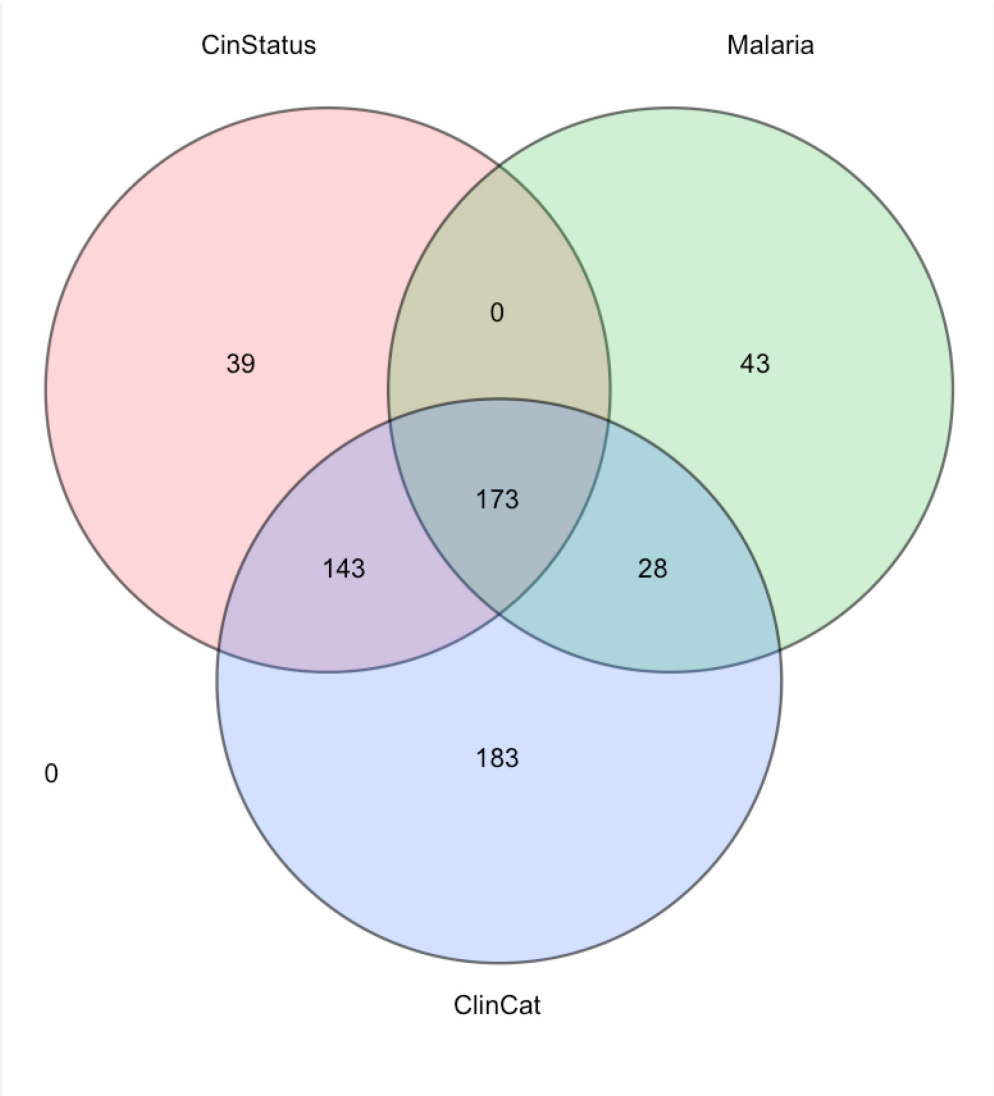
We inferred differences in the proportions of subtypes of peripheral blood mononuclear cells (PBMCs)[200] by obtaining cell type specific expression modules[139] that are constructed based on genomic signatures of flow-cytometry-sorted immune cell types. We calculated the average transcript abundance across all individuals for each module and compared the means in each SCD clinical category to see if they were significantly different. This analysis showed significant differences between SCD clinical categories for all six cell type-specific expression profiles investigated.

Cat Comp	Bali_PBMC_B_cells	Bali_PBMC_mDC	Bali_PBMC_pDC	Bali_PBMC_NK	Bali_PBMC_Tcells	Bali_PBMC_Monocytes	Bali_B_cells_Naive	Bali_Bcells_Plasma	Bali_Bcells_GC	Bali_Bcells_Memory
C - E			<0.0001		0.0155	0.0343	0.0027	<0.0001		
C- SSS			0.006		0.0121			<0.0001	0.0103	
C-SSU	0.0056		0.0311		<0.0001	0.0002	0.0042	<0.0001	<0.0001	<0.0001
C-A			0.0134		<0.0001	<0.0001		0.0036	0.0198	0.0149
E-SSS										
E-SSU	<0.0001				0.0427				<0.0001	<0.0001
E-A	0.0089	0.0172		0.0191	0.0103	0.0118		0.0261		
SSS-SSU	0.0113					0.036			0.0116	0.0045
SSS-A		0.0084				0.0054				
SSU-A										

Appendix XVI: Characterisation of HP alleles.

Characterisation of Hp alleles in acute SCD patients and testing for association with gene expression. Specific primers that differentiate between Hp1 and 2 alleles [201] were used to genotype acute SCD patients in PCR reactions. A multiple linear regression analysis to test for the association between HP gene expression and rs742898 SNP genotype conditioning on Hp allele was performed using JMP Genomics v5.0 (SAS).

Appendix XVII: Venn diagram SCD and Malaria



Appendix XVIII: GEO Accession Numbers

All expression data are available at NCBI Gene Expression Ominbus (GEO) under the series number GSE35007. The individual expression arrays are listed as GSM860207 through GSM860517.

Appendix XIX: Drugs identified in DrugBank that target 14 eSNP genes

Drug Generic Name	Partner Gene Name
3,4-Dichloroisocoumarin	CFD
Isatoic Anhydride	CFD
Aminoglutethimide	CYP19A1
Anastrozole	CYP19A1
Exemestane	CYP19A1
Letrozole	CYP19A1
Testolactone	CYP19A1
Guanosine-5'-Diphosphate	EEF1A1
4-{[(CYCLOHEXYLAMINO)CARBONYL]AMINO}BUTANOIC ACID	EPHX2
6-{[(CYCLOHEXYLAMINO)CARBONYL]AMINO}HEXANOIC ACID	EPHX2
7-{[(CYCLOHEXYLAMINO)CARBONYL]AMINO}HEPTANOIC ACID	EPHX2
N-[(CYCLOHEXYLAMINO)CARBONYL]GLYCINE	EPHX2
N-Cyclohexyl-N'-(4-Iodophenyl)Urea	EPHX2
N-Cyclohexyl-N'-(Propyl)Phenyl Urea	EPHX2
N-Cyclohexyl-N'-Decylurea	EPHX2
(11S)-8-CHLORO-11-[1-(METHYLSULFONYL)PIPERIDIN-4-YL]-6-PIPERAZIN-1-YL-11H-BENZO[5,6]CYCLOHEPTA[1,2-B]PYRIDINE	FNTB
(20S)-19,20,21,22-TETRAHYDRO-19-OXO-5H-18,20-ETHANO-12,14-ETHENO-6,10-METHENO-18H-BENZ[D]IMIDAZO[4,3-K][1,6,9,12]OXATRIAZA-CYCLOOCTADECOSINE-9-CARBONITRILE	FNTB
(20S)-19,20,22,23-TETRAHYDRO-19-OXO-5H,21H-18,20-ETHANO-12,14-ETHENO-6,10-METHENOBENZ[D]IMIDAZO[4,3-L][1,6,9,13]OXATRIAZACYCLONOADECOSINE-9-CARBONITRILE	FNTB
[(3,7,11-TRIMETHYL-DODECA-2,6,10-TRIENYLOXYCARBAMOYL)-METHYL]-PHOSPHONIC ACID	FNTB
2-CHLORO-5-(3-CHLORO-PHENYL)-6-[(4-CYANO-PHENYL)-(3-METHYL-3H-IMIDAZOL-4-YL)- METHOXYMETHYL]-NICOTINONITRILE	FNTB

4-[(5-{[4-(3-CHLOROPHENYL)-3-OXOPIPERAZIN-1-YL]METHYL}-1H-IMIDAZOL-1-YL)METHYL]BENZONITRILE	FNTB
ALPHA-HYDROXYFARNESYLPHOSPHONIC ACID	FNTB
FARNESYL	FNTB
FARNESYL DIPHOSPHATE	FNTB
GERANYLGERANYL DIPHOSPHATE	FNTB
Glycine	GCAT
Pyridoxal Phosphate	GCAT
Glutathione	GPX4
Glutathione	GSTA5
(9r,10r)-9-(S-Glutathionyl)-10-Hydroxy-9,10-Dihydrophenanthrene	GSTM1
(9s,10s)-9-(S-Glutathionyl)-10-Hydroxy-9,10-Dihydrophenanthrene	GSTM1
Fluorotryptophane	GSTM1
Glutathione	GSTM1
Glutathione S-(2,4 Dinitrobenzene)	GSTM1
Zinc Trihydroxide	GSTM1
Glutathione	GSTT1
Cyanocobalamin	MTRR
Hydroxocobalamin	MTRR
L-Methionine	MTRR
Alpha,Beta-Methyleneadenosine-5'-Triphosphate	NMNAT3
Deamido-Nad+	NMNAT3
Nicotinamide Mononucleotide	NMNAT3
Nicotinamide-Adenine-Dinucleotide	NMNAT3
Chymostatin	PAM
N-Alpha-Acetyl-3,5-Diiodotyrosylglycine	PAM
Threonine Derivative	PAM
Vitamin C	PAM
2-methyl-3,5,7,8-tetrahydro-4H-thiopyrano[4,3-d]pyrimidin-4-one	PARP3

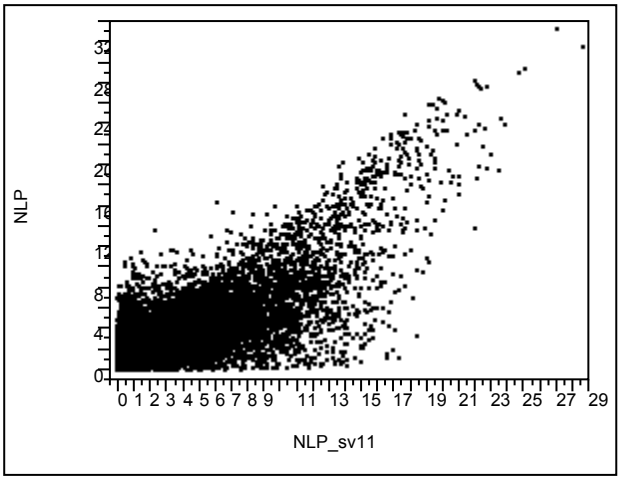
4-[3-(1,4-diazepan-1-ylcarbonyl)-4-fluorobenzyl]phthalazin-1(2H)-one	PARP3
N~2~,N~2~-DIMETHYL-N~1~-(6-OXO-5,6-DIHYDROPHENANTHRIDIN-2-YL)GLYCINAMIDE	PARP3
1-Phenylsulfonamide-3-Trifluoromethyl-5-Parabromophenylpyrazole	PTGS2
1-Phenylsulfonamide-3-Trifluoromethyl-5-Parabromophenylpyrazole	PTGS2
Acetaminophen	PTGS2
Acetylsalicylic acid	PTGS2
Aminosalicylic Acid	PTGS2
Antipyrine	PTGS2
Antrafenine	PTGS2
Balsalazide	PTGS2
Bromfenac	PTGS2
Carprofen	PTGS2
Celecoxib	PTGS2
Diclofenac	PTGS2
Diflunisal	PTGS2
Etodolac	PTGS2
Etoricoxib	PTGS2
Fenoprofen	PTGS2
Flufenamic Acid	PTGS2
Flurbiprofen	PTGS2
gamma-Homolinolenic acid	PTGS2
Ginseng	PTGS2
Heme	PTGS2
Ibuprofen	PTGS2
Icosapent	PTGS2
Indomethacin	PTGS2
Ketoprofen	PTGS2
Ketorolac	PTGS2

Lenalidomide	PTGS2
Lornoxicam	PTGS2
Lumiracoxib	PTGS2
Magnesium salicylate	PTGS2
Meclofenamic acid	PTGS2
Mefenamic acid	PTGS2
Meloxicam	PTGS2
Mesalazine	PTGS2
Nabumetone	PTGS2
Naproxen	PTGS2
Nepafenac	PTGS2
Niflumic Acid	PTGS2
Nimesulide	PTGS2
Oxaprozin	PTGS2
Phenylbutazone	PTGS2
Piroxicam	PTGS2
Pomalidomide	PTGS2
Prostaglandin G2	PTGS2
Resveratrol	PTGS2
Rofecoxib	PTGS2
Salicyclic acid	PTGS2
Salicylate-sodium	PTGS2
Salsalate	PTGS2
Sulfasalazine	PTGS2
Sulindac	PTGS2
Suprofen	PTGS2
Tenoxicam	PTGS2
Thalidomide	PTGS2
Tiaprofenic acid	PTGS2

Tolmetin	PTGS2
Trisalicylate-choline	PTGS2
Valdecoxib	PTGS2
Formic Acid	PVALB

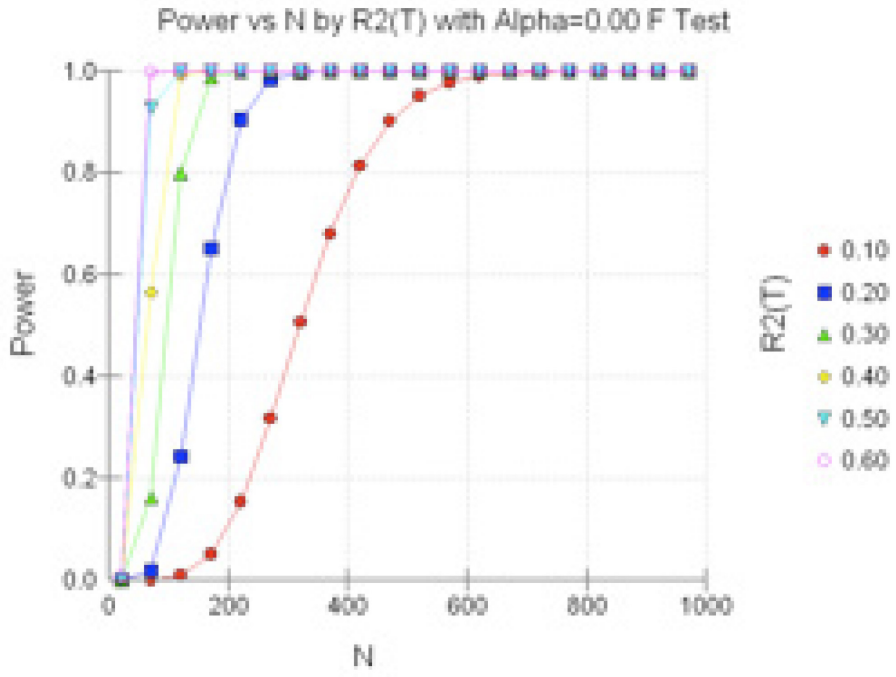
Appendix XX Correlation between ANCOVA results before and after accounting for surrogate variables.

Correlation plot of the significance level (NLP= -log10 (p-value)) for differentially expressed probes for the clinical status effect before and after accounting for 11 surrogate variables. The full ANCOVA model was run (sex, clinical status, RBC, WBC) for the Combined data set II (HbSS non-acutes SCD + controls) with and without the 11 surrogate variables. For each model, the NLP values were extracted for each probe for the clinical status effect and correlated with each other, as seen below. Based on the high correlation, the gene expression analysis is believed to be robust and not significantly influenced by non-accounted for variables.



Appendix XXI Relative power and sample size requirements

Genotype-gene expression association power analysis using SPSS (linear trend). R2(T) are regression coefficients. N is sample size. A sample size of 200 individuals was calculated to be sufficient to identify eSNPs that explain 20% of the variation across the genome with 80% power.



Appendix XXII Contribution, source of funding, and publications

Contribution:

The present study was a collaborative effort. The PhD candidate was involved in all stages of the study, from design, ethics approval, sampling, genomic experiments, analysis, interpretation of results, and publication. The samples were collected in Benin by Elias Gbeha, Selma Gomez, Mohamed Cherif Rahimy, Youssef Idaghdour, and Jacklyn Quinlan. Mohamed Cherif Rahimy followed the SCD patients and oversaw characterisation of SCD patient clinical categories. All haematological analysis was performed at the National Sickle Cell Disease Centrer under Mohamed Cherif Rahimy's direction. Elias Gbeha, Youssef Idaghdour, and Jacklyn Quinlan processed the samples and performed the genomic experiments in Montreal. Vanessa Bruat, Thibault de Malliard, and Jean-Christophe Grenier provided bioinformatics support for statistical analysis of the data by Jacklyn Quinlan. Jean-Philippe Goulet and Jacklyn Quinlan performed enrichment analysis. Interpretation of the results was performed by Jacklyn Quinlan with help from Philip Awadalla and Youssef Idaghdour. Philip Awadalla, Youssef Idaghdour and Jacklyn Quinlan wrote the paper (accepted Feb. 14, 2014 by Frontiers in Genetics).

Source of funding:

The research study operations were funded by a Human Frontiers in Science Program Grant RGP0054/2006-C awarded to Philip Awadalla. Jacklyn Quinlan was supported by doctoral Fellowships from the Fonds de la Recherche en Santé du Québec, the Réseau de médecine génétique appliquée (the Louis-Dallaire fellowship) and by the "Bourse de Fins d'Études" from the Département de Médecine Sociale et Préventive from Université de Montréal.

Publications:

Articles

Quinlan J, Idaghdour Y, Goulet JP, Gbeha E, Bruat V, de Malliard T, Grenier JC, Gomez S, Sanni A, Rahimy MC, Awadalla P. Genomic architecture of Sickle Cell Disease in West African children. Frontiers in Genetics. Accepted.

Idaghdour Y, **Quinlan J**, Goulet JP, Berghout J, Gbeha E, Bruat V, de Malliard T, Grenier JC, Gomez S, Gros P, Rahimy MC, Sanni A, Awadalla P. Evidence foradditive and interaction effects of host genotype and infection in malaria. Proc Natl Acad Sci U S A. 2012 Oct 16;109(42):16786-93.

Myers RA, Casals F, Gauthier J, Hamdan FF, Keebler J, Boyko AR, Bustamante CD, Piton AM, Spiegelman D, Henrion E, Zilversmit M, Hussin J, **Quinlan J**, Yang Y, Lafrenière RG, Griffing AR, Stone EA, Rouleau GA, Awadalla P. A population genetic approach to mapping neurological disorder genes using deep resequencing. PLoS Genet. 2011 Feb;7(2):e1001318.

Abstracts

Quinlan J, *et al.* Genomic Architecture of SCD, RMGA, Mai 2012.

Quinlan J, *et al*. L'Architecture génomique de l'anémie falciforme en Afrique de l'Ouest, XXVIIe congrès de la recherche des étudiants gradués et post-gradués du Centre de recherche du CHU Sainte-Justine, Mai 2012.

Quinlan J, *et al.* Genomic Architecture of Sickle Cell in Children from West Africa. The National Conference on Blood Disorders in Public Health and the Global Sickle Cell Disease Network, GA, March 2012.

Quinlan J, *et al*. Determining the genomic factors of sickle cell disease severity among West African children. American Society of Human Genetics, Montreal, QC. October, 2011.

Quinlan J, *et al*. Genetic Architecture of sickle cell disease in Benin, West Africa. Biology of Genomes. Cold Spring Harbor, NY, May, 2011.

Quinlan J. *et al.* L'architecture génomique de l'anemie falciforme en Afrique de l'Ouest, Colloque des étudiants en Santé Publique, Montréal, QC, Feb 2011.

Quinlan J, *et al.* L'Architecture génomique de l'anémie falciforme et du paludisme en Afrique de l'Ouest. Centre de recherche du CHU Sainte-Justine, May 26, 2010 (First prize).

Quinlan J, Awadalla P. L'anémie falciforme et le paludisme au Benin, Hopital Notre Dame,Montreal, QC May 18, 2010.

Quinlan J, *et al*. Genetic and environmental architecture of sickle cell disease and malaria in Benin. 5th Annual Canadian Genetic Epidemiology and Statistical Genetics Meeting. King City, ON. 2010.