This electronic thesis or dissertation has been downloaded from the King's Research Portal at https://kclpure.kcl.ac.uk/portal/



Genotype-Phenotype Correlation in Sickle Cell Disease

Gardner, Catherine Joanne

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT



Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. https://creativecommons.org/licenses/by-nc-nd/4.0/

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. Oct. 2023

Genotype-Phenotype Correlation in Sickle Cell Disease

Kate Gardner

Thesis submission for the degree of
Doctor of Philosophy

Student Number 1416607

Supervised by
Professor Swee Lay Thein and Dr Stephan Menzel

Contents

Acknowledgements		3
Abstract		4
Ethics		4
Abbreviations		5
Chapter 1	Introduction	7
Chapter 2	Phenotyping in Sickle Cell Disease	26
Chapter 3	Genotyping: creating a genome-wide dataset of common genetic variants	68
Chapter 4	Genome-Wide Association Studies in Sickle Cell Disease	127
Chapter 5	Candidate Gene Studies in Sickle Cell Disease using Two Severity Indices	197
Chapter 6	KLF1 as a Candidate Gene for Fetal Haemoglobin Levels in Sickle Cell Disease	229
Chapter 7	Conclusions	253

Acknowledgements

This work was made possible by the help and support of many individuals.

First, a big thanks to all the patients with sickle cell disease who donated their time and blood samples, without which this work would not have been possible.

I am very grateful to both of my supervisors. Swee Lay Thein has mentored me for 10 years and it has been a privilege to be her PhD student. Stephan Menzel has educated me in genetics and focused my mind during writing-up. Both have dedicated inordinate amounts of personal time to me over the last four years, for which I am very thankful. I hope to have an ongoing research relationship with them both. Amandine Breton and Helen Rooks from the KCL Red Cell research team must both be acknowledged for their support and advice in the laboratory.

I must also thank multiple other individuals who gave their time and support to specific projects. Hamel Patel and Stephen Newhouse (the Social, Genetic and Developmental Centre, Institute of Psychiatry, King's College London) for doing the genetic variant calling and offering (extensive) bioinformatics advice (chapter 3). Charles Curtis, for his work on the MEGA chip (the Social, Genetic and Developmental Centre, Institute of Psychiatry, King's College London). Paul O'Reilly (the Social, Genetic and Developmental Centre, Institute of Psychiatry, King's College London) must be thanked for his statistical genetics guidance. I also acknowledge the help of Clive Stringer (ex-Deputy ICT Director and System Delivery Manager at King's College Hospital) for getting raw clinical data for the King's College Hospital sickle cell patients (chapter 2). Tony Fulford (Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ) must be thanked for his statistical input into the HbF modelling study (chapter 4).

Finally, this PhD would not have been possible without my partner Melissa. She has provided boundless support and kept me in excellent supply of karaage chicken.

Abstract

Sickle cell disease (SCD) has a complex pathophysiology initiated by the polymerisation of deoxy-sickle-haemoglobin. The single nucleotide change underpinning SCD does not account for the vast range and severity of SCD complications. This clinical heterogeneity is only partly explained by the genetic variability of fetal haemoglobin gene levels and co-inheritance of α-thalassemia. Although environmental factors also contribute to the clinical complexity of SCD, further genetic modifiers of SCD severity exist but are yet to be determined.

Genetic association studies have been boosted recently not only with the advent of new genotyping tools, but also with the development of increasingly sophisticated analytical methods. New developments in phenotyping, genotyping and genotype-phenotype association approaches allow us to disentangle true genetic associations from hits due to chance. This thesis seeks to investigate biomarkers of sickle severity and to use these clinical markers in genotype-phenotype correlation studies.

I have investigated three key markers of disease severity: haemolysis, frequency of acute pain episodes and mortality. Estimated median survival of 67 years in HbSS disease in our UK cohort is a significant improvement in survival compared to other recent estimates in the USA and Jamaica. I have undertaken genome-wide micro-array scanning and created an imputed genotype dataset of over 15,000,000 genetic variants. I have used these phenotype and genotype datasets to conduct genetic association studies, both genome-wide and candidate gene association studies. These analyses are based on linear mixed modelling to account for relatedness (including population stratification) within the cohort. In addition to the severity outcomes, I have also evaluated the known genetic loci for HbF and created a genetic "summary statistic" to quantify the effects of these three loci. Finally, I have also assessed the role of the erythroid regulator *KLF1* in HbF levels in SCD with two laboratory-based projects.

Ethical basis of research

Written informed consent was obtained through three approved study protocols (LREC 01-083, 07/H0606/165, and 12/LO/1610) and research conducted in accordance with the Helsinki Declaration (1975, as revised 2008).

Abbreviations

2,3-DPG	2,3-DiPhosphoGlycerate
ACS	Acute Chest Syndrome
ADAMTS13	A Disintegrin And Metalloproteinase with a ThromboSpondin type 1 motif, member 13
AFR	African
aHUS	Atypical Haemolytic Uraemic Syndrome
ALP	Alkaline phosphatase
ALT	Alanine transaminase
APE	Acute Painful Episode
AST	Aspartate transaminase
AVN	Avascular Necrosis
BRC	Biomedical Research Centre
CAAPA	Consortium on Asthma among African-ancestry Populations in the Americas
CFB	Complement Factor B
CFH	Complement Factor H
CKD	Chronic kidney disease
DARC	Duffy Antigen Receptor for Chemokines
DGKE	DiacylGlycerol Kinase Epsilon
DNA	DeoxyriboNucleic Acid
eGFR	Estimated glomerular filtration rate
EPO	Erythropoietin
EPR	Electronic patient record
EUR	European
GC	Genomic Control
GFR	Glomerular filtration rate
GGT	Gamma-glutamyl transferase
GRM	Genetic Relatedness Matrix
GSTT	Guys and St Thomas' Hospital NHS Trust
GWAS	Genome-Wide Association Study
Hb	Haemoglobin
HbSC	Hb SC disease (compound heterozygous inheritance of HbS and HbC)
HbS/HPFH	Compound heterozygous inheritance of HbS and HPFH
HbSS	Homozygous inheritance of HbS
HbC	Haemoglobin C
HbE	Haemoglobin E
HbF	Haemoglobin F (fetal haemoglobin)
HbS	Haemoglobin S (sickle haemoglobin)
HbSβ thalassaemia	Compound heterozygous inheritance of HbS and β-thalassaemia ($Hb\beta^{+}$ or $Hb\beta^{0}$)
HMIP	*HBS1L-MYB* intergenic polymorphisms
HRC	Haplotype Reference Consortium
HPFH	Hereditary persistence of fetal haemoglobin
HPLC	High Performance Liquid Chromatography
HIV	Human Immunodeficiency Virus
HRC	Haplotype Reference Consortium
HWE	Hardy-Weinberg Equilibrium
IQR	Inter-Quartile Range
KCH	King's College Hospital NHS Trust
KCL	King's College London

LD	Linkage disequilibrium
LDH	Lactate dehydrogenase
LH	Lewisham Hospital
LMM	Linear mixed model(ling), form of regression analysis used in genetic association studies
LOCO	Leave one chromosome out (statistical analysis method during association analysis)
MAF	Minor Allele Frequency
MAHA	Micro-angiopathic Haemolytic Anaemia
MCP	Membrane Cofactor Protein
MCV	Mean Cell Volume (of red blood cells)
MEGA	Multi-Ethnic Genotyping Array
MLMA	Mixed linear modelling association (file format)
mRNA	Messenger ribonucleic acid
NO	Nitric oxide
PC	Principal Components generated from PCA
PCA	Principal Components Analysis (statistical technique to identify independent determinants of a set of variables – used in a myriad of settings including to reduce a dataset or in population genetics to infer ethnicity)
qPCR	Quantitative Polymerase Chain Reaction
QC	Quality Control
QEH	Queen Elizabeth Hospital, Woolwich
QQ plots	Quantile-Quantile plots (QC)
QTL	Quantitative trait locus(i)
RAM	Random Access Memory
RBC	Red Blood Cell
RNA	RiboNucleic Acid
RT	Reverse Transcription
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SCD	Sickle Cell Disease (all genotypes involving homozygote/heterozygote inheritance of haemoglobin S causing a clinical phenotype)
SNP	Single Nucleotide Polymorphism
TCD	TransCranial Doppler (ultrasound imaging technique used to calculate cerebral vessel velocities used to infer stroke risk in children with SCD)
THBD	Thrombomodulin
TTP	Thrombotic Thrombocytopenic Purpura
UCSC	University of California Santa Cruz (owner of genome browser)
Urinary ACR	Urinary albumin to creatinine ratio
UTR	Untranslated Region
VCF	Variant Call Format (file type)
VEGF	Vascular Endothelial Growth Factor
WBC	White Blood Cell count
WHO	World Health Organisation

Chapter 1: Introduction

List of figures 7

1.1. Outline of my PhD 8

1.2. Haemoglobin (Hb) 9
1.2.1. Background 9
1.2.2. Hb variants and haemoglobinopathies 10

1.3. Sickle cell disease 11
1.3.1. Haemoglobin S: primary cause of sickle cell disease 11
1.3.2. Sickle cell disease: pathophysiology 11
1.3.3. Clinical aspects of sickle cell disease 12
1.3.4. Genetic modifiers of sickle cell disease 13
1.3.4.1. Genetic modifiers of "global" sickle cell disease severity 14
1.3.4.2. Causative sickle genotype 14
1.3.4.3. Alpha globin genotype 15
1.3.4.4. Fetal haemoglobin (HbF) 15
1.3.4.5. Other genetic modifiers of "global" sickle cell disease severity 18

1.4. Genetic methodologies 18
1.4.1. Background 18
1.4.2. Association studies: candidate gene studies and genome-wide association studies 19
1.4.3. Development of genome-wide association studies: summary 20

1.5. Aims of my PhD 20

References 21

List of figures

Figure 1 Globin-switching in normal humans 10
Figure 2 Pathophysiology of sickle cell disease adapted from (Rees et al., 2010) 12
Figure 3 Multi-organ complications of sickle cell disease, modified from http://nurseslabs.com/sickle-cell-anemia-nursing-management/ 13

1.1. Outline of my PhD

The broad aim of my PhD is to investigate genotype-phenotype correlation in the haemoglobin disorder sickle cell disease (SCD). Sickle cell disease is a complex multi-system disorder which results from a single nucleotide change in the β-globin gene. This simple genetic defect does not explain the multi-faceted pathophysiological mechanisms and clinical manifestations of SCD. To investigate the genetic determinants of SCD severity, I defined and implemented "*phenotypes"* – that is, clinical outcomes or biomarkers of disease severity – which are both meaningful clinically *and* markers reasonably expected to have genetic modifiers. In addition to using these phenotypes in genetic association studies, characterising clinical characteristics of a large SCD cohort is valuable in and of itself (and has led to off-shoot clinical projects that I have published during my PhD). I have defined four key markers of global severity in our cohort: fetal haemoglobin (HbF) levels, haemolytic index, hospitalisation rate, and mortality (chapter 2).

"Genotypes" for many genetic polymorphisms are now easier to produce than ever. Genome-wide arrays are readily available and provide vast genotyping data rapidly and relatively cheaply: I used the African-heritage specific "MEGA" chip from Illumina which generates data on 1.7 million variants. Statistical imputation can be used to expand the genotyping dataset to create a massive genotype database: imputation of our dataset resulted in dataset of over 15 million variants (chapter 3).

Analysis of genotype-phenotype association has become more sophisticated over the last ten years in response to the availability of new analytical approaches. Contemporary genetic association studies must make use of genome-wide data to infer information about "*relatedness"* of individuals within the cohort, and use this information to account for this relatedness when undertaking analysis. This avoids the pitfalls of many older studies, especially those where *population structure* (known or unknown) was not controlled for and resulted in false positive results that report genotype- phenotypes associations that are confounded by ethnicity. Assessment and utilisation of relatedness is crucial in genotype-phenotype correlation in SCD because (a) many cohorts contain cryptic "near relateds" due to the Mendelian inheritance of this autosomal recessive disorder (b) statistical issues of population stratification ("far relateds") are particularly acute in SCD in the UK where there are both multiple ethnic groups migrating from different regions of Africa to the UK and issues of admixture in our African-Caribbean patients.

I used *linear mixed modelling* approaches which takes account of relatedness (both near and far), as well as incorporating individual-specific covariates (e.g. age, sex). I studied both genome wide association ("GWAS") in chapter 4 and also assessed candidate genes in chapter 5.

Finally, I considered the key erythroid factor *KLF1* as a candidate gene for modulating HbF% in SCD (chapter 6). Rare variants in *KLF1* have been associated with different HbF-boosting phenotypes, mainly in the non-sickle setting. I performed gene expression studies on one very rare variant not previously associated with very high HbF% (and concomitant phenotype which is virtually disease-free). I also investigated a common *KLF1* intronic variant for association with HbF% levels.

1.2. Haemoglobin (Hb)

1.2.1. Background

Haemoglobin (Hb) has been the subject of some of the earliest molecular and genetic research. Its pivotal role in oxygen transport is well-learned in school: Hb binds to oxygen in the lungs, delivers it to body tissues where it is exchanged for carbon dioxide; carboxy-Hb then returns to the lungs, offloads the carbon dioxide in exchange for oxygen, and the cycle continues.

Hb is a tetramer of two α-like globin chains and two β-like globin chains, with each chain containing an oxygen-carrying heme group (Perutz et al., 1960). Different Hb protein forms are synthesised at different stages of fetal development. The three embryonic haemoglobins are Gower I ($\zeta_2\varepsilon_2$), Gower II ($\alpha_2\varepsilon_2$), and Portland ($\zeta_2\gamma_2$) (Huehns et al., 1964). The first "globin-switch" from embryonic to fetal Hb ($\alpha_2\gamma_2$, HbF) occurs at 6-8 weeks' gestation. Around birth, the second "globin-switch" occurs from fetal to adult Hb ($\alpha_2\beta_2$, HbA) and is completed by 6 months of age. HbA represents more than 95% total Hb and remains the dominant Hb throughout life (Kunkel and Wallenius, 1955). The globin-switch process is pictured in Figure 1. HbA_2 ($\alpha_2\delta_2$) is a minor form of Hb which makes up 2-3% of total Hb (Stamatoyannopoulos, 1972). The switch from fetal to adult, however, is not total, as residual HbF at <1% total Hb is retained in healthy adults. The genes coding for ε-, γ-, δ- and β-globin chains are localised in a cluster on chromosome 11p while those encoding the α-like globins are found in another cluster on chromosome 16 and consists of a single ζ-globin gene and two co-expressed α-globin genes (α2 and α1).

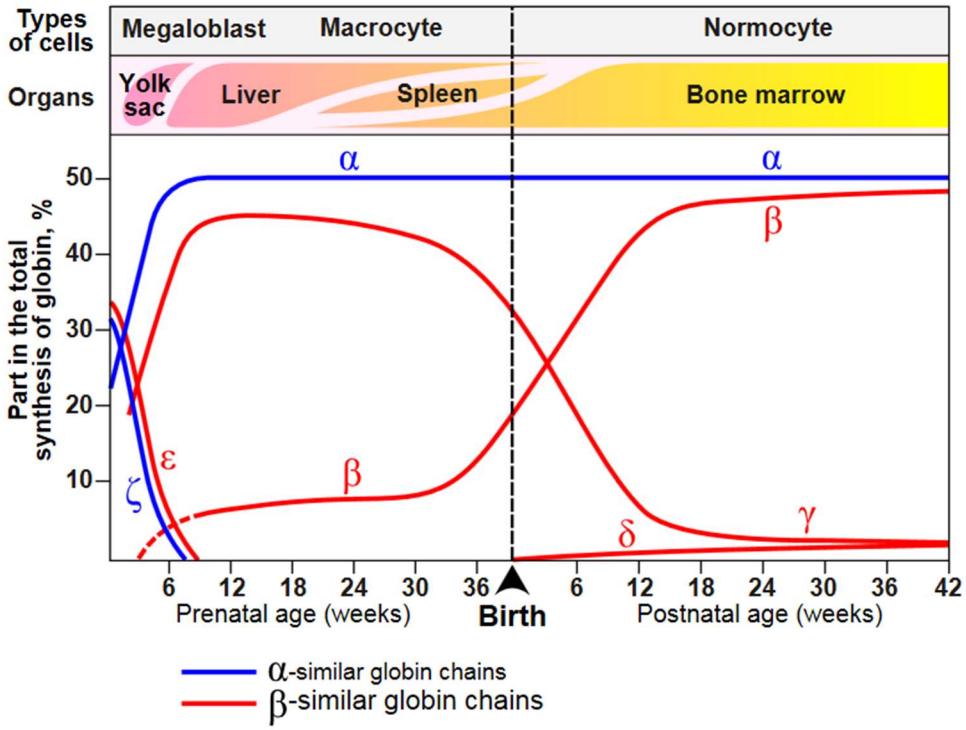


Figure 1 Globin-switching in normal humans

By Postnatal_genetics.svg: original: Furfur, File:Haemoglobin-Ketten.svg, derivation/translation:Leonid 2 derivative work: Leonid 2 (Postnatal_genetics.svg) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0) or GFDL (http://www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons

1.2.2. Hb variants and haemoglobinopathies

The globin genes are genetically heterogeneous leading to many Hb variants. More than 1200 Hb variants have been described to date (http://globin.cse.psu.edu/globin/hbvar/) with variable clinical significance; from no clinical effect to significant biochemical Hb changes and pathology. Diseases associated with clinically significant Hb abnormalities are termed *haemoglobinopathies*. Either the α-globin and β-globin chains can be affected, but it is not unusual for individuals to have inherited Hb variants affecting both α and β globin genes, as these variants are often prevalent in the same regions. Haemoglobinopathies include both *quantitative* and *qualitative* changes to Hb.

Hb variants that result in *quantitative* changes are termed thalassaemias, the most common of which are α- or β-thalassaemia depending on whether there is reduced production of α- (α-thalassaemia) or β- (β-thalassaemias) respectively, globin chains(Steinberg et al., 2009, Huisman et al., 1996).

Qualitative Hb variants produce structurally abnormal Hb, but usually in normal quantities. This structural change frequently does not have functional consequences. Clinically significant Hb variants include those that result in a change in charge and solubility (HbS $\alpha_2\beta_2^{6Glu\rightarrow Val}$, HbC

$\alpha_2\beta_2^{6Glu\rightarrow Lys}$) and oxygen affinity (HbChesapeake $\alpha_2^{92Arg\rightarrow Leu}$ β_2)(Huisman et al., 1996). Some qualitative Hb variants (e.g. HbE) can also be reduced in amounts, and such variants are sometimes termed thalassaemic haemoglobinopathies.

1.3. Sickle cell disease

1.3.1. Haemoglobin S: primary cause of sickle cell disease

The discovery of "sickle haemoglobin" or "Haemoglobin S" (HbS) represents an important step in molecular medicine; sickle cell disease (SCD) has been heralded as the first "molecular disease". Pauling ascribed the basis of SCD to the presence of an abnormal haemoglobin in 1949 (Pauling et al., 1949). In 1957, Ingram (Ingram, 1957) described the abnormal haemoglobin as being caused by a single amino acid substitution (glutamic acid to valine) at position 6 of the β-globin chain of haemoglobin and in 1963, Goldstein (Goldstein et al., 1963) demonstrated that this resulted from a single base change of thymine to adenine.

Functionally, the change from glutamic acid (negative charge) to valine (neutral charge) facilitates polymerisation of HbS when deoxygenated (Brittenham et al., 1985). HbS polymerisation deforms red blood cells (RBC) into the classical "sickle" shape. HbS polymerisation is dependent upon: degree of (de)oxygenation of the RBC, intracellular pH, and intracellular HbS concentration (Bunn, 1997). Sickle carriers (HbS 35-40%, HbA 60-65%) are asymptomatic except under extreme conditions (very high altitude, dehydration from extreme exercise).

Homozygous inheritance of HbS or co-inheritance with specific haemoglobin variants (compound heterozygotes) results in a clinically significant "sickling" haemoglobinopathy: sickle cell disease. HbSS is the most common genotype; in patients of African heritage, HbSS comprises 60% to 70%, with most of the remaining cases made up of compound heterozygotes HbSC (co-inheritance of β^S and β^C alleles) in up to 35% or HbSβ thalassaemia (co-inheritance of β^S and either β^0 or β^+ mutations) in 5%, based on World Health Organisation (WHO) international figures (Modell and Darlison, 2008). Very rare causes of SCD in the UK include HbSE, $HbSO^{Arab}$ and $HbSD^{Punjab}$.

1.3.2. Sickle cell disease: pathophysiology

HbS polymerisation and sickling of RBCs results in two major pathological processes: vaso-occlusion of micro-vasculature and haemolytic anaemia, see Figure 2.

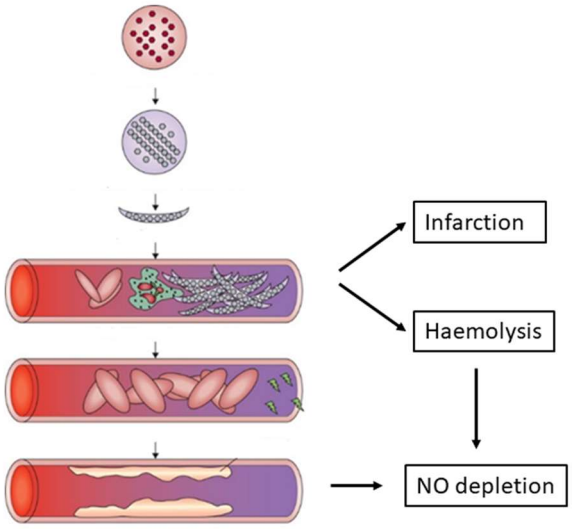


Figure 2 Pathophysiology of sickle cell disease adapted from (Rees et al., 2010)

Patients with a higher haematocrit/total Hb tend to have a higher incidence of "vaso-occlusive" problems including frequent pain episodes, acute chest syndrome (ACS) and avascular necrosis (AVN). Patients with more haemolysis (evidenced by higher lactate dehydrogenase, lower Hb, and higher bilirubin) tend to have increased incidence of leg ulcers, priapism and pulmonary hypertension. Haemolytic-related pathologies in SCD are postulated to be mediated by nitric oxide (NO) bioavailability (Kato et al., 2007, Rees et al., 2010). Physiologically, NO controls vasodilatation, inhibits platelet activation and reduces adhesion molecule expression. Plasma Hb release during intravascular haemolysis leads to NO consumption. Furthermore, haemolysis also releases arginase which breaks down the L-arginine (a substrate for NO production), further reducing endothelial NO. The consequences of reduced NO bioavailability in SCD (both through reduced production and increased consumption of NO) are: vasoconstriction, endothelial activation, inflammation and vasculopathy.

1.3.3. Clinical aspects of sickle cell disease

The simple genetic basis of SCD does not manifest as identical clinical presentations in all patients. Indeed, there is a striking range *and* severity of complications in SCD, even within the same sickle genotype. Patients may develop all or none of the complications described, or in a variety of different combinations with varying severities.

Acute pain episodes are the most common presentation of SCD and are the result of tissue ischaemia secondary to vaso-occlusion. Pain is extremely variable in location, severity, duration and the triggers appear to be different in patients (Ballas, 2005, Platt et al., 1991).

Patients who have a higher baseline Hb appear to be at increased risk of painful episodes. Prompt treatment with appropriate analgesics and individualised care plans are used to tailor treatment to individual patients (Rees et al., 2003).

Chronic haemolytic anaemia is a key feature of SCD, although there is significant inter-genotypic, and intra-genotypic, variability in degrees of anaemia. Furthermore, anaemia can be exacerbated acutely; an increase in the rate of HbS polymerisation and sickling compared to steady-state can occur in acute vaso-occlusive crises triggered by a variety of causes including infection/inflammation, cold or decreased oxygen tension. The increased sickling of RBCs increases haemolysis. Other important causes of acute anaemia include acute splenic sequestration (Solanki et al., 1986, Emond et al., 1985), Parvovirus B19-associated red cell aplasia, and haemolytic transfusion reactions. When anaemia of any cause becomes symptomatic, transfusion may be required.

There is a plethora of end-organ complications in SCD summarised in Figure 3.

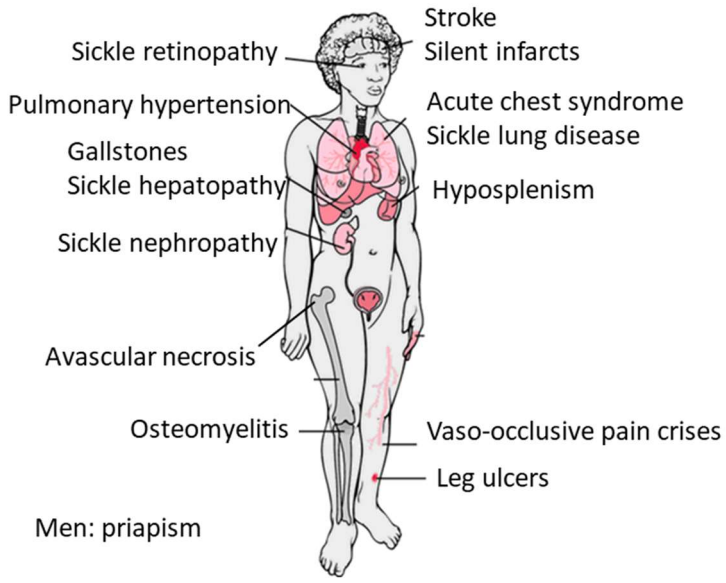


Figure 3 Multi-organ complications of sickle cell disease, modified from http://nurseslabs.com/sickle-cell-anemia-nursing-management/

1.3.4. Genetic modifiers of sickle cell disease

As described in section 1.3.1, the syndrome of SCD includes homozygosity for the β^S allele (HbSS), as well as HbSC disease, HbSβ thalassaemia and rarer compound heterozygotes. Generally, the compound heterozygotes have a milder disease than HbSS patients, but even within each genotypic and ethnic group, a spectrum of clinical variability is the recurring theme. For example, mild patients with HbSS are virtually asymptomatic, while at the severe

end of the clinical spectrum presentation at an early age, frequent hospital admissions with acute pain crises, childhood strokes, other end-organ damage and early mortality is seen.

Both environmental and genetic factors contribute to this clinical variability. Environmental factors (Tewari et al., 2015) have long been recognised as significant in SCD including: weather changes (cold, rain), and air pollution. Environmental factors also include nutritional state, access to social support and medical care, all of which influence risk factors such as infections. The impact of environmental factors is demonstrated most graphically on the differences in the natural history and outcomes of SCD between high- and low-income countries.

1.3.4.1. Genetic modifiers of "global" sickle cell disease severity

Since the central pathology of SCD is HbS polymerisation and formation of sickled RBCs, factors that impact this primary event will have a global effect on the disease phenotype. The three major genetic modifiers that affect this are: the causative sickle genotype, co-existing α-thalassaemia and the innate ability to produce HbF.

1.3.4.2. Causative sickle genotype

While presence of HbS is fundamental to the pathobiology, the likelihood of HbS polymerisation and sickling is highly dependent on the concentration of intra-cellular HbS, and the presence of non-HbS haemoglobin(Noguchi et al., 1983). Thus, individuals with HbSS or HbSβ^0 thalassaemia, where the intracellular Hb is almost all HbS, tend to have the most severe disease, followed by HbSC and HbSβ^+ thalassaemia.

HbA ($\alpha_2\beta_2$) or HbA_2 ($\alpha_2\delta_2$) do not participate in HbS polymerisation. Since the β^+ thalassaemia alleles in African-heritage populations are normally of the milder type with minimal deficit in β globin production, Africans with HbSβ^+ thalassaemia have substantial proportions of intra-cellular HbA and the SCD tends to be very mild. In contrast, individuals with HbSβ^+ thalassaemia in the Mediterranean, have SCD almost as severe as that of HbSS(Serjeant GR, 2001). Subjects with sickle cell trait (HbAS) with HbS of 35-40%, rarely suffer from symptoms of SCD.

The HbS gene is found on a genetic background of four common β^S-globin haplotypes: Senegal, Benin, Bantu (or Central African Republic), and Arab-Indian. Clinical studies demonstrate variation in SCD severity between the β^S haplotypes, with decreasing severity from the Bantu > Benin > Senegal > Arab-Indian haplotypes. Disease severity correlates inversely with the HbF% levels seen in these groups; lowest HbF% seen in individuals with Bantu haplotype, and highest HbF% in individuals with Arab-Indian haplotype(Nagel et al., 1985, Nagel et al., 1987, Nagel et

al., 1991, Powars, 1991, Figueiredo et al., 1996). The differences in clinical severity were ascribed to the difference in HbF% levels implicating the *Xmn*1-*HBG2* site which is linked to the Senegal and Arab-Indian β^S haplotype but not to the Bantu haplotype (Labie et al., 1985) (see below for further discussion on the modifying effects of HbF% on SCD).

1.3.4.3. Alpha globin genotype

About one-third of African-descended patients with SCD have co-existing α-thalassaemia (Steinberg and Embury, 1986, Vasavda et al., 2008). Most commonly, this is due to the deletion variant ($-\alpha^{3.7}$); 30-35% of patients are heterozygous ($\alpha\alpha/-\alpha^{3.7}$) with 3–5% homozygous for the deletion ($-\alpha^{3.7}/-\alpha^{3.7}$) (Steinberg and Embury, 1986, Vasavda et al., 2007). Co-inheritance of α-thalassaemia affects SCD red cell phenotype; it reduces intracellular HbS concentration and thus the propensity of HbS polymerisation, reducing the number of irreversibly sickled cells and decreasing haemolysis(Embury et al., 1982, Ballas, 2001).

Clinically, co-inherited α-thalassaemia protects against complications related to severe haemolysis including pulmonary hypertension, leg ulceration, priapism and albuminuria (Steinberg, 2009, Buchanan et al., 2004, Day et al., 2012, Higgs et al., 1982). Conversely, the increased haematocrit and associated blood viscosity in α-thalassaemia predispose patients to an increased likelihood of developing osteonecrosis, acute chest syndrome (ACS), retinopathy and acute painful vaso-occlusive episodes(Embury et al., 1994). Several studies have also demonstrated association of α-thalassaemia with lower Trans-Cranial Doppler (TCD) measurements and, by implication, reduced risk for stroke (Bernaudin et al., 2008, Rees et al., 2009, Flanagan et al., 2011, Cox et al., 2014) while another study could not demonstrate association between α-thalassaemia and magnetic resonance angiography (MRA)-defined vasculopathy in paediatric patients with HbSS disease (Thangarajh et al., 2012). Co-existing α-thalassaemia also reduces bilirubin with a quantitative effect that is independent to that of the *UGT1A1* promoter polymorphism (Vasavda et al., 2007). Co-inheritance of α-thalassaemia blunts the response to hydroxycarbamide therapy in SCD; this may be explained by its effect on HbF% levels and MCV, two key parameters associated with hydroxycarbamide response (Vasavda et al., 2008).

1.3.4.4. Fetal haemoglobin (HbF)

Fetal haemoglobin (HbF, $\alpha_2\gamma_2$) is a major ameliorating factor in SCD. Understanding fetal haemoglobin control and its therapeutic reactivation (via pharmacological and genetic approaches) remains a top research priority. HbF% reduces the propensity for HbS polymerisation and its sequelae in two major ways: 1) the hybrid tetramers ($\alpha_2\gamma\beta^S$) do not partake in HbS polymerisation, and 2) the presence of intra-RBC HbF dilutes the concentration of HbS (Noguchi et al., 1988). The protective effect of HbF in SCD becomes evident within six

months to two years of age as HbF levels decline and clinical manifestations of SCD become evident.

In SCD, high HbF% levels are a major predictor of survival(Platt et al., 1994), and pain rates (Platt et al., 1991, Dampier et al., 2004); conversely, low levels of HbF% have been associated with increased risk of brain infarcts in young children (Wang et al., 2008). At the sub-phenotype level, there appear to be disparities in its effects on complications such as renal impairment, retinopathy and priapism (Thein, 2011, Steinberg and Sebastiani, 2012); this may relate to the small sample sizes in genetic studies with specific sub-phenotypes.

Although the γ-globin genes are autonomously silenced in adults, genetic variants lead to natural variation in γ-globin expression of over 20-fold. These variants account for 89% of the quantitative variation but the genetic aetiology is complex with no clear Mendelian inheritance patterns. The three known quantitative trait loci (QTLs) for the common HbF% variation in adults are: *Xmn1-HBG2* (rs782144) within the β-globin gene cluster on chromosome 11p, *HBS1L-MYB* intergenic polymorphism (*HMIP*) on chromosome 6q23, and *BCL11A* on chromosome 2p16.

Variants in the *HBB*, *HMIP* and *BCL11A* loci account for 10% - 50% of the variation in HbF% levels in adults, healthy or with SCD or β-thalassaemia, depending on the population studied (Menzel et al., 2007, Lettre et al., 2008, Galanello et al., 2009, Bhatnagar et al., 2011, Makani et al., 2011, Badens et al., 2011, Bae et al., 2012, Mtatiro et al., 2014). The remaining variation ('missing heritability') is likely to be accounted for by many loci with relatively small effects, and/or rare variants with significant quantitative effects on γ-globin gene expression that are typically missed.

HBB cluster on chromosome 11p

Xmn1-HBG2 (rs782144) in the *HBB* cluster was the first known QTL for HbF% and long-implicated by clinical genetic studies (Labie et al., 1985). The differences in clinical severity of SCD was ascribed to the difference in HbF% levels implicating the *Xmn1-HBG2* site which is linked to the Senegal and Arab-Indian β^S haplotype but not to the Bantu haplotype (Labie et al., 1985). Recent high-resolution genotyping, however, suggests that rs782144 is not likely to be the variant itself, but in tight linkage disequilibrium with causal element(s) in the β-globin cluster.

BCL11A on chromosome 2p16

Functional studies in primary human erythroid progenitor cells and transgenic mice demonstrated that BCL11A acts as a repressor of γ-globin gene expression that is affected by variants in intron 2 of this gene (Sankaran et al., 2008). Fine-mapping demonstrated that these HbF%-associated variants, in particular rs1427407, localised to an enhancer that is erythroid-specific and not functional in lymphoid cells (Bauer et al., 2013).

HMIP on chromosome 6q23

High resolution genetic mapping and resequencing refined the 6q QTL to a group of variants in tight linkage disequilibrium (LD) in a 24-kb block between the *HBS1L* and *MYB* gene, referred to as *HMIP-2* (Thein et al., 2007). The causal variants are likely to reside in two clusters within the block that coincide with conserved core enhancer elements at -84 and -71 kb respectively, upstream of *MYB* (Stadhouders et al., 2014, Menzel et al., 2014). A three-base pair (3-bp) deletion in *HMIP-2* -84 region is one functional element in the *MYB* enhancers accounting for increased HbF expression in individuals who have the sentinel variant rs9399137 that was found to be common in European and Asian populations, although less frequently in African populations (Farrell et al., 2011). The *HBS1L-MYB* intergenic enhancers do not appear to affect expression of *HBS1L*, the other flanking gene.

KLF1 on chromosome 19p13

KLF1 (previously termed *EKLF*), an erythroid "master regulator" and discovered by Jim Bieker in 1993 (Miller and Bieker, 1993), re-emerged as a key transcription factor controlling HbF through genetic studies in a Maltese family with β-thalassaemia and hereditary persistence of HbF (HPFH). Linkage studies identified a locus for the HPFH that segregated independently of the *HBB* locus on chromosome 19p13 which encompassed *KLF1* (Borg et al., 2010). Subsequent studies, which included expression profiling of erythroid progenitor cells, confirmed *KLF1* as the γ-globin gene modifier in this family. Family members with HPFH were heterozygous for the nonsense K288X mutation in *KLF1* that disrupted the DNA-binding domain of KLF1. Multiple studies have now confirmed that *KLF1* is key in the globin-switch from *HBG* to *HBB* expression; it not only activates *HBB* directly, providing a competitive edge, but also silences the γ-globin genes indirectly via activation of *BCL11A* (Siatecka and Bieker, 2011, Zhou et al., 2010, Esteghamat et al., 2013). KLF1 has three zinc finger domains, which mediate sequence-specific binding to DNA and are essential for activation of KLF1 target genes.

There have been numerous reports of association of rare *KLF1* variants with increased HbF% either as a primary phenotype, or in association with other red cell disorders (Borg et al., 2011). Recently, *KLF1* mutations have been noted not just to be relatively more common in,

but also to ameliorate the severity of, β-thalassaemias in China. However, to date there are no reports of *KLF1* mutations in SCD patients. Furthermore, several genome-wide association studies of HbF% (including ones in SCD patients of African descent) have investigated common variants at the KLF1 locus but no influence on HbF% levels was detected (Bhatnagar et al., 2011, Mtatiro et al., 2014). *KLF1* continues to be actively explored as a genetic modifier of HbF% levels in SCD. It should be noted that the KLF1 variants in the Chinese population were discovered by target resequencing of a candidate gene (i.e. KLF1), not by GWAS.

1.3.4.5. Other genetic modifiers of "global" sickle cell disease severity

Causative sickle genotype, α-globin status and HbF remain the only established genetic determinants of SCD severity. Other researchers have used alternate phenotypes and/or scoring systems to establish new genetic loci. Using the global severity index propounded by El-Hazmi (el-Hazmi, 1992), Nishank identified three *eNOS* gene polymorphisms associated with SCD severity (Nishank et al., 2013)**.** A genome-wide association study (Sebastiani et al., 2010) utilised the global severity score devised by their own group (Sebastiani et al., 2007) to identify associations with variants in *KCNK6* (potassium channel gene) and *TNKS* (telomere length regulator gene). The same group went on to demonstrate variants in *NPRL3* and *HBA1*/*HBA2* regulatory elements were associated with their own haemolytic score derived through principal components analysis (Gordeuk et al., 2009, Milton et al., 2013). All these are tentative findings and are not established genetic determinants of SCD: they have not been significantly repeated in other cohorts.

1.4. Genetic methodologies

1.4.1. Background

Approaches to locating genetic variants relevant to human disease have evolved over time from linkage analysis to association studies (Hirschhorn and Daly, 2005). Historically, linkage analysis studies were used to establish linkage between genes that co-segregate with a trait/disease within a family. This technique has been successful in highly penetrant single gene disorders, but has had limited success in detecting the common, low effect variants in complex traits. Association studies look for differences in the frequencies of genetic variants between cases and controls to find genetic variants that are strongly associated with a trait/disease, and is applied in cohorts with no family structure. If a variant is more common in cases than controls, an association is described. Quantitative traits in cohorts, instead of case-control studies, can be used in linear rather than logistic regression models. Association studies require large sample numbers and until recently have not been feasible due to genotyping cost. Crucially, variants identified in pilot studies ("discovery cohort") should

always be replicated in additional independent populations ("replication" or "validation" cohort).

1.4.2. Association studies: candidate gene studies and genome-wide association studies

Prerequisites for any genetic association study include: the trait must be heritable (correlation of sibling pairs, good r value); and there must be a clear distinction between cases and controls (or sufficient variability in a quantitative trait). Adequate patient numbers are essential to allow robust statistical analysis and replication. Again, this presents problems in SCD genetic association studies; most institutions have small numbers of patients (in contrast to hypertensive or diabetic cohorts). Admixture of different ethnic groups is a confounder when different cohorts are pooled for analysis unless population stratification is accounted for prior to association analysis.

Two types of association studies have been utilised in SCD: candidate gene and genome-wide association studies (GWAS).

Candidate gene association studies look for differences in the frequencies of genetic variants in targeted genes between cases and controls. Candidate genes can be derived from a variety of sources. They may have been proven in one disease or population, and a new study transfers this knowledge onto a different disease or population. Alternatively, putative genes associated with a particular mechanism relevant to a disease may be studied. In SCD, while the primary aetiology is HbS polymerisation, multiple different (but inter-related) downstream pathological mechanisms also contribute to SCD phenotype: haemolysis/heme damage, inflammation, oxidant injury, nitric oxide biology, vaso-regulation, cell adhesion and blood coagulation. All of these downstream pathways suggest candidate genes that could plausibly affect the different sickle-related complications. Critics of candidate gene studies argue that our limited knowledge of SCD pathophysiology is inadequate to predict functional candidate genes (Manolio, 2013). I will pursue a candidate gene study with multiple candidate genes against global severity indices in SCD in chapters 5 and 6.

GWAS involve an unbiased scan of the human genome and therefore are more likely to reveal unsuspected interactions (Manolio, 2013). It thus delivers a "hypothesis free" method that could reveal new genes controlling SCD, thereby exposing novel pathophysiological pathways. The complexity of SCD pathophysiology, and the dearth of knowledge about the genetic determinants of SCD global severity makes a GWAS approach attractive. I will pursue a GWAS in SCD with multiple outcome measures in chapter 4.

1.4.3. Development of genome-wide association studies: summary

Design of genome-wide marker panels requires data on linkage disequilibrium (LD) for the "whole" genome; this was the aim of the HapMap project, www.hapmap.org. HapMap gave us evidence that LD can persist between loci several thousand bases apart. However, this depends upon the age of the population of interest. Notably, in African populations, LD encompasses smaller or less extensive regions than "younger" populations (European, Asian) so studies of African populations have less power to detect associated alleles. Conversely, African 'hits' may locate causal variants more precisely. Extensive LD permits design of efficient marker panels i.e. reasonably few tag variants. This motivated commercial companies to develop generic "chips" for GWAS analysis. The market is now dominated by two companies, Illumina and Affymetrix.

The first wave of GWAS (2005-7) used 100,000 to 500,000 variants. Current chips have coverage of over two million variants (1 variant per 2kb) and power fairly close to the theoretical maximum for common variants in European populations.

Once an association has been made, it is crucial to confirm the association through replication in another cohort, and then to recheck the association across multiple populations; causal variants may differ due to many factors including environmental influences and population genetic background.

GWAS, including the theory behind the process as well as pros and cons, are discussed in detail in chapter 4.

1.5. Aims of my PhD

- Create and curate a clinical dataset of a large British adult SCD cohort: a "phenotype database"
    - Use this for clinical research e.g. mortality
    - Define meaningful 'global' SCD phenotypes:
        - HbF%, hospitalisation rate, haemolytic index, mortality
- Perform genome-wide MEGA genotyping chip array on a South East London sickle cohort, and use statistical imputation to create an extended "genotype database"
- Analyse phenotype-genotype associations for both HbF% and newly created (albeit experimental) global phenotypes in SCD using:
    - Mixed linear modelling including a "genetic relatedness matrix" to take account of near- and far- relatedness.

- Candidate gene analysis for putative regions of interest, considering a modified p-value based on regional linkage disequilibrium
  - Assess the known HbF% modifiers [*BCL11A*, *HMIP2*, *HBG2-Xmn1*] using MEGA dataset
  - Explore other biologically plausible markers in *SCD* severity mining the MEGA dataset
- Create software which allows other laboratory users for future genotype / phenotype analysis for new phenotypes (both genome-wide and candidate gene analysis)
- Examine the role of *KLF1* variants in determining HbF% levels in SCD patients (laboratory work)

References

BADENS, C., JOLY, P., AGOUTI, I., THURET, I., GONNET, K., FATTOUM, S., FRANCINA, A., SIMEONI, M. C., LOUNDOU, A. & PISSARD, S. 2011. Variants in genetic modifiers of beta-thalassemia can help to predict the major or intermedia type of the disease. *Haematologica,* 96**,** 1712-4.

BAE, H. T., BALDWIN, C. T., SEBASTIANI, P., TELEN, M. J., ASHLEY-KOCH, A., GARRETT, M., HOOPER, W. C., BEAN, C. J., DEBAUN, M. R., ARKING, D. E., BHATNAGAR, P., CASELLA, J. F., KEEFER, J. R., BARRON-CASELLA, E., GORDEUK, V., KATO, G. J., MINNITI, C., TAYLOR, J., CAMPBELL, A., LUCHTMAN-JONES, L., HOPPE, C., GLADWIN, M. T., ZHANG, Y. & STEINBERG, M. H. 2012. Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood,* 120**,** 1961-2.

BALLAS, S. K. 2001. Effect of alpha-globin genotype on the pathophysiology of sickle cell disease. *Pediatr Pathol Mol Med,* 20**,** 107-21.

BALLAS, S. K. 2005. Pain management of sickle cell disease. *Hematol Oncol Clin North Am,* 19**,** 785-802, v.

BAUER, D. E., KAMRAN, S. C., LESSARD, S., XU, J., FUJIWARA, Y., LIN, C., SHAO, Z., CANVER, M. C., SMITH, E. C., PINELLO, L., SABO, P. J., VIERSTRA, J., VOIT, R. A., YUAN, G. C., PORTEUS, M. H., STAMATOYANNOPOULOS, J. A., LETTRE, G. & ORKIN, S. H. 2013. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science,* 342**,** 253-7.

BERNAUDIN, F., VERLHAC, S., CHEVRET, S., TORRES, M., COIC, L., ARNAUD, C., KAMDEM, A., HAU, I., GRAZIA NEONATO, M. & DELACOURT, C. 2008. G6PD deficiency, absence of alpha-thalassemia, and hemolytic rate at baseline are significant independent risk factors for abnormally high cerebral velocities in patients with sickle cell anemia. *Blood,* 112**,** 4314-7.

BHATNAGAR, P., PURVIS, S., BARRON-CASELLA, E., DEBAUN, M. R., CASELLA, J. F., ARKING, D. E. & KEEFER, J. R. 2011. Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J Hum Genet,* 56**,** 316-23.

BORG, J., PAPADOPOULOS, P., GEORGITSI, M., GUTIERREZ, L., GRECH, G., FANIS, P., PHYLACTIDES, M., VERKERK, A. J., VAN DER SPEK, P. J., SCERRI, C. A., CASSAR, W., GALDIES, R., VAN IJCKEN, W., OZGUR, Z., GILLEMANS, N., HOU, J., BUGEJA, M., GROSVELD, F. G., VON LINDERN, M., FELICE, A. E., PATRINOS, G. P. & PHILIPSEN, S. 2010. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat Genet,* 42**,** 801-5.

BORG, J., PATRINOS, G. P., FELICE, A. E. & PHILIPSEN, S. 2011. Erythroid phenotypes associated with KLF1 mutations. *Haematologica,* 96**,** 635-8.

BRITTENHAM, G. M., SCHECHTER, A. N. & NOGUCHI, C. T. 1985. Hemoglobin S polymerization: primary determinant of the hemolytic and clinical severity of the sickling syndromes. *Blood,* 65**,** 183-9.

BUCHANAN, G. R., DEBAUN, M. R., QUINN, C. T. & STEINBERG, M. H. 2004. Sickle cell disease. *Hematology Am Soc Hematol Educ Program***,** 35-47.

BUNN, H. F. 1997. Pathogenesis and treatment of sickle cell disease. *N Engl J Med,* 337**,** 762-9.

COX, S. E., MAKANI, J., SOKA, D., L'ESPERENCE, V. S., KIJA, E., DOMINGUEZ-SALAS, P., NEWTON, C. R., BIRCH, A. A., PRENTICE, A. M. & KIRKHAM, F. J. 2014. Haptoglobin, alpha-thalassaemia and glucose-6-phosphate dehydrogenase polymorphisms and risk of abnormal transcranial Doppler among patients with sickle cell anaemia in Tanzania. *Br J Haematol,* 165**,** 699-706.

DAMPIER, C., ELY, E., EGGLESTON, B., BRODECKI, D. & O'NEAL, P. 2004. Physical and cognitive-behavioral activities used in the home management of sickle pain: a daily diary study in children and adolescents. *Pediatr Blood Cancer,* 43**,** 674-8.

DAY, T. G., DRASAR, E. R., FULFORD, T., SHARPE, C. C. & THEIN, S. L. 2012. Association between hemolysis and albuminuria in adults with sickle cell anemia. *Haematologica,* 97**,** 201-5.

EL-HAZMI, M. A. 1992. Clinical and haematological diversity of sickle cell disease in Saudi children. *J Trop Pediatr,* 38**,** 106-12.

EMBURY, S., HEBBEL, R., MOHANDAS, N. & STEINBERG, M. 1994. Sickle cell anemia: basic principles and clinical practice. *New York, NY, Raven*.

EMBURY, S. H., DOZY, A. M., MILLER, J., DAVIS, J. R., JR., KLEMAN, K. M., PREISLER, H., VICHINSKY, E., LANDE, W. N., LUBIN, B. H., KAN, Y. W. & MENTZER, W. C. 1982. Concurrent sickle-cell anemia and alpha-thalassemia: effect on severity of anemia. *N Engl J Med,* 306**,** 270-4.

EMOND, A. M., COLLIS, R., DARVILL, D., HIGGS, D. R., MAUDE, G. H. & SERJEANT, G. R. 1985. Acute splenic sequestration in homozygous sickle cell disease: natural history and management. *J Pediatr,* 107**,** 201-6.

ESTEGHAMAT, F., GILLEMANS, N., BILIC, I., VAN DEN AKKER, E., CANTU, I., VAN GENT, T., KLINGMULLER, U., VAN LOM, K., VON LINDERN, M., GROSVELD, F., BRYN VAN DIJK, T., BUSSLINGER, M. & PHILIPSEN, S. 2013. Erythropoiesis and globin switching in compound Klf1::Bcl11a mutant mice. *Blood,* 121**,** 2553-62.

FARRELL, J. J., SHERVA, R. M., CHEN, Z. Y., LUO, H. Y., CHU, B. F., HA, S. Y., LI, C. K., LEE, A. C., LI, R. C., YUEN, H. L., SO, J. C., MA, E. S., CHAN, L. C., CHAN, V., SEBASTIANI, P., FARRER, L. A., BALDWIN, C. T., STEINBERG, M. H. & CHUI, D. H. 2011. A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood,* 117**,** 4935-45.

FIGUEIREDO, M. S., KERBAUY, J., GONCALVES, M. S., ARRUDA, V. R., SAAD, S. T., SONATI, M. F., STOMING, T. & COSTA, F. F. 1996. Effect of alpha-thalassemia and beta-globin gene cluster haplotypes on the hematological and clinical features of sickle-cell anemia in Brazil. *Am J Hematol,* 53**,** 72-6.

FLANAGAN, J. M., FROHLICH, D. M., HOWARD, T. A., SCHULTZ, W. H., DRISCOLL, C., NAGASUBRAMANIAN, R., MORTIER, N. A., KIMBLE, A. C., AYGUN, B., ADAMS, R. J., HELMS, R. W. & WARE, R. E. 2011. Genetic predictors for stroke in children with sickle cell anemia. *Blood,* 117**,** 6681-4.

GALANELLO, R., SANNA, S., PERSEU, L., SOLLAINO, M. C., SATTA, S., LAI, M. E., BARELLA, S., UDA, M., USALA, G., ABECASIS, G. R. & CAO, A. 2009. Amelioration of Sardinian beta-zero thalassemia by genetic modifiers. *Blood,* 114**,** 3935-7.

GOLDSTEIN, J., KONIGSBERG, W. & HILL, R. J. 1963. The structure of human hemoglobin. VI. The sequence of amino acids in the tryptic peptides of the beta chain. *J Biol Chem,* 238**,** 2016-27.

GORDEUK, V. R., CAMPBELL, A., RANA, S., NOURAIE, M., NIU, X., MINNITI, C. P., SABLE, C., DARBARI, D., DHAM, N., ONYEKWERE, O., AMMOSOVA, T., NEKHAI, S., KATO, G. J., GLADWIN, M. T. & CASTRO, O. L. 2009. Relationship of erythropoietin, fetal

hemoglobin, and hydroxyurea treatment to tricuspid regurgitation velocity in children with sickle cell disease. *Blood,* 114**,** 4639-44.

HIGGS, D. R., ALDRIDGE, B. E., LAMB, J., CLEGG, J. B., WEATHERALL, D. J., HAYES, R. J., GRANDISON, Y., LOWRIE, Y., MASON, K. P., SERJEANT, B. E. & SERJEANT, G. R. 1982. The interaction of alpha-thalassemia and homozygous sickle-cell disease. *N Engl J Med,* 306**,** 1441-6.

HIRSCHHORN, J. N. & DALY, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet,* 6**,** 95-108.

HUEHNS, E. R., DANCE, N., BEAVEN, G. H., HECHT, F. & MOTULSKY, A. G. 1964. Human Embryonic Hemoglobins. *Cold Spring Harbor Symposia on Quantitative Biology,* 29**,** 327-&.

HUISMAN, T. H. J., CARVER, M. F. H. & EFREMOV, G. D. (eds.) 1996. *A Syllabus of Human Hemoglobin Variants* Augusta.

INGRAM, V. M. 1957. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature,* 180**,** 326-8.

KATO, G. J., GLADWIN, M. T. & STEINBERG, M. H. 2007. Deconstructing sickle cell disease: reappraisal of the role of hemolysis in the development of clinical subphenotypes. *Blood Rev,* 21**,** 37-47.

KUNKEL, H. G. & WALLENIUS, G. 1955. New hemoglobin in normal adult blood. *Science,* 122**,** 288.

LABIE, D., PAGNIER, J., LAPOUMEROULIE, C., ROUABHI, F., DUNDA-BELKHODJA, O., CHARDIN, P., BELDJORD, C., WAJCMAN, H., FABRY, M. E. & NAGEL, R. L. 1985. Common haplotype dependency of high G gamma-globin gene expression and high Hb F levels in beta-thalassemia and sickle cell anemia patients. *Proc Natl Acad Sci U S A,* 82**,** 2111-4.

LETTRE, G., SANKARAN, V. G., BEZERRA, M. A., ARAUJO, A. S., UDA, M., SANNA, S., CAO, A., SCHLESSINGER, D., COSTA, F. F., HIRSCHHORN, J. N. & ORKIN, S. H. 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**,** 11869-74.

MAKANI, J., MENZEL, S., NKYA, S., COX, S. E., DRASAR, E., SOKA, D., KOMBA, A. N., MGAYA, J., ROOKS, H., VASAVDA, N., FEGAN, G., NEWTON, C. R., FARRALL, M. & THEIN, S. L. 2011. Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. *Blood,* 117**,** 1390-2.

MANOLIO, T. A. 2013. Bringing genome-wide association findings into clinical use. *Nat Rev Genet,* 14**,** 549-58.

MENZEL, S., GARNER, C., GUT, I., MATSUDA, F., YAMAGUCHI, M., HEATH, S., FOGLIO, M., ZELENIKA, D., BOLAND, A., ROOKS, H., BEST, S., SPECTOR, T. D., FARRALL, M., LATHROP, M. & THEIN, S. L. 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet,* 39**,** 1197-9.

MENZEL, S., ROOKS, H., ZELENIKA, D., MTATIRO, S. N., GNANAKULASEKARAN, A., DRASAR, E., COX, S., LIU, L., MASOOD, M., SILVER, N., GARNER, C., VASAVDA, N., HOWARD, J., MAKANI, J., ADEKILE, A., PACE, B., SPECTOR, T., FARRALL, M., LATHROP, M. & THEIN, S. L. 2014. Global Genetic Architecture of an Erythroid Quantitative Trait Locus, HMIP-2. *Ann Hum Genet*.

MILLER, I. J. & BIEKER, J. J. 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the *Krüppel* family of nuclear proteins. *Molecular and Cellular Biology,* 13**,** 2776-2786.

MILTON, J. N., ROOKS, H., DRASAR, E., MCCABE, E. L., BALDWIN, C. T., MELISTA, E., GORDEUK, V. R., NOURAIE, M., KATO, G. R., MINNITI, C., TAYLOR, J., CAMPBELL, A., LUCHTMAN-JONES, L., RANA, S., CASTRO, O., ZHANG, Y., THEIN, S. L., SEBASTIANI, P., GLADWIN, M. T., WALK, P. I. & STEINBERG, M. H. 2013. Genetic determinants of haemolysis in sickle cell anaemia. *Br J Haematol,* 161**,** 270-8.

MODELL, B. & DARLISON, M. 2008. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull World Health Organ,* 86**,** 480-7.

MTATIRO, S. N., SINGH, T., ROOKS, H., MGAYA, J., MARIKI, H., SOKA, D., MMBANDO, B., MSAKI, E., KOLDER, I., THEIN, S. L., MENZEL, S., COX, S. E., MAKANI, J. & BARRETT, J. C. 2014. Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One,* 9**,** e111464.

NAGEL, R. L., ERLINGSSON, S., FABRY, M. E., CROIZAT, H., SUSUKA, S. M., LACHMAN, H., SUTTON, M., DRISCOLL, C., BOUHASSIRA, E. & BILLETT, H. H. 1991. The Senegal DNA haplotype is associated with the amelioration of anemia in African-American sickle cell anemia patients. *Blood,* 77**,** 1371-5.

NAGEL, R. L., FABRY, M. E., PAGNIER, J., ZOHOUN, I., WAJCMAN, H., BAUDIN, V. & LABIE, D. 1985. Hematologically and genetically distinct forms of sickle cell anemia in Africa. The Senegal type and the Benin type. *N Engl J Med,* 312**,** 880-4.

NAGEL, R. L., RAO, S. K., DUNDA-BELKHODJA, O., CONNOLLY, M. M., FABRY, M. E., GEORGES, A., KRISHNAMOORTHY, R. & LABIE, D. 1987. The hematologic characteristics of sickle cell anemia bearing the Bantu haplotype: the relationship between G gamma and HbF level. *Blood,* 69**,** 1026-30.

NISHANK, S. S., SINGH, M. P., YADAV, R., GUPTA, R. B., GADGE, V. S. & GWAL, A. 2013. Endothelial nitric oxide synthase gene polymorphism is associated with sickle cell disease patients in India. *J Hum Genet,* 58**,** 775-9.

NOGUCHI, C. T., RODGERS, G. P., SERJEANT, G. & SCHECHTER, A. N. 1988. Levels of fetal hemoglobin necessary for treatment of sickle cell disease. *N Engl J Med,* 318**,** 96-9.

NOGUCHI, C. T., TORCHIA, D. A. & SCHECHTER, A. N. 1983. Intracellular polymerization of sickle hemoglobin. Effects of cell heterogeneity. *J Clin Invest,* 72**,** 846-52.

PAULING, L., ITANO, H. A. & ET AL. 1949. Sickle cell anemia, a molecular disease. *Science,* 109**,** 443.

PERUTZ, M. F., ROSSMANN, M. G., CULLIS, A. F., MUIRHEAD, H., WILL, G. & NORTH, A. C. 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature,* 185**,** 416-22.

PLATT, O. S., BRAMBILLA, D. J., ROSSE, W. F., MILNER, P. F., CASTRO, O., STEINBERG, M. H. & KLUG, P. P. 1994. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med,* 330**,** 1639-44.

PLATT, O. S., THORINGTON, B. D., BRAMBILLA, D. J., MILNER, P. F., ROSE, W. F., VICHINSKY, E. & KINNEY, T. R. 1991. Pain in sickle cell disease: Rates and risk factors. *New England Journal of Medicine,* 325**,** 11-6.

POWARS, D. R. 1991. Beta s-gene-cluster haplotypes in sickle cell anemia. Clinical and hematologic features. *Hematol Oncol Clin North Am,* 5**,** 475-93.

REES, D. C., LAMBERT, C., COOPER, E., BARTRAM, J., GOSS, D., DEANE, C. & THEIN, S. L. 2009. Glucose 6 phosphate dehydrogenase deficiency is not associated with cerebrovascular disease in children with sickle cell anemia. *Blood,* 114**,** 742-3; author reply 743-4.

REES, D. C., OLUJOHUNGBE, A. D., PARKER, N. E., STEPHENS, A. D., TELFER, P. & WRIGHT, J. 2003. Guidelines for the management of the acute painful crisis in sickle cell disease. *Br J Haematol,* 120**,** 744-52.

REES, D. C., WILLIAMS, T. N. & GLADWIN, M. T. 2010. Sickle-cell disease. *Lancet,* 376**,** 2018-31.

SANKARAN, V. G., MENNE, T. F., XU, J., AKIE, T. E., LETTRE, G., VAN HANDEL, B., MIKKOLA, H. K., HIRSCHHORN, J. N., CANTOR, A. B. & ORKIN, S. H. 2008. Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor BCL11A. *Science,* 322**,** 1839-42.

SEBASTIANI, P., NOLAN, V. G., BALDWIN, C. T., ABAD-GRAU, M. M., WANG, L., ADEWOYE, A. H., MCMAHON, L. C., FARRER, L. A., TAYLOR, J. G. T., KATO, G. J., GLADWIN, M. T. & STEINBERG, M. H. 2007. A network model to predict the risk of death in sickle cell disease. *Blood,* 110**,** 2727-35.

SEBASTIANI, P., SOLOVIEFF, N., HARTLEY, S. W., MILTON, J. N., RIVA, A., DWORKIS, D. A., MELISTA, E., KLINGS, E. S., GARRETT, M. E., TELEN, M. J., ASHLEY-KOCH, A., BALDWIN, C. T. & STEINBERG, M. H. 2010. Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study. *Am J Hematol,* 85**,** 29-35.

SERJEANT GR, S. B. 2001. *Sickle Cell Disease,* Oxford, Oxford University Press.

SIATECKA, M. & BIEKER, J. J. 2011. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood,* 118**,** 2044-54.

SOLANKI, D. L., KLETTER, G. G. & CASTRO, O. 1986. Acute splenic sequestration crises in adults with sickle cell disease. *Am J Med,* 80**,** 985-90.

STADHOUDERS, R., AKTUNA, S., THONGJUEA, S., AGHAJANIREFAH, A., POURFARZAD, F., VAN IJCKEN, W., LENHARD, B., ROOKS, H., BEST, S., MENZEL, S., GROSVELD, F., THEIN, S. L. & SOLER, E. 2014. HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J Clin Invest,* 124**,** 1699-710.

STAMATOYANNOPOULOS, G. 1972. The molecular basis of hemoglobin disease. *Annu Rev Genet,* 6**,** 47-70.

STEINBERG, M. H. 2009. Genetic etiologies for phenotypic diversity in sickle cell anemia. *ScientificWorldJournal,* 9**,** 46-67.

STEINBERG, M. H. & EMBURY, S. H. 1986. Alpha-thalassemia in blacks: genetic and clinical aspects and interactions with the sickle hemoglobin gene. *Blood,* 68**,** 985-90.

STEINBERG, M. H., FORGET, B. G., HIGGS, D. R. & WEATHERALL, D. J. 2009. *Disorders of hemoglobin: genetics, pathophysiology, and clinical management*, Cambridge University Press.

STEINBERG, M. H. & SEBASTIANI, P. 2012. Genetic modifiers of sickle cell disease. *Am J Hematol,* 87**,** 795-803.

TEWARI, S., BROUSSE, V., PIEL, F. B., MENZEL, S. & REES, D. C. 2015. Environmental determinants of severity in sickle cell disease. *Haematologica,* 100**,** 1108-16.

THANGARAJH, M., YANG, G., FUCHS, D., PONISIO, M. R., MCKINSTRY, R. C., JAJU, A., NOETZEL, M. J., CASELLA, J. F., BARRON-CASELLA, E., HOOPER, W. C., BOULET, S. L., BEAN, C. J., PYLE, M. E., PAYNE, A. B., DRIGGERS, J., TRAU, H. A., VENDT, B. A., RODEGHIER, M. & DEBAUN, M. R. 2012. Magnetic resonance angiography-defined intracranial vasculopathy is associated with silent cerebral infarcts and glucose-6-phosphate dehydrogenase mutation in children with sickle cell anaemia. *Br J Haematol,* 159**,** 352-9.

THEIN, S. L. 2011. Genetic modifiers of sickle cell disease. *Hemoglobin,* 35**,** 589-606.

THEIN, S. L., MENZEL, S., PENG, X., BEST, S., JIANG, J., CLOSE, J., SILVER, N., GEROVASILLI, A., PING, C., YAMAGUCHI, M., WAHLBERG, K., ULUG, P., SPECTOR, T. D., GARNER, C., MATSUDA, F., FARRALL, M. & LATHROP, M. 2007. Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc Natl Acad Sci U S A,* 104**,** 11346-51.

VASAVDA, N., BADIGER, S., REES, D., HEIGHT, S., HOWARD, J. & THEIN, S. L. 2008. The presence of alpha-thalassaemia trait blunts the response to hydroxycarbamide in patients with sickle cell disease. *Br J Haematol,* 143**,** 589-92.

VASAVDA, N., MENZEL, S., KONDAVEETI, S., MAYTHAM, E., AWOGBADE, M., BANNISTER, S., CUNNINGHAM, J., EICHHOLZ, A., DANIEL, Y., OKPALA, I., FULFORD, T. & THEIN, S. L. 2007. The linear effects of alpha-thalassaemia, the UGT1A1 and HMOX1 polymorphisms on cholelithiasis in sickle cell disease. *Br J Haematol,* 138**,** 263-70.

WANG, W. C., PAVLAKIS, S. G., HELTON, K. J., MCKINSTRY, R. C., CASELLA, J. F., ADAMS, R. J. & REES, R. C. 2008. MRI abnormalities of the brain in one-year-old children with sickle cell anemia. *Pediatr Blood Cancer,* 51**,** 643-6.

ZHOU, D., LIU, K., SUN, C. W., PAWLIK, K. M. & TOWNES, T. M. 2010. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. *Nat Genet,* 42**,** 742-4.

Chapter 2: Phenotyping in Sickle Cell Disease

Figures 27

2.1. Introduction: defining endpoints for genetic association studies 28

2.2. Subjects 29

2.2.1. Regional research sickle gene bank 29
2.2.2. KCH clinical cohort 30

2.3. The KCH clinical cohort: creating and curating a clinical dataset 31

2.3.1. Background 31
2.3.2. Raw clinical data 31
2.3.3. Data cleaning pipeline 32
2.3.4. Results 32
2.3.5. Adding more clinical data 35

2.4. Fetal haemoglobin (HbF) 35

2.4.1. Introduction 35
2.4.2. Methods 36
2.4.3. Results 36
2.4.3.1. KCH clinical cohort 36
2.4.3.2. Regional sickle gene bank 37
2.4.3.3. Missing/lack of validated HbF data 37
2.4.4. Validation of the KCH clinical dataset 37

2.5. Haemolysis: 10-year data 38

2.5.1. Introduction: a haemolytic index 38
2.5.2. Methods 39
2.5.2.1. Principal component analysis (PCA) 39
2.5.3. Results 39
2.5.4. Discussion 40

2.6. Acute pain frequency: hospitalisation rate from 10-year data 40

2.6.1. Introduction 40
2.6.2. Methods 42
2.6.3. Results 42

2.7. Survival | Mortality 43

2.8. Sickle nephropathy 43

2.8.1. Introduction 43
2.8.2. Methods 44
2.8.3. Results 44

2.9. Discussion 45

References 45

Appendix 1 49
Appendix 2 54
Appendix 3 56
Appendix 4 57
Appendix 5 58
Appendix 6 59
Appendix 7 61
Appendix 8 62
Appendix 9 63
Appendix 10 64

Figures

Figure 1 Overlap of two cohorts: research sickle gene bank and King's College Hospital adult clinic 29
Figure 2 Demographics of the regional sickle gene bank: panel (a) demonstrates the distribution of sickle genotype across age; panel (b) demonstrates the α-globin status. 30
Figure 3 Demographics of the KCH clinical cohort: panel (a) demonstrates the distribution of sickle genotype across age; panel (b) demonstrates the alpha globin status. 31
Figure 4 Raw clinical data format 32
Figure 5 Basic laboratory results for HbSS and HbSβ0 patients 33
Figure 6 Basic laboratory results for HbSC patients..... 34
Figure 7 validated HbF% values for KCH clinical cohort: panel a - HbSS/SB0 (N=393); panel b - HbSC (N=220) 37
Figure 8 validated HbF% values for the regional sickle gene bank: panel a - HbSS/SB0 (N=572); panel b - HbSC (N=170) 37
Figure 9 Histogram of haemolytic index values for (a) HbSS/HbSβ0 and (b) HbSC patients 40
Figure 10 Association between haemolytic index and haemoglobin (all SCD patients)..... 40
Figure 11 Hospitalisation rate (i.e. hospital admission frequency) for the KCH clinical cohort, broken down by sickle genotype. Panel A – stacked histogram; panel B – box plot..... 43
Figure 12 Association between hospitalisation rate and haemoglobin (all SCD patients) 43
Figure 13 Urinary albumin creatinine ratio (mg/mmol) in (a) HbSS patients (b) HbSC patients 45
Figure 14 Histograms to compare HbF% when pregnant/not pregnant 62
Figure 15 Distribution of HbF% in steady state and acute settings for all sickle genotypes 63
Figure 16 Distribution of HbF% in steady state and acute settings for HbSS/HbSβ0 patients .. 63
Figure 17 Distribution of HbF% in steady state and acute settings for HbSC patients..... 64

2.1. Introduction: defining endpoints for genetic association studies

Accurate phenotyping is crucial for genetic association studies. It is imperative to define clear, reproducible, meaningful endpoints so that the true genetic component can be teased out. In case-control genetic studies, this means being able to distinguish "cases" and "controls". In sickle cell disease (SCD), the clinical heterogeneity makes defining accurate phenotypes complex (Rees et al., 2010, Ballas et al., 2012, Smith-Whitley et al., 2007). Disease severity is hugely variable in SCD, even within a genotype. 50 years ago SCD was considered by most as a "disease of childhood" (Dacie, 1960), but even then there were reports of long survival (>59 years) with infrequent or absent pain episodes and little or no end organ impairment (Charache and Richardson, 1964, Sydenstricker et al., 1962).

Phenotypes may be specific end-organ complications, "sub-phenotypes". Many patients develop isolated SCD-related clinical complications – stroke, proteinuria, osteonecrosis, pulmonary hypertension – but are otherwise well.

Phenotypes can also include clinical and laboratory parameters. While laboratory parameters are simple to measure, many values vary with the clinical state of the patient; for example, lactate dehydrogenase (LDH) is normally elevated during steady-state, and further increases during acute clinical events. Laboratory "intermediate" phenotypes (such as HbF%) are measurable and reproducible, and disease-related, and have proven much more successful in genetic association in SCD studies than clinical endpoints.

Markers of global SCD severity (as opposed to single organ complications) represent the "holy grail" of accurate clinical phenotyping. However, global severity scores have proved particularly difficult to define – there is no accepted, validated severity index. The difficulties arise from the clinical complexity of SCD: in order to create an index of global severity one needs to take account of the severity of each individual organ dysfunction within a patient. Examples of previously proposed global severity scores in SCD include:

- number of clinical presentations with acute pain episodes per year;
- transfusion requirements;
- a "severity index" based on frequency of painful crises, hospitalisation, blood transfusion, infection and specific complications (el-Hazmi, 1992);
- dactylitis in infants, white cell count and Hb (Miller et al., 2000);
- a global severity score using a Bayesian network model (a "statistical" phenotype) (Sebastiani et al., 2007).

I have defined four markers of "global" SCD severity for use in genotype-phenotype association analysis: fetal haemoglobin, frequency of acute pain episodes, level of haemolysis and mortality. I have assembled a large clinical database which includes these parameters, as well as other demographic and laboratory data.

2.2. Subjects

My work focused on two overlapping cohorts: a **regional research sickle gene bank** and the **King's College Hospital (KCH) clinical cohort**. Figure 1 demonstrates the overlap between these two cohorts.

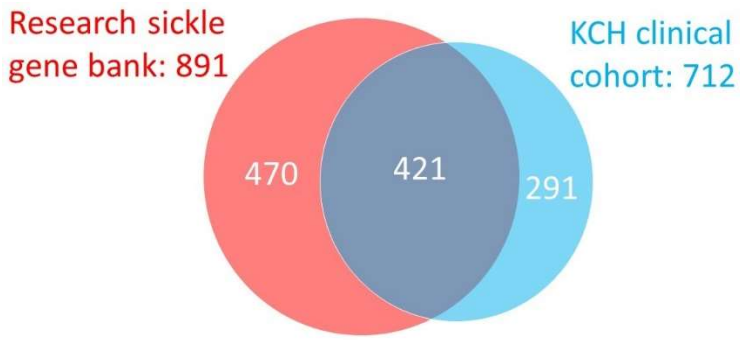


Figure 1 Overlap of two cohorts: research sickle gene bank and King's College Hospital adult clinic

Recruitment into the research sickle gene bank is ongoing in the hope that many of the KCH clinical cohort will become subjects in genetic research.

2.2.1. Regional research sickle gene bank

The regional **research sickle gene bank** comprises 891 SCD patients for which we have performed genome-wide genotyping ("MEGA" array, see chapter 3). This comprises patients from four adult clinics and one paediatric clinic in South East London: KCH adult clinic, KCH paediatric clinic, Guys and St Thomas' Hospital (GSTT) adult clinic, Lewisham Hospital adult clinic (LH) and Queen Elizabeth Hospital (Woolwich, QEH) adult clinic. Written informed consent was obtained through three approved study protocols (LREC 01-083, 07/H0606/165, and 12/LO/1610). Demographics of the regional sickle cohort are displayed in Figure 2.

Of the 891 (516 female, 375 male) patients with MEGA data, 666 (75%) were HbSS (375 female, 291 male), 195 (22%) were HbSC (126 female, 69 male), 20 (2%) were HbSβ^+ thalassaemia (10 female, 11 male), 9 (1%), HbSβ^0 thalassaemia (4 female, 5 male) and 1 (<1%) was HbS/HPFH.

The median recruitment age for HbSS/Sβ^0 patients was 28 years (range 4-71 years, IQR: 20-38); for HbSC, 37 years (range 7-77, IQR: 27-48); and HbSβ^+ thalassemia, 47 years (range 17-81

years, IQR: 32-55). α-globin genotypes were available in 377 (45%) patients of which 63% were αα/αα, 31% αα/α-, and 5% α-/α- genotypes.

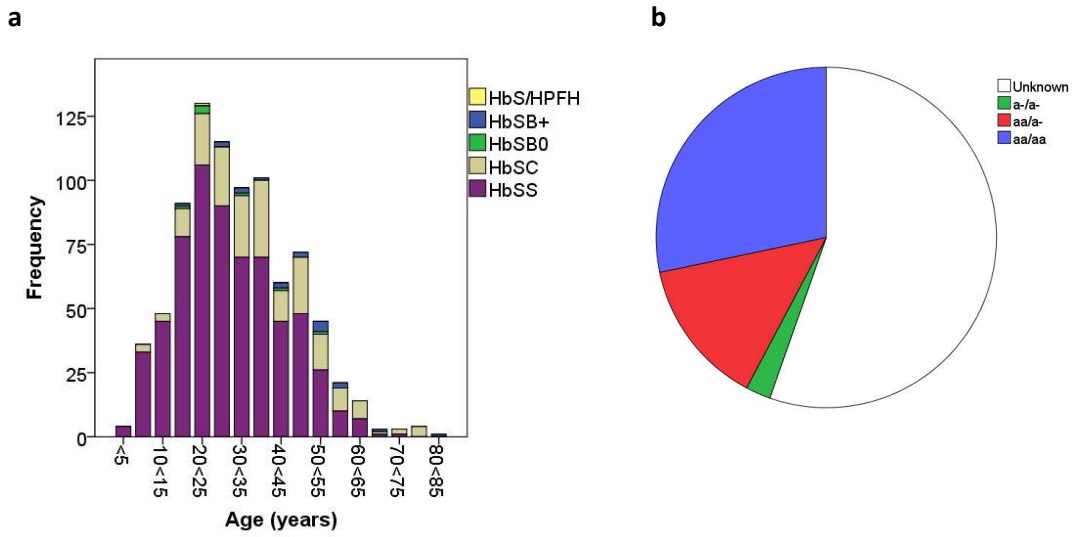


Figure 2 Demographics of the regional sickle gene bank: panel (a) demonstrates the distribution of sickle genotype across age; panel (b) demonstrates the α-globin status.

2.2.2. KCH clinical cohort

The **KCH clinical cohort** includes 712 patients seen in the adult clinic at King's College Hospital (London, United Kingdom), observed over a 10-year period (2004-2013 inclusive). Demographics of the KCH clinical cohort are displayed in Figure 3. Of the 712 patients, 444 (62%) were HbSS, 229 (32%) were HbSC, 33 (5%) were HbSβ^+ thalassemia, and 6 (1%) were HbSβ^0 patients. The median age for HbSS/Sβ^0 patients was 32 years (IQR: 25-43); HbSC, 39 years (IQR: 29-48); and HbSβ^+ thalassemia, 40 years (IQR: 31-58). α-globin genotypes were available in 542 (76%) patients of which 62% were αα/αα, 32% αα/α-, and 5% α-/α- genotypes. During the 10-year study period, 72 patients (all HbSS) had received hydroxycarbamide therapy, and 71 patients had received regular blood transfusions. All patients, except for one, were of African or African-Caribbean heritage. I have looked at the KCH clinical cohort (or a subset) for multiple clinical audit projects(Birkeland et al., 2016, Gardner et al., 2016, Gardner and Thein, 2015, van Hamel Parsons et al., 2016, Vidler et al., 2015). The clinical research was all done as an audit of clinical practice and therefore informed consent was not required.

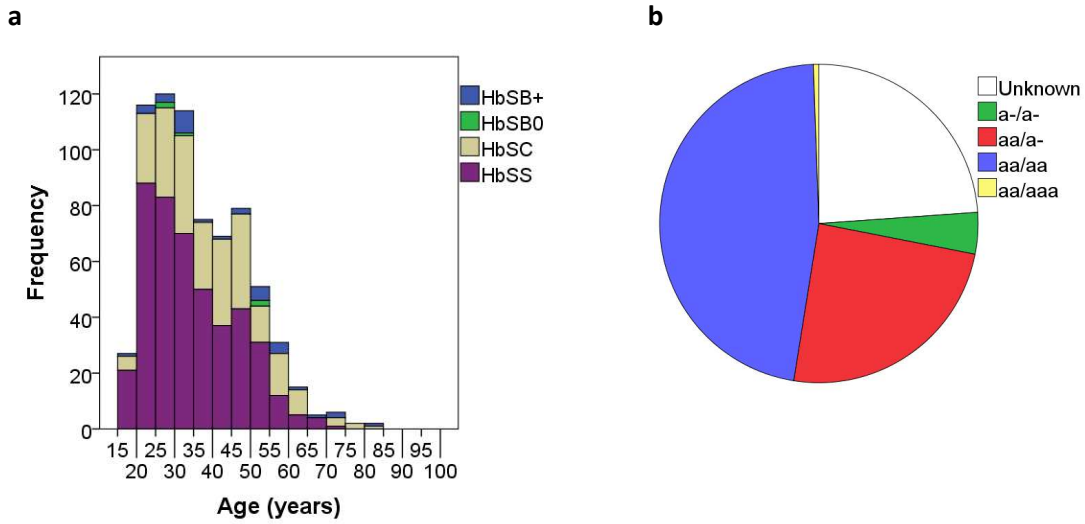


Figure 3 Demographics of the KCH clinical cohort: panel (a) demonstrates the distribution of sickle genotype across age; panel (b) demonstrates the alpha globin status.

2.3. The KCH clinical cohort: creating and curating a clinical dataset

2.3.1. Background

Clinical data management is critical for research which requires high-quality and reliable collection, integration and availability of data. Depth and breadth of clinical data allows us to understand and create better phenotypes. A comprehensive phenotype of laboratory and clinical data allows us to characterise the clinical outcome more accurately, as other values can be used as covariates. In my genotype/phenotype analyses these will be used:

- as clinical endpoints (phenotypes) directly
- to create new endpoints indirectly e.g. as a composite score (haemolytic index)
- as covariates.

2.3.2. Raw clinical data

Clinical data for the 712 adult patients in the KCH clinical cohort were obtained from the "electronic patient record" (EPR) which included all investigations undertaken from 2004. This included laboratory results (e.g. full blood count, microbiology), radiology results (e.g. computed tomography scan, echocardiogram), and other tests and procedures (e.g. electrocardiogram, lung function testing, oesophago-gastro-duodenoscopy). A "test" represents anything with a *single* result – so a full blood count includes 17 separate "tests" (e.g. haemoglobin, platelet count, white cell count, neutrophil count). See Appendix 1 for the full list of 2546 tests.

A total of 1,762,163 tests were provided as raw data for 712 patients (mean 2475 tests per patient). Raw data were provided as csv files for all medical results available (e.g. laboratory results, radiology, lung function testing), see Figure 4. This included each individual test result by line, so for any one date, one patient has many rows. However, it only included the actual

results for tests that have *numerical* results (mainly laboratory results) and not text-based tests (which return as NULL).

[redacted]								
Sodium	139	13/03/2003	NULL	Results	12/06/2000	12/06/2000	Results	
Potassium	4.3	13/03/2003	NULL	Results	12/06/2000	12/06/2000	Results	
Creatinine	61	13/03/2003	NULL	Results	12/06/2000	12/06/2000	Results	
Glucose (f	5.6	13/03/2003	NULL	Results	28/11/2001	28/11/2001	Results	
High Flour	4.1	13/03/2003	NULL	Results	13/07/2000	12/07/2000	Results	
Medium F	16.9	13/03/2003	NULL	Results	13/07/2000	12/07/2000	Results	
MCV	60.1	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Comment	NULL	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
MCH	20	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
MCHC	33.2	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
RDW	17.1	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
PLT	237	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
MPV	6	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Neutroph	15.73	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Lymphocy	1.03	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Monocyte	0.75	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Eosinophi	0.02	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Basophils	0.06	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Nucleatec	<0.2%	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
%HYPO	2.8	24/09/2007	24/09/2007	Accident and Emergency	25/09/2007	24/09/2007	Emergency	
Urine - mi	C&S	NULL	24/09/2007	44:00.0	Accident and	25/09/2007	44:45.0	Emergency
Urine - mi	C&S	NULL	24/09/2007	44:00.0	Accident and	26/09/2007	44:45.0	Emergency
Hb	9	13/03/2003	NULL	Results	12/06/2000	12/06/2000	Results	
Neutroph	3.77	22/06/2004	NULL	Haematology OPD	22/06/2004	22/06/2004	Outpatient	
Lymphocy	2.79	22/06/2004	NULL	Haematology OPD	22/06/2004	22/06/2004	Outpatient	
Histology	NULL	13/03/2003	NULL	Results	02/11/2004	19/10/2004	Results	
Histology	NULL	13/03/2003	NULL	Results	02/11/2004	19/10/2004	Results	
Histology	NULL	13/03/2003	NULL	Results	02/11/2004	19/10/2004	Results	
Histology	NULL	13/03/2003	NULL	Results	02/11/2004	19/10/2004	Results	

Figure 4 Raw clinical data format
Hospital number (redacted), test name, rest result, visit start date (e.g. clinic visit date / hospital admission date), visit end date (e.g. hospital discharge date, NULL if clinic visit only), location (e.g. clinic name, ward name, "results" if this was an externally requested test), date test result released, date test requested, visit type

2.3.3. Data cleaning pipeline

The data was batched and each batch subjected to the same data cleaning and processing pipeline, see Appendix 2. The data are presented on a per patient, per clinic visit level. The final clinical information available in the clinical dataset is in Appendix 7. As well as laboratory results, these data include demographic information. Analysis from subsequent sections on hospital admission data and haemolytic indices has also been added to this central database.

2.3.4. Results

712 patients had 6839 outpatient visits and 2819 inpatient admissions during the 10-year study period of 2004-13 inclusive. Overall, there were 4246 patient-years of observation in the 10-year study period. Not all patients were observed for all 10 years; some patients entered or left the adult haematology clinic during the study period (e.g. through death, house move or simply lost to follow up). See the infographic demonstrating the range of an individual's years of observation over the 10-year study period in Appendix 3.

Examples of basic laboratory results (as raw data) are summarised for 450 HbSS/HbSβ^0 thalassaemia patients in Figure 5 and for 229 HbSC patients in Figure 6.

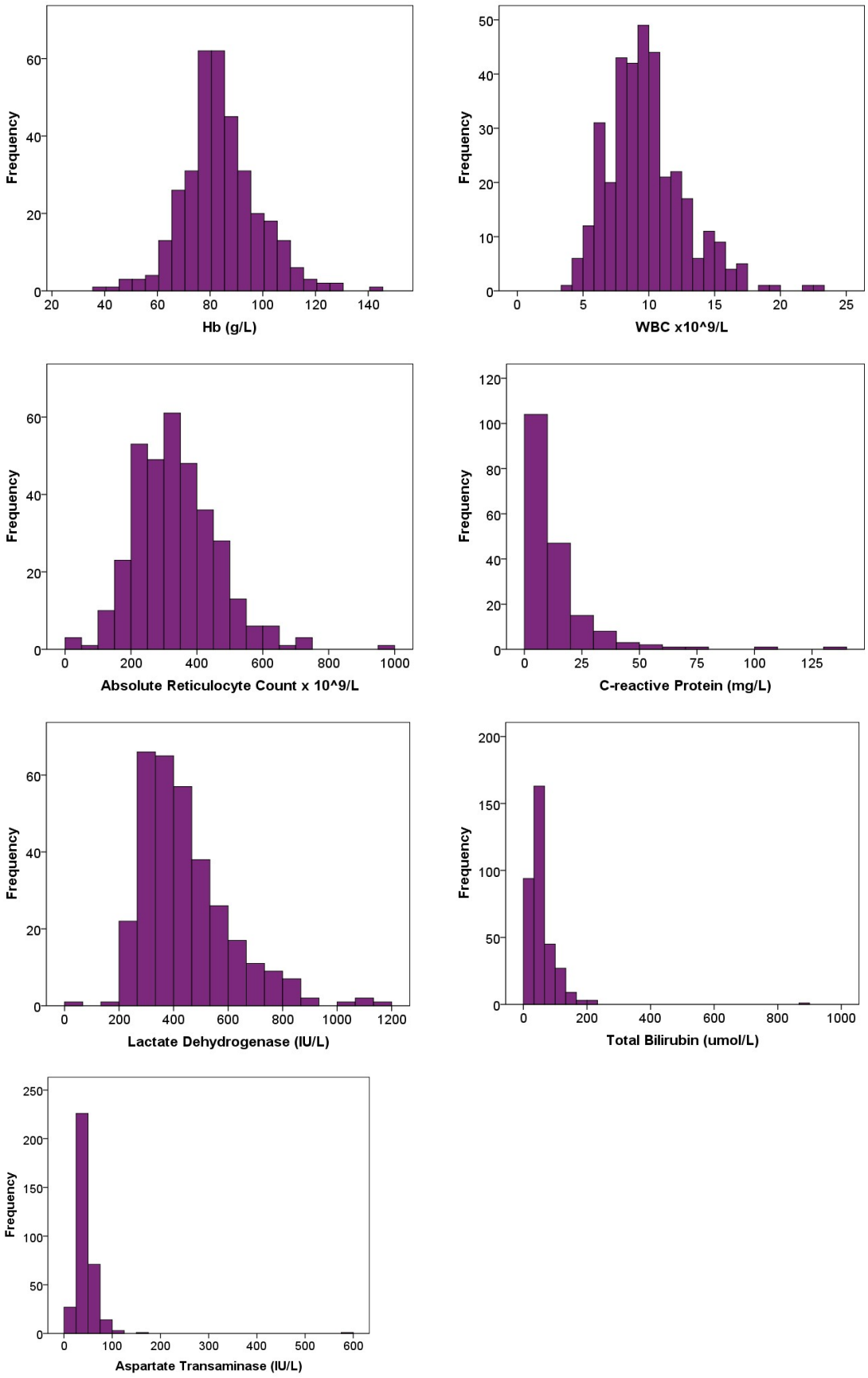


Figure 5 Basic laboratory results for HbSS and HbSβ0 patients

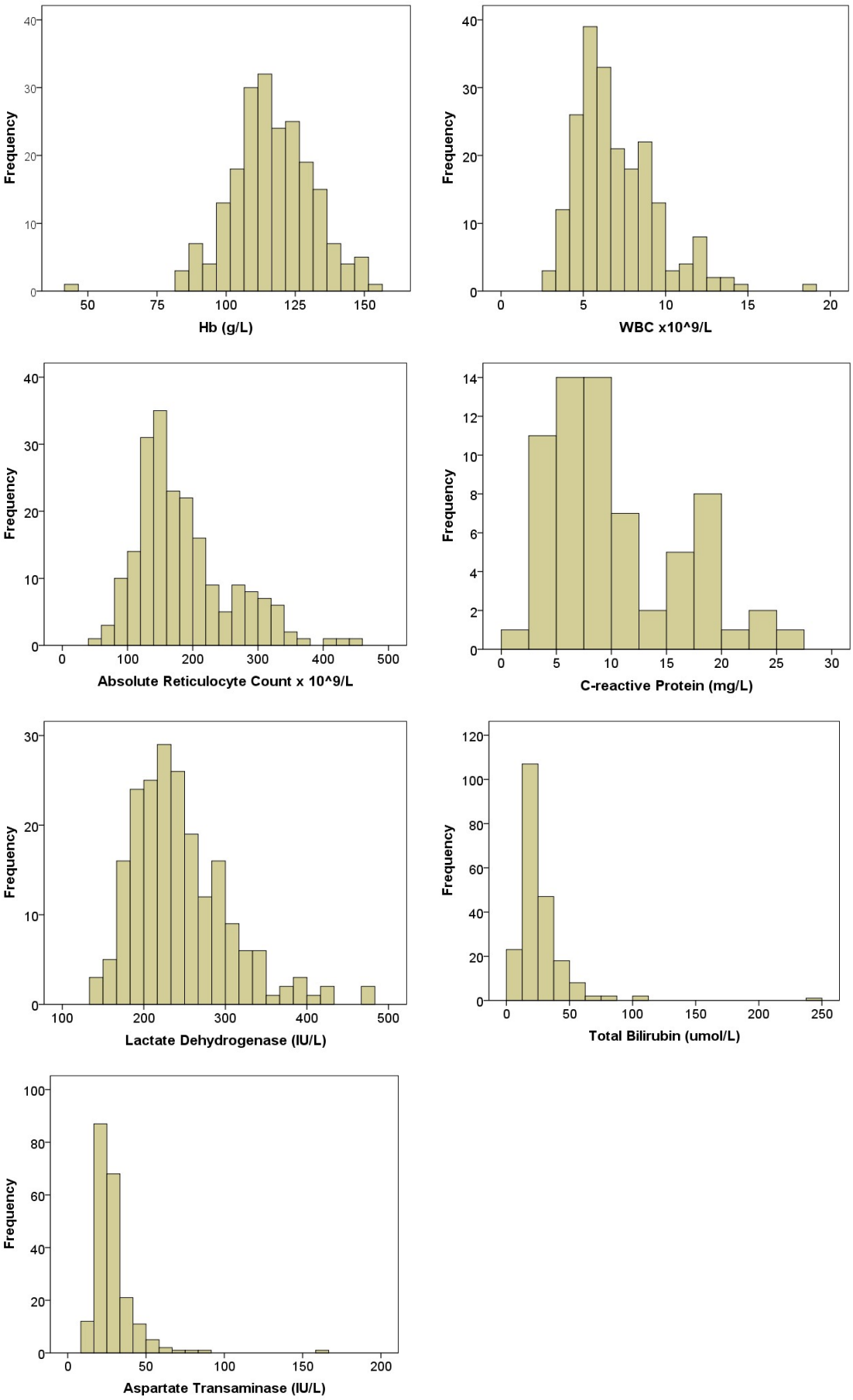


Figure 6 Basic laboratory results for HbSC patients

2.3.5. Adding more clinical data

The clinical database is not static. More information and results are added as they are collected, so that the subsequent sections have all been added to the original database; this is a dynamic process.

2.4. Fetal haemoglobin (HbF)

2.4.1. Introduction

Fetal haemoglobin (HbF, α2γ2) is a major ameliorating factor in SCD. Its use as a phenotype in SCD in genetic association studies is well-established. Investigation into causes of HbF% variability and genetic regulation is an ongoing endeavour: its therapeutic reactivation (via pharmacological and genetic approaches) remains a top research priority. See section 1.2.4.4 for a longer description of HbF% as a phenotype in SCD, as well as its well-established three known quantitative trait loci.

In both SCD and non-SCD populations, HbF% is influenced by both sex and age at sampling. Unlike non-SCD populations in whom adult levels of HbF% are reached by 2 years of age, subjects with SCD achieve adult levels much later, and adult HbF% levels are higher in SCD than in a healthy population (Mason et al., 1982). In SCD, an HbF% age cut-off of 5 years has been used by researchers; after this age, there is a negative, roughly linear, correlation (Mason et al., 1982). Sex also plays a role: females have been found to have higher HbF% levels than males in both healthy individuals(Miyoshi et al., 1988) and in SCD(Steinberg et al., 1995, Nagel, 1991, Morris et al., 1991).

Pregnancy is also associated with elevated HbF% levels in both healthy women (Pembrey et al., 1973) and women with SCD (Dunn et al., 1989), probably mediated via progesterone and physiology of increased erythropoiesis. In the Jamaican SCD cohort, there was a significant HbF% increase in the first two trimesters above steady state, followed by decrease to below steady state in the third trimester(Dunn et al., 1989). I confirmed the finding of significantly increased HbF% in the first trimester compared to non-pregnancy state in our own SCD cohort ($p<0.0001$), see Appendix 8.

In SCD, HbF% levels are further influenced by drugs and transfusion status. Drugs include hydroxycarbamide, the key disease-modifying therapy in SCD used precisely because of its therapeutic induction of HbF%. Other HbF%–inducing agents include: hypomethylating agents (5-azacitidine, decitabine), pomalidomide, and butyrate. Transfusions alter HbF% levels by infusing healthy (HbAA) blood, which typically has reduced HbF% (<1%) compared to SCD

patients. Finally, it seems feasible that an acute pain episode could affect HbF% levels in SCD; theoretically, increased stress erythropoiesis could cause faster release of immature erythrocytes (with higher HbF%) from the bone marrow. I investigated this in 127 patients with paired (steady state versus inpatient) and otherwise valid HbF% samples which revealed no statistical difference in the mean HbF% in steady state versus during an acute pain episode, see Appendix 9.

2.4.2. Methods

HbF% levels are typically measured by high performance liquid chromatography (HPLC). In healthy individuals HbF% is typically low, with the majority (85-90%) having <1% HbF% levels. However, in SCD, HbF% levels are 1% - 30%, which can be measured accurately by HPLC.

All haemoglobin profiles (including HbF% levels) were generated with HPLC (Variant II Hemoglobin Testing System). Before using an HbF% value, I confirmed that it reflected "native" (baseline) HbF% i.e. that it was obtained when the patient was transfusion-free for at least 3 months, not taking hydroxycarbamide or other HbF%-inducing drug for at least 3 months, and not pregnant. Samples taken in steady state were preferred to those taken during episodes of acute pain, however, samples from acute pain were acceptable if no other values were available.

Regional sickle gene bank

HbF% values were assayed at KCH for all patients except for GSTT patients where they were tested locally at GSTT.

2.4.3. Results

2.4.3.1. KCH clinical cohort

Of 712 patients, 645 patients had a validated HbF% level available. The median HbF% was 3.2% (range: 0.1-31.6%, IQR: 1.4-7.25%) for all genotypes. For 450 HbSS/SB0 patients, 393 had a validated HbF% with a median of 5.4% (range: 0.2-29.5%, IQR: 2.6-9.5%). For 229 HbSC patients, 220 had a validated HbF%, with a median of 1.1% (range: 0.1-11.1%, IQR: 0.6-2.2%). The distribution of HbF% levels in the KCH clinical cohort is shown in Figure 7.

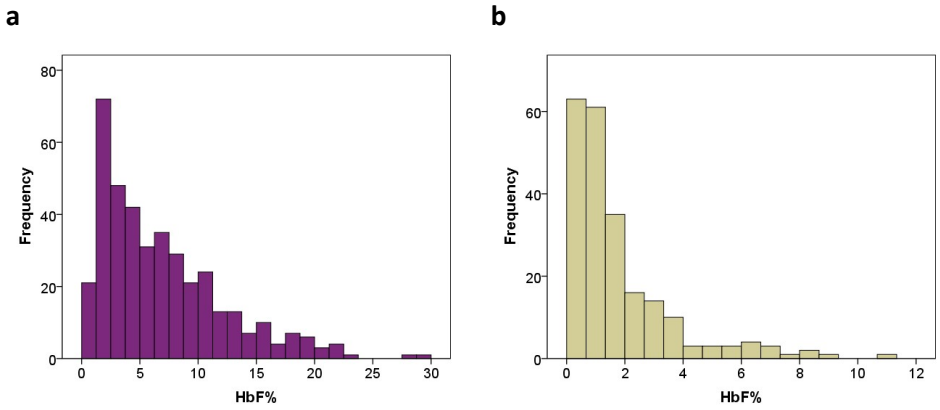


Figure 7 validated HbF% values for KCH clinical cohort: panel a - HbSS/SB0 (N=393); panel b - HbSC (N=220)

2.4.3.2. Regional sickle gene bank

Validated HbF% levels for the research cohort are shown in Figure 8. Of 891 patients, 760 patients had a validated HbF% level available, median HbF% was 4.6% (range: 0.2-31.6%, IQR: 1.9-8.9%) for all genotypes. 572 HbSS/HbSβ^0 thalassaemia patients had a validated HbF%, with median 6.2% (range: 0.2-29.5%, IQR: 3.1-10.5%). 170 HbSC patients had a validated HbF%, median 1.3% (range: 0.2-13.4%, IQR: 0.79-2.5%).

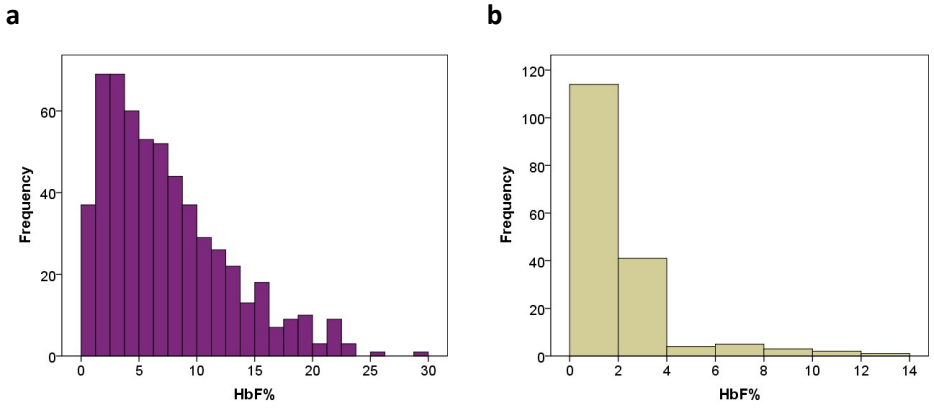


Figure 8 validated HbF% values for the regional sickle gene bank: panel a - HbSS/SB0 (N=572); panel b - HbSC (N=170)

2.4.3.3. Missing/lack of validated HbF data

For both cohorts, not all patients had validated clinical data available (i.e. not taken post-transfusion, while on hydroxycarbamide or while pregnant). The large majority of those without a valid HbF% available were on either long-term hydroxycarbamide or transfusions; thus the "missing" HbF% data is skewed towards clinically severe patients.

2.4.4. Validation of the KCH clinical dataset

"Validation" of results has been fed back into the KCH clinical dataset. A flag has been added to denote the "validation" process of samples that are not affected by transfusion, hydroxycarbamide or pregnancy.

2.5. Haemolysis: 10-year data

2.5.1. Introduction: a haemolytic index

Haemolysis and sickle vaso-occlusion are the two major pathophysiological drivers for SCD clinical manifestations (Rees et al., 2010). Patients with more haemolysis (evidenced by higher lactate dehydrogenase, lower Hb, and higher bilirubin) tend to have increased incidence of leg ulcers, priapism and pulmonary hypertension. Haemolytic-related pathologies in SCD are postulated to be mediated by nitric oxide (NO) bioavailability (Kato et al., 2007, Rees et al., 2010). Plasma free haemoglobin is a specific marker of intravascular haemolysis; red blood cell survival is the definitive haemolysis measurement. These measurements are not feasible in large cohorts. Instead haemolysis, in clinical practice and in research, is estimated by the reticulocyte count, lactate dehydrogenase (LDH), aspartate aminotransferase (AST) and bilirubin levels, all of which are commonly measured in cohort studies, although none is specific for haemolysis (Hebbel, 2011). Ameliorators of sickled red cell lifespan include high HbF% and α-thalassaemia (de Ceulaer et al., 1983, Steinberg and Sebastiani, 2012) but other genes are likely to modify red cell life span i.e. mediate haemolysis.

I created a "haemolytic index" – a single continuous variable that quantitates haemolysis by using a principal component analysis of the commonly measured markers of haemolysis(Minniti et al., 2009, Gordeuk et al., 2009) – reticulocyte count, LDH, AST and total bilirubin levels. The development of a haemolytic index resolves the problem of dealing with correlated predictors in multivariate analyses, and has been used previously in a genome wide association study(Milton et al., 2013). Milton *et al* identified an association between haemolysis and a single nucleotide polymorphism in *NPRL3* on chromosome 16 (and associated with $-\alpha^{3.7}$ thalassaemia gene deletion). The principal components analysis yielded one component – the *haemolytic index* – which was associated with intravascular haemolysis as measured by plasma haemoglobin and red cell micro-particles(Nouraie et al., 2013). Therefore, the score can be used as a robust quantitation of haemolytic rate.

I used a principal components analysis approach to transform four clinical laboratory markers of haemolysis (reticulocyte count, LDH, AST and total bilirubin levels) into one "haemolytic index". This essentially reduced the data from four parameters down to one. The haemolytic index can then be used as a phenotype in a genotype-phenotype association studies.

2.5.2. Methods

2.5.2.1. Principal component analysis (PCA)

Principal Component Analysis (PCA) is a dimension reduction technique for quantitative variables invented in 1901 by Karl Pearson(Pearson, 1901). It results in one or more components which summarise the information available in the data. The components are linear combinations of the original variables: the analysis transforms a set of *correlated* variables into a set of *uncorrelated* theoretical variables termed "principal components". The resultant number of principal components is a maximum of the number of original variables. PCA is useful for studying underlying mechanisms reflected in individual biological measurements(Genser et al., 2007).

Assumptions are made when performing this technique, and need to be validated prior to performing the analysis. The PCA can only perform a compression of the available information if there is redundancy within the original variables – that is, the variables are correlated. Two techniques can be applied to check this: Bartlett's test and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. Bartlett's test of sphericity confirms that the correlation matrix diverges significantly from the identity matrix; i.e. the extent to which the original variables are correlated. The KMO measure of sampling adequacy also checks if one can factorise efficiently the original variables. The KMO index uses the partial correlation between variables (to measure the relation between two variables by removing the effect of the remaining variables) and compares the values of correlations between variables and those of the partial correlations. The PCA can act efficiently if the KMO index is high (~1) but not low (~0).

I performed the PCA in SPSS version 22, on the KCH clinical cohort of 712 patients, using the four variables reticulocyte count, LDH, AST and total bilirubin levels.

2.5.3. Results

544 of 712 patients had validated four variables for analysis available on the same day. Both Bartlett's test of sphericity (chi-square = 471, df=6, $p<0.001$) and the KMO measure of sampling adequacy (0.692) suggest that the available data could be compressed.

Principal components analysis resulted in a first component with eigenvalue 2.159 (explaining 54% of variability) and subsequent components having eigenvalues<1. Thus I extracted the only first component as the "haemolytic index". The new haemolytic index has a mean 0 and standard deviation of 1 (normalised by design), see Figure 9 for a distribution of haemolytic

index values. This haemolytic index had correlations of r=-0.755 with Hb (p<0.0001), r=0.439 with absolute reticulocyte count (p<0.0001), r=0.870 with lactate dehydrogenase levels (p<0.0001), and r=0.608 with total bilirubin levels (p<0.0001).

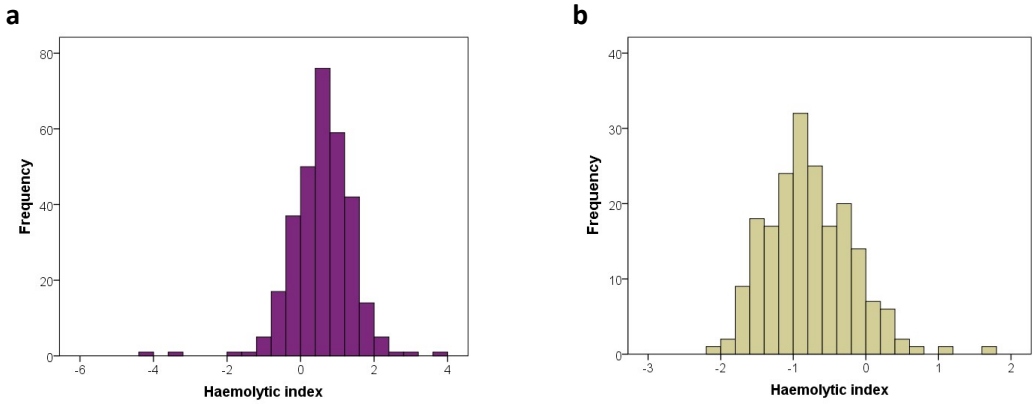


Figure 9 Histogram of haemolytic index values for (a) HbSS/HbSβ0 and (b) HbSC patients

The haemolytic index is correlated with haemoglobin levels (Pearson's r=-0.585, p<0.001), see Figure 10.

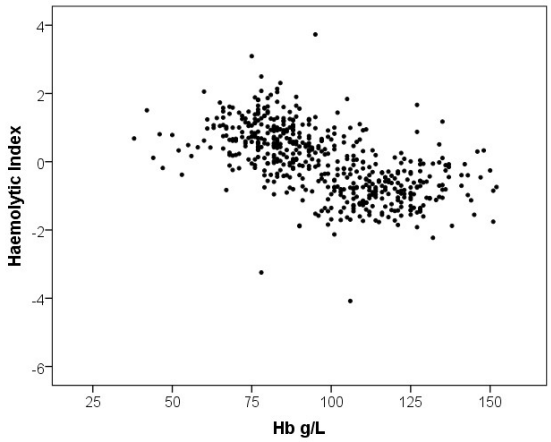


Figure 10 Association between haemolytic index and haemoglobin (all SCD patients)

2.5.4. Discussion

The haemolytic index can now be used as a theoretical variable quantitating haemolysis having incorporated information from the reticulocyte count, lactate dehydrogenase (LDH), aspartate aminotransferase (AST) and bilirubin levels.

2.6. Acute pain frequency: hospitalisation rate from 10-year data

2.6.1. Introduction

Acute pain episodes (APE) are the hallmark clinical feature in SCD. They are a measure of disease severity and a predictor of early mortality (Platt et al., 1991). Frequency of APE varies widely in SCD patients, with highest pain rates seen in those with high haematocrit and low HbF% (Platt et al., 1991). Outwith these associations, there is no concrete further understanding of the genetic basis of APE frequency in SCD. It is probably the complication

most affected by environmental factors. A compounding problem with pain studies is the clinical definitions of phenotypes. Nearly all patients with SCD have pain, and it is often difficult to quantitate objectively both frequency and severity of individual APEs. Furthermore, the standard treatment for pain in APEs is parenteral opioids, and individual response to opioid analgesia is itself related to genetic variability of their metabolism (Ballas, 2007), making it harder still to dissect and measure APE accurately. As a result of these complicating features, many genetic studies on pain in SCD are poor, in particular because of lack of clear-cut definitions of *cases* versus *controls* required to make objective associations. Furthermore, some of the studies described are poorly conducted and not corrected for other key modifying factors including genotype and HbF% levels. In African-American patients and patients from Cameroon, the presence of alleles at the 3 HbF% loci (*BCL11A*, *HBS1L-MYB*, and *Xmn*I-*HBG2*) that increased HbF% levels also led to a corresponding reduction in APEs and hospitalisation (Lettre et al., 2008, Wonkam et al., 2014).

Studies into genetic determinants of APEs have focused on candidate genes based on APE pathology, itself a complex event involving: red cell deformation, enhancement of white cell adhesion, inflammation, endothelial injury and activation of the coagulation and complement pathways. These genetic studies are generally uncorroborated and not replicated in secondary cohorts. Examples of studies relating to APE in SCD include genes related to:

- *Oxidative stress*. SCD complications, and notably APE, are associated with oxidative stress. Glutathione S-transferases (GSTs) are a group of enzymes that protect against oxidative stress. Shiba *et al* found the *GSTM1* null genotype to be associated with increased risk of severe APE in Egyptian SCD patients (Shiba et al., 2014)
- *Vasculopathy*. Vascular endothelial growth factors (VEGF) are known to contribute to the pathogenesis of APE in SCD. A study in Bahrain associated multiple VEGF gene polymorphisms with the risk of APE (Al-Habboubi et al., 2012). Unfortunately, the differences between cases and controls were not clear cut (comparing whether patients with SCD had a recent APE or not).
- *Thrombosis*. Cystathionine beta-synthase (CBS) enzyme gene mutations are a risk factor for thromboembolic disorders. CBS 844ins68 was three times more frequent among SCD patients with APE (Alves Jacob et al., 2011). Again, there was poor clarification of the difference between “severe” and “mild” individuals with APE.
- *Infections*. *MBL2* codes for mannose-binding lectin (MBL), and is associated with modifications in the progression of infectious and inflammatory vascular diseases. Using better definitions of APE severity (using APE frequency), *MBL2* polymorphisms have been associated with APE in children with SCD (Oliveira et al., 2009, Mendonca et

al., 2010). Unexpectedly, studies have observed no association of *MBL2* variants with susceptibility to infections (Oliveira et al., 2009)(Dossou-Yovo et al., 2009).

While the broad idea of frequency of APE in SCD as a marker of severity is well-accepted, the implementation of a precise, quantitative variable has been more difficult to implement. Questions surrounding APE definition (e.g. minor crisis versus major crisis), consideration of duration of APE, background issues of non-medical causal factors to APE precipitation, all contribute to an indistinct quantity. As have others, I have chosen to look at frequency of hospitalisation – admissions to haematology wards (and not emergency department visits). This definition allows us not only to use a precise event to count *per* patient but also makes the decision to admit (as a proxy of clinical severity) physician-determined rather than patient-determined. I reviewed 10 year patient records to create a "hospitalisation rate *per* year" variable, as has been used previously, quoted as *per* 100 patient by researchers in the STOP (Miller et al., 2001) and SITS (DeBaun et al., 2014) trials.

2.6.2. Methods

Hospitalisation rates were collected for KCH adults over 10 years (2004-2013) using data gleaned in section 2.3. "Hospitalisation" included any admission to a *haematology* or *general medical* ward (and excluded admissions to the emergency department and orthopaedic, surgical, obstetric and paediatric wards), as per Appendix 6. The reason for hospital admission was not identified (therefore patients with medical non-sickle related problems were included e.g. myocardial infarction). An individual's mean hospitalisation rate (haematology hospital admissions only) was calculated by dividing the number of haematology hospital admissions by the number of observed years.

2.6.3. Results

Of the 712 patients (all genotypes), 465 had been hospitalised at least once, and 247 patients had no hospital admissions during the 10-year observation period. The distribution is heavily positively skewed and is also related to the underlying sickle genotype, see Figure 11.
For the whole cohort, median mean hospitalization rate was 0.25/year (range 0-11.25/year, IQR: 0-0.71). For the 450 patients with HbSS/HbSβ^0, median mean hospitalization rate was 0.35/year (range 0-11.25/year, IQR: 0-1.0); and for the 229 HbSC patients, median mean hospitalization rate was 0.1/year (range 0-2.11/year, IQR: 0-0.4).

a **b**

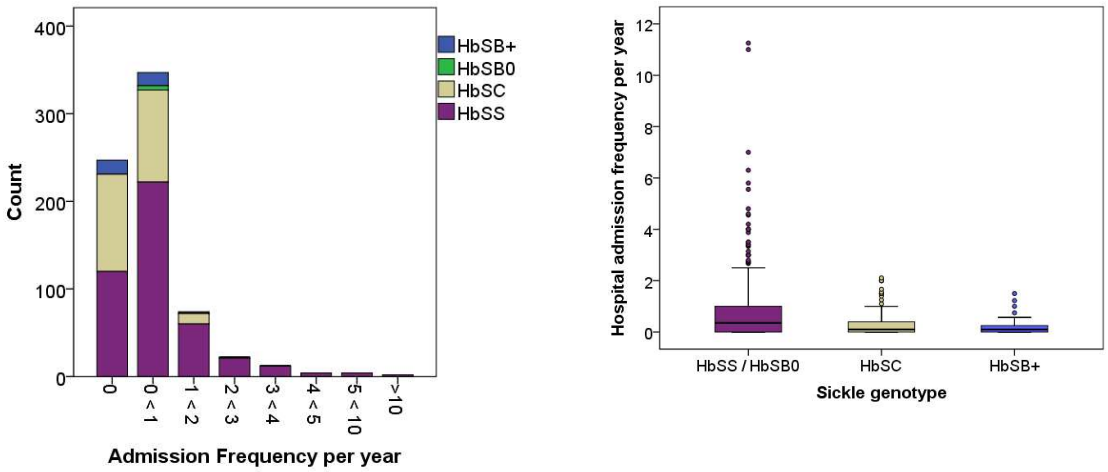


Figure 11 Hospitalisation rate (i.e. hospital admission frequency) for the KCH clinical cohort, broken down by sickle genotype. Panel A – stacked histogram; panel B – box plot.

Hospitalisation rate is very weakly correlated with haemoglobin levels (Spearman's ρ=-0.185, p<0.001), see Figure 12.

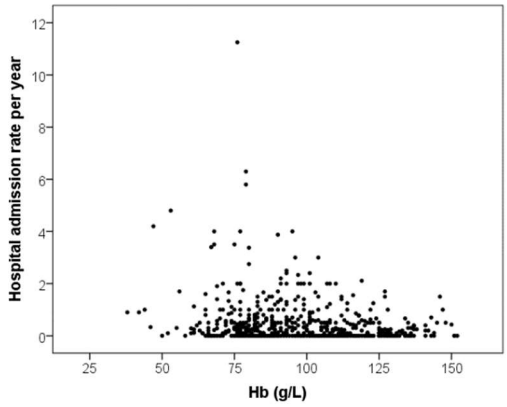


Figure 12 Association between hospitalisation rate and haemoglobin (all SCD patients)

2.7. Survival | Mortality

Mortality is the ultimate arbiter of any disease severity. We have published survival data, including looking at a Cox regression analysis of risk factors for mortality, for the KCH clinical cohort of 712 adults(Gardner et al., 2016), see Appendix 10.

2.8. Sickle nephropathy

2.8.1. Introduction

Renal impairment – as measured by either proteinuria or glomerular filtration rate (GFR) – is a common complication of SCD (Sharpe and Thein, 2014, Nath and Hebbel, 2015), and in some cases sickle renal disease progresses to end-stage renal failure. Renal damage is due to the underlying environment of the renal medulla; where low partial pressure of oxygen, low pH and high osmolality combine to create optimum conditions for HbS polymerisation and sickling (Sharpe and Thein, 2011). This leads to recurrent vaso-occlusion and chronic ischaemia

resulting in papillary necrosis and medullary fibrosis (focal segmental glomerulosclerosis). Renal dysfunction is also associated with severity of haemolysis (Becton et al., 2010, Maier-Redelsperger et al., 2010, Day et al., 2012). Thus, it is not too unexpected that co-inheritance of α-thalassaemia which reduces haemolysis, is protective against albuminuria (Nebor et al., 2010).

Renal impairment begins with glomerular hyperfiltration (seen in childhood and early adulthood) and protein loss in the urine (Scheinman, 2009, Becton et al., 2010). Microalbuminuria (urinary albumin creatinine ratio repeatedly >3.5 mg/mmol or 30 mg/g) marks the onset of sickle nephropathy, and its prevalence increases with age(McPherson Yee et al., 2011). In the KCH adult cohort, microalbuminuria was detectable in 28% of patients aged 16-25 years, 38% in 26-35 years, 50% in 36-45 years, and >60% in those aged at least 46 years(Day et al., 2012). In a small minority of patients, sickle nephropathy progresses to end stage renal failure but the natural history of renal disease is not characterised. Being able to stratify (the large number of) SCD patients with albuminuria into those at high risk of disease progression would impact on clinical management and may direct therapy.

The *APOL1* locus, an important genetic risk factor for end-stage renal failure in non-SCD populations of African ancestry (Genovese et al., 2010), has been shown to be associated with sickle cell nephropathy (Ashley-Koch et al., 2011). The original association of nephropathy with *MYH9* has been attributed to the strong linkage disequilibrium between *MYH9* and *APOL1*.

2.8.2. Methods

Urinary albumin creatinine ratio (uACR) results were collected in steady state for KCH adults over the 10 year period (2004-2013) using data gleaned in section 2.3. An individual's mean uACR level was calculated across the 10-year study period, and used in analysis.

2.8.3. Results

606 of 712 patients had uACR results available. Median uACR across all genotypes was 3.1 mg/mmol (range 0.2 to 1351, IQR 1.2-8.9). For the 403 HbSS/HbSβ^0 thalassaemia patients with uACR available, median uACR was 4.6 mg/mmol (range 0.3 to 1351, IQR 1.5-11.7), see Figure 13a. 240/403 (60%) of these patients have microalbuminuria (uACR≥3.5mg/mmol). For the 176 HbSC patients with uACR available, median uACR was 1.6 mg/mmol (range 0.2 to 134, IQR 0.8-3.4). 43/176 (24%) of these patients have microalbuminuria (uACR≥3.5mg/mmol), see Figure 13b.

a **b**

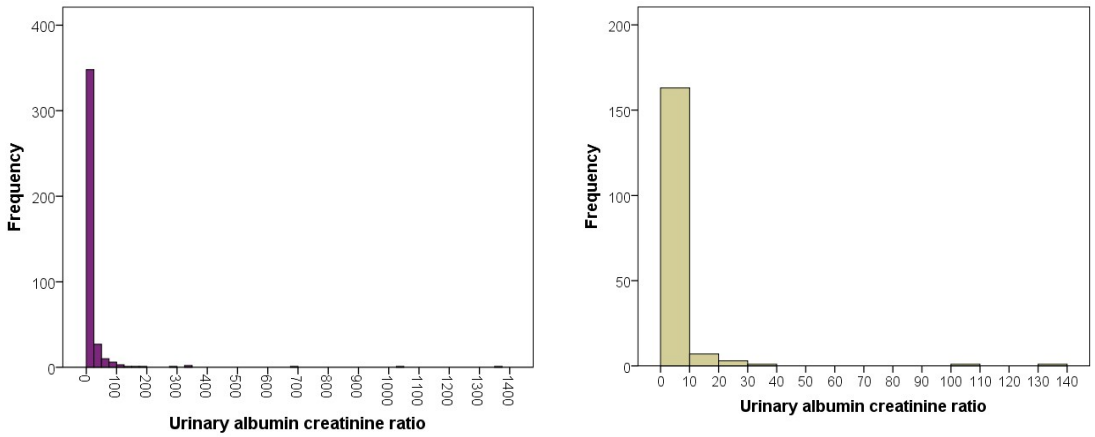


Figure 13 Urinary albumin creatinine ratio (mg/mmol) in (a) HbSS patients (b) HbSC patients

2.9. Discussion

There is no perfect marker of severity in sickle cell disease. I have defined several clinical endpoints, as others have done. There is strength in multiple phenotypes – to define severity as different traits allows us not only to understand better the complexity of phenotype, but strengthens our genotype/phenotype association studies. One can determine not only the genetic architecture of specific phenotypes but, by looking at the overlap, gain some understanding of global mechanisms contributing towards severity in SCD. Furthermore, there is more statistical power with multiple definitions of "severity".

References

AL-HABBOUBI, H. H., MAHDI, N., ABU-HIJLEH, T. M., ABU-HIJLEH, F. M., SATER, M. S. & ALMAWI, W. Y. 2012. The relation of vascular endothelial growth factor (VEGF) gene polymorphisms on VEGF levels and the risk of vasoocclusive crisis in sickle cell disease. *Eur J Haematol,* 89**,** 403-9.

ALVES JACOB, M., DA CUNHA BASTOS, C. & REGINA BONINI-DOMINGOS, C. 2011. The 844ins68 cystathionine beta-synthase and C677T MTHFR gene polymorphism and the vaso-occlusive event risk in sickle cell disease. *Arch Med Sci,* 7**,** 97-101.

ASHLEY-KOCH, A. E., OKOCHA, E. C., GARRETT, M. E., SOLDANO, K., DE CASTRO, L. M., JONASSAINT, J. C., ORRINGER, E. P., ECKMAN, J. R. & TELEN, M. J. 2011. MYH9 and APOL1 are both associated with sickle cell disease nephropathy. *Br J Haematol,* 155**,** 386-94.

BALLAS, S. K. 2007. Current issues in sickle cell pain and its management. *Hematology Am Soc Hematol Educ Program***,** 97-105.

BALLAS, S. K., KESEN, M. R., GOLDBERG, M. F., LUTTY, G. A., DAMPIER, C., OSUNKWO, I., WANG, W. C., HOPPE, C., HAGAR, W., DARBARI, D. S. & MALIK, P. 2012. Beyond the definitions of the phenotypic complications of sickle cell disease: an update on management. *ScientificWorldJournal,* 2012**,** 949535.

BECTON, L. J., KALPATTHI, R. V., RACKOFF, E., DISCO, D., ORAK, J. K., JACKSON, S. M. & SHATAT, I. F. 2010. Prevalence and clinical correlates of microalbuminuria in children with sickle cell disease. *Pediatr Nephrol,* 25**,** 1505-11.

BIRKELAND, P., GARDNER, K., KESSE-ADU, R., DAVIES, J., LAURITSEN, J., ROM POULSEN, F., TOLIAS, C. M. & THEIN, S. L. 2016. Intracranial Aneurysms in Sickle-Cell Disease Are Associated With the Hemoglobin SS Genotype But Not With Moyamoya Syndrome. *Stroke,* 47**,** 1710-3.

CHARACHE, S. & RICHARDSON, S. N. 1964. PROLONGED SURVIVAL OF A PATIENT WITH SICKLE CELL ANEMIA. *Arch Intern Med,* 113**,** 844-9.

DACIE, J. V. 1960. *The Haemolytic Anaemias: The congenital anaemias*, Grune & Stratton.

DAY, T. G., DRASAR, E. R., FULFORD, T., SHARPE, C. C. & THEIN, S. L. 2012. Association between hemolysis and albuminuria in adults with sickle cell anemia. *Haematologica,* 97**,** 201-5.

DE CEULAER, K., HIGGS, D. R., WEATHERALL, D. J., HAYES, R. J., SERJEANT, B. E. & SERJEANT, G. R. 1983. alpha-Thalassemia reduces the hemolytic rate in homozygous sickle-cell disease. *N Engl J Med,* 309**,** 189-90.

DEBAUN, M. R., GORDON, M., MCKINSTRY, R. C., NOETZEL, M. J., WHITE, D. A., SARNAIK, S. A., MEIER, E. R., HOWARD, T. H., MAJUMDAR, S., INUSA, B. P., TELFER, P. T., KIRBY-ALLEN, M., MCCAVIT, T. L., KAMDEM, A., AIREWELE, G., WOODS, G. M., BERMAN, B., PANEPINTO, J. A., FUH, B. R., KWIATKOWSKI, J. L., KING, A. A., FIXLER, J. M., RHODES, M. M., THOMPSON, A. A., HEINY, M. E., REDDING-LALLINGER, R. C., KIRKHAM, F. J., DIXON, N., GONZALEZ, C. E., KALINYAK, K. A., QUINN, C. T., STROUSE, J. J., MILLER, J. P., LEHMANN, H., KRAUT, M. A., BALL, W. S., JR., HIRTZ, D. & CASELLA, J. F. 2014. Controlled trial of transfusions for silent cerebral infarcts in sickle cell anemia. *N Engl J Med,* 371**,** 699-710.

DOSSOU-YOVO, O. P., ZACCARIA, I., BENKERROU, M., HAUCHECORNE, M., ALBERTI, C., RAHIMY, M. C., ELION, J. & LAPOUMEROULIE, C. 2009. Effects of RANTES and MBL2 gene polymorphisms in sickle cell disease clinical outcomes: association of the g.In1.1T>C RANTES variant with protection against infections. *Am J Hematol,* 84**,** 378-80.

DUNN, D. T., PODDAR, D., SERJEANT, B. E. & SERJEANT, G. R. 1989. Fetal haemoglobin and pregnancy in homozygous sickle cell disease. *Br J Haematol,* 72**,** 434-8.

EL-HAZMI, M. A. 1992. Clinical and haematological diversity of sickle cell disease in Saudi children. *J Trop Pediatr,* 38**,** 106-12.

GARDNER, K., DOUIRI, A., DRASAR, E., ALLMAN, M., MWIRIGI, A., AWOGBADE, M. & THEIN, S. L. 2016. Survival in adults with sickle cell disease in a high-income setting. *Blood,* 128**,** 1436-8.

GARDNER, K. & THEIN, S. L. 2015. Super-elevated LDH and thrombocytopenia are markers of a severe subtype of vaso-occlusive crisis in sickle cell disease. *Am J Hematol,* 90**,** E206-7.

GENOVESE, G., FRIEDMAN, D. J., ROSS, M. D., LECORDIER, L., UZUREAU, P., FREEDMAN, B. I., BOWDEN, D. W., LANGEFELD, C. D., OLEKSYK, T. K., USCINSKI KNOB, A. L., BERNHARDY, A. J., HICKS, P. J., NELSON, G. W., VANHOLLEBEKE, B., WINKLER, C. A., KOPP, J. B., PAYS, E. & POLLAK, M. R. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science,* 329**,** 841-5.

GENSER, B., COOPER, P. J., YAZDANBAKHSH, M., BARRETO, M. L. & RODRIGUES, L. C. 2007. A guide to modern statistical analysis of immunological data. *BMC Immunol,* 8**,** 27.

GORDEUK, V. R., CAMPBELL, A., RANA, S., NOURAIE, M., NIU, X., MINNITI, C. P., SABLE, C., DARBARI, D., DHAM, N., ONYEKWERE, O., AMMOSOVA, T., NEKHAI, S., KATO, G. J., GLADWIN, M. T. & CASTRO, O. L. 2009. Relationship of erythropoietin, fetal hemoglobin, and hydroxyurea treatment to tricuspid regurgitation velocity in children with sickle cell disease. *Blood,* 114**,** 4639-44.

HEBBEL, R. P. 2011. Reconstructing sickle cell disease: a data-based analysis of the "hyperhemolysis paradigm" for pulmonary hypertension from the perspective of evidence-based medicine. *Am J Hematol,* 86**,** 123-54.

KATO, G. J., GLADWIN, M. T. & STEINBERG, M. H. 2007. Deconstructing sickle cell disease: reappraisal of the role of hemolysis in the development of clinical subphenotypes. *Blood Rev,* 21**,** 37-47.

LETTRE, G., SANKARAN, V. G., BEZERRA, M. A., ARAUJO, A. S., UDA, M., SANNA, S., CAO, A., SCHLESSINGER, D., COSTA, F. F., HIRSCHHORN, J. N. & ORKIN, S. H. 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**,** 11869-74.

MAIER-REDELSPERGER, M., LEVY, P., LIONNET, F., STANKOVIC, K., HAYMANN, J. P., LEFEVRE, G., AVELLINO, V., PEROL, J. P., GIROT, R. & ELION, J. 2010. Strong association between a new marker of hemolysis and glomerulopathy in sickle cell anemia. *Blood Cells Mol Dis,* 45**,** 289-92.

MASON, K. P., GRANDISON, Y., HAYES, R. J., SERJEANT, B. E., SERJEANT, G. R., VAIDYA, S. & WOOD, W. G. 1982. Post-natal decline of fetal haemoglobin in homozygous sickle cell disease: relationship to parenteral Hb F levels. *Br J Haematol,* 52**,** 455-63.

MCPHERSON YEE, M., JABBAR, S. F., OSUNKWO, I., CLEMENT, L., LANE, P. A., ECKMAN, J. R. & GUASCH, A. 2011. Chronic kidney disease and albuminuria in children with sickle cell disease. *Clin J Am Soc Nephrol,* 6**,** 2628-33.

MENDONCA, T. F., OLIVEIRA, M. C., VASCONCELOS, L. R., PEREIRA, L. M., MOURA, P., BEZERRA, M. A., SANTOS, M. N., ARAUJO, A. S. & CAVALCANTI, M. S. 2010. Association of variant alleles of MBL2 gene with vasoocclusive crisis in children with sickle cell anemia. *Blood Cells Mol Dis,* 44**,** 224-8.

MILLER, S. T., SLEEPER, L. A., PEGELOW, C. H., ENOS, L. E., WANG, W. C., WEINER, S. J., WETHERS, D. L., SMITH, J. & KINNEY, T. R. 2000. Prediction of adverse outcomes in children with sickle cell disease. *N Engl J Med,* 342**,** 83-9.

MILLER, S. T., WRIGHT, E., ABBOUD, M., BERMAN, B., FILES, B., SCHER, C. D., STYLES, L. & ADAMS, R. J. 2001. Impact of chronic transfusion on incidence of pain and acute chest syndrome during the Stroke Prevention Trial (STOP) in sickle-cell anemia. *J Pediatr,* 139**,** 785-9.

MILTON, J. N., ROOKS, H., DRASAR, E., MCCABE, E. L., BALDWIN, C. T., MELISTA, E., GORDEUK, V. R., NOURAIE, M., KATO, G. R., MINNITI, C., TAYLOR, J., CAMPBELL, A., LUCHTMAN-JONES, L., RANA, S., CASTRO, O., ZHANG, Y., THEIN, S. L., SEBASTIANI, P., GLADWIN, M. T. & STEINBERG, M. H. 2013. Genetic determinants of haemolysis in sickle cell anaemia. *Br J Haematol,* 161**,** 270-8.

MINNITI, C. P., SABLE, C., CAMPBELL, A., RANA, S., ENSING, G., DHAM, N., ONYEKWERE, O., NOURAIE, M., KATO, G. J., GLADWIN, M. T., CASTRO, O. L. & GORDEUK, V. R. 2009. Elevated tricuspid regurgitant jet velocity in children and adolescents with sickle cell disease: association with hemolysis and hemoglobin oxygen desaturation. *Haematologica,* 94**,** 340-7.

MIYOSHI, K., KANETO, Y., KAWAI, H., OHCHI, H., NIKI, S., HASEGAWA, K., SHIRAKAMI, A. & YAMANO, T. 1988. X-linked dominant control of F-cells in normal adult life: characterization of the Swiss type as hereditary persistence of fetal hemoglobin regulated dominantly by gene(s) on X chromosome. *Blood,* 72**,** 1854-60.

MORRIS, J., DUNN, D., BECKFORD, M., GRANDISON, Y., MASON, K., HIGGS, D., DE CEULAER, K., SERJEANT, B. & SERJEANT, G. 1991. The haematology of homozygous sickle cell disease after the age of 40 years. *Br J Haematol,* 77**,** 382-5.

NAGEL, R. L. 1991. Severity, pathobiology, epistatic effects, and genetic markers in sickle cell anemia. *Semin Hematol,* 28**,** 180-201.

NATH, K. A. & HEBBEL, R. P. 2015. Sickle cell disease: renal manifestations and mechanisms. *Nat Rev Nephrol,* 11**,** 161-171.

NEBOR, D., BROQUERE, C., BRUDEY, K., MOUGENEL, D., TARER, V., CONNES, P., ELION, J. & ROMANA, M. 2010. Alpha-thalassemia is associated with a decreased occurrence and a delayed age-at-onset of albuminuria in sickle cell anemia patients. *Blood Cells Mol Dis,* 45**,** 154-8.

NOURAIE, M., LEE, J. S., ZHANG, Y., KANIAS, T., ZHAO, X., XIONG, Z., ORISS, T. B., ZENG, Q., KATO, G. J., GIBBS, J. S., HILDESHEIM, M. E., SACHDEV, V., BARST, R. J., MACHADO, R. F., HASSELL, K. L., LITTLE, J. A., SCHRAUFNAGEL, D. E., KRISHNAMURTI, L., NOVELLI, E., GIRGIS, R. E., MORRIS, C. R., ROSENZWEIG, E. B., BADESCH, D. B., LANZKRON, S., CASTRO, O. L., GOLDSMITH, J. C., GORDEUK, V. R. & GLADWIN, M. T. 2013. The

relationship between the severity of hemolysis, clinical manifestations and risk of death in 415 patients with sickle cell anemia in the US and Europe. *Haematologica,* 98**,** 464-72.

OLIVEIRA, M. C., MENDONCA, T. F., VASCONCELOS, L. R., MOURA, P., BEZERRA, M. A., SANTOS, M. N., ARAUJO, A. S. & CAVALCANTI, M. S. 2009. Association of the MBL2 gene EXON1 polymorphism and vasoocclusive crisis in patients with sickle cell anemia. *Acta Haematol,* 121**,** 212-5.

PEARSON, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6,* 2**,** 559-572.

PEMBREY, M. E., WEATHERALL, D. J. & CLEGG, J. B. 1973. Maternal synthesis of haemoglobin F in pregnancy. *Lancet,* 1**,** 1350-4.

PLATT, O. S., THORINGTON, B. D., BRAMBILLA, D. J., MILNER, P. F., ROSE, W. F., VICHINSKY, E. & KINNEY, T. R. 1991. Pain in sickle cell disease: Rates and risk factors. *New England Journal of Medicine,* 325**,** 11-6.

REES, D. C., WILLIAMS, T. N. & GLADWIN, M. T. 2010. Sickle-cell disease. *Lancet,* 376**,** 2018-31.

SCHEINMAN, J. I. 2009. Sickle cell disease and the kidney. *Nat Clin Pract Nephrol,* 5**,** 78-88.

SEBASTIANI, P., NOLAN, V. G., BALDWIN, C. T., ABAD-GRAU, M. M., WANG, L., ADEWOYE, A. H., MCMAHON, L. C., FARRER, L. A., TAYLOR, J. G. T., KATO, G. J., GLADWIN, M. T. & STEINBERG, M. H. 2007. A network model to predict the risk of death in sickle cell disease. *Blood,* 110**,** 2727-35.

SHARPE, C. C. & THEIN, S. L. 2011. Sickle cell nephropathy - a practical approach. *Br J Haematol,* 155**,** 287-97.

SHARPE, C. C. & THEIN, S. L. 2014. How I treat renal complications in sickle cell disease. *Blood,* 123**,** 3720-6.

SHIBA, H. F., EL-GHAMRAWY, M. K., SHAHEEN, I. A., ALI, R. A. & MOUSA, S. M. 2014. Glutathione S-transferase gene polymorphisms (GSTM1, GSTT1, and GSTP1) in Egyptian pediatric patients with sickle cell disease. *Pediatr Dev Pathol,* 17**,** 265-70.

SMITH-WHITLEY, K., PACE, B. & PACE, B. 2007. Sickle cell disease: a phenotypic patchwork. *Renaissance of sickle cell disease research in the Genome Era*.

STEINBERG, M. H., HSU, H., NAGEL, R. L., MILNER, P. F., ADAMS, J. G., BENJAMIN, L., FRYD, S., GILLETTE, P., GILMAN, J., JOSIFOVSKA, O. & ET AL. 1995. Gender and haplotype effects upon hematological manifestations of adult sickle cell anemia. *Am J Hematol,* 48**,** 175-81.

STEINBERG, M. H. & SEBASTIANI, P. 2012. Genetic modifiers of sickle cell disease. *Am J Hematol,* 87**,** 795-803.

SYDENSTRICKER, V. P., KEMP, J. A. & METTS, J. C. 1962. Prolonged Survival in Sickle Cell Disease. *American Practitioner and Digest of Treatment,* 13**,** 584-&.

VAN HAMEL PARSONS, V., GARDNER, K., PATEL, R. & THEIN, S. L. 2016. Venous thromboembolism in adults with sickle cell disease: experience of a single centre in the UK. *Ann Hematol,* 95**,** 227-32.

VIDLER, J. B., GARDNER, K., AMENYAH, K., MIJOVIC, A. & THEIN, S. L. 2015. Delayed haemolytic transfusion reaction in adults with sickle cell disease: a 5-year experience. *Br J Haematol,* 169**,** 746-53.

WONKAM, A., NGO BITOUNGUI, V. J., VORSTER, A. A., RAMESAR, R., COOPER, R. S., TAYO, B., LETTRE, G. & NGOGANG, J. 2014. Association of variants at BCL11A and HBS1L-MYB with hemoglobin F and hospitalization rates among sickle cell patients in Cameroon. *PLoS One,* 9**,** e92506.

Appendix 1

List of 2546 tests in raw data dump from the electronic patient record

% 3-0-m-Glucose excreted
% CD19+ CD20+ Lymphocytes
% CD19+ Lymphocytes
% CD20+ Lymphocytes
% D-Xylose excreted
% Fe Saturation
% Hypo
% Iron Saturation
% Lactulose excreted
% L-Rhamnose excreted
% Melibiose excreted (5h)
%HYPO
*CD3 CELLS
*CD3%
*CD4
*CD4%
*CD8 CELLS
*CD8%
+
++
+++
17-Hydroxyprogesterone
19D
1st Addendum Diagnosis
1st Addendum Microscopy
24 hr ECG Tape
24hr ALA excretion
24hr PBG excretion
24hr total porphyrin excretion
24Hr Urine 5HIAA
24Hr Urine 5HIAA.
2nd Addendum Diagnosis
2nd Addendum Microscopy
3 Phase Bone Imaging - NM
50/50 APTR Correction
50/50 INR Correction
5-HIAA (24 hour urine)
6PGA
6PGD assay.
9OH-Risperidone
a-amylase
Abdomen
Abdomen - CT
Abdomen - CT (Contrast)
Abdomen - CT (Enhanced)
Abdomen - Mobile
Abdomen - MRI
Abdomen - MRI (Contrast)
Abdomen - Pelvis - CT (Contrast)
Abdomen - X Ray
Abdomen - X Ray (Erect/Supine/Decubitus)
Abdomen - X Ray (Mobile)
Abdomen - X Ray (Plain film)
Abdomen & pelvis - CT
Abdomen Guided Aspiration - US
Abdomen(Thorax) - CT (Contrast)
Abdominal Xray (Mobile) - Neuro
ABE
ABPI & Doppler Waveform Analysis - Vasc
Abscess swab - M
Absolute Reticulocyte Count
Absolute Reticulocyte Count.
Achilles Tendon L. - US
Acinetobacter lwoffi
Acinetobacter Screen
Acromio-Clavicular L. - X Ray
ADAMTS13
Additional Information
Adenovirus
Adenovirus CFT result
Adenovirus DNA:
Adjusted Calcium
ADPCR
Adrenal - MRI
Adrenal - MRI (Contrast)
Adrenal Antibodies
Adrenal Both - CT
Adrenal Imaging - 123I - NM
Adrenaline (24 hour urine)
Adrenocorticotrophic Hormone
Adrnal Imaging - 123I - NM.
Aerobic bottle
Age Band (Yrs)
AI dec slope
AI max PG
AI max vel
AI P1/2t
Albumin
Albumin Excretion Rate
Aldosterone
Aldosterone/Renin Ratio
Alkaline Phosphatase
Allergen Grade
Almond
Alpha-1-antitrypsin phenotype
Alpha-1-antitrypsin phenotype
Alpha-Feto Protein
ALT
Aluminium (Serum)
Amikacin (Post-dose)
Amikacin level - post-dose
Amikacin level - random
Amino-Acid (urine)
Amino-Acids (plasma)
Aminolaevulinic Acid (ALA)
Ammonia (Plasma)
Amoxicillin.
Amphetamine Class
AMPO
AMPO - EXPIRED
Amylase
Amylase (random urine)
ANA Pattern
Anaerobic bottle
ANCA IIF Pattern
ANCP
Androstenedione
Angiography - CT
Angiography (head vessels) - CT
Angiography (neck vessels) - CT
Angioplasty
Angiotensin Converting Enzyme
Anion Gap
Ankle L.
Ankle L. - X Ray
Ankle R.
Ankle R. - X Ray
Ankle - USAnkle Bilateral - X Ray
Ankle L. - MRI
Ankle R. - CT
Ankle R. - MRI
Ankle R. Inversion - X Ray
Ankle Series
Ankles - MRI
Ankles + Feet - MRI
Anti 21-Hydroxylase Antibodies
Anti Acetylcholine Receptor Antibody
Anti Gastric Parietal Cell Antibody
Anti Glomerular Basement Membrane Ab
Anti Haemophilus Antibody
Anti HB core IgM
Anti Jo-1
Anti La
Anti MPO Antibodies
Anti MPO Antibodies (Sensitive)
Anti Neutrophil Cytoplasmic Antibody
Anti Nuclear Antibody
Anti PR3 Antibodies
Anti PR3 Antibodies (Sensitive)
Anti RN
Anti RNP
Anti Ro
Anti Scl-70
Anti Sm
Anti Smooth Muscle Antibody
Anti Thrombin III
Anti Thyroid Microsomes
Anti Tissue Transglutaminase
Antibody interpretation (1)
Antibody interpretation (2)
Antibody interpretation (3)
Antibody interpretation (4)
Antibody interpretation (5)
Antibody Report
Antibody screen.
Anti-Factor Xa
Anti-GBM Ab (FIDIS)
Anti-glutamic acid decarboxylase
Anti-IA-2
Anti-insulin antibody
Anti-Liver Cytosol-1
Antimicrobial substances
Anti-MPO Ab (FIDIS)
Anti-PR3 Ab (FIDIS)
Anti-soluble liver antigen
Antithrombin III
Anti-Xa
Ao max PG
Ao max PG (full)
Ao mean PG
Ao mean PG (full)
Ao root area
Ao root diam
Ao V2 max
Ao V2 mean
Ao V2 VTI
Aorta - US
Aortic Valve
APC Resistance-V
Appearance
Appearance (APPF)
Appearance (Sputum)
Appendix - US
APTR (Heparin)
APTT Control
APTT Ratio
APTT ratio (heparin)
APTT.
AREP2
Arm Bilateral - Vein Duplex - Vasc
Arm L. - Vein Duplex - Vasc
Arm R. - Vein Duplex - Vasc
ARR
Arthralgia serology
AS max vel
asc Aorta
Ascites - blood culture set
Ascites - M
Ascites - US
Ascites TB
Ascorbic Acid Screen (urine)
ASM
ASMT
Aspartate Transaminase
Aspergillus EIA.
Aspergillus fumigatus.
Aspergillus IgG (ImmunoCAP)
Aspirate (other) - M
Aspirate TB
Aspiration - US
Atria
Auramine AFB stain
Authorised by
AVA(I
AVA(V
Axilla L. - US
Axilla R. - US
Axilla swab - C&S
Azithromycin.
B Cells %
B Cells Absolute Counts
B12
B12 (Vitamin B12 assay)
B19R
B2GG
B2GM
Bacillus sp
Bacillus species
Bacteria
Barbiturates
Barium Enema (MS)
Barium Meal
Bartonella Serology
BASO
Basophils
Basophils (from film)
BBV incident (donor source)
BBV incident (recipient)
BCG
Benzodiazepine
Beta Haemolytic Streptococcus Group A
Beta Haemolytic Streptococcus Group B
Beta-2 Microglobulin
BHAM1
BHAT1
BHBC2
BHBS1
B-HCG
BHCG (molar pregnancy)
BHCV1
Bicarbonate
Bicarbonate.
Bile Acids
Bile Acids (Total)
Bilirubin (Conjugated)
Bilirubin (random urine)
Bilirubin (Total)
Biochemistry (Glucose)
Biopsy (other) - C&S
Biopsy (other) - C&S & AFB
Biopsy TB
Biopsy/Drain in Recovery - US
Bladder - US
Blast Count
Blasts (from film)
Blood Culture
Blood Culture for TB
Blood group (rapid).
Blood group.
Blood Lactate
Blood pH
Blood Products are available to collect
Blue mussel
BMI
BMWP
Body temp corrected pCO2
Body temp corrected pO2
Bone Biopsy
Bone Densitometry - Clin H+S - NM
Bone Imaging - 2 Phase - NM
Bone Imaging - 3 Phase - NM
Bone Imaging - NM
Bone Imaging - Whole Body - NM
Bone Marrow Aspirate
Bone Marrow TB
Bone sample - C&S & AFB
Bowel - US
Brain - MRI
Brain - MRI (Contrast)
Brain CT - Neuro
Brain Natriuretic Peptide
Brazil nut
Breast - bilateral - US
Breast (Bilateral) - US
Breast L. - US
Breast R. - US
Breast R. C/B - US
Bronchial washings - M
Bronchoscopy Test
Brucella ELISA IgG
Brucella ELISA IgM
Brucella serology
Brucellacapt (total IgG/IgM)
BSA
B-type Natriuretic Peptide
Burkholderia cepacia
Buttock - X Ray (Foreign Body)
-C
C1 Esterase Inhibitor
C3D
-C3d
C3D1
CA125
CA-125
CA199
CA-199
Caeruloplasmin (Serum)
Calcaneum L. - X Ray
Calcaneum R. - X Ray
Calcium
Calcium (random urine)
Calculated Creatinine Clearance
Campylobacter
Candida albicans
Candida species
Cannabis
Carbamazepine levels
Carcino-Embryonic Antigen
Cardiac - 12 Lead ECG
Cardiac - 24 hr ECG
Cardiac - 24hr BP Monitor
Cardiac - Cardiac MRI
Cardiac - Echo Reports
Cardiac - Exercise Tolerance Test Dr Led
Cardiac - MRI
Cardiac & Hepatic Iron Load - MRI
Cardiac CT
Cardiac - ExerciseToleranceTest Nurse Led
Cardiac_Echo_Reports
Cardiolipin Antibodies
Cardiolipin Antibodies IgG
Cardiolipin Antibodies IgM
Carotid - Angiogram (Bilateral)
Carotid - Angiogram (Left)
Carotid & Vertebral Artery Duplex - Vasc
Carotid Artery Both Doppler (US)
Carotid Artery Both Doppler (US) - Vasc
Carotids/Vertebral Angiography - MRI
Casts
Cat Dog-dander GuineaPig Epith Rat Mouse
Cat epithelium and dander
Catheter site swab - C&S
Catheter tip - C&S
Catheter tip (other) - C&S
Ca-Thrombin Control
Ca-Thrombin Time.
Caval Filter Insertion
CCA
CCP Antibodies
CD19 Absolute
CD19+CD20 Absolute
CD20 Absolute
CD3 CELLS
CD3%
CD4
CD4 CELLS
CD4%
CD8
CD8 CELLS
CD8%
CE MRA
Cefalexin.
Cefotaxime.
Cefoxitin.
Cefuroxime
Central line site swab - C&S
Central Venous Access Patency - Vasc
Centromere Abs (FIDIS)
Cerebral - Angioplasty
Cerebral Aneurysm - Pc coil Emb
Cerebral Angiography - 4 Vessel
Cerebral Venogram - CT
Cerebrospinal fluid - M
Cerebrospinal fluid - virology
Cerebrospinal fluid shunt - M
Cerebrospinal fluid virology
cH+
Chest
Chest - CT
Chest - CT (Contrast)
Chest - CT (Enhanced)
Chest - Mobile
Chest - PA + Lateral
Chest - US
Chest - X ray
Chest - Xray (Mobile)
Chest & abdomen - CT
Chest & abdomen & pelvis - CT
Chest (AP) - X Ray
Chest (PA + Lateral) - X Ray
Chest + Ribs - X Ray
Chest CT (Enhanced)
Chest High Resolution Supine - CT
Chest Unit - Lung Function Report
Chest Xray - Neuro
Chest Xray (Mobile) - Neuro
Chickenpox past exposure status
Chickenpox past exposure status.
Chickenpox status
CHK
Chlamydia - swabs
Chlamydia - urine
Chlamydia genus CFT result
Chlamydia pneumoniae
Chlamydia psittaci
Chlamydia trachomatis
Chlamydia trachomatis eye scrapings
Chlamydia trachomatis NAAT.
Chlamydia trachomatis urine
Chloramphenicol.
Chloride
Chloride concentration in blood
Cholesterol
CHR
Ciprofloxacin.
Circulating Immune Complexes
Citrobacter koseri
CKMB (cardiac markers)
Clauss Fibrinogen
Clavicle Bilateral - X Ray
Clavicle L. - X Ray
CLO Test
CLO Test1
Clostridium difficile report
Clostridium difficile toxin
Clue Cells
CMV antibody (IgG)
CMV antibody (IgG).
CMV DNA - log copies/ml
CMV DNA:
CMV IgG Avidity
CMV PCR
CMVL
CMVPCR
CMVV1
Coagulase Negative Staphylococcus
Co-amoxiclav.
Cocaine
Coccyx - X Ray
Coconut
Collagen/ADP
Collagen/EPI
Collection - US
Colon - MRI
Colonoscopy
Combined Chlamydia & Gonorrhea
Combined throat & nasal swab in VTM
Comment
Comment (APBS)
Comment (BCOM)
Comment (GYCO)
Comment (TBC1)
Comment (TBCO)
Comment (UCOM)
COMMENT CD20
Comment Electrophoresis
Comment Enzymes
Comment Film
Comment G6PD
Comment_
Comments (FCOM)
Comments (FLCO)
Comments (MCOM)
Comments (MYCO)
Comments (RCOM)
Comments (WCOM)
Common silver birch
COMPHENO
Complement C3
Complement C4
Concentration of Haemoglobin in blood
Conclusion
Contrast - US
Copper (Serum)
Coroner
Corrected Calcium
Cortisol
Cortisol (0 mins)
Cortisol (120 mins)
Cortisol (150 mins)
Cortisol (180 mins)
Cortisol (20 mins)
Cortisol (210 mins)
Cortisol (240 mins)
Cortisol (30 mins)
Cortisol (60 mins)
Cortisol (90 mins)
Cotrimoxazole.
Coxiella burnetii phase 1 IgA
Coxiella burnetii phase 1 IgG
Coxiella burnetii phase 2 IgA
Coxiella burnetii phase 2 IgG
Coxiella burnetii phase 2 IgM

Coxiella burnetii/Q-fever
C-peptide
C-reactive Protein
Creatine Kinase
Creatinine
Creatinine (enzymatic)
Creatinine (serum)
Creatinine (urine - enzymatic)
Creatinine (urine)
Creatinine Clearance.
Creatinine Kinase
Creatinine Urine
Cross match requested
Cryoglobulin Studies
Cryptococcal Antigen Test
Cryptosporidia ZN Stain
CSF Glucose
CSF Oligoclonal bands
CSF Protein
CSF TB
CSF Xanthochromia Screen
CT Chest
CT Chest Enhanced
CT Chest Hr Sup
CT Foreign Film
CT Head
CT Head Enhanced
CT Neck Enhanced
Ctl
CTR
Culture (UCL)
Culture Status
Cyclosporin Assay - Liver Unit
Cyclosporin levels
Cystatin C
Cytogenetics
Cytomegalovirus - blood
Cytomegalovirus - urine
Cytomegalovirus IgM
Cytomegalovirus IgM.
DAGT
DAT1
Date result received (TBRE)
Date sent to MRU
DCTR
DCTR1
D-Dimer
D-Dimer (Auto)
D-Dimer (Use for DIC Only)
Dehydroepiandrosterone SO4
Delta Ab (IgM)
Delta Ab (Total)
Delta RNA Quantitative
DFTM
DHS - Theatres
Dialysis Fistula Duplex - Vasc
Diff. Wtd. MRI - Neuro
Diffusion - MRI
Diffusion MRI - Neuro
Dilute RVV
Dilute RVV Confirm
Dilute RVV Test
Direct Antiglobulin Test
Dog dander
Doppler - US
Doppler Measurements & Calculations
Double-stranded DNA Antibodies
Doxycycline.
DPDC
DPDR
Drain exit site swab - C&S
Drain fluid - C&S
Drainage - US
Drug Screen (Serum)
Drug Screen (Urine)
DRVV % Correction
DVV Confirm
E. coli 0157
EBM Donor - virology
EBV DNA - Log copies/ml
EBV DNA Post-Transplant: BMT/Liver
EBV DNA:
EBV past exposure status (EBV VCA IgG)
EBV past exposure status (EBV VCA IgG).
EBV VCA IgM.
EBVL1
EDV(MOD-sp2)
EDV(MOD-sp4)
EDV(sp4-el)
EF(MOD-bp)
EF(MOD-sp2)
EF(MOD-sp4)
EF(sp2-el)
EF(sp4-el)
Egg white Milk Fish Wheat Peanut Soybean
EIA Index
Elbow - MRI
Elbow - US
Elbow Both - US
Elbow L.
Elbow L. - US
Elbow L. - X Ray
Elbow R.
Elbow R. - US
Elbow R. - X Ray
Elbow Rt - MRI
Electroencephalogram
ELU1
Elution of bound antibody from Red Cells
Embolisation
EMG Report
ENA JO-1
ENA Scl-70
ENA Sm
ENA Sm/RNP
ENA SSA (Ro52)
ENA SSA (Ro60)
ENA SSB (La)
End Point
Endocervical swab - M
Endocrinology Urine Steroid Profile
Endoscopy Report PDF
Endoscopy Result PDF
Endotracheal secretions - M
Enrichment culture
Enrichment culture (RCMF)Enrichment culture (RCMS)
Enterobacter cloacae
Enteroclysis - CT
Enterococcus faecalis
Enterovirus RNA:
EOS
Eosinophils
Eosinophils (EOSI)
Eosinophils (from film)
Eosinophils.
Epithelial Cells
Epstein Barr virus VCA IgM
ERCP
ERCP - Endoscopy
ERCP & Sphincter
Erythrocyte Sedimentation Rate
Erythromycin.
Erythropoietin
Erythropoietin Level
Escherichia coli
ESR
ESR.
ESRE
Estimated GFR
ESV(MOD-sp2)
ESV(MOD-sp4)
ESV(sp4-el)
Ethambutol.
Excercise Echo
Exercise Echo
External Organisation Result
Extractable Nuclear Antigens
Eye swab - M
F425
F428
F8
FABDOM
Facet Injection - X Ray
Facet Joint Injection
Facial - CT
Facial - CT (Contrast)
Facial Bones
Facial Bones - X Ray
Factor V Leiden
FAEC
Faecal Calpotectin
Faecal Calprotectin
Faecal Elastase
Faecal Occult Blood
Faeces - Culture
Faeces virus detection
FAI
FBR
FDP D-Dimer
FDP D-Dimer (Use for ?PE DVT Only)
Feet - Standing - X-Ray
Femora - MRI
Femoral Line - X Ray
Femur Bilateral - X Ray
Femur L. - X Ray
Femur R.
Femur R. - X Ray
Ferriscan - MRI
Ferritin
Ferritin (Serum Ferritin assay)
FEV 1
FEV 1 % Predicted
FEV 1 % VC MAX
FEV 1 % VC MAX Post
FEV 1 Post
FEV 1 Post % Predicted
FEV1
FEV1 % Predicted
FEV1 / VC Ratio
FEV1 Post Bronchodilator
FEV1 Post Bronchodilator % Change
FFCR
FFCT
FFMR
FFNM
FFUS
FHIPLM
Fibrinogen
Film Comment.
Finger L. Index - X Ray
Finger L. Little - X Ray
Finger L. Middle - X Ray
Finger L. Ring - X Ray
Finger R. Index - X Ray
FIO21
FISH
Fish Shrimp Blue Mussel Tuna Salmon
Fish cod
Fistulogram (MS)
Fit to Fly
Flexible Sigmoidoscopy
FLOLRM
Flucloxacillin.
FLUI
Fluid (other) - M
Fluid Albumin.
Fluid Amylase
Fluid Amylase.
Fluid Bilirubin
Fluid Bilirubin.
Fluid Crystal Examination
Fluid Glucose
Fluid Glucose.
Fluid LDH
Fluid LDH.
Fluid Total Protein
Fluid Total Protein.
Fluid Type
Fluoro.Only Image Intensifier (Theatre)
Fluroscopic Embolisation
FLUS
FMMX
Folate
Folate (Serum Folate assay)
Follicle Stimulating Hormone
Foot - US
Foot Both - US
Foot L.
Foot L. - US
Foot L. - X Ray
Foot L. (Lateral) - X Ray
Foot R.
Foot R. - X Ray
Foot R. (Lateral) - X Ray
Foot Series
FOPI
Forearm - MRI (Contrast)
Forearm L.
Forearm L. - X Ray
Forearm R. - X Ray
Foreign Body - X Ray (Demo)
Foreign Body - X Ray (Localisation)
Foreign Films
FRE1
Free Androgen Index
Free Protein S
Free Thyroxine
Free Tri-Iodothyronine
FREP
Frozen Section Diagnosis
FS
FSH (0 mins)
FSH (20 mins)
FSH (60 mins)
FT4 (0 mins)
Full culture (inc ANO2)
Full culture (inc ANO2) (SCU)
Full culture (Sterile Site) (FCUS)
Full culture (Sterile Site) (SCUS)
Fungal Culture
Fungal Precipitins serology
FUPLRM
Fusidic acid.
FV Loops
FXC - Vibrio Culture
G6PD assay (WHO).
G6PD assay two-stage.
G6PD..
Gallbladder - US
Gallbladder + Liver - US
Gallium Imaging - Infection - NM
Gallium Imaging - Lymph - NM
Gamma-glutamyl Transferase
Gastric Parietal Cell Abs paediatric
Gastrograffin Enema
Gastrograffin Meal
GCR
GCYTO
GDHX
Genital culture
Gentamicin (Post-dose)
Gentamicin (Pre-dose)
Gentamicin level
Gentamicin level - post-dose
Gentamicin level - pre-dose
Gentamicin level - random
Gentamicin.
GHPV
Gilbert's Genetics
Glandular Fever Test
Globulin
Glomerular Filtration Rate - NM
Glucose
Glucose (120 mins)
Glucose (150 mins)
Glucose (180 mins)
Glucose (20 mins)
Glucose (30 mins)
Glucose (60 mins)
Glucose (90 mins)
Glucose (Fasting)
Glucose (Plasma)
Glucose (serum)
Glucose (whole blood)
Glucose.
Glutamic Acid Decarox Abs
Glycated Hb
GNOS
GPC
GPDA
Granulocyte PNH clone
Grass Pollen Mix (standard)
GX1
Great Vessels
Groin swab - C&S
Gross Description
Group + Save Rejected
Group and Screen
Growth Hormone
Growth Hormone (0 mins)
Growth Hormone (120 mins)
Growth Hormone (150 mins)
Growth Hormone (180 mins)
Growth Hormone (20 mins)
Growth Hormone (210 mins)
Growth Hormone (240 mins)
Growth Hormone (60 mins)
Growth Hormone (90 mins)
GRP2
Guided Drainage - CT
Guided Injection - US
Gynae Clinical History
Gynae Comment
Gynae Cytopathology Report
Gynae FHSA
Gynae Infection
Gynae Recall Status
Gynae Report part 1
Gynae Report part 2
Gynae Report part 3
Gynae Specimen Description
Gynae Suggested Management
H+
H+ (venous)
H+.
H2CO3 mmol/l
H2CO3 on 1l O2
H2CO3 on 2l O2
Haem Other External Results
Haem Rayne External Results
Haematocrit (packed cell volume %)
Haematology Disorders - Virology
Haematology disorders virology
Haemoglobin A%
Haemoglobin A2 % (antenatal)
Haemoglobin A2 % (HbA2)
Haemoglobin A2 % (HbS)
Haemoglobin A2 % (pre-op)
Haemoglobin A2 % (screen)
Haemoglobin C % (HbC)
Haemoglobin C % (HbS)
Haemoglobin D %
Haemoglobin DNA Investigation
Haemoglobin F % (antenatal)
Haemoglobin F % (HbF)
Haemoglobin F % (pre-op)
Haemoglobin F % (screen)
Haemoglobin F% (screen)
Haemoglobin S % (antenatal)
Haemoglobin S % (HbS)
Haemoglobin S % (pre-op)
Haemoglobin S % (screen)
Haemophilus influenzae
Haemophilus parainfluenzae
Hand Bilateral - X Ray
Hand Both - US
Hand L.
Hand L. - US
Hand L. - X Ray
Hand R.
Hand R. - X Ray
Haptoglobin
HAV Ab (IgM)
HAV Ab (Total)
Hazel nut
HAZY
Hb
HB Elect.pH 6.0
Hb electrophoresis pH 6.0 (antenatal)
Hb electrophoresis pH 6.0 (pre-op)
Hb electrophoresis pH 6.0 (screen)
Hb electrophoresis pH 8.6 (antenatal)
Hb electrophoresis pH 8.6 (pre-op)
Hb electrophoresis pH 8.6 (screen)
Hb Electrophoresis pH8.6
Hb.
HB1
HBA0%
HBA1c (DCCT)
HBA1c (IFCC)
HBcAb
HBcAb (IgM)
HBCB1
HBeAb
HBeAg
HBEC
HBGN
HBGX
HbO2 saturation (oximeter)
HbO2% saturation
HBsAb
HBsAg
HBsAg Confirmatory
HBV DNA
HBV DR Mutation Screen
HBV Genotyping
HCAX
HCO3-
HCV Ab
HCV Genotype
HCV RNA Qualitative
HCV RNA Quantitative
HDL-Cholesterol
HDV RNA Quantitative
Head - Angiography - MRI
Head - CT
Head - CT (Enhanced)
Head - MRI
Head + Angiography MRI - Neuro
Head MRI - Neuro
HeartRate
Helicobacter 13c-H Breath Test
Helicobacter serology
Helicobacter Stool Antigen
Helicobacter Stool Antigen.
Hep_
Hep_HBsAgQuant
Hep_HCVAg
Hep_IL-28B Genotype
Hepatitis (acute) with raised LFTs
Hepatitis A exposure status
Hepatitis A status
Hepatitis A total antibody.
Hepatitis A virus IgM
Hepatitis A virus IgM.
Hepatitis A virus total Ab
Hepatitis B - known carrier - markers
Hepatitis B core total Ab
Hepatitis B core total Ab.
Hepatitis B e antibody
Hepatitis B e antigen.
Hepatitis B immune status
Hepatitis B known carrier markers
Hepatitis B past exposure status
Hepatitis B post immunisation status
Hepatitis B status
Hepatitis B surface Ag Confirmation
Hepatitis B surface antibody
Hepatitis B Surface Antigen (HBGN)
Hepatitis B surface antigen screen
Hepatitis B surface antigen status
Hepatitis B surface antigen status.
Hepatitis B surface antigen..
Hepatitis B/C infection status
Hepatitis B/C status
Hepatitis C RNA
Hepatitis C virus antibody
Hepatitis C virus antibody..
Hepatitis C virus IgG.
Hepatitis C virus RNA
Hepatitis C virus RNA load
Hepatitis E Send Away
Hepatitis Markers Comment
Hepatitis with raised LFT's - virology
Hepato-Biliary Imaging - NM
HEV Ab (IgG)
HEV Ab (IgM)
HEV IgM
HEV total antibody
Hexokinase assay.
HFE Liver Unit
HFR
HHV-6 DNA.
Hickman Line
High Flourescence Ratio
High Fluorescence Ratio
High Sensitivity CRP
Hip - Bilateral - MRI
Hip - Bilateral - US
Hip - Bilateral - X Ray
Hip (Adult) - US
Hip Both - MRI
Hip Both - X Ray
Hip Frog Leg - X Ray
Hip L (AP & Lateral) - X Ray
Hip L.
Hip L. - MRI
Hip L. - X Ray
Hip L. - Xray
Hip R (AP & Lateral) - X Ray
Hip R.
Hip R. - MRI
Hip R. - US
Hip R. - X Ray
Hip R. - Xray
IOP Research - Neuro
PCV
Pus - M
Serum Creatinine

Hips - Both
Histology
Histology Block Description
Histology Clinical History
Histology Diagnosis
Histology Macroscopy
Histology Microscopy
Histology Report
Histology Specimen
Histone
Histopathology Result Document
Histopathology Results Document
HIV 1&2 antibody/HIV-1 p24 antigen
HIV 1&2 antibody/HIV-1 p24 antigen.
HIV Antibody
HIV type 1 RNA viral load quantification
HIV-1 Genotypic Resistance Test
HIV-1 proviral DNA on children <18mths
HIV-2 Genotypic Resistance Test
HLA B51 Genotype
HLA B57 Genotype1
HMPR
Homocysteine (Plasma)
House dust mite
HPLC
HPLC (antenatal)
HPLC (pre-op)
HPLC (screen)
HRNA1
HSV 1 AND 2 DNA
HSV IgG
HT Ratio
HTLV 1/2
HTLV type 1 and 2 antibody
HTLV type 1&2 antibody.
Human metapneumovirus RNA
Humerus L. - X Ray
Humerus R. - X Ray
HVC
Hydatid Disease serology
HYPO
I ABOAUTO
I KLE
IA2 Antibodies
IAMs - MRI
IGA
IgE concentration
IGF BP-1
IGF BP-3
IGF-BP1
IGF-BP3
IGG
IgG Pneumococcal Antibody
IgG subclass 1
IgG subclass 2
IgG subclass 3
IgG subclass 4
IGG.
IGG1
IgG2 Pneumococcal Antibody
IGLA
IGLG
IGLM
IGM
II:C
IIGE1
IL-28B Genotype (rs 12979860)
Immunoglobulin A
Immunoglobulin A.
Immunoglobulin D
Immunoglobulin E
Immunoglobulin G
Immunoglobulin G.
Immunoglobulin M
Immunoglobulin M.
Immunology & Allergy External Results
Immunophenotyping
Immunophenotyping lab. comments
Immunophenotyping: Comment
Immunophenotyping: Conclusion
Immunophenotyping: Sample Type
Immunotyping Comments
Immunotyping Concentration
INC1
Incorrect HospNo on Group and Save Samp
Influenza A virus CFT result
Influenza A virus RNA
Influenza B virus CFT result
Influenza B virus RNA
INR
INR (warfarin)
INR 50/50
Insulin
Insulin Antibodies
Insulin-Levels
Insulin-like growth factor 1
Interim HSV DNA result
Interim influenza B virus RNA
Interp HBcore Total
Interpretation Summary
Interpreting Physician
Intrinsic Factor Antibody
IOG Integrated Report
IRF
Iron (Serum)
Iron (Urine)
Iron Urine (24)
Isoelectric Focusing
Isoelectric Focusing (pre-op)
Isoelectric Focusing (screen)
Isoniazid.
IVP
IVSd
IVSs
IX:C
Joint - US
Joint Fluid Crystals.
Jugular L. Int. - Venogram
K.U.B.
Kappa:Lambda Ratio
KCO
KCOc.
KCOc. % Predicted
Ketones (random urine)
Kidney (Transplant) - US
Kidney (transplant) US
Kidney Bilateral - US
Kidney Imaging (DMSA) - NM
Kidneys + Bladder - US
Klebsiella pneumoniae
Klebsiella species
Kleihauer % Fetal Cells
Kleihauer result
Kleihauer Test
Knee - MRI
Knee - Patella L - X Ray
Knee - Patella R. - X Ray
Knee - US
Knee - Xray (Weight Bearing)
Knee Bilateral - X Ray
Knee Both - MRI
Knee Both - US
Knee L.
Knee L. - MRI
Knee L. - US
Knee L. - X Ray
Knee L. - X Ray (Weight Bearing)
Knee R.
Knee R. - CT
Knee R. - MRI
Knee R. - US
Knee R. - X Ray
Knee R. - X Ray (Weight Bearing)
KUB
L. Spine
LA dimension
LABCODE
Lactate (CSF)
Lactate (plasma)
Lactate concentration in blood
Lactate Dehydrogenase
Lactulose/Rhamnose ratio
Lamotrgine.
Lamotrigine levels
Latex
LC1 Antibodies blot
LDL-Cholesterol
Left Ventricle
Leg - Lower R. - MRI
Leg (Lower) Both - CT
Leg Bilateral - DVT Duplex - Vasc
Leg Bilateral -Venous Incompetence- Vasc
Leg L. - DVT Duplex - Vasc
Leg L. - Venogram
Leg L. - Venous Incompetence - Vasc
Leg R. - DVT Duplex - Vasc
Leg R. - Venous Incompetence - Vasc
Legionella
Levetiracetam
LFR
LH (0 mins)
LH (20 mins)
LH (60 mins)
Limb - CT
Linear EUS - Endoscopy
Linezolid.
Linogram
Lipase
LIT
Lithium (Serum)
Little Finger R.
Liver - Biphasic - CT
Liver - CT (Contrast)
Liver - CT (Portogram)
Liver - MRI
Liver - MRI (Contrast)
Liver - US
Liver - US (Paediatric)
Liver Biopsy - US
Liver Biopsy - US (low risk/daycase)
Liver Kidney Microsomal Abs paediatric
Liver Kidney Microsomal Antibody
LKM
LKM Antibodies blot
Long Leg Standing View - X Ray
Low Flourescence Ratio
Low Fluorescence Ratio
Lower leg Both - MRI
Lower limb Both - Pinning
LSUM
Lumbar Puncture Screening - Neuro
Lumbo-Sacral Spine
Lung Function Comments
Lung Perfusion Only - NM
Lung V/Q Imaging (VQ DTPA) - NM
Lung V/Q Imaging (VQ Kr) - NM
Lung V/Q Scan
Lung Ventilation Only - NM
Lung Ventilation/Perfusion (VQ) - NM
Lupus Anticoagulant
Luteinising Hormone
LV max PG
LV mean PG
LV V1 max
LV V1 mean
LV V1 VTI
LVAd ap4
LVAs ap4
LVIDd
LVIDs
LVLd ap4
LVLs ap4
LVOT area
LVOT diam
LVPWd
LVPWs
LYM
Lyme Disease serology
Lymphadenopathy/Glandular Fever screen
Lymphocyte Count.
Lymphocytes
Lymphocytes (from film)
Lymphocytes (LYMP)
M2 Antibodies Blot
Macroprolactin
Macular rash serology
Magnesium
Malaria Parasitaemia
Malaria serology
Malarial Parasites
Malarial Parasites Percentage
Malarial Parasites Species
Mammogram - Bilateral
Mammogram - FV R.
Mammogram L. - X Ray
Mammogram R. - X Ray
Mammography R.
Mandible - CT
Mandible - X Ray
MANKRC
Mantoux Range
Mantoux Result
Manual Retics
Mass - US
Mastoids
MCH
MCH1
MCHC
MCHC.
MCHC1
MCV
MCV1
Measles IgG (Immunity only)
Medium Flouresence Ratio
Medium Fluoresence Ratio
Melibiose/Rhamnose ratio
Meropenem Res GNR screen.
Meropenem.
Metamyelocyte Count
Metamyelocytes (from film)
Methadone Metab
Met-haemoglobin
Methylmalonic acid
Metronidazole.
MFR
Microscopy comment
Midline Insertion
Misc. Fluid Type
Misc. Fluid Type I
Miscellaneous Biochemistry (2)
Miscellaneous Comments
Miscellaneous Immunology Test
Miscellaneous sample - M
Miscellaneous tissue sample - C&S
MITO
Mitochondrial antibodies paediatric
Mitochondrial Antibody
Mitral Valve
Mixed anaerobes
Mixed Coagulase Negative Staphylococci
MKNERC
MMode 2D Measurements & Calculations
MON
Monocyte PNH clone
Monocytes
Monocytes (from film)
Mono-nuclear cells
Morganella morganii
Mouth swab - C&S
MPV
MPV1
MRCP - MRI
MRI
MRI .....
MRI Brain
MRI Foreign Film
MRI Head - Angiography
MRI Hips
MRI Knee
MRI Lumbar Spine
MRI Shoulder
MRI Whole Spine
MRSA Admission Screen
MRSA Culture
MRSA PCR Result
MRSA Pre-admission Screen
MRSA Screen
MRSA to VITEK
MRSA..
MRU TB Culture
MTHF Reductase
MTHIRC
MUGA Imaging (Cardiac - Rest) - NM
Mumps virus IgG
Mupirocin 5
Mupirocin.
MURCHL
Muscle - US
Muscle Specific Kinase Antibodies
MV A point
MV dec slope
MV dec time
MV E point
MV E/A
MV max PG
MV mean PG
MV P1/2t
MV P1/2t max vel
MV V2 max
MV V2 mean
MV V2 VTI
MVA(P1/2t)
MVA(traced)
MYC1
MYC2
Mycobacterium abscessus
Mycobacterium fortuitum
Mycobacterium tuberculosis complex
Mycology culture
Mycology KOH microscopy
Mycology Results
Mycophenolate Assay - Liver Unit
Mycoplasma pneumoniae CFT
Myelocyte Count
Myelocytes (from film)
Myoglobin (cardiac markers)
NAG
Nail clippings - fungal culture
Nasopharyngeal aspirate - virology
Nasopharyngeal aspirate virology
NCAC
Neck - CT (Contrast)
Neck - CT (Enhanced)
Neck - MRI
Neck - MRI (Contrast)
Neck - Thorax - CT (Contrast)
Neck - US
Neck & chest - CT
Neck & chest & abdomen & pelvis - CT
Neck (Lateral Soft Tissue) - X Ray
Neck-Thorax-Abdomen-Pelvis-CT (Contrast)
Neisseria Gonorrhoea
Neisseria Gonorrhoea - Culture
Neisseria gonorrhoeae NAAT.
Neomycin.
NER
Nerve Root Injection (Cervical) -Fluoro
Neuro Pathology Result
Neurophysiology - EEG Result
Neutrophils
Neutrophils (from film)
Nitrofurantoin.
NK %
NK CELLS
NM BD CLIN H&S&L
NM Bone Imaging - 2 Phase
NM Lung (V/Q) Imaging - VQ Kr
NM Parathy MIBI Imaging
No Name on Group and Save Sample
No of samples received
No req form received for GS only sample
No sample received for GS only request
Nocturnal Oximetry
Non Gynae Report
Non-gynae Clinical History
Non-gynae Cytopathology Report
Non-gynae Specimen
Norovirus RNA.
Nose swab - C&S
NPTHY
NRBC
NRBC (from film)
NRH
NSR
NTX (C)
NTX (R)
Nuclear Antibodies
Nuclear Antibodies Paediatric
Nuclear antibodies titre
Nucleated RBC
NULL
Nuts-Peanut Hazel Brazil Almond Coconut
OBSERVATION
Occupational Health - needle injury
Oestradiol
OGD - Diagnostic
OGD (Gastroscopy)
On T4 replacement therapy?
Oncology External Results
OP Chole (Theatre) - X Ray (Mobile)
Opiate
Orbits - CT
Orbits - CT (Contrast)
Orbits - MRI (Contrast)
Orbits - X Ray
Orbits CT - Neuro
ORDFILM
Organic Acids (random urine)
Organisms (FGR1)
Organisms (GRO2)
Organisms (GROR)
Organisms (SGO1)
Organisms (SGO2)
Orthopantomogram - X Ray
Osmolality (random urine)
Osmolality Plasma
Other Foreign Film
OUTC
Ova cysts and parasites
Ovarian Antibodies
Oxacillin.
PA acc slope
PA acc time
PA max PG
PA mean PG
PA pr(Accel)
PA V2 max
PA V2 mean
PA V2 VTI
PaCO2 (kPA)
PaCO2 on 1l O2
PaCO2 on 2l O2
PACSCD
PACSIL
Paed Haem External Test Results
Paed Resp Sleep Study Results
Paediatric bottle
Pancreas - MRI
Pancreas - US
Pandemic influenza A virus H1N1 PCR
PaO2 (kPa)
PaO2 on 1l O2
PaO2 on 2l O2
Paracetamol Levels
Parainfluenza virus type 1 RNA
Parainfluenza virus type 2 RNA
Parainfluenza virus type 3 RNA
Parathyroid - US
Parathyroid Hormone
Parathyroid Imaging (MIBI) - NM
Parotid - US
Parovirus B19 IgM
Parvovirus B19 DNA
Parvovirus B19 IgG
Parvovirus B19 IgG:
Parvovirus B19 IgM.
Parvovirus B19 IgM:
Parvovirus B19 past exposure status
Parvovirus Status
PASSB1
Patella R. - X Ray
PatientHeight
PatientWeight
PCNA
PCO2
PCO21
PCV1
PD Fluid 1 Creatinine
PD Fluid 1 Glucose
PD Fluid 1 Urea
PD Fluid 2 Creatinine
PD Fluid 2 Glucose
PD Fluid 2 Urea
PD Fluid 3 Creatinine
PD Fluid 3 Glucose
PD Fluid 3 Urea
PD Fluid Creatinine
PD Fluid o/night creatinine
PD Fluid o'night glucose
PD Fluid o'night urea
PD Fluid Protein
PD Fluid Urea
PDW
Peanut
PEF
PEF % Predicted
PEF Post
PEF Post % Predicted
PEF Post Bronchodilator
PEF Post Bronchodilator % Change
Pelvis - CT
Pelvis - CT (Contrast)
Pelvis - CT (Enhanced)
Pelvis - MRI
Pelvis - MRI (Contrast)
Pelvis - US
Pelvis (AP) - X Ray
Pelvis (TV) - US
Pelvis AP
Penicillin.
Penile swab - M
Penis - US
Performed date

Perfusion - CT
Pericardium/Pleural
Peritoneal dialysis fluid - M
Peritoneal fluid - M
Peritoneal swab - C&S
PET CT FDG - NM
PF1R
PF2R
PF3R
pH
pH on 1l
pH on 2l
pH(T)
pH1
Phenytoin levels
Phosphate
Phosphate (random urine)
PI end-d vel
PICC Exchange/Re-Wiring
PICC Line Exchange
PICC Line Insertion
Piperacillin.
Piperacillin-tazo
Pituitary - MRI
Pituitary - MRI (Contrast)
Pituitary Fossa - CT (Contrast)
Placental tissue - M
Plain Film Foreign Film
Plasma Adrenaline
Plasma Noradrenaline
Platelet Morphology
Pleural fluid - M
Pleural Fluid TB
PLT
PM Clinical Information
PM Patient Details
PM-SCL
Pneumococcal serology
Pneumonia (atypical) - virology
Pneumonia (atypical) serology
PNH
Pnuem Cap Polysac IgG Ab Total
PO2
pO21
Polymorphs
Porphobilinogen (PBG)
Porphyrins 24hr Urine
Portacath
Post
Post Liver Biopsy - US
Post Nasal Space - X Ray
Potassium
Potassium concentration in blood
PR31
Pre CEA/TCD Window Assessment - Vasc
Pre Fistula Mapping Bilateral - Vasc
Pre Fistula Mapping L. - Vasc
Pre Fistula Mapping R. - Vasc
Pre Renal Transplant Assessment - Vasc
Pregnancy booking - virology
Pregnancy booking virology
Pregnancy Glucose Test
Pregnancy Test
Procollagen III Np
Progesterone
Prolactin
Prolactin (0 mins)
Prolactin (20 mins)
Prolactin (60 mins)
Prolactin (dil x5)
Prolonged Anaerobic Culture (FPAC)
Promyelocyte Count
Promyelocytes (from film)
Prostate - MRI
Prostate Specific Antigen
Protein (urine)
Protein C Activity
Protein Electrophoresis
Protein/Creatinine Ratio
Proteus species
Prothrombin
Prothrombin Genotype
Prothrombin Time.
Pseudomonas aeruginosa
Pseudomonas stutzeri
PT Control
Pulmonary - Angiogram - CT
Pulmonary - DSA
Pulmonary Angiogram CT
Pulmonic Valve
Pus cells (FGRP)
Pus cells (GRPU)
Pus cells (SGRP)
Pus Cells (UMWB)
Pus cells (Wet prep)
Pus swab - M
Pus TB
Pyrazinamide.
Pyruvate
QPCR
Quantitation / Titre
Radial EUS - Endoscopy
Radial Head R. - X Ray
Random ALA excretion
Random PBG excretion
Random total porphyrin excret.
Random Urine Bilirubin
Random Urine Blood
Random Urine Creatinine.
Random Urine Glucose
Random urine HMMA
Random urine HVA
Random Urine Ketones
Random Urine Leucocytes
Random Urine Nitrite
Random Urine pH
Random Urine Protein
Random Urine Protein.
RAST
RBC
RBC count (CSF)
RBC Count (CSF).
RCDN
RDW
RDW1
Reconstruction Limb - CT
Rectal swab - C&S
Red Blood Cells
Red Cell Antigen Typing
Red Cell Folate
Red Cell Morphology
Red cell PNH clone
Red Cells Elution Studies
Ref Lab Referral
Reference Lab (TO)
Reference Laboratory Comment (MRSC)
Referred to (MRE2)
REJ
REJC
Renal - MRI
Renal Artery/Vein Duplex (Native) - Vasc
Renal Artery/Vein Duplex (Tx) - Vasc
Renal dialysis virology- new patient
Renal Imaging (MAG3 + Frusemide) - NM
Renal Pathology Specimen
Renals - DSA
Renin
Renography Imaging (DTPA) - NM
REPT1
Research - US - VASC
Respiratory PCR Flag.
Respiratory Viral PCR
Result Comment
Reticulocyte HB Content
Reticulocyte Percent
RETP
Rh Phenotype
Rheumatoid Factor
RHIN
Rhinovirus RNA
RHNR
Ribavirin Assay - Liver Unit
Ribosomal Antibodies (FIDIS)
Rifampicin.
Right Ventricle
Ring Finger L.
RINV
Risperidone.
RNP
RORG
Routine culture (no ANO2)
Routine culture (no ANO2) (SCS)
RPR
RRBC
RSAR1
RSBR
RSV CFT result
RSV subgroup A RNA
RSV subgroup B RNA
Rubella status
Rubella virus IgG
Rubella virus IgG.
Rubella virus IgM
Rubella virus IgM.
RUGV1
RV
RV % Predicted
RV max PG
RV mean PG
RV Predicted
RV V1 max
RV V1 mean
RV V1 VTI
RVDd
Sacro-Iliacs - X Ray
Sacrum - X Ray
Salicylate
Salmon
Salmonella
SAM
Sample Quality
Sample sent to
SaO2 on 1l O2
SaO2 on 2l O2
Saturated oxygen % in blood
Save Sample
Scanned Flexible Cystoscopy
Scanned Flow Rate and Residual
Scanty
Scaphoid - Bone Imaging - NM
Scaphoid L. - X Ray
Scaphoid R. - X Ray
Schilling Test Part 1
Screening (H10)
Scrotum - US
Selenium (Serum)
Serratia marscesens
Serum Electrophoresis
Serum Iron
Serum Kappa Light Chains
Serum lambda Light Chains
Serum Saved temporarily
Sex Hormone Binding Globulin
Shigella
Shoulder - MRI
Shoulder - US
Shoulder Bilateral - X Ray
Shoulder Both - MRI
Shoulder Both - US
Shoulder L.
Shoulder L. - MRI
Shoulder L. - US
Shoulder L. - X Ray
Shoulder R.
Shoulder R. - MRI
Shoulder R. - US
Shoulder R. - X Ray
Shoulders - CT
Shrimp
Sickle Cell Crisis - virology
Sickle cell crisis serology
Sickle Solubility Test
Sickle Solubility Test (antenatal)
Sickle Solubility Test (pre-op)
Sickle Solubility Test (screen)
Sinus - CT
Sinuses
Sinuses - X Ray
Skeletal Survey (Metabolic) - X Ray
Skin Biopsy - Direct Immunofluorescence
Skin Prick Test Results
Skin scrapings - fungal culture
Skin swab - C&S
Skull
Skull - X Ray
SLA antibodies blot
Sleep Studies
Sm/RNP
Smooth Muscle antibodies paediatric
SMUM
Sodium
Sodium concentration in blood
Soft Tissue - US
Soluble Transferrin Receptor
Species and comment
Sphingomonas paucimobilis
Spine - Thoracolumbar - MRI
Spine - Whole - MRI
Spine (Cervical) - X Ray
Spine (Cervical + Thor) - MRI (Contrast)
Spine (Cervical) - CT
Spine (Cervical) - MRI
Spine (Cervical) - MRI (Contrast)
Spine (Cervical) CT - Neuro
Spine (Cervical) Lateral - X Ray
Spine (Cervical) Odontoid Peg - X Ray
Spine (Lumbar + Sacral) - X Ray
Spine (Lumbar) - CT
Spine (Lumbar) - MRI
Spine (Lumbar) - MRI (Contrast)
Spine (Lumbar) - X Ray
Spine (Thoracic) - CT
Spine (Thoracic) - MRI
Spine (Thoracic) - X Ray
Spine (Whole) - MRI
SPJUNDR
Spleen - US
Sputum - C&S
Sputum - Cystic Fibrosis
Staphylococcus aureus
Staphylococcus epidermidis
Staphylococcus hominis
Stapylococcus epidermidis
Stenotrophomonas maltophilia
Sterno-Clavicular Joint L. - X Ray
Sternum - X Ray
Sticky Label on Sample
Streptococcal serology
Streptococcus viridans group
Submandibular - US
SUMMARY
SUMMARY1
SV(LVOT)
SV(MOD-bp)
SV(MOD-sp2)
SV(MOD-sp4)
SV(sp4-el)
Synovial fluid - C&S
Synovial fluid - C&S & AFB
Synovial Fluid TB
T. Tube Cholangiogram
Tacrolimus Assay - Liver Unit
TB Test - Cystic Fibrosis Sputum
TB Test - Sputum & TB Test
TCEL CALCULATION
TCOM
Teicoplanin Pre level
Teicoplanin.
Tempero Mandibular Joint L. - X Ray
Tempero Mandibular Joint R. - X Ray
Tempero Mandibular Joints - X Ray
Test Comment
Testo/SHBG Ratio
Testosterone
Tetracycline.
Thiopurine Methyl Transferase
Thoracic Inlet - X Ray
Thoracolumbar Spine - X Ray
Thorax - MRI
Thorax - MRI (Contrast)
Thorax (Abdomen + Pelvis) - CT(Contrast)
THR
Throat swab - C&S
Thrombin time (control)
Thrombin time (patient)
Thrombin Time Control
Thrombin Time.
Thumb L.
Thumb L. - X Ray
Thumb R. - X Ray
Thyroid - US
Thyroid Imaging + Uptake (Tc) - NM
Thyroid Peroxidase antibodies
Thyroid Stimulating Hormone
Thyroid Stimulating Hormone Receptor Abs
Thyrotoxicosis Therapy - 131 I - NM
Tib & Fib L.
Tib & Fib R.
TIBC Result
Tibia & fibula Bilateral - X Ray
Tibia + Fibula - MRI
Tibia + Fibula L. - X Ray
Tibia + Fibula R. - X Ray
Tibula + Fibula L. - X Ray
Tibula + Fibula R. - X Ray
Time of collection (mins)
TLC
TLC % Predicted
TLC Predicted
TLCO
TLCO % Predicted
TLCOc SB
TLCOc SB % Predicted
TLDL
Tobramycin (Pre-dose)
Toe R. Great - X Ray
Toes L. - X Ray
Toes R. - X Ray
TORCH screen - Adult
Total Cholesterol-HDL Ratio
Total Iron Binding Capacity
Total Porphyrin
Total Protein
Toxoplasma Dye Test.
Toxoplasma IgG
Toxoplasma IgM (EIA)
TR Max PG
TR Max vel
Trans Jugular Liver Biopsy
Transcranial Imaging (Paediatric) - Vasc
Transplant - Adult Haem Donor
Transplant - Adult Haem Recipient
Transplant - Adult Liver Donor
Transplant - Adult Post BMT CMV/Adeno
Transplant - Adult Post Liver CMV load
Transplant - Adult Pre-Liver Recipient
Transplant - Adult Renal
Transplant - Paed Liver CMV/EBV load
Trauma Series - X-Ray
Tree Pollen (Standard) Mix
Treponemal antibody (TREG)
Treponemal antibody.
Treponemal serology
Trichomonas
Tricuspid Valve
Triglyceride
Tri-iodothyronine
Trimethoprim.
Troponin I
Troponin I (cardiac markers)
Troponin I.
Tryptase
TSDL
TSH (0 mins)
TSH (20 mins)
TSH (60 mins)
TSH Receptor Binding Antibody
T-Spot TB
TTCN
TTSC
Tuna
TV max PG
TV mean PG
TV V2 max
TV V2 mean
TV V2 VTI
Tx Reaction Investigation Report
Type II PNH clone
Type III PNH clone
U OX
U1RNP
U-albumin/creat. ratio
UGT1
UGT2
Ulcer swab - C&S
Ultrasound Foreign Film
UMYR
Under General Anaesthetic
UNIT
Unknown HB%
Unlabelled Sample
UOX/CR
Urea
Urea (Post dialysis)
Urea (random urine)
Urea (serum)
Urea Reduction Ratio
Urethral swab - M
Urethrogram
Uric Acid
Urinary Albumin
Urinary Albumin (24 hour)
Urine - bag specimen - M
Urine - catheter specimen - M
Urine - clean catch - M
Urine - dipslide - C&S
Urine - early morning - TB culture
Urine - early morning urine TB culture
Urine - Legionella Antigen Test
Urine - M
Urine - mid stream specimen - M
Urine - other specimen - M
Urine - Schistosome Ova
Urine Albumin conc.
Urine Blood Screen
Urine Calcium
Urine Calcium excr.
Urine Calcium per unit time
Urine Copper
Urine Copper per 24hrs
Urine Cortisol
Urine Cortisol over 24hrs
Urine Creatinine
Urine Creatinine excr.
Urine Creatinine per unit time
Urine Creatinine/ 24 hour (enzymatic)
Urine Dopamine (24 hour)
Urine Glucose
Urine Haemosiderin
Urine HMMA (24 hour)
Urine HVA (24 hour)
Urine Legionella antigen
Urine Leucocytes
Urine Nitrite.
Urine Noradrenaline (24 hour)
Urine Oxalate per unit time
Urine pH
Urine Phosphate
Urine Phosphate per unit time
Urine Pneumococcal antigen
Urine Potassium
Urine Potassium excr.
Urine protein
Urine Protein Electrophoresis
Urine Protein per unit time
Urine Protein Screen
Urine Sodium
Urine Sodium excr.
Urine Steroids
Urine Tot. Protein excr.
Urine Urate
Urine Urate per unit time
Urine Urea per unit time
Urine Urobilinogen
Urine virology (specify which virus)
Urine vol
Urine Volume
Urine volume.
USS Aspiration
USS Bowel TP
USS Collection
USS Gallbladder
USS Gallbladder + Liver
USS Kidneys
USS Liver
USS Liver Paediatric
USS Pancreas
USS Pelvis
USS Penis
USS Scrotum
USS Soft Tissue
UURO
V:C
VA
VA % Predicted
Vaginal (High) swab - M
Vaginal (Low) swab - M
Vaginal swab - M
Valproate levels
Vancomycin (Pre-dose)
Vancomycin level - pre-dose
Vancomycin level - random
Vancomycin resistant E. faecium
Vancomycin.
Varicella zoster virus DNA:
Vascular (Other Investigation) - Vasc
Vascular Line Tip
VC
VC % Predicted
VC Post
VC Post % Predicted
VC Post Bronchodilator
VC Post Bronchodilator % Change
Venogram - MRV
Venous Bicarbonate
Venous H+
Venous PCO2
Venous pH

Venous PO2
Vertex WB Imaging
Vesicle fluid - virus isolation
Vesicle fluid molecular virology
Video - Cystourethrography
Video Capsule Enteroscopy Report
VII:C
VIII:C
Virology Comment
Virology Results
Virus Isolation - miscellaneous sample
Virus Isolation - non blood sample
Vitamin A
Vitamin D
Vitamin E
Vitamin K
Volume (0-5 hrs)
Voriconazole level
VRE Screen
VRE Screen.
Vulval swab - M
VWAG
VWAT
VWF Activity
VWF Antigen
VWF Antigen.
WACT
Ward Unit Summary
WB 123-I Imaging - NM
WBC
WBC (Labelled) Imaging (HMPAO) - NM
WBC (Labelled) Imaging (Indium) - NM
WBC count (CSF)
WBC count (Fluid)
WBC Imaging - Indium - NM
WBC1
White Cell Morphology
Wound swab - C&S
Wrist - MRI
Wrist - MRI (Contrast)
Wrist - US
Wrist + Hand - MRI
Wrist Bilateral - X Ray
Wrist L.
Wrist L. - Arthrogram
Wrist L. - MRI
Wrist L. - US
Wrist L. - X Ray
Wrist R. - MRI
Wrist R. - US
Wrist R. - X Ray
Wrong DOB on Sample
WSUM1
X:C
XI:C
XII:C
Yeast (GYYC)
Yeast Cells
Yeast Cells
Yeast.
Yeasts (WPYE)
Yersinia
Yersinia serology
Zinc (Serum)
1-25 Di-Hydroxy Vitamin D
19G1
7 day lifecard
AB INT 5
Actinomyces
Adrenal - US
Adrenaline (random urine)
Alk. Phos. Isoenzymes
Alpha-1-antitrypsin genotype
Amfetamine Class
Amikacin (Pre-dose)
Amikacin level - pre-dose
Amikacin.
Anaerobic Streptococci
Angiogram CT - Neuro
Angiography 4 Vessel - Neuro
Ankle L. - US
Ankle R. - US
Ankles + Feet - MRI (Contrast)
Anti Striated Muscle Antibody
Anti Thyroglobulin
Anti-liver kidney microsomal ELISA
Anti-mitochondrial M2
Anus + Rectum - MRI
Aortic arch
APTR
Arm - Upper R. - MRI
Arm (Upper) Both - CT
ASPEIA
Aspergillus flavus
Aspergillus fumigatus
Aspergillus niger
Aspergillus terreus
Axilla L. Biopsy - US
Barium Meal (MS)
Barium Swallow
Bile Acids (urine)
BK virus DNA
BK virus DNA - GSTT.
Blood Culture
Blood Cobalt
Blood Results - Rheumatology
Bone - CT (Guided Biopsy)
Bone Densitometry - NM
Bone Marrow Report
Bone Marrow Sample - C&S
Booked Admissions
Breast Abscess L. - US
Breast Abscess R. - US
Breast L. - Fine Needle Aspiration - US
Bronchial Washing TB
Bronchitis External Results
Bronchoalveolar Lavage
CA153
Calcitonin
Candida glabrata
Cardiac - 7 day Lifecard
Cardiac - CPET
Cardiac - Pacemaker Check
Carotid Angiogram L. - Neuro
Carrot Potato Spinach Cucumber
CCASC
CD19 %
CD19 CELLS
CD4C
Ceftazidime.
Ceftriaxone.
Central line tip - C&S
Cervical Spine
Cheese cheddar type
Cheese mould type
Chest - AP
Chest (Lat Decub) - X Ray
Chest (Lateral) - X Ray
Chest High Resolution Prone - CT
Chest Screening
Chlamydia trachomatis swabs
Chromium
Chromium (Serum)
Chromogranin A
Chromogranin B
CK Iso-enzymes
Clarithromycin.
Clavicle R. - X Ray
Cobalt (Serum)
Colonoscopy - CT
Combined Chlamydia & Gonorrhoea
COMBO
Comment..
Cough Monitoring
Cough swab - C&S
Cow's Milk
CRESER
Cryoglobulins
CSF AFP
Cystic Fibrosis Genotype
Dehydro-epiandrosterone
desc Aorta
Dialysis Fistula - PTA
Dialysis Line
Digoxin level
Digoxin.
Diphtheroids
Drain site swab - C&S
Dry Weight
Egg White
Egg yolk
Embolisation - Neuro
Enterobacter aerogenes
Enterococcus faecium
Enterococcus species
Enterovirus RNA
Epidural Injection - X Ray
Ertapenem.
Exner KCT
F.B. Demo.
Facial Bones CT - Neuro
Faeces - follow up culture
Faeces - virus isolation
Feet - MRI
FEV 1 % VC MAX Best
FEV 1 % VC MAX Best % Predicted
FHIPRM
Finger R. Little - X Ray
Finger R. Middle - X Ray
Finger R. Ring - X Ray
Fistulogram
Fluid TB
Fluid Triglyceride.
Fluid Urate
Foetus - MRI
Foot R. - MRI
Foot R. - US
Forearm R.
Forearm R. - MRI
Fosfomycin.
FTHOXM
FTYP
Fungal Culture (FFUN)
Fungal Culture (FUNC)
FUPLLM
Gallium Imaging - NM
Gastrin
GAWK (B)
Glucagon
Gram positive cocci
GSR1
H2CO3 on 1/2l O2
Hand L. - CT
Hand R. - US
HBGA
Heel R. - X Ray
Helicobacter pylori serology
Hep_HBVDRM
Hep_HBVGen
Hep_HDVRNA
Hepatic Iron Index
Hickman Line - X Ray
Histology Supplementary Report 1
HIV type 1&2 antibody/antigen
HIV-1 proviral DNA Send Away
HLA broad genotype
HLA-B27
HSV 1/2 DNA (Taqman)
Immunofixation
Index Finger R.
Infection screen - bacteriology
Inferior Vena Cava Filter - Temporary
Interim adenovirus DNA result
Interim VZV DNA result
IVP/IVU
KHMDC Integrated Report
Klebsiella Screen
Kluyvera species
Lateral Foot L. - X-Ray
Lateral Soft Tissue Neck
Leg - Lower L. - MRI
Leg L. - Arterial Duplex - Vasc
Liver - CT
Liver Biopsy - US (high risk/inpatient)
Liver Biopsy - US Guided Test Only
LKMT
Lower GI Endoscopy - Fluoro
Lower limb pinning Left
Lung V/Q Imaging (Per MAA) - NM
Metabolic Skelatal Survey
Methotrexate levels
MIBG Therapy - I 131 - NM
Micturating Cystourethrogram
Micturating Cystourethrogram (MS)
Milk (boiled)
Miscellaneous Biochemistry (1)
MITT
MLOLLC
MLOLRC
MMEF 75/25 Best
MMEF 75/25 Best % Predicted
Mobile - US
MPA area
MPA diam
MRI Pelvis
MSHRLC
MTHILC
Mumps Serology
Myositis Specific ENA
N123W2
Needlestick Injury - virology
Neurotensin
NM Gallium Sarcoid
NM Hepato-Biliary Imaging
NM I-123 MIBG WITH SPECT
NM Kidney Imaging (DMSA)
NM Labelled WBC Imaging - Ind
NM Lung (V/Q) Imaging - VQ DTPA
NM Renography (DTPA) Imaging
NM Thyroid Imaging - Tc I + Uptake
NM Whole Body FDG PET CT (King's College
NM131W
NMIN04
NMIN24
NMINJ
NMPINJ
Non Gynae Supplementary Report 1
NPA - bordetella pertussis
Nuclear Medicine Foreign Film
Octreotide Imaging - NM
OGD - Additional Procedures (Therapeutic
Opiates Screen
Organisms (GYO1)
Organisms (GYO2)
Organisms (GYO3)
P/T Amylase Ratio
PaCO2 on 1/2l O2
Pancreatic Amylase
Pancreatic Polypeptide
PaO2 on 1/2l O2
Paracetamol
Paraprotein
Parathyroid SPECT CT - NM
Parvovirus B19 DNA Send Away
Parvovirus B19 genotype
Parvovirus B19 past exposure status.
PEF Best
PEF Best % Predicted
Perfusion - CT Neuro
pH on 1/2l
Phenobarbitone levels
PI dec slope
PI max PG
PI max vel
PI P1/2t
PNH Test
Polymixin.
Popliteal fossa R. - US
Porphyrins Random Urine
Post Transplant - Olt - US
Propionibacterium species
Proteus mirabilis
Pseudo-aneurysm R.- US
Duplex - Vasc
Pseudomonas species
Pus cells (GYPU)
PVA(I
PVA(V
Q fever Send Away.
R92
Random Urine 5HIAA.
Random Urine Creatinine
Random Urine Creatinine__
RAP systole
Reference Laboratory Comment (FRFC)
Referred to (FRE1)
Renal - MRI (Contrast)
Renal Biopsy - US
Reptilase Control
Reptilase Time.
Retrograde Pyelogram - X Ray (Mobile)
Rivaroxaban
RVOT diam
RVSP
Sacro-Iliacs
Sacrum - MRI
Salivary Amylase
Salmonella enteritidis
Salmonella species ISOLATED
SaO2 on 1/2l O2
SaO2Ox on 1/2l O2
SaO2Ox on 1l O2
SaO2Ox on 2l O2
Scapula L. - X Ray
Serotype 1
Serotype 14
Serotype 18C
Serotype 19F
Serotype 23F
Serotype 4
Serotype 5
Serotype 6B
Serotype 7F
Serotype 9V
Shoulder Axial - X Ray
Shoulder Rt - CT
Sinus - CT (Contrast)
Skelatal Survey (Metabolic) - X Ray
Skeletal Survey (Myeloma) - X Ray
Skin Immunofluorescence
Skull - Lateral Only - X Ray
Somatostatin
Sphingomonas spiritivorum
Spine (Whole) - MRI (Contrast)
Steroid Profile (24 hour urine)
Streptococcus mitis
SU2
Subphrenic - US
Sulphamethoxazole.
Sweat chloride
TE Miscellaneous
Tempero Mandibular Joints - MRI
Thoracic Spine
Three Dimensional - CT
Time since last dose
Tissue Copper
Tissue Iron
Tissue TB
TNUN
Toe L. Great - X Ray
Toes L.
Tomogram - Bone Imaging - NM
Total Amylase
Toxoplasma serology SA
Transcranial Imaging (Adult) - Vasc
Transplant - Paed Haem Recipient
TSDD
UMC2
Under Sedation
UPELV
Upper GI Paeds
Upper limb pinning Left
Urine - Pneumococcal antigen
Urine - terminal specimen
Urine Adrenaline (random)
Urine Creatinine.
Urine Dopamine (Random)
Urine Noradrenaline (Random)
US Renal Biopsy
USAL
USS Breasts
USS Kidneys/Bladder
USS Neck
USS Thyroid
Vancomycin level
Vancomycin level - post-dose
Vasoactive Intestinal Polypeptide
VC MAX Best
VC MAX Best % Predicted
Venoplasty
Video Fluoroscopy Swallow
Visual Field
WB 1-131 Imaging - NM
WB 123-I mIBG Imaging - NM
WBC count (PD fluid)
WBC Imaging (Technitium HMPAO) - NM
Weight of sweat
Whole Body - MIBG Therapy - I 131 - NM
Whole Body - MRI
Wrist Both - US
Wrist L. - CT
Wrist R.
Yeast
157

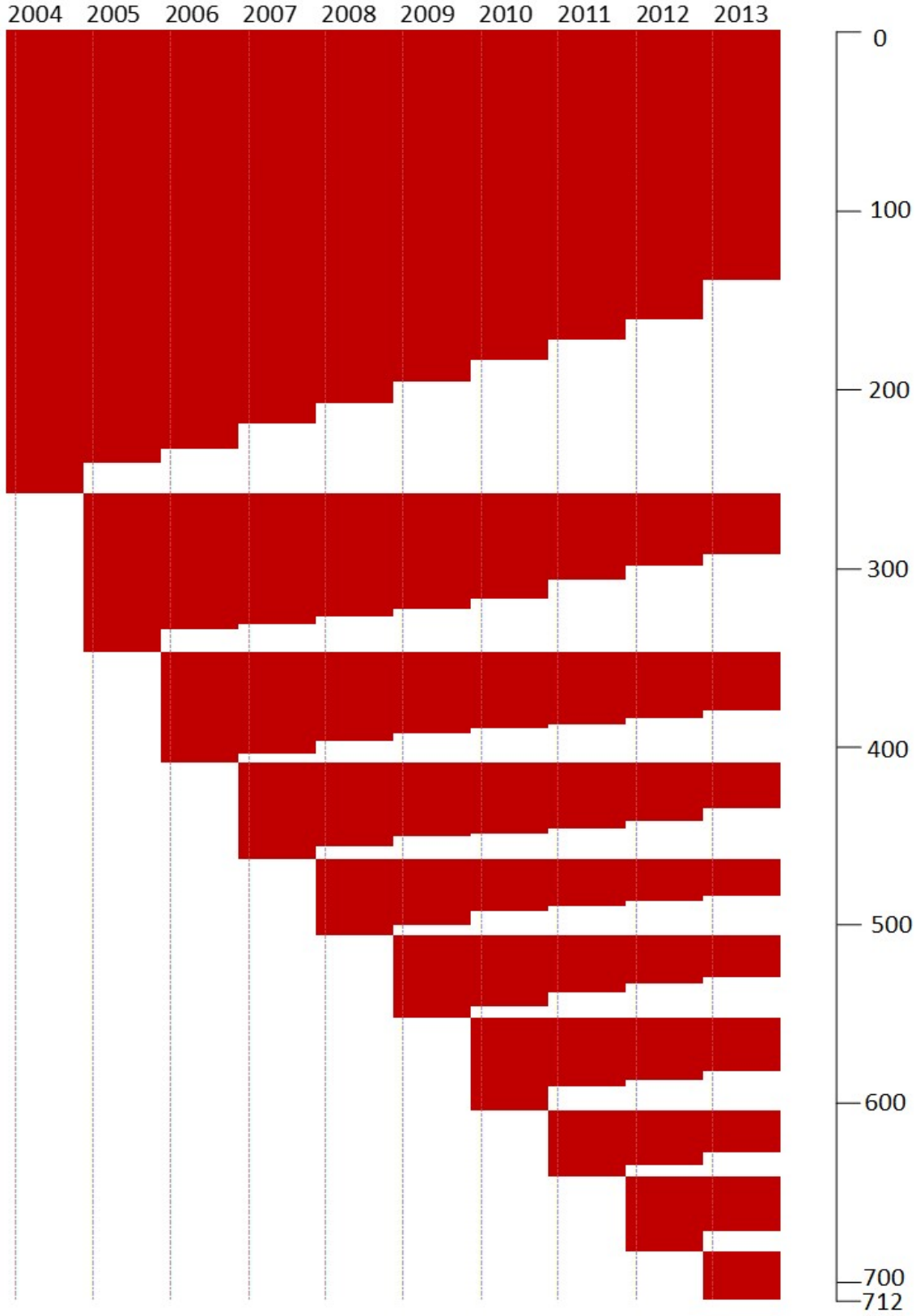
Appendix 2
Data cleaning pipeline

- Modify: rows which have incorrectly formatted columns e.g. instead of 9 columns, there are 10 or 11 columns, see Figure 4 Raw clinical data format
    - Organism names include multiple "+"
    - Oestradiol or parathormone results
    - Culture results
- Delete:
    - Meaningless test names (often comments or addendums to the actual test which remains in the dataset), see Appendix 4.
    - Meaningless test results (often reflect problems with the sample), see Appendix 5.
- Categorise visit: based on visit type code and location, see Appendix 6. This creates two notable subgroups:
    - **all adult haematology outpatients**
    - **all adult inpatient admissions (to haematology or medical wards only)**
- Correct:
    - Errors where data = "NULL" for either date of "test requested" or date "results released"
    - All variable formats (date, string, number)
- Delete:
    - any entry outside the "admission dates". This ensures that we only capture tests done on the same day as a haematology clinic day, or during a haematology admission. Date errors can occur when clinicians select the incorrect visit when requesting a test.
- Manipulate data:
    - Re-configure the data to create a table with:
        - ROWS: unique hospital numbers /clinic dates as rows
        - COLUMNS: laboratory tests
    - Transform values containing '<' or '>' to numerical values: eGFR, CRP, NRBC, Urine ACR, haptoglobin, folate
- Merge: laboratory variables with changes to name during study period:
    - 'Hb' and 'Hb.', 'MCHC' and 'MCHC.', '%hypo' and '% hypo'; 'eosinophil' and 'eosinophil.'; and 'iron (serum)' and 'serum iron'
- Final laboratory results: see Appendix 7
- Add:
    - Demographic data:
        - date of birth
        - age at test (exact)
        - sex
        - sickle genotype
        - alpha globin genotype
        - G6PD status
        - ethnicity: whether African / African-Caribbean or not (one patient is Yemeni)
        - alpha globin status: using the KCH red cell laboratory customised genetics panel which includes the $-\alpha^{3.7}$ (common African deletion) and $-\alpha^{4.2}$ alpha thalassaemic variants only.
    - Study ID numbers
    - HbFg: see chapter 4
        - HbFg = 1.89 +0.14×rs6545816+0.3 ×rs1427407+0.13 ×rs66650371+0.1×rs7482144

        - Genotype = 0,1,2 depending on number of HbF boosting alleles present
    - Calculated MDRD_GFR (mL/min/1.73 m^2) = 175 × (S_{cr} /88.4)$^{-1.154}$ × (Age)$^{-0.203}$ × (0.742 if female) × (1.212 if African American)
    - Proteinuria category – three categories as per NICE: uACR <3, <30, over 30
    - Haemolytic index for the validated values – based on a principal components analysis of lactate dehydrogenase (LDH), aspartate transaminase (AST), absolute reticulocyte count, total bilirubin. See section 2.5.
    - First and last dates of haematology review
    - Hospitalisation rates, see section 2.6.
        - Per patient: number of admissions/ observation period
    - Mortality and date of death, see section 2.7.

Appendix 3
Infographic demonstrating years of attendance of 712 KCH adult patients

Patient years of attendance to the KCH haematology 2004-13 inclusive (10 years). Total 712 patients, each (narrow) row represents one patient over the 10 years, with a scale (number of patients) on the right hand side.



Appendix 4
List of meaningless test names

Meaningless test names deleted during data cleaning (often comments or addendums to the actual test which remains in the dataset).

1st Addendum Diagnosis	Comments (MCOM)	Gynae Suggested Management	Non Gynae Report
1st Addendum Microscopy	Comments (RCOM)	Interpretation	Non-gynae Clinical History
2nd Addendum Diagnosis	Comments (WCOM)	Summary	Non-gynae Cytopathology Report
2nd Addendum Microscopy	compheno	Interpreting Physician	Non-gynae Specimen
Additional Information	Conclusion	LABCODE	OUTC
Age Band (Years)	Date result received (TBRE)	Misc. Fluid Type I	Paed Haem External Test Results
Age Band (Yrs)	Date sent to MRU	Miscellaneous Comments	REPT1
Authorised by	Group + Save Rejected	Miscellaneous Immunology Test	SAM
AVA(I	Gynae Clinical History	Miscellaneous sample - M	Sample sent to
AVA(V	Gynae Comment	Miscellaneous tissue sample - C&S	Save Sample
Comment	Gynae Cytopathology Report	NM BD CLIN H&S&L	Sticky Label on Sample
Comment (APBS)	Gynae FHSA	No Name on Group and Save Sample	Under General Anaesthetic
Comment (TBC1)	Gynae Infection	No of samples received	Unknown HB%
Comment (UCOM)	Gynae Recall Status	No req form received for GS only sample	Unlabelled sample
Comment Electrophoresis	Gynae Report part 1	No sample received for GS only request	Wrong DOB on sample
Comment Enzymes	Gynae Report part 3		
Comment G6PD	Gynae Specimen Description		
Comment_			

Appendix 5
List of meaningless test results

Meaningless test results deleted during data cleaning (often reflecting problems with the sample).

aged
B
D)
DUPLICAT
EDTA
HAEM
HEMOLY
INS
INSU
INSUFF
N/A
NA
NOT DONE
RECEIVED
UNS
UNSUIT
x-nores

Appendix 6
Categorisation of clinical visit type for clinical dataset

Categorisation into clinical visit type: based on visit type code (column 9) and location (column 6)

Categorisation	Visit Type Code	Location (see below for definitions of wards)
Adult Inpatient	Inpatient or Emergency	Include Haematology, Clinical Decision Unit, Medical, Guthrie, ITU, CCU/Acute Coronary Wing HDU.
Excluded Inpatient	Inpatient	Obstetric, Apheresis, chemo, chest, endo, medihome, Cardiac Catheter Suite, Frank Cooksey, Haemodialysis Unit, Clinical Research, Renal satellite unit
Paed Inpatient	Inpatient	Paed wards
Haematology Day Case	Inpatient	Haematology Day Case only
AdultHaemOutpatient	Outpatient	Haematology outpatient
Paed Outpatient	Outpatient	Paed haem OP or Paed OP
Adult Other Outpatient	Outpatient	Any other OP clinic
AdultEmergency	Emergency	A&E or CDU (not required)
Results	Results	

Ward definitions

ITU	Paediatric	Surgical	Medical	Obstetric	Cusp of IP/OP	Haem
Kinnier Wilson Liver ICU Medical Critical Care Unit	Butlin Lion Philip Isaacs Day Care Princess Elizabeth Rays Of Sunshine Toni & Guy Mountbatten	Acute Surgical Unit Brunel Christine Brown (but medical historically) Cotton Katherine Monk Lister Lonsdale Mary Ray Matthew Whiting Twining Coptcoat Victoria & Albert	Annie Zunz Dawson Donne Fisk & Cheere David Marsden Leighton Oliver The Friends Stroke Unit Trundle Trundle & Waddington Victor Parsons Unit Medical AAU Todd Guthrie Fisk Cheere Howard Medical Assessment Centre Sam Oram Murray Falconer	Antenatal William Gilliat Maternal Assessment Unit Midwifery Centre Nightingale Birth Centre Postnatal William Gilliat William Gilliat Sylvia Henley	Apheresis Room Chemotherapy Unit Chest Unit Clinical Decision Unit (A&E) Endoscopy Suite Haematology Day Case Haematology OPD Medihome Day Surgery Cardiac Catheter Suite Frank Cooksey COOK	Davidson Derek Mitchell Unit Elf & Libra Guthrie R D Lawrence Waddington

Appendix 7
Clinical information in the clinical dataset

Basic demographics	Attendance and admissions date	Haematology	Biochemistry	Calculated values
HospitalNumber	FirstDate	WBC	Lactate Dehydrogenase	Haemolytic Index
StudyID	LastDate	Hb	Erythropoietin	HbFg
Flag_DNA	YearsObserved	MCV	Level	FlagValidated
Flag_MEGAdata	NumberOfOutpatients	MCH	sTfR	MDRD_GFR
DOB	NumberOfAdmissions	MCHC	C-reactive Protein	NICE proteinuria category
AgeAtTest	AvgLengthOfStay	PCV	Creatinine	MaxUACRcategory
Genotype	TotalLengthOfStay	RBC	Urea	
Alpha	AdmissionFreqPerYear	RDW	Estimated GFR	
Sex	FlagMoreThan2Yrs	RRBC	CystatinC	
G6PD	YrsObs	% Hypo	U-albumin/creat. ratio	
AfricanOrCaribbeanFlag		Neutrophils	uPCR	
DateAdmission		Lymphocytes	Potassium	
DateDischarge		Monocytes	Sodium	
Location		Eosinophils	Phosphate	
DateTestRequested		Basophils	Bilirubin (Total)	
TypeCode		PLT	Bilirubin (Conjugated)	
		MPV	Alkaline Phosphatase	
		PDW	ALT	
		Absolute Reticulocyte Count	Aspartate Transaminase	
		Reticulocyte Percent	Gamma-glutamyl Transferase	
		Reticulocyte HB Content	Total Protein	
		IRF	Albumin	
		NRBC	Globulin	
		Nucleated RBC	Calcium	
		Haemoglobin F % (HbF)	Corrected Calcium	
			Ferritin	
			Serum Iron	
			corrected	
			TIBC Result	
			% Iron Saturation	
			B12	
			Folate	

Appendix 8
Comparing HbF% in pregnancy/not in pregnancy

Of 712 patients with SCD at KCH, 25 women (all sickle genotypes) had paired HbF% values both *during first trimester* pregnancy and in steady state (non-pregnant). None of these values were taken within three months of either a transfusion or hydroxycarbamide. Median non-pregnant HbF%=7.0 (range 0.2-22.5, IQR 4.7-10.15), median pregnant HbF%=8.1 (range 0.3-22, IQR 5.7-13.75), for distribution of HbF% in these two states, see Figure 14.

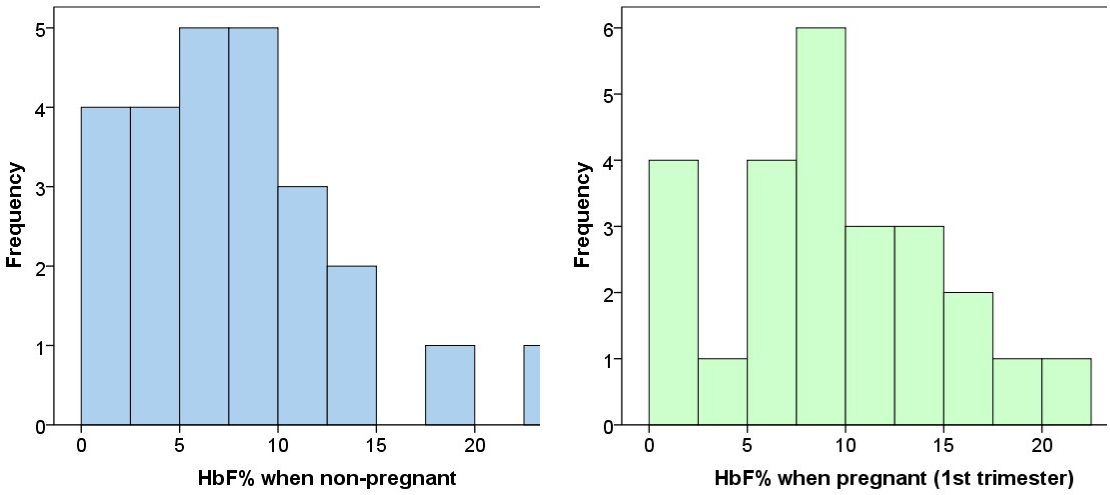


Figure 14 Histograms to compare HbF% when pregnant/not pregnant

HbF% values were normalised by applying a natural logarithm (Ln) and then compared using a paired student's t-test. Pregnancy Ln HbF% were significantly higher than non-pregnant values ($t=4.835$, $df=24$, $p<0.0001$).

Appendix 9
Comparing HbF% in the acute setting versus steady state

Of 712 patients with SCD at KCH, 127 (all sickle genotypes) had paired HbF% values available both in steady state and during acute hospital admission. None of these values were taken within three months of a transfusion or being on hydroxycarbamide. Median steady-state HbF%=4.0 (range 0.2-27.7, IQR 1.8-8.3), median acutely unwell HbF%=4.5 (range 0.2-28, IQR 1.8-8.2), for distribution of HbF% in these two states, see Figure 15.

For HbSS/SB0 thalassaemia only (N=99), median steady-state HbF%=5.850 (range 0.4-27.7, IQR 2.5-9.25), median acutely unwell HbF%=5.5 (range 0.3-28, IQR 3-9.275), for distribution of HbF% in these two states, see Figure 16.

For HbSC only (N=25), median steady-state HbF%=0.7 (range 0.2-8.7, IQR 0.6-1.45), median acutely unwell HbF%=0.7 (range 0.2-9.5, IQR 0.55-1.6), for distribution of HbF% in these two states, see Figure 17.

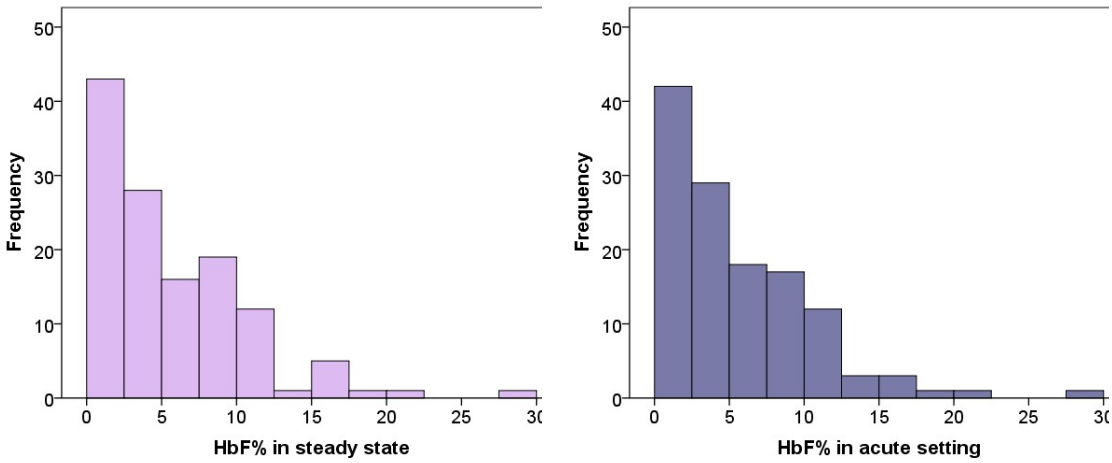


Figure 15 Distribution of HbF% in steady state and acute settings for all sickle genotypes

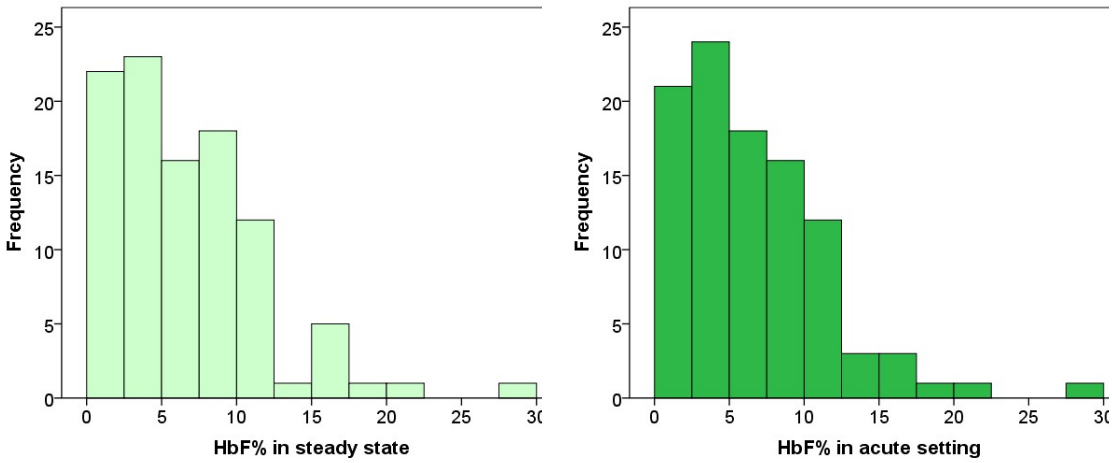


Figure 16 Distribution of HbF% in steady state and acute settings for HbSS/HbSβ0 patients

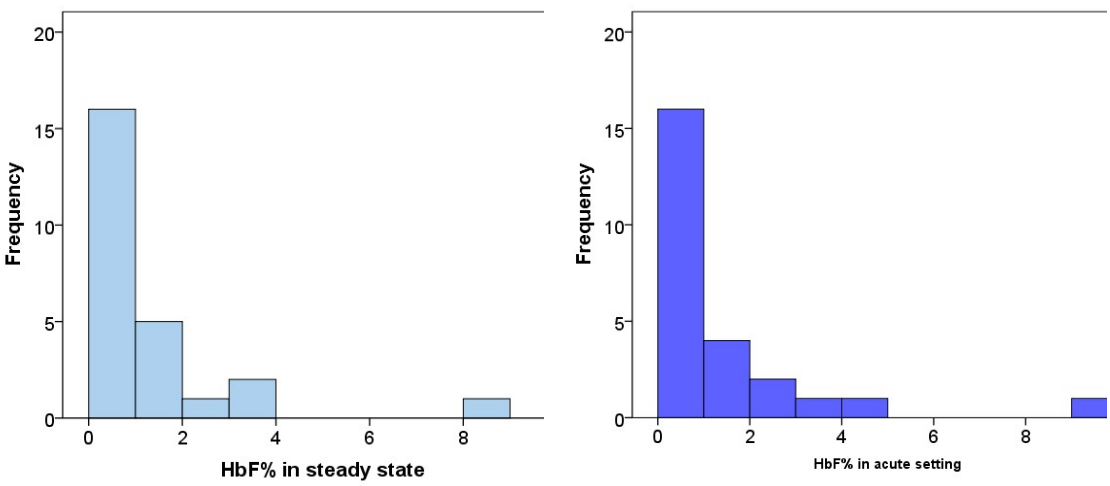


Figure 17 Distribution of HbF% in steady state and acute settings for HbSC patients

HbF% values were normalised by applying a natural logarithm (Ln) and then compared using a paired student's t-test. There was no statistical different between steady state and acute Ln HbF% (t=1.340, p=0.183).

For HbSS/SB0 thalassaemia (N=99), there was no statistical different between steady state and acute Ln HbF% (t=1.267, p=0.208).

For HbSC (N=25), there was no statistical different between steady state and acute Ln HbF% (t=0.579, p=0.568).

Appendix 10
Survival in adults with sickle cell disease in a high-income setting

From www.bloodjournal.org by guest on September 3, 2017. For personal use only.

To the editor:

Survival in adults with sickle cell disease in a high-income setting

Kate Gardner,[1,2] Abdel Douiri,[3-5] Emma Drasar,[1,2] Marlene Allman,[2] Anne Mwirigi,[2] Moji Awogbade,[2] and Swee Lay Thein[1,2]

[1]Molecular Haematology, Division of Cancer Studies, King's College London, London, United Kingdom; [2]Department of Haematological Medicine, King's College Hospital National Health Service (NHS) Foundation Trust, London, United Kingdom; [3]Division of Health and Social Care, King's College London, London, United Kingdom; [4]National Institute for Health Research (NIHR) Biomedical Research Centre, Guy's and St. Thomas' NHS Trust and King's College London, London, United Kingdom; and [5]NIHR Collaboration for Leadership in Applied Health Research and Care, King's College Hospital NHS Foundation Trust, London, United Kingdom

Survival of patients with sickle cell disease (SCD) in high-income countries has improved greatly in the last 60 years. In 1960, it was described as a "disease of childhood"[1] whereas 25 years later, the Cooperative Study of Sickle Cell Disease reported that 85% of hemoglobin SS (HbSS) patients lived to adulthood. More recently, the estimate is 99% in London,[2] 97% in Paris,[3] and 94% in the United States.[4]

Survival estimates have continued to improve; in 1994, the median survival for patients with HbSS/Sβ^0 thalassemia was estimated at 42 to 48 years,[5] increasing to 53 to 58 years in Jamaica in 2001[6] and 58 years in the United States in 2014.[7] Nonetheless, the life expectancy of patients with SCD is still shortened by >2 decades compared with the general population.[8-10]

This study evaluates survival among adult patients with SCD followed at a single center in the United Kingdom. The study was an audit of clinical practice, and involved analysis of data collected in routine clinical care. All procedures followed were in accordance with the ethical standards of the Helsinki Declaration of 1975, as revised in 2008. Seven hundred twelve adult patients with SCD (16-80 years of age) at King's College Hospital (London, United Kingdom) were observed over 10 years (2004-2013 inclusive) and mortality outcome was identified (5268 patient-years of observation; median, 8 years of observation per patient).

All patients, except for 1, were of African or African-Caribbean heritage. Of the 712 patients, 444 (62%) were HbSS, 229 (32%) were HbSC, 33 (5%) were HbSβ^+ thalassemia, and 6 (1%) were HbSβ^0 patients. For subanalysis, we considered HbSS and HbSβ^0 thalassemia patients as a group. The median age for HbSS/Sβ^0 patients was 32 years (interquartile range [IQR], 25-43 years); HbSC, 39 years (IQR, 29-48 years); and HbSβ^+ thalassemia, 40 years (IQR, 31-58 years). α-Globin genotypes were available in 542 patients (76%) of which 62% were αα/αα, 32% αα/α−, and 5% α−/α− genotypes. During the study period, 72 patients (all HbSS) had received hydroxyurea therapy, and 71 patients had received regular blood transfusion. We underline the low uptake of hydroxyurea therapy in our cohort. Oxygen saturations by pulse oximetry and laboratory data collected during outpatient clinic attendance were documented. Laboratory results were averaged over the 10-year period to create a "steady-state" value for each patient. The mean number of hospital admissions under hematology for each patient was calculated from the total admissions/number of observed years of admissions. Local hospitals were contacted to identify outcome in patients not seen in 2012 or 2013; despite this, 104 (14.6%) were not reviewed in 2012 or 2013. Data collection finished on July 31, 2015.

IBM SPSS Statistics 22 was used for statistical analyses. Continuous variables were log-transformed where necessary to obtain normalized distributions. Kaplan-Meier survival analysis considered

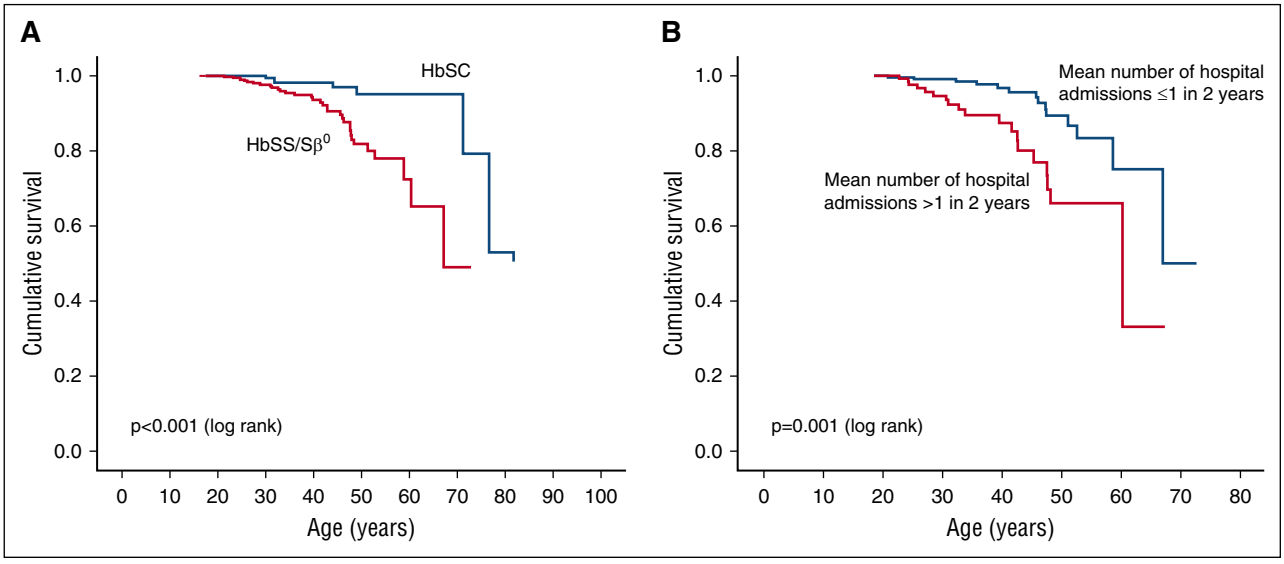


Figure 1. Kaplan-Meier survival curves. (A) Survival curve by sickle genotype. (B) Survival curve for HbSS/Sβ^0, by hospitalization frequency.

From www.bloodjournal.org by guest on September 3, 2017. For personal use only.

Table 1. Univariate Cox regression analysis for HbSS/HbSβ^0 thalassemia

HbSS/HbSβ^0 thalassemia	Hazard ratio (95% CI)	*P* value
Demographics		
α-Thalassemia*	1.34 (0.67-2.71)	.411
Sex†	0.67 (0.34-1.34)	.261
Admissions		
High admission rate, >0.5/y‡	*3.13 (1.57-6.26)*	*.001*
Hydroxyurea use	1.48 (0.57-3.86)	.42
Transfusions	1.00 (0.35-2.87)	.99
Steady-state O_2 saturations		
Oxygen saturations <95%§	*2.84 (1.36-5.92)*	*.005*
Hematology		
White blood cell count, $\times 10^9$/L	*1.18 (1.04-1.35)*	*.01*
Hemoglobin, g/L	0.98 (0.96-1.00)	.07
Platelets, $\times 10^9$/L	1.00 (0.99-1.00)	.16
Reticulocytes, $\times 10^9$/L	1.00 (1.00-1.00)	.29
Fetal hemoglobin high/low‖	*0.44 (0.20-0.96)*	*.04*
Biochemistry		
Lactate dehydrogenase, IU/L	*1.00 (1.00-1.00)*	*.04*
Ln C-reactive protein, mg/L	*1.98 (1.16-3.38)*	*.013*
Ferritin >1000 μg/L¶	*2.52 (1.21-5.23)*	*.013*
Liver enzymes		
Ln total bilirubin, μmol/L	1.78 (1.02-3.10)	.04
Ln aspartate transaminase, IU/L	*3.84 (2.11-7.00)*	*<.001*
Ln alanine transaminase, IU/L	2.37 (0.91-6.22)	.08
Ln γ-glutamyl transferase, IU/L	1.44 (0.97-2.14)	.07
Ln alkaline phosphatase, IU/L	*3.24 (1.93-5.45)*	*<.0001*
Renal function		
Ln creatinine, μmol/L	*2.11 (1.31-3.40)*	*.002*
Ln urinary albumin creatinine ratio, mg/mmol	*1.34 (1.07-1.68)*	*.01*

Significant hazards (risk factors) are italicized. Nonnormal continuous variables were natural logged for statistical comparison (marked "Ln"). Dichotomous variables were handled as follows:

*α-Thalassemia, default = no.

†Sex, default = male.

‡Mean hospitalization rate, default ≤0.5 admissions per year.

§Oxygen saturations, default = normal (≥95%).

‖HbF based on median split of validated HbF values, default HbF <5.5%.

¶Iron overload based on ferritin >1000 μg/L, default = no.

nonfatal cases as censored at their last clinic visit. Univariate Cox regression analysis was undertaken for the HbSS/Sβ^0 subgroup only to identify risk factors for mortality. Dichotomous variables were handled as follows: α-thalassemia, default = no; sex, default = male; fetal hemoglobin (HbF) based on median split of validated HbF values, default HbF <5.5%; iron overload based on ferritin >1000 μg/L, default = no; mean hospitalization rate, default ≤0.5 admissions per year. We chose this cutoff based on the very skewed data distribution: it is clinically meaningful (equivalent to 1 admission every 2 years) and to ensure we had large enough numbers in the "high admission rate" group for statistical analyses.

During the study period, 43 of the 712 patients (6.0%) died at a median age of 42 years (IQR, 31-48 years). They included 33 deaths in the 450 HbSS/HbSβ^0 group (7.3%), at a median age of 41 years (IQR, 30-47 years), and 8 deaths in 229 HbSC patients (3.5%) at a median age of 46 years (31-72 years). For the HbSS/HbSβ^0 group, Kaplan-Meier analysis gave an estimated median survival of 67 years (confidence interval [CI], 55-78 years), significantly lower than in HbSC ($P < .001$; Figure 1A). For HbSS/HbSβ^0, there was a 90% estimated survival to 45 years (39-51 years), 80% to 51 years (CI, 44-57 years), and 70% to 60 years (CI, 51-69 years).

Subanalysis was undertaken for the HbSS/HbSβ^0 subgroup; the sample size in the HbSC subgroup was too small. Median survival in patients with high hospital admission rates (>0.5 admissions per year) was 60 years (CI, 43-77 years), significantly lower than that in patients with low admission rates (≤0.5 per year) ($P = .001$; Figure 1B).

Univariate Cox regression analysis (Table 1) revealed that neither α-thalassemia nor sex were significant risk factors for death. Lack of difference in survival between the sexes may be due to the low numbers of deaths. Hospitalization frequency was a simple but strong predictor of survival in SCD; the risk of death was more than threefold if patients had high-frequency admissions compared with those with low admission rate. Neither hydroxyurea nor blood transfusion was associated with mortality. This likely reflects both the relatively low use of these therapies in our cohort and also the disproportionate use of these therapeutic strategies in our younger patients, confounding the data. Risk of death was increased nearly threefold if baseline oxygen saturations were low (<95%).

For steady-state laboratory results, risk of death was increased if there was: increased white blood cell count, low baseline HbF level, higher lactate dehydrogenase, higher C-reactive protein, or iron overload (ferritin >1000 μg/L). The correlation of disease severity with iron overload is likely via transfusion rate; it is unclear whether iron overload in itself is an independent risk factor. For hepatic enzymes, risk of death was increased if total bilirubin, aspartate transaminase (AST), or alkaline phosphatase were raised, but neither alanine transaminase nor γ-glutamyl transferase affected mortality risk. This may reflect red cell rather than hepatic origin of bilirubin and AST. Conspicuously, AST provides more dramatic hazard ratios than lactate dehydrogenase as a marker of hemolysis. Both measures of renal dysfunction (creatinine and urinary albumin creatinine ratio) demonstrated significant associations with mortality.

Multivariate Cox regression analysis (Table 2) was based on combining variables associated with risk of death in the univariate analysis, plus sex and age at the start of the study. Variables that remained independently significant after multivariate analysis were high admission rate (>0.5 per year), Ln creatinine, and Ln aspartate transaminase, each associated with striking hazard ratios (Table 2), suggesting that poor renal function, excess hemolysis, and frequent hospital admissions can all contribute independently to mortality risk in SCD.

In this retrospective analysis, we have demonstrated a high estimated survival (median, 67 years) for adults with HbSS/HbSβ^0 at a single UK center, which is markedly higher than recent estimates from other institutions. We speculate the reasons: close monitoring of patients in a specialist hematology clinic, plus regular joint care with other specialists (renal, hepatology, neurology, cardiology, obstetrics, and orthopedics); inpatient management by a dedicated health-care team; on-site erythrocytapheresis; and a focused "transition program" to ensure safe transition of teenagers to the adult service. Four of the 43 deaths were in patients under the age of 25 years: 1 from hemopericardium due to stab wound, 1 from cerebral hemorrhage, and 2 from fulminant hepatic failure. We did not assess the socioeconomic class of each patient, but they were from a broad spectrum of social backgrounds. All of these features are similar to other large sickle centers in the United Kingdom.

We acknowledge some study limitations. As an adult-only study, exclusion of pediatric patients may have inflated survival estimates; however, the vast majority of SCD patients reach adulthood in the

Table 2. Multivariate Cox regression analysis for HbSS/HbSβ^0 thalassemia

HbSS/HbSβ^0 thalassemia	Hazard ratio (95% CI)	*P* value
High admission rate, >0.5/y	2.09 (1.02-4.29)	.04
Ln creatinine	3.13 (1.83-5.33)	<.0001
Ln aspartate transaminase	5.82 (2.93-11.54)	<.0001

Age and sex as cofactors.

From www.bloodjournal.org by guest on September 3, 2017. For personal use only.

United Kingdom.[2] We concede that we did not model for those "lost in transition" between pediatric and adult care. However, all 100 patients who turned 19 years of age in 2008 to 2013 inclusive (data from the King's Pediatric Sickle database) have been seen in the adult clinic. We also recognize some missing data for those not reviewed at the end of the study period, despite repeated attempts to obtain information. We also acknowledge the low uptake of hydroxyurea in our cohort (72 of 450 of HbSS/HbSβ^0 patients).

Although life expectancy for a patient with SCD in the United Kingdom continues to improve, it still falls behind that in the general population in London, where it is 80.3 years for men, and 84.2 years for women.[11] We confirmed known predictors of mortality in SCD including markers of cardiorespiratory dysfunction, renal impairment, and hemolysis as well as frequent hospitalization rate.[5,6,12-14] Although these risk factors are not causative, they certainly contribute to the mortality and morbidity in SCD. These risk factors identify higher risk patients who perhaps should be prioritized for therapies including hydroxyurea and hematopoietic stem cell transplantation.

The current affiliation for S.L.T. is Sickle Cell Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD.

Acknowledgments: The authors thank Clive Stringer (King's College Hospital) for the help in data extraction from the Electronic Patient Record system (EPR). The authors thank David Rees and Sandra O'Driscoll from the King's College Hospital Pediatric Sickle Service for data on pediatric patients.

This work was supported by the Medical Research Council (MRC) UK (MRC nos. G0001249 and ID62593) (S.L.T.). A.D. acknowledges financial support from the National Institute for Health Research (NIHR) Biomedical Research and from the NIHR Collaboration for Leadership in Applied Health Research and Care South London at King's College Hospital National Health Service (NHS) Foundation Trust.

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the UK Department of Health.

Contribution: K.G. and S.L.T. designed the research study; K.G., E.D., and M. Allman collected data; K.G., E.D., M. Allman, A.M., M. Awogbade, and S.L.T. provided patient care and follow-up; K.G. and A.D. analyzed the data; K.G. and S.L.T. wrote the paper; and all authors participated in editing the final version of paper.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Swee Lay Thein, Sickle Cell Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Building 10-CRC, Room 5E-5142, 10 Center Dr, Bethesda, MD 20892; e-mail: sl.thein@nih.gov; and Kate Gardner, Molecular Haematology, King's College London, James Black Centre, 125 Coldharbour Ln, London SE5 9NU, United Kingdom; e-mail: kate.gardner@doctors.org.uk.

References

1. Dacie J. The hereditary haemoglobinopathies. Sickle cell disease and allied syndromes. In: The Haemolytic Anaemias: Congenital and Acquired Part I—The Congenital Anaemias. New York, NY: Grune & Stratton; 1960:243-330.
2. Telfer P, Coen P, Chakravorty S, et al. Clinical outcomes in children with sickle cell disease living in England: a neonatal cohort in East London. *Haematologica.* 2007;92(7):905-912.
3. Couque N, Girard D, Ducrocq R, et al. Improvement of medical care in a cohort of newborns with sickle-cell disease in North Paris: impact of national guidelines. *Br J Haematol.* 2016;173(6):927-937.
4. Quinn CT, Rogers ZR, McCavit TL, Buchanan GR. Improved survival of children and adolescents with sickle cell disease. *Blood.* 2010;115(17):3447-3452.
5. Platt OS, Brambilla DJ, Rosse WF, et al. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med.* 1994;330(23):1639-1644.
6. Wierenga KJ, Hambleton IR, Lewis NA. Survival estimates for patients with homozygous sickle-cell disease in Jamaica: a clinic-based population study. *Lancet.* 2001;357(9257):680-683.
7. Elmariah H, Garrett ME, De Castro LM, et al. Factors associated with survival in a contemporary adult sickle cell disease cohort. *Am J Hematol.* 2014;89(5):530-535.
8. Lanzkron S, Carroll CP, Haywood C Jr. Mortality rates and age at death from sickle cell disease: U.S., 1979-2005. *Public Health Rep.* 2013;128(2):110-116.
9. Hassell KL. Population estimates of sickle cell disease in the U.S. *Am J Prev Med.* 2010;38(suppl 4):S512-S521.
10. Pleasants S. Epidemiology: a moving target. *Nature.* 2014;515(7526):S2-S3.
11. UK Office of National Statistics. Life expectancy at birth and at age 65 by local areas in England and Wales, 2012 to 2014 regional life expectancy at birth. http://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/lifeexpectancyatbirthandatage65bylocalareasinenglandandwales/2015-11-04#regional-life-expectancy-at-birth. Accessed July 31, 2016.
12. Miller ST, Sleeper LA, Pegelow CH, et al. Prediction of adverse outcomes in children with sickle cell disease. *N Engl J Med.* 2000;342(2):83-89.
13. Powars DR, Chan LS, Hiti A, Ramicone E, Johnson C. Outcome of sickle cell anemia: a 4-decade observational study of 1056 patients. *Medicine (Baltimore).* 2005;84(6):363-376.
14. van der Plas EM, van den Tweel XW, Geskus RB, et al. Mortality and causes of death in children with sickle cell disease in the Netherlands, before the introduction of neonatal screening. *Br J Haematol.* 2011;155(1):106-110.

DOI 10.1182/blood-2016-05-716910

© 2016 by The American Society of Hematology

Chapter 3: Genotyping: creating a genome-wide dataset of common genetic variants

Figures 69

Tables 69

3.1. Introduction 70
3.1.1. Selecting “tag” markers and genome-wide genotyping arrays 70
3.1.2. Common disease, common variant hypothesis: basis of epidemiological genetic studies, including GWAS 71
3.1.3. Towards a “MEGA” sickle genetic dataset: a massive database of quality controlled, imputed genotypes 73

3.2. Materials 73
3.2.1. Illumina Infinium MEGA chip 73
3.2.2. Data 73
3.2.2.1. Genome wide variant data 73
3.2.2.2. Reference panel: 1000 genomes project 73
3.2.3. Computational requirements: hardware and software 74
3.2.3.1. Hardware 74
3.2.3.2. Software 75

3.3. Methods 75
3.3.1. Study subjects 75
3.3.2. Sample processing pre-MEGA 75
3.3.3. Variant calling 76
3.3.4. Workflow: statistics-based quality control and imputation 76
3.3.5. Data quality control 77
3.3.5.1. Background 77
3.3.5.2. Basic quality control 79
3.3.5.3. Linkage disequilibrium- (LD-) pruned set of genotypes 81
3.3.5.4. Sex discordancy 81
3.3.5.5. Genetic Relatedness Matrix: GRM 82
3.3.5.6. Duplicates 83
3.3.6. Imputation 83
3.3.6.1. Background 83
3.3.6.2. Reference panels 85
3.3.6.3. Imputation accuracy 85
3.3.6.4. Process of imputation 86
3.3.6.5. Pre-imputation quality control 87
3.3.6.6. Strand alignment 87
3.3.6.7. Phasing 88
3.3.6.8. Post-phasing QC 89
3.3.6.9. Imputation on the online Michigan imputation server 89
3.3.6.10. Post imputation QC 90
3.3.6.11. Merge imputation dataset with raw genotypes 90
3.3.6.12. Post-merge QC 93

3.4. Results 93

3.4.1. Quality control and imputation pipeline 93
3.4.1.1. Four scripts to manage the quality control and imputation pipeline 93
3.4.1.2. run_QC.sh 93
3.4.1.3. PreMichiganProcessing_SHAPEIT.sh 94
3.4.1.4. PostMichiganProcessing_SHAPEIT.sh 94
3.4.1.5. PostImputationAnalysis.sh 94
3.4.2. Results for my data 94
3.5. Discussion 95
References 96
Appendix 1 99
Appendix 2 100
Appendix 3 101
Appendix 4 102
Appendix 5 105
Appendix 6 109
Appendix 7 120
Appendix 8 125

Figures

Figure 1 LD blocks and disease-causing genetic variants (D) versus indirect markers (M) 70
Figure 2 Direct and indirect association testing, adapted from (Hirschhorn and Daly, 2005): .. 70
Figure 4 Common versus rare variants, from (McCarthy et al., 2008) 72
Figure 5 allele signal intensity plots, from (Weale, 2010) 76
Figure 6 workflow 77
Figure 7 Imputation process, from (Marchini and Howie, 2010) 84
Figure 8 benefit of imputed genotype calls (right panel) to overcome fuzzy raw genotype calls in pale blue (left panel), from (Marchini et al., 2007) 85

Tables

Table 1 Imputation quality by chromosome 92

3.1. Introduction

3.1.1. Selecting "tag" markers and genome-wide genotyping arrays

Genetic association studies test whether different alleles of a gene are associated with a trait: in case-control studies, whether one allele is more frequent among cases compared to controls, and in quantitative trait studies, whether trait values are higher among carriers of one allele than another. Identification of a locus associated with a trait or disease is then followed by high resolution mapping to identify the causal variant. This relies on the concept of linkage disequilibrium (LD), where the mapped variant acts as a "tag" for the causal variant as they are both inherited as a block due to the close physical proximity. Genetic association studies are motivated by this concept of LD between alleles at two loci: carrying a specific allele at the first locus gives information on the allele carried at the second. A significant result for testing association between disease and marker implies the association is either (Figure 1):

- **Direct**: the variant allele directly affects disease risk (D)
- **Indirect**: the tested marker (M) is in linkage disequilibrium with the causal disease mutation – tends to occur on the same ancestral chromosome: a 'tag' variant
- **Spurious**: due to confounding or chance

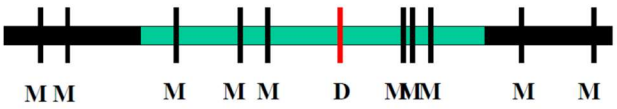


Figure 1 LD blocks and disease-causing genetic variants (D) versus indirect markers (M)

An ideal whole-genome assay system would genotype *every* polymorphic site in the genome (i.e. sequencing), but this is still too expensive for large samples. Instead, one can select variants (or "tag markers") to use, based on LD information, in the hope that causal variants will be in LD with some of them. The first successful genome-wide association study (GWAS) used less than 100,000 markers and identified one variant for myocardial infarction(Ozaki et al., 2002). An association between a tag variant and disease implies the existence of LD between the marker and the causal locus i.e. that the causal variant is near the marker, see Figure 2.

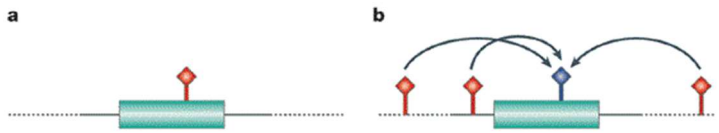


Figure 2 Direct and indirect association testing, adapted from (Hirschhorn and Daly, 2005):

(a) candidate variant (red) is directly tested for association with a disease phenotype. (b) the variants to be genotyped (red) are chosen on the basis of linkage disequilibrium to provide information about as many variants as possible. In this case, the blue variant is tested for association indirectly, as it is in LD with the other three (red) variants.

When creating a micro-array, then, the aim is to select the *minimum* number of tag variants that are in sufficient LD with the *maximum* number of unselected variants. Genome-wide genotyping projects (HapMap project, 1000 Genomes Project) were developed with precisely the aim to create a reference set of LD data on the "whole" genome, in order to identify not just polymorphic sites, but haplotype structures (LD blocks) in different representative populations. These public data can be used as "reference panels" for a variety of bioinformatics purposes. These projects showed that LD can persist between loci kilobases apart, but that this is dependent upon the age of the population; there is more LD in European and Asian populations compared to African populations. Therefore, good coverage of the human genome can be achieved with reasonably few tag variants but African populations will require more tag variants to achieve good coverage of the genome. As a result, commercial companies have developed generic "chips" or "micro-arrays" for genome-wide analysis. The market is now dominated by two companies: Illumina (tag variants are chosen for reliable genotyping) and Affymetrix ("random" variants spread more evenly across the genome). The first wave of genome-wide association studies (2005-7) used 100,000 to 500,000 markers. Chips then started using over 1,000,000 variants (equivalent to one variant per <3kb).

More recently, the advent of imputation to infer non-genotyped variants based on densely genotyped reference panels, means many millions of variants can now be used in genome-wide analyses despite the lower density of markers on chips. Imputation is a method for estimating unobserved genotypes; by combining study data with a publicly available reference panel, unobserved variants values are predicted (imputed) based on haplotype data in the reference panel. This is based on linkage disequilibrium information from the reference panel (data such as the *1000 Genomes Project*).

3.1.2. Common disease, common variant hypothesis: basis of epidemiological genetic studies, including GWAS

The "common disease, common variant" hypothesis states that common diseases, present in all major human populations, are caused by universally common alleles. In contrast, rare variants cannot be distinguished by epidemiological data. This concept drives the success of genome wide association studies (GWAS) which have identified common variants contributing to the inherited component of common diseases. In contrast, rare variants (Mendelian diseases) have historically been identified through family studies; and now are more amenable to whole genome or whole exome sequencing approaches(McCarthy et al., 2008). See Figure 3 for common versus rare variants, and how the allele frequency and penetrance dictates study approaches.

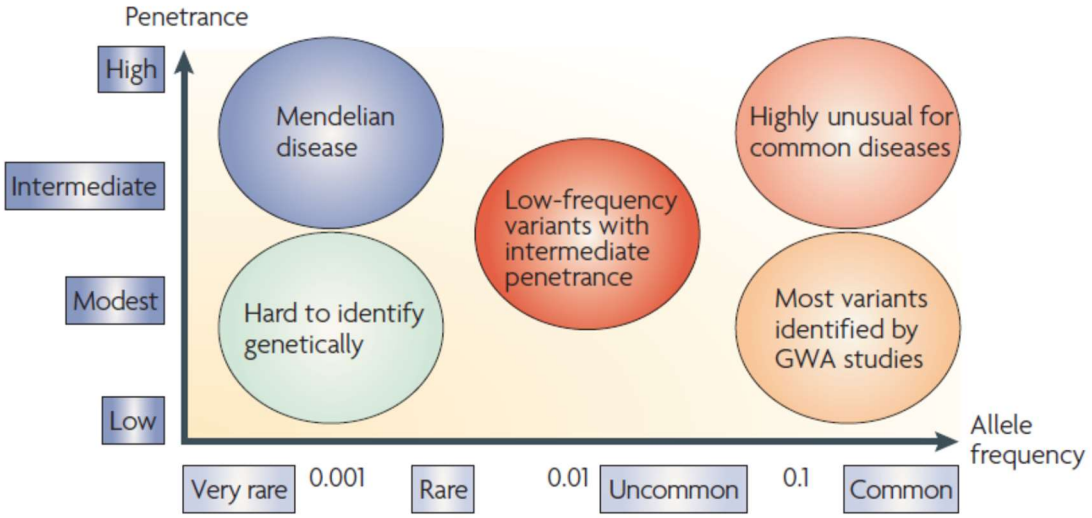


Figure 3 Common versus rare variants, from (McCarthy et al., 2008)

However, there remains a disparity between the observed heritability for many common diseases and the relatively small amount of increased risk currently attributable to known disease-relevant variants. This is often referred to as “missing heritability” which is explained by common and uncommon variants, at both already identified loci and unknown other loci. Previously, it was broadly accepted that rare variants played a major role in this deficit, but more recent analysis in the diabetes setting suggested that most associated variants were overwhelmingly common, and mostly at loci already identified(Fuchsberger et al., 2016). The advent of larger, multi-ethnic reference panels, in the era of imputation, allows larger-scale association testing to better characterise known loci.

Consideration of the population studied is also important. Since LD extends further in younger populations, there is longer LD in Europeans than Africans. On a given genotyping chip, a causal variant is likely to be better tagged in a European population than in an African population. As a result, for a chip with N markers, studies of African populations have less power to detect associated alleles than those performed on younger populations. Imputation (using densely typed and ethnically diverse reference panels) circumvents this to some extent with the resultant denser genotyping capturing smaller LD blocks. The reduced LD blocks in African populations means “hits” are nearer the true causal variant. Now that research into African-heritage populations is better served by both ethnic specific micro-arrays and population-matched imputation methods, it could be argued that African-heritage populations are the best research cohorts precisely because of their small LD blocks.

3.1.3. Towards a "MEGA" sickle genetic dataset: a massive database of quality controlled, imputed genotypes

I have created a "database" of genotypes in our sickle cohort for use in statistical analysis. This has required: good quality DNA, an appropriate micro-array (the "MEGA" chip), and bioinformatic expertise in interpreting and using micro-array results: variant-calling, data quality control, and imputation. The final, imputed dataset contains many millions of variants ready for use in association testing.

3.2. Materials

3.2.1. Illumina Infinium MEGA chip

Illumina's Infinium "MEGA" chip (Multi Ethnic Genotyping Array, Illumina, San Diego, CA, USA, see Appendix 1). The MEGA chip was in beta-testing phase, and comprised about 1.7m markers. For the set of African-heritage population specific markers, it leveraged genotyping content from the Consortium on Asthma among African-Ancestry Populations (CAAPA, https://www.caapa-project.org/). The chip includes content from historical Illumina commercial arrays, plus African diaspora content identified through the sequencing of 700 individuals by the CAAPA.

The new (expanded) MEGA array is now commercially available from Illumina at: https://www.illumina.com/science/consortia/human-consortia/multi-ethnic-genotyping-consortium.html.

3.2.2. Data

3.2.2.1. Genome wide variant data

I received the MEGA chip (genome-wide) variant data (post-variant calling) in PLINK format.

3.2.2.2. Reference panel: 1000 genomes project

The 1000 Genomes Project (1000G, http://www.internationalgenome.org/data/) developed a large public database of human variation and genotype data. The goal of the 1000 Genomes Project was to identify genetic variants with frequencies ≥1% in populations studied. The project ran between 2008 and 2015. It was the first project to sequence the genomes of a large population. Data were made available freely through public databases.

Cost limited the sequencing depth. However, since any genomic region contains a limited number of haplotypes, data was combined across samples to allow efficient detection of most of the variants in a region. The project planned to sequence each sample to 4x genome coverage; at this depth, sequencing allows the detection of most variants with frequencies ≥1%. The multi-sample approach combined with genotype imputation allowed the project to determine a sample's genotype, even in variants not covered by sequencing reads in that sample.

The final dataset (1000G phase 3) contains data for 2,504 individuals from 26 populations; phase 3 analyses were published at the project conclusion in 2015 (Sudmant et al., 2015, Auton et al., 2015). 661 of these individuals were from seven African (AFR) populations:

- YRI: Yoruba in Ibadan, Nigeria
- LWK: Luhya in Webuye, Kenya
- GWD: Gambian in Western Divisions in the Gambia
- MSL: Mende in Sierra Leone
- ESN: Esan in Nigeria
- ASW: Americans of African Ancestry in South-Western USA
- ACB: African Caribbean in Barbados

These seven groups reflect populations in West and East Africa, as well as admixed populations in the Caribbean and North America.

After the completion of the 1000G project, the international genome sample resource (IGSR) has been setup to provide ongoing support for the 1000G data. It aims to ensure the future accessibility of the data as well as to extend the dataset (both new data on existing samples *and* new populations).

Since my work, the Haplotype Reference Consortium (HRC, http://www.haplotype-reference-consortium.org/) is beginning to supersede the 1000G data as the gold standard in reference genomes. The HRC project aims to create a large reference panel of human haplotypes by combining data from multiple cohorts.

3.2.3. Computational requirements: hardware and software

Computational requirements for whole genome data are significant and include adequate hardware and software.

3.2.3.1. Hardware

Imputing variants in datasets of hundreds to thousands of samples using reference sets with millions of variants (e.g. 1000 Genomes Project), up to several tens of millions of variants cannot be done on a desktop computer, as that would take months/years and would require more memory (RAM) than is available. Instead, the imputation process is divided into smaller chunks and these sub-tasks are then run on a computer cluster. The work described in this PhD thesis was done on such a cluster.

I made use of King's College London's (KCL) server cluster ("super computer") *Rosalind* which runs a Linux platform (http://rosalind.kcl.ac.uk/). *Rosalind* is a heterogeneous cluster that consists of 4 machines with 10 cores each, each with 192-384 GB of RAM, running Scientific

Linux release 6.6 (https://www.scientificlinux.org/). *Rosalind* uses the Open Grid Scheduler (Grid Engine), a batch-queuing system to schedule tasks across the nodes to allow for distributed resource management (http://gridscheduler.sourceforge.net/). Producing the genetic relatedness matrix in GCTA is computationally the most expensive step in my workflow: it requires consideration of NxNxM (N study individuals, M number of pruned variants) – around 80G RAM for my project.

3.2.3.2. Software

I wrote my own bash scripts to manipulate the data and made use of multiple open-source Linux-platform based genetic software:

- Unix/Linux text manipulation commands (e.g., awk, grep, head, tail, sort, join, uniq) for manipulating text files
- R statistical package for additional analysis and graphics (www.r-project.org) (Team, 2011).
- PLINK v1.90b3.38(Purcell et al., 2007) for genetic data management, data quality control, summary statistics, see http://zzz.bwh.harvard.edu/plink/index.shtml.
- SHAPEIT v2.r837 (Delaneau et al., 2011) for phasing, see https://mathgen.stats.ox.ac.uk/genetics_software/SHAPEIT/SHAPEIT.v2.r790.RHELS_5.4.static.tar.gz
- GCTA v1.26.0(Yang et al., 2011) for statistical genetics
- Michigan imputation online server (Das et al., 2016): https://imputationserver.sph.umich.edu/index.html. The Michigan Imputation Server is free genotype imputation service using minimac3 and is a so-called "next generation" imputation server because it retains imputation accuracy with computational efficiency.

3.3. Methods

3.3.1. Study subjects

891 samples from patients with sickle cell disease (516 female, 375 male), as described in chapter 2.2.1, were utilised. These patients comprised:

- 666 HbSS (375 female, 291 male)
- 195 HbSC (126 female, 69 male)
- 21 HbSβ^+ thalassaemia (10 female, 11 male)
- 9 HbSβ^0 thalassaemia (4 female, 5 male)

3.3.2. Sample processing pre-MEGA

DNA had previously been extracted from peripheral blood white cells and was held at 100ng/µl concentrations in the sickle gene bank. All useable samples from the gene bank were prepared for MEGA processing: ten 96-well plates were filled with at least 1µg genomic DNA on each sample (mostly 15µl of 100ng/µl).

Samples were genotyped at the Institute of Psychiatry, Psychology & Neuroscience BRC Genomics Facility on Illumina's Infinium MEGA chip (Multi Ethnic Genotyping Array), see section 3.2.1. Demographic information (including clinical sex and relationship information) was also supplied to the BRC Genomics Facility.

3.3.3. Variant calling

Variant-calling was not done by me: it was performed externally by Hamel Patel/Stephen Newhouse (at the Social, Genetic and Developmental Centre, Institute of Psychiatry, King's College London) using GenomeStudio software (Illumina, San Diego, CA, USA) and a standard operating protocol.

In summary, the genotyping assays for any given variant produce a quantitative signal intensity for each of the two possible alleles (A, B). A single marker can then be represented as a scatter-plot of signal intensities for allele A versus B (each point represents a different individual). If the process works, individuals can be separated into three distinct clusters representing the three possible genotypes AA, AB, BB, with AB in the middle, see Figure 4, panel (a): the coloured clusters represent individuals with AA (*blue*), AB (*green*) and BB (*red*) genotypes with genotype clusters well separated while panel (b) a problem marker where the AA and AB genotype clusters overlap. Individuals in the grey zone are impossible to call and so are labelled as missing. The ultimate assessment of genotype quality is manual inspection of cluster plots; these should be inspected after association testing for any positive variants.

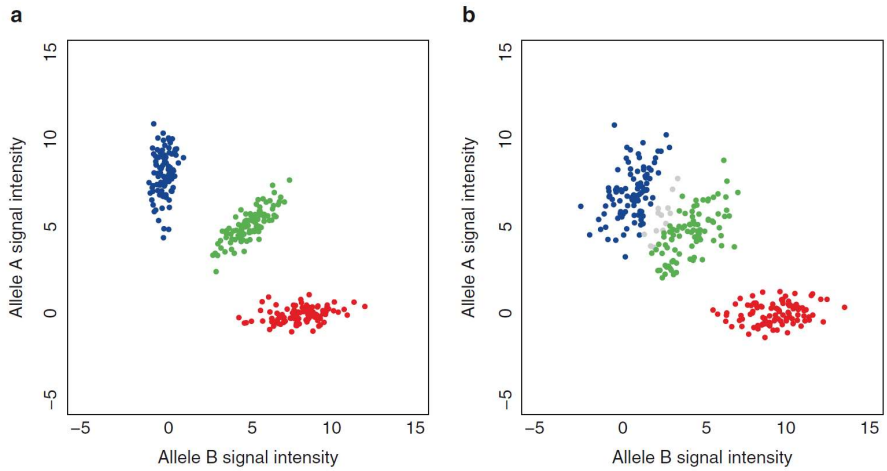


Figure 4 allele signal intensity plots, from (Weale, 2010)

3.3.4. Workflow: statistics-based quality control and imputation

I received the data post-variant calling. I undertook quality control and imputation based on the workflow in Figure 5.

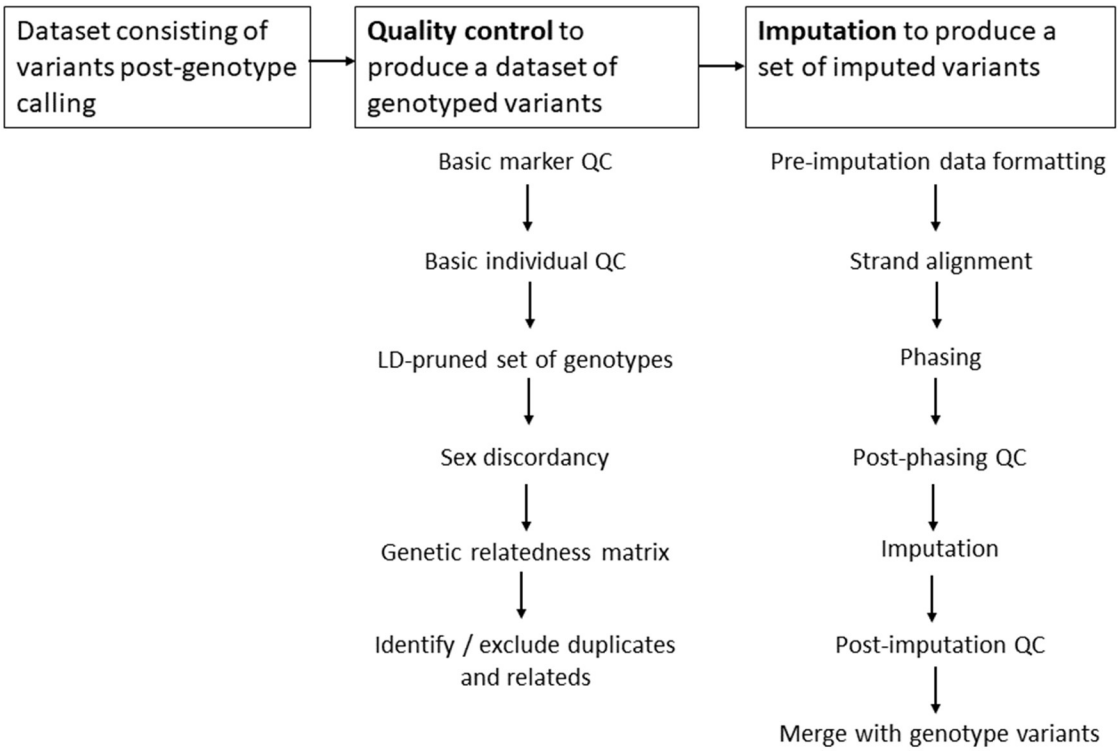


Figure 5 workflow

3.3.5. Data quality control

3.3.5.1. Background

Errors in genotypes or DNA sample identification ("sample mix-ups") lead to errors in association studies. If these errors are systematic (e.g. in a case control study, cases and controls are genotyped separately), then this can introduce systematic bias leading to an increase in false positive findings and a decrease in power. Current genotyping panels have impressive genotyping quality (reports of accuracy ≥99.9%), partly because variants have been replaced if they work poorly by better markers in linkage disequilibrium. However, variant call quality is rarely as high in practice – the sample DNA is usually less carefully prepared. The large number of markers now tested in genome-wide association studies means even when errors are infrequent, they can be detrimental. For example, if 1,000,000 markers are tested for association and 1/1,000 markers are poorly genotyped, and the inaccurate calling results in detection of a spurious association, then there may be up to 1,000 false-positive results.

Potential false positive findings need to be managed by correcting for bias (often deleting samples and/or markers from the analysis). False negative findings, on the other hand, must also be avoided: while deleting samples and markers that are error prone, good markers or samples must not be deleted (these may contribute to true positive findings). False positive errors are easier to manage than false negative results, and as it currently stands, most data quality control is about controlling positive false positive than false negative results.

Quality control (QC) measures are therefore imperative to remove samples and variants prior to association analysis. QC involves the identification and removal of DNA samples and markers that introduce bias. These QC steps are necessary before statistical association testing and critical for a successful study. However, the criteria used to filter out low quality markers is a balance: care must be taken to delete poor quality markers only because every removed marker is potentially a missed disease variant.

Principles of QC were taken from previously published protocols (Weale, 2010, Anderson et al., 2010), and guidance from UK Biobank cut-offs (based on more recent genotyping array quality) at http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf. These were not absolute and instead each QC step was considered in the context of the study data, rather than an automatic application of arbitrary QC thresholds. Markers were identified and removed using UK Biobank guidelines (BioBank, 2015).

Quality control includes assessment of both individuals (samples) and markers (variants): issues relating to both sample and variant problems must be identified and removed. In a large study population, the impact of removing one marker is potentially greater than the removal of one individual and as a result *sample QC* before *marker QC* is advocated (Anderson et al., 2010). However, in a smaller study population like ours, implementation of marker QC prior to sample QC optimises sample number over variant number. This approach prevents individuals being removed due to a subset of poorly genotyped markers, but is susceptible to markers being falsely removed on the basis of a poorly genotyped subset of individuals. Genotype imputation recovers markers potentially lost in this process(Marchini et al., 2007).

Variants were removed based on sample quality - low call rate, minor allele frequency (poor quality marker), deviation from Hardy-Weinberg equilibrium (genotyping error), study-specific variant QC filters (such as differences in allele frequencies between multiple control cohorts). Low frequency variants are more prone to bias due to genotyping error and low power to detect association. Therefore, these are often removed.

Samples were removed based on DNA quality, sample identity (e.g. phenotypic sex does not match genotypic sex), and patient relatedness (duplicates). This is based on: low call rate (poor DNA quality), outlying heterozygosity across autosomes (DNA sample contamination or consanguinity), duplication (or relatedness) based on calculated genetic relatedness, mismatches with external information (sample mix-up). Historically, samples were also

removed in large studies based on “near-relatedness” (cryptic or known) or “far-relatedness” (issues of population structure); in the latter case, sub-analyses based on one ethnic group only was conducted. The issues of relatedness can now be addressed using a “genetic relatedness matrix” which itself can identify duplicates and near (cryptic) relateds, and subsequently be used as a covariate in a mixed linear statistical model in order to manage both near and far relatedness. Duplicate samples are identified and removed as part of this process.

I used PLINK (v1.90b3.38), a tool for handling genetic data, to perform assessments of failure rate per individual and per marker, and GCTA (v 1.26.0) to assess the degree of relatedness between individuals (both near relatedness – close relationships – and far – issues of population stratification). Both these software platforms are widely-used and computationally efficient.

3.3.5.2. Basic quality control

Quality control measures for pre-imputation steps generally apply stricter thresholds than in the non-imputation setting. Broadly speaking, stricter marker QC measures improves imputation quality (Marchini and Howie, 2010) but imputation accuracy is a balance between study variant density and variant quality (see section 3.3.6.3).

Genetics variants (i.e. markers) were identified and removed if:

- **Genotyping rate <95%.** Missingness (missing calls) can be a big problem; generally, missingness is not randomly distributed among genotypes but is overrepresented in some. There is a correlation between missing genotype rate and variant quality. Missingness can generate both false positive and false negative signals of association. Classically, variants with a call rate of less than 95% have been removed (Silverberg et al., 2009, Pompanon et al., 2005).
- **Minor allele frequency (MAF) <0.1%.** This is a useful step before imputation to remove monomorphic and very rare variants. These variants can be reintroduced later (which is particularly useful if comparing our cohort to other populations. Data quality tends to decrease with decreasing MAF (a low MAF is synonymous with rare genotypes which can be problematic at the variant calling stage). Missingness can affect low-MAF variants more strongly and thereby increase the chances of a false positive signal. Furthermore, the power to detect an association signal decreases with decreasing MAF. Finally, given that rare variants virtually never detect an association signal and that their inclusion in the study increases the overall number of tests performed, including these variants decreases the power to detect signals in *other* variants (Morris and Zeggini, 2010).

- **Genotypes were out of Hardy-Weinberg equilibrium (HWE), with a threshold of *p*<5x10-7.** Genotypes for a marker are considered in HWE if their frequencies do not deviate significantly from those predicted based on the frequency of its alleles. So, given a MAF of q, the probabilities of the three possible genotypes (aa, Aa, AA) at a bi-allelic locus are ((1−q)2, 2q(1−q), q2). In a large, randomly mating, homogenous population these probabilities should be satisfied in each generation. Significant deviation from HWE can be due to consanguinity, population structure, and non-random selection of study subjects. If deviation from HWE occurs with individual markers, it can indicate a genotyping error or calling error. It may also occur at a disease locus where it would clearly be counter-productive to remove these data. Recommended significance thresholds for declaring variants to be in HWE vary greatly because of this balance between removing potentially erroneous variants while retaining valid variants. Suggested p-value thresholds suggested have ranged from 0.001 to 5.7 × 10−7)(Anderson et al., 2009, Wittke-Thompson et al., 2005) with lower values in more recent studies e.g. the UK BioBank applied a threshold of p<1x10-7.

Individuals (samples) were identified and removed if:

- Genotyping rate <99%. There are large variations in DNA sample quality and these can have large effects on genotype call rate and genotype accuracy. False positives results can arise if DNA quality differs with phenotype, leading to differences in the frequency of called genotypes. This presents a particular issue for case-control studies if cases are collected and/or processed separately to controls, and there is a systematic difference in genotyping between the two groups. In our cohort, the requirement for an individual's genotyping rate to be >99% did not delete any samples.
- **Outlying heterozygosity rates.** For a given individual, their heterozygosity rate is the *proportion of heterozygous genotypes excluding those of the sex chromosomes*. Individuals with anomalously high heterozygosity can indicate sample contamination, and with anomalously low heterozygosity can indicate membership of a different population or inbreeding. After extracting the autosomes for analysis only, there were no issues with excess heterozygosity in my population. [If there had been problems, there are tools available to remove extreme outliers e.g. https://github.com/JoniColeman/gwas_scripts/blob/master/IdHets.R.]
- **Discordant sex**: phenotypic sex is inconsistent with the genetic sex (as measured by X-chromosome homozygosity). Discordant sex was identified and managed after construction of an LD-pruned dataset: see sections 3.3.5.3 and 3.3.5.4.

- **Relatedness: duplicated individuals, closely related individuals, issues of population structure.** Relatedness was identified after creation of a *genetic relatedness matrix*: see sections 3.3.5.5 and 3.3.5.6.

3.3.5.3. Linkage disequilibrium- (LD-) pruned set of genotypes

A "pruned" genotype dataset is required for multiple downstream processing steps including sex checking (section 3.3.5.4), and constructing a genetic relatedness matrix (section 3.3.5.5). The aim is to remove variants to get one variant per linkage disequilibrium (LD) block, producing the "LD-pruned dataset". Thinning to a set of 50,000-100,000 variants has been advocated consistently, and appears adequate for identification of cryptic relatedness and population outliers (isolated individuals from different ethnic populations to the main population) (Weale, 2010). I aimed for a pruned dataset at the upper limit of this range given the African origins of our population (with smaller LD blocks).

In PLINK, markers were compared pairwise in windows of 1000 kb markers, and one of each pair removed if r^2 (multiple correlation coefficient for a marker being regressed on all other markers simultaneously) > 0.04, and the procedure repeated after an interval shift of 100 markers. This resulted in a pruned dataset with 98,299 variants.

3.3.5.4. Sex discordancy

Samples were excluded if phenotypic gender was inconsistent with genotypic sex, as measured by X-chromosome homozygosity. X chromosome homozygosity was calculated from genotypes: male samples are expected to be homozygous for X chromosome variants, and females heterozygous. Thus, males have a homozygosity rate of 1 and females 0 (though there is some variation due to genotyping error). Comparing the calculated homozygosity rate across all X-chromosome variants for each individual to the expected rate for phenotypic sex, individuals with discrepancies were detected. This is an important administrative check that the DNA sample has not been mislabelled and that the wrong phenotype data has been associated with the genotypes. The discrepant samples can then be removed.

PLINK implements this by calculating the inbreeding *F statistic* which measures the severity of departure from HWE on the X chromosome: females have X-variants broadly in HWE, whereas males will depart severely from this (no heterozygous genotypes). For genetic females, F is ~0 and for genetic males, F is ~1. There are, however, some individuals who are intermediate between these two groups – this may be evidence of sample contamination, membership of a different population, or X chromosome mosaicism in females. The most pragmatic solution is to exclude all intermediate-*F statistic* individuals.

This procedure must be performed after dataset pruning. Final cut-offs can depend upon specific study data. This involved checking different cut-offs for the data – the final definitions of sex cut-offs were:

- F statistic: female maximum 0.3, male minimum 0.9

No samples were outside these thresholds, but 12 samples had their *genotypic* sex discordant with *phenotypic* sex and these were removed from both the QC'd original dataset and the pruned dataset. It should be noted that this procedure will flag sample mix-ups only where the gender is different. All male-male and female-female mix-ups occur at the same rate but will remain undetected.

3.3.5.5. Genetic Relatedness Matrix: GRM

A genetic relatedness matrix (GRM) is a NxN correlation matrix which is a quantification of relatedness of all possible pairings between individuals in the study group (sample size N). GRMs are a relatively new concept and developed to be a finer-scale method of detecting relatedness and population structure compared to principal components analysis.

Construction of a GRM occurs in two steps. First, individuals are scored at each variant according to how different the number of reference alleles is from the cohort average, and weighted by the variant's heterozygosity (Yang et al., 2010, Yang et al., 2011, Kang et al., 2010). These variant scores are summed to give a total score for a given individual. Second, for each pairwise relationship within the cohort, these scores are then compared to assess genetic similarity between each pairwise relationship, thus constructing an NxN correlation matrix of all pairwise relationships.

Two identical samples have GRM correlation value of around 1 (values comprise a normal distribution around 1 so are not exact). First degree relationships (parent/child) would have a value of 0.5, and second degree 0.25.

The GRM is useful in and of itself. It identifies sample duplicates (with a cut-off >0.9). These may represent monozygotic twins, sample errors, or sample duplications which need to be removed. Quantification of the relationships allows us to (a) exclude one of a pair of close relatives prior to analysis by having a "cut-off" GRM and (b) to use this quantitative relatedness data as part of statistical modelling.

In our study, recruitment did not include documentation of family history and therefore, given the nature of an autosomal recessive condition in an ethnic minority community distributed densely in SE London, cryptic relatedness (e.g. cousins) may be a significant issue.

I created a GRM using GCTA(Yang et al., 2011) using pruned, un-imputed variants (see section 3.3.5.3).

The GRM cannot be visualised directly, but using R packages *devtools* and *OmicKriging*, one can create a readable format of the GRM. This can then be viewed manually in csv format. In chapter 4, as part of GWAS analysis, the GRM is fitted into a mixed linear model as a fixed effect to estimate the population variance explained by the genome-wide markers (as an improvement on principal components).

3.3.5.6. Duplicates

Identification and deletion of duplicate samples is paramount to prevent overcalling of genetic information based on these duplicated individuals. The GRM revealed 9 pairs with GRM>0.9. I returned to the clinical data to assess these duplicates and consider which sample to delete. For one pair of monozygotic twins, one individual in the pair was deleted (the individual with less clinical data available). For each of five duplicate-pairs that were identified clinically as the same individual, but from different study sites, one of each pair was deleted (the individual with less clinical data available). Finally, for three pairs which appeared to be different people, it was assumed to be a sample mix-up and all six individuals were deleted.

3.3.6. Imputation

3.3.6.1. Background

Genotype imputation is a method for estimating (*imputing*) unobserved genotypes i.e. genotypes that are not directly genotyped in a sample. Imputation is performed with a statistical algorithm extracting LD and haplotype information from a *reference panel* (e.g. 1000 Genomes Project) of individuals who have been densely genotyped and applying it to a *study sample* of less densely genotyped individuals. Imputation will increase the number of variants that can be tested in association analyses up to the size of the reference panel. This may include markers with (1) missing genotypes at typed variant sites (2) genotypes at un-typed variant sites that are present in an external high-density reference panel. Imputation provides many benefits. Imputed datasets increase study power: the reference panel is statistically more likely to contain the causal variant than the original dataset. Imputation can also fine-map any association signal: it provides a high-resolution view of an association signal across a locus. Finally, imputation facilitates meta-analysis: it can bridge the gap between different

platforms so that different studies genotyped with different arrays can be combined up to variants in the reference panel.

The process of imputation is shown in Figure 6. The raw dataset consists of a set of genotypes including many unobserved variants, as in panel (a): association testing at just these variants may not lead to a significant association (b). Imputation predicts these missing genotypes. Modern pipelines first strand-align and phase each individual at the typed variants, and then perform imputation. Panel (c) demonstrates three phased individuals. These haplotypes are then compared to the reference panel which contains dense haplotypes (d). The haplotype of a given individual is modelled as a mosaic of (the limited number of) haplotypes of other individuals. Missing genotypes in the study sample are then predicted (imputed) using those matching haplotypes in the reference panel (e). Statistically, the uncertainty with which genotypes are imputed can be modelled with a probability distribution over all three possible genotypes. This uncertainty (imputation quality) information must be taken into account in any downstream analysis of the imputed data. Testing the imputed dataset can lead to more significant associations (f) with a detailed view of the associated locus.

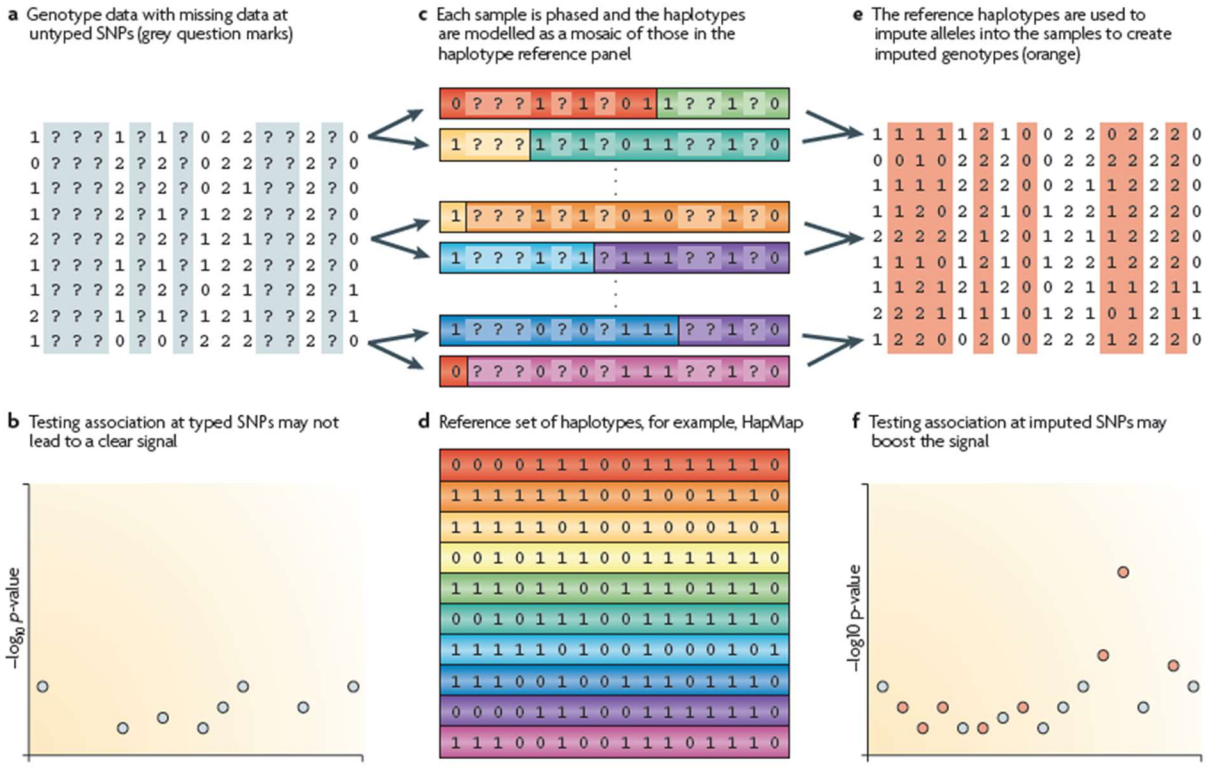


Figure 6 Imputation process, from (Marchini and Howie, 2010)

In this way, imputation methods attempt to identify sharing between the underlying haplotypes of the study individuals and the haplotypes in the reference set, and use this sharing to impute the missing alleles in study individuals. The methods used in imputation are strongly connected to those used to infer haplotype phase (Excoffier and Slatkin, 1995,

Stephens et al., 2001), tagging marker-based approaches (Johnson et al., 2001) and methods in linkage studies (Elston and Stewart, 1971).

Imputation can add variants back that were previously removed at the variant calling stage due to the issue of "fuzzy calls" in raw genotyping (section 3.3.3, also see Figure 7).

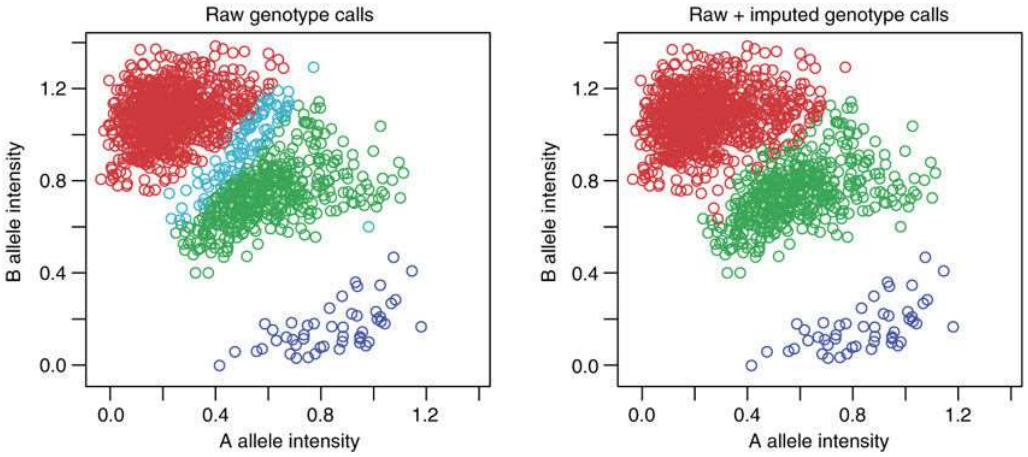


Figure 7 benefit of imputed genotype calls (right panel) to overcome fuzzy raw genotype calls in pale blue (left panel), from (Marchini et al., 2007)

3.3.6.2. Reference panels

The first large-scale reference panels were made publicly available through the 1000 Genomes Project, see section 3.2.2.2. This is being superseded by the HRC but at the time of my project, the 1000 Genomes Project dataset comprised the largest dataset (by population, ethnic diversity and number of variants). Using a large, multi-ethnic reference panel improves imputation accuracy for rare variants. This can come at the expense of computational speed which, however, is less of an issue with the advent of "next generation" online imputation servers.

3.3.6.3. Imputation accuracy

Many factors affect imputation accuracy(Marchini and Howie, 2010). The first determinant is the study population itself with both the number of individuals and ethnicity having an effect. Presently, imputation accuracy declines across the world from: Europe→ Asia→ America→ Oceania→ Middle East→ Africa(Huang et al., 2009). This situation is improving slowly due to modern attempts at reflecting worldwide genetic diversity, with strategies such as new chips (like the MEGA chip) which reflect polymorphisms in non-European populations, as well as more diverse reference populations (including the 1000 Genomes Project). However, imputation will remain less accurate for African populations rather than "newer" populations like European populations: African populations are "old", and the many generations that have passed under stable conditions have provided the recombination events to break down LD, resulting in them being genetically heterogeneous, with shorter linkage disequilibrium (LD)

blocks. This makes the imputation process less robust, as the mosaic blocks are shorter and more varied so that prediction of genotypes is less confident.

The second factor influencing imputation accuracy is the chip used for genotyping i.e. how well it covers variation in study population. The ethnic-specificity of the new MEGA chip reflects present efforts of the genetics community to capture polymorphisms in all human populations.

The third determinant of imputation accuracy is the reference panel used. In the reference panel, the number of individuals (better accuracy with larger panels), density of typing (denser is better as there will be more overlap with chip markers) and the issue of matching ancestry, all influence imputation. Historically, same-ethnicity reference panels were used (possibly because phasing was undertaken simultaneously with imputation), but, presently, using as many diverse haplotypes as possible is advocated, to capture as many uncommon and rare variants as possible.

The fourth factor determining imputation accuracy is good data quality control. Notably, error rates in imputation are higher as the minor allele frequency decreases (Marchini and Howie, 2010).

The reality of imputation is that all thresholds used in pre-imputation steps (i.e. quality control and phasing) are study-specific. This means an iterative process of trial imputation/re-imputation cycles to identify appropriate criteria is used.

3.3.6.4. Process of imputation

The practical process of imputation is summarised in Figure 5. After pre-imputation QC, strand alignment is performed using PLINK to ensure all genotypes are expressed relative to the same (positive) strand. Phasing is then undertaken in SHAPEIT (v2.r837) in order to align genotypes to the correct haplotype. After phasing, further QC is performed. Imputation follows, on the online Michigan imputation server(Das et al., 2016), and then post-imputation QC is carried out before the imputed genotypes are merged with raw genotype data.

Throughout this process, it is imperative to maintain the same nomenclature between study and reference panel data. In particular, the same genome build must be used(GRCh37/hg19) and the same variant names (I converted all study and reference panel variant names to format “ChromosomeNumber_PositionNumber”).

3.3.6.5. Pre-imputation quality control

There are strict data format requirements for phasing and imputation software, so I undertook data pre-processing steps. This included identification and deletion of: variants with non-ACGT format (e.g. “B”, “I/D”), duplicate variant names and duplicate variant positions. Non-bi-allelic markers were also removed. The files were then processed by single chromosomes for computational efficiency.

3.3.6.6. Strand alignment

The process of strand alignment expresses all alleles relative to the same strand on a chromosome. Genotyping technologies and variant calling algorithms can be expressed relative to either the positive (forward) or negative (backward) strand on a chromosome e.g. a variant may have alleles T and C situated within the coding strand but complementary alleles A and G to the template strand. During imputation, it is imperative that each variant on a chromosome has its genotypes expressed relative to the same strand in both the reference and study populations. All the 1000 Genomes Panels have their variant alleles expressed relative to the forward strand. Thus, the cohort data must be strand-aligned relative to the forward strand. It is also crucial that marker positions for both datasets use the same co-ordinate system (i.e. the same build).

Historically, strand alignment (as well as phasing) was done *during* imputation using a single reference panel. For many imputation servers, including the Michigan server, this process has evolved and is now performed in three steps: strand alignment, phasing and imputation steps. This not only speeds up the imputation process (Howie et al., 2012) but it also separates out the need for population matching of the reference panel at the phasing stage (i.e. using only match-populations), versus the benefits of maximal genetic diversity at the imputation stage (i.e. a reference panel with all ethnic populations). For the Michigan server, it is a pre-requisite that strand alignment is performed in pre-phasing.

After downloading the reference panel (1000G phase 3 in hg19 build), duplicates of variant name / position were removed, and markers were renamed in format "ChromosomeNumber_PositionNumber". Strand-ambiguous variants (A/T and C/G) were removed.
(C/G and A/T variants are *ambiguous* or *cryptic* as their complementary alleles are G/C and T/A, respectively. The ambiguity means it is much more difficult to detect and resolve strand alignment issues for these variants).

I assessed whether my data was strand-aligned indirectly. I used PLINK's *merge* function by merging the cohort data with the strand aligned reference panel. By using a "dummy merge" of the two datasets in PLINK, merge errors were identified. The merge errors can have a trial "flip" to flip the errors and then a reattempt at merging. Those variants that could not be merged with the reference panel were deleted and those that were able to be merged were retained in the format of the successful merge (see http://pngu.mgh.harvard.edu/~purcell/plink/dataman.shtml#merge).

3.3.6.7. Phasing

The process of phasing assigns alleles (A/C/T/G) to the paternal / maternal chromosomes, i.e. the correct haplotype, so that now each individual has, instead of a single series of genotypes, two series of haplotypes, each containing alleles that sit together on the same chromosome. For a pair of genotypes e.g. A/a and B/b, along one chromosome in unphased data, it is not clear if the genotypes are AB and ab *or* aB and Ab. Phasing determines which the most likely combination is in the respective individual, based on known haplotypes in a reference panel.

SHAPEIT v2.r837 was used to phase the dataset to haplotype format(Delaneau et al., 2011). SHAPEIT requires haplotype genetic maps (African populations downloaded from the 1000G project) as a reference data panel.

I initially trialled and failed using two other phasing software programmes: CONFORM-GT and fastPHASE. CONFORM-GT was unsuccessful due to allele frequency differences with the HapMap reference data (not African samples). fastPHASE successfully phased but the phased data subsequently failed to impute, possibly because of the quality of phasing. Also, fastPHASE requires the removal of missing genotypes and this might have deleted too many variants. Both approaches required significant data formatting pre- and post-running as the input/output data formats were different.

Increasing phasing accuracy comes at the cost of reasonable running times. The accuracy can be improved by increasing the number of states (per variant) on which haplotype estimation is based. 100 states across the dataset gives good accuracy while maintaining reasonable running times. The final haplotype estimate was found by averaging across the 20 main iterations.

Finally, phased data was visually inspected in readable format to compare with pre- and post-strand alignment.

3.3.6.8. Post-phasing QC

First, I compared African samples from 1000 Genomes Project with our cohort to identify potentially flipped alleles based on allele frequencies in the populations. "Flipping" of alleles is a data error where the names for the two alleles of a marker have been swapped during allele calling. This generates a serious problem for subsequent imputation procedures. I assessed and trial-flipped those alleles with a minor allele frequency <0.35 plus an allele frequency difference between the cohort and reference data of at least 0.15. These alleles were flipped back.

Second, I used Will Rayner's QC perl script(Rayner, 2017) to identify and delete variants which did not fulfil QC criteria: the variant was not in the 1000 Genomes Project, in-del variants, palindromic variants, and variants with non-matching alleles with the reference panel.

I used a tight threshold (0.1) to identify significant differences in allele frequencies between the cohort data and the African populations in 1000 Genomes Project. When optimising this step, it is a balance between the *number* and the *quality* of variants in the pre-imputation set in order to increase imputation accuracy: that is, a balance between fewer quality variants and more, but poorer quality, variants in the pre-imputation dataset.

A comparison of different pre-imputation QC is displayed in Appendix 2, in relation to ultimate imputation concordance rates. When comparing imputation concordance rates, this needs to be viewed in relation to the absolute number of variants pre- and post- imputation datasets.

3.3.6.9. Imputation on the online Michigan imputation server

The online Michigan imputation server is a fast, efficient method that now surpasses imputation software used locally, see section 3.2.3.1. The server allows upload of a user's dataset (that has been strand aligned and phased); this is then imputed using a reference panel of the user's choice. Prior to imputation, an extensive and compulsory QC is performed, including a check to ensure no excess of "strand flips" (strand alignment errors) or "allele switches" (phasing errors). This is displayed in Appendix 3.

The choice of reference panel and population is important. The server offers imputation from a variety of reference panels: HapMap, 1000 Genomes Phase 1 and 3, CAAPA and (since my work), the new HRC reference panel.

Historically, populations were imputed against reference panels of similar ethnicity (possibly because to improve phasing when carried out as part of imputation, requiring appropriate

haplotypes), but more recently, a more diverse (and larger) reference dataset has been advocated. Thus, using all populations in the 1000 Genomes Project was considered the ideal approach, however, part of the Michigan pre-QC requires allele frequencies that are similar enough to the reference population. Using ALL populations led to mismatches between allele frequencies in our sample set versus reference populations which did not pass Michigan QC. Unfortunately, I was therefore forced to restrict the choice to African populations only (N=1018 samples).

3.3.6.10. Post imputation QC

After imputation, I performed further data processing and quality control. First, duplicated variants were identified and removed and family ID and sex needed to be reinserted as they were lost in the vcf file / imputation step respectively.

The quality of imputation was then evaluated via two considerations. The first consideration was the correlation between imputed and raw genotype calls. See Table 1 for concordance rates by chromosome; mean whole-genome concordance rate was 0.99835685.

The second consideration was "imputation quality". Imputation quality measure is specific to different software. On the Michigan server, it is represented by r^2 in the .info file. However it is presented, the quality score α is in a range 0-1. α for N individuals has equivalent power to αN perfectly genotyped individuals. So, in a cohort with 1000 samples, an α (quality score) of 0.4 at a marker indicates that the amount of accurate information obtained is equivalent to perfect genotype data in a sample of size 400. There is no standardised threshold for quality control as it is study population-dependent (especially in relation to study sample size). In the literature, many people have used a cut-off 0.3-0.5; these levels have mostly been applied in typical cohort sizes of a few thousand. I established a cut-off by assessing the distribution of info scores. I plotted the density of the info r^2 in R, chromosome by chromosome, see Appendix 4. Prior to marker removal, mean r^2 was 0.5244643. Based on this and the density plots, I used a cut-off info r^2 rate of 0.5. 26,124,664 of 46,925,116 variants (55.9%) were "well-imputed" (i.e. info score > 0.5) and carried forward.

3.3.6.11. Merge imputation dataset with raw genotypes

The imputed data were then carefully merged with the original genotyped dataset (in strand-aligned, phased mode), using PLINK's merge function. First, the data was "dummy" merged to identify potential merge issues, and these variants were corrected or deleted. Variants with problems were reviewed and stored separately. A full merge was then undertaken to keep raw calls (i.e. the original MEGA chip calls) and add in imputed variant calls not found in the original dataset, see Table 1 for a breakdown by chromosome of variants during the imputation

process. In summary, there were 980,120 variants pre-imputation and 25,740,653 post-imputation. Of 833,979,266 non-missing overlapping calls (all variants and all samples), 832,608,911 were concordant, giving a concordance rate of 0.99835685. 335,915 variants had a total of 1,370,377 call differences between pre- and post- imputation. After merging pre- and post-imputation files (including deletion of problem variants), there were 25,740,653 variants remaining.

Table 1 Imputation quality by chromosome

Chr	Number of variants pre-imputation	Number of variants post-imputation	Number of replicates	Number of badly imputed variants i.e. imputation .info r^2 <=0.5	Number of post-imputation variants after removing duplicates and badly imputed variants	Concordance information: all variants / all samples				Number of variants post-merge in final file
						Overlapping calls (no missing data)	Concordance rate	All differences in variant calls between pre- and post-imputation	Differences in unique variants between pre- / post-imputation	
1	77307	3738755	113	1685603	2038872	65654770	0.99848	99803	24535	2010379
2	85503	4058012	106	1790533	2251824	72826146	0.99838	118010	29194	2219618
3	72399	3356229	90	1441689	1901335	61699715	0.998328	103168	24788	1873392
4	65157	3338493	84	1424115	1901238	55546455	0.99837	90515	21460	1872498
5	58081	3032670	73	1299059	1721382	49498606	0.99835	81677	19922	1695440
6	67474	2954719	98	1248586	1694290	57481561	0.998473	87797	20834	1668254
7	54732	2753719	100	1203505	1539370	46643082	0.998349	77009	18826	1517165
8	51873	2651783	69	1147727	1493046	44201489	0.998249	77407	18640	1470402
9	42186	2063294	58	936349	1118563	35910014	0.998322	60267	15560	1102223
10	48093	2334316	67	1028296	1296733	40954316	0.998323	68679	16570	1277881
11	48518	2333584	65	1043239	1280890	41187420	0.998348	68062	16808	1262359
12	45583	2242990	73	982135	1251698	38760396	0.9984	62017	15074	1233090
13	34607	1661784	43	703452	951909	29556216	0.998344	48949	11687	937524
14	31612	1535740	45	675613	853981	26911514	0.998331	44922	10959	841635
15	29729	1404332	29	645659	753017	25284071	0.998272	43696	10767	742248
16	32847	1549576	47	737328	805187	27852500	0.998269	48202	12634	792854
17	30232	1346169	68	627919	712352	25554026	0.998362	41854	10585	702323
18	28207	1319723	32	581988	732546	24065417	0.998261	41850	10262	721959
19	23415	1084851	48	515177	564860	19729854	0.998509	29415	7862	556445
20	24172	1047749	36	468430	575128	20556002	0.998264	35690	8867	567099
21	13583	653867	13	302749	348356	11547218	0.998354	19012	4632	342808
22	14810	652313	24	311301	338087	12558478	0.998218	22376	5449	333057
Total	**980120**	**47114668**	**1381**	**20800452**	**26124664**	**833979266**	**0.99835685**	**1370377**	**335915**	**25740653**

3.3.6.12. Post-merge QC

Once merged, final quality control measures are applied (prior to statistical analysis). I used thresholds based on UK Biobank procedures (http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf). These are broadly less strict criteria for variants than pre-imputation QC as greater stringency for imputation accuracy is required. I again performed the QC in PLINK in the following order to optimise number of samples over number of variants:

- Marker genotyping rate ('geno') > 0.02
- Minor allele frequency ('maf') > 0.01
- Hardy Weinberg equilibrium ('hwe') > 0.0000001
- Individual genotyping rate ('mind') > 0.01

Notably, many of the variants were monomorphic (i.e. the minor allele was not present in our dataset) so applying the maf > 0.01 threshold dramatically reduced the dataset down to 15,977,584 variants.

3.4. Results

3.4.1. Quality control and imputation pipeline

3.4.1.1. Four scripts to manage the quality control and imputation pipeline

In addition to my goal of enabling and performing genetic analysis of our SCD population, an important objective of my work at King's was to enable other team members to follow my footsteps and analyse new data and additional patients. For this purpose, I have written 4 scripts to manage the pipeline from raw variant calling to quality controlled, imputed data: run_QC.sh (in Appendix 5). PreMichiganProcessing_SHAPEIT.sh (Appendix 6Appendix 6), PostMichiganProcessing_SHAPEIT.sh (Appendix 7), and PostIMputationAnalysis.sh (Appendix 8). Future samples from our sickle genebank will be able to be run through this pipeline relatively quickly. In addition, to accommodate the queueing scheduling system on Rosalind, all of these scripts were paired with an equivalent "outer" script that sends them to the job scheduling system, chromosome by chromosome.

3.4.1.2. run_QC.sh

run_QC.sh requires post variant-calling data in PLINK binary format. The script applies basic quality control using PLINK, as described in methods section 3.3.5.

The script produces (1) a QC'd raw genotyped dataset in PLINK binary format that can be used for analysis and (2) a pruned version of the genotyped dataset which is required for downstream bioinformatic manoeuvres.

3.4.1.3. PreMichiganProcessing_SHAPEIT.sh

PreMichiganProcessing_SHAPEIT.sh requires post-QC in PLINK binary format. It also requires reference panel data: both the download of 1000 Genomes phase 3 data (chromosome by chromosome) and HapMap genetic map for strand alignment, phasing and QC steps pre-imputation. These must be in the same build as the study data (GRCh37/hg19). The script performs the work of pre-imputation QC including: data processing, strand alignment, and phasing, using basic Linux, awk, PLINK and SHAPEIT, as described in methods sections 3.3.6.5-3.3.6.8.

The script produces bgzipped vcf files (with chromosome-by-chromosome data in strand aligned, phased haplotype form) suitable for upload to the Michigan imputation server at: https://imputationserver.sph.umich.edu/

3.4.1.4. PostMichiganProcessing_SHAPEIT.sh

PostMichiganProcessing_SHAPEIT.sh requires post-imputation data (chromosome-by-chromosome) from the Michigan imputation server in zipped vcf format. These are password-protected files so the password is required to inflate the files. The script applies basic data formatting to the files and converts them to PLINK format, then removes poor-quality imputed markers, before merging all well-imputed variants with pre-imputation (but post-phasing/strand alignment) variants, as described in sections 3.3.6.10.

The script produces chromosome-by-chromosome imputed data in binary PLINK format. It also produces a log file documenting different quality measures including imputation concordance rate.

3.4.1.5. PostImputationAnalysis.sh

PostImputationAnalysis.sh requires 22 sets of PLINK imputed data files by chromosome. The script merges the 22 sets of chromosome files to one set of whole-genome files, updates family ID (lost in translation to vcf files), updates sex (lost during imputation file conversion) and then applies QC thresholds as per UK Biobank criteria. This is described in sections 3.3.6.11-3.3.6.12.

Thus, one final quality-controlled, imputed genotype dataset is produced, in binary PLINK format, ready for association analysis (chapters 4 and 5).

3.4.2. Results for my data

Individuals were removed with excess missingness (0.01), outlying heterozygosity, relatedness (GRM score>0.9), and discordant phenotypic/genotypic sex information. Genetic variants with excess missingness (2%), deviation from Hardy-Weinberg equilibrium (0.0000001), or low

minor allele frequency (0.01) were removed. All quality control was performed using PLINK v1.90b3.38. This produced a raw genotype dataset of 832 individuals with 980,120 variants. Data were then imputed to 1000 Genomes phase 3 reference set using SHAPEIT (v2.r837) for pre-phasing and the online Michigan server. After post-imputation quality control, an imputed dataset of 832 individuals with 15,977,584 quality controlled variants was produced.

3.5. Discussion

Production of a large, imputed dataset with good quality genotyping takes significant effort and time. The genetic data itself (including accuracy of variant calls, QC thresholds, imputation quality information) must be understood so that the limitations of any data analysis undertaken are appreciate and results are not misinterpreted. Vigilance should be maintained during the subsequent genetic analysis: to re-assess variant call plots for positive signals, to be cautious when interpreting results of variants with low allele frequencies, and to re-evaluate the origin of variants (raw data versus imputed data) for positive association signals.

It must be reiterated here that QC criteria are subjective and vary from one study to another. Sample QC filters should not be so stringent as to remove the majority of the analysis cohort. Variant QC filters should eliminate the worst quality markers without removing good quality (and potentially significant) markers. This involves some "trial and error" where the quality of downstream manoeuvres (e.g. imputation) can be assessed based on parameters chosen further upstream.

While genotype imputation is computationally demanding (and used to require access to a high-performance computing cluster), the recent introduction of online imputation servers like the Michigan server I used has facilitated imputation for those without large computer resources.

The final imputed dataset produced is not only useful for direct genotyping of variants of interest but, more importantly, to enlarge an association analysis (increasing power), getting more significant variants and thus fine-mapping to get closer to the causative mutation. Imputation can also facilitate future meta-analysis: this has already proved useful in forming collaborations with a Tanzanian group who have performed a (similar) micro-array on a sickle cohort.

Our group have also benefited more widely from producing this massive genotype dataset. These processes have yielded information that can be used not only directly in subsequent

analyses, but can be fed back into our sickle research genebank database, including evidence of genetic relatedness, potential duplicate samples that can be cross-checked, and genome-wide scores of genetic ancestry.

References

ANDERSON, C. A., MASSEY, D. C., BARRETT, J. C., PRESCOTT, N. J., TREMELLING, M., FISHER, S. A., GWILLIAM, R., JACOB, J., NIMMO, E. R., DRUMMOND, H., LEES, C. W., ONNIE, C. M., HANSON, C., BLASZCZYK, K., RAVINDRARAJAH, R., HUNT, S., VARMA, D., HAMMOND, N., LEWIS, G., ATTLESEY, H., WATKINS, N., OUWEHAND, W., STRACHAN, D., MCARDLE, W., LEWIS, C. M., LOBO, A., SANDERSON, J., JEWELL, D. P., DELOUKAS, P., MANSFIELD, J. C., MATHEW, C. G., SATSANGI, J. & PARKES, M. 2009. Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. *Gastroenterology,* 136**,** 523-9.e3.

ANDERSON, C. A., PETTERSSON, F. H., CLARKE, G. M., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2010. Data quality control in genetic case-control association studies. *Nat Protoc,* 5**,** 1564-73.

AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A. & ABECASIS, G. R. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

BIOBANK, U. 2015. *Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource* [Online]. Available: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf [Accessed].

DAS, S., FORER, L., SCHONHERR, S., SIDORE, C., LOCKE, A. E., KWONG, A., VRIEZE, S. I., CHEW, E. Y., LEVY, S., MCGUE, M., SCHLESSINGER, D., STAMBOLIAN, D., LOH, P. R., IACONO, W. G., SWAROOP, A., SCOTT, L. J., CUCCA, F., KRONENBERG, F., BOEHNKE, M., ABECASIS, G. R. & FUCHSBERGER, C. 2016. Next-generation genotype imputation service and methods. *Nat Genet,* 48**,** 1284-7.

DELANEAU, O., MARCHINI, J. & ZAGURY, J. F. 2011. A linear complexity phasing method for thousands of genomes. *Nat Methods,* 9**,** 179-81.

ELSTON, R. C. & STEWART, J. 1971. A general model for the genetic analysis of pedigree data. *Hum Hered,* 21**,** 523-42.

EXCOFFIER, L. & SLATKIN, M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol,* 12**,** 921-7.

FUCHSBERGER, C., FLANNICK, J., TESLOVICH, T. M., MAHAJAN, A., AGARWALA, V., GAULTON, K. J., MA, C., FONTANILLAS, P., MOUTSIANAS, L., MCCARTHY, D. J., RIVAS, M. A., PERRY, J. R. B., SIM, X., BLACKWELL, T. W., ROBERTSON, N. R., RAYNER, N. W., CINGOLANI, P., LOCKE, A. E., TAJES, J. F., HIGHLAND, H. M., DUPUIS, J., CHINES, P. S., LINDGREN, C. M., HARTL, C., JACKSON, A. U., CHEN, H., HUYGHE, J. R., VAN DE BUNT, M., PEARSON, R. D., KUMAR, A., MULLER-NURASYID, M., GRARUP, N., STRINGHAM, H. M., GAMAZON, E. R., LEE, J., CHEN, Y., SCOTT, R. A., BELOW, J. E., CHEN, P., HUANG, J., GO, M. J., STITZEL, M. L., PASKO, D., PARKER, S. C. J., VARGA, T. V., GREEN, T., BEER, N. L., DAY-WILLIAMS, A. G., FERREIRA, T., FINGERLIN, T., HORIKOSHI, M., HU, C., HUH, I., IKRAM, M. K., KIM, B. J., KIM, Y., KIM, Y. J., KWON, M. S., LEE, J., LEE, S., LIN, K. H., MAXWELL, T. J., NAGAI, Y., WANG, X., WELCH, R. P., YOON, J., ZHANG, W., BARZILAI, N., VOIGHT, B. F., HAN, B. G., JENKINSON, C. P., KUULASMAA, T., KUUSISTO, J., MANNING, A., NG, M. C. Y., PALMER, N. D., BALKAU, B., STANCAKOVA, A., ABBOUD, H. E., BOEING, H., GIEDRAITIS, V., PRABHAKARAN, D., GOTTESMAN, O., SCOTT, J., CAREY, J., KWAN, P., GRANT, G., SMITH, J. D., NEALE, B. M., PURCELL, S., BUTTERWORTH, A. S., HOWSON, J. M. M., LEE, H. M., LU, Y., KWAK, S. H., ZHAO, W., DANESH, J., LAM, V. K. L., PARK, K. S.,

SALEHEEN, D., et al. 2016. The genetic architecture of type 2 diabetes. *Nature,* 536**,** 41-47.
HIRSCHHORN, J. N. & DALY, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet,* 6**,** 95-108.
HOWIE, B., FUCHSBERGER, C., STEPHENS, M., MARCHINI, J. & ABECASIS, G. R. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet,* 44**,** 955-959.
HUANG, L., LI, Y., SINGLETON, A. B., HARDY, J. A., ABECASIS, G., ROSENBERG, N. A. & SCHEET, P. 2009. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet,* 84**,** 235-50.
JOHNSON, G. C., ESPOSITO, L., BARRATT, B. J., SMITH, A. N., HEWARD, J., DI GENOVA, G., UEDA, H., CORDELL, H. J., EAVES, I. A., DUDBRIDGE, F., TWELLS, R. C., PAYNE, F., HUGHES, W., NUTLAND, S., STEVENS, H., CARR, P., TUOMILEHTO-WOLF, E., TUOMILEHTO, J., GOUGH, S. C., CLAYTON, D. G. & TODD, J. A. 2001. Haplotype tagging for the identification of common disease genes. *Nat Genet,* 29**,** 233-7.
KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S. Y., FREIMER, N. B., SABATTI, C. & ESKIN, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet,* 42**,** 348-54.
MARCHINI, J. & HOWIE, B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet,* 11**,** 499-511.
MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. & DONNELLY, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet,* 39**,** 906-13.
MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69.
MORRIS, A. P. & ZEGGINI, E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol,* 34**,** 188-93.
OZAKI, K., OHNISHI, Y., IIDA, A., SEKINE, A., YAMADA, R., TSUNODA, T., SATO, H., SATO, H., HORI, M., NAKAMURA, Y. & TANAKA, T. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet,* 32**,** 650-4.
POMPANON, F., BONIN, A., BELLEMAIN, E. & TABERLET, P. 2005. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet,* 6**,** 847-59.
PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.
RAYNER, W. 2017. *HRC or 1000G Imputation preparation and checking* [Online]. Available: http://www.well.ox.ac.uk/~wrayner/tools/ [Accessed].
SILVERBERG, M. S., CHO, J. H., RIOUX, J. D., MCGOVERN, D. P., WU, J., ANNESE, V., ACHKAR, J. P., GOYETTE, P., SCOTT, R., XU, W., BARMADA, M. M., KLEI, L., DALY, M. J., ABRAHAM, C., BAYLESS, T. M., BOSSA, F., GRIFFITHS, A. M., IPPOLITI, A. F., LAHAIE, R. G., LATIANO, A., PARE, P., PROCTOR, D. D., REGUEIRO, M. D., STEINHART, A. H., TARGAN, S. R., SCHUMM, L. P., KISTNER, E. O., LEE, A. T., GREGERSEN, P. K., ROTTER, J. I., BRANT, S. R., TAYLOR, K. D., ROEDER, K. & DUERR, R. H. 2009. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet,* 41**,** 216-20.
STEPHENS, M., SMITH, N. J. & DONNELLY, P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet,* 68**,** 978-89.
SUDMANT, P. H., RAUSCH, T., GARDNER, E. J., HANDSAKER, R. E., ABYZOV, A., HUDDLESTON, J., ZHANG, Y., YE, K., JUN, G., HSI-YANG FRITZ, M., KONKEL, M. K., MALHOTRA, A., STUTZ, A. M., SHI, X., PAOLO CASALE, F., CHEN, J., HORMOZDIARI, F., DAYAMA, G., CHEN, K.,

MALIG, M., CHAISSON, M. J. P., WALTER, K., MEIERS, S., KASHIN, S., GARRISON, E., AUTON, A., LAM, H. Y. K., JASMINE MU, X., ALKAN, C., ANTAKI, D., BAE, T., CERVEIRA, E., CHINES, P., CHONG, Z., CLARKE, L., DAL, E., DING, L., EMERY, S., FAN, X., GUJRAL, M., KAHVECI, F., KIDD, J. M., KONG, Y., LAMEIJER, E.-W., MCCARTHY, S., FLICEK, P., GIBBS, R. A., MARTH, G., MASON, C. E., MENELAOU, A., MUZNY, D. M., NELSON, B. J., NOOR, A., PARRISH, N. F., PENDLETON, M., QUITADAMO, A., RAEDER, B., SCHADT, E. E., ROMANOVITCH, M., SCHLATTL, A., SEBRA, R., SHABALIN, A. A., UNTERGASSER, A., WALKER, J. A., WANG, M., YU, F., ZHANG, C., ZHANG, J., ZHENG-BRADLEY, X., ZHOU, W., ZICHNER, T., SEBAT, J., BATZER, M. A., MCCARROLL, S. A., THE GENOMES PROJECT, C., MILLS, R. E., GERSTEIN, M. B., BASHIR, A., STEGLE, O., DEVINE, S. E., LEE, C., EICHLER, E. E. & KORBEL, J. O. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature,* 526**,** 75-81.

TEAM, R. D. C. 2011. R: A language and environment for statistical computing. . Foundation for Statistical Computing, Vienna, Austria. .

WEALE, M. E. 2010. Quality control for genome-wide association studies. *Methods Mol Biol,* 628**,** 341-72.

WITTKE-THOMPSON, J. K., PLUZHNIKOV, A. & COX, N. J. 2005. Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet,* 76**,** 967-86.

YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. & VISSCHER, P. M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-9.

YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82.

Appendix 1

Preliminary Information, Subject To Change

illumina

Infinium Multi-Ethnic Genotyping Array (MEGA)

The Multi-Ethnic Genotyping Array (MEGA) leverages content from Phase 3 of the 1000 Genomes Project (1KGP), the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA), Population Architecture in Genomics and Epidemiology (PAGE), and OMIM/Clinvar to create a true multi-ethnic array. To date, there has been a significant investment in detecting common genetic variants associated with complex disease and replicating associations across populations. However, functional and medically relevant variation might be rare or population-specific, requiring studies of diverse human populations to identify new risk factors. These studies are essential for enabling precision medicine, adding value to biobank repositories, empowering the next generation of genetic studies, and demonstrating the importance of understanding and measuring fine-scale population structure and its associations with biomedical traits. In total, these studies will shed light on modern human origins, population history, and the genetic basis of non-pathologic traits, as well as traits that affect disease susceptibility.

		% Covered*	Mean r^2
EUR	(non-Singletons)	0.503	0.585
	(MAF > 1%)	0.625	0.719
	(MAF > 5%)	0.725	0.813
AFR	(non-Singletons)	0.244	0.422
	(MAF > 1%)	0.281	0.479
	(MAF > 5%)	0.392	0.617
EAS	(non-Singletons)	0.534	0.601
	(MAF > 1%)	0.643	0.719
	(MAF > 5%)	0.739	0.814
AMR	(non-Singletons)	0.438	0.590
	(MAF > 1%)	0.535	0.684
	(MAF > 5%)	0.656	0.790

The MEGA contains content from the most popular Illumina commercial arrays and markers expertly selected by our consortium partners, combining backwards compatibility with the most up-to-date genomic discoveries.

Specifically MEGA contains:

- The Infinium HumanCore content with highly informative genome-wide tag SNPs
- African Diaspora content identified through the sequencing of 700 individuals by CAAPA
- Genome-wide Hispanic coverage selected from Phase 3 of 1KGP by PAGE
- The Infinium HumanExome content with exonic content selected from over 12,000 individuals
- Additional multi-ethnic exome and functional variant content designed by PAGE

High Value Content	% Regions Covered**	Number of Markers
ADME Core and Extended genes	99.00%	19,594
ADME Core and Extended genes +/- 10kb	99.34%	23,720
Blood Group Genes	100.00%	2,433
COSMIC Genes	98.76%	954,127
Finger Print SNPs	100.00%	841
HLA	100.00%	2,182
MHC	100.00%	25,143
NHGRI GWAS Catalog	74.36%	9,637
refGene_3UTRs	42.57%	44,474
refGene_5UTRs	24.28%	31,914
refGene	91.89%	1,054,045
refGene_promoters	62.43%	46,777
refGene_splice regions***	2.39%	10,662

*at R² > 0.8
**Coverage defined as having at least one marker in the region
***A splice "region" is defined as the two bases before or after an exon with the exception of exon 1 and the last exon

© 2014 Illumina, Inc. All rights reserved

Appendix 2
A comparison of post-phasing, pre-imputation quality control

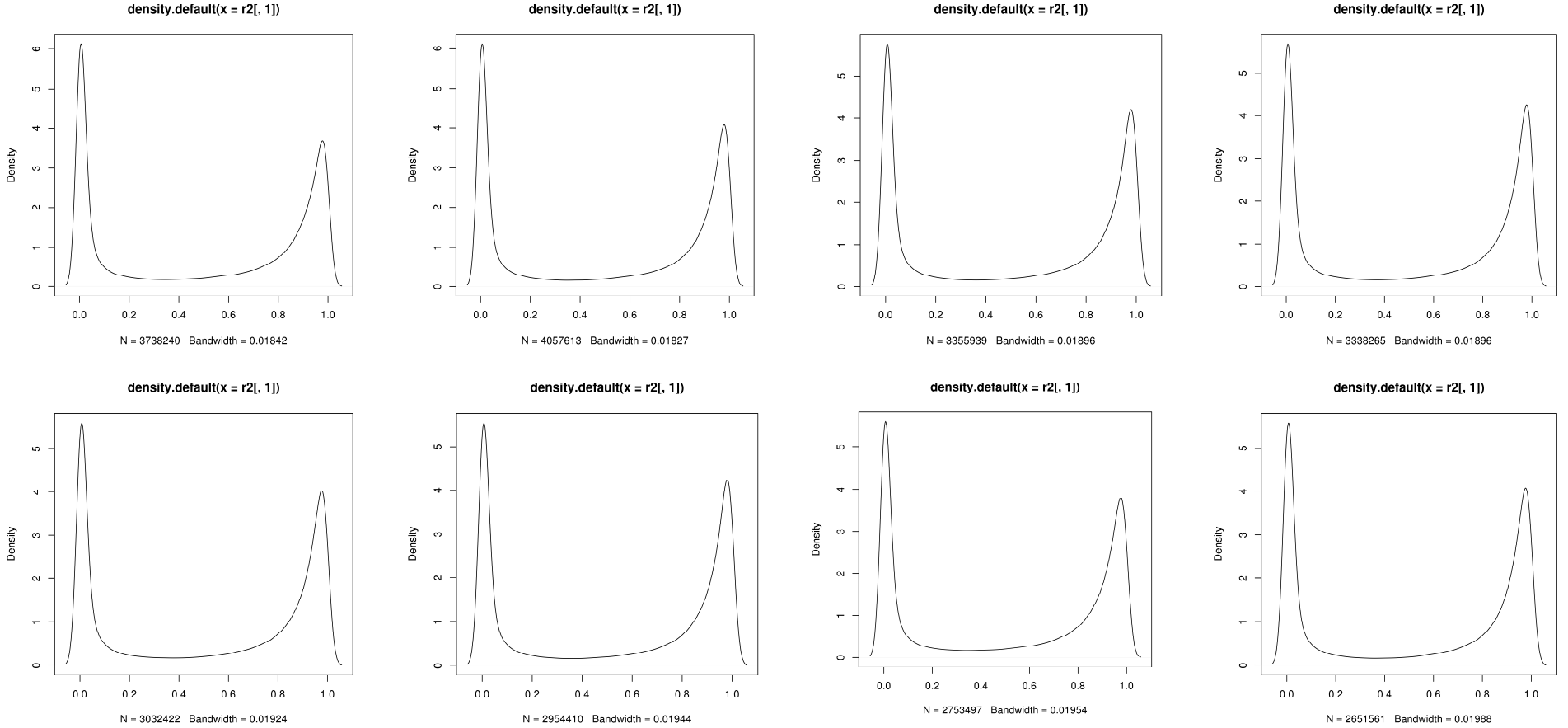
Post phasing, pre-imputation parameter modifications: chromosome 22. Final decision line highlighted in grey.

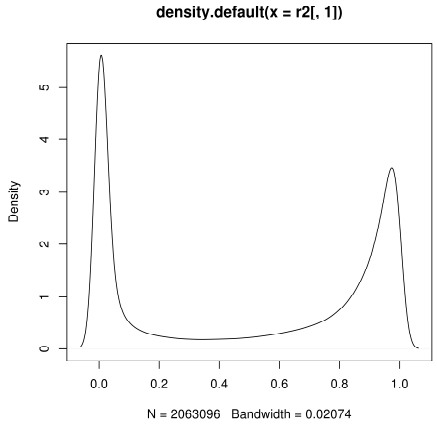
Haplotypes	Flipped alleles by MAF threshold	Action – exclude or flip	Allele difference threshold	Number variants pre imputation	Imputation ref panel source	Imputation population	Number of variants post imputation	Concordance rate	All differences in variant calls between pre / post imputation	Differences in unique variants between pre / post imputation
ALL	-		0.1	14194	1000G Ph3v5	AFR	638148	0.99648	42483	6023
ALL	-		0.2	14231	1000G Ph3v5	AFR	638148	0.995874	49932	6035
ALL	-		0.3	14271	1000G Ph3v5	AFR	638148	0.994955	61231	6080
ALL	-		0.4	14305	1000G Ph3v5	AFR	638148	0.994162	71022	6100
ALL	0.05	Flip	0.1	14518	1000G Ph3v5	AFR	638108	0.996461	43681	6132
ALL	0.1	Flip	0.1	14605	1000G Ph3v5	AFR	638087	0.996432	44303	6228
ALL	0.15	Flip	0.1	14658	1000G Ph3v5	AFR	638075	0.996408	44761	6235
ALL	0.15	Exclude	no	15163	1000G Ph3v5	AFR	638146	0.991074	109094	6293
AFR	0.1	Exclude	no	15188	1000G Ph3v5	AFR	638146	0.992615	90395	6245
AFR	0.15	Exclude	no	15178	1000G Ph3v5	AFR	638146	0.993003	85603	6262
AFR	0.1	Flip	no	15797	1000G Ph3v5	AFR	638144	0.992603	94374	6537
AFR	0.15	Flip	no	15794	1000G Ph3v5	AFR	638141	0.992931	90172	6517

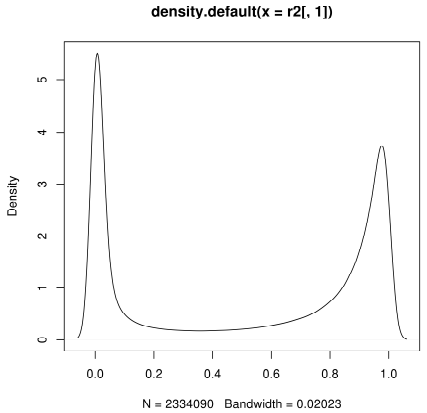
Appendix 3
Michigan imputation server QC (pre-imputation)

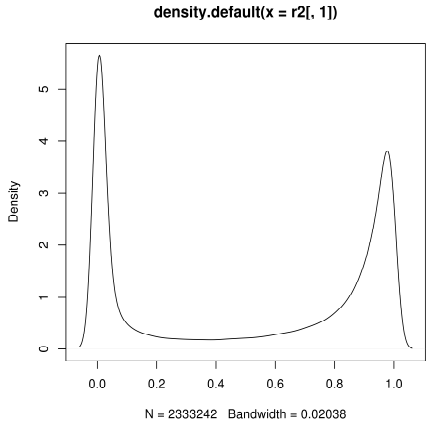
		Chunk 1	Chunk 2	Chunk 3	Chunk 4
Input Validation	Number of valid VCF file(s)	3	4	5	10
	Samples:	857	857	857	857
	Chromosomes:	1 2 3	4 5 6 7	8 9 10 11 12	13 14 15 16 17 18 19 20 21 22
	SNPs:	235209	245444	236253	263214
	Chunks: 4	36	37	37	44
	Reference Panel:	phase3	phase3	phase3	phase3
	Phasing:	SHAPEIT	SHAPEIT	SHAPEIT	SHAPEIT
Quality Control:	**Statistics:**				
	Alternative allele frequency > 0.5 sites:	0	0	0	0
	Reference Overlap:	99.49%	99.59%	99.47%	99.34%
	Match:	196,571	203,803	197,310	218,994
	Allele switch:	35,331	38,318	35,580	40,065
	Strand flip:	47	39	46	43
	Strand flip and allele switch:	4	7	6	6
	A/T, C/G genotypes:	2,052	2,270	2,053	2,372
	Filtered sites:				
	Filter flag set:	0	0	0	0
	Invalid alleles:	0	0	0	0
	Duplicated sites:	0	0	0	0
	NonSNP sites:	0	0	0	0
	Monomorphic sites:	0	0	0	0
	Allele mismatch:	0	0	0	0
	SNPs call rate < 90%:	0	0	0	0
Summary	Excluded sites in total:	51	46	52	49
	Remaining sites in total:	233,954	244,391	234,943	261,431

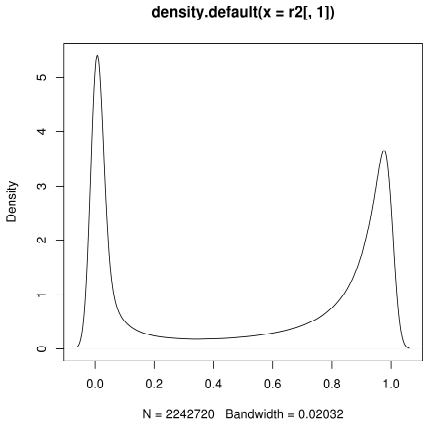
Appendix 4
Imputation call rates (.info) by chromosome

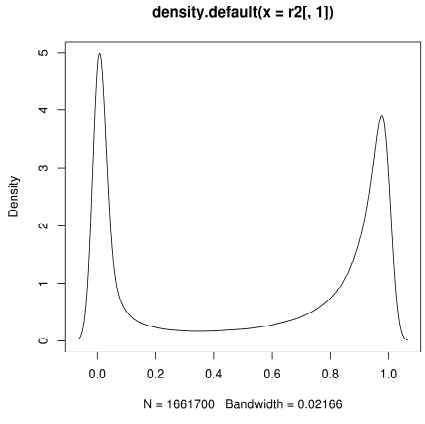


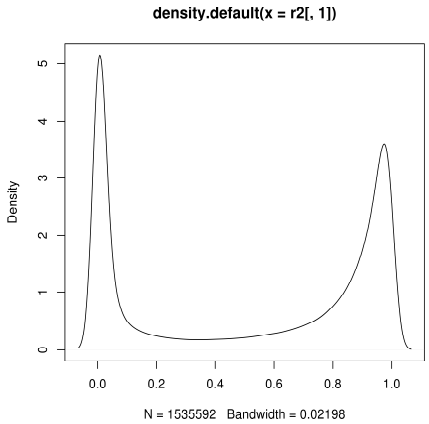

density.default(x = r2[, 1])
Density
N = 2063096 Bandwidth = 0.02074

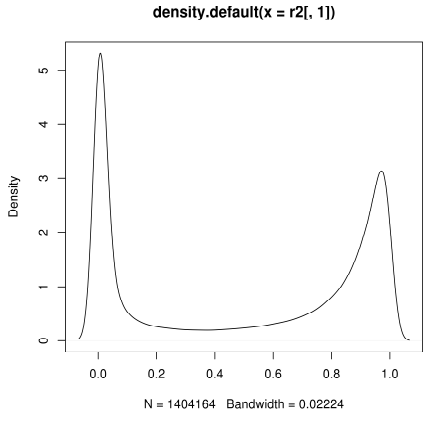

density.default(x = r2[, 1])
Density
N = 2334090 Bandwidth = 0.02023

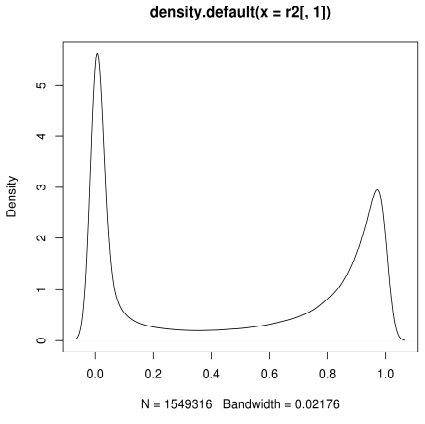

density.default(x = r2[, 1])
Density
N = 2333242 Bandwidth = 0.02038

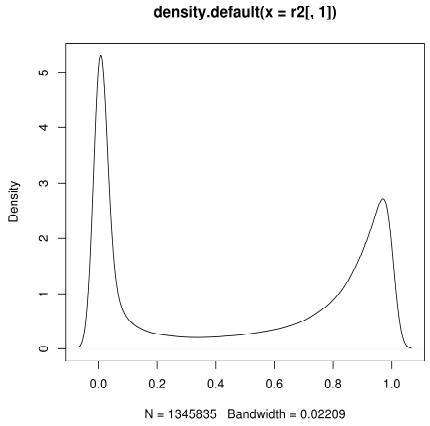

density.default(x = r2[, 1])
Density
N = 2242720 Bandwidth = 0.02032

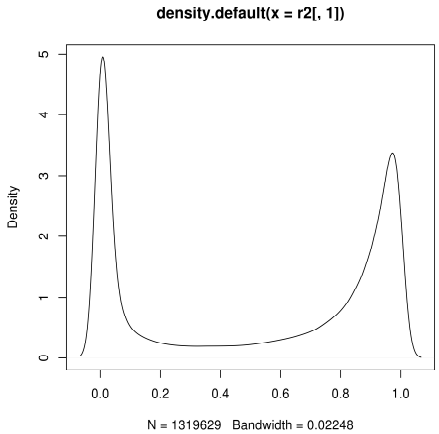

density.default(x = r2[, 1])
Density
N = 1661700 Bandwidth = 0.02166

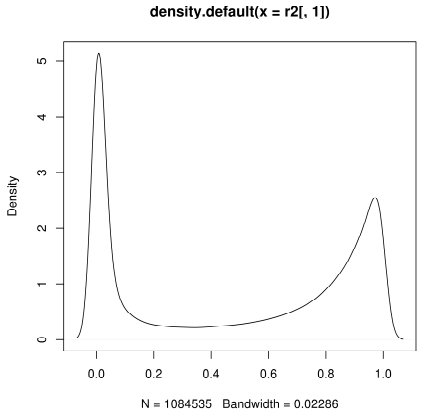

density.default(x = r2[, 1])
Density
N = 1535592 Bandwidth = 0.02198

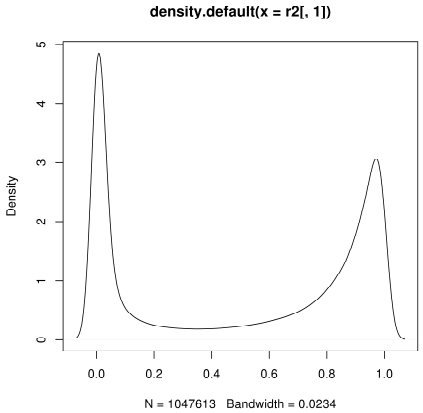

density.default(x = r2[, 1])
Density
N = 1404164 Bandwidth = 0.02224

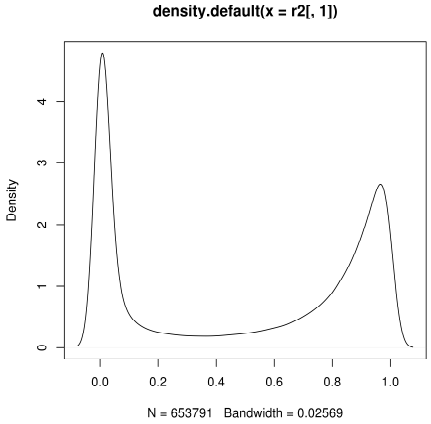

density.default(x = r2[, 1])
Density
N = 1549316 Bandwidth = 0.02176

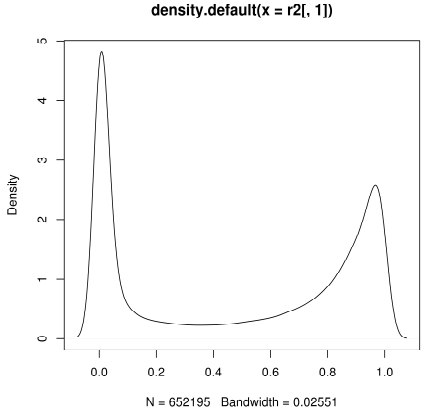

density.default(x = r2[, 1])
Density
N = 1345835 Bandwidth = 0.02209


density.default(x = r2[, 1])
Density
N = 1319629 Bandwidth = 0.02248


density.default(x = r2[, 1])
Density
N = 1084535 Bandwidth = 0.02286


density.default(x = r2[, 1])
Density
N = 1047613 Bandwidth = 0.0234


density.default(x = r2[, 1])
Density
N = 653791 Bandwidth = 0.02569


density.default(x = r2[, 1])
Density
N = 652195 Bandwidth = 0.02551

Appendix 5
run_QC.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -q HighMemLongterm.q,LowMemLongterm.q
#$ -M kate.gardner@doctors.org.uk
#$ -m beas
#$ -l h_vmem=20G
##########
# BASIC QUALITY CONTROL #
#########################

module add bioinformatics/plink2/1.90b3.38

# remove non sickle samples
plink --bfile SickleCell_MEGA --remove SickleMEGA_nonSCD.txt --make-bed --out SickleMEGA_noNonSCD

# remove sample with mixup ID-0195
plink --bfile SickleMEGA_noNonSCD --remove DodgySCD.txt --make-bed --out SickleMEGA_noNonSCD2

# change FID (known updates according to clinical not genetic FID)
plink --bfile SickleMEGA_noNonSCD2 --update-ids FIDIIDrecode2.txt --make-bed --out SickleMEGA_noNonSCD_newFID

# QC genotyping rate>95% (ie <5% missing genotypes for a SNP, across all individuals)
plink --bfile SickleMEGA_noNonSCD_newFID --geno 0.05 --make-bed --out SickleMEGA_noNonSCD_newFID_geno0_05

# didn't do MAF = 0 in end as
# QC MAF<0.1% ie <1/1000 and since population<1000 this is equivalent to MAF=0 ie monomorphic SNPs. This is a
useful pre-imputation step to remove monomorphic SNPs but we may want to reintroduce them later eg as ethnicity
marker when comparing to different populations
```

```
plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05 --maf 0.001 --make-bed --out
SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001

# QC genotyping rate>95% for individuals (ie <5% missing genotypes for an individual, across all SNPs)
# note the QC for individuals done last (after SNP QC) to maximise number of individuals kept in
plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001 --mind 0.05 --make-bed --out
SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05

# HWE p<5x10^-8
plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05 --hwe 0.00000005 --make-bed --out
SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05_hwe0_00000005

# correct 4 with wrong sex attributed
plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05_hwe0_00000005 --update-sex CorrectSex.txt --
make-bed --out SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05_hwe0_00000005_sexCorrect



##################################
# PRUNING USING --INDEP-PAIRWISE #
##################################

# (aside: there are other PLINK pruning strategies using either --INDEP or --INDEP-PAIRPHASE)
# Aim: prune SNPs in two parts to get one SNP/LD block. Need pruned rather than all SNPs for downstream processing
steps including sex checking, PCA, ...
# --indep-pairwise has 3 parameters to define: {N M r2}
# N SNP window size in kb, large it is, more SNPs removed
# M SNP intervals to shift (step size to shift window at end of each step)
# r2 (NOT VIF!): VIF = 1/(1-R^2) where R^2 is the multiple correlation coefficient for a SNP being regressed on all
other SNPs simultaneously, lower it is, more SNPs removed
# Paul suggests r2~0.1 => VIF~1.1, or go lower to aim for pruned SNPs ~50k-100k

plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05_hwe0_00000005_sexCorrect --indep-pairwise 1000
100 0.04 --out SickleMEGA_PrunedSNPs
```

```
plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05_hwe0_00000005_sexCorrect  --extract
SickleMEGA_PrunedSNPs.prune.in --make-bed --out SickleMEGA_Pruned_1000_100_0.04


############################
# CHECK-SEX ON PRUNED DATA #
############################

# comparing clinical sex with genotypic sex
# this is a QC step in itself, but also much of the haploid heterozygote genotypes are a result of sex mismatches -
so discarding the sex mismatches removes hap heterozygotes (over 90% in my case)
# Needs to be done AFTER PRUNING. Final cutoffs depend on own data and also PRUNING THRESHOLDS above - cutoffs vary
wildly!
# Final definitions of sex cut-offs after I played around with the data:
# 1. F statistic: female max 0.3, male min 0.9
# 2. ycount: female max 6, male min 30
# anyone outside of these definition cannot create a "genotypic sex" (noone in my dataset)
# anyone where this genotypic sex is discordant with clinical sex, go back to collection site to confirm clinical
sex correctly recorded (hence addition of above --update-sex step where 4 samples incorrectly labelled
# if genotypic sex still =/= clinical sex, have to discard the sample - assume sample error
# this will have removed most haploid heterozygote genotypes


plink --bfile SickleMEGA_Pruned_1000_100_0.04 --check-sex ycount 0.3 0.9 6 30 --out
SickleMEGA_Pruned_1000_100_0.04_checkSex_ycount
grep PROBLEM SickleMEGA_Pruned_1000_100_0.04_checkSex_ycount.sexcheck | awk '{print $1"\t"$2}' > SexMismatches.txt

#remove sex mismatches from pruned data
plink --bfile SickleMEGA_Pruned_1000_100_0.04 --remove SexMismatches.txt --make-bed --out
SickleMEGA_Pruned_NoSexMismatch

#remove sex mismatches from unpruned data
plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05_hwe0_00000005_sexCorrect --remove
SexMismatches.txt --make-bed --out SickleMEGA_QC_NoSexMismatch
```

```
##############################################
# MANAGE ONGOING GENOME WIDE HETEROZYGOSITY #
##############################################

# Initially, I used Joni Coleman's R script to check for genome wide heterozygosity and remove extreme outliers>3sd
# However, these represent precisely the males in the sample and therefore haploid heterozygotes only an issue for
sex chromosomes. Since I am interested in autosomes only, I have extracted the autosomes and hap heterozygosity no
longer an issue
# Thus extracting autosomes is better for me than identifying / removing samples (6 samples) with heterozygosity
only in X chromosome.

# extract autosomes only, from both pruned and unpruned data
plink --bfile SickleMEGA_Pruned_NoSexMismatch --chr 1-22 --make-bed --out SickleMEGA_Pruned_NoSexMismatch.autosomes
plink --bfile SickleMEGA_QC_NoSexMismatch --chr 1-22 --make-bed --out SickleMEGA_QC_NoSexMismatch.autosomes
```

Appendix 6
PreMichiganProcessing_SHAPEIT.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -q HighMemLongterm.q,LowMemLongterm.q
#$ -M kate.gardner@doctors.org.uk
#$ -m beas
#$ -l h_vmem=40G
##########

# This script performs work of Pre imputation QC including: data processing, strand alignment, phasing
# To use queueing on the cluster, use a different script to call this one so different chromosomes can be processed as separate jobs for cluster resource efficiency: call_PreMichiganProcessing_SHAPEIT.sh
# However, this script started out as sequential for loops through the chromosomes which is where it still is: to be changed

# Summary:
# Aim is to upload sickle MEGA data and haplotype data (from 1000G) in vcf format to Michigan minimac server
# Data for Michigan minimac must be in strand aligned, hased haplotypes format - use plink to strand align and then shapeit to phase
# SNP names in sickle MEGA are non-standard so I will convert both sickle MEGA and haplotype SNP names to new SNP name of format "ChrNum_PosnNum"
# Finally, we are slicing the data by chromosome BUT I HAVE ONLY DONE AUTOSOMES

# Input and output files:
# Input: binary PLINK files (.bed, .bim, .fam) with basic QC (eg mind, geno, maf cutoffs): named SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25
# Output: bgzipped vcf files suitable for upload to the Michigan imputation server minimac at: https://imputationserver.sph.umich.edu/
```

```
# Assumes further downloads of a reference panel (1000G) an genetic map (HapMap) for strand alignment & phasing
steps:
# 1. Reference panel from 1000G: make sure same build as own data:
ALL.chr${x}.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz
# 2. Genetic map from HapMap for correct assembly (GRCh37 for MEGA data): genetic_map_HapMapII_GRCh37.tar.gz

#Add software modules required
            module add bioinformatics/htslib/1.3
            module add bioinformatics/plink/1.90b3.31
      module add bioinformatics/shapeit


#####################################################################
# DATA PROCESSING OF OWN DATA PRIOR TO STRAND ALIGNMENT AND PHASING #
#####################################################################

# 1. See what ref and alt allele names are in bim file
# VCF files only allowed to include {A, C, G, T, N, a, c, g, t, n} so will need to remove all other characters
# first see what characters there are in bim file, checking both ref and alt allele columns (5 and 6)
 cut -f5 SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25.bim |sort |uniq -c
 cut -f6 SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25.bim |sort |uniq -c


# 2. Delete SNPs that won't be accepted later either for shapeit, michigan server, or during formatting (all snps
need to be unique for plink2 manipulation)
# (i) random names ('B' in my case)
 plink2 --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25 --exclude SNPsToRemove.txt --make-bed --out
SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB


# (ii) I and D
# SNPS with I / D which comprise 2995 of 1145883 at this stage (0.26137% of all SNPs)
# Realised afterwards should have combined this and above removal of 'B' alleles. In fact, should have gone further
to do an if *not* ACGTacgtNn (ie vcf format) then produce this text file.
# HOWEVER, may actually want to rename/recode these SNPs rather than delete them
```

```
 awk '{if ($5=="I" || $5=="D" || $6=="I" || $6=="D") print $2}'
SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25.bim > SickleMEGA_VCF.txt
 plink2 --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB --exclude SNPsWithDI.txt --make-bed --out
SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB_noDI

# (iii) duplicates (difficulty later otherwise) - we are finding duplicate SNP names and duplicate SNP positions
since later we will update the SNP name as the SNP position
        cut -f2 SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB_noDI.bim | sort | uniq -c | awk
'{if($1>1) print $2}' > SickleMEGADuplicates.txt
        plink2 --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB_noDI --exclude
SickleMEGADuplicates.txt --make-bed --out SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB_noDI_noDups
        cut -f4 SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB_noDI.bim | sort | uniq -c | awk
'{if($1>1) print $2}' > SickleMEGADuplicates2.txt
        plink2 --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB_noDI_noDups --exclude
SickleMEGADuplicates2.txt --make-bed --out SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_25_noB_noDI_noDups2

     # Caution with ID in VCF files: VCF just contain sample IDs, instead of the distinct family and within-family
IDs tracked by PLINK. We have converted these by using  --double-id causes both family and within-family IDs to be
set to the sample ID.


#####################################
# CONVERT TO SINGLE CHROMOSOME FILES #
#####################################

# i. Convert to single chromosome files and also do some further processing:
# ii. Convert SNP names so can compare to reference data
# iii. remove duplicates

# CHROMOSOMES="1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22"
# CHROMOSOMES=$1

# for x in $CHROMOSOMES;
#         do
     # i Get sickle data into binary plink format by chromosome
```

```
      plink --bfile SickleMEGA_noNonSCD_newFID_geno0_05_maf0_001_mind0_05_hwe0_00000005 --chr ${x} --make-bed --out
Chr${x}

      # ii. convert snp names of sickle data into "ChrNum_PositionNum" format. This is because lots of the given
SNP names are non-standard in the sickle MEGA set
      # first use awk create text file of old SNP name / new SNP name (of format chr:position)
      awk '{print $2"\t"$1"_"$4}' Chr${x}.bim > Chr${x}newName.txt
      # second feed this file into plink as update-SNP name command
       plink --bfile Chr${x} --update-name Chr${x}newName.txt --make-bed --out Chr${x}_2

#     # iii. remove duplicates
      # shouldn't have to do this next step as have already removed duplicate position numbers above, however,
since duplicates remained I have added it in
         cut -f2 Chr${x}_2.bim | sort | uniq -c | awk '{if($1>1) print $2}' > Chr${x}Duplicates.txt
         plink --bfile Chr${x}_2 --exclude Chr${x}Duplicates.txt --make-bed --out Chr${x}_3


#        done


#############################################################
# DOWNLOAD AND DATA PROCESSING OF REFERENCE PANEL FROM 1000G #
#############################################################

# 1. Download sample area hg19 from 1000G in  gzipped format BY CHROMOSOME, and copy to Rosalind, this takes time
# WARNING: ensure using same build of reference panel as your own data


# 2. Data formatting of 1000G haplotype data before doing strand alignment

# CHROMOSOMES=$1

#  for x in $CHROMOSOMES;
#        do
      # 1. gunzip haplotype files to vcf format
```

```
        chmod 777 ALL.chr${x}*
        gunzip ALL.chr${x}.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz
      # 2. convert vcf to binary plink format
      plink --vcf ALL.chr${x}.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf --double-id --make-bed
--out Chr${x}Haplo

      # 3. re-gzip the (very large) vcf file
      gzip ALL.chr${x}.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf

      # 4. Remove duplicate SNPs by both SNP name
        cut -f2 Chr${x}Haplo.bim | sort | uniq -c | awk '{if($1>1) print $2}' > Chr${x}HaploDuplicates.txt
        plink --bfile Chr${x}Haplo --exclude Chr${x}HaploDuplicates.txt --make-bed --out Chr${x}Haplo_noDups


      # 5. re-name SNPs in format 'chr_position'
      # first use awk create text file of old SNP name / new SNP name [use semi colon as delimiter bw Chr and SNP
as this is how minimac server outputs SNP names so this format make it easier for pre- to post- imputation
comparison down the line]
      awk '{print $2"\t"$1"_"$4}' Chr${x}Haplo_noDups.bim > Chr${x}newNameHaplo.txt
      # Second, make this file unique
      sort -k2 "Chr${x}newNameHaplo.txt" | uniq > "Chr${x}newNameHaplo2.txt"
      # third, feed this file into plink as update-SNP name command
      plink --bfile Chr${x}Haplo_noDups --update-name Chr${x}newNameHaplo2.txt --make-bed --out
Chr${x}Haplo_noDups2

        # 6. Remove duplicate SNPs by both SNP name (previously position)
         cut -f2 Chr${x}Haplo_noDups2.bim | sort | uniq -c | awk '{if($1>1) print $2}' >
Chr${x}HaploDuplicates2.txt
         plink --bfile Chr${x}Haplo_noDups2 --exclude Chr${x}HaploDuplicates2.txt --make-bed --out
Chr${x}Haplo_noDups3

      # 7. remove "<" as not accepted allele
      fgrep "<" Chr${x}Haplo_noDups3.bim | awk '{print $0}' > Chr${x}WrongChar.txt
        plink --bfile Chr${x}Haplo_noDups3 --exclude Chr${x}WrongChar.txt --make-bed --out Chr${x}Haplo_noDups4
```

```
      # 8. Remove large, unnecessary files
      rm Chr${x}Haplo_noDups.*
        rm Chr${x}Haplo_noDups2.*
        rm Chr${x}Haplo_noDups3.*


#        done

####################
# STRAND ALIGNMENT #
####################

# Alignment comparison between reference 1000G and our phased data
# Aside: strand alignment can be done either in prephasing or in imputation steps, but not both. For Michigan
server, strand align in pre-phasing; and for impute2, strand align with imptutation
# Aside 2: Alternative to plink strand alignment, as below, is GTOOL
# In PLINK, achieve alignment using --merge and both --flip and --flipscan
# There are several possible causes for merge failures: the variant could be known to be triallelic (which plink
can't handle); there could be a strand flipping issue, or a sequencing error, or a previously unseen variant
# see merging files at http://pngu.mgh.harvard.edu/~purcell/plink/dataman.shtml#merge  or  https://www.cog-
genomics.org/plink2/data
# make strict bi allelic otherwise will have merge headaches


#  CHROMOSOMES="1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22"
#CHROMOSOMES=$1

#  for x in $CHROMOSOMES;
#        do

        # 1. Make strict bi allelic otherwise will have merge headaches
       plink --bfile Chr${x}_3 --biallelic-only strict --make-bed --out Chr${x}_4
       plink --bfile Chr${x}Haplo_noDups4 --biallelic-only strict --make-bed --out Chr${x}Haplo_noDups5
```

```
      # 2. Identify and flip obvious strand errors using --bmerge and --flip
        # i. do first (dry run) merge using merge-mode 7 : Report mismatching non-missing calls (diff mode 7 -- do
not merge)
       plink  --bfile Chr${x}Haplo_noDups5  --bmerge Chr${x}_4.bed Chr${x}_4.bim Chr${x}_4.fam  --merge-mode 7 --
out Chr${x}_Merge

        # ii. use Chr${x}_Merge.missnp file as input to flip alleles in our data
       plink  --bfile Chr${x}_4 --flip Chr${x}_Merge.missnp --make-bed --out Chr${x}_5

        # iii. re-merge (again, dry run) our data (with flipped alleles) and haplotype data to create : Report
mismatching non-missing calls (diff mode -- do not merge)
       plink --bfile Chr${x}Haplo_noDups5  --bmerge Chr${x}_5.bed Chr${x}_5.bim Chr${x}_5.fam  --merge-mode 7 --out
Chr${x}_Merge2

      # iv. flip errors back again (weren't strand errors)
       plink  --bfile Chr${x}_5 --flip Chr${x}_Merge2.missnp --make-bed --out Chr${x}_6


      # 3. Exclude remaining errors using --exclude
      #i. remerge to find remaining missnps
       plink  --bfile Chr${x}Haplo_noDups5  --bmerge Chr${x}_6.bed Chr${x}_6.bim Chr${x}_6.fam  --merge-mode 7 --
out Chr${x}_Merge3

         #ii. exclude these missnps that still exist from both our data and haplo data
       plink --bfile Chr${x}_6 --exclude Chr${x}_Merge3.missnp --make-bed --out Chr${x}_7
       plink --bfile Chr${x}Haplo_noDups5 --exclude Chr${x}_Merge3.missnp --make-bed --out Chr${x}Haplo_noDups6


        # 4. Use --flip-scan to find those in high negative LD and exclude these
      # i. re-merge our data again : Report mismatching non-missing calls (diff mode -- do not merge)
       plink  --bfile Chr${x}Haplo_noDups6  --bmerge Chr${x}_7.bed Chr${x}_7.bim Chr${x}_7.fam  --merge-mode 2 --
out Chr${x}_MergeTemp

      # ii. make phenotypes of cases (our data) versus controls (so --flipscan will know source of data)
```

```
       plink --bfile Chr${x}_MergeTemp --make-pheno Chr${x}_6.fam '*' --make-bed --out Chr${x}_6_fakepheno

      # iii. do flipscan
       plink --bfile Chr${x}_6_fakepheno --allow-no-sex --flip-scan --out Chr${x}_6_fakepheno_flipscan

      # iv. identify those with negative LD and exclude from dataset (WE COULD FLIP THEM INSTEAD???)
#     awk '{if ($9>0) print $2}' Chr${x}_6_fakepheno_flipscan.flipscan > Chr${x}_NegLD.txt
       plink --bfile Chr${x}_7 --exclude Chr${x}_NegLD.txt --make-bed --out Chr${x}_8


      # 5. Remove unnecessary, big files
      rm Chr${x}_6_fakepheno*
      rm Chr${x}_MergeTemp*

#        done


########################
# PHASING WITH SHAPEIT #
########################
# Use SHAPEIT to phase the data to haplotype format: input is haplotype genetic maps (from 1000G) and our data in
binary plink format
# following scripts cover phasing, checks on phasing, and uploading to minimac server for imputation
# Assumes download of genetic map from HapMap for correct assembly


#  CHROMOSOMES="1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22"
# CHROMOSOMES=$1

#  for x in $CHROMOSOMES;
#        do
        #1. Download genetic map (correct version)
        #Extract 2nd to 4th columns from this
       cut -f2-4 genetic_map_GRCh37_chr${CHROMOSOMES}.txt > genetic_map_GRCh37_chr${CHROMOSOMES}_3cols.txt
```

```
        #2. use shapeit to phase
        shapeit --input-bed Chr${CHROMOSOMES}_8 --input-map genetic_map_GRCh37_chr${CHROMOSOMES}_3cols.txt --
output-max Chr${CHROMOSOMES}_Phased.haps Chr${CHROMOSOMES}_Phased.sample --thread 8 --states 100

        #PARAMETERS:
        #input-bed: input files in binary plink format
        #input-map: genetic map using same db assembly as the input files. Not necessary to have full overlap with
SNPs, but helpful to have some overlap - For the SNPs that don't have a genetic position, SHAPEIT internally
determines its genetic position using linear interpolation.
            #If the intersection between your SNP map (BIM, MAP or GEN file) and your genetic map (GMAP file) is
poor, you should verify that the positions in both files are from the same build of the Human genome (b36 or b37
for example).
        #output-max: output file name for 2 files: haplotypes (haps) and samples
        #STATES: You can increase accuracy of SHAPEIT by increasing the number of conditioning states on which
haplotype estimation is based. By default, SHAPEIT uses 100 states per SNP across the dataset which gives good
accuracy while maintaining reasonable running times. The complexity of the algorithm       is linear with the
number of conditioning states, so feel free to increase this number if your computational resources allow it. For
instance, if you set this number to 200, it will take times longer than with 100 states.
        #burn X : where X is the number of the burn-in iterations. The default value is 7 which is sufficient for
most application. Those iterations are used to find a good starting haplotype estimate.
        #prune Y : where Y is the number of iterations of the pruning stage. The default value is 8. Transition
probabilities are stored across pruning iterations and used to prune the genotype graphs in order to get more
parsimonious graphs.
        #main Z : where Z is the number of main iterations. The default value is 20. The final haplotype estimate
is found by averaging across the main iterations.

        #3. Convert shapeit output to vcf and plink formats (using shapeit convert function)
       shapeit -convert --input-haps Chr${CHROMOSOMES}_Phased --output-vcf Chr${CHROMOSOMES}_Phased.vcf
        #       Convert vcf to binary plink format as will need this post-imputation
        plink --vcf Chr${CHROMOSOMES}_Phased.vcf --double-id --make-bed --out Chr${CHROMOSOMES}_Phased
        #[Aside: can also convert shapeit output to plink using gtool, see https://www.biostars.org/p/17266/]

        #4. Visual checks on phased data
        # i. convert plink binary to non binary format for visual comparison with pre-strand alignment and post-
strand alignment/pre-phasing
```

```
       plink --bfile Chr${x}_Phased --recode --out Chr${x}_Phased

#        done

################################################################################################
# Post phasing QC to identify difference in allele frequency between own data and reference data (1000G) #
################################################################################################

# CHROMOSOMES=$1

#  for x in $CHROMOSOMES;
#        do
      #1. identify and keep AFR samples only from 1000G only, and get allele frequencies (needed for below)
      awk 'NR==FNR { n[$1] = $1; next } ($1 in n) {print $0}' 1000G_Ph3_AFRsamples.txt Chr${x}Haplo_noDups6.fam >
Chr${x}Haplo_AFRsamples.txt
      plink --bfile Chr${x}Haplo_noDups6 --keep Chr${x}Haplo_AFRsamples.txt --make-bed --out Chr${x}Haplo_AFRonly
        plink --bfile Chr${x}Haplo_AFRonly --freq --out Chr${x}Haplo_AFRonly

        #2. identify potentially flipped alleles, "y=-x" based on comparing allele frequencies in our data and
1000G AFR data using a MAF <0.35 and a difference between ours and ref panel of 0.15 (does not appear to be
concordance issue if threshold is lower than this - may try higher threshold at a later date)
       plink --bfile Chr${x}_Phased --freq --out Chr${x}_Phased

       awk 'NR==FNR {a[$2]=$0;next} {if ($2 in a) print $2,$3,$4,$5,a[$2]}' Chr${x}_Phased.frq
Chr${x}Haplo_AFRonly.frq > Chr${x}_compareFreqAFR.txt
      awk '{print $1,$2,$3,$4,$7,$8,$9}' Chr${x}_compareFreqAFR.txt > Chr${x}_compareFreqAFR2.txt
      awk '{if($2==$6 && $3==$5) print $0}' Chr${x}_compareFreqAFR2.txt > Chr${x}_compareFreqAFR3.txt
      awk '{if($4<0.35 && $7<0.35 && sqrt(($4-$7)^2)<0.15) print $1}' Chr${x}_compareFreqAFR3.txt >
Chr${x}_flippedAllelesAFR.txt

      #3. flip these alleles, and get allele frequencies (needed for below)
      plink --bfile Chr${x}_Phased --flip Chr${x}_flippedAllelesAFR.txt --make-bed --out Chr${x}_Phased0.15_AFR
        plink --bfile Chr${x}_Phased0.15_AFR --freq --out Chr${x}_Phased0.15_AFR

      rm Chr${x}_compareFreqAFR.txt
```

```
        rm Chr${x}_compareFreqAFR2.txt
        rm Chr${x}_compareFreqAFR3.txt


      #4. use Will Rayner's QC perl script with tight threshold for difference between allele frequencies (t=0.1)
as this led to best imputation performance
      # http://www.well.ox.ac.uk/~wrayner/tools/
      # Removes those with one of 5 QC problems: no match to 1000G, indels, allele freq difference > threshold,
palindromic SNPs, non matching alleles
      # (most are either no match to 1000G or allele freq difference)

      gunzip 1000GP_Phase3_combined.legend.gz

        perl HRC-1000G-check-bim.pl -b Chr${x}_Phased0.15_AFR.bim -f Chr${x}_Phased0.15_AFR.frq  -r
1000GP_Phase3_combined.legend -g  -t 0.1 -p AFR
        # parameters:
        # -p <population> default is ALL. Choose 7 black populations in 1000G with -p AFR
        # -t <difference> : allele frequency thresholds. -t 0.2 sets allele difference threshold to 0.2 (default)
                #Use this to change the allele frequency difference used to exclude SNPs in the final file, range 0
- 1, the larger the difference the fewer SNPs will be excluded.
        # -n flag to specify that you do NOT wish to exclude any SNPs based on allele frequency difference, if -n
is used -t has no effect.

      gzip 1000GP_Phase3_combined.legend
        plink --bfile Chr${x}_Phased0.15_AFR --exclude Exclude-Chr${x}_Phased0.15_AFR-1000G.txt --make-bed --out
Chr${x}_Phased0.15_AFR_QC


      #5. convert QC files to vcf and bgzip
       plink --bfile Chr${x}_Phased0.15_AFR_QC --recode vcf --out Chr${x}_Phased0.15_AFR_QC
        bgzip Chr${x}_Phased0.15_AFR_QC.vcf

#        done
```

Appendix 7
PostMichiganProcessing_SHAPEIT_r2.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -q HighMemLongterm.q,LowMemLongterm.q
#$ -M kate.gardner@doctors.org..uk
#$ -m beas
#-l h_vmem=60G
##########

 module add bioinformatics/htslib/1.3
 module add bioinformatics/plink2/1.90b3.38

########################################
# DATA PROCESSING POST-IMPUTATION FILES #
########################################

# Download files from minimac imputation server and copy to Rosalind
# Also copy pre-imputation (but post-strand alignment/phasing) files across, named: Chr${CHROMOSOMES}_Phased.vcf

#  echo "***************sickle MEGA data formatting**************"

# CHROMOSOMES="1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22"
 CHROMOSOMES=$1

# for x in $CHROMOSOMES;
#              do

#     FILES PRE IMPUTATION (BUT POST STRAND ALIGNMENT, PHASING AND FURTHER QC) ARE CALLED
Chr${x}_Phased0.15_AFR_QC, then renamed Chr${CHROMOSOMES}_PreImputation.bed
#       FILES POST IMPUTATION WE WILL NAME Chr${x}_Imputed
```

```
      #1. unzip imputation files from minimac server, then make into plink format, so we have a set of imputed
plink files called Chr${CHROMOSOMES}_Imputed
      # will need PASSWORD (emailed) from minimac to inflate the files
      unzip -P 0tbEGAAOua696 chr_${CHROMOSOMES}.zip
      gunzip chr${CHROMOSOMES}.info.gz #need to take account of this later
      gunzip chr${CHROMOSOMES}.dose.vcf.gz

      plink --vcf chr${CHROMOSOMES}.dose.vcf --biallelic-only strict --make-bed --out Chr${CHROMOSOMES}_Imputed
        gzip chr${CHROMOSOMES}.dose.vcf


      # 2. remove duplicate variants from imputed SNPs
      plink --bfile Chr${CHROMOSOMES}_Imputed --list-duplicate-vars  --out Chr${CHROMOSOMES}_Imputed_Duplicates
      plink --bfile Chr${CHROMOSOMES}_Imputed --exclude Chr${CHROMOSOMES}_Imputed_Duplicates.dupvar --make-bed --
out Chr${CHROMOSOMES}_Imputed_noDups


      # 3. remove badly imputed SNPs using .info average call rate
      # we plotted the density (frequency) of call rates in R - see AvgCallRate.sh
      # this suggests we can be as stringent as call rate > 0.95 and not really lose anything above call rate > 0.4
(graph saved on external hard disc)
      awk '{if ($7<=0.4) print $1}' chr${CHROMOSOMES}.info > Chr${CHROMOSOMES}_PoorImpQualSNPs.txt
      plink --bfile Chr${CHROMOSOMES}_Imputed_noDups --exclude Chr${CHROMOSOMES}_PoorImpQualSNPs.txt --make-bed --
out Chr${CHROMOSOMES}_Imputed_noDups_GoodImp

       # 4.move and rename pre-imputation plink files over for comparison checks
      cp
/users/k1343761/brc_scratch/sickle_new/PreMichiganImpProcessing_SHAPEIT/Chr${CHROMOSOMES}_Phased0.15_AFR_QC.bed
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit
      cp
/users/k1343761/brc_scratch/sickle_new/PreMichiganImpProcessing_SHAPEIT/Chr${CHROMOSOMES}_Phased0.15_AFR_QC.bim
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit
```

```
      cp
/users/k1343761/brc_scratch/sickle_new/PreMichiganImpProcessing_SHAPEIT/Chr${CHROMOSOMES}_Phased0.15_AFR_QC.fam
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit

      mv
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit/Chr${CHRO
MOSOMES}_Phased0.15_AFR_QC.bed
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit/Chr${CHRO
MOSOMES}_PreImputation.bed
      mv
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit/Chr${CHRO
MOSOMES}_Phased0.15_AFR_QC.bim
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit/Chr${CHRO
MOSOMES}_PreImputation.bim
      mv
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit/Chr${CHRO
MOSOMES}_Phased0.15_AFR_QC.fam
/users/k1343761/brc_scratch/sickle_new/PostMichiganImpProcessing_SHAPEIT/AllChr/1000G_Ph3v5_MIXED_shapeit/Chr${CHRO
MOSOMES}_PreImputation.fam

      # 5. compare pre and post imputation data using plink's --merge function
      # i. merge the two datasets two create a new dataset of pre-imputation values where available, plus imputed
SNPS --merge-mode 7 which is a "diff" merge mode (ie don't actually perform merge, report mismatching non-missing
calls only)
      plink --bfile Chr${CHROMOSOMES}_PreImputation --bmerge Chr${CHROMOSOMES}_Imputed_noDups_GoodImp.bed
Chr${CHROMOSOMES}_Imputed_noDups_GoodImp.bim Chr${CHROMOSOMES}_Imputed_noDups_GoodImp.fam  --merge-mode 7 --out
Chr${CHROMOSOMES}_merge


      #ii. Merge fails - we strand aligned (with 1000G) and phased (with 1000G genetic map). Michigan server will
fail pre-imputation QC if th3ere are more than 100 obvious strand flips
      #for safety, I am deleting the merge fails
      plink --bfile Chr${CHROMOSOMES}_PreImputation --exclude Chr${CHROMOSOMES}_merge.missnp --make-bed --out
Chr${CHROMOSOMES}_PreImputation_NoMergeFails
```

```
      plink --bfile Chr${CHROMOSOMES}_Imputed_noDups_GoodImp --exclude Chr${CHROMOSOMES}_merge.missnp --make-bed --
out Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails
      plink --bfile Chr${CHROMOSOMES}_PreImputation_NoMergeFails --bmerge
Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.bed Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.bim
Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.fam  --merge-mode 7 --out Chr${CHROMOSOMES}_merge2

      #iii. inspect differences in SNP calls in Chr${CHROMOSOMES}_merge2.diff
      # review SNPs that have different SNP calls - include a count of number of individuals with different SNPs
      awk '{print $1}' Chr${CHROMOSOMES}_merge2.diff | sort | uniq --count | sort -k1 -n -r  >
Chr${CHROMOSOMES}_merge2_uniq.diff
      #This list of differences is important for future QC: Chr${CHROMOSOMES}_merge2_uniq.diff contains SNPs with
mismatches in reverse order of mismatch frequency. This is not an issue for SNPs where we have MEGA data (ie these
very SNPs), but what about surrounding imputed SNPs?

      #6. Perform actual merge to keep original (MEGA chip) calls and add in imputed SNP calls not found in
original dataset (merge mode 2 with first file as pre-imputation)
      plink --bfile Chr${CHROMOSOMES}_PreImputation_NoMergeFails --bmerge
Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.bed Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.bim
Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.fam  --merge-mode 2 --out Chr${CHROMOSOMES}_FINAL
      # check that this has been merged correctly for your data

      #7. Write this post imputation QC to a log file Chr${CHROMOSOMES}_QC.log
      #review concordance rates
      numberOfPreImputationSNPs=$(wc Chr${CHROMOSOMES}_PreImputation_NoMergeFails.bim | awk '{print $1}')
        numberOfImputedSNPs=$(wc Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.bim | awk '{print $1}')
        numberOfPreImputationSNPs=$(wc Chr${CHROMOSOMES}_PreImputation.bim | awk '{print $1}')
        numberOfImputedSNPs=$(wc Chr${CHROMOSOMES}_Imputed.bim | awk '{print $1}')
      numberOfDuplicatesPlus1=$(wc Chr${CHROMOSOMES}_Imputed_Duplicates.dupvar | awk '{print $1}')
      numberOfDuplicates=$(($numberOfDuplicatesPlus1-1))
      numberOfBadlyImputedSNPS=$(wc Chr${CHROMOSOMES}_PoorImpQualSNPs.txt | awk '{print $1}')
      numberOfImputedGoodSNPsPreMerge=$(wc Chr${CHROMOSOMES}_Imputed_noDups_GoodImp.bim | awk '{print $1}')
      concordanceInfo=$(tail -n4 Chr${CHROMOSOMES}_merge2.log | head -n2)
      diffAllIndivs=$(wc Chr${CHROMOSOMES}_merge2.diff | awk '{print $1}')
       diffSNPsOnly=$(wc Chr${CHROMOSOMES}_merge2_uniq.diff | awk '{print $1}')
      awk '{if ($1>100) print $2}' Chr${CHROMOSOMES}_merge2_uniq.diff > Chr${CHROMOSOMES}_merge2_over100.txt
```

```
      diffSNPsOver100=$(wc Chr${CHROMOSOMES}_merge2_over100.txt | awk '{print $1}')
        awk '{if ($1>50) print $2}' Chr${CHROMOSOMES}_merge2_uniq.diff > Chr${CHROMOSOMES}_merge2_over50.txt
        diffSNPsOver50=$(wc Chr${CHROMOSOMES}_merge2_over50.txt | awk '{print $1}')
        numberOfSNPsPostMerge=$(wc Chr${CHROMOSOMES}_FINAL.bim | awk '{print $1}')

      echo "QC checking post imputation for chromosome ${CHROMOSOMES}:
      Number of SNPs pre-imputation: $numberOfPreImputationSNPs
      Number of SNPs post-imputation: $numberOfImputedSNPs
      Number of duplicates: $numberOfDuplicates
      Number of badly imputed SNPs ie imputation call quality <=0.95: $numberOfBadlyImputedSNPS (little difference
between using 0.4 and 0.95 as cutoff hence higher stringency)
      Number of post-imputation SNPs after removing duplicates and badly imputed SNPs, pre-merging with pre-
imputation genotypes: $numberOfImputedGoodSNPsPreMerge
      Concordance info: $concordanceInfo
      All differences in SNP calls between pre- and post-imputation: $diffAllIndivs
      Differences in unique SNPs between pre- and post-imputation: $diffSNPsOnly
      Number of SNPs with differences in more than 100 individuals: $diffSNPsOver100
        Number of SNPs with differences in more than 50 individuals: $diffSNPsOver50
      Number of SNPs Post Merge in final file: $numberOfSNPsPostMerge
      " > Chr${CHROMOSOMES}_QC.log



      #8. remove large files BUT ONLY AT END
        rm Chr${CHROMOSOMES}_Imputed_Duplicates.*
        rm Chr${CHROMOSOMES}_Imputed_noDups.*
        rm Chr${CHROMOSOMES}_Imputed_noDups_GoodImp.*
        rm Chr${CHROMOSOMES}_Imputed_noDups_GoodImp_NoMergeFails.*
      rm Chr${CHROMOSOMES}_PreImputation_NoMergeFails.*
       rm Chr${CHROMOSOMES}_merge.*
       rm Chr${CHROMOSOMES}_merge2.*

#        done
```

Appendix 8
PostImputationAnalysis.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -q HighMemLongterm.q,LowMemLongterm.q
#$ -M kate.gardner@doctors.org..uk
#$ -m beas
#-l h_vmem=40G
##########


 module add bioinformatics/plink2/1.90b3.38
 module add bioinformatics/htslib/1.3

      #1. merge Chr to create 1 file
      plink --bfile Chr1_FINAL --merge-list IndivChrFiles.txt --make-bed --out Sickle_Imputed

      #2. Post imputation QC
      #a. update FID (vcf can't deal with FID so all FID are IID, to return to known clinical not genetic FID)
      plink --bfile Sickle_Imputed --update-ids FIDIIDrecode2.txt --make-bed --out Sickle_Imputed_updateFID

      #b. reinsert original sex (removed during imputation file conversion)
      plink --bfile Sickle_Imputed_updateFID --update-sex OriginalSex.txt --make-bed --out
Sickle_Imputed_updateFID_sexIncl

      #c. correction to sex (should have happened before imputation but didn't)
            #i. update sex: correct 4 with wrong sex attributed (should have happened pre imputation but didn't)
            plink --bfile Sickle_Imputed_updateFID_sexIncl --update-sex CorrectSex.txt --make-bed --out
Sickle_Imputed_updateFID_sexIncl_updateSex

            #ii. remove sex mismatches from imputed samples (should have happened pre imputation but didn't)
```

```
           plink --bfile Sickle_Imputed_updateFID_sexIncl_updateSex --remove SexMismatches.txt --make-bed --out
Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch

     #d. QC using strict cutoffs (as per UK Biobank criteria). Also do in this order so only removing samples (if
any) at the end
     plink --bfile Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch --geno 0.02 --make-bed --out
Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch_geno0_02
     plink --bfile Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch_geno0_02 --maf 0.01 --make-bed --out
Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch_geno0_02_maf0_01
       plink --bfile Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch_geno0_02_maf0_01 --hwe 0.0000001 --
make-bed --out Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch_geno0_02_maf0_01_hwe10-7
       plink --bfile Sickle_Imputed_updateFID_sexIncl_updateSex_NoSexMismatch_geno0_02_maf0_01_hwe10-7 --mind 0.01
--make-bed --out Sickle_Imputed_QC_strict
```

Chapter 4: Genome-Wide Association Studies in Sickle Cell Disease

Figures 129

Tables 129

4.1. Introduction 130
4.1.1. Background on genome wide association studies 130
4.1.2. Tagging and linkage disequilibrium 131
4.1.3. Selection of phenotypes for association studies 131
4.1.4. Statistical tests 131
4.1.5. Consideration of population ethnicity 132
4.1.6. Confounding in GWAS 133
4.1.7. Population structure and relatedness 133
4.1.8. Managing relatedness 135
4.1.8.1. Background 135
4.1.8.2. Study design 135
4.1.8.3. Family studies 135
4.1.8.4. Genomic control 136
4.1.8.5. Principal components analysis 136
4.1.8.6. Regression analysis using linear mixed modelling 138
4.1.8.7. Assessing a model: Lambda GC and QQ plots 139
4.1.9. Power and sample size considerations 140

4.2. Methods 141
4.2.1. Genotyping 141
4.2.2. Phenotyping 141
4.2.3. Statistical analysis 141
4.2.3.1. Scripting in Linux 141
4.2.3.2. Significance levels in GWAS 141
4.2.3.3. Checks on positive signals 142
4.2.3.4. Replication 142

4.3. Results 142
4.3.1. GWAS analysis scripts: a resource for genetic association analysis traits in sickle cell disease using linear mixed modelling 142
4.3.2. Fetal haemoglobin (HbF%) as phenotype 144
4.3.2.1. HbFg project 147
4.3.3. Hospitalisation rate as phenotype 147
4.3.4. Haemolytic index as phenotype 150
4.3.5. Mortality/survival as phenotype 153
4.3.6. Urinary albumin creatinine ratio as phenotype 153

4.4. Conclusions 156

References 156

Appendix 1 159

Appendix 2 160

Appendix 3 162
Appendix 4 163
Appendix 5 184
Appendix 6 184

Figures

Figure 1 GWAS studies to July 2017, from https://www.ebi.ac.uk/gwas..... 130
Figure 2 Models for statistical tests: panel (a) case-control study (b) quantitative trait study 132
Figure 3 Confounding in genotype / phenotype association studies: what if ethnicity is a confounder, associated with both the phenotype and the genotype?..... 133
Figure 4 Evidence of population stratification: different populations (red/blue) in cases versus controls 134
Figure 5 European population structure, from Novembre (2008) 137
Figure 6 QQ plot demonstrating genomic inflation..... 140
Figure 7 Linear mixed modelling script interface for user to answer questions. This means the parameters are decided at the command line without the user having to alter the code. Parameters that the user can modify (after answering the questions) include: patient population (e.g. HbSS, HbSC, ALL, nonHbSS); relatedness cut-off; outcome (phenotype) name; whether imputed or chip (raw) genotype data to be used; upper and lower age limits to consider..... 143
Figure 8 Demographic details for the HbF% GWAS (N=690): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of HbF% values 145
Figure 9 Linear mixed model results for Ln(HbF%): (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome 146
Figure 10 Demographic details for the hospitalisation rate GWAS (N=354): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of hospitalisation rate per year 147
Figure 11 Linear mixed model results for hospitalisation rate: (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome..... 149
Figure 12 Demographic details for the haemolytic index GWAS (N=321): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of haemolytic index values 150
Figure 13 Linear mixed model results for the haemolytic index: (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome..... 152
Figure 14 Demographic details for the uACR GWAS (N=326): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of urinary albumin creatinine ratios..... 153
Figure 15 Linear mixed model results for uACR: (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome 155

Tables

Table 1 Suggestive novel loci associated with hospitalisation rate for all sickle genotypes (A1 is the effect allele)..... 148
Table 2 Suggestive novel loci associated with haemolytic index (averaged over 10 years) for all sickle genotypes (A1 is the effect allele)..... 151
Table 3 Suggestive novel loci associated with UACR (averaged over 10 years) for all sickle genotypes (A1 is the effect allele) 154

4.1. Introduction

4.1.1. Background on genome wide association studies

Genome-wide association studies (GWAS) analyse the association between many millions of genetic markers (across the whole genome) with a "phenotype" or "trait". As such, the GWAS approach is a "hypothesis-free" search for positive signals: it involves an unbiased scan of the whole human genome. In addition to detecting direct association between genes and traits, GWAS can also reveal unsuspected interactions of several loci affecting the same outcomes (Manolio, 2013). GWAS data will also serve to confirm previous findings if the association is robust (Menzel et al., 2007, Milton et al., 2012). GWAS have transformed our understanding of the underlying causes of common diseases and complex phenotypes. A case in point is the application of GWAS in the β-globin field: the unexpected discovery of *BCL11A* (an oncogene that, hitherto, was not known to have a role in erythropoiesis) as a quantitative trait locus (QTL) controlling HbF% (Menzel et al., 2007, Uda et al., 2008). GWAS also confirmed association of the other two loci – *Xmn*1-*HBG2* (*rs782144*) on chromosome 11p and *HBS1L-MYB* (HMIP) on chromosome 6q – with HbF%, QTLs that were previously discovered through candidate gene (Labie et al., 1985b) and genetic linkage studies (Craig et al., 1996). Similarly, GWAS has confirmed the association between bilirubin level and *UGT1A1* polymorphism in SCD (Milton et al., 2012).

In July 2017, the GWAS Catalogue contained 3043 publications and 38708 unique genetic variant-trait associations ($p \leq 5 \times 10^{-8}$) for 17 trait categories (http://www.ebi.ac.uk/gwas/), see Figure 1.



Figure 1 GWAS studies to July 2017, from https://www.ebi.ac.uk/gwas

I performed genome-wide analysis of multiple phenotypes of sickle cell disease in individuals from South London, United Kingdom, to identify genetic variants associated with disease severity in mixed African and African-Caribbean population. I used mixed linear model association analysis (LMM): this procedure involves using a genomic relationship matrix (GRM) from genome-wide data as a random effect in the mixed linear model to model relatedness / population structure of the cohort. The ability of the GWAS to replicate previous findings was also assessed.

4.1.2. Tagging and linkage disequilibrium

The premise of genetic association studies is to test whether different alleles of a gene are associated with a trait: in case-control studies, whether one allele more frequent among cases compared to controls, and in quantitative trait studies, whether trait values are higher among carriers of one allele than another. Significant association can mean that this locus itself has a direct effect on the disease or trait, or that this marker is physically near, or "tags", the DNA variant that has a direct effect on the disease or trait.

4.1.3. Selection of phenotypes for association studies

A phenotype is suitable for association analysis if it is both *variable* and *heritable*. Continuous phenotype data offer more statistical power than those obtained by re-coding observations into binary or categorical variables. A mathematical transformation of phenotype data (e.g. log-transformation of positively skewed data) may be appropriate to create as "normal" a trait distribution as possible.

4.1.4. Statistical tests

GWAS are similar to randomised controlled trials or experiments in principle, but there are some key differences: (i) the *number* of tests, (ii) the *model* of association, and (iii) the importance of *quality control* checks (discussed in chapter 3).

The number of tests performed depends, among other factors, on the number of genetic markers tested, which now runs into millions in the era of large scale imputation. Statistically, repeat testing has implications for setting genome-wide significance thresholds (see section 4.2.3.2). On a practical level, there are also computational efficiency implications: most GWAS are now run on computer server clusters, rather than single machines.

There are multiple models of association in genetic association studies. Data can be analysed as genotypes (counting AA, Aa, aa), as alleles (counting A or a), or using a dominant or recessive mode of inheritance (Clarke et al., 2011), see Figure 2.

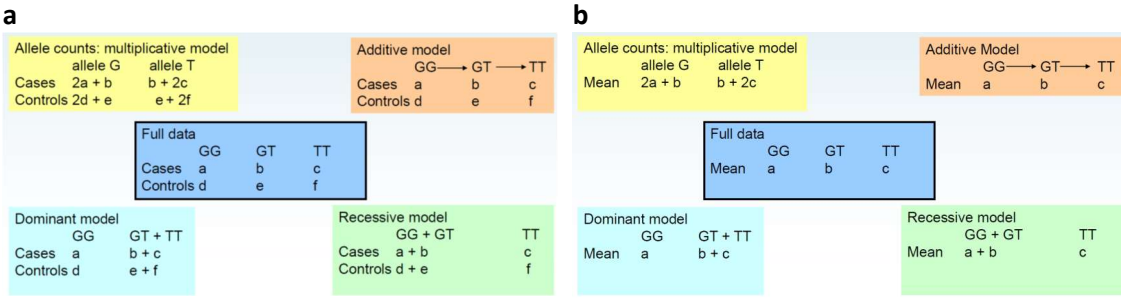


Figure 2 Models for statistical tests: panel (a) case-control study (b) quantitative trait study

The differences between these analysis methods are important enough to severely affect the ability to detect an association (significant *versus* non-significant results). Since the mode of inheritance for the trait under study is usually unknown, in GWAS, either a genotypic or allelic model is applied.

Regression models (whether linear or logistic) allow covariates to be built in, to control for a variety of parameters – fixed effects related both to the individual (age, sex) or random effects related to the population (relatedness, via principal components or a GRM, see below).

4.1.5. Consideration of population ethnicity

Genetic heterogeneity within and between populations is a major issue in genotype-phenotype association studies. Genetic variation can be specific to an ethnic group or unrecognised sub-groupings within known ethnic entities. The failure to deal with the ethnic make-up of a study population can lead to either false-positive findings (due to population stratification) or to loss of power (i.e. failure to capture genetic variability). Until recently, barriers to performing GWAS in non-Europeans populations included the lack of appropriate micro-arrays (genotyping chips) that capture ethnicity-specific genetic variability, scarce reference population data, and sub-optimal analysis procedures.

Most previous genotyping arrays used in GWAS had made use of data based on European populations. Using ethnically-relevant assay systems is particularly important with our patients, all but one of whom are of African or African-Caribbean heritage. The advent of Illumina's MEGA chip and other "African" weighted arrays (which include African-specific genetic variants) overcomes this.

The lack of suitable reference panels meant a failure to accommodate the genetic characteristics of African populations. With the release of 1000 Genomes Project phase 3 (which contains seven African-heritage populations), this barrier has been removed.

To date, studies in European populations (which includes European Americans) significantly outnumber those in other populations. African populations have been particularly neglected, hampering research not only into diseases prevalent in Africa, but also preventing the study of the increased burden for cardiovascular disease, type 2 diabetes mellitus and renal disease among members of the African diaspora in Europe. Sickle cell disease in the UK is a condition of individuals of African heritage. Therefore, there is considerable scope for identifying the genetic component to severity of sickle cell disease in African heritage populations.

4.1.6. Confounding in GWAS

Compared to other studies, GWAS are subject to fewer sources of bias: genotypes don't change in an individual's lifetime. There are two key sources of confounding in GWAS: *technical artefacts* and *relatedness*. Technical artefacts arise from sample handling or lab processing that are correlated with phenotype e.g. case samples are stored differently or processed differently to control samples. This can also, indirectly, affect quantitative trait studies such as ours. Our careful genotyping and phenotyping on the same platforms (chapters 2 and 3) together with strict quality control reduce the likelihood of technical artefacts.

Population stratification, or confounding by genetic ancestry, is a key cause of false positive results in genotype-phenotype association studies. In statistics, confounding is a type of bias causing spurious or distorted findings caused by a correlation between a third variable (the confounding variable) and both the exposure variables (e.g. the genotype) and the outcome variable (e.g. quantitative trait, case-control status), see Figure 3. In genetic association studies, the concern is that recognised or hidden ethnic heterogeneity could represent a confounding factor.

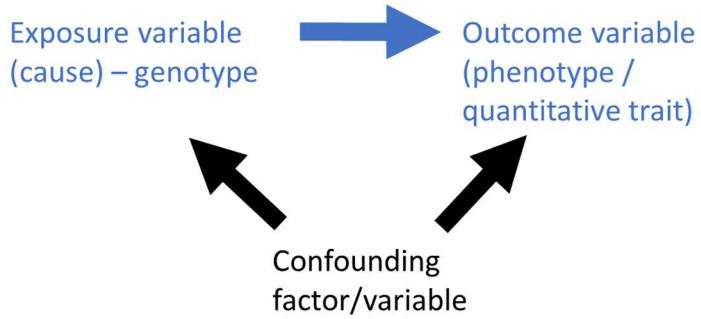


Figure 3 Confounding in genotype / phenotype association studies: what if ethnicity is a confounder, associated with both the phenotype and the genotype?

4.1.7. Population structure and relatedness

Population structure can generate spurious genotype–phenotype associations. Population stratification implies some degree of population substructure in the cohort under analysis. For example, sleeping sickness (*Trypanosoma brucei*) occurs more frequently in Africans than in

Europeans. In a hypothetical case-control study of sleeping sickness, which included European and African populations, false-positive association will occur at genetic markers that differ in genotype frequencies between the two subpopulations (i.e. markers of ethnicity not the disease under investigation), see Figure 4. This is because cases are drawn preferentially from the African sub-population which also has a higher frequency of the variant and so the difference in genotypes between cases and controls reflects their population origin, *not* their disease status. Population stratification only arises if: (1) different proportion of cases/controls are from each population *and* (2) populations differ in allele frequency. Population stratification can cause both false positive or false negative results.

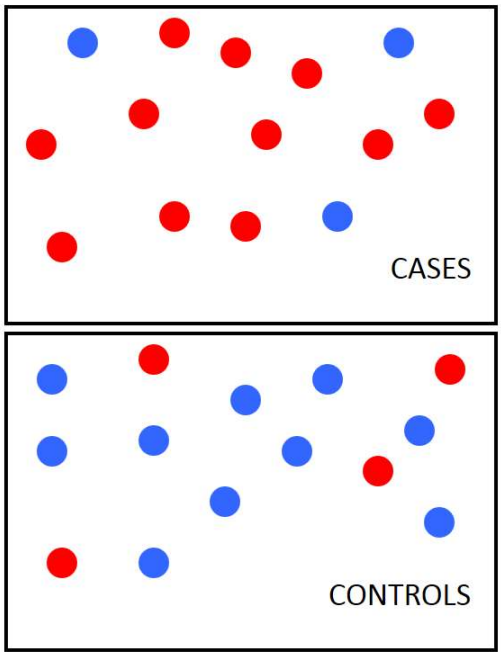


Figure 4 Evidence of population stratification: different populations (red/blue) in cases versus controls

Even a small degree of population stratification can adversely affect a study due to the large sample sizes required to detect common variants underlying most complex diseases(Risch and Merikangas, 1996). Historically, population geneticists have been sceptical of many case-control studies for this reason.

Cryptic relatedness occurs when pairs or groups of individuals are more closely related to each other than the population average – thus indicating they are family members. Individuals that are closely related need to be accounted for in association analyses as they induce unforeseen correlation structures. Cryptic relatedness may introduce false positive and false negative results.

The terms “population structure” and “cryptic relatedness” can together be re-conceptualised as *degrees of relatedness*. The example of different sub-populations within a study cohort are good way to describe patterns of (distant) relatedness. The problem of spurious associations arises if cases are on average *more closely related* with each other than with controls. Thus, it is not a problem of population structure but one of *degree of relatedness*, and this insight is more accurate, and also leads to more precise, more powerful analysis approaches.

4.1.8. Managing relatedness

4.1.8.1. Background

Relatedness has been considered in genetic association studies for decades, and strategies to manage this have become more sophisticated over time. All modern solutions to relatedness require some degree of genome-wide genotyping in individuals, in addition to the candidate genes. Or, in GWAS, one can make use of this genome-wide data to estimate relatedness and then use this as part of the association analysis.

4.1.8.2. Study design

An historical solution to population stratification was to allow for structure during study design, by matching cases and controls for ethnic group (so when a case is selected of given ethnicity, a matched control is selected of the same ethnicity). However, there may be fine-scale structure within ethnic groups or population admixture that cannot be accounted for by matching. This is a particular issue in our cohort because of: (a) multiple tribes in West African populations (b) the issue of European admixture in our African Caribbean population. This issue of admixture is exemplified by the frequently cited case of an initially apparent association between variants and type 2 diabetes in Pima Indians. Type 2 diabetes occurs with greater prevalence in Pima Indians than European heritage individuals. The association identified was in fact due to population admixture: cases tended to have a lower proportion of Caucasian ancestry, and allele frequencies of the confounding genotype vary between the ancestral populations(Knowler et al., 1988).

4.1.8.3. Family studies

The problem of population structure can be eliminated by collecting family data and this approach was adopted historically. Family-based association studies ascertain affected cases and their unaffected parents. The parents form “internal” controls from alleles not transmitted from the parents to the child, effectively matching for ancestry. This is much less powerful since two parents are required to form a single matched control. Furthermore, parental data may not always be available, especially for late-age onset diseases. The family study approach for genetic association studies is mostly obsolete now, since one can make use of genome-wide data to infer and adjust for population ancestry.

4.1.8.4. Genomic control

Genomic control was the first quantitative method to control relatedness; it was developed by Devlin and Roeder in 1999 (Devlin and Roeder, 1999). In this method, the (genome-wide) non-candidate gene variants (the "null" variants) are assessed, a theoretical λ_{GC} is calculated to estimate the genomic inflation (across the genome) in χ^2 statistics as: λ_{GC} = median(χ^2)/0.455. Then all χ^2 tests in analysis are corrected by dividing by λ_{GC} i.e. λ_{GC} corrects for genomic inflation. Few (if any) of the null variants will be associated with the phenotype, so if λ_{GC} >1, this is likely to be due to population stratification, and dividing by λ_{GC} cancels this effect for the candidate variants. GC performs well under many but not all scenarios(Marchini et al., 2004). Notably, it corrects only for false positive results, not false negative results.
While using genomic control methods has been superseded by the strategies below, for each model, λ_{GC} can be calculated and used to assess if genomic inflation is present.

4.1.8.5. Principal components analysis

Principal component analysis (PCA) is a data reduction technique used widely in statistics and introduced in chapter 2. In the last ten years, PCA has become a standard tool in genetics to study ethnic variation(Patterson et al., 2006, Price et al., 2006). In genetics, principal components reflect the genome-wide variability i.e. genetic ancestry(Price et al., 2006). PCA calculates axes of genetic variation that maximise the variability between individuals. Plotting principal components including reference panel samples can be used to identify population outliers. A subset of the PC can then be used in association studies as covariates as they represent ethnic variation between individuals within the study population i.e. PCs "correct for" finer scale population structure within the cohort. In the last 10 years, PC have been widely used in genome-wide association studies.

European population structure mirrored by genetics (PC) was elegantly shown by Novembre(Novembre et al., 2008), see Figure 5 for the graph of principal components 1 and 2 (rotated) reflecting the geographical origins of individuals (1387 samples, ~200,000 variants).

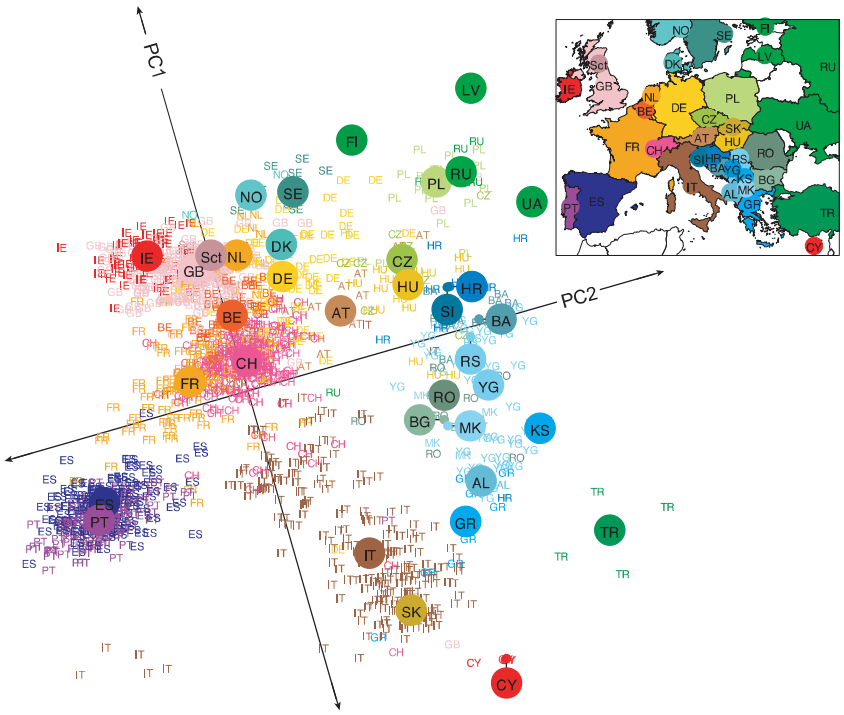


Figure 5 European population structure, from Novembre (2008)

A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasise the similarity to the geographic map of Europe. AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia.

PCs often do not separate out the finer population substructure. For example, our population contained a majority of individuals of West African heritage and minority of African-Caribbean heritage. These two separate ancestry groups will be separated, but within the majority West African group, it is difficult to separate the substructure (e.g. tribal heritage) by PCs. One approach could be to re-run the analysis on the West African only subgroup, with PC reflecting this group West African-specific principal components for use in the final analysis.

In practice, PCs are estimated based on genotyped, independent variants i.e. the LD-pruned dataset chapter 3. When using PCA to detect ancestry the observations are the individuals and the genetic variants are the potentially correlated variables. PCA techniques are computationally efficient and can be applied in the context of whole genome association studies.

PCA was performed in GCTA (Yang et al., 2011) on a pruned dataset for linkage disequilibrium (Weale, 2010). See

Appendix 1 for the top 10 PCs in pairs.

4.1.8.6. Regression analysis using linear mixed modelling

Most recently, linear mixed modelling (LMM) has emerged as *the* method of choice for conducting genetic association studies (Yang et al., 2014, Kang et al., 2010, Yang et al., 2010, Yang et al., 2011). LMM association analyses in genome-wide data provided a better alternative to control for background genetic similarity between individuals than PCA (Yang et al., 2014).

More generally, in statistics, linear mixed models incorporate both *fixed* and *random* effects as covariates in a regression model for a quantitative outcome. Fixed effects are covariates fixed to the individual (e.g. age, sex, other genotypes). In contrast, random effects are not fixed to the individual but explain the effects from the underlying population (i.e. population structure). In genetics, a LMM tests the effect at a locus (the genotype or allele) by controlling for both fixed effects (factors related to the individual and added to the model as covariates) as well as random factors to explain population structure. The random effect is derived from the genetic relatedness matrix, GRM, see chapter 3. Because the GRM is a quantitative measure of each and every pairwise genetic relationship, it takes account of both near- and far-relatedness.

Not only do LMM methods prevent false positive associations due to relatedness (population structure) but they also increase power because the correction applied (the genetic relatedness matrix, GRM) is specific to this structure. LMM can manage geographic population structure (far relatedness), family relatedness and/or cryptic relatedness (near relatedness) (Wang et al., 2013, Yu et al., 2006, Kang et al., 2010). The GRM models genome-wide sample structure, estimating the contribution of the GRM to phenotypic variance using a random-effects model (with or without additional fixed effects) and computing association statistics that account for this component of phenotypic variance. Each variant is then assessed in the context of the gross genetic similarity between individuals(Kang et al., 2010, Yang et al., 2014). Of note, LMM can also be used to estimate phenotypic heritability explained by genotyped markers(Yang et al., 2010, Zaitlen and Kraft, 2012).

There have been two key modifications to standard LMM approaches. First, the “leave one chromosome out” method. Recent work has shown that inclusion of the candidate variant in the GRM can lead to loss in power(Listgarten et al., 2012), due to “double-fitting” of the candidate variant in the model (both as a fixed effect tested for association and as a random effect as part of the GRM). This phenomenon has been termed “proximal contamination,” and

it has been shown that a LMM excluding the candidate variant works, and increases power(Listgarten et al., 2012, Yang et al., 2014). However, this is computationally expensive; in GCTA it is implemented with the "leave one chromosome out" method so that the whole chromosome is not part of the GRM when assessing markers on that chromosome(Yang et al., 2011). This prevents any effect of the genetic variant of interest being captured by the GRM (thereby reducing the measured effect of the variant). The LOCO analysis is computationally less efficient but more powerful compared with the original analysis which included the candidate. This is because the genetic variance is re-estimated each time a chromosome is excluded from calculating the GRM. The second modification involves the number of markers used to create the GRM. Using a too-small subset of markers in the GRM (e.g. a few thousand only) can also compromise results(Yang et al., 2014). I used the LD-pruned subset of ~98,000 markers to create the GRM.

4.1.8.7. Assessing a model: Lambda GC and QQ plots

For an association analysis, part of model assessment includes calculation of the genomic control parameter λ_{GC} to assess genome-wide inflation of the model, and generation of QQ plots. λ_{GC} is discussed in section 4.1.8.4.

QQ plots are used to evaluate if *p*-values are drawn from a normal distribution between 0 and 1. If the *observed p*-values are ordered from lowest to highest, then the *n*th ordered item should on average be equal to the corresponding *expected p*-values from a uniform distribution. Thus, if the observed p values are plotted against the expected ones, one would expect to see a roughly straight line through the origin with a unit slope, albeit with some random variation. If the *p*-values are broadly skewed away from this expected line (towards a distribution with lower *p*-values) then the line will be "inflated" away from the expected straight line. Log-scaling emphasises these low *p*-values so QQ plots are plotted on a negative logarithmic scale. Potential true hits represent the very low p-values at the top-right of the plot, and genomic inflation (as calculated by λ_{GC}) appears as points raised above the y=x line, see Figure 6.

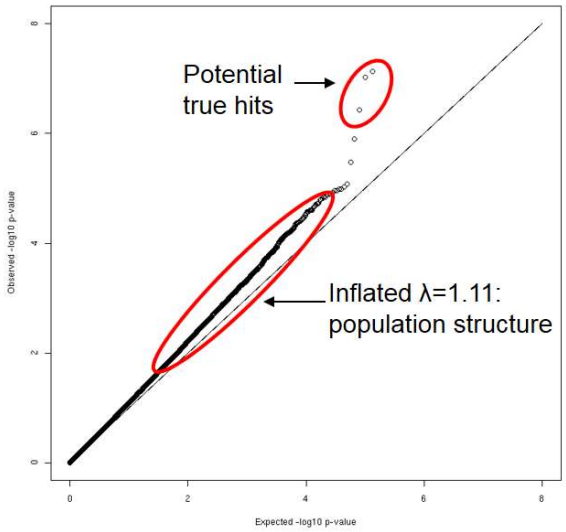


Figure 6 QQ plot demonstrating genomic inflation

Thus, QQ plots and λ_{GC} were used in an iterative process to improve the model for genetic analysis. This led to a progressive improvement in model performance, based on QQ plot and λ_{GC}, in progressively more sophisticated models: from simple linear regression, to linear regression incorporating age and sex as covariates, to adding in first 10 PC and finally the linear mixed model including the GRM. See Appendix 2 where this is demonstrate: Ln(HbF%) is used as the exemplar phenotype.

4.1.9. Power and sample size considerations

The small size of our population means it cannot be expected to identify variants in an hypothesis-free genome-wide analysis for the highly complex (and presumably polygenic) phenotypes of sickle cell disease severity. However, as in previous GWAS of HbF%, an oligogenic trait, we anticipated identifying known and potentially novel "hits" using HbF% as an outcome. SCD is not common in the UK, so, despite linking up with local hospital trusts in a regional network to collect samples, our cohort remains small. For this reason, the main aim of the project was to use the data with "global" and organ-specific severity phenotypes for candidate gene analysis (see chapter 5).

Due to the small size, rather than doing sub-analyses by sickle genotype, I have included all patients of all sickle genotypes in the analyses, in order to maximise sample numbers (hence power) of the study.

I have summarised the considerations to be made in power calculations for GWA studies in Appendix 3.

4.2. Methods

4.2.1. Genotyping

Genotyping was described in detail in chapter 3. In summary, I used Illumina's Infinium MEGA array to get raw variant calls, and imputed the genotypes on the Michigan imputation server using 1000 Genome Phase III data up to over 15 million markers. Data were fully quality controlled prior to analysis.

4.2.2. Phenotyping

Phenotyping was described in detail in chapter 2. I considered fetal haemoglobin (HbF%) and three SCD global severity indices – haemolytic index, hospitalisation rate, mortality – and single organ dysfunction – proteinuria (using urinary albumin creatinine ratio, uACR). For HbF%, only "validated" HbF% values were used (i.e. samples obtained when patients were not on hydroxycarbamide, not transfused for at least 3 months, and not pregnant). For both the haemolytic index and uACR, I used results averaged over the 10 year study period. Traits were normalised prior to analysis using a natural logarithmic transformation (Ln).

4.2.3. Statistical analysis

4.2.3.1. Scripting in Linux

I used the Linux bash commands to generate scripts to automate data handling, and calling of suites of genetic analysis tools and presentation of results.

I used GCTA to create/manipulate the GRM and perform linear mixed modelling (LMM) (Yang et al., 2011). As well as the genetic relatedness matrix as the random effect, I added age, sex and sickle genotype of each individual as a fixed effect (covariate) in the model. GCTA analysis is similar to that implemented in other LMM software tools such as EMMAX, FaST-LMM and GEMMA. I used R (*qqman* package) to perform checks on the model (λ_{GC}, QQ plots) as well as to create the Manhattan plots (www.r-project.org) (Team, 2011).

4.2.3.2. Significance levels in GWAS

There is no absolute significance level for accepting GWAS results (Wellcome_Trust_Case_Control_Consortium, 2007). Interpreting the strength of evidence in an association study depends on several factors: the likely number of true associations, the power to detect them (which, itself depends on both effect sizes and sample size). In a less-well-powered study, more stringent p-value thresholds should be adopted to control the false-positive rate. Because of this, I have presented results tentatively and used the term "suggestive association". Proof is in replication of results in a different cohort (by different genotyping methods). McCarthy compiled best practice guidelines for performing GWAS, and advocated a threshold of 5×10^{-8} for significance in GWAS(McCarthy et al., 2008).

4.2.3.3. Checks on positive signals

Further checks must be made on any identified positive signals of association before moving towards replication of results in other cohorts. The first is to check that the results do not depend solely on one variant: ensure there is some evidence for association at neighbouring variants in linkage disequilibrium. The second is to ensure the minor allele frequency is reasonable i.e. results don't depend upon a handful of individuals.

After these checks, one must refer back to the variant signal intensity plots (chapter 3, figure 4) to inspect for genotype calling problems i.e. that the clusters for each of the three genotypes are well-defined, and therefore that there is confidence in genotyping calling.

4.2.3.4. Replication

Once positive results are checked, the association must be replicated in one or more independent studies. This could be either *in silico* replication using existing GWAS data for the same phenotype in independent samples or *de novo* replication genotype variants of interest in independent samples using a different genotyping method.

Our group has set up collaborations with international researchers in the genetics of SCD in order to replicate our results.

4.3. Results

4.3.1. GWAS analysis scripts: a resource for genetic association analysis traits in sickle cell disease using linear mixed modelling

I wrote a bash script to encapsulate the analysis, image generation and data formation, see Appendix 4. This user-friendly script has been made available to colleagues unfamiliar with Linux/programming.

The script has a user-friendly command line interface to run where a user can stipulate a variety of parameters on request, see Figure 7. This allows users to choose new phenotypes and different subgroups to analyse without having to alter the code.

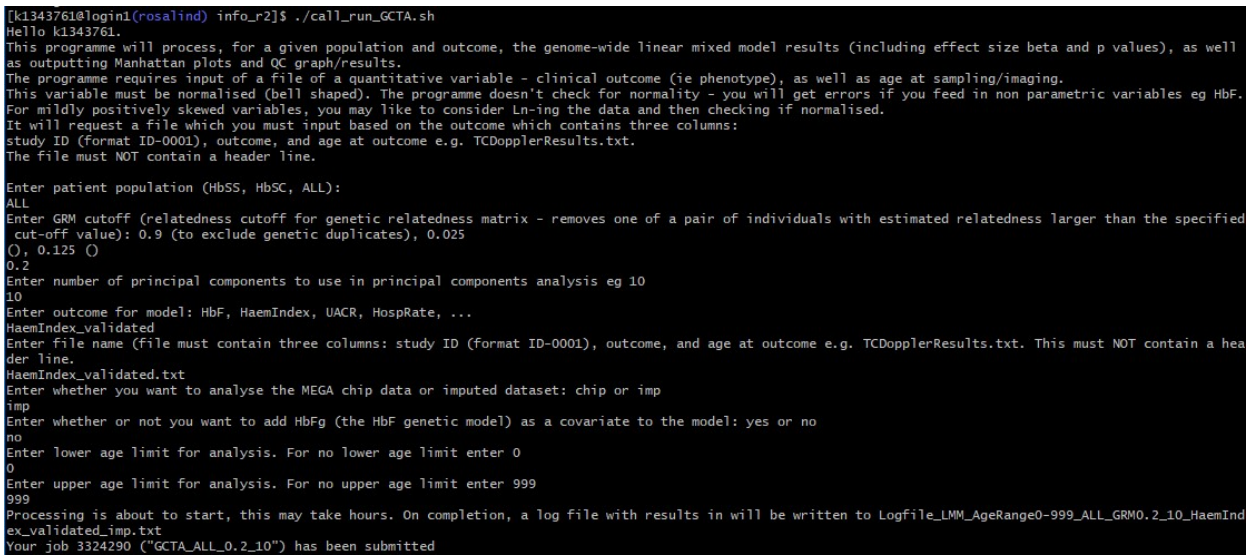


Figure 7 Linear mixed modelling script interface for user to answer questions. This means the parameters are decided at the command line without the user having to alter the code. Parameters that the user can modify (after answering the questions) include: patient population (e.g. HbSS, HbSC, ALL, nonHbSS); relatedness cut-off; outcome (phenotype) name; whether imputed or chip (raw) genotype data to be used; upper and lower age limits to consider

The user provides one file with the trait data in, and chooses various parameters in answer to the questions:

- Outcome (user provides name of a file that they have uploaded which contains phenotypic details, in the format):
    - ID number, quantitative trait value, age at quantitative trait
    - Quantitative trait must be normalised
- Whether to use raw data only, or imputed data (big computational efficiency difference)
- Population (all sickle genotypes versus specific sickle genotype) – if "all" is used, then sickle genotype itself is added as a categorical covariate (implemented as four binary variables HbSS, HbSC, HbSβ^0 thalassaemia and HbSβ^+ thalassaemia).
- Age range (i.e. specify lower and upper age limits)
- Relatedness cut-offs to exclude (also termed the *GRM cut-off*):
    - Expected values
        - 1 for monozygotic twins / duplicated samples
        - 0.5 for first degree relatives (e.g. full-sibs or parent-offspring)
        - 0.25 for second degree relatives (e.g. grandparent-grandchild)
        - 0.125 for third degree relatives (e.g. cousins)
    - These are the *expected* relatedness values; there will be some variation around these numbers. The procedure automatically removes the least number of people to optimise the population size

The script then selects the specific population, modifies the GRM to exclude people more related than the specified relatedness cut-off[1]. LMM is then performed using GCTA, using mlma-loco (see section 4.1.8.6) with user-determined parameters, as above.

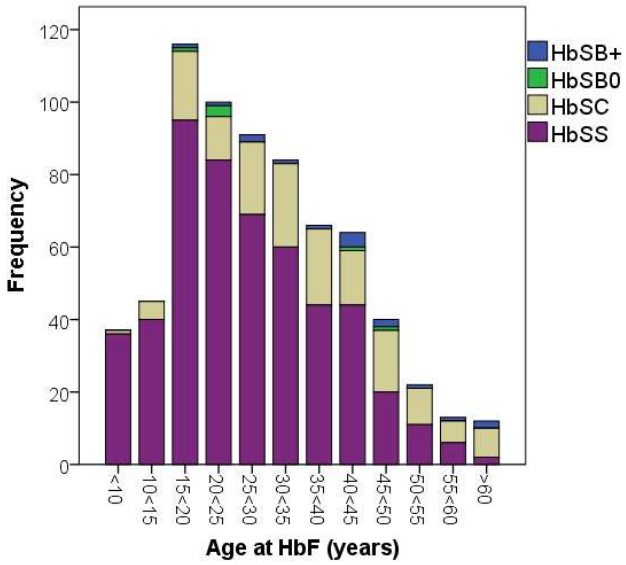
Model evaluation parameters are then generated using R: QQ plots and λ_{GC}. These to allow the user to assess the model, see section 4.1.8.7. A Manhattan plot is also drawn using R, based on non-missing data from the mlma file. The results are written to a log file for the user to interpret, see Appendix 5.

In each analysis, λ_{GC}<1.01, indicating that inflation due to population structure, was well-controlled.

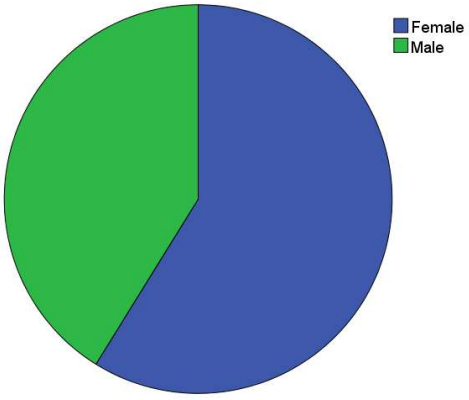
4.3.2. Fetal haemoglobin (HbF%) as phenotype

I performed a genome-wide association study for Ln(HbF%) in a discovery cohort of 690 patients (no duplicate samples, 1st or 2nd degree relatives) with SCD (all sickle genotypes) of African Caribbean or West African heritage. Demographic details are presented in Figure 8, all aged at least 5 years old.

a



b



c

[1] I treated grm-cutoff=0.9 differently. GRM cut-offs are used to exclude people more related than the cut-off. Where GRM_CUTOFF=0.9 we want to exclude patients manually after identifying duplicates/monozygotic twins and choosing appropriately (this might depend on the volume of phenotype data we have on each in the pair, see chapter 3)

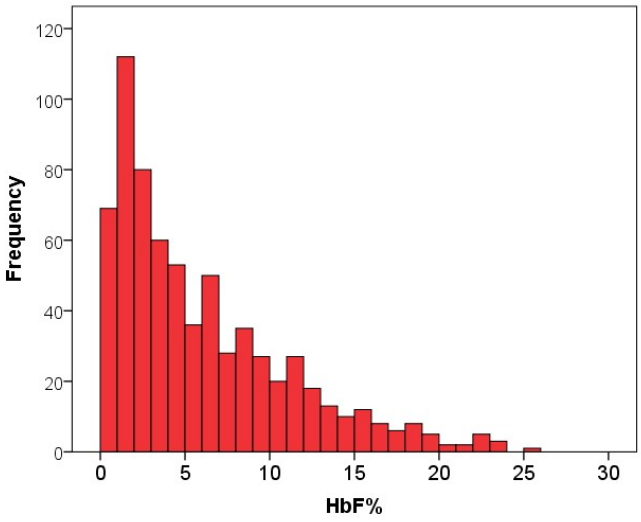


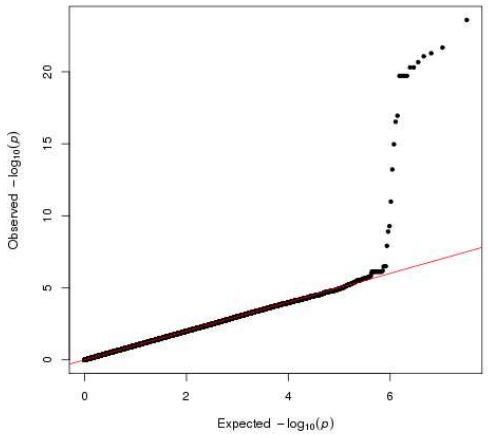
Figure 8 Demographic details for the HbF% GWAS (N=690): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of HbF% values

I used age, sex and sickle genotype as fixed covariates in the linear mixed model. The genomic control (λ_{GC}) for the analysed variants was 0.999118 and a QQ plot of the observed versus expected p-values is shown in Figure 9a. The absence of an early departure of the observed p-values suggests that our data are not affected by problems with genotyping, imputation, and uncontrolled sample relatedness or population stratification. The Manhattan plot (distribution of association p-values) for Ln(HbF%) is shown in Figure 9b.

I have replicated known HbF% modifier loci: *BCL11A* on chromosome 2 and *HMIP* on chromosome 6. The peak *BCL11A* signal was at *rs1427407* (β=0.5158, p=2.526x10^{-24}) plus a second *BCL11A* signal at *rs11692396* (β=0.390053, p=6.02798x10^{-14}). At *HMIP*, the peak signals were *rs116460276* (β=0.959567, p=3.166x10^{-7}) and rs61028892 (β=0.888939 p=6.5453x10^{-7}). *rs66650371* is not in the imputed dataset. In the *HBB* region, the peak signal was *rs10564838* (β=0.229908, p=6.3242x10^{-5}). *rs7482144* (*Xmn*1- *HBG2)* is not in the imputed dataset. Notably, no variants situated at the *KLF1* or *ZBTB7A* loci were significant in our study.

I have detected no novel loci, confirming previous findings that genetic HbF% variability in human populations is dominated by three major loci.

a



b

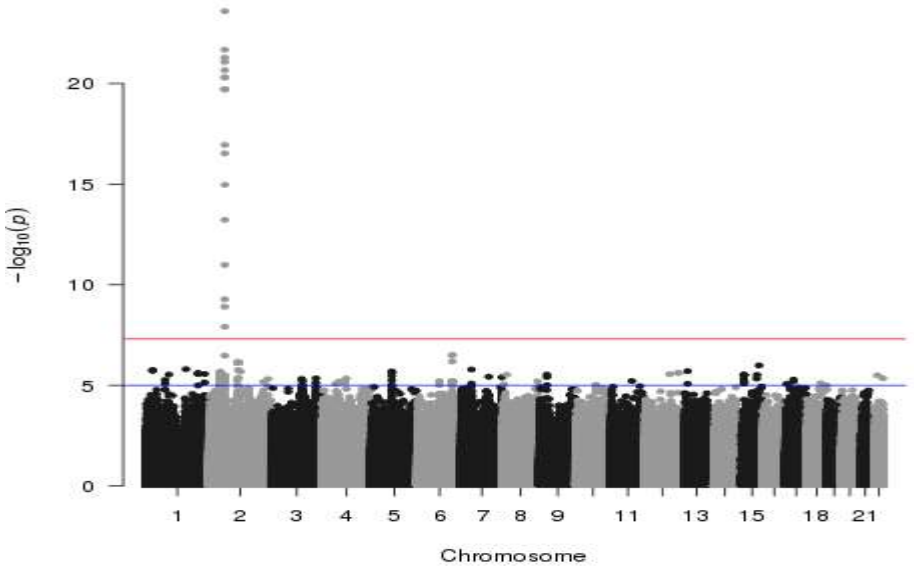


Figure 9 Linear mixed model results for Ln(HbF%): (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome

4.3.2.1. *HbFg project*

I utilised these data to create a summary variable to quantify the genetic component of HbF% based on the three trait loci, see Appendix 6.

4.3.3. Hospitalisation rate as phenotype

I performed a genome-wide association study for hospitalisation rate in a discovery cohort of 354 adult patients (no duplicate samples, 1st or 2nd degree relatives) with SCD (all sickle genotypes) of African Caribbean or West African heritage. Demographic details are presented in Figure 10.

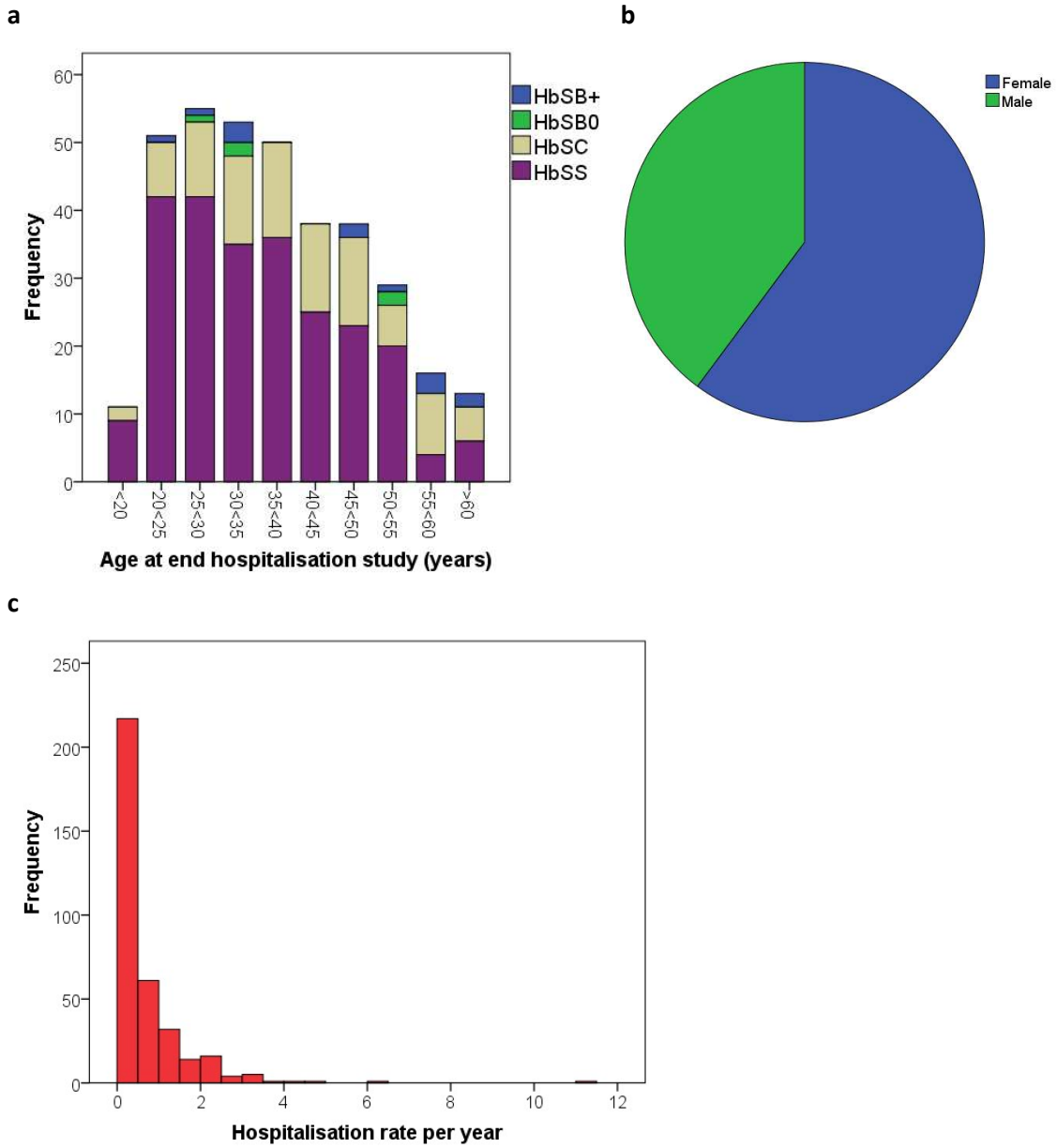


Figure 10 Demographic details for the hospitalisation rate GWAS (N=354): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of hospitalisation rate per year

I used age, sex and sickle genotype as fixed covariates in the linear mixed model. The genomic control (λ_{GC}) for the analysed variants was 0.995384 and a QQ plot of the observed versus expected p-values is shown in Figure 11a. The absence of an early departure of the observed p-values suggests that our data are not affected by problems with genotyping, imputation, and uncontrolled sample relatedness or population stratification. The Manhattan plot (distribution of association p-values) for hospitalisation rate is shown in Figure 11b.

HbF% and HbF% quantitative trait loci have been found to be associated with hospitalisation rate in other cohorts(Wonkam et al., 2014). I have replicated two *HMIP* (chromosome 6) variants that have been associated with hospitalisation rates in a Cameroonian cohort with SCD: both in the *HMIP* region on chromosome 6 known to be a quantitative trait locus for HbF%. HMIP1 (rs28384513 , β=-0.243966, p=2.46346x10^{-05}) and HMIP2 (rs9494142, β=0.279293, p=0.000140198).

There were no significant genome-wide results but I have detected some suggestive novel loci associated with hospitalisation rate, see Table 1. Top loci include a region on chromosome 5 (2913808-2948858) with peak signal at *rs75904749* (β= 0.436632, p=1.36045x10^{-7}); a region on chromosome 11 (80277911-80295720) with peak signal at *rs10792490* (β=0.179749, p=6.79482x10^{-7}); and a region on chromosome 12 (351711-364599) with peak signal *rs510384* (β=0.111138, p=3.90022x10^{-7}). The chromosome 12 region is an intronic site within *SLC6A13* (SLC6A13 is a sodium-dependent GABA and taurine transporter. In presynaptic terminals, it regulates GABA signalling termination through GABA uptake. It may also be involved in beta-alanine transport).

Table 1 Suggestive novel loci associated with hospitalisation rate for all sickle genotypes (A1 is the effect allele)

Chr	Variant	Position (hg19)	A1	A2	Gene	MAF	β-value	p-value
5	rs75904749	2947971	A	G		0.014	0.437	1.360 x10^{-7}
11	rs10792490	80283002	C	T		0.066	0.180	6.795 x10^{-7}
12	rs510384	357319	G	A	SLC6A13	0.247	0.111	3.900 x10^{-7}

All of these were imputed variants, after returning to view the cluster plots for the nearest raw genotyped variants, all were well clustered.

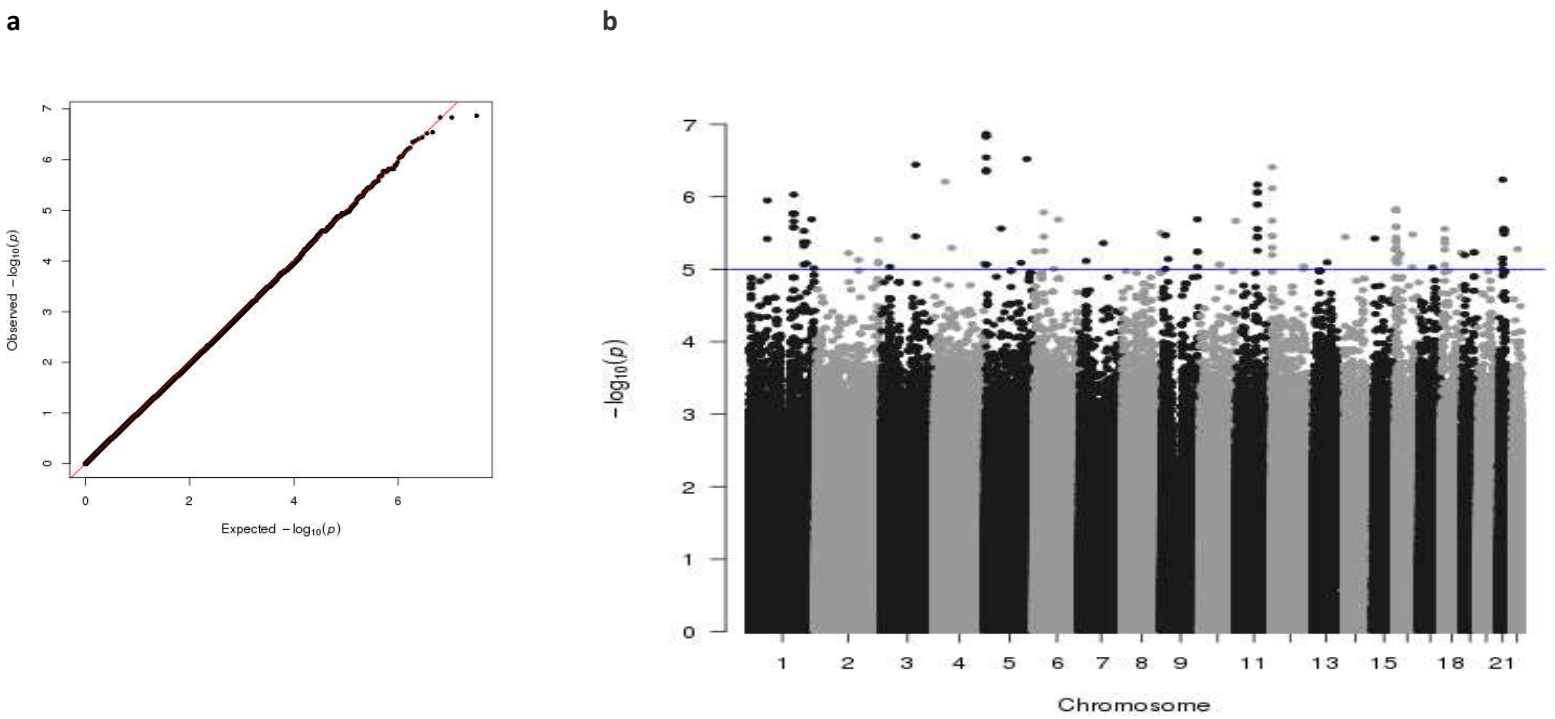


Figure 11 Linear mixed model results for hospitalisation rate: (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome

4.3.4. Haemolytic index as phenotype

I performed a genome-wide association study for the haemolytic index in a discovery cohort of 321 adult patients with SCD (all sickle genotypes) of African Caribbean or West African heritage. Demographic details are presented in Figure 12.

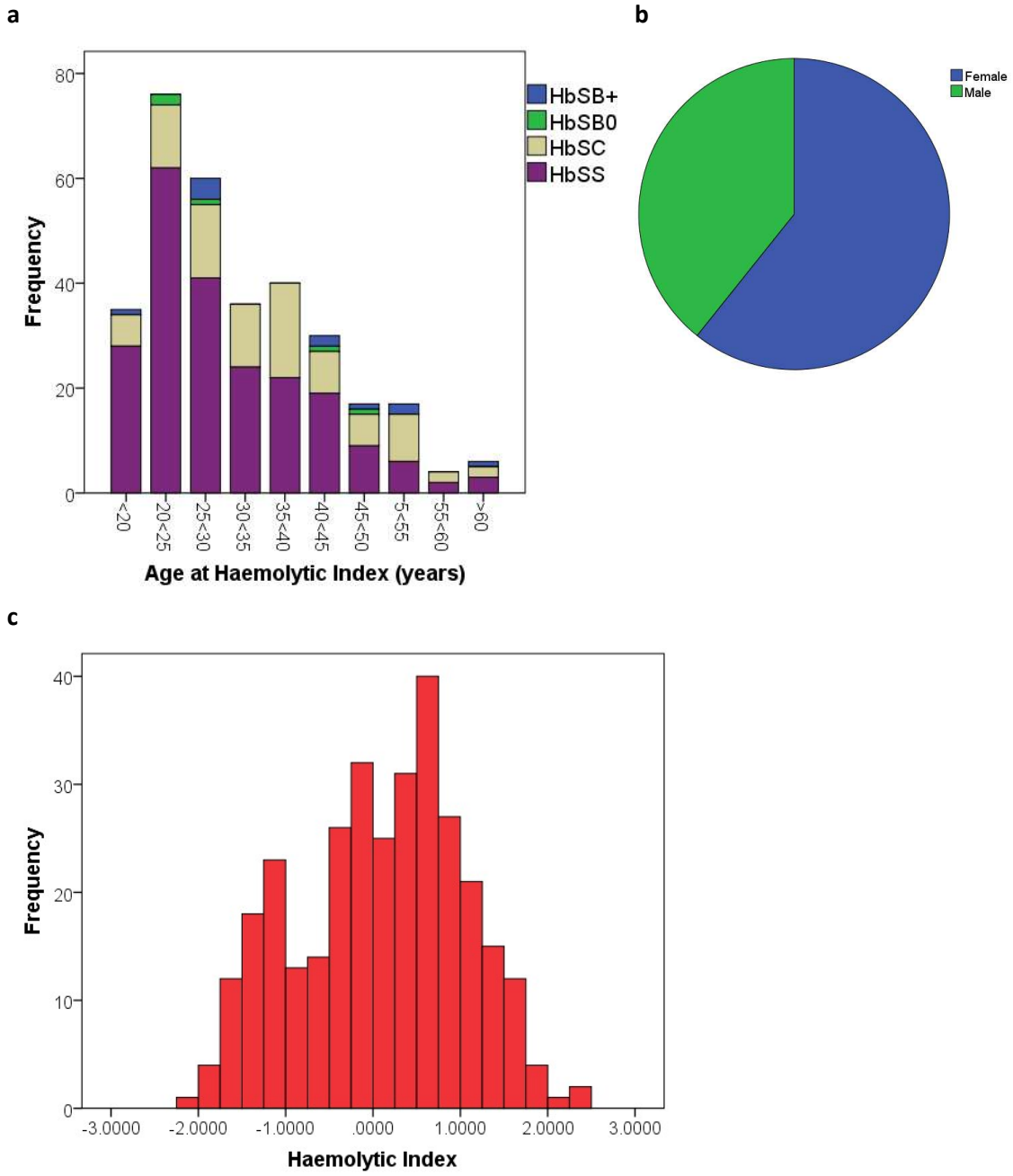


Figure 12 Demographic details for the haemolytic index GWAS (N=321): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of haemolytic index values

I used age, sex and sickle genotype as fixed covariates in the linear mixed model. The genomic control (λ_{GC}) for the analysed variants was 0.983 and a QQ plot of the observed versus expected p-values is shown in Figure 13a. The absence of an early departure of the observed p-values suggests that our data are not affected by problems with genotyping, imputation, and uncontrolled sample relatedness or population stratification. The distribution of association p-

values (Manhattan plot) for haemolytic index is shown in Figure 13b. The low QQ plot suggests the model has somewhat overcorrected for population structure. This gives conservative results (p values not as low in our analysis). The likely cause is low sample size leading to under-powering of analysis.

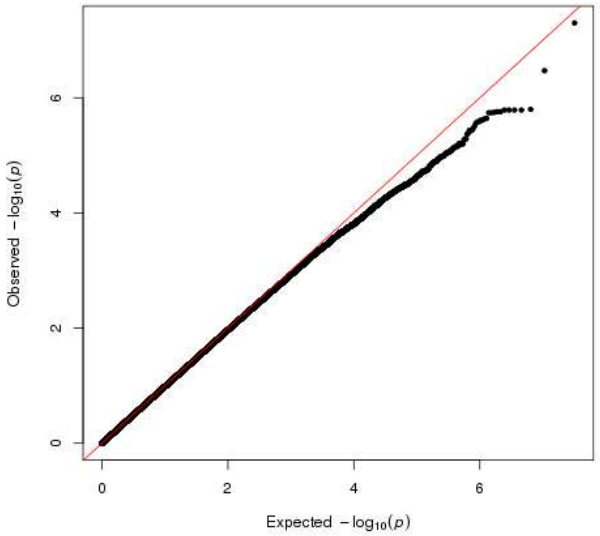
There were no significant genome-wide results but I have detected some suggestive novel loci, see Table 2. The second locus is a copy number variant in *HBA2* (the second α-globin gene). Co-inheritance of α-thalassaemia with SCD reduces haemolysis(Embury et al., 1982, Ballas, 2001). I replicated the previous association between *NPRL3* variant *rs7203560* on chr16:184390 and haemolysis in SCD(Milton et al., 2013), see Table 2. Of note, in this manuscript, a subset of our current cohort (N=213) comprised part of the "replication cohort". *NPRL3* is likely to be in LD with the functional *HBA2* variants.

Table 2 Suggestive novel loci associated with haemolytic index (averaged over 10 years) for all sickle genotypes (A1 is the effect allele)

Chr	Variant	Position (hg19)	A1	A2	Gene	MAF	β-value	s.e.	p-value
16	*rs7203560*	184390	G	T	*NPRL3*	0.107	-0.247	0.079	0.0017
4	rs4695226	47382920	C	A		0.246	-0.303	0.056	5.006×10^{-8}
16		223678	Del	C	*HBA2*	0.209	-0.298	0.058	3.380×10^{-7}

On review of the cluster plots, *rs7203560* is well clustered. The remaining two variants were both imputed but raw genotype calls local to these were well clustered, except for one (of several) markers assessed near the copy number variant at 16:223678.

a



b

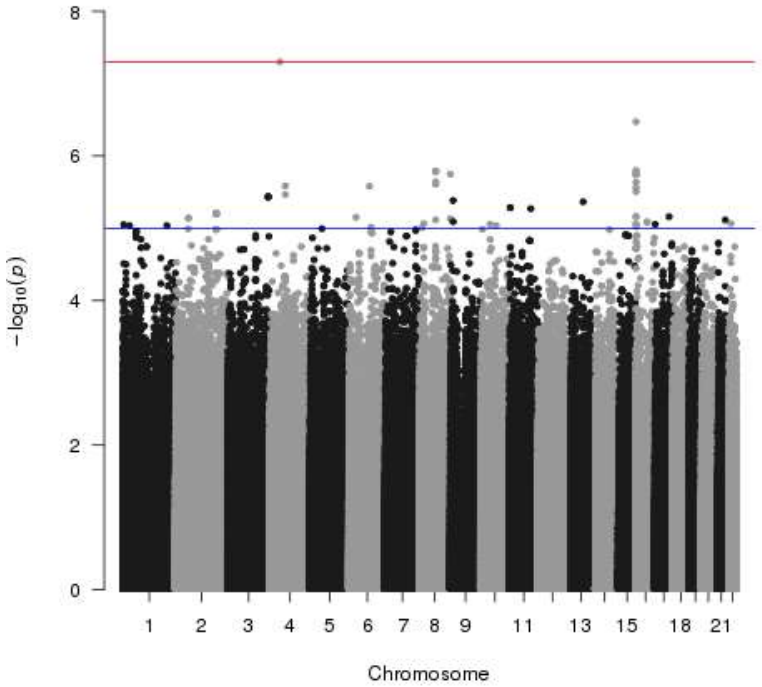


Figure 13 Linear mixed model results for the haemolytic index: (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome

4.3.5. Mortality/survival as phenotype

I abandoned the GWAS with mortality/survival as the trait because of the low numbers: of 354, only 21 died, which is too small to produce a robust model.

4.3.6. Urinary albumin creatinine ratio as phenotype

I performed a genome-wide association study for urinary albumin creatinine ratio (UACR, averaged over 10 years) in a discovery cohort of 326 adult patients with SCD (all sickle genotypes) of African Caribbean or West African heritage. Demographic details are presented in Figure 14.

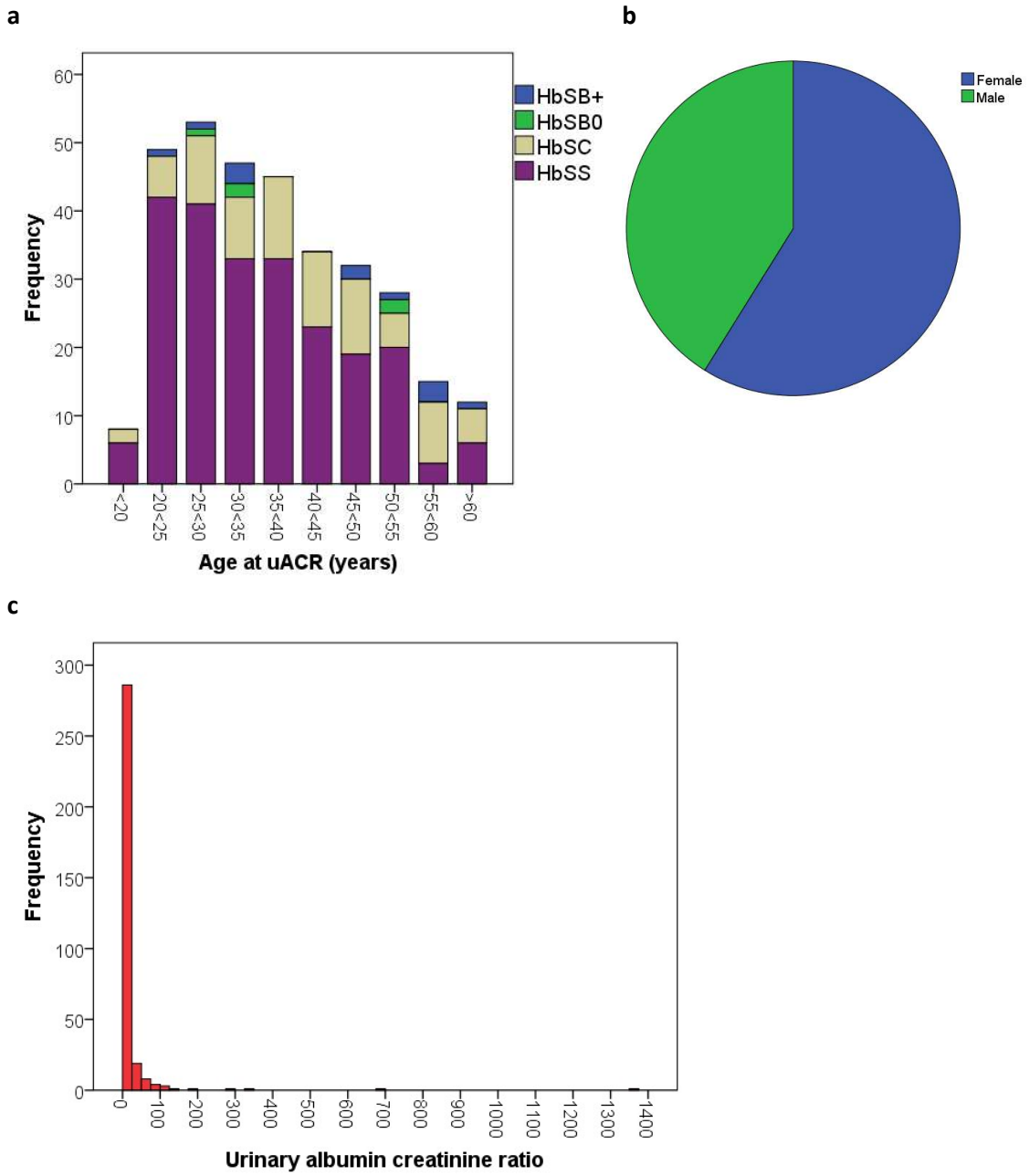


Figure 14 Demographic details for the uACR GWAS (N=326): (a) histogram of age and sickle genotype (b) pie chart of sex (c) histogram of urinary albumin creatinine ratios

I used age, sex and sickle genotype as fixed covariates in the linear mixed model. The genomic control (λ_{GC}) for the analysed variants was 0.995 and a QQ plot of the observed versus expected p-values is shown in Figure 15a. Again, the low QQ plot suggests the model has somewhat overcorrected for population structure. This gives conservative results (p values not as low in our analysis). The likely cause is low sample size leading to under-powering of analysis. The distribution of association p-values (Manhattan plot) for UACR is shown in Figure 15b.

I replicated the previous associations between *APOL1 G1* (both *rs73885319* and *rs60910145*) and proteinuria in SCD(Saraf et al., 2017): *rs73885319* (β=0.0784576, p=0.00815811) and *rs60910145* (β= 0.0870071, p=0.00371177). [Unfortunately, *APOL1 G2*, another locus associated in previous studies, is not in the imputed dataset: *rs71785313* N388 deletion and Y389 deletion chr22:36662051]

There were no significant genome-wide results but I have detected some suggestive novel loci, see Table 3. The chromosome 12 site is within an intron within gene *RIC8B*, and the chromosome 21 region is within an intron in gene *GRIK1*.

Table 3 Suggestive novel loci associated with UACR (averaged over 10 years) for all sickle genotypes (A1 is the effect allele)

Chr	Variant	Position (hg19)	A1	A2	Gene	MAF	β-value	s.e.	p-value
12	rs112494669	107172971	A	G	*RIC8B*	0.101	0.214	0.045	1.915×10^{-6}
19	rs7249863	23730717	C	G		0.466	0.132	0.027	1.415×10^{-6}
21	rs546244357	31043724	A	AT	*GRIK1*	0.512	0.135	0.028	1.497×10^{-6}

However, on review of the cluster plots, rs112494669 clustered poorly so should be rejected. rs7249863 and rs546244357 were both imputed but raw genotype calls local to these were well clustered.

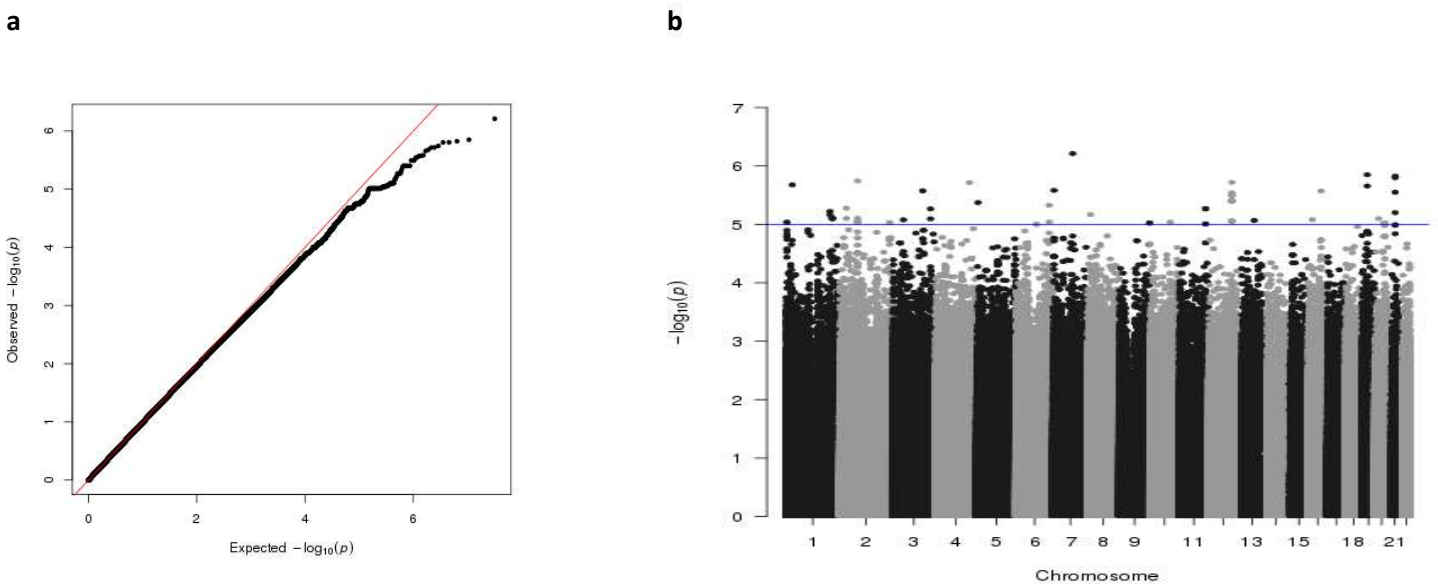


Figure 15 Linear mixed model results for uACR: (a) QQ plot of the observed versus the expected p-values (b) Manhattan plot for association results (-log10p) plotted against the position on each chromosome

4.4. Conclusions

After the limited success of linkage and candidate gene association studies to find genes in complex diseases, genome wide association studies boosted by developments of genotyping platforms have emerged as promising alternative approaches. The HapMap project showed that extensive linkage disequilibrium allowed the design of marker panels ("tag" variants). Quality control is crucial to reduce genotyping errors (especially those that can cause false positive associations). Imputed variants can be more significant than genotyped variants, and can therefore help to "fine-map" the location of the causal mutation. However, relatedness (both near and far – in the form of population stratification) can confound association studies and this must be taken into account in analysis.

I have created a software resource which encapsulates all genome-wide analysis in one user-friendly script. Non-technical-savvy colleagues can now supply their own files of phenotypes to analyse different sickle traits, and as part of the package are able to define different parameters (e.g. sickle subgroup analyses, age ranges, relatedness cut-offs).

I have evaluated fetal haemoglobin, hospitalisation rates, a haemolytic index and urinary albumin creatinine ratio as quantitative markers of severity of SCD. Genome wide studies have replicated previous findings for HbF%, hospitalisation rates, haemolytic index and UACR. There are also some tentative novel loci identified for hospitalisation rates and haemolytic index. For hospitalisation rates, peak signals for three regions were at *rs75904749* on chromosome 5, *rs10792490* on chromosome 11 and *rs510384* on chromosome 12. The latter is within an intron of gene *SLC6A13*, a sodium-dependent GABA and taurine transporter. For haemolytic index, peak signals were a copy number variant in *HBA2* and rs4695226 on chromosome 4.

The genome-wide analyses remain under-powered for the evaluation of the SCD severity indices hospitalisation rate, haemolytic index and uACR, so instead I will move forward to candidate gene analysis in the next chapter where we will have greater power to detect new risk variants.

References

BALLAS, S. K. 2001. Effect of alpha-globin genotype on the pathophysiology of sickle cell disease. *Pediatr Pathol Mol Med,* 20**,** 107-21.

CLARKE, G. M., ANDERSON, C. A., PETTERSSON, F. H., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2011. Basic statistical analysis in genetic case-control studies. *Nat Protoc,* 6**,** 121-33.

CRAIG, J. E., ROCHETTE, J., FISHER, C. A., WEATHERALL, D. J., MARC, S., LATHROP, G. M., DEMENAIS, F. & THEIN, S. L. 1996. Dissecting the loci controlling fetal haemoglobin

production on chromosomes 11p and 6q by the regressive approach. *Nature Genetics,* 12**,** 58-64.

DEVLIN, B. & ROEDER, K. 1999. Genomic control for association studies. *Biometrics,* 55**,** 997-1004.

EMBURY, S. H., DOZY, A. M., MILLER, J., DAVIS, J. R., JR., KLEMAN, K. M., PREISLER, H., VICHINSKY, E., LANDE, W. N., LUBIN, B. H., KAN, Y. W. & MENTZER, W. C. 1982. Concurrent sickle-cell anemia and alpha-thalassemia: effect on severity of anemia. *N Engl J Med,* 306**,** 270-4.

KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S. Y., FREIMER, N. B., SABATTI, C. & ESKIN, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet,* 42**,** 348-54.

KNOWLER, W. C., WILLIAMS, R. C., PETTITT, D. J. & STEINBERG, A. G. 1988. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet,* 43**,** 520-6.

LABIE, D., PAGNIER, J., LAPOUMEROULIE, C., ROUABHI, F., DUNDA-BELKHODJA, O., CHARDIN, P., BELDJORD, C., WAJCMAN, H., FABRY, M. E. & NAGEL, R. L. 1985. Common haplotype dependency of high G gamma-globin gene expression and high Hb F levels in beta-thalassemia and sickle cell anemia patients. *Proceedings of the National Academy of Sciences, USA,* 82**,** 2111-4.

LISTGARTEN, J., LIPPERT, C., KADIE, C. M., DAVIDSON, R. I., ESKIN, E. & HECKERMAN, D. 2012. Improved linear mixed models for genome-wide association studies. *Nat Methods,* 9**,** 525-6.

MANOLIO, T. A. 2013. Bringing genome-wide association findings into clinical use. *Nat Rev Genet,* 14**,** 549-58.

MARCHINI, J., CARDON, L. R., PHILLIPS, M. S. & DONNELLY, P. 2004. The effects of human population structure on large genetic association studies. *Nat Genet,* 36**,** 512-7.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69.

MENZEL, S., GARNER, C., GUT, I., MATSUDA, F., YAMAGUCHI, M., HEATH, S., FOGLIO, M., ZELENIKA, D., BOLAND, A., ROOKS, H., BEST, S., SPECTOR, T. D., FARRALL, M., LATHROP, M. & THEIN, S. L. 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet,* 39**,** 1197-9.

MILTON, J. N., ROOKS, H., DRASAR, E., MCCABE, E. L., BALDWIN, C. T., MELISTA, E., GORDEUK, V. R., NOURAIE, M., KATO, G. R., MINNITI, C., TAYLOR, J., CAMPBELL, A., LUCHTMAN-JONES, L., RANA, S., CASTRO, O., ZHANG, Y., THEIN, S. L., SEBASTIANI, P., GLADWIN, M. T. & STEINBERG, M. H. 2013. Genetic determinants of haemolysis in sickle cell anaemia. *Br J Haematol,* 161**,** 270-8.

MILTON, J. N., SEBASTIANI, P., SOLOVIEFF, N., HARTLEY, S. W., BHATNAGAR, P., ARKING, D. E., DWORKIS, D. A., CASELLA, J. F., BARRON-CASELLA, E., BEAN, C. J., HOOPER, W. C., DEBAUN, M. R., GARRETT, M. E., SOLDANO, K., TELEN, M. J., ASHLEY-KOCH, A., GLADWIN, M. T., BALDWIN, C. T., STEINBERG, M. H. & KLINGS, E. S. 2012. A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. *PLoS One,* 7**,** e34741.

NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S., NELSON, M. R., STEPHENS, M. & BUSTAMANTE, C. D. 2008. Genes mirror geography within Europe. *Nature,* 456**,** 98-101.

PATTERSON, N., PRICE, A. L. & REICH, D. 2006. Population structure and eigenanalysis. *PLoS Genet,* 2**,** e190.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38**,** 904-9.

RISCH, N. & MERIKANGAS, K. 1996. The future of genetic studies of complex human diseases. *Science,* 273**,** 1516-7.

SARAF, S. L., SHAH, B. N., ZHANG, X., HAN, J., TAYO, B. O., ABBASI, T., OSTROWER, A., GUZMAN, E., MOLOKIE, R. E., GOWHARI, M., HASSAN, J., JAIN, S., COOPER, R. S., MACHADO, R. F., LASH, J. P. & GORDEUK, V. R. 2017. APOL1, alpha-thalassemia, and BCL11A variants as a genetic risk profile for progression of chronic kidney disease in sickle cell anemia. *Haematologica,* 102**,** e1-e6.

TEAM, R. D. C. 2011. R: A language and environment for statistical computing. . Foundation for Statistical Computing, Vienna, Austria. .

UDA, M., GALANELLO, R., SANNA, S., LETTRE, G., SANKARAN, V. G., CHEN, W., USALA, G., BUSONERO, F., MASCHIO, A., ALBAI, G., PIRAS, M. G., SESTU, N., LAI, S., DEI, M., MULAS, A., CRISPONI, L., NAITZA, S., ASUNIS, I., DEIANA, M., NAGARAJA, R., PERSEU, L., SATTA, S., CIPOLLINA, M. D., SOLLAINO, C., MOI, P., HIRSCHHORN, J. N., ORKIN, S. H., ABECASIS, G. R., SCHLESSINGER, D. & CAO, A. 2008. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc Natl Acad Sci U S A,* 105**,** 1620-5.

WANG, K., HU, X. & PENG, Y. 2013. An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. *Hum Hered,* 76**,** 1-9.

WEALE, M. E. 2010. Quality control for genome-wide association studies. *Methods Mol Biol,* 628**,** 341-72.

WELLCOME_TRUST_CASE_CONTROL_CONSORTIUM 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

WONKAM, A., NGO BITOUNGUI, V. J., VORSTER, A. A., RAMESAR, R., COOPER, R. S., TAYO, B., LETTRE, G. & NGOGANG, J. 2014. Association of variants at BCL11A and HBS1L-MYB with hemoglobin F and hospitalization rates among sickle cell patients in Cameroon. *PLoS One,* 9**,** e92506.

YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. & VISSCHER, P. M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-9.

YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82.

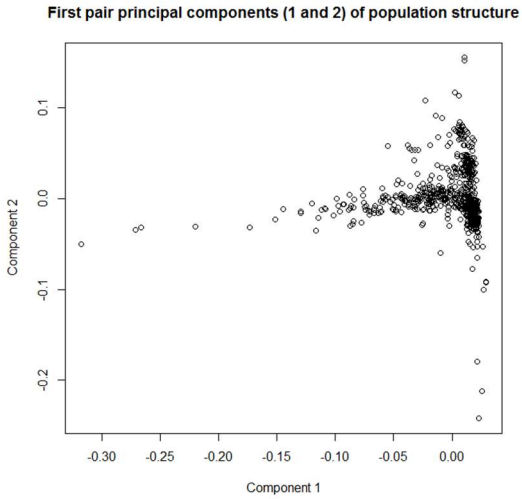
YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. & PRICE, A. L. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet,* 46**,** 100-6.

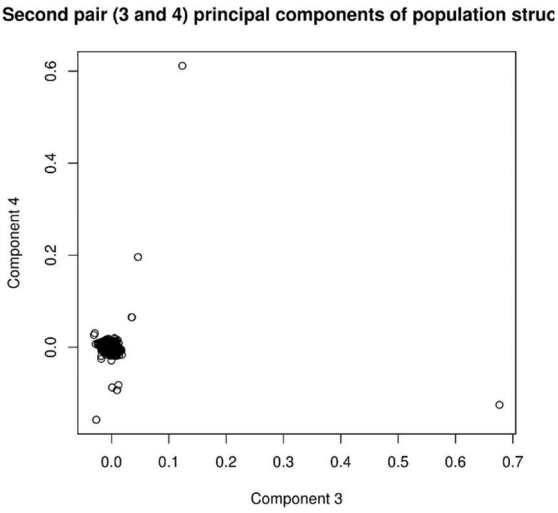
YU, J., PRESSOIR, G., BRIGGS, W. H., VROH BI, I., YAMASAKI, M., DOEBLEY, J. F., MCMULLEN, M. D., GAUT, B. S., NIELSEN, D. M., HOLLAND, J. B., KRESOVICH, S. & BUCKLER, E. S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet,* 38**,** 203-8.

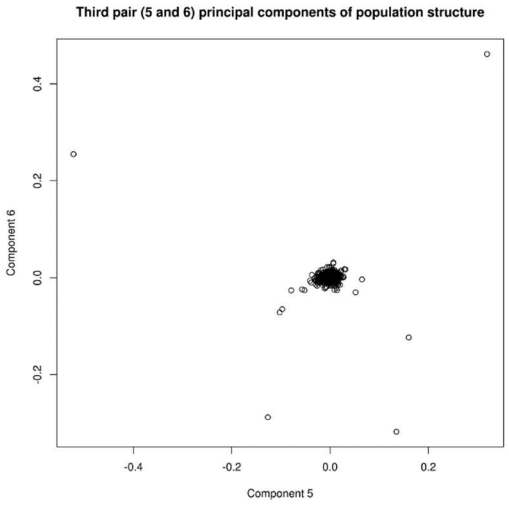
ZAITLEN, N. & KRAFT, P. 2012. Heritability in the genome-wide association era. *Hum Genet,* 131**,** 1655-64.

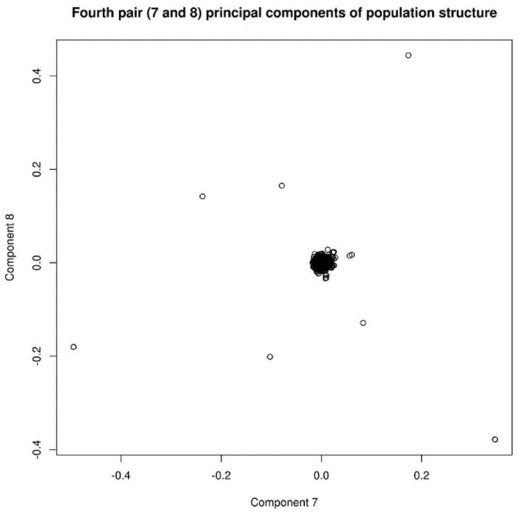
Appendix 1

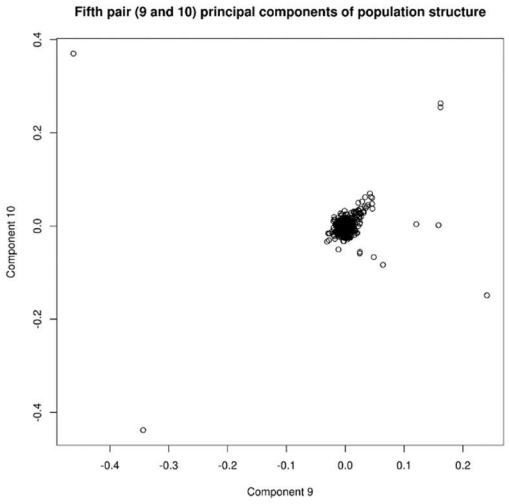
First 10 principal components





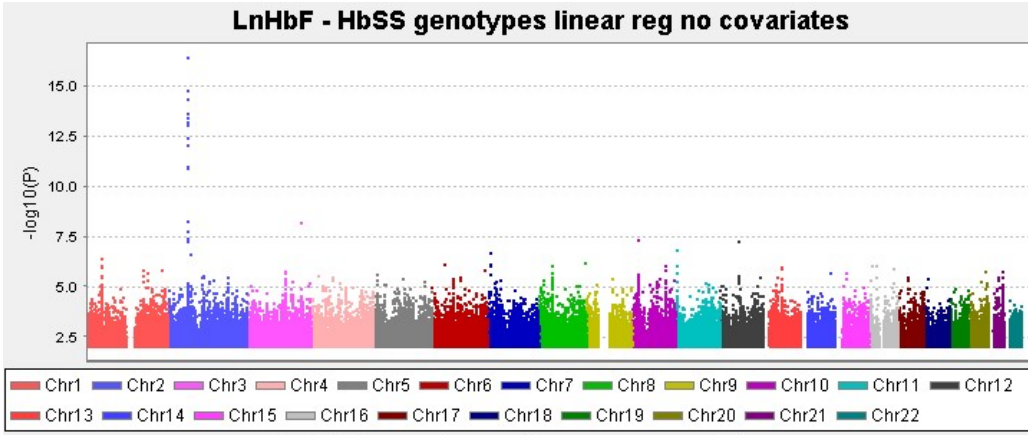
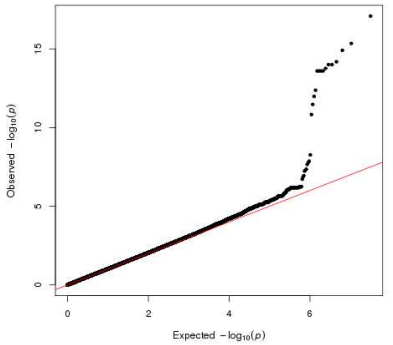
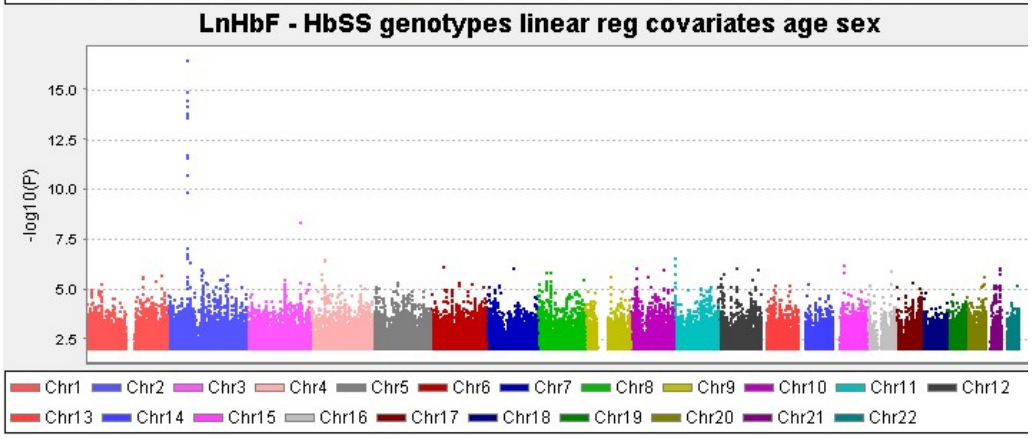
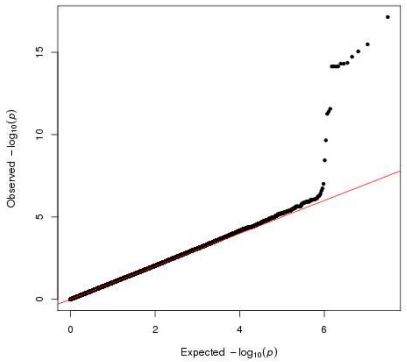


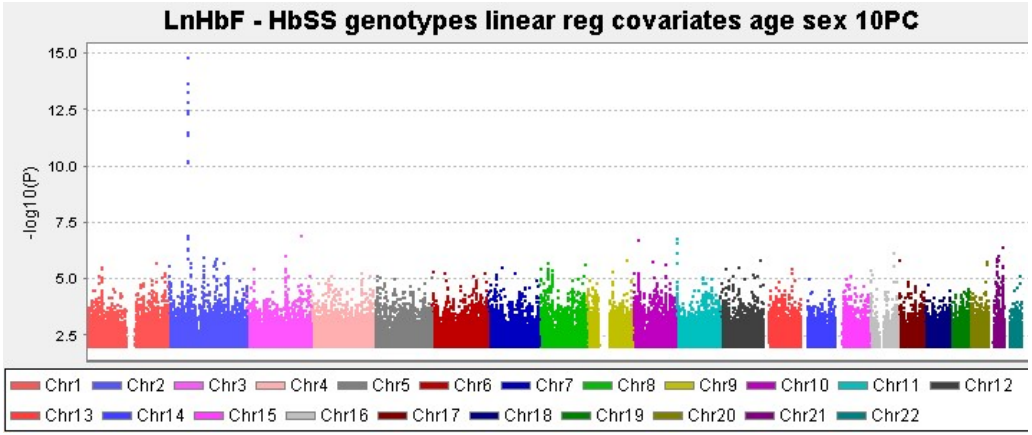
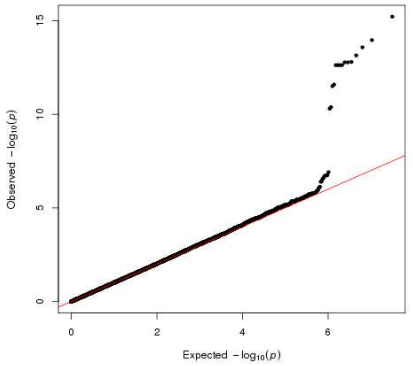
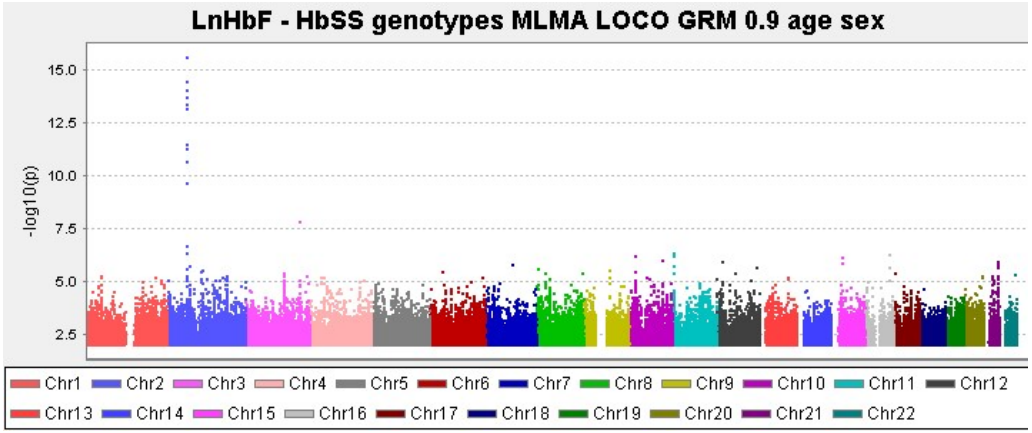
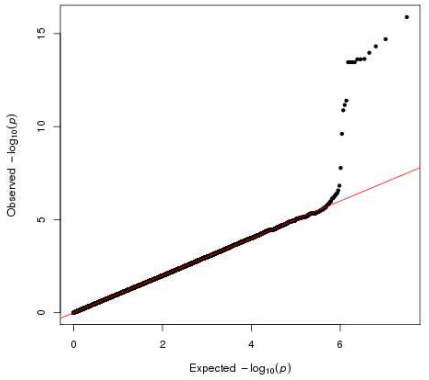




Appendix 2

Improvement in analysis parameters as model becomes more sophisticated. Table of QQ plots and λGC in statistical models of association with progressive sophistication: using linear regression: from basic association with no covariates, to adding age and sex only, then adding the first ten principal components, then using LMM with GRM as well as sex and age.

Model	**Covariates**	**HbSS N=545**	**λ_{GC}**	**QQ plot**
Linear	None		1.040	
Linear (fixed effects)	Sex Age		1.035	

Linear (fixed effects)	Sex Age First 10		1.018	
GRM with cut-off 0.9	Sex Age GRM 0.9		1.004	

Appendix 3

Considerations of power and sample size in in genetic association studies

There are multiple determinants of power in genetic associations studies. First, odds ratio (for logistic regression) or expected beta values – small odds ratios require larger samples (typical range 1.1 – 1.3 in complex disease). Minor Allele Frequencies (MAF) – smaller MAF requires larger sample size however the "common disease, common variant" hypothesis assumes a range 5% to 50%. Third, reasonable linkage disequilibrium between the marker variant and causal variant – most studies assume a genotyped variant has r^2>0.8 with the causal variant. Finally, some consideration of the significance threshold to account for multiple testing.

Appendix 4

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -q HighMemLongterm.q,LowMemLongterm.q
#$ -M kate.gardner@doctors.org.uk
#$ -m beas
#$ -l h_vmem=80G
##########


# First select subpopulation (if any) based on sickle genotype

# Second remove patients that are inappropriate eg in HbF analysis, those aged less than 5

# Third do analysis. Three levels of modelling:
# A. basic association analysis (with/without fixed covariates)
# B. association analysis using PCA
# C. Linear mixed modelling to take account of close relatedness
# Use GCTA to do linear mixed modelling to take account of family relatedness, population substructure and
admixture
# http://gcta.freeforums.net/board/2/gcta-user-manual
# http://cnsgenomics.com/software/gcta/index.html

#####################################
# FOR SPECIFIC SICKLE GENOTYPES ONLY #
#####################################
#Input population HbSS, HbSC etc. Else no
POPULATION=$1 # requires a text file with two columns FID/IID so we can extract that population eg HbSS.txt,
HbSC.txt
GRM_CUTOFF=$2 #relatedness cutoff for genetic relatedness matric eg 0.9 to exclude genetic duplicates
PCA=$3 #number of principal components to analyse eg 10
```

```
OUTCOME=$4 # "HbF" or "HaemIndex" or
IMP_OR_CHIP_DATA=$5 #"chip" or "imp"
HbFg=$6
OUTCOMEFILE=$7
LOWER_AGE=$8
UPPER_AGE=$9

        module add bioinformatics/plink2/1.90b3.38
        module add bioinformatics/plink/1.90b3.31
        module add bioinformatics/gcta/1.26.0
        module add bioinformatics/R/3.3.0

#Create phenotype files
cp ${OUTCOMEFILE} Pheno_${OUTCOMEFILE}
~/dos2unix Pheno_${OUTCOMEFILE} # otherwise  function doesn't work properly
echo "FID IID ${OUTCOME} AgeAt${OUTCOME}" > ${OUTCOME}WithFID.txt
join -1 2 -2 1 -o1.1,1.2,2.2,2.3 <(sort -k2 SickleMEGA_QC_NoSexMismatch.autosomes.fam) <(sort -k1
Pheno_${OUTCOMEFILE}) >> ${OUTCOME}WithFID.txt
awk '{print $1, $2, $3}' ${OUTCOME}WithFID.txt > Phenotype_${OUTCOME}.txt
awk '{print $1, $2, $4}' ${OUTCOME}WithFID.txt > Covariate_AgeAt${OUTCOME}.txt

#add HbFg to coavriatefile if required
if [[ "${HbFg}" = "yes" ]]; then
        HbFg_FILE="_withHbFgenetic"
      HbFg_NAME=", HbFg (HbF genetic model)"
      #Make phenotype file with HbFgenetic data
      echo "FID IID ${OUTCOME} HbFg" > Covariate_AgeAt${OUTCOME}_withHbFgenetic.txt
      tail -n +2 HbFgenetic.txt > HbFgenetic_noHeader.txt
      join -j2 -o1.1,1.2,1.3,2.3 <(sort -k2 Covariate_AgeAt${OUTCOME}.txt) <(sort -k2 HbFgenetic_noHeader.txt) >>
Covariate_AgeAt${OUTCOME}_withHbFgenetic.txt
       else
        HbFg_FILE=""
      HbFg_NAME=""
fi
```

```
if [[ "$IMP_OR_CHIP_DATA" = "chip" ]]; then
                #chip plink data
      IMP_CHIP_FILE="SickleMEGA_QC_NoSexMismatch.autosomes"
      IMP_OR_CHIP_DATA_NAME="Chip data"
        else
      #imputed plink data
      IMP_CHIP_FILE="Sickle_Imputed_QC_strict"
      IMP_OR_CHIP_DATA_NAME="Imputed data"
fi


if [[ "$POPULATION" = "ALL" || "$POPULATION" = "nonHbSS" || "$POPULATION" = "HbSSHbSC" || "$POPULATION" =
"HbSSHbSCHbSBplus"  ]]; then
      COVARIATES="Sex, age, sickle genotype"
      COVARIATES_FILE="_age_sex_sickle"
      COVARIATES_NONQUANT_FILE="SexSickleCovariates.txt"
       else
      COVARIATES="Sex, age"
        COVARIATES_FILE="_age_sex"
        COVARIATES_NONQUANT_FILE="SexCovariate.txt"
fi


#Create log file
echo "Log file for linear mixed modelling for STSTN sickle cohort
Outcome: ${OUTCOME}
Population: ${POPULATION}
Data source: ${IMP_OR_CHIP_DATA}
LMM GRM cutoff: ${GRM_CUTOFF}
Covariates used in model: ${COVARIATES}${HbFg_NAME}
(PCA: ${PCA} components)
" > Logfile_LMM_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_${PCA}_${OUTCOME}_${IMP_OR_CHIP_DATA}${HbFg_FILE}.txt
```

```
#####################
#1. Remove patients #
#####################

#(i) remove HPFH
#plink --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes --remove HbSHPFH.txt --make-bed --out
SickleMEGA_Pruned_NoSexMismatch.autosomes_NoSHPFH
#if [ "$IMP_OR_CHIP_DATA" = "chip" ]; then
#     plink --bfile SickleMEGA_QC_NoSexMismatch.autosomes --remove HbSHPFH.txt --make-bed --out
SickleMEGA_QC_NoSexMismatch.autosomes_NoSHPFH
#else
#     plink --bfile Sickle_Imputed_QC_strict --remove HbSHPFH.txt --make-bed --out Sickle_Imputed_QC_strict_NoSHPFH
#fi
#
#
##(ii) remove genetic duplicates (identified from previous cycling of this and seeing those individuals in genetic
relatedness matrix with value > 0.9. Then went back to see if these were the same person in twice, MZ twin (in
which cases fine to delete one in the duplicate pair) or if there wasn't a reason for duplicate (in which case
delete both)
#plink --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes_NoSHPFH --remove GeneticDuplicates.txt --make-bed --out
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved
#
#if [ "$IMP_OR_CHIP_DATA" = "chip" ]; then
#     plink --bfile SickleMEGA_QC_NoSexMismatch.autosomes_NoSHPFH --remove GeneticDuplicates.txt --make-bed --out
SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved
#else
#     plink --bfile Sickle_Imputed_QC_strict_NoSHPFH --remove GeneticDuplicates.txt --make-bed --out
Sickle_Imputed_QC_strict_PtsRemoved
#fi

#########################################
#2. Make GRM (genetic relatedness matrix) #
#########################################
```

```
# GRM required for genetic duplicate analysis, but also needed for MLMA and PCA below.
# Creating GRM before filtering out population we want to get GRM on largest dataset
# using pruned QC file: also required to create PCA
#(i) make initial GRM using pruned, unimputed SNPs
#       gcta --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved --autosome --maf 0.01 --thread-num 6 --
make-grm --out SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved

############################################################
# 3. Select specific population by sickle genotype and age #
############################################################


# (i) choose age category (user defined)
 awk -v LOWER_AGE="$LOWER_AGE" -v UPPER_AGE="$UPPER_AGE" '{if ($11>=LOWER_AGE && $11<UPPER_AGE) print $1, $2}'
AllCovariates.txt > AgeRange${LOWER_AGE}-${UPPER_AGE}.txt

#plink --bfile TROY_CNV --keep AgeRange${LOWER_AGE}-${UPPER_AGE}.txt --make-bed --out
TROY_CNV_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL
 plink --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved --keep AgeRange${LOWER_AGE}-${UPPER_AGE}.txt --
make-bed --out SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL
 if [ "$IMP_OR_CHIP_DATA" = "chip" ]; then
       plink --bfile SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved --keep AgeRange${LOWER_AGE}-${UPPER_AGE}.txt
--make-bed --out SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL
else
       plink --bfile Sickle_Imputed_QC_strict_PtsRemoved  --keep AgeRange${LOWER_AGE}-${UPPER_AGE}.txt --make-bed -
-out Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL
fi

#(ii) choose sickle genotype population (user defined)

# population = ${POPULATION} from both pruned, unimputed dataset AND FROM unpruned, imputed dataset

if [ "$POPULATION" != "ALL" ]; then
        if [ "$POPULATION" = "nonHbSS" ]; then
```

```
#           plink --bfile TROY_CNV_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL --remove HbSS.txt --make-bed --out
TROY_CNV_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}
                plink --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_ALL --remove HbSS.txt --make-bed --out
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}
            if [ "$IMP_OR_CHIP_DATA" = "chip" ]; then
                 plink --bfile SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_ALL --remove HbSS.txt --make-bed --out
SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}
            else
                 plink --bfile Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL --remove
HbSS.txt --make-bed --out Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}
            fi
        else
#                plink --bfile TROY_CNV_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL --keep ${POPULATION}.txt --make-bed -
-out TROY_CNV_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}
               plink --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_ALL --keep ${POPULATION}.txt --make-bed --out
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}
               if [ "$IMP_OR_CHIP_DATA" = "chip" ]; then
                plink --bfile SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_ALL --keep ${POPULATION}.txt --make-bed --out
SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}
               else
                plink --bfile Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_ALL --keep
${POPULATION}.txt --make-bed --out Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}
               fi

        fi
fi


#################
#4. Do analysis #
```

```
#################
#
##A. BASIC ASSOC CHECK compare basic plink assoc pre and post imputation. plink's assoc can have binary or
quantitative outcome. Plink's --linear or --logistic allows for multiple covariates
#                       #(i) assoc
#
##                      plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno AllCovariates.txt --
pheno-name LnHbF --assoc --out Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved
##                                 head -n1 Sickle_Imputed${GRM_CUTOFF}_QC_strict_${POPULATION}_PtsRemoved.qassoc >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.qassoc.pLT0.01
#                                awk '{if($9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.qassoc >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.qassoc.pLT0.01
#
                       #(ii) linear, no covariates or one covariate individually


#                 if [[ "$POPULATION" = "ALL" || "$POPULATION" = "nonHbSS" || "$POPULATION" = "HbSSHbSC" ||
"$POPULATION" = "HbSSHbSCHbSBplus" ]]; then
#
#
#                       plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno AllCovariates.txt -
-pheno-name LnHbF --linear --out Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved
#                        head -n1 Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear.pLT0.01
#                        awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear.pLT0.01

 #                              plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno
AllCovariates.txt --pheno-name LnHbF --linear --covar AllCovariates.txt --covar-name SexCode --out
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex
 #                               head -n1 Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex.assoc.linear.pLT0.01
```

```
 #                                awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex.assoc.linear.pLT0.01
#
#                       plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno AllCovariates.txt --
pheno-name LnHbF --linear --covar AllCovariates.txt --covar-name AgeAtHbF --out
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age
 #                                head -n1 Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear.pLT0.01
 #                                awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear.pLT0.01

#                       plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno AllCovariates.txt --
pheno-name LnHbF --linear --covar AllCovariates.txt --covar-name HbSS,HbSC,HbSBplus --out
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sickle_SSSCSBplus
 #                                head -n1
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sickle_SSSCSBplus.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sickle_SSSCSBplus.assoc.linear.pLT0.01
  #                                awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sickle_SSSCSBplus.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sickle_SSSCSBplus.assoc.linear.pLT0.01

#                else
#                       plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno AllCovariates.txt -
-pheno-name LnHbF --linear --out Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved
#                        head -n1 Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear.pLT0.01
#                        awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved.assoc.linear.pLT0.01
#
#                       plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno AllCovariates.txt -
-pheno-name LnHbF --linear --covar AllCovariates.txt --covar-name SexCode --out
```

```
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex_sickle_Imputed_QC_strict_ALL_GRM0.9_GCTA_HbF${COVARIATES_FILE
}.loco.mlma
#                         awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_sex.assoc.linear.pLT0.01

 #                       plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno AllCovariates.txt -
-pheno-name LnHbF --linear --covar AllCovariates.txt --covar-name AgeAtHbF --out
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age
#                         head -n1 Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear.pLT0.01
#                         awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_age.assoc.linear.pLT0.01
#
#                fi

                      #(iii) linear, with covariates

 #                     if [[ "$POPULATION" = "ALL" || "$POPULATION" = "nonHbSS" || "$POPULATION" = "HbSSHbSC" ||
"$POPULATION" = "HbSSHbSCHbSBplus"  ]]; then

 #                              plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno
AllCovariates.txt --pheno-name LnHbF --linear --covar AllCovariates.txt --covar-name
SexCode,AgeAtHbF,HbSS,HbSC,HbSBplus --out Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}
 #                                #this is a big file - only need the ADD beta/p value (not the covariates beta/p)
 #                              head -n1
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear.pLT0.01
 #                              awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear.pLT0.01
 #                      else
```

```
 #                              plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno
AllCovariates.txt --pheno-name LnHbF --linear --covar AllCovariates.txt --covar-name SexCode,AgeAtHbF --out
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}
 #                                #this is a big file - only need the ADD beta/p value (not the covariates beta/p)
 #                              head -n1
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear.pLT0.01
 #                              awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved${COVARIATES_FILE}.assoc.linear.pLT0.01
 ##                         fi



#B. Make GRM (genetic relatedness matrix) using pruned QC file: also required to create PCA
#(i) make initial GRM using pruned, unimputed SNPs
#     gcta --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes_${POPULATION}_PtsRemoved --autosome --maf 0.01 --
thread-num 6 --make-grm --out SickleMEGA_Pruned_NoSexMismatch.autosomes_${POPULATION}_PtsRemoved



#(ii) Do PCA using first n=${PCA} components using GCTA
      # (a) get the first N=${PCA} PCA components - requires GRM (GCTA) based on pruned, unimputed SNPs as above
#     gcta  --grm SickleMEGA_Pruned_NoSexMismatch.autosomes_${POPULATION}_PtsRemoved --pca ${PCA}  --out
SickleMEGA_Pruned_NoSexMismatch.autosomes_${POPULATION}_PtsRemoved

      #plot first two components in R
#     module add bioinformatics/R/3.3.0
#     Can run R scripts (from within R) with source("exampleScript.R") : script must be in same folder
#     EigenVec10PC_ALL <- read.table("SickleMEGA_Pruned_NoSexMismatch.autosomes_ALL.eigenvec",header=FALSE)
#     png("1and2PC.png", width=6, height=6, units="in", res=1000)
#     plot(EigenVec10PC_ALL$V3,EigenVec10PC_ALL$V4,main="First pair principal components (1 and 2) of population
structure",xlab= "Component 1", ylab="Component 2")
#        dev.off()
```

```
#     png("3and4PC.png", width=6, height=6, units="in", res=1000)
#     plot(EigenVec10PC_ALL$V5,EigenVec10PC_ALL$V6,main="Second pair principal components (3 and 4) of population
structure",xlab= "Component 1", ylab="Component 2")
#     dev.off()


      #(b) do analysis using PCA components as covariates (plink)
#     if [[ "$POPULATION" = "ALL" || "$POPULATION" = "nonHbSS" || "$POPULATION" = "HbSSHbSC" || "$POPULATION" =
"HbSSHbSCHbSBplus" ]]; then
#           #  create a file of covariates combining (1) the N=${PCA} PCA components taken directly from the above
eigenvectors and (2) the demographic covariates
#     _AgeRange${LOWER_AGE}-${UPPER_AGE}echo "FID IID PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 SexCode AgeAtHbF
HbSS HbSC HbSBplus" > SexAgeAtHbFSickle${PCA}PCcovariates_${POPULATION}.${PCA}.txt
#           join -j2 -o1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9,1.10,1.11,1.12,2.4,2.11,2.12,2.13,2.14 <(sort -k2
SickleMEGA_Pruned_NoSexMismatch.autosomes_${POPULATION}_PtsRemoved.eigenvec) <(sort -k2 AllCovariates_noHead.txt)
>> SexAgeAtHbFSickle${PCA}PCcovariates_${POPULATION}.${PCA}.txt
##          # do linear analysis in plink using these covariates
#           plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno LnHbF_updated2.txt --linear --
covar SexAgeAtHbFSickle${PCA}PCcovariates_${POPULATION}.${PCA}.txt --out
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}
#           #this is a big file - only need the ADD beta/p value (not the covariates beta/p)
#           head -n1 Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear.pLT0.01
#           awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear.pLT0.01
#
#     else
#                 # create a file of covariates combining (1) the N=${PCA} PCA components taken directly from the
above eigenvectors and (2) the demographic covariates
#             echo "FID IID PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 SexCode AgeAtHbF" >
SexAgeAtHbF${PCA}PCcovariates_${POPULATION}.${PCA}.txt
#             join -j2 -o1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9,1.10,1.11,1.12,2.4,2.11 <(sort -k2
SickleMEGA_Pruned_NoSexMismatch.autosomes_${POPULATION}_PtsRemoved.eigenvec) <(sort -k2 AllCovariates_noHead.txt)
>> SexAgeAtHbF10PCcovariates_${POPULATION}.${PCA}.txt
```

```
#             # do linear analysis in plink using these covariates
#             plink --bfile Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved --pheno LnHbF_updated2.txt --linear -
-covar SexAgeAtHbF${PCA}PCcovariates_${POPULATION}.${PCA}.txt --out
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}
#           #this is a big file - only need the ADD beta/p value (not the covariates beta/p)
#             head -n1 Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear >
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear.pLT0.01
#             awk '{if($5=="ADD" && $9<0.01) print $0}'
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear >>
Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_PCA${COVARIATES_FILE}.assoc.linear.pLT0.01
#     fi

#########################################################################
# C. modify GRM to exclude people more related than cutoff= $GRM_CUTOFF #
#########################################################################
#This is treated differently for grm-cutoff=0.9. GRM cutoffs are used to exclude people more related than the
cutoff. Where GRM_CUTOFF=0.9 we want to exclude patients manually after identifying duplicates/MZ twins and
choossing appropriately (might depend on how much phenotype data we have on them)

if [[ "$GRM_CUTOFF" = "0.9" ]]; then
#if GRM_CUTOFF=0.9, simple copy over the files with genetic duplicates already manually removed to the new
nomenclature
#     #pruned plink and GRM data
      cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved.grm.bin
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcuto
ff${GRM_CUTOFF}.grm.bin
        cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved.grm.id
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcuto
ff${GRM_CUTOFF}.grm.id
        cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved.grm.N.bin
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcuto
ff${GRM_CUTOFF}.grm.N.bin
```

```
      cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${L
OWER_AGE}-${UPPER_AGE}_${POPULATION}.bim
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${L
OWER_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bim
        cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${L
OWER_AGE}-${UPPER_AGE}_${POPULATION}.bed
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${L
OWER_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bed
        cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${L
OWER_AGE}-${UPPER_AGE}_${POPULATION}.fam
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${L
OWER_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.fam

#     cp /users/k1343761/brc_scratch/sickle_new/LMM/info_r2/TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.bim /users/k1343761/brc_scratch/sickle_new/LMM/info_r2/TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bim
#        cp /users/k1343761/brc_scratch/sickle_new/LMM/info_r2/TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.bed /users/k1343761/brc_scratch/sickle_new/LMM/info_r2/TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bed
#        cp /users/k1343761/brc_scratch/sickle_new/LMM/info_r2/TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.fam /users/k1343761/brc_scratch/sickle_new/LMM/info_r2/TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.fam

      if [ "$IMP_OR_CHIP_DATA" = "chip" ]; then
            #chip plink data
            cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER
_AGE}-${UPPER_AGE}_${POPULATION}.bim
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER
_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bim
```

```
                cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER
_AGE}-${UPPER_AGE}_${POPULATION}.bed
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER
_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bed
                cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER
_AGE}-${UPPER_AGE}_${POPULATION}.fam
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER
_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.fam

      else
            #imputed plink data
            cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.bim
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bim
                cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.bed
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bed
                cp
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.fam
/users/k1343761/brc_scratch/sickle_new/LMM/info_r2/Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.fam

      fi

else
#if GRM_CUTOFF<0.9, use ./GCTA to exclude individuals based on GRM_CUTOFF
       gcta --grm SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved --grm-cutoff ${GRM_CUTOFF} --thread-num 6  -
-make-grm  --out SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcutoff${GRM_CUTOFF}
```

```
      #keep only those individuals left in GRM from plink datasets
      plink --bfile SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION} --keep
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcutoff${GRM_CUTOFF}.grm.id --make-bed --out
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}
#     plink --bfile TROY_CNV_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION} --keep
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcutoff${GRM_CUTOFF}.grm.id --make-bed --out
TROY_CNV_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}

        if [ "$IMP_OR_CHIP_DATA" = "chip" ]; then
               #chip plink data
           plink --bfile SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION} --keep
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcutoff${GRM_CUTOFF}.grm.id --make-bed --out
SickleMEGA_QC_NoSexMismatch.autosomes_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}
        else
               #imputed plink data
             plink --bfile Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION} --
keep SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcutoff${GRM_CUTOFF}.grm.id --make-bed --out
Sickle_Imputed_QC_strict_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}
      fi
fi

#############################################################################################
#D. (GCTA) Do mixed linear model association (MLMA) using leave-one-chromosome-out (LOCO) #
#############################################################################################
        #input: imputed file, GRM from above, covariate file and phenotype file
        # covariates:
      # discrete covariates (--covar): sex plus sickle if population = ALL
      # quantitative covariates (--qcovar): Age

#GRM cutoffs used to exclude people too related. The process excludes both individuals (not one) - this seems odd
```

```
#we only want to exclude genetic identical individuals (GRM>0.9) and importantly to keep the data - we don't want
to blanket delete individuals.

#########################################################################
#temporary step to merge TROY_CNV data
#plink --bfile ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF} --bmerge TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bed TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.bim TROY_CNV_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.fam --make-bed --out
${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_TROY_CNV
#
#
#gcta --mlma-loco --reml-maxit 1000 --bfile ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_TROY_CNV --grm
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcutoff${GRM_CUTOFF} --pheno Phenotype_${OUTCOME}.txt --
covar ${COVARIATES_NONQUANT_FILE} --qcovar Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt --thread-num 6 --out
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV
#
#
#
###create results file only containing results wiht p<0.01 (for easy transfer off Rosalind, eg to local computer to
view in Haploview)
## CHANGE FILE NAMES IF USING#
#head -n1 ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma >
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma.pLT0.0
1
#awk '{if($9<0.01) print $0}' ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma >>
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma.pLT0.0
1
```

```
#
###create results file with only valid p values (no NA etc) - subsequent R processing cannot cope with missing
values
#head -n1 ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma >
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma.tick
#awk '{if($9<=1 && $9>=0) print $0}' ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma >>
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma.tick
#
#
##QC and create plots via R script (manhattan plots, QQ plots and calculate lambda(GC))
#Rscript LambdaAndQQ_TROY_CNV.R ${POPULATION} ${GRM_CUTOFF} ${OUTCOME} ${IMP_OR_CHIP_DATA} ${HbFg}
${COVARIATES_FILE} ${LOWER_AGE} ${UPPER_AGE}
#
##Find number of patients in common in the files
##(a) create file of people with non-missing phenotype AND non missing HbFg data (ie this equates to complete
phenotype and covariate set)
#if [[ "${HbFg}" = "yes" ]]; then
#        #check both outcome and HbFg are non missing and create file of non missing data
#        awk '{if($3>-9 && $4>-9) print $0}' Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt >
AgeAt${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${IMP_OR_CHIP_DATA}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_noMissing.txt
#       else
#        #check outcome is non missing and create file of non missing data
#        awk '{if($3>-9) print $0}' Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt >
AgeAt${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${IMP_OR_CHIP_DATA}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_noMissing.txt
#fi
#
#
##(b) those in common between genetic and phenotypic data
```

```
#awk 'NR==FNR {a[$2];next}$2 in a{print $1,$2}' ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_TROY_CNV.fam AgeAt${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${IMP_OR_CHIP_DATA}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_noMissing.txt >
Pts_LMManalysis_${IMP_OR_CHIP_DATA}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${OUTCOME}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_TROY_CNV.txt
#
#
#N=$(wc Pts_LMManalysis_${IMP_OR_CHIP_DATA}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${OUTCOME}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_TROY_CNV.txt | awk '{print $1}')
#Lambda=$(awk '{print $1}' ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV_lambda.txt)
#
#
#echo "Number of patients in analysis (N): $N
#Lambda (GC)=$Lambda
#
#Files saved:
#Patients used in analysis: Pts_LMManalysis_${IMP_OR_CHIP_DATA}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${OUTCOME}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_TROY_CNV.txt
#Binary PLINK files used for analysis: ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_TROY_CNV
#Covariate files used (non-quantitative and quantitative): ${COVARIATES_NONQUANT_FILE}
Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt
#MLMA results: ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma
#MLMA results without missing data: ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_TROY_CNV.loco.mlma.tick
#Manhattan plot: ${IMP_OR_CHIP_DATA}_${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}${COVARIATES_FILE}_GRM${GRM_CUTOFF}${HbFg_FILE}_TROY_CNV_manhattan.png
#QQ plot: ${IMP_OR_CHIP_DATA}_${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}${COVARIATES_FILE}_GRM${GRM_CUTOFF}${HbFg_FILE}_TROY_CNV_qq.png
#" >> Logfile_LMM_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_${PCA}_${OUTCOME}_${IMP_OR_CHIP_DATA}${HbFg_FILE}_TROY_CNV.txt
#
```

```
#########################################################################

gcta --mlma-loco --reml-maxit 1000 --bfile ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF} --grm
SickleMEGA_Pruned_NoSexMismatch.autosomes_PtsRemoved_postGRMcutoff${GRM_CUTOFF} --pheno Phenotype_${OUTCOME}.txt --
covar ${COVARIATES_NONQUANT_FILE} --qcovar Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt --thread-num 6 --out
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}



##create results file only containing results wiht p<0.01 (for easy transfer off Rosalind, eg to local computer to
view in Haploview)
# CHANGE FILE NAMES IF USING#
head -n1 ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma >
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma.pLT0.01
awk '{if($9<0.01) print $0}' ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma >>
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma.pLT0.01

##create results file with only valid p values (no NA etc) - subsequent R processing cannot cope with missing
values
head -n1 ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma >
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma.tick
awk '{if($9<=1 && $9>=0) print $0}' ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma >>
${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma.tick
```

```
#QC and create plots via R script (manhattan plots, QQ plots and calculate lambda(GC))
Rscript LambdaAndQQ.R ${POPULATION} ${GRM_CUTOFF} ${OUTCOME} ${IMP_OR_CHIP_DATA} ${HbFg} ${COVARIATES_FILE}
${LOWER_AGE} ${UPPER_AGE}

#Find number of patients in common in the files
#(a) create file of people with non-missing phenotype AND non missing HbFg data (ie this equates to complete
phenotype and covariate set)
if [[ "${HbFg}" = "yes" ]]; then
      #check both outcome and HbFg are non missing and create file of non missing data
      awk '{if($3>-9 && $4>-9) print $0}' Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt >
AgeAt${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${IMP_OR_CHIP_DATA}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_noMissing.txt
       else
      #check outcome is non missing and create file of non missing data
      awk '{if($3>-9) print $0}' Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt > AgeAt${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${IMP_OR_CHIP_DATA}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_noMissing.txt
fi


#(b) those in common between genetic and phenotypic data
awk 'NR==FNR {a[$2];next}$2 in a{print $1,$2}' ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}.fam AgeAt${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${IMP_OR_CHIP_DATA}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}_noMissing.txt >
Pts_LMManalysis_${IMP_OR_CHIP_DATA}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${OUTCOME}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}.txt


N=$(wc Pts_LMManalysis_${IMP_OR_CHIP_DATA}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${OUTCOME}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}.txt | awk '{print $1}')
Lambda=$(awk '{print $1}' ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}_lambda.txt)


echo "Number of patients in analysis (N): $N
Lambda (GC)=$Lambda
```

```
Files saved:
Patients used in analysis: Pts_LMManalysis_${IMP_OR_CHIP_DATA}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_${OUTCOME}_postGRMcutoff${GRM_CUTOFF}${HbFg_FILE}.txt
Binary PLINK files used for analysis: ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}
Covariate files used (non-quantitative and quantitative): ${COVARIATES_NONQUANT_FILE}
Covariate_AgeAt${OUTCOME}${HbFg_FILE}.txt
MLMA results: ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma
MLMA results without missing data: ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}${COVARIATES_FILE}${HbFg_FILE}.loco.mlma.tick
Manhattan plot: ${IMP_OR_CHIP_DATA}_${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}${COVARIATES_FILE}_GRM${GRM_CUTOFF}${HbFg_FILE}_manhattan.png
QQ plot: ${IMP_OR_CHIP_DATA}_${OUTCOME}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}${COVARIATES_FILE}_GRM${GRM_CUTOFF}${HbFg_FILE}_qq.png
" >> Logfile_LMM_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_${PCA}_${OUTCOME}_${IMP_OR_CHIP_DATA}${HbFg_FILE}.txt
```

Appendix 5

Logfile output for user post-LMM

```
Outcome (phenotype)
Population: ALL, HbSS, HbSC, …
Data source: raw or imputed data
LMM GRM cutoff


Number of patients in analysis: <number of patients>
Lambda (GC): <calculated lambda GC>

Files saved:
Patients used in analysis <file name>
Binary PLINK files used for analysis <file name>
Covariates used in mode, together with files used (non-quantitative and
quantitative) <file name>
MLMA results (.mlma file) <file name>
Manhattan plot <file name>
QQ plot <file name>
```

Appendix 6

Manuscript:

HbF$_G$: a Genetic Model of Fetal Hemoglobin in Sickle Cell Disease

HbF$_G$: a Genetic Model of Fetal Hemoglobin in Sickle Cell Disease

Kate Gardner[1,2], Tony Fulford[3], Nicholas Silver[1], Helen Rooks[1], Nikolaos Angelis[1], Marlene Allman[2], Siana Nkya[4], Julie Makani[4], Jo Howard[5], Rachel Kesse-Adu[5], David C Rees[1,2], Sara Stuart-Smith[2,6], Tullie Yeghen[7], Moji Awogbade[2], Raphael Z Sangeda[4], Josephine Mgaya[4], Hamel Patel[8,9], Stephen Newhouse[8,9,10], Stephan Menzel[1,%], Swee Lay Thein[1,2,*,%]

1. King's College London, Division of Cancer Studies, London, UK
2. King's College Hospital NHS Foundation Trust, London, UK
3. University of Cambridge, Behavioural Ecology, Department of Zoology, Cambridge, UK
4. Muhimbili University of Health and Allied Sciences, Dar-es-Salaam, Tanzania
5. Guy's and St Thomas' hospitals NHS Foundation Trust, Department of Haematology, London, UK,
6. Queen Elizabeth Hospital, Lewisham and Greenwich NHS Trust, London, UK
7.University Hospital Lewisham, Lewisham and Greenwich NHS Trust, London, UK
8. Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.
9. NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, UK
10. Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, London, United Kingdom.

* Present address: National Heart, Lung and Blood Institute / NIH, Sickle Cell Branch, Bethesda, USA

% Equal senior authors

Correspondence:

Swee Lay Thein
Sickle Cell Branch
National Heart, Lung and Blood Institute
The National Institutes of Health
Building 10-CRC, Room 6S241
10 Center Drive, Bethesda, MD 20892
Office Line: +1-301-435-2345
Direct Line: +1-301-402-6699
Fax: +1-301-451-7091
Email: sl.thein@nih.gov

Kate Gardner
Red Cell Biology Programme
King's College London
Rayne Institute
123 Coldharbour Lane
London SE5 9NU
Tel:+44-207-848-0740
Email: kate.gardner@doctors.org.uk

Word count (excluding references and abstract): 1198

Abstract count: 200

Reference count: 24

Key points

1. The three established HbF genetic loci can be summarized into one quantitative variable, ***HbF$_G$***, in SCD.

2. ***HbF$_G$*** has been replicated in other SCD cohorts and demonstrated to influence markers of SCD severity.

3. ***HbF$_G$*** provides a quantitative marker for the "genetic component" of HbF variability, potentially useful in future genetic and clinical studies in SCD.

Abstract

Fetal hemoglobin (HbF) is a strong modifier of sickle cell disease (SCD) severity, and associated with three common genetic loci. Quantifying the genetic effects of the 3 loci would specifically address the beneficial side of HbF increases in patients. Here, we have applied statistical methods using the most representative variants - *rs1427407* and *rs6545816* in *BCL11A, rs66650371* (3bp deletion) and rs9376090 in *HMIP-2A, rs9494142* and *rs9494145* in *HMIP-2B,* and *rs7482144* (*Xmn*1- *HBG2* in the β-globin locus), to create ***HbF$_G$***, a genetic quantitative variable for HbF in SCD.

Only patients aged ≥ 5 years with complete genotype and HbF data were studied. 581 patients with HbSS or HbSβ^0 thalassemia formed the "discovery" cohort. Multiple linear regression modelling rationalized the 7 variants down to 4 markers (*rs6545816, rs1427407, rs66650371* and *rs7482144*) each *independently* contributing HbF-boosting alleles; together accounting for 21.8% of HbF variability (r^2) in the HbSS or HbSβ^0 patients. The model was replicated with consistent r^2 in two different cohorts: 27.5% in HbSC patients (N=186) and 23% in 994 Tanzanian HbSS patients. ***HbF$_G$***, our 4-variant model, provides a robust approach to account for the genetic component of HbF in SCD, and is of potential utility in sickle genetic and clinical studies.

Introduction

High HbF levels are clinically beneficial in sickle cell disease (SCD), being associated with longer survival(Platt et al., 1994) and reduced pain rates (Platt et al., 1991). Patients with SCD have higher HbF levels compared to non-affected adults and, within SCD, HbF levels are higher in HbSS compared to HbSC individuals(Steinberg, 2009). One component of HbF variability relates to the expanded erythron secondary to chronic hemolysis, and preferential survival of HbF-containing red cell precursors (F cells)(Quinn et al., 2016, Franco et al., 2006). A second component is the innate ability for HbF synthesis based on genetic variants at three quantitative trait loci (QTLs): *BCL11A* on chromosome 2p, *HMIP-2* on chromosome 6q and *Xmn1-HBG2 (rs7482144)* on chromosome 11p. Dependent upon the genetic variants investigated and analysis performed, such variants were found to account for between 8 and ~20% of the HbF variability in SCD in studies from the UK, USA, Brazil, Tanzania and Cameroon(Lettre et al., 2008, Bhatnagar et al., 2011, Bae et al., 2012, Makani et al., 2011, Mtatiro et al., 2014, Wonkam et al., 2014, Sebastiani et al., 2010, Cardoso et al., 2014). This genetic component is likely to account for much of the variable HbF levels in SCD patients. Consequently, it is desirable to quantify and summarize the effects of the respective genetic loci into a single genetic variable to capture the essence of genetic disease alleviation through the HbF mechanism. Here, we present such a genetic HbF summary variable – $\boldsymbol{HbF_G}$ - which will be a useful parameter to use as a covariate in genetic, biological, and clinical studies in diverse SCD populations.

Subjects and Methods

British patients are part of the South East London sickle gene bank (King's College Hospital "KCH", Guys and St Thomas' Hospitals Trust, Lewisham Hospital and Queen Elizabeth Hospital Woolwich). Written informed consent was obtained through three approved study protocols (LREC 01-083, 07/H0606/165, and 12/LO/1610) and research conducted in accordance with the Helsinki Declaration (1975, as revised 2008).

892 patients consented to the study, of whom 785 aged over 5 years had a full dataset (genotypes and phenotype). These 785 comprised: 581 with HbSS or HbSβ^0 (the "discovery cohort" used for the primary analysis), 186 HbSC (the "validation" cohort) and 18 HbSβ^+ thalassemia (figure 1). Additional validation was performed in the Muhimbili HbSS cohort, Tanzania (N=994)(Mtatiro et al., 2014).

Genetic Variants and Genotyping

We assembled an initial set of seven known (and widely replicated) HbF modifier variants, prioritizing those where additional functional evidence had been generated (Table 1): *BCL11A* -

rs1427407(Bauer et al., 2013) and *rs6545816*(Mtatiro et al., 2014); *HMIP-2A - rs66650371*(Farrell et al., 2011, Stadhouders et al., 2014) and *rs9376090(Menzel et al., 2014); HMIP-2B-rs9494145(Mtatiro et al., 2015b, Menzel et al., 2014)* and *rs9494142(Stadhouders et al., 2014);* Xmn1-*HBG2 - rs7482144*(Labie et al., 1985b, Labie et al., 1985a, Lettre et al., 2008).

A combination of three genotyping methodologies was used: (1) "manual" genotyping in the laboratory (all variants) by the TaqMan procedure, except *rs66650371* which was assayed by capillary electrophoresis (Applied Biosystems, Foster City, CA), as previously described(Menzel et al., 2014), (2) a genome-wide chip (Illumina Infinium Multi-Ethnic Genotyping Array), (3) imputation with public and in-house reference haplotypes (online supplementary methods). Alpha-thalassemia status was not associated with HbF levels in our cohort (data not shown).

Phenotypes

HbF levels (measured by HPLC, BioRad Variant II) - no red cell transfusion or hydroxyurea for at least 3 months, and not pregnant - were retrospectively collected. For the 581 HbSS/HbSβ^0 discovery set, median HbF level was 4.5% (IQR: 1.9-8.8%) (supplementary figure s1).

We estimated global disease severity using "hospitalization rate" as a measure of pain frequency, mortality and laboratory results. Mean hospitalization rates were calculated for KCH adults over 10 years (2004-2013), dividing an individual's number of hematology hospital admissions by the number of observed years. For the 302 patients with HbSS/HbSβ^0, median mean hospitalization rate was 0.25/year (IQR: 0-0.71) (supplementary figure s2). Mortality outcome was available for the 302 adults (1 January 2004-31 July 2015). Steady state laboratory values (hemoglobin, white blood cells (WBC)) over a 10-year period (2004-2013) were averaged for 278 patients.

Building and validating the genetic model for HbF%

Genetic association between the 7 genetic variants (as normalized genotype scores) with HbF (ln[%HbF]) was investigated by linear regression (using STATA12) under an additive allelic model. Manual linear regression modelling was carried out in the HbSS/HbSβ^0 thalassemia "discovery group" (see supplementary methods). We then validated the model – $\boldsymbol{HbF_G}$ – in two replication groups: (1) our own HbSC subgroup (N=186) and (2) a Tanzanian HbSS cohort (N=994)(Mtatiro et al., 2014).

Testing for association of $\boldsymbol{HbF_G}$ with clinical severity

See supplementary methods

Results and Discussion

Summary variables combining genotypes across HbF modifier loci have been found to be associated with clinical severity in β-thalassemia(Danjou et al., 2014), and have also been explored in SCD(Mtatiro et al., 2014, Milton et al., 2014, Leonardo et al., 2016, Mtatiro et al., 2015a). To represent the relationship between genetic factors and HbF more accurately and to build a summary variable that is robust across diverse SCD cohorts, we used regression modeling of the effect of seven known modifier variants (Table 1) on HbF levels in 581 SCD patients with HbSS and HbSβ^0 genotypes. We targeted genetic variants at the three major HbF loci that have been widely replicated and have been implicated as causative genetic variants. Preliminary analysis using basic regression with age/sex only yielded a model with r^2=0.1082, and with the 7 genetic variants only produced a model with r^2=0.2256. Putting age, sex and the 7 genetic variants together in the model increased the r^2 to 0.3167. As age and sex are roughly orthogonal to the variants, our subsequent analyses did not control for age/sex.

Final regression analysis resulted in a model utilizing 4 variants: *rs1427407*, *rs6545816* (both *BCL11A*), *rs66650371* (*HMIP-2A*) and *rs7482144* (*Xmn1-HBG2*), see table 1. *rs9376090*, *rs9494142* or *rs9494145* (all at *HMIP-2*) did not improve the model and were considered redundant. Applying this model, the predicted ln[HbF%]) - ***HbF$_G$*** – would be calculated:

$$\boldsymbol{HbF_G} = 1.89 + 0.14 \times \mathrm{rs6545816} + 0.3 \times \mathrm{rs1427407} + 0.13 \times \mathrm{rs66650371} + 0.1 \times \mathrm{rs7482144},$$

(genotype for each variant = 0, 1, or 2, according to the number of HbF-boosting alleles).
HbF$_G$ underlies 22% (r^2= 0.2178, p<0.0001) of the variability in HbF levels in our discovery group, and confirming its robustness, 23% in the Muhimbili 'replication group' (N=994) and 27.5% in HbSC patients (Table 1). In HbSC disease, the comparatively large effect of ***HbF$_G$*** is likely due to the less severe pathology and thus smaller influence of non-genetic hemolysis-related factors.

HbF levels affect the severity of SCD; patients with higher levels of HbF have fewer complications and live longer(Platt et al., 1994, Platt et al., 1991). We tested the influence of ***HbF$_G$*** on hospitalization rate in HbSS and HbSβ^0 patients, and detected tentative association (N=304, Beta=0.47, p=0.031), suggesting that a 2.7-fold increase in ***HbF$_G$*** would result in a 38% decrease in hospitalization frequency. Nevertheless, the ***HbF$_G$*** for frequently-admitted patients was not significantly changed. ***HbF$_G$*** was, however, associated with hemoglobin (N=278, Beta=17.871, p<0.001). We found no association of ***HbF$_G$*** with mortality or WBC.

Our cohort has potential power to investigate the influence of ***HbF$_G$*** on global measures of disease severity. International collaboration, larger sample sizes, adding new loci as they are discovered, and development of the formula will be required to realize the utility of the ***HbF$_G$*** variable. We saw no significant benefit for including the *HMIP-2B* locus(Menzel et al., 2014) in ***HbF$_G$***. This will be re-visited once the underlying functional variant has been identified.

We believe that estimating ***HbF$_G$***, or similar genetic summary variables, will add significant value to genetic and clinical studies, either to test the influence of genetic modifiers on outcomes, or to act as a co-variate to adjust for such effects. The strength of ***HbF$_G$*** is that it isolates the genetic component of HbF from the component *reactive* to disease severity. Using a pre-set formula for HbF such as ***HbF$_G$***, will be especially useful in smaller and medium-size cohorts or clinical trials, where *de novo* modelling is meaningless.

Acknowledgements

We thank Clive Stringer (system Delivery Manager, King's College Hospital) for the help in the data extraction from the EPR. We thank Charles Curtis and Sanghyuck Lee for their work processing the samples for the Illumina MEGA chip.

This work was supported by the Medical Research Council, UK to SLT (MRC No: G0001249 and ID62593) and a grant from Shire Pharmaceuticals to SM and SLT. SNewhouse is also supported by the National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre, and by awards establishing the Farr Institute of Health Informatics Research at UCL Partners, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, National Institute for Health Research, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1).

Authorship contributions

SM, TF and SLT designed the research study; KG, HR, JH and SLT collected data; KG, HR, NA, and NS performed experiments; TF, KG, SM, HP and SNewhouse analysed the data; KG, SM and SLT wrote the paper. SNkya, JMakani, RZS, JMgaya provided data from the Muhimbili Sickle cell biorepository (Dar es Salaam, Tanzania) for analysis. All authors participated in editing the final version of paper.

Conflict of interest

The authors declare that they have no competing interests.

References:

BAE, H. T., BALDWIN, C. T., SEBASTIANI, P., TELEN, M. J., ASHLEY-KOCH, A., GARRETT, M., HOOPER, W. C., BEAN, C. J., DEBAUN, M. R., ARKING, D. E., BHATNAGAR, P., CASELLA, J. F., KEEFER, J. R., BARRON-CASELLA, E., GORDEUK, V., KATO, G. J., MINNITI, C., TAYLOR, J., CAMPBELL, A., LUCHTMAN-JONES, L., HOPPE, C., GLADWIN, M. T., ZHANG, Y. & STEINBERG, M. H. 2012. Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood,* 120**,** 1961-2.

BALLAS, S. K. 2001. Effect of alpha-globin genotype on the pathophysiology of sickle cell disease. *Pediatr Pathol Mol Med,* 20**,** 107-21.

BAUER, D. E., KAMRAN, S. C., LESSARD, S., XU, J., FUJIWARA, Y., LIN, C., SHAO, Z., CANVER, M. C., SMITH, E. C., PINELLO, L., SABO, P. J., VIERSTRA, J., VOIT, R. A., YUAN, G. C., PORTEUS, M. H., STAMATOYANNOPOULOS, J. A., LETTRE, G. & ORKIN, S. H. 2013. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science,* 342**,** 253-7.

BHATNAGAR, P., PURVIS, S., BARRON-CASELLA, E., DEBAUN, M. R., CASELLA, J. F., ARKING, D. E. & KEEFER, J. R. 2011. Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J Hum Genet,* 56**,** 316-23.

CARDOSO, G. L., DINIZ, I. G., SILVA, A. N., CUNHA, D. A., SILVA JUNIOR, J. S., UCHOA, C. T., SANTOS, S. E., TRINDADE, S. M., CARDOSO MDO, S. & GUERREIRO, J. F. 2014. DNA polymorphisms at BCL11A, HBS1L-MYB and Xmn1-HBG2 site loci associated with fetal hemoglobin levels in sickle cell anemia patients from Northern Brazil. *Blood Cells Mol Dis,* 53**,** 176-9.

CLARKE, G. M., ANDERSON, C. A., PETTERSSON, F. H., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2011. Basic statistical analysis in genetic case-control studies. *Nat Protoc,* 6**,** 121-33.

CRAIG, J. E., ROCHETTE, J., FISHER, C. A., WEATHERALL, D. J., MARC, S., LATHROP, G. M., DEMENAIS, F. & THEIN, S. L. 1996. Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nature Genetics,* 12**,** 58-64.

DANJOU, F., FRANCAVILLA, M., ANNI, F., SATTA, S., DEMARTIS, F. R., PERSEU, L., MANCA, M., SOLLAINO, M. C., MANUNZA, L., MEREU, E., MARCEDDU, G., PISSARD, S., JOLY, P., THURET, I., ORIGA, R., BORG, J., FORNI, G. L., PIGA, A., LAI, M. E., BADENS, C., MOI, P. & GALANELLO, R. 2014. A genetic score for the prediction of beta-thalassemia severity. *Haematologica*.

DEVLIN, B. & ROEDER, K. 1999. Genomic control for association studies. *Biometrics,* 55**,** 997-1004.

EMBURY, S. H., DOZY, A. M., MILLER, J., DAVIS, J. R., JR., KLEMAN, K. M., PREISLER, H., VICHINSKY, E., LANDE, W. N., LUBIN, B. H., KAN, Y. W. & MENTZER, W. C. 1982. Concurrent sickle-cell anemia and alpha-thalassemia: effect on severity of anemia. *N Engl J Med,* 306**,** 270-4.

FARRELL, J. J., SHERVA, R. M., CHEN, Z. Y., LUO, H. Y., CHU, B. F., HA, S. Y., LI, C. K., LEE, A. C., LI, R. C., YUEN, H. L., SO, J. C., MA, E. S., CHAN, L. C., CHAN, V., SEBASTIANI, P., FARRER, L. A., BALDWIN, C. T., STEINBERG, M. H. & CHUI, D. H. 2011. A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood,* 117**,** 4935-45.

FRANCO, R. S., YASIN, Z., PALASCAK, M. B., CIRAOLO, P., JOINER, C. H. & RUCKNAGEL, D. L. 2006. The effect of fetal hemoglobin on the survival characteristics of sickle cells. *Blood,* 108**,** 1073-6.

KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S. Y., FREIMER, N. B., SABATTI, C. & ESKIN, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet,* 42**,** 348-54.

KNOWLER, W. C., WILLIAMS, R. C., PETTITT, D. J. & STEINBERG, A. G. 1988. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet,* 43**,** 520-6.

LABIE, D., DUNDA-BELKHODJA, O., ROUABHI, F., PAGNIER, J., RAGUSA, A. & NAGEL, R. L. 1985a. The -158 site 5' to the ^{G}g gene and ^{G}g expression. *Blood,* 66**,** 1463-1465.

LABIE, D., PAGNIER, J., LAPOUMEROULIE, C., ROUABHI, F., DUNDA-BELKHODJA, O., CHARDIN, P., BELDJORD, C., WAJCMAN, H., FABRY, M. E. & NAGEL, R. L. 1985b. Common haplotype dependency of high G gamma-globin gene expression and high Hb F levels in beta-

thalassemia and sickle cell anemia patients. *Proceedings of the National Academy of Sciences, USA,* 82**,** 2111-4.

LEONARDO, F. C., BRUGNEROTTO, A. F., DOMINGOS, I. F., FERTRIN, K. Y., DE ALBUQUERQUE, D. M., BEZERRA, M. A., ARAUJO, A. S., SAAD, S. T., COSTA, F. F., MENZEL, S., CONRAN, N. & THEIN, S. L. 2016. Reduced rate of sickle-related complications in Brazilian patients carrying HbF-promoting alleles at the BCL11A and HMIP-2 loci. *Br J Haematol,* 173**,** 456-60.

LETTRE, G., SANKARAN, V. G., BEZERRA, M. A., ARAUJO, A. S., UDA, M., SANNA, S., CAO, A., SCHLESSINGER, D., COSTA, F. F., HIRSCHHORN, J. N. & ORKIN, S. H. 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**,** 11869-74.

LISTGARTEN, J., LIPPERT, C., KADIE, C. M., DAVIDSON, R. I., ESKIN, E. & HECKERMAN, D. 2012. Improved linear mixed models for genome-wide association studies. *Nat Methods,* 9**,** 525-6.

MAKANI, J., MENZEL, S., NKYA, S., COX, S. E., DRASAR, E., SOKA, D., KOMBA, A. N., MGAYA, J., ROOKS, H., VASAVDA, N., FEGAN, G., NEWTON, C. R., FARRALL, M. & THEIN, S. L. 2011. Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. *Blood,* 117**,** 1390-2.

MANOLIO, T. A. 2013. Bringing genome-wide association findings into clinical use. *Nat Rev Genet,* 14**,** 549-58.

MARCHINI, J., CARDON, L. R., PHILLIPS, M. S. & DONNELLY, P. 2004. The effects of human population structure on large genetic association studies. *Nat Genet,* 36**,** 512-7.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69.

MENZEL, S., GARNER, C., GUT, I., MATSUDA, F., YAMAGUCHI, M., HEATH, S., FOGLIO, M., ZELENIKA, D., BOLAND, A., ROOKS, H., BEST, S., SPECTOR, T. D., FARRALL, M., LATHROP, M. & THEIN, S. L. 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet,* 39**,** 1197-9.

MENZEL, S., ROOKS, H., ZELENIKA, D., MTATIRO, S. N., GNANAKULASEKARAN, A., DRASAR, E., COX, S., LIU, L., MASOOD, M., SILVER, N., GARNER, C., VASAVDA, N., HOWARD, J., MAKANI, J., ADEKILE, A., PACE, B., SPECTOR, T., FARRALL, M., LATHROP, M. & THEIN, S. L. 2014. Global Genetic Architecture of an Erythroid Quantitative Trait Locus, HMIP-2. *Ann Hum Genet*.

MILTON, J. N., GORDEUK, V. R., TAYLOR, J. G. T., GLADWIN, M. T., STEINBERG, M. H. & SEBASTIANI, P. 2014. Prediction of fetal hemoglobin in sickle cell anemia using an ensemble of genetic risk prediction models. *Circ Cardiovasc Genet,* 7**,** 110-5.

MILTON, J. N., ROOKS, H., DRASAR, E., MCCABE, E. L., BALDWIN, C. T., MELISTA, E., GORDEUK, V. R., NOURAIE, M., KATO, G. R., MINNITI, C., TAYLOR, J., CAMPBELL, A., LUCHTMAN-JONES, L., RANA, S., CASTRO, O., ZHANG, Y., THEIN, S. L., SEBASTIANI, P., GLADWIN, M. T. & STEINBERG, M. H. 2013. Genetic determinants of haemolysis in sickle cell anaemia. *Br J Haematol,* 161**,** 270-8.

MILTON, J. N., SEBASTIANI, P., SOLOVIEFF, N., HARTLEY, S. W., BHATNAGAR, P., ARKING, D. E., DWORKIS, D. A., CASELLA, J. F., BARRON-CASELLA, E., BEAN, C. J., HOOPER, W. C., DEBAUN, M. R., GARRETT, M. E., SOLDANO, K., TELEN, M. J., ASHLEY-KOCH, A., GLADWIN, M. T., BALDWIN, C. T., STEINBERG, M. H. & KLINGS, E. S. 2012. A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. *PLoS One,* 7**,** e34741.

MTATIRO, S. N., MAKANI, J., MMBANDO, B., THEIN, S. L., MENZEL, S. & COX, S. E. 2015a. Genetic variants at HbF-modifier loci moderate anemia and leukocytosis in sickle cell disease in Tanzania. *Am J Hematol,* 90**,** E1-4.

MTATIRO, S. N., MGAYA, J., SINGH, T., MARIKI, H., ROOKS, H., SOKA, D., MMBANDO, B., THEIN, S. L., BARRETT, J. C. & MAKANI, J. 2015b. Genetic association of fetal-hemoglobin levels in

individuals with sickle cell disease in Tanzania maps to conserved regulatory elements within the MYB core enhancer. *BMC medical genetics,* 16**,** 1.

MTATIRO, S. N., SINGH, T., ROOKS, H., MGAYA, J., MARIKI, H., SOKA, D., MMBANDO, B., MSAKI, E., KOLDER, I., THEIN, S. L., MENZEL, S., COX, S. E., MAKANI, J. & BARRETT, J. C. 2014. Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One,* 9**,** e111464.

NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S., NELSON, M. R., STEPHENS, M. & BUSTAMANTE, C. D. 2008. Genes mirror geography within Europe. *Nature,* 456**,** 98-101.

PATTERSON, N., PRICE, A. L. & REICH, D. 2006. Population structure and eigenanalysis. *PLoS Genet,* 2**,** e190.

PLATT, O. S., BRAMBILLA, D. J., ROSSE, W. F., MILNER, P. F., CASTRO, O., STEINBERG, M. H. & KLUG, P. P. 1994. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med,* 330**,** 1639-44.

PLATT, O. S., THORINGTON, B. D., BRAMBILLA, D. J., MILNER, P. F., ROSE, W. F., VICHINSKY, E. & KINNEY, T. R. 1991. Pain in sickle cell disease: Rates and risk factors. *New England Journal of Medicine,* 325**,** 11-6.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38**,** 904-9.

QUINN, C. T., SMITH, E. P., ARBABI, S., KHERA, P. K., LINDSELL, C. J., NISS, O., JOINER, C. H., FRANCO, R. S. & COHEN, R. M. 2016. Biochemical surrogate markers of hemolysis do not correlate with directly measured erythrocyte survival in sickle cell anemia. *Am J Hematol,* 91**,** 1195-1201.

RISCH, N. & MERIKANGAS, K. 1996. The future of genetic studies of complex human diseases. *Science,* 273**,** 1516-7.

SARAF, S. L., SHAH, B. N., ZHANG, X., HAN, J., TAYO, B. O., ABBASI, T., OSTROWER, A., GUZMAN, E., MOLOKIE, R. E., GOWHARI, M., HASSAN, J., JAIN, S., COOPER, R. S., MACHADO, R. F., LASH, J. P. & GORDEUK, V. R. 2017. APOL1, alpha-thalassemia, and BCL11A variants as a genetic risk profile for progression of chronic kidney disease in sickle cell anemia. *Haematologica,* 102**,** e1-e6.

SEBASTIANI, P., SOLOVIEFF, N., HARTLEY, S. W., MILTON, J. N., RIVA, A., DWORKIS, D. A., MELISTA, E., KLINGS, E. S., GARRETT, M. E., TELEN, M. J., ASHLEY-KOCH, A., BALDWIN, C. T. & STEINBERG, M. H. 2010. Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study. *Am J Hematol,* 85**,** 29-35.

STADHOUDERS, R., AKTUNA, S., THONGJUEA, S., AGHAJANIREFAH, A., POURFARZAD, F., VAN IJCKEN, W., LENHARD, B., ROOKS, H., BEST, S., MENZEL, S., GROSVELD, F., THEIN, S. L. & SOLER, E. 2014. HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J Clin Invest,* 124**,** 1699-710.

STEINBERG, M. H. 2009. Genetic etiologies for phenotypic diversity in sickle cell anemia. *ScientificWorldJournal,* 9**,** 46-67.

TEAM, R. D. C. 2011. R: A language and environment for statistical computing. . Foundation for Statistical Computing, Vienna, Austria. .

UDA, M., GALANELLO, R., SANNA, S., LETTRE, G., SANKARAN, V. G., CHEN, W., USALA, G., BUSONERO, F., MASCHIO, A., ALBAI, G., PIRAS, M. G., SESTU, N., LAI, S., DEI, M., MULAS, A., CRISPONI, L., NAITZA, S., ASUNIS, I., DEIANA, M., NAGARAJA, R., PERSEU, L., SATTA, S., CIPOLLINA, M. D., SOLLAINO, C., MOI, P., HIRSCHHORN, J. N., ORKIN, S. H., ABECASIS, G. R., SCHLESSINGER, D. & CAO, A. 2008. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc Natl Acad Sci U S A,* 105**,** 1620-5.

WANG, K., HU, X. & PENG, Y. 2013. An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. *Hum Hered,* 76**,** 1-9.
WEALE, M. E. 2010. Quality control for genome-wide association studies. *Methods Mol Biol,* 628**,** 341-72.
WELLCOME_TRUST_CASE_CONTROL_CONSORTIUM 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.
WONKAM, A., NGO BITOUNGUI, V. J., VORSTER, A. A., RAMESAR, R., COOPER, R. S., TAYO, B., LETTRE, G. & NGOGANG, J. 2014. Association of variants at BCL11A and HBS1L-MYB with hemoglobin F and hospitalization rates among sickle cell patients in Cameroon. *PLoS One,* 9**,** e92506.
YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. & VISSCHER, P. M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-9.
YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82.
YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. & PRICE, A. L. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet,* 46**,** 100-6.
YU, J., PRESSOIR, G., BRIGGS, W. H., VROH BI, I., YAMASAKI, M., DOEBLEY, J. F., MCMULLEN, M. D., GAUT, B. S., NIELSEN, D. M., HOLLAND, J. B., KRESOVICH, S. & BUCKLER, E. S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet,* 38**,** 203-8.
ZAITLEN, N. & KRAFT, P. 2012. Heritability in the genome-wide association era. *Hum Genet,* 131**,** 1655-64.

Legends:

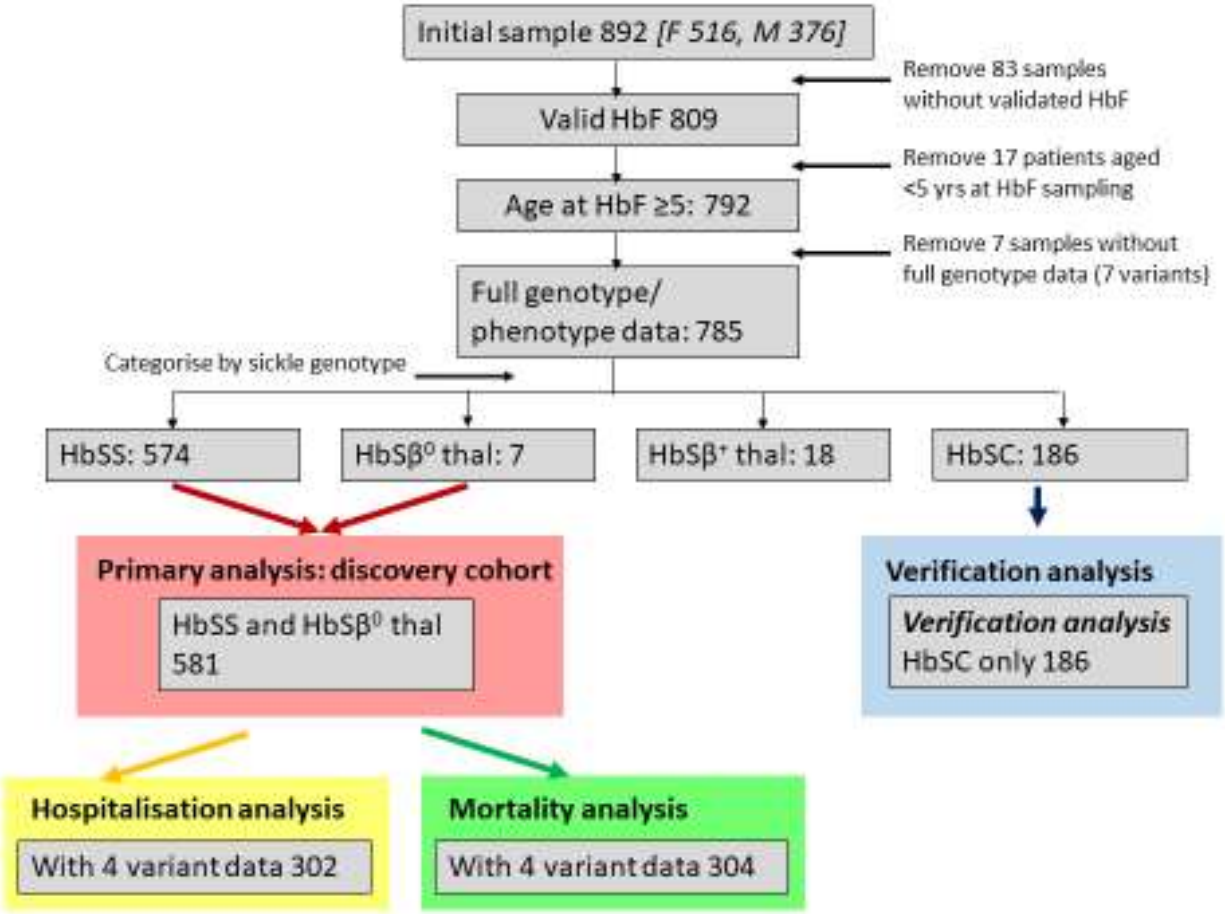
Table 1: Seven variants across the three major HbF QTLs representing the most recent biological understanding of the role of *BCL11A* and *HMIP* variants. The table also includes HbF_G, the new 4-variant model to represent the genetic component of HbFin HbSS and $HbS\beta^0$ thalassemia

Figure 1: Study sample fate

Table 1: Seven variants across the three major HbF QTLs representing the most recent biological understanding of the role of BCL11A and HMIP variants. The table also includes HbF_G, the new 4-variant model to represent the genetic component of HbF

Gene		Variant	HbF-boosting allele	Allele frequency		Results: **HbF_G** model	
				London UK	Dar es Salaam Tanzania	Coefficient (Beta) (95% CI)	p-value
BCL11A (Chr 2)		rs6545816 (A>C)	C	0.34	0.36	0.14 (0.08- 0.19)	<0.001
		rs1427407 (G>T)	T	0.26	0.21	0.30 (0.26-0.35)	<0.001
HMIP (HBSL1-MYB intergenic polymorphism on Chr 6)	2A	rs9376090 (T>C)	C	0.01			
		rs66650371 (TAC>---)	3bp del	0.04	0.02	0.13 (0.08-0.17)	<0.001
	2B	rs9494142 (T>C)	C	0.05			
		rs9494145 (T>C)	C	0.05			
HBG2-Xmn1 (β-globin gene cluster on Chr 11)		rs7482144 (G>A)	A	0.07	0.01	0.10 (0.05-0.15)	<0.001

Figure 1: Study sample fate



Chapter 5: Genetics of severity of sickle cell disease: a candidate gene study in sickle cell disease using two severity indices

Figures 198

Tables 198

5.1. Introduction 199
5.1.1. Background 199
5.1.2. Phenotypes 201
5.1.3. Regions of interest 201
5.1.3.1. Choosing regions of interest 201
5.1.3.2. Pyruvate kinase 202
5.1.3.3. Adenosine A_{2B} receptor 204
5.1.3.4. Complement related genes (MCP, CFH, CFB, ADAMTS13) 204
5.1.4. Specific variants previously implicated in sickle cell disease severity 205
5.1.4.1. APOL1 205
5.1.4.2. DARC 205
5.1.4.3. MAPK8 206

5.2. Methods 207
5.2.1. Summary of methodological approach 207
5.2.2. Defining regions of interest 208
5.2.3. Defining significance levels 210
5.2.3.1. Significance levels for regions of interest 210
5.2.3.2. Significance levels for specific genetic variants 210
5.2.4. Statistical analysis 210
5.2.5. Post-analysis steps 211

5.3. Results 212
5.3.1. Regions of interest 212
5.3.1.1. Pyruvate kinase 212
5.3.1.2. 2,3-diphosphoglycerate 214
5.3.1.3. Adenosine A_{2B} receptor 214
5.3.1.4. Complement related genes 214
5.3.2. Replication of specific variants 214
5.3.2.1. APOL1 214
5.3.2.2. DARC 214
5.3.2.3. MAPK8 215

5.4. Discussion 215

References 218

Appendix 1 222

Appendix 2 227

Figures

Figure 1 The glycolytic pathway, taken from http://laboratoryinfo.com/glycolysis-steps-diagram-energy-yield-and-significance/. 2,3-DPG is formed in stage 7. 203
Figure 2 Region of interest script interface for user to answer questions. 211
Figure 3 Hospitalisation rate (double logarithm) by genotype for the significant variants in the linear mixed modelling: (a) rs8177970 (b) rs116244351 (c) rs114455416 (d) rs8177964 213
Figure 4 UCSC genome browser: Layered H3K27Ac track (K562 cells) for PK-LR 216

Tables

Table 1 Established APOL1 variants associated with sickle- and non-sickle renal dysfunction 205
Table 2 Allele frequencies for rs2814778 in African (AFR) and European (EUR) in 1000 Genomes Phase 3 dataset 206
Table 3 Allele frequencies for rs10857560 in African (AFR) and European (EUR) in 1000 Genomes Phase 3 dataset 207
Table 4 Candidate regions of interest 209
Table 5 Variants associated with sickle phenotypes to be replicated in our linear mixed model 209
Table 6 Association of PK-LR variants with hospitalisation rate (LnLnHospRate) in HbSS and HbSC 213
Table 7 Association of APOL1 G1 variants with urinary albumin creatinine ratio (uACR) in HbSS and HbSC 214

5.1. Introduction

5.1.1. Background

Two paradigms exist for assessing genotype-phenotype correlation: GWAS and candidate gene association studies. Both are based on genotyping of human polymorphisms. Candidate gene studies are based on testing an *a priori* hypothesis that a specific genetic region of interest is associated with trait risk (either disease status or a quantitative trait). Better than a hypothesis is the availability of actual prior evidence, whether that is genetic, biological or clinical. Candidate gene studies thus provide a focused view of genomic regions of interest, hypothesised to be associated with the trait.

Candidate gene studies have been at the forefront of genetic association studies for as long as the genetic era. These studies evaluate genetic variants within a gene or region of interest that has been in some way related to the disease previously. The candidate region may be motivated by prior functional studies (the gene is known from basic science (especially pathway analysis) or clinical studies) or by genetic position (the gene lies within a broad region found by linkage analysis).

The candidate gene approach begins with selection of a candidate region of interest based on its relevance to the mechanism of the trait being evaluated. The genetic variant is then analysed for association with the trait.

Candidate gene studies have multiple advantages. Because the approach is hypothesis-driven, it allows for targeted assessment of selected alleles relevant to the hypothesis, in the chosen study population. Within the targeted candidate genes / genetic markers, this approach may confer inferential advantages in comparison with GWAS which is untargeted, and where coverage is genome-wide and typically does not specifically target functional variants or regions of interest. The low number of markers tested provides superior power to identify significant associations. This enhanced power to detect associations is particularly important when allele frequencies are low, effect sizes are small, or the study population is small. This makes candidate gene studies a good approach for asking research questions in our (relatively) small cohort. Because the candidate variants have been specifically selected, the credibility and interpretability of significant associations are frequently greater than in GWA studies. Furthermore, candidate gene approaches are relatively cheap and quick to perform (for well-focused research questions). Candidate gene analysis is also valuable for replicating previous reports of genetic associations with disease in different populations.

Candidate gene studies also have disadvantages, primarily its limitation from the accuracy of the candidate selection. Precisely because it is hypothesis-driven, it cannot identify anything unexpected precluding the discovery of novel associations. It is constrained by our understanding of the trait(s) being studied: this understanding is particularly poor for sickle pathophysiology. Until recently, candidate gene association studies have been considered somewhat unreliable, after findings often could not be replicated in independent follow-up studies. Several factors have contributed to this. First, the lack of management of relatedness and population stratification. By using a statistical model which encompasses genome-wide data to capture relatedness in all forms, this can be circumvented. I have done this using the MEGA data set, constructing a genetic relatedness matrix and then using a linear mixed model to control for all forms of relatedness, both near and far. Second, the necessity of replication of positive findings in an independent cohort has often been disregarded (with different population characteristics e.g. different ethnicity, admixture frequency). Third, multiple testing has often been ignored during the estimation of statistical significance, leading to false positive findings. Careful consideration must be given to an appropriate significance level. This can be managed by adjusting the p-value to take account of multiple testing. Testing *m* multiple variants within a gene is not equivalent to *m* independent tests, since genotypes are dependent upon each other due to linkage disequilibrium. One can assess the linkage disequilibrium between variants to calculate an effective number of independent variants (and hence effective number of tests).

The HapMap and 1000 Genomes Projects have both provided rich data by thoroughly cataloguing human polymorphisms, more recently encompassing diverse ethnic groups, including African-heritage populations. As well as enabling a GWAS approach, this has also made it possible to interrogate candidate genes in finer detail, leading to a revival of candidate gene strategies for the dissection of complex genetic disorders.

Our sample size has been too small to power a successful GWAS for sickle severity indices in our cohort, but we have reasonable expectations of meaningful association results by carefully targeting candidate genes to a limited but thoroughly curated set of phenotypes recorded in our patient cohort.

I have assessed several regions of interest consisting of the candidate gene and immediately adjacent DNA sequence. By using a linear mixed model to account for relatedness, I have avoided potential problems with population stratification that occur with simpler statistical association models. I have used two quantitative traits as measures of sickle severity:

hospitalisation rate and a haemolytic index. For each of these severity indices, I considered two groups of genes. First, those which increase red cell intracellular 2,3-DPG levels and so increase deoxy-HbS thus promoting sickling. Of particular note here is the gene for the erythrocyte form of pyruvate kinase (*PK-LR*). Pyruvate kinase deficiency has been associated with a sickle cell phenotype in patients with sickle trait (HbAS) (Alli et al., 2008, Cohen-Solal et al., 1998). Second, I evaluated a group of complement-related genes, because of the parallels between atypical haemolytic uraemic syndrome (aHUS, a recurrent micro-angiopathic haemolytic anaemia) and sickle cell disease.

As well as evaluating regions of interest, I also assessed whether specific genetic variants previously associated with morbidity in sickle cell disease could be replicated in our cohort: *APOL1* and proteinuria, *DARC* and both indices of sickle severity, *MAPK8* and haemolysis. Notably, many historical studies do not utilise a linear mixed model approach and there are specific concerns about population stratification and admixture in these studies which I have addressed.

5.1.2. Phenotypes

I have evaluated two quantitative markers of severity of sickle cell disease: a haemolytic index and hospitalisation rate. I have investigated if these two "severity" indices are associated with the described candidate "regions of interest" as well as other specific variants previously associated with sickle severity in *DARC* and *MAPK8*. I have also assessed the relationship between urinary albumin creatinine ratio (uACR) and specific variants in *APOL1*.

5.1.3. Regions of interest

5.1.3.1. Choosing regions of interest

A precondition of the selection of regions of interest is that there is *strong* prior evidence of their connection with the phenotype of interest (or a similar phenotype). Therefore, genetic association seen with a particular phenotype (e.g. proteinuria) in the non-sickle setting may provide ideal candidates for assessing proteinuria in SCD. Evidence for the importance of genes or regions of interest may come from biological, clinical or genetic studies that link the candidate with sickle severity of specific complications of sickle cell disease. Thus, the most common source of candidate genes and/or polymorphisms is the existing literature.

This choice must be objective, there must be strong prior independent evidence, otherwise one should adopt genome-wide significance thresholds.

To reduce a set of variants down to a smaller number, databases and predictive software tools can be used to choose missense, nonsense and splice site variants over those without immediate obvious functional consequence.

5.1.3.2. Pyruvate kinase

2,3-diphosphoglycerate (2,3-DPG) decreases oxygen affinity of red cells, including of those from patients with HbSS(Charache et al., 1970). Sickling occurs in conditions favouring deoxygenation and subsequent polymerisation of HbS in red blood cells. Theoretically, then, any mechanism that increases intra-cellular deoxy-HbS concentration facilitates sickling. Increased 2,3-DPG concentration and decreased intracellular red cell pH have both been shown to boost deoxy-HbS polymerisation(Poillon and Kim, 1990, Poillon et al., 1998).

2,3-DPG is synthesized in RBCs as an intermediate substrate in the glycolytic pathway(Rapoport and Luebering, 1952), see Figure 1. Enzyme deficiencies in the glycolytic pathway (including pyruvate kinase (PK) deficiency) lead to the accumulation of upstream enzyme substrates, including 2,3-DPG.

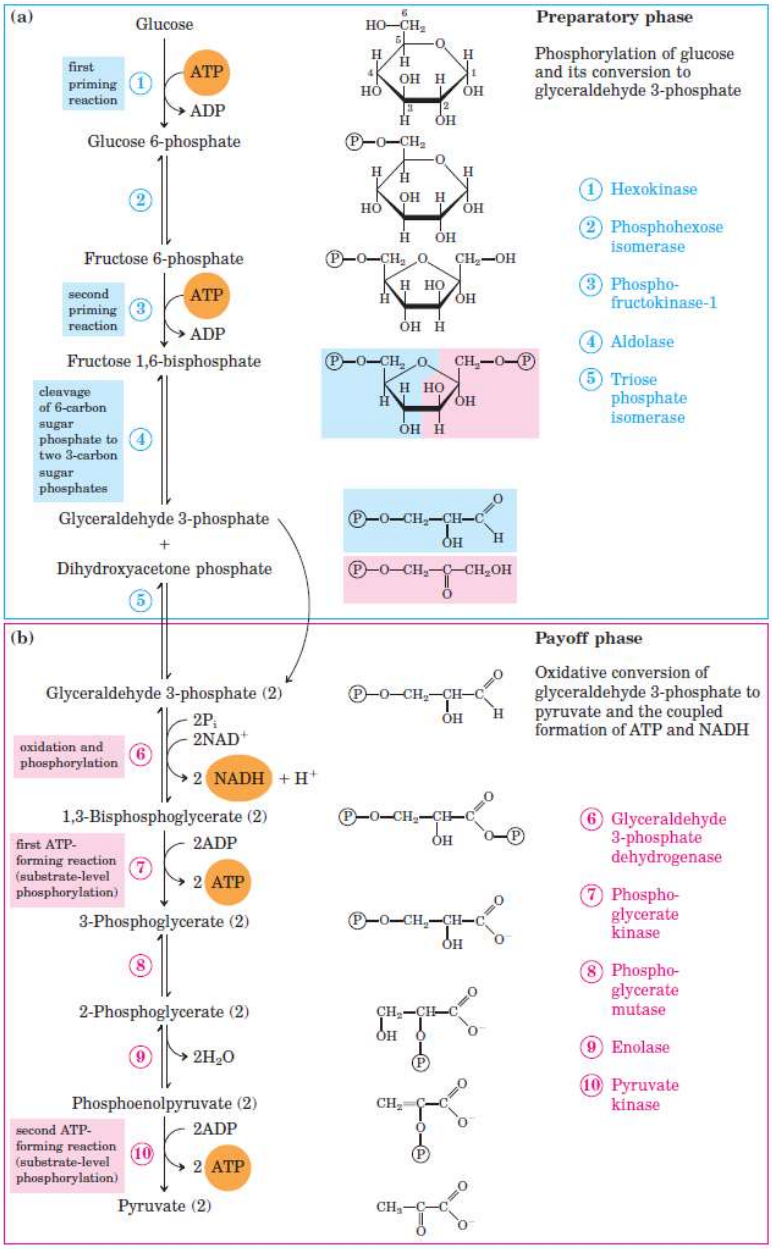


Figure 1 The glycolytic pathway, taken from http://laboratoryinfo.com/glycolysis-steps-diagram-energy-yield-and-significance/. 2,3-DPG is formed in stage 7.

PK deficiency is a rare red cell disorder which manifests as a haemolytic anaemia. The anaemia is partly compensated by a right-shift of the oxygen dissociation curve to increase tissue oxygenation. Mechanistically, this happens via an increase of upstream glycolytic pathway substrates which includes 2,3-DPG concentration; the rise in 2,3-DPG causes reduced haemoglobin oxygen affinity. In SCD, and even in sickle trait, reduced oxygen affinity will favour deoxy-HbS polymerisation - and thus sickling. The combination of PK deficiency and sickle cell trait causing a sickling syndrome has been reported in two cases(Alli et al., 2008, Cohen-Solal et al., 1998). PK levels are a spectrum, and in SCD, in theory, this could represent a quantitative trait which modifies the risk of sickling to induce a more severe SCD phenotype.

In red blood cells, PK is produced by *PK-LR* (PK in liver and red cells). 2,3-DPG is produced via *BPGM* (biphosphoglycerate mutase).

5.1.3.3. Adenosine A_{2B} receptor

Adenosine A_{2B} receptor can also act to induce 2,3-DPG and decrease oxygen-binding affinity of haemoglobin, ultimately inducing sickling. Adenosine A_{2B} receptor abnormalities have been associated with "haemolytic" type complications of SCD. The first association between adenosine A_{2B} receptor and pulmonary hypertension is seen in both the non-sickle settings (Karmouty-Quintana et al., 2012, Karmouty-Quintana et al., 2013), and in SCD(Desai et al., 2012). The second association is seen with murine priapism in a mouse model of SCD: excess adenosine contributed to priapism via adenosine A_{2B} receptor signalling(Mi et al., 2008).

There are high adenosine concentrations in the blood of patients with SCD. Increased adenosine levels promote sickling and haemolysis(Zhang et al., 2011). The adenosine A_{2B} receptor mediates induction of 2,3-DPG, thus decreasing the oxygen-binding affinity of HbS. Adenosine signalling (via the adenosine A_{2B} receptor) therefore has a pathological role in inducing sickling by excess adenosine.

The Adenosine A_{2B} receptor is encoded on the gene *ADORA2B*.

5.1.3.4. Complement related genes (MCP, CFH, CFB, ADAMTS13)

Arguments for the potential involvement of the complement system in acute pain episodes of SCD stem from the parallels with micro-angiopathic haemolytic anaemias, especially in more severe cases of acute pain episodes. In addition to the reported explicit cases of thrombotic thrombocytopenic purpura (TTP) in SCD(Prichard et al., 1988, Chinowsky, 1994, Geigel and Francis, 1997, Bolanos-Meade et al., 1999, Lee et al., 2003, Venkata Sasidhar et al., 2010, Shelat, 2010), we and others have described cases of "extreme haemolysis" during acute pain episodes, associated with markedly elevated LDH and thrombocytopenia(Shome et al., 2013, Gardner and Thein, 2015). Furthermore, plasma exchange has historically been used successfully in acute multi-organ failure syndromes associated with gross haemolysis in SCD (Geigel and Francis, 1997). Moreover, standard treatment of atypical haemolytic uraemic syndrome (aHUS) using eculizumab has been translated to the sickle setting for delayed haemolytic transfusion reactions(Pirenne et al., 2017, Dumas et al., 2016, Boonyasampant et al., 2015), transplant-associated thrombotic microangiopathy(Abusin et al., 2017) as well as frank aHUS(Chonat et al., 2016).

In the setting of aHUS, multiple complement mutations have been associated with the disease, including variants in *MCP* (CD46), *CFH*, *CFB*, thrombomodulin (*THBD*), diacylglycerol kinase epsilon (*DGKE*), and C3(Phillips et al., 2016, Kavanagh and Goodship, 2010). These variants contribute to a variety of severity and frequency of aHUS events as well as risk of progression to end stage kidney disease.

5.1.4. Specific variants previously implicated in sickle cell disease severity

5.1.4.1. APOL1

Several polymorphisms at the *APOL1* locus have been associated with kidney disease in African-heritage individuals in both the non-sickle(Genovese et al., 2010, Kopp et al., 2011, Larsen et al., 2013) and sickle settings (Ashley-Koch et al., 2011, Saraf et al., 2017, Saraf et al., 2015, Schaefer et al., 2016, Kormann et al., 2017). In the non-sickle setting, it is proposed that *APOL1* risk variants function as a "second hit" that promotes kidney disease progression secondary to specific chronic kidney injury(Freedman and Skorecki, 2014) – namely, focal segmental glomerulosclerosis, HIV-associated nephropathy and severe lupus nephritis in order of reference above.

Two *APOL1* alleles, G1 and G2, have been associated with nephropathy, see Table 1. G1 has two missense variants (rs73885319 / S342G and rs60910145 / I384M). G2 comprises a deletion of six base pairs (rs71785313 is both N388 *and* Y389 deletion). In some populations, positive selection for these mutations occurred due to their association with Trypanosome resistance (Genovese et al., 2010): those who have at least one copy of the alternate allele of G1/G2 are resistant to Trypanosome infection. In African-Americans, the prevalence of G1 *or* G2 variants is 10-15%.

Table 1 Established APOL1 variants associated with sickle- and non-sickle renal dysfunction

	Variant	hg19 coordinates	Ref> alt allele	Protein effect
G1	rs73885319	36661906	A>G	S342G
	rs60910145	36662034	T>G	I384M
G2	rs71785313	36662051	TTATAA>-	N388 deletion Y389 deletion

5.1.4.2. DARC

"Benign ethnic neutropenia" is common in African-heritage individuals. This is caused by the mutation rs2814778 within the *DARC* (Duffy Antigen Receptor for Chemokines) gene promoter, resulting in both a lower white cell count and null Duffy expression on red blood cells(Reich et al., 2009).

In SCD, studies investigating the association between rs2814778 and a variety of sickle end-organ complications have given inconsistent results, with positive results made *and* refuted, both by our group and others(Drasar et al., 2013, Afenyi-Annan et al., 2008, Mecabo et al., 2010, Schnog et al., 2000, Nebor et al., 2010, Araujo et al., 2015). The unresolved question of association may be due to small sample sizes, however, all six studies fail to account for population structure in their analyses; only the promoter +/- other DARC variants have been genotyped. Controlling for population structure is crucial for this variant; population genetics demonstrate widely varying allele frequencies in different ethnic populations, see Table 2. Thus, population stratification present in our cohort is likely to be tracked by rs2814778, with the T allele associated with European admixture (in our African-Caribbean patients) and the C allele over-represented in our African patients due to lack of such admixture in this group. It is paramount that statistical modelling takes account of relatedness (including population stratification) when assessing the variant given its population genetics. This current study will provide the first robust assessment of rs2814778 in SCD.

Table 2 Allele frequencies for rs2814778 in African (AFR) and European (EUR) in 1000 Genomes Phase 3 dataset

	C	T
AFR	96%	4%
EUR	1%	99%

Despite these cautions surrounding interpretation of previous genetic association analyses, it must also be acknowledged that functional work has associated DARC red cell expression and the inflammatory response in SCD(Durpes et al., 2010, Durpes et al., 2011), strengthening the underlying biology. Furthermore, it is long-established that higher white cell counts are associated with increased disease severity in SCD(Platt et al., 1994, Miller et al., 2000): recall that rs2814778 causes benign ethnic neutropenia.

5.1.4.3. MAPK8

Zhang *et al* identified a *MAPK8* variant (rs10857560) to be associated with pre-capillary pulmonary hypertension in SCD (Zhang et al., 2014). The authors concluded this after identifying *MAPK8* via a gene expression approach which compared variation of expression in peripheral blood mononuclear cells from SCD and Chuvash polycythaemia cohorts, and then using standard regression modelling (with consideration of population structure) looking at multiple *MAPK8* variants.

In SCD, pulmonary hypertension is considered to be a consequence of excess haemolysis(Kato et al., 2017). Hence, we postulate that *MAPK8* is a candidate gene as a quantitative trait locus

for haemolytic rate in SCD. There is diversity of allele frequencies of rs10857560 in African versus European populations, see Table 3. In order to assess the significance of rs10857560, a statistical modelling approach that accounts for relatedness is important.

Table 3 Allele frequencies for rs10857560 in African (AFR) and European (EUR) in 1000 Genomes Phase 3 dataset

	A	C
AFR	66%	34%
EUR	39%	61%

5.2. Methods

5.2.1. Summary of methodological approach

I have identified regions of interest using a systematic approach to decide which genomic sequence to interrogate, see section 5.2.2. The next step is to define significance levels against which to assess association, see section 5.2.3. Finally, statistical analysis can be undertaken, see section 5.2.4.

For each specific region of interest, for the relevant phenotype, I extracted existing linear mixed model results from the genome-wide analysis described in chapter 4. I considered the analysis of the HbSS group as a "discovery cohort" and the HbSC group as a "replication set". Relatedness cut-off for exclusion was >0.2 (i.e. one of a pair of any first or second degree relatives were excluded from analysis). As before, age and sex were used as fixed covariates in the linear mixed model described. As in the genome-wide studies for these outcomes, only patients in the KCH adult clinic were considered. I considered two "global severity" clinical phenotypes:

- Hospitalisation rate (using the double natural logarithm to improve normalisation of the data): N=242 HbSS patients, 94 HbSC patients, 354 total patients (all genotypes)
- Haemolytic index (using average results over 10 years): N=216 HbSS patients, 89 HbSC patients, 328 (all genotypes)

I also assessed previous reports of specific variants and their association with phenotypes in our cohort, again used the linear mixed model analysis. I investigated specific *APOL1* variants and urinary Albumin Creatinine Ratio (uACR) levels (using the double natural logarithm of the average results over 10 years): N=229 HbSS patients, N=80 HbSC patients. I also assessed *DARC* and both indices of sickle severity (numbers as above), and *MAPK8* and the haemolytic index.

5.2.2. Defining regions of interest

Regions of interest were defined as the transcribed transcript for the gene – the "canonical transcript"[1] – which includes the untranslated regions upstream and downstream (i.e. 5' UTR and 3' UTR). The exact length of a promoter can often only be defined experimentally so I added to the transcribed region 500bp directly upstream of the transcription initiation site to capture the promoter (taking note of the strand). Precise regions of interest are in Table 4.

Specific variants to be assessed for replication in our cohort are defined in Table 5.

[1] The canonical transcript is set according to the hierarchy: 1. Longest CCDS translation with no stop codons. 2. If no (1), choose the longest Ensembl/Havana merged translation with no stop codons. 3. If no (2), choose the longest translation with no stop codons. 4. If no translation, choose the longest non-protein-coding transcript.

Table 4 Candidate regions of interest

	Gene	Chr	Canonical transcript	Strand	hg19 coordinates	Region of interest	References
2,3-DPG – boosting candidates	2,3-bisphosphoglycerate mutase, *BPGM*	7	ENST00000393132	+	134331560-134364565	134331060-134364565	
	Pyruvate kinase, liver and RBC, *PK-LR*	1	ENST00000342741	-	155259630-155271225	155259630-155271725	(Alli et al., 2008, Cohen-Solal et al., 1998)
	Adenosine A2b receptor, *ADORA2B*	17	ENST00000304222	+	15848231-15879060	15847731-15879060	(Desai et al., 2012, Mi et al., 2008)
Complement/ MAHA related	*MCP* (CD46)	1	ENST00000322875	+	207925402-207968858	207924902-207968858	(Phillips et al., 2016, Kavanagh and Goodship, 2010)
	CFH	1	ENST00000367429	+	196621008-196716634	196620508-196716634	(Phillips et al., 2016, Kavanagh and Goodship, 2010)
	CFB	6	ENST00000425368	+	31895475-31919861	31894975-31919861	(Phillips et al., 2016, Kavanagh and Goodship, 2010)
	Thrombomodulin *THBD*	20	ENST00000377103	-	23026270-23030378	23026270-23030878	(Kavanagh and Goodship, 2010)
	Diacylglycerol kinase, epsilon, *DGKE*	17	ENST00000284061	+	54911460-54946036	54910960-54946036	(Kavanagh and Goodship, 2010)
	Complement component 3	21	ENST00000245907	-	6677715-6730573	6677715-6731073	(Kavanagh and Goodship, 2010)
	ADAMTS13	9	ENST00000371929	+	136279478-136324508	136278978-136324508	(Lotta et al., 2010)

Table 5 Variants associated with sickle phenotypes to be replicated in our linear mixed model

Gene	Chr		Variant	hg19 coordinates	References
Apolipoprotein L, 1, *APOL1*	22	G1	rs73885319	36661906	(Ashley-Koch et al., 2011, Saraf et al., 2017, Saraf et al., 2015, Schaefer et al., 2016, Kormann et al., 2017)
			rs60910145	36662034	
		G2	rs71785313	36662051	
Duffy atypical chemokine receptor, *DARC*	1		rs2814778	159174683	(Drasar et al., 2013, Afenyi-Annan et al., 2008, Mecabo et al., 2010, Schnog et al., 2000, Nebor et al., 2010)
Mitogen-activated protein kinase 8, *MAPK8*	10		rs10857560	49594240	(Zhang et al., 2014)

5.2.3. Defining significance levels

5.2.3.1. Significance levels for regions of interest

Defining significance levels in larger scale genetic analysis is not trivial (see also chapter 4 for discussing similar issues in genome-wide analysis). I corrected for multiple testing at a region of interest using an approach described by Cheverud(Cheverud et al., 2001). In this method, the linkage disequilibrium within a genomic region is quantitatively assessed to produce an "effective number of tests" rather than the total number of tests (=number of genetic variants). The method is implemented with the following four steps:

1) Calculation of the correlation matrix for the variants in PLINK
2) Estimation of the effective number (t.eff) of independent tests from the eigenvalues of the correlation matrix using R.
    a) t.eff=1+(t-1)*(1-var(λ_i)/t)
    b) t is the number of tests (equal to the number of genetic variants) and λ_i (i=1,..., t) are the eigenvalues of the correlation matrix
3) Adjustment of the test criteria as though there were t.eff independent tests with the Sidak correction(Šidák, 1967):
    a) Modified Sidak p = 1-(1-p)^(1/t.eff)
4) Evaluating the association results between genotypes and phenotypes variant by variant. If the *p*-value of any test is lower than *modified Sidak p*-value, the test is accepted as statistically significant.

Further consideration should also be given to testing for multiple (two) phenotypes in our regions of interest. A Bonferroni correction (divide the p-value by 2) is the most cautious approach. One should also be mindful of the multiple testing of different regions.

5.2.3.2. Significance levels for specific genetic variants

Established, specific genetic variants already associated with phenotypes in SCD can be considered in our dataset by applying a p-value 0.05 as I am not performing repeat testing. In this case, a specific hypothesis exists, so correction for multiple testing is not warranted as the result is simply a replication of the original finding.

5.2.4. Statistical analysis

The analysis requires genome-wide linear mixed modelling results from chapter 4. For each analysis, I evaluated the linear mixed modelling results for the HbSS subgroup as the discovery cohort, and then attempted to replicate any positive findings in the HbSC subgroup as a validation cohort.

Requires LMM results which have already been computed genome-wide.

I wrote a bash script to encapsulate the analysis, image generation and data formation. The bash script has a user-friendly command line interface to run where a user can stipulate a variety of parameters on request, see Figure 2. This allows users to choose new phenotypes and new regions of interest without having to alter the code.

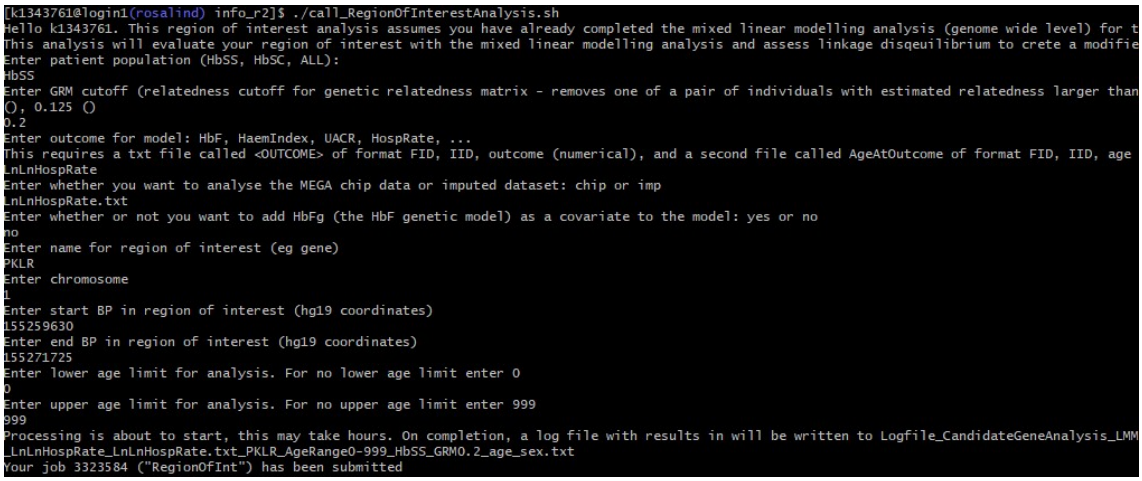


Figure 2 Region of interest script interface for user to answer questions.

This means the parameters are decided at the command line without the user having to alter the code. Parameters that the user can modify (after answering the questions) include: patient population (e.g. HbSS, HbSC, ALL, nonHbSS); relatedness cut-off; outcome (phenotype) name; whether imputed or chip (raw) genotype data to be used; region of interest name; region chromosome; region coordinates (hg19 build)

Any region of the genome can be inputted. Since I am interested in candidate genes, I used canonical transcripts of these genes plus 500bp upstream (as described in section 5.2.20.

In summary, it extracts the region-specific plink data (using PLINK) and mlma results (using AWK) and then computes the correlation (*r*) between all variants using PLINK. The correlation matrix of linkage disequilibrium statistics is read into R where the matrix is reduced down to calculate the "effective number statistical tests, t.eff". A modified Sidak p-value is then calculated using t.eff (as described in 5.2.3.1). The scripts involved are in Appendix 1 (the main script) and Appendix 2 (the calculation of the modified p value in R).

5.2.5. Post-analysis steps

Several post-analysis steps can help evaluate the plausibility of any significant results. Statistical checks can be made by confirming the cluster plots in any raw genotyping data available, plus imputation quality. If previous associations have been reported in the literature, consistency is reassuring. In the absence of previous publications, the direction of effect can be evaluated for consistency with the expected biological effect of the polymorphism. Other data can support the associations e.g. variation in gene expression or protein levels, consistent with the biological mechanism suggested by the association. None of these are substitutes for

replication of positive findings in an independent cohort (+/- with different genotyping methodologies).

5.3. Results

5.3.1. Regions of interest

5.3.1.1. Pyruvate kinase

47 genetic variants were evaluated in the *PK-LR* region. For hospitalisation rate (LnLnHospRate, N=242), the modified Sidak p-value was 0.001268. Seven variants in *PK-LR* were associated with LnLnHospRate in HbSS disease, of which 4 variants were replicated in the HbSC cohort: rs8177970, rs116244351, rs114455416 and rs8177964, see Table 6. No other clinical outcomes were associated with genetic variants in *PK-LR*.

I then re-assessed the genotyping quality of these variants to ensure there were no concerns with the results. Since they are all imputed variants, I have reviewed the cluster plots of surrounding raw genotyped variants; they all cluster well. I also revisited the imputation quality of those variants which have both raw genotypes and imputed genotypes available: all have imputation concordance info score (r2)>99.9%.

For rs8177970, 167 were wildtype homozygotes TT, 72 heterozygotes CT, and 3 homozygotes CC. For rs116244351, 168 were wildtype homozygotes GG, 71 heterozygotes AG, and 3 homozygotes AA. For rs114455416, 168 were wildtype homozygotes GG, 71 heterozygotes AG, and 3 homozygotes AA. For rs8177964, 169 were wildtype homozygotes GG, 70 heterozygotes AG, and 3 homozygotes AA. Hospitalisation rate (double logarithm) by genotype are displayed in Figure 3.

Thus four new risk variants for sickle severity (hospitalisation rate) has been identified and replicated in our study: rs8177970, rs116244351, rs114455416 and rs8177964. We are beginning further work to assess functional implications of these variants.

Table 6 Association of PK-LR variants with hospitalisation rate (LnLnHospRate) in HbSS and HbSC

Variant	Coordinates chr:position (hg19)	A1	A2	HbSS (discovery) N=242				HbSC (validation) N=94				Meta-analysis: all sickle genotypes (N=354)			
				Freq	Beta	S.E	P value	Freq	Beta	S.E	P value	Freq	Beta	S.E	P value
rs2071053	1:155265177	A	G	0.37	-0.0883	0.0268	0.00097	0.39	-0.0124	0.0299	0.67877	0.37	-0.0631	0.0199	0.00156
rs8177970	**1:155265661**	**C**	**T**	**0.16**	**0.1299**	**0.0364**	**0.00036**	**0.13**	**0.1074**	**0.0448**	**0.01657**	**0.15**	**0.1181**	**0.0280**	**0.00003**
rs116244351	**1:155266935**	**A**	**G**	**0.16**	**0.1247**	**0.0365**	**0.00064**	**0.13**	**0.1074**	**0.0448**	**0.01657**	**0.15**	**0.1142**	**0.0281**	**0.00005**
rs114455416	**1:155267389**	**A**	**G**	**0.16**	**0.1247**	**0.0365**	**0.00064**	**0.13**	**0.1212**	**0.0475**	**0.01075**	**0.15**	**0.1177**	**0.0286**	**0.00004**
rs12741350	1:155268425	C	T	0.38	-0.0864	0.0266	0.00115	0.39	-0.0033	0.0295	0.91181	0.38	-0.0592	0.0198	0.00276
rs3020781	1:155269776	A	G	0.38	-0.0864	0.0266	0.00115	0.39	-0.0033	0.0295	0.91181	0.38	-0.0592	0.0198	0.00276
rs8177964	**1:155269780**	**A**	**G**	**0.16**	**0.1241**	**0.0367**	**0.00071**	**0.13**	**0.0890**	**0.0453**	**0.04948**	**0.15**	**0.1096**	**0.0282**	**0.00010**

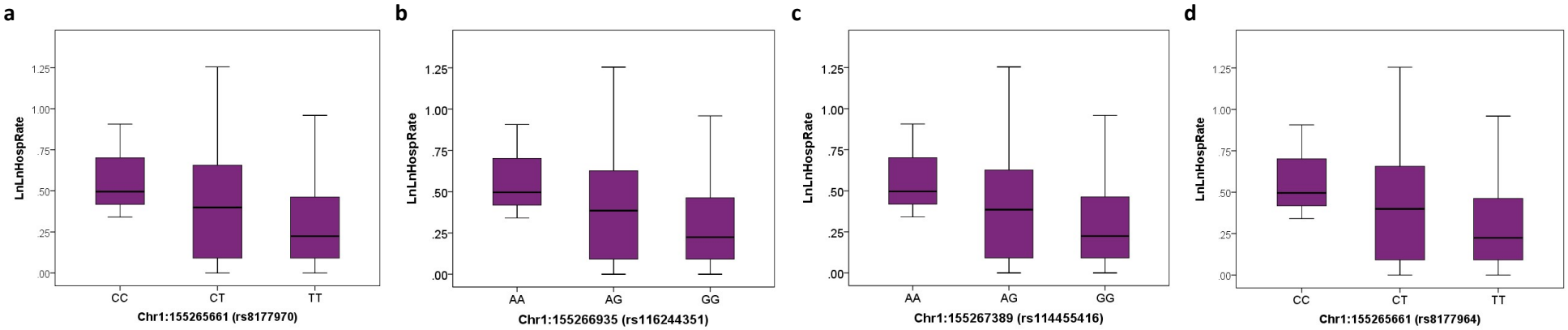


Figure 3 Hospitalisation rate (double logarithm) by genotype for the significant variants in the linear mixed modelling: (a) rs8177970 (b) rs116244351 (c) rs114455416 (d) rs8177964

5.3.1.2. 2,3-diphosphoglycerate

188 genetic variants were assessed in the region *BPGM*. For both severity outcomes (LnLnHospRate, HaemIndex_avg), there were no significant region-wide variants.

5.3.1.3. Adenosine A_{2B} receptor

252 genetic variants were evaluated in the region *ADORA2B*. For both severity outcomes (LnLnHospRate, HaemIndex_avg), there were no significant region-wide variants.

5.3.1.4. Complement related genes

All genetic variants within a complement-associated gene were analysed for both severity outcomes (LnLnHospRate, HaemIndex_avg). There were no significant, replicated region-wide variants for any region.

- 156 variants in the region *MCP*.
- 563 in the region *CFH*.
- 87 genetic variants in the region *CFB*.
- 16 variants in the region *THBD*.
- 167 variants in the region *DGKE*.
- No variants in C3.
- 251 variants in the region *ADAMTS13*.

5.3.2. Replication of specific variants

5.3.2.1. APOL1

94 genetic variants were evaluated in the region *APOL1*. For urinary albumin creatinine ratio, there were no significant region-wide variants (genome wide p value = 0.0005797).

I have replicated the association between two G1 variants and proteinuria in SCD, see Table 7.

Table 7 Association of APOL1 G1 variants with urinary albumin creatinine ratio (uACR) in HbSS and HbSC

Variant	Ch	Position (hg19)	A1	A2	HbSS (N=229)				HbSC (N=80)			
					MAF	β-value	s.e.	p-value	MAF	β-value	s.e.	p-value
rs73885319	22	36661906	G	A	0.32	0.078	0.03	0.008	0.29	0.05	0.05	0.28
rs60910145	22	36662034	G	T	0.32	0.087	0.03	0.003	0.28	0.06	0.05	0.22

Unfortunately, G2 is not in the imputed dataset so this analysis is not a full assessment of current APOL1 risk variants. Furthermore, a full investigation should include covariates in analysis (e.g. blood pressure, diabetes status, HIV status) – all causes of renal dysfunction that have been associated with APOL1 renal disease.

5.3.2.2. DARC

9 genetic variants in the *DARC* region were evaluated. For both severity outcomes (LnLnHospRate, HaemIndex_avg), there were no significant region-wide variants.

I have *not* replicated the previous association between rs2814778 and either of our two SCD severity markers.

I did, however, replicate its association with average neutrophil count. This analysis demonstrates region-wide significance of rs2814778 with neutrophil count: MAF 0.053, Beta 0.81, p value 0.0027.

The reason for the disparity between my results and others may well be due to controlling for relatedness in my analysis (see section 5.1.4.2). The diversity of European admixture in the populations studied (from 0% in some African populations to up to 60% in Brazilian populations) presumably accounts for the discrepancies in the reported findings between different cohorts (Drasar et al., 2013, Afenyi-Annan et al., 2008, Mecabo et al., 2010, Schnog et al., 2000, Nebor et al., 2010, Araujo et al., 2015).

5.3.2.3. MAPK8

628 genetic variants in the *MAPK8* region were assessed. For both severity outcomes (LnLnHospRate (N=242), HaemIndex_avg (N=216)), there were no significant variants.
I have not replicated locus rs10857560 that has been associated with pulmonary hypertension in SCD (in any of the four outcomes). The diversity of allele frequencies again may mean the previous association identified was confounded by ethnicity. Unfortunately, we did not have enough echocardiogram data to perform an analysis using tricuspid regurgitant jet velocity as an outcome.

5.4. Discussion

Candidate gene studies are powerful, hypothesis-driven approaches to genotype-phenotype correlation that can contribute much to our understanding of the genetics of common diseases and traits, particularly when pre-existing data strengthen the hypothesis. Furthermore, the candidate gene approach remains the only feasible approach for studying small populations. Nevertheless, the method does not always work (both false negative and false positive association signals) because of inefficient study design and suboptimal gene or variant selection strategies.

Using a candidate gene approach, I have identified four risk variants in the red blood cell pyruvate kinase gene *PK-LR* for frequent hospitalisation in SCD. I used our HbSS patients as the discovery set, and validated four variants in *PK-LR* in our HbSC cohort. A meta-analysis of *all* SCD patients showed improved p-values. These four risk variants are all in intron 2 of *PK-LR*. Notably, this region overlaps with a regulatory element active in K562 cells on the UCSC genome browser (hg19) using the layered H3K27Ac track (Kent et al., 2002), see Figure 4.

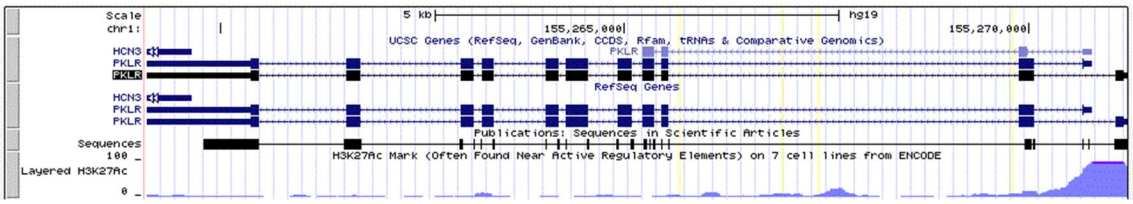


Figure 4 UCSC genome browser: Layered H3K27Ac track (K562 cells) for PK-LR

In the literature, only one of the four variants, rs8177970, has reported clinical associations. These are all negative findings in GWA studies assessing: Alzheimer's disease, bipolar disorder, schizophrenia, response to antipsychotic therapy, tardive dyskinesia and ulcerative colitis. If this association is corroborated, these intronic risk variants are not likely to represent the causal element in itself, but to be in linkage disequilibrium with the causal variant. It is interesting to note, however, that rs8177970 is strongly predicted to create a new 5' splice site only when the variant is present.

PK deficiency is associated with *Plasmodium falciparum* malaria prevalence; it provides protection against infection and replication of *P. falciparum* in human red cells(Ayi et al., 2008). Moreover, genetic variants in *PK-LR* have been shown to modify malaria phenotypes(van Bruggen et al., 2015). This group demonstrated that human erythrocytes infected *ex vivo* with *P. falciparum* with host PK-deficiency alleles reduces infection phenotypes. Furthermore, on a population level, heterozygosity for one coding variant, R41Q, was shown to be associated with reduced frequency of *P. falciparum* infections(van Bruggen et al., 2015).

Malarial protection seen in PK deficiency helps to explain the population genetics of the four *PK-LR* variants we identified. All four have a minor allele frequency of 14-15% in African populations in 1000 Genomes Project, while they have maximum 1% in non-African populations. Because I constructed our analysis model to manage relatedness (including admixture), I do not believe these differences reflect population stratification issues in the results. Furthermore, the HbSC group in which I replicated the data is, broadly, an ethnically distinct region, reflecting Ghanaian / Ivory Coast heritage.

Our study has some limitations. Our sample numbers are small. Frequency of acute pain episodes is used by clinicians informally as a marker of disease severity, and it is broadly a trait which appears familial, but it is psychosocially as well as biologically determined. Using "hospitalisation rate" as a measure of acute painful crisis frequency does not account for inter-individual differences in decision to attend hospital which is influenced by a range of cultural and social issues.

We plan to continue to continue to investigate these *PK-LR* risk variants. We plan replication studies, as well as sequencing studies to identify functional variants. Measuring protein levels of pyruvate kinase or 2,3-DPG in the red blood cells of patients with SCD may prove difficult as formal reference ranges are undefined. In SCD, red cells with low PK are not removed by the spleen. This also means that, for SCD, a genetic diagnosis of PK deficiency (however mild) is likely to be better than a proteomic approach.

We identified no risk variants in complement-related genes. Other strategies to evaluate this region could be to consider the case series we reported of episodes of "superhaemolysis" and thrombocytopenia in 18 patients – which we postulated could represent an acute MAHA(Gardner and Thein, 2015). These 18 patients were identified precisely from the KCH adult cohort, so a case-control study using these patients (of whom 14 have genetic data available) may provide further insight into this association.

I also investigated specific genetic variants and their association with sickle complications. I have replicated the finding of both G1 variants in *APOL1* being associated with degree of proteinuria. Unfortunately, I did not have genetic data on *APOL1* G2 available to assess this.

I found no association between the *DARC* promoter variant rs2814778 and either severity index (hospitalisation rate or haemolysis rate). I suggest that this is because, in contrast with all previous studies, I accounted for relatedness (including population structure) in our statistical modelling, and rs2814778 seems to represent an ethnicity marker only seen in African-heritage populations. I therefore suggest that previous studies are confounded by admixture and the false association has occurred because of confounding due to ethnicity. (Notably, I did confirm the expected association between the variant and neutrophil level).

I found no association between the *MAPK8* variant rs10857560 and haemolysis in our cohort. The original paper associating rs10857560 and pulmonary hypertension did account for population structure. Unfortunately, we do not have a large enough dataset of echocardiographic results to assess the association between the variant and tricuspid regurgitant velocities in our own cohort.

In summary, the targeted approach of candidate gene analysis identified a promising new risk region in the *PK-LR* gene associated with severity in sickle cell disease, namely hospitalisation rate. The approach is robust against the old flaws of candidate gene studies – notably

population stratification and admixture – because I accounted for relatedness in the statistical modelling. The same approach has also allowed us to argue that previous findings of association between *DARC* and sickle severity are likely spurious on the basis of more sophisticated statistical modelling.

In the future, we aim to expand candidate gene analysis for our cohort. The existing pipelines make it very easy for new phenotypes, and new regions of interest, to be evaluated by users not experienced in either bioinformatics tools or statistical genetics.

Future studies can be improved further by better choice of regions of interest, we can: (a) choose pathways not genes (b) use algorithms to prioritise variants to choose. There are multiple bioinformatics tools to do this. We could enrich our dataset by just choosing variants more likely to have functional sequelae - missense/nonsense/splice site variants rather than intronic or synonymous variants. Variants could also be prioritised based on allele frequency, so that using the "common trait, common variant" hypothesis, we only consider common alleles – such an approach would confer greater statistical power to detect associations. Conversely, for rare traits, we may be more interested in rare variants (with large effect sizes).

References

ABUSIN, G. A., ABU-ARJA, R., BAJWA, R. P. S., HORWITZ, E. M., AULETTA, J. J. & RANGARAJAN, H. G. 2017. Severe transplant-associated thrombotic microangiopathy in patients with hemoglobinopathies. *Pediatr Blood Cancer,* 64.

AFENYI-ANNAN, A., KAIL, M., COMBS, M. R., ORRINGER, E. P., ASHLEY-KOCH, A. & TELEN, M. J. 2008. Lack of Duffy antigen expression is associated with organ damage in patients with sickle cell disease. *Transfusion,* 48**,** 917-24.

ALLI, N., COETZEE, M., LOUW, V., VAN RENSBURG, B., ROSSOUW, G., THOMPSON, L., PISSARD, S. & THEIN, S. L. 2008. Sickle cell disease in a carrier with pyruvate kinase deficiency. *Hematology,* 13**,** 369-72.

ARAUJO, N. B., DOMINGOS, I. F., MEDEIROS, F. S., HATZLHOFER, B. L., MENDONCA, T. F., VASCONCELOS, L. R., CAVALCANTI MDO, S., ARAUJO, A. S., OLIVEIRA MDO, C., LUCENA-ARAUJO, A. R. & BEZERRA, M. A. 2015. Lack of association between the Duffy antigen receptor for chemokines (DARC) expression and clinical outcome of children with sickle cell anemia. *Immunol Lett,* 166**,** 140-2.

ASHLEY-KOCH, A. E., OKOCHA, E. C., GARRETT, M. E., SOLDANO, K., DE CASTRO, L. M., JONASSAINT, J. C., ORRINGER, E. P., ECKMAN, J. R. & TELEN, M. J. 2011. MYH9 and APOL1 are both associated with sickle cell disease nephropathy. *Br J Haematol,* 155**,** 386-94.

AYI, K., MIN-OO, G., SERGHIDES, L., CROCKETT, M., KIRBY-ALLEN, M., QUIRT, I., GROS, P. & KAIN, K. C. 2008. Pyruvate kinase deficiency and malaria. *N Engl J Med,* 358**,** 1805-10.

BOLANOS-MEADE, J., KEUNG, Y. K., LOPEZ-ARVIZU, C., FLORENDO, R. & COBOS, E. 1999. Thrombotic thrombocytopenic purpura in a patient with sickle cell crisis. *Ann Hematol,* 78**,** 558-9.

BOONYASAMPANT, M., WEITZ, I. C., KAY, B., BOONCHALERMVICHIAN, C., LIEBMAN, H. A. & SHULMAN, I. A. 2015. Life-threatening delayed hyperhemolytic transfusion reaction in a patient with sickle cell disease: effective treatment with eculizumab followed by rituximab. *Transfusion,* 55**,** 2398-403.
CHARACHE, S., GRISOLIA, S., FIEDLER, A. J. & HELLEGERS, A. E. 1970. Effect of 2,3-diphosphoglycerate on oxygen affinity of blood in sickle cell anemia. *Journal of Clinical Investigation,* 49**,** 806-812.
CHEVERUD, J. M., VAUGHN, T. T., PLETSCHER, L. S., PERIPATO, A. C., ADAMS, E. S., ERIKSON, C. F. & KING-ELLISON, K. J. 2001. Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice. *Mamm Genome,* 12**,** 3-12.
CHINOWSKY, M. S. 1994. Thrombotic thrombocytopenic purpura associated with sickle cell-hemoglobin C disease. *South Med J,* 87**,** 1168-71.
CHONAT, S., CHANDRAKASAN, S., KALINYAK, K. A., INGALA, D., GRUPPO, R. & KALFA, T. A. 2016. Atypical haemolytic uraemic syndrome in a patient with sickle cell disease, successfully treated with eculizumab. *Br J Haematol,* 175**,** 744-747.
COHEN-SOLAL, M., PREHU, C., WAJCMAN, H., POYART, C., BARDAKDJIAN-MICHAU, J., KISTER, J., PROME, D., VALENTIN, C., BACHIR, D. & GALACTEROS, F. 1998. A new sickle cell disease phenotype associating Hb S trait, severe pyruvate kinase deficiency (PK Conakry), and an alpha2 globin gene variant (Hb Conakry). *Br J Haematol,* 103**,** 950-6.
DESAI, A. A., ZHOU, T., AHMAD, H., ZHANG, W., MU, W., TREVINO, S., WADE, M. S., RAGHAVACHARI, N., KATO, G. J., PETERS-LAWRENCE, M. H., THIRUVOIPATI, T., TURNER, K., ARTZ, N., HUANG, Y., PATEL, A. R., YUAN, J. X., GORDEUK, V. R., LANG, R. M., GARCIA, J. G. & MACHADO, R. F. 2012. A novel molecular signature for elevated tricuspid regurgitation velocity in sickle cell disease. *Am J Respir Crit Care Med,* 186**,** 359-68.
DRASAR, E. R., MENZEL, S., FULFORD, T. & THEIN, S. L. 2013. The effect of Duffy antigen receptor for chemokines on severity in sickle cell disease. *Haematologica,* 98**,** e87-e89.
DUMAS, G., HABIBI, A., ONIMUS, T., MERLE, J. C., RAZAZI, K., MEKONTSO DESSAP, A., GALACTEROS, F., MICHEL, M., FREMEAUX BACCHI, V., NOIZAT PIRENNE, F. & BARTOLUCCI, P. 2016. Eculizumab salvage therapy for delayed hemolysis transfusion reaction in sickle cell disease patients. *Blood,* 127**,** 1062-4.
DURPES, M. C., HARDY-DESSOURCES, M. D., EL NEMER, W., PICOT, J., LEMONNE, N., ELION, J. & DECASTEL, M. 2011. Activation state of alpha4beta1 integrin on sickle red blood cells is linked to the duffy antigen receptor for chemokines (DARC) expression. *J Biol Chem,* 286**,** 3057-64.
DURPES, M. C., NEBOR, D., DU MESNIL, P. C., MOUGENEL, D., DECASTEL, M., ELION, J. & HARDY-DESSOURCES, M. D. 2010. Effect of interleukin-8 and RANTES on the Gardos channel activity in sickle human red blood cells: role of the Duffy antigen receptor for chemokines. *Blood Cells Mol Dis,* 44**,** 219-23.
FREEDMAN, B. I. & SKORECKI, K. 2014. Gene-gene and gene-environment interactions in apolipoprotein L1 gene-associated nephropathy. *Clin J Am Soc Nephrol,* 9**,** 2006-13.
GARDNER, K. & THEIN, S. L. 2015. Super-elevated LDH and thrombocytopenia are markers of a severe subtype of vaso-occlusive crisis in sickle cell disease. *Am J Hematol,* 90**,** E206-7.
GEIGEL, E. J. & FRANCIS, C. W. 1997. Reversal of multiorgan system dysfunction in sickle cell disease with plasma exchange. *Acta Anaesthesiol Scand,* 41**,** 647-50.
GENOVESE, G., FRIEDMAN, D. J., ROSS, M. D., LECORDIER, L., UZUREAU, P., FREEDMAN, B. I., BOWDEN, D. W., LANGEFELD, C. D., OLEKSYK, T. K., USCINSKI KNOB, A. L., BERNHARDY, A. J., HICKS, P. J., NELSON, G. W., VANHOLLEBEKE, B., WINKLER, C. A., KOPP, J. B., PAYS, E. & POLLAK, M. R. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science,* 329**,** 841-5.
KARMOUTY-QUINTANA, H., WENG, T., GARCIA-MORALES, L. J., CHEN, N. Y., PEDROZA, M., ZHONG, H., MOLINA, J. G., BUNGE, R., BRUCKNER, B. A., XIA, Y., JOHNSTON, R. A., LOEBE, M., ZENG, D., SEETHAMRAJU, H., BELARDINELLI, L. & BLACKBURN, M. R. 2013. Adenosine A2B

receptor and hyaluronan modulate pulmonary hypertension associated with chronic obstructive pulmonary disease. *Am J Respir Cell Mol Biol,* 49**,** 1038-47.
KARMOUTY-QUINTANA, H., ZHONG, H., ACERO, L., WENG, T., MELICOFF, E., WEST, J. D., HEMNES, A., GRENZ, A., ELTZSCHIG, H. K., BLACKWELL, T. S., XIA, Y., JOHNSTON, R. A., ZENG, D., BELARDINELLI, L. & BLACKBURN, M. R. 2012. The A2B adenosine receptor modulates pulmonary hypertension associated with interstitial lung disease. *Faseb j,* 26**,** 2546-57.
KATO, G. J., STEINBERG, M. H. & GLADWIN, M. T. 2017. Intravascular hemolysis and the pathophysiology of sickle cell disease. *J Clin Invest,* 127**,** 750-760.
KAVANAGH, D. & GOODSHIP, T. 2010. Genetics and complement in atypical HUS. *Pediatric Nephrology,* 25**,** 2431-2442.
KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.
KOPP, J. B., NELSON, G. W., SAMPATH, K., JOHNSON, R. C., GENOVESE, G., AN, P., FRIEDMAN, D., BRIGGS, W., DART, R., KORBET, S., MOKRZYCKI, M. H., KIMMEL, P. L., LIMOU, S., AHUJA, T. S., BERNS, J. S., FRYC, J., SIMON, E. E., SMITH, M. C., TRACHTMAN, H., MICHEL, D. M., SCHELLING, J. R., VLAHOV, D., POLLAK, M. & WINKLER, C. A. 2011. APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. *J Am Soc Nephrol,* 22**,** 2129-37.
KORMANN, R., JANNOT, A. S., NARJOZ, C., RIBEIL, J. A., MANCEAU, S., DELVILLE, M., JOSTE, V., PRIE, D., POUCHOT, J., THERVET, E., COURBEBAISSE, M. & ARLET, J. B. 2017. Roles of APOL1 G1 and G2 variants in sickle cell disease patients: kidney is the main target. *Br J Haematol*.
LARSEN, C. P., BEGGS, M. L., SAEED, M. & WALKER, P. D. 2013. Apolipoprotein L1 risk variants associate with systemic lupus erythematosus-associated collapsing glomerulopathy. *J Am Soc Nephrol,* 24**,** 722-5.
LEE, H. E., MARDER, V. J., LOGAN, L. J., FRIEDMAN, S. & MILLER, B. J. 2003. Life-threatening thrombotic thrombocytopenic purpura (TTP) in a patient with sickle cell-hemoglobin C disease. *Ann Hematol,* 82**,** 702-4.
LOTTA, L. A., GARAGIOLA, I., PALLA, R., CAIRO, A. & PEYVANDI, F. 2010. ADAMTS13 mutations and polymorphisms in congenital thrombotic thrombocytopenic purpura. *Hum Mutat,* 31**,** 11-9.
MECABO, G., HAYASHIDA, D. Y., AZEVEDO-SHIMMOTO, M. M., VICARI, P., ARRUDA, M. M., BORDIN, J. O. & FIGUEIREDO, M. S. 2010. Duffy-negative is associated with hemolytic phenotype of sickle cell anemia. *Clin Immunol,* 136**,** 458-9; author reply 460-1.
MI, T., ABBASI, S., ZHANG, H., URAY, K., CHUNN, J. L., XIA, L. W., MOLINA, J. G., WEISBRODT, N. W., KELLEMS, R. E., BLACKBURN, M. R. & XIA, Y. 2008. Excess adenosine in murine penile erectile tissues contributes to priapism via A2B adenosine receptor signaling. *J Clin Invest,* 118**,** 1491-501.
MILLER, S. T., SLEEPER, L. A., PEGELOW, C. H., ENOS, L. E., WANG, W. C., WEINER, S. J., WETHERS, D. L., SMITH, J. & KINNEY, T. R. 2000. Prediction of adverse outcomes in children with sickle cell disease. *N Engl J Med,* 342**,** 83-9.
NEBOR, D., DURPES, M. C., MOUGENEL, D., MUKISI-MUKAZA, M., ELION, J., HARDY-DESSOURCES, M. D. & ROMANA, M. 2010. Association between Duffy antigen receptor for chemokines expression and levels of inflammation markers in sickle cell anemia patients. *Clin Immunol,* 136**,** 116-22.
PHILLIPS, E. H., WESTWOOD, J. P., BROCKLEBANK, V., WONG, E. K., TELLEZ, J. O., MARCHBANK, K. J., MCGUCKIN, S., GALE, D. P., CONNOLLY, J., GOODSHIP, T. H., KAVANAGH, D. & SCULLY, M. A. 2016. The role of ADAMTS-13 activity and complement mutational analysis in differentiating acute thrombotic microangiopathies. *J Thromb Haemost,* 14**,** 175-85.
PIRENNE, F., BARTOLUCCI, P. & HABIBI, A. 2017. Management of delayed hemolytic transfusion reaction in sickle cell disease: Prevention, diagnosis, treatment. *Transfus Clin Biol*.
PLATT, O. S., BRAMBILLA, D. J., ROSSE, W. F., MILNER, P. F., CASTRO, O., STEINBERG, M. H. & KLUG, P. P. 1994. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med,* 330**,** 1639-44.

POILLON, W. N. & KIM, B. C. 1990. 2,3-Diphosphoglycerate and intracellular pH as interdependent determinants of the physiologic solubility of deoxyhemoglobin S. *Blood,* 76**,** 1028-36.
POILLON, W. N., KIM, B. C. & CASTRO, O. 1998. Intracellular hemoglobin S polymerization and the clinical severity of sickle cell anemia. *Blood,* 91**,** 1777-83.
PRICHARD, J. G., CLARK, H. G. & JAMES, R. E., 3RD 1988. Abdominal pain and sicklemia in a patient with sickle cell trait. *South Med J,* 81**,** 1312-4.
RAPOPORT, S. & LUEBERING, J. 1952. An optical study of diphosphoglycerate mutase. *J Biol Chem,* 196**,** 593-8.
REICH, D., NALLS, M. A., KAO, W. H., AKYLBEKOVA, E. L., TANDON, A., PATTERSON, N., MULLIKIN, J., HSUEH, W. C., CHENG, C. Y., CORESH, J., BOERWINKLE, E., LI, M., WALISZEWSKA, A., NEUBAUER, J., LI, R., LEAK, T. S., EKUNWE, L., FILES, J. C., HARDY, C. L., ZMUDA, J. M., TAYLOR, H. A., ZIV, E., HARRIS, T. B. & WILSON, J. G. 2009. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet,* 5**,** e1000360.
SARAF, S. L., SHAH, B. N., ZHANG, X., HAN, J., TAYO, B. O., ABBASI, T., OSTROWER, A., GUZMAN, E., MOLOKIE, R. E., GOWHARI, M., HASSAN, J., JAIN, S., COOPER, R. S., MACHADO, R. F., LASH, J. P. & GORDEUK, V. R. 2017. APOL1, alpha-thalassemia, and BCL11A variants as a genetic risk profile for progression of chronic kidney disease in sickle cell anemia. *Haematologica,* 102**,** e1-e6.
SARAF, S. L., ZHANG, X., SHAH, B., KANIAS, T., GUDEHITHLU, K. P., KITTLES, R., MACHADO, R. F., ARRUDA, J. A., GLADWIN, M. T., SINGH, A. K. & GORDEUK, V. R. 2015. Genetic variants and cell-free hemoglobin processing in sickle cell nephropathy. *Haematologica,* 100**,** 1275-84.
SCHAEFER, B. A., FLANAGAN, J. M., ALVAREZ, O. A., NELSON, S. C., AYGUN, B., NOTTAGE, K. A., GEORGE, A., ROBERTS, C. W., PICCONE, C. M., HOWARD, T. A., DAVIS, B. R. & WARE, R. E. 2016. Genetic Modifiers of White Blood Cell Count, Albuminuria and Glomerular Filtration Rate in Children with Sickle Cell Anemia. *PLoS One,* 11**,** e0164364.
SCHNOG, J. B., KELI, S. O., PIETERS, R. A., ROJER, R. A. & DUITS, A. J. 2000. Duffy phenotype does not influence the clinical severity of sickle cell disease. *Clin Immunol,* 96**,** 264-8.
SHELAT, S. G. 2010. Thrombotic thrombocytopenic purpura and sickle cell crisis. *Clin Appl Thromb Hemost,* 16**,** 224-7.
SHOME, D. K., RAMADORAI, P., AL-AJMI, A., ALI, F. & MALIK, N. 2013. Thrombotic microangiopathy in sickle cell disease crisis. *Ann Hematol,* 92**,** 509-15.
ŠIDÁK, Z. 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association,* 62**,** 626-633.
VAN BRUGGEN, R., GUALTIERI, C., ILIESCU, A., LOUICHAROEN CHEEPSUNTHORN, C., MUNGKALASUT, P., TRAPE, J.-F., MODIANO, D., SIRIMA, B. S., SINGHASIVANON, P., LATHROP, M., SAKUNTABHAI, A., BUREAU, J.-F. & GROS, P. 2015. Modulation of Malaria Phenotypes by Pyruvate Kinase (PKLR) Variants in a Thai Population. *PloS one* [Online], 10. [Accessed 2015].
VENKATA SASIDHAR, M., TRIPATHY, A. K., VISWANATHAN, K. & SHUKLA, M. 2010. Thrombotic thrombocytopenic purpura and multiorgan system failure in a child with sickle cell-hemoglobin C disease. *Clin Pediatr (Phila),* 49**,** 992-6.
ZHANG, X., ZHANG, W., MA, S. F., DESAI, A. A., SARAF, S., MIASNIAKOVA, G., SERGUEEVA, A., AMMOSOVA, T., XU, M., NEKHAI, S., ABBASI, T., CASANOVA, N. G., STEINBERG, M. H., BALDWIN, C. T., SEBASTIANI, P., PRCHAL, J. T., KITTLES, R., GARCIA, J. G., MACHADO, R. F. & GORDEUK, V. R. 2014. Hypoxic response contributes to altered gene expression and precapillary pulmonary hypertension in patients with sickle cell disease. *Circulation,* 129**,** 1650-8.
ZHANG, Y., DAI, Y., WEN, J., ZHANG, W., GRENZ, A., SUN, H., TAO, L., LU, G., ALEXANDER, D. C., MILBURN, M. V., CARTER-DAWSON, L., LEWIS, D. E., ZHANG, W., ELTZSCHIG, H. K., KELLEMS, R. E., BLACKBURN, M. R., JUNEJA, H. S. & XIA, Y. 2011. Detrimental effects of adenosine signaling in sickle cell disease. *Nat Med,* 17**,** 79-86.

Appendix 1

RegionOfInterestAnalysis.sh

```
#!/bin/bash
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -q HighMemLongterm.q,LowMemLongterm.q
#$ -M kate.gardner@doctors.org.uk
#$ -m beas
#$ -l h_vmem=20G
##########

POPULATION=$1 # requires a text file with two columns FID/IID so we can extract that population eg HbSS.txt,
HbSC.txt
GRM_CUTOFF=$2 #relatedness cutoff for genetic relatedness matric eg 0.9 to exclude genetic duplicates
OUTCOME=$3 # "HbF" or "HaemIndex" or
IMP_OR_CHIP=$4 #"chip" or "imp"
HbFg=$5
REGION_OF_INTEREST=$6 # name of region of interest
REGION_CHR=$7 # region of interest chromosome
REGION_FROMBP=$8 # start BP in region of interest (hg19 coordinates)
REGION_TOBP=$9 # end BP in region of interest (hg19 coordinates)
LOWER_AGE=${10}
UPPER_AGE=${11}


module add bioinformatics/plink2/1.90b3.38
module add bioinformatics/plink/1.90b3.31
module add bioinformatics/R/3.3.0

if [[ "$IMP_OR_CHIP" = "chip" ]]; then
        #chip plink data
```

```
      IMP_CHIP_FILE="SickleMEGA_QC_NoSexMismatch.autosomes"
      #SickleMEGA_QC_NoSexMismatch.autosomes_${POPULATION}_PtsRemoved_postGRMcutoff${GRM_CUTOFF}
else
        #imputed plink data
      IMP_CHIP_FILE="Sickle_Imputed_QC_strict"
      ##Sickle_Imputed_QC_strict_${POPULATION}_PtsRemoved_postGRMcutoff${GRM_CUTOFF}
fi

if [[ "$POPULATION" = "ALL" || "$POPULATION" = "nonHbSS" || "$POPULATION" = "HbSSHbSC" || "$POPULATION" =
"HbSSHbSCHbSBplus"  ]]; then
      COVARIATES_FILE="age_sex_sickle"
      COVARIATES="Age at sample, sex, sickle genotype,"
else
        COVARIATES_FILE="age_sex"
        COVARIATES="Age at sample, sex,"
fi

if [[ "$HbFg" = "yes" ]]; then
      HbFg_FILE="_withHbFgenetic"
        HbFgCOVARIATE="HbFg (HbF genetic model)"
       else
        HbFg_FILE=""
        HbFgCOVARIATE="(no HbFg)"
fi




########################################
#get MLMA results for region of interest
head -n1 ${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}_${COVARIATES_FILE}${HbFg_FILE}.loco.mlma.tick >
```

```
${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}.txt
awk -v REGION_CHR="$REGION_CHR" -v REGION_FROMBP="$REGION_FROMBP" -v REGION_TOBP="$REGION_TOBP" -v
IMP_CHIP_FILE="$IMP_CHIP_FILE" -v LOWER_AGE="$LOWER_AGE" -v UPPER_AGE="$UPPER_AGE" -v POPULATION="$POPULATION"
-v GRM_CUTOFF="$GRM_CUTOFF" -v OUTCOME="$OUTCOME" -v REGION_OF_INTEREST="$REGION_OF_INTEREST"
'{if($1==REGION_CHR && $3>=REGION_FROMBP && $3<=REGION_TOBP) print $0}'
"${IMP_CHIP_FILE}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_GCTA_${OUTCOME}_${COVARIATES_FILE}${HbFg_FILE}.loco.mlma.tick" >>
"${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}.txt"



#get corrected p values using (1) LD in region of interest to create an effective number of tests plus (2)
Sidak's correction
#first get LD matrix for region of interest
#maf 0.001 added otherwise the LD matrix subsequently will contain nan values
plink --bfile ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF} --maf 0.001 --chr ${REGION_CHR} --from-bp
${REGION_FROMBP} --to-bp ${REGION_TOBP} --make-bed --out ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_${REGION_OF_INTEREST}

plink --bfile ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_${REGION_OF_INTEREST} --r  --matrix --out
${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_LD_r

#second use R to reduce LD matrix to a number and also apply Sidak's correction to get a new p value
Rscript modifiedPvalueCalculation.R ${POPULATION} ${GRM_CUTOFF} ${REGION_OF_INTEREST} ${IMP_OR_CHIP}
${LOWER_AGE} ${UPPER_AGE}

d=$(awk 'NR==2 {print $1}' ${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_modifiedPvalues.txt)
```

```
bonferroni_p=$(awk 'NR==2 {print $2}' ${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_modifiedPvalues.txt)
sidak_p=$(awk 'NR==2 {print $3}' ${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_modifiedPvalues.txt)
d_eff=$(awk 'NR==2 {print $4}' ${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_modifiedPvalues.txt)
corr_p=$(awk 'NR==2 {print $5}' ${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_modifiedPvalues.txt)
#ImputationSNPs=$(wc Chr${CHROMOSOMES}_PreImputation_NoMergeFails.bim | awk '{print $1}')

#What SNPs satisfy corrected p value threshold?
head -n1 ${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}.txt >
${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}_signifVariants.txt
awk -v CORR_P="$corr_p" '{if($9<CORR_P) print $0}'
"${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}.txt" >>
"${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}_signifVariants.txt"

#Create log file
        echo "Candidate Gene Analysis using linear mixed modelling for outcome ${OUTCOME}
     Summary of results and files generated:

     Data source: ${IMP_OR_CHIP} data
     Region of interest: ${REGION_OF_INTEREST}, Chromosome ${REGION_CHR}, from bp ${REGION_FROMBP} to bp
${REGION_TOBP} (hg19 coordinates)
     Population: ${POPULATION}
     Age range: ${LOWER_AGE} to ${UPPER_AGE} years
     GRM cutoff: ${GRM_CUTOFF}
     Covariates: ${COVARIATES} ${HbFgCOVARIATE}
```

```
     Number of genetic variants in region (d) = $d (number of tests performed - as we performed one test for
each genetic variant)
     Bonferroni p = p/number tests = ${bonferroni_p}
     Sidak p = 1-(1-p)^(1/d) = ${sidak_p} (Sidak's corrected p value (at least as big as Bonferroni's p))
     'd.eff' = effective d = ${d_eff} (based on LD - if tests are correlated with each other then d.eff<d)
     Fully corrected p = ${corr_p} (based on d.eff and Sidak's modified p, at least as big as Sidak's p
without d.eff)

     Files are saved:
     LMM results for region of interest:
${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}.txt
     LMM results - significant results only (satisfy the fully corrected p as above, based on d.eff and
Sidak's modification): ${OUTCOME}_${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}.MLMA_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}_signifVariants.txt
     PLINK files for region of interest: ${IMP_CHIP_FILE}_PtsRemoved_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_${REGION_OF_INTEREST}
     LD matrix for region of interest: ${REGION_OF_INTEREST}_${IMP_OR_CHIP}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_postGRMcutoff${GRM_CUTOFF}_LD_r.ld
       " >
Logfile_CandidateGeneAnalysis_LMM_${OUTCOME}_${IMP_OR_CHIP}_${REGION_OF_INTEREST}_AgeRange${LOWER_AGE}-
${UPPER_AGE}_${POPULATION}_GRM${GRM_CUTOFF}_${COVARIATES_FILE}${HbFg_FILE}.txt
```

Appendix 2
modifiedPvalueCalculation.R

```
#!/usr/bin/env Rscript

args = commandArgs(trailingOnly=TRUE)

if (length(args)==0) {
      POPULATION<-"ALL"
      GRM_CUTOFF<-0.2
      REGION_OF_INTEREST<-"MAPK8"
         IMP_OR_CHIP<-"chip" #"Imputed"
      LOWER_AGE<-0
      UPPER_AGE<-999
      HbFg_FILE<-"no"
      } else {
      POPULATION<-args[1]
      GRM_CUTOFF<-args[2]
         REGION_OF_INTEREST<-args[3]
         IMP_OR_CHIP<-args[4]
         LOWER_AGE<-args[5]
         UPPER_AGE<-args[6]
      HbFg_FILE<-args[7]
}


p=0.05

inputFile<-paste(REGION_OF_INTEREST,"_",IMP_OR_CHIP,"_AgeRange",LOWER_AGE,"-
",UPPER_AGE,"_",POPULATION,"_postGRMcutoff",GRM_CUTOFF,"_LD_r.ld",sep="")
inputFile
LDtable<-read.table(inputFile,header=FALSE)
```

```
d=nrow(LDtable) #d is number of tests performed (so if you perform one test for each SNP then that's the
number of SNPs)
bonferroni.p=p/d
sidak.p=1-(1-p)^(1/d) # Sidak's corrected p value (smaller than Bonferroni's p)
d.eff=1+(d-1)*(1-var(eigen(LDtable)$values)/d) #effective d - based on LD - if tests are correlated with each
other then d.eff<d
corr.p=1-(1-p)^(1/d.eff)

col1<-c("Candidate Gene Analysis using linear mixed modelling with sex, age(, sickle genotype, HbFg) as
covariates","Region of interest","Data source:","Population:","Lower age limit:","Upper age limit:","GRM
cutoff:","Number of genetic variants","Bonferroni p","Sidak p","d.eff","Fully corrected p")
col2<-
c("",REGION_OF_INTEREST,IMP_OR_CHIP,POPULATION,LOWER_AGE,UPPER_AGE,GRM_CUTOFF,d,bonferroni.p,sidak.p,d.eff,cor
r.p)
outputExtended<-data.frame(col1,col2)
outputExtendedFile<-paste(REGION_OF_INTEREST,"_",IMP_OR_CHIP,"_AgeRange",LOWER_AGE,"-
",UPPER_AGE,"_",POPULATION,"_postGRMcutoff",GRM_CUTOFF,"_extendedModifiedPvalues.txt",sep="")
write.table(outputExtended,outputExtendedFile,row.names=FALSE)

c<-c(d,bonferroni.p,sidak.p,d.eff,corr.p)
output<-data.frame(d,bonferroni.p,sidak.p,d.eff,corr.p)
outputFile<-paste(REGION_OF_INTEREST,"_",IMP_OR_CHIP,"_AgeRange",LOWER_AGE,"-
",UPPER_AGE,"_",POPULATION,"_postGRMcutoff",GRM_CUTOFF,"_modifiedPvalues.txt",sep="")
write.table(output,outputFile,row.names=FALSE)
```

Chapter 6: *KLF1* as a Candidate Gene for Fetal Haemoglobin levels in Sickle Cell Disease

Figures 230

Tables 230

6.1. Introduction: 231
6.1.1. Background 231
6.1.2. Previous work in our laboratory on *KLF1* in SCD 233
6.1.3. Summary of my KLF1 projects 234

6.2. KLF1 intron 1 project: rs10407416 and HbF% in SCD 234
6.2.1. Introduction 234
6.2.2. Subjects and Methods 235
6.2.2.1. Subjects and phenotyping 235
6.2.2.2. Genotyping 235
6.2.2.3. Statistical analysis 236
6.2.3. Results 236
6.2.4. Conclusions 240

6.3. KLF1 zinc finger 2 variant p.His329Tyr 240
6.3.1. Introduction 240
6.3.2. Subjects 241
6.3.3. Methods 241
6.3.3.1. Summary of methodological approach 241
6.3.3.2. Genotyping 242
6.3.3.3. RNA from reticulocytes: preservation and extraction 242
6.3.3.4. Reverse transcription 242
6.3.3.5. Quantitative PCR 242
6.3.4. Results 243
6.3.4.1. Summary of cohort 243
6.3.4.2. Genotyping 243
6.3.4.3. qPCR results 246
6.3.5. Conclusions 248

6.4. Discussion 248

References 249

Appendix 1 251

Appendix 2 252

Figures

Figure 1 KLF1 regulates globin switching. 232
Figure 2 Map of KLF1 with known mutations associated with high HbF% phenotypes 233
Figure 3 Schematic representation of the genetic sequence of *KLF1*, plus positions of the two variants of interest. 234
Figure 4 Details of 285 patients in the rs10407416 analysis: panel (a) demographics of patients (b) HbF% levels 237
Figure 5 Chromatograms from Sanger sequencing, variant of interest is marked black. Panel (a) demonstrates homozygous wildtype CC (b) heterozygous GC (c) homozygous variant GG 239
Figure 6 Family tree for KLF1 zinc finger 2 family study 241
Figure 7 Chromatograms from Sanger sequencing, variant of interest (c.986A>G) is marked black. 245
Figure 8 Relative expression of KLF1, BCL11A and HBG2 in reticulocytes in our cohort (reference gene=Beta actin). 247

Tables

Table 1 Primer pairs used in KLF1 intron 1 project 235
Table 2 HbF% levels and genotyping results 237
Table 3 Multivariate regression results for a statistical model (first analysis) incorporating the KLF1 variant rs10407416 in association with Ln(HbF%) 237
Table 4 Multivariate regression results for a statistical model (refined analysis) incorporating the KLF1 variant rs10407416 in association with Ln(HbF%) 238
Table 5 Primer pairs used Sanger sequencing in KLF1 zinc finger 2 project 242
Table 6 Final primers used in qPCR for the KLF1 zinc finger 2 project 243
Table 7 Demographic and genotype details for major HbF% loci for zinc finger 2 project. After the proband and her daughter, the remainder are arranged in ascending order of HbF% level 244

6.1. Introduction:

6.1.1. Background

The clinical diversity of sickle cell disease (SCD) is unexplained by its defining single base change in the β-globin gene, *HBB.* HbF levels remain a key moderator of disease severity; increased HbF values are associated with a milder SCD phenotype.

While the three known quantitative trait loci (QTLs) for HbF (*BCL11A*, *HMIP* and *HBB)* have been reported to account for 20–50% of HbF variation in non-anaemic Europeans(Menzel et al., 2007), the contribution of these loci is estimated to be lower in African populations (Solovieff et al., 2010, Wonkam et al., 2014). This may be because other loci may be more important in African populations due to allele frequency differences at both known and unknown loci. Or, since many of the studies in African populations are in individuals with SCD, erythropoietic stress is a key determinant of HbF levels, thus "diluting out" the genetic contribution. This has been described in detail in my manuscript in Appendix 6, chapter 4.

KLF1 is an important candidate gene for HbF% levels in SCD. *KLF1* (previously termed *EKLF*) has been termed a 'master regulator' of erythropoiesis, playing a key role in globin-switching from HbF to HbA. It was discovered by Jim Bieker in 1993 (Miller and Bieker, 1993), is the key transcription factor controlling HbF identified through genetic studies in a Maltese family with β-thalassaemia and hereditary persistence of HbF (HPFH). Linkage studies with this family identified a locus (on chromosome 19p13) for the HPFH that segregated independently of the *HBB* locus. This linked region encompassed *KLF1* (Borg et al., 2010). Subsequent studies, which included expression profiling of erythroid progenitor cells, confirmed *KLF1* as the γ-globin gene modifier in this family. Family members with HPFH were heterozygous for the nonsense mutation K288X in *KLF1* that disrupts the DNA-binding domain of KLF1. Several studies have now confirmed that *KLF1* is key in the switch from *HBG* to *HBB* expression; it not only activates *HBB* directly, providing a competitive edge, but also silences the γ-globin genes indirectly via activation of the γ-globin repressor *BCL11A*, Figure 1 (Siatecka and Bieker, 2011, Zhou et al., 2010, Esteghamat et al., 2013). KLF1 has three zinc finger domains, which mediate sequence-specific binding to DNA and are essential for activation of KLF1 target genes.

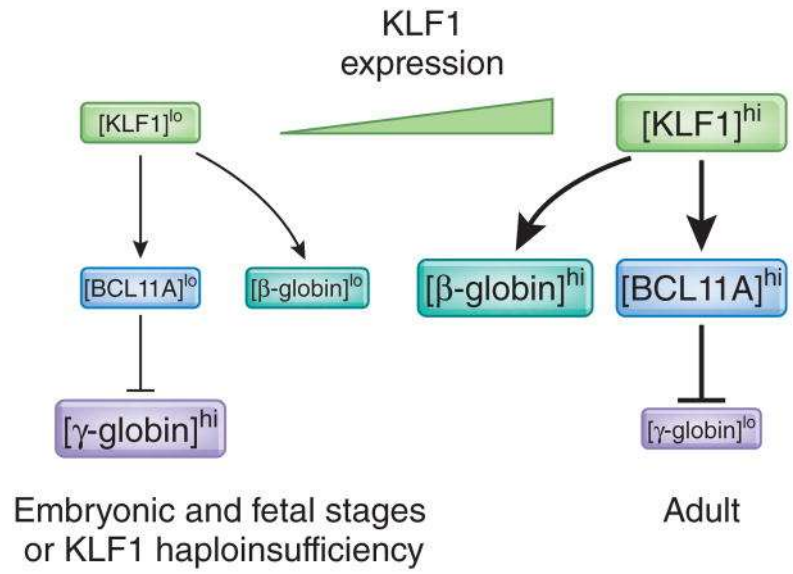


Figure 1 KLF1 regulates globin switching.

On the left, during embryonic and fetal development (or in adults with KLF1 mutations), KLF1 levels are low, resulting in low levels of adult β-globin and BCL11A and high levels of γ-globin. On the right, in adults with two functional copies of KLF1, increased expression of KLF1 in definitive red blood cells promotes high levels of adult β-globin and BCL11A expression, which in turn represses γ-globin expression. Taken from (Bieker, 2010)

There have been numerous reports of association of *KLF1* variants with increased HbF either as a primary phenotype, or in association with other red cell disorders (Borg et al., 2011), Figure 2. More recently, *KLF1* mutations have been noted not just to be more common in, but also to ameliorate the severity of β-thalassaemias(Liu et al., 2014). There is one report associating a *KLF1* mutation in a patient with SCD, the patient in question had high HbF% and a mild sickle phenotype(Gallienne et al., 2012). Such is *KLF1*'s pivotal role in globin switching that genetic therapeutic strategies (using the CRISPR/Cas9 system) are now being explored to disrupt *KLF1* in order to overexpress γ-globin(Shariati et al., 2016). Notably, several GWAS of HbF% (including ones in SCD patients of African descent) have failed to identify common *KLF1* variants (Bhatnagar et al., 2011, Mtatiro et al., 2014). Thus, the role of *KLF1* in SCD remains unclear and it is being actively explored as a genetic modifier of HbF% levels (and severity) in SCD.

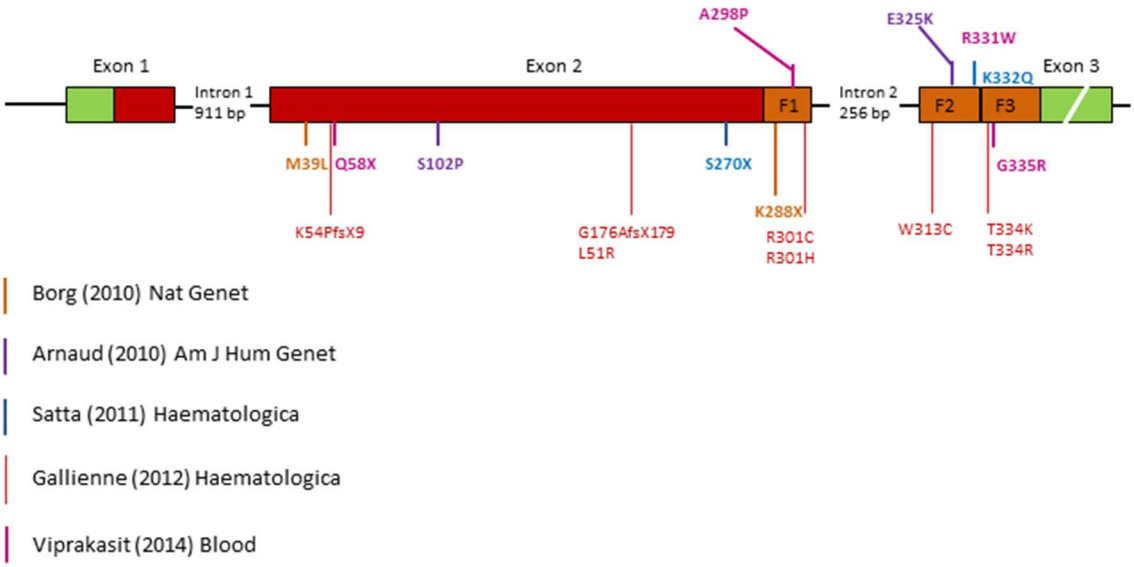


Figure 2 Map of KLF1 with known mutations associated with high HbF% phenotypes

(Borg et al., 2010, Arnaud et al., 2010, Satta et al., 2011, Gallienne et al., 2012, Viprakasit et al., 2014)

6.1.2. Previous work in our laboratory on *KLF1* in SCD

Our laboratory began evaluating *KLF1* in patients with SCD in 2012, prior to my recruitment. 250 patients from the sickle research gene bank were genotyped for four variants in the three HbF% QTLs (*BCL11A*, *HMIP* and *HBB*). Using a statistical model, "HbF%-residuals" were calculated for each patient; that is, HbF% not accounted for by either age, sex or the three known QTLs for HbF%. Thus, an ordered list of patients by "residual HbF%" was created. This parameter describes the component of a patient's high HbF% level that is not explained by currently known HbF% loci. The extreme highest and lowest 10% of HbF%-residuals (25 in each group) plus 10 intermediate patients were selected to have their *KLF1* genes fully sequenced using a modified method of Sanger cycle-sequencing and capillary electrophoresis on ABI3130. First, a variant in intron 1 of *KLF1* (rs10407416) was found to be over-represented in the high residual-HbF% group (9/25) compared to the low residual-HbF% group (1/25). rs10407416 is not currently known to be associated with any phenotype. Second, an unreported variant (c.986A>G) in the zinc finger 2 domain (exon 3) of *KLF1* was detected. This is a novel missense variant (p.His329Tyr) that affects a critical domain that is highly conserved across vertebrates. Crucially, in *KLF1*, the invariant histidine at position 329 coordinates the zinc atom (within the zinc finger) therefore its mutation is highly likely to be pathogenic. Figure 3 shows a schematic representation of the genetic sequence of *KLF1* plus the two variants of interest.

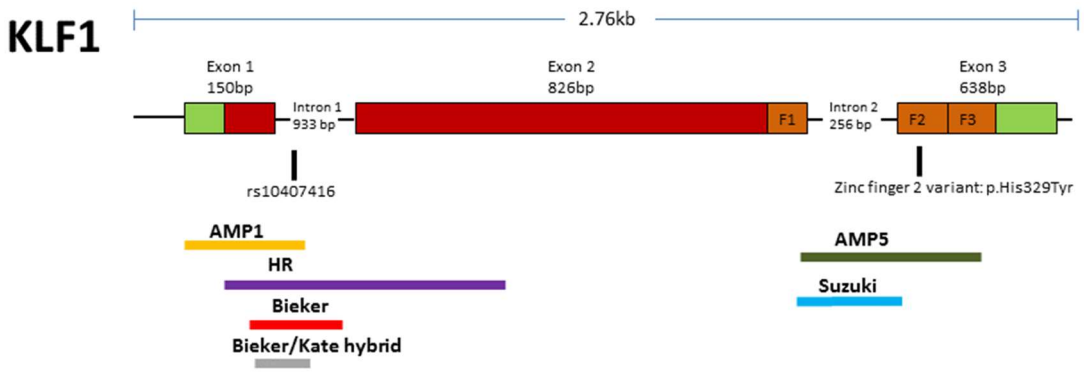


Figure 3 Schematic representation of the genetic sequence of *KLF1*, plus positions of the two variants of interest.

Amplicons from the primers used in the two projects are shown as coloured bars.

6.1.3. Summary of my KLF1 projects

Prompted by previous work in our laboratory, I undertook two studies to investigate if *KLF1* variants are associated with increased HbF% levels in our SCD cohort. Since most functional variants reported are (very) rare, one cannot adopt cohort-based GWAS or regions of interest studies using our MEGA chip data – these approaches aim to identify common variants. Instead, one must implement an individualised genotyping approach (not reliant on "tagging"). I Sanger-sequenced DNA to identify variants and attempted the use of family studies in which to undertake functional methods.

6.2. KLF1 intron 1 project: rs10407416 and HbF% in SCD

6.2.1. Introduction

The aim of my project was to investigate the *KLF1* candidate variant rs10407416 to evaluate its putative association with HbF% in SCD. I genotyped rs10407416 in 285 patients with HbSS or HbSβ^0-thalassaemia across the full spectrum of "HbF%-residuals" using Sanger sequencing. I used multiple regression testing with known HbF% loci results, age and sex as covariates, to test if the addition of rs10407416 could improve variability of HbF% values explained by genetics.

rs10407416 has a minor allele frequency of 4-10% across the African populations in the 1000 Genomes Project. While it is not currently associated with any clinical phenotype in the literature, it is very close to an intronic regulatory element (an intronic enhancer) and therefore may be of biological significance (Siatecka 2010).

6.2.2. Subjects and Methods

6.2.2.1. Subjects and phenotyping

285 patients with sickle cell disease (HbSS or HbSβ^0-thalassaemia) from the sickle genebank, who had already been genotyped for variants in the three HbF% QTLs *BCL11A*, *HMIP* and *HBB* were included. HbF% results were collected as described in chapter 2 (in summary, a baseline HbF% was used that was taken at least three months post-transfusion, not on hydroxycarbamide, and not while pregnant), and the natural logarithm was taken to normalise the parameter (i.e. Ln(HbF%)).

6.2.2.2. Genotyping

I used Sanger sequencing to genotype the variant rs10407416, in five steps:

Step 1: polymerase chain reaction (PCR) of amplicons including the variant rs10407416

Extensive PCR optimisation was undertaken prior to doing high throughput work. Primer pairs trialled are listed in Table 1 and pictorially represented in Figure 3. I needed to re-design primers and optimise PCR conditions for quality high throughput sequencing. Final thermocycling conditions were 5 minutes at 95°C, 33x (30 seconds at 95°C, 30 seconds at 60°C, 1 minute at 72°C), 5 minutes at 72°C, 10 minutes at 4°C.

Table 1 Primer pairs used in KLF1 intron 1 project

Amplicon name	Primer details	Amplicon length (bp)	Comments
Amp1	TACCCAGCACCTGGACCCT GAACCTCAAACCCCTAGACCACC	566	Rejected due to its proximity to variant of interest.
HR	TTGCCCTCCATCAGCACACTG CGAGTGATCCTCCGAACCCAAAA	1289	Developed by a colleague to incorporate a second, tag variant ~800 bases away. Rejected because of difficulty sequencing through a polyT tail.
Bieker	ACACAGGATGACTTCCTCAAGGT CCAGAACATCCCTCTCCTTCC	593	Rejected because it included the same polyT tail.
KG	Bieker forward GGCTACCTTCGTTTTCTATTACCG	255	Comprise the Bieker forward primer plus in-house designed new reverse primer at the variant side of the polyT tail.

Step 2: Purification of the PCR product using AMPure beads: on a robotic workstation Biomek® NXP Laboratory Automation Workstation

Step 3: Thermocycling DYE term step: Two separate experiments for each sample (forward and reverse primer). Again, extensive optimisation. Final thermocycling conditions for each plate (forward and reverse): 40x (30 secs at 95°C, 15 secs at 52°C, 2 min at 60°C), 10 min at

4°C. The DYE term step also took time to optimise: I added betaine, increased PCR volumes and changed PCR conditions.

Step 4: Purification of the DYE term product using CleanSEQ beads: on a robotic workstation Biomek® NXP Laboratory Automation Workstation

Step 5: Sequencing on ABI3130: ABI3130 sequencer using a standard sequencing run.

Results from the final step were analysed as chromatograms in the software Sequencher. Genotypes were only called if they were clear, unclear results were re-run.

6.2.2.3. Statistical analysis

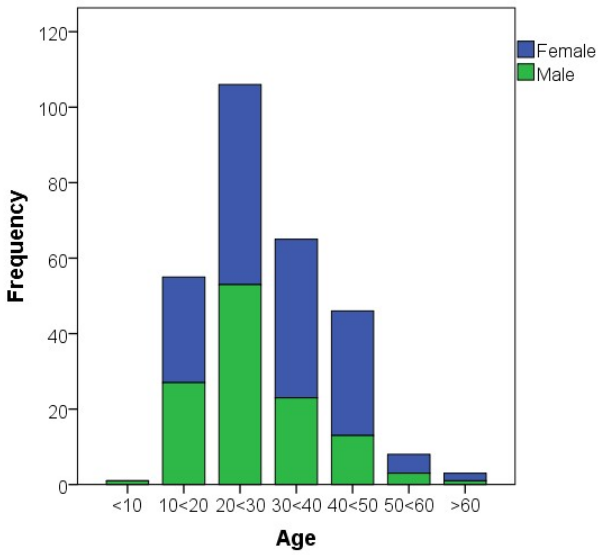
I conducted simple multiple regression in IBM SPSS version 22, with outcome Ln(HbF); and co-variates: age, sex, and four variant results to represent the above three QTLs: *BCL11A* (rs11886868), 2x *HMIP* (rs9399137, rs9402686), and *HBB* (rs7482144).

I then performed a second phase of statistical analysis with two refinements to the statistical model. First, following other groups' use, I considered using age-squared as a covariate rather than age. Second, I considered recent understanding of the genomic architecture of the HMIP-2 locus to build an improved model for the HbF% associated variants(Menzel et al., 2014). This is based on the difference between European and African haplotypes at HMIP-2: blocks A and B are in cis in European haplotypes, therefore one should use only one of the two blocks for regression analysis. European haplotypes can be identified by the separate marker rs9376090 alternate allele.

6.2.3. Results

285 patients aged at least 9 years old with SCD (HbSS or HbSβ^0-thalassaemia) were genotyped for rs10407416, see Figure 4 for demographic details and HbF% levels. Example chromatograms of the three genotypes are displayed in Figure 5. In total, 243 were wildtype (CC), 41 heterozygous for the variant (CG) and 1 homozygous (GG), see Table 2.

a



b

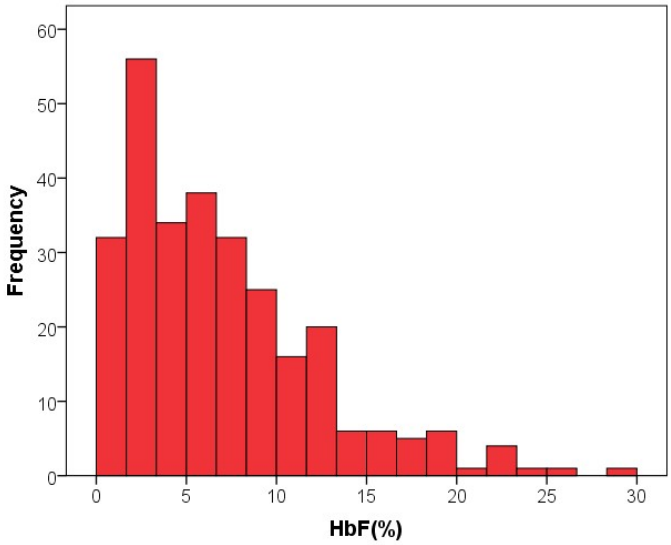


Figure 4 Details of 285 patients in the rs10407416 analysis: panel (a) demographics of patients (b) HbF% levels

Table 2 HbF% levels and genotyping results

	"-/-" C C	**"+/-" G C**	**"+/+" G G**	**"+/-" or "+/+" G C or G G**
Number	243	41	1	42
Mean HbF%	7.142+/-5.474	7.366+/-5.239	4	7.272+/-5.195
Mean HbF%-residuals	3.447+/-0.741	3.696+/-0.706	3.974	3.704+/-0.698

The results of the first multivariate regression analysis showed that rs10407416 did not statistically significantly improve the genetic model for estimating Ln(HbF%) levels (Table 3). By including rs10407416, the model reduced the variability of Ln(HbF%) explained, from r^2=0.261 without rs10407416 to r^2=0.245 with rs10407416.

After I performed the refined statistical analysis using (1) age-squared and (2) managing the genetic architecture of HMIP, the analysis maintained the variability of Ln(HbF%) explained at R^2=0.245 and rs10407416 remained statistically non-significant (Table 4).

Table 3 Multivariate regression results for a statistical model (first analysis) incorporating the KLF1 variant rs10407416 in association with Ln(HbF%)

	Beta (95% confidence intervals)	p-value
(Constant)	2.482 (1.319 to 3.645)	<0.001
Sex	0.548 (0.358 to 0.738)	<0.001
Age	0.007 (-0.001 to 0.016)	0.092
rs7482144	-0.001 (-0.352 to 0.349)	0.994
rs11886868	-0.412 (-0.553 to -0.271)	<0.001
rs9399137	-0.319 (-0.647 to 0.010)	0.057
rs9402686	-0.408 (-0.805 to -0.011)	0.044
rs10407416	0.226 (-0.027 to 0.479)	0.080

Table 4 Multivariate regression results for a statistical model (refined analysis) incorporating the KLF1 variant rs10407416 in association with Ln(HbF%)

	Beta (95% confidence intervals)	p-value
(Constant)	2.322 (1.223 to 3.421)	<0.001
Sex	0.550 (0.352 to 0.748)	<0.001
Age Squared	0.000 (0.000 to 0.000)	0.025
rs7482144	-0.080 (-0.489 to 0.329)	0.700
rs11886868	-0.429 (-0.577 to -0.282)	<0.001
HMIP aggregate score	-0.257 (-0.416 to -0.099)	0.002
rs10407416	0.212 (-0.052 to 0.476)	0.115

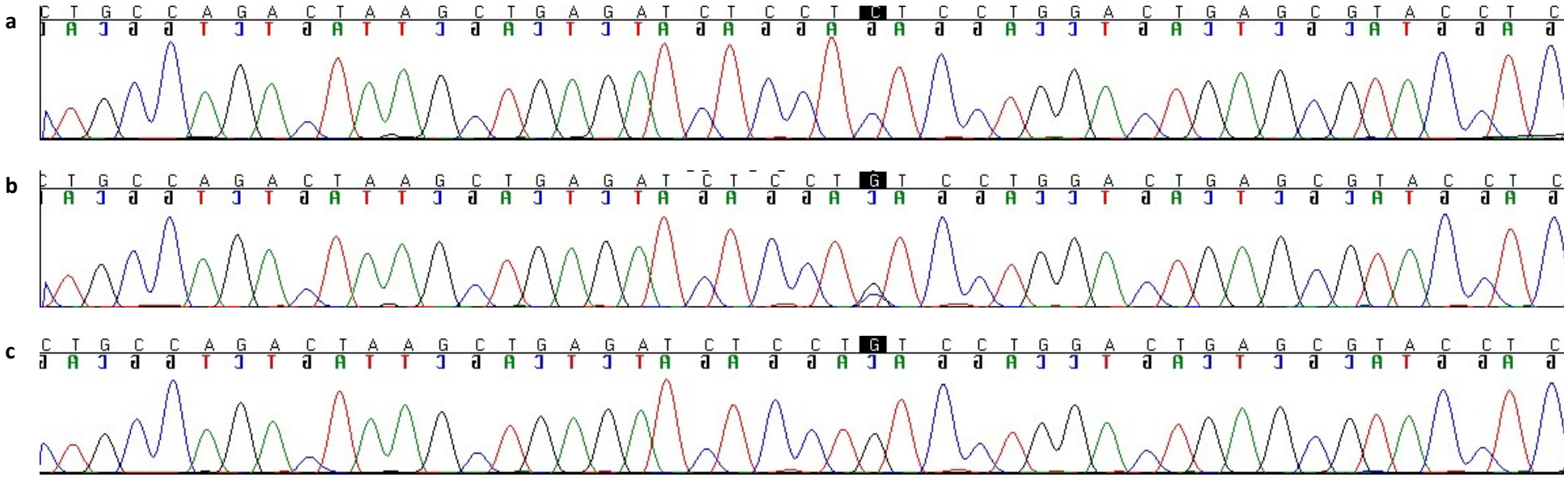


Figure 5 Chromatograms from Sanger sequencing, variant of interest is marked black. Panel (a) demonstrates homozygous wildtype CC (b) heterozygous GC (c) homozygous variant GG

6.2.4. Conclusions

In conclusion, despite refinements, using the KLF1 intronic variant rs10407416 does not improve the existing model for predicting HbF% in patients with SCD, with co-variates age, sex and 4 variants for 3 known QTLs for HbF%. Of note, this analysis has *not* taken account of population stratification and this is important for this variant: in African populations, minor allele frequency for rs10407416 is 4-10%, and for non-African populations is 0%. Therefore, the variant could represent ethnicity and our analysis is confounded by issues of admixture. Furthermore, the small sample size precludes us being able to reject categorically the association of this variant with Ln(HbF%). Therefore, evaluating the role of *KLF1* in SCD remains open.

6.3. KLF1 zinc finger 2 variant p.His329Tyr

6.3.1. Introduction

The aim of my project was to investigate the *KLF1* zinc finger 2 variant to see if it is a genetic modifier of SCD severity (via high HbF%) by assessing the role of this mutation in a patient with SCD with a very high HbF% (in the absence of Mendelian HPFH) accompanied by a very mild sickle phenotype. I performed gene expression studies on reticulocytes in this patient with high HbF% levels to investigate the profile of *BCL11A*, *KLF1* and *HBG* expression.

The purpose of a family study is to show co-segregation of a trait and variant: in this setting, that the *KLF1* zinc finger 2 variant co-segregates with very high HbF%/mild SCD.

The variant (c.986A>G) is a missense variant (p.His329Tyr) in KLF1's second zinc finger. There are no published reports to date of associations between this variant and high HbF% phenotypes. Waye and Eng have declared unpublished data with no phenotype reported(Waye and Eng, 2015). In another PhD student's project using the same SCD cohort, Matthew Shannon (co-supervised by SLT) discovered the same *KLF1* variant as one of the variants in exome sequencing as associated with "mildness" of disease severity.

There are multiple reasons that this variant could be causative of the elevated HbF% phenotype. There are chemical differences between histidine (which is basic) and tyrosine (hydrophobic, polar uncharged). Multiple mutation prediction software programmes suggest that this variant produces a deleterious change (using Align GVGD, SIFT and MutationTaster). Furthermore, its location within the second zinc finger is in a cluster of 11 contiguous amino acids where six different variants have already been associated with high HbF% levels(Arnaud et al., 2010, Satta et al., 2011, Gallienne et al., 2012, Viprakasit et al., 2014). This includes the

mutation c.973G>A (p.Glu325Lys) manifesting as congenital dyserythropoietic anaemia (CDA) type IV; a rare CDA with high HbF% levels (Arnaud et al., 2010, Jaffray et al., 2013).

6.3.2. Subjects

The family tree, with sickle status demonstrated, is in Figure 6.

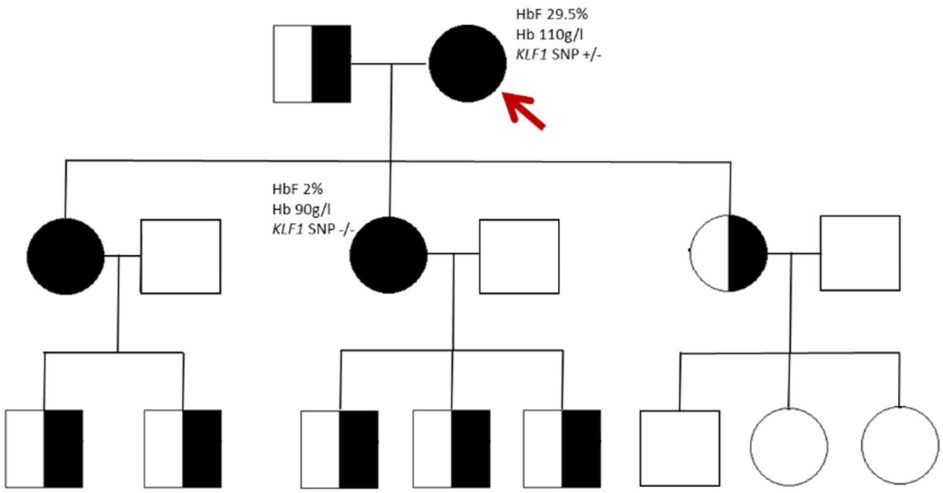


Figure 6 Family tree for KLF1 zinc finger 2 family study

SCD patients are shown with filled symbols, heterozygous carriers are half-filled.

I only managed to recruit the proband and one daughter; the remainder of the family declined recruitment. The proband is a 75-year-old with HbSS disease and no history of sickle acute pain events. Her only chronic end-organ complication is very mild albuminuria not requiring medication; her Hb is 110g/l and HPLC reveals HbF% 29.5%, HbS 64.4%, HbA2 2.1%. She does not have Mendelian HPFH (deletions in the β-globin cluster or mutations in the γ-globin gene promoters were excluded). The daughter is a 49-year-old with HbSS disease, she has a low HbF% and a relatively mild SCD phenotype; her Hb is 90g/l and HPLC reveals HbF% 2%, HbS 90.1%, and HbA2 3.8%. I used 20 sickle genebank HbSS/HbSβ^0 thalassaemia patients as controls. At the time of blood sampling, no patients were: taking hydroxycarbamide, had a recent transfusion (< 3 months) and or were pregnant.

6.3.3. Methods

6.3.3.1. Summary of methodological approach

First, I genotyped the variant using Sanger sequencing, and second, I used reverse transcription PCR (RT-PCR) to assess mRNA expression of KLF1, BCL11A and HBG in reticulocytes. The sickle genebank includes genomic DNA but not RNA samples so I preserved patients' fresh red cells in TRI reagent prior to RNA extraction.

6.3.3.2. Genotyping

I performed Sanger sequencing using the same method as above (section 6.2.2.2), using primers courtesy of Vip Viprakasit in Bangkok, see the schematic diagram of *KLF1* with amplicons in Figure 3 and primer details in Table 5.

Table 5 Primer pairs used Sanger sequencing in KLF1 zinc finger 2 project

Amplicon name	Primer details	Amplicon length (bp)
Amp5	TGTAAAACGACGGCCAGTGCGGCAAGAGCTACACCA CAGGAAACAGCTATGACCTTGTCCCATCCCCAGTCACT	gDNA 536 cDNA 88

Red indicates M13 sequences

6.3.3.3. RNA from reticulocytes: preservation and extraction

Reticulocytes were isolated, and RNA *preserved* from reticulocytes based on a standard protocol for RNA preservation from buffy coat using TRI reagent and modified for use for RNA preservation from reticulocytes in SCD. A key modification included a higher ratio of TRI reagent to reticulocyte. See Appendix 1 for the RNA preservation protocol. RNA was *extracted* based on standard methods. I optimised this for our SCD samples; increased isopropranol (1.5x) was used for RNA precipitation, see Appendix 2.

6.3.3.4. Reverse transcription

Reverse transcription was conducted after DNase treatment of 2µg RNA using standard methods. Invitrogen's reagents were used for reverse transcription and the manufacturer's protocol was followed. In brief, for every 11µl experiment (post-DNase STOP with 2µg RNA), first add random primers and dNTPs and place on the thermocycler at 65°C for 5 minutes. Second, add forward strand buffer, DTT 0.1, RNAsin and RT superscript and place on the thermocycler: for 25°C for 5 minutes, 50°C for 60°C minutes and 70°C for 15 minutes.

6.3.3.5. Quantitative PCR

Quantitative PCR (qPCR) was conducted once the reverse transcription method was optimised and after proof of cDNA synthesis using standard PCR/gel electrophoresis. Since *KLF1* acts via *BCL11A* as well as directly, I evaluated gene expression of: *KLF1*, *BCL11A*, γ-globin gene *HBG2,* and housekeeping gene β-*actin*. I trialled multiple KLF1, BCL11A and housekeeping primers, the final (best-performing) primers are listed in Table 6. Notably, KLF1 primers performed very variably between standard PCR and qPCR.

Positive and negative controls were used for all runs.

Table 6 Final primers used in qPCR for the KLF1 zinc finger 2 project

Primer name	Primer details	cDNA bp	gDNA bp
bActin_IntS_	AAATCGTGCGTGACATTAAGG	229	324
	ATGATGGAGTTGAAGGTAGTT		
BCL11A exon1&2	AACCCCAGCACTTAAGCAAA	114	None
	GGAGGTCATGATCCCCTTCT		
KLF1 Suzuki	GTTGCGGCAAGAGCTACAC	80	336
	GCAGGCGTATGGCTTCTC		
gamma globin_LS	GAGAAACCCTGGGAAGGCTC	334	334
	CCAGTCACCATCTTCTGCCA		

6.3.4. Results

6.3.4.1. Summary of cohort

Demographic details for the proband and her daughter, plus 20 HbSS/HbSβ^0 thalassaemia controls, are displayed in Table 7. The 20 control samples are ordered in increasing HbF% values. Also in this table are the genotypes for major HbF% genetic loci across BCL11A, HMIP-2 and HBG2-Xmn1.

6.3.4.2. Genotyping

The proband had heterozygous genotype CT for the *KLF1* c.986A>G variant and her daughter homozygous wildtype TT, see chromatograms in Figure 7. All other samples tested were wildtype homozygous.

Table 7 Demographic and genotype details for major HbF% loci for zinc finger 2 project. After the proband and her daughter, the remainder are arranged in ascending order of HbF% level

	Sex	Sickle genotype	Alpha genotype	HbF %%	Age	KLF1 variant	BCL11A		HMIP-2					Xmn1-HBG2
							rs1427407	rs6545816	rs9494142	rs6920211	rs9494145	rs9376090	rs66650371	rs7482144
Proband	F	HbSS	aa/aa	29.5	75	+/-	G G	C C	T T	C T	T T	T T	I I	G G
Provand's daughter	F	HbSS	aa/aa	2	49	-/-	G G	A C	T T	T T	T T	T T	I I	G G
1	F	HbSS	aa/aa	1	20	-/-	G G	A C	T T	T T	T T	T T	I I	G G
2	M	HbSβ^0	?	1.8	23		G T	A A	C T	C T	C T	T T	I I	G G
3	F	HbSS	aa/aa	2	36	-/-	G G	A C	T T	T T	T T	T T	I I	G G
4	M	HbSS	aa/aa	2.5	17		?	?	?	?	?	?	?	G G
5	F	HbSS	aa/aa	2.7	23		G G	A C	T T	C T	T T	T T	I I	G G
6	F	HbSS	aa/aa	2.7	18		G T	A A	C T	C T	T T	T T	I I	G G
7	F	HbSS	aa/aa	2.9	20		G G	A C	T T	C T	T T	T T	I I	G G
8	F	HbSS	aa/aa	3.3	19		G G	A A	T T	T T	T T	T T	I I	G G
9	F	HbSS	aa/aa	3.3	22		G G	C C	T T	C C	T T	T T	I I	G G
10	F	HbSS	aa/a-	3.9	23		G G	A C	C T	C C	T T	T T	I I	G G
11	F	HbSS	?	4.1	19		?	?	?	?	?	?	?	?
12	M	HbSS	aa/a-	6.2	45		T T	A A	T T	C T	T T	T T	I I	A G
13	F	HbSS	?	7.7	16		G T	A C	T T	C T	T T	T T	I I	G G
14	F	HbSS	aa/aa	8.3	18		G G	A C	T T	C T	T T	T T	I I	G G
15	M	HbSS	aa/aa	8.7	44	-/-	G G	A A	T T	C T	T T	T T	D I	G G
16	M	HbSS	aa/aa	8.7	28		T T	A A	C T	C T	T T	T T	I I	G G
17	F	HbSS	aa/aa	11.9	42	-/-	G G	A C	T T	T T	T T	T T	I I	G G
18	F	HbSS	aa/aa	13.6	46		G T	A C	C T	C T	C T	T T	I I	A G
19	M	HbSS	aa/aa	15.6	20	-/-	G G	A C	T T	C T	T T	T T	D I	A G
20	F	HbSS	aa/aa	19.2	14		G T	A C	C T	C T	C T	T T	I I	G G

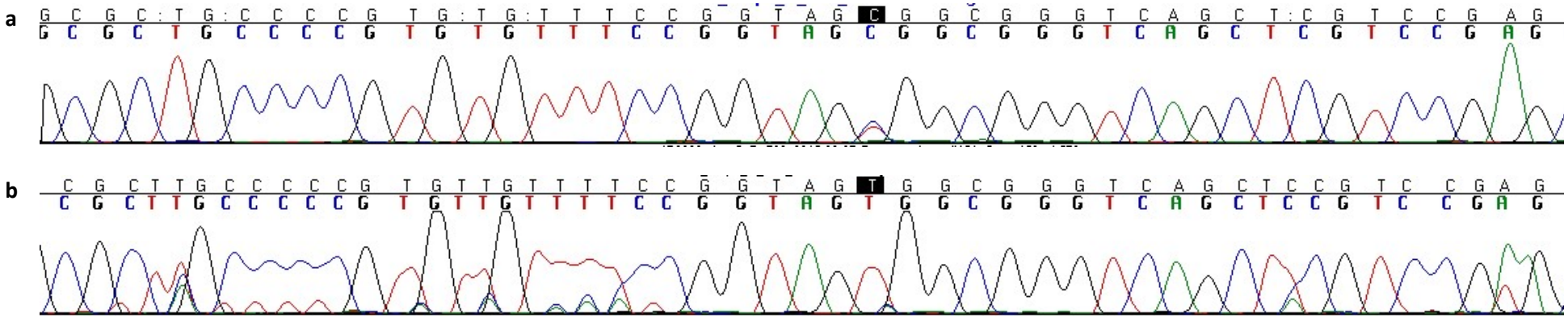


Figure 7 Chromatograms from Sanger sequencing, variant of interest (c.986A>G) is marked black.

I have used the negative strand as the reference strand. Panel (a) proband heterozygous CT (b) proband's daughter homozygous TT [Note the reference strand used was positive so the quoted variant A>G becomes T>C.]

6.3.4.3. qPCR results

Quantitating cDNA amounts from qPCR is based on C_t values representing the number of PCR cycles to reach a threshold[1]. The relative expression of any given gene is quoted relative to a housekeeping gene (β-actin). See Figure 8 for graphs of the relative expression of *KLF1, BCL11A* and *HBG2*. The results of the HbSS controls are ordered 1-20 in increasing HbF% baseline results.

[1] There is a negative linear relationship between C_t value and log (initial cDNA concentration). Thus, sample cDNA amount can be quantitated relative to one sample. Aberrant C_t values were discarded, and since assays were performed in triplicate, a mean C_t value for each sample/primer pair was used. To quote a relative expression, the C_t for a given primer/sample was normalised to the housekeeping gene, and then exponentiated to produce a quantitative result

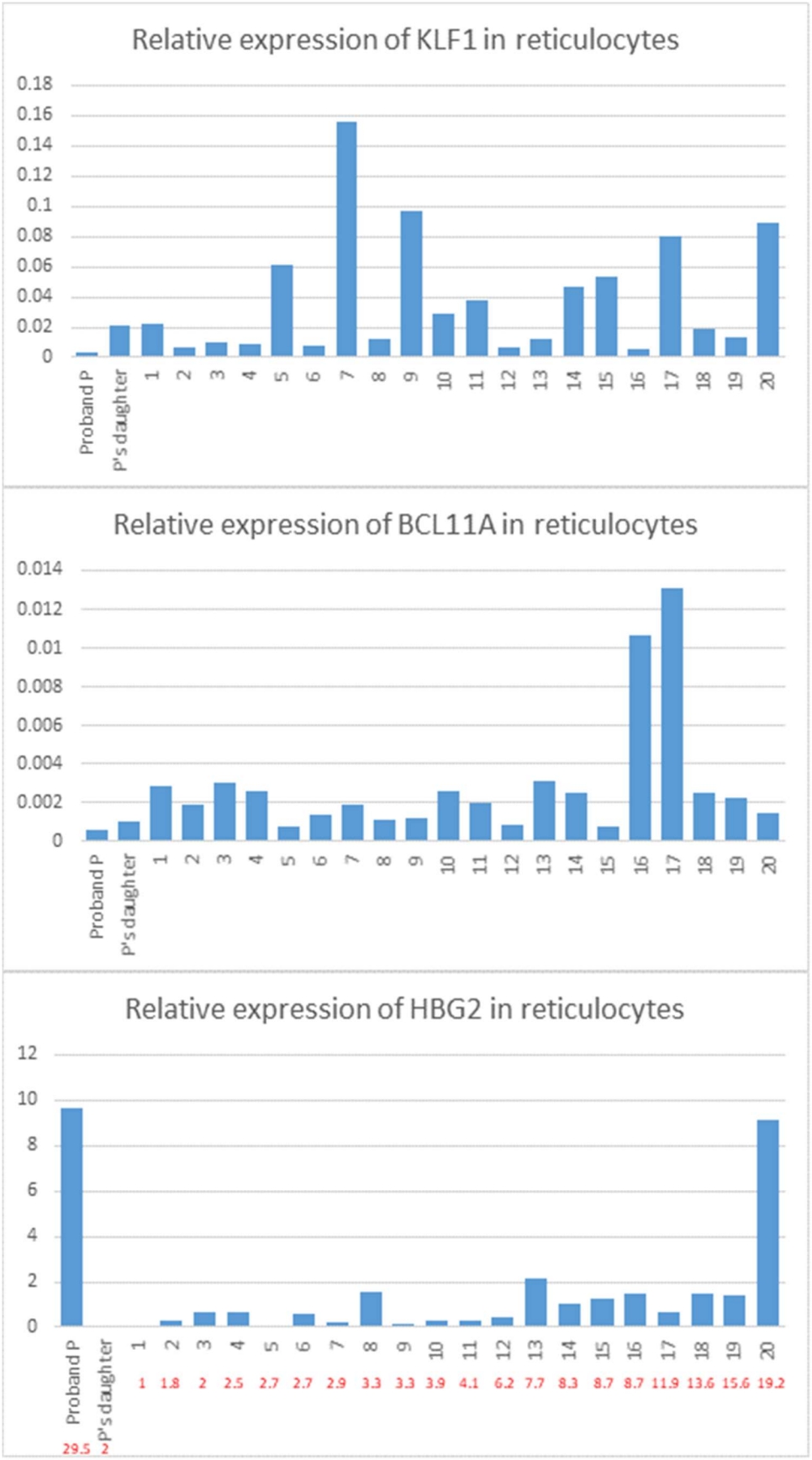


Figure 8 Relative expression of KLF1, BCL11A and HBG2 in reticulocytes in our cohort (reference gene=Beta actin).

The control samples are displayed in order of increasing HbF% levels, and absolute HbF% are also displayed in red.

6.3.5. Conclusions

In our cohort of 22 patients with SCD, I identified a large diversity of *KLF1* expression levels in red cells. The proband (with very high HbF and mild SCD phenotype) has the lowest expression of both *KLF1* and the γ-globin repressor *BCL11A* of any patient. However, this is not markedly less than other individuals. The results are consistent with a role of the mutation in increasing HbF% and potentially reducing sickle severity, but this is inconclusive.

Notably, the qPCR approach enables us to measure the consequences of a *quantitative* change in KLF1 caused by the KLF1 variant but does not allow us to assess any possible *qualitative* change in KLF1. It is possible that the KLF1 mutation mediates its effect through a qualitative change to the protein.

As with previous *KLF1* mutations identified, this is a heterozygous variant, indicating that a single *KLF1* allele can boost HbF%.

In the project, I used RNA from a mixed population of reticulocytes where levels may already be changing; it may be preferable to have a more homogeneous red cell population e.g. to use primary erythroid progenitor cells instead.

The disparity between the proband and her daughter's HbF% levels, in the setting of mild disease in both subjects, raises the question of whether another (non-HbF% related) factor may be contributing to the mild phenotype in both patients in this family.

Finally, as previously discussed, it is more difficult to assess the genetic component of HbF% levels in SCD as in this setting stress erythropoiesis is also determinant of HbF% levels.

6.4. Discussion

HbF% is a well-established quantitative trait which has been extensively researched and much is now understood about its oligo-genic basis. Recent work has more accurately characterised the functional determinants of its variability. However, there remains "missing heritability".

Multiple GWAS have not identified a further locus contributing to common variation of HbF% apart from the known three trait loci. Instead, it is likely that many rare variants individually contribute significantly to an individual's HbF%, but that the low frequency of the HbF%-boosting allele does not affect cohorts. Therefore, rather than cohort-based genetic

association studies one must adopt strategies for research that allow us to identify rare variants, such as whole exome sequencing and family studies.

KLF1 is an ideal candidate gene to search for high HbF% variants. I investigated two variants in KLF1 and their association with high HbF% in SCD. KLF1 – a "master erythroid regulator" – controls the globin switch and multiple (rare) variants have been identified and associated with high HbF% phenotypes. I have presented data consistent with a new, and rare, missense variant in association with reduced KLF1 expression and increased γ-globin expression in association with a phenotype with high HbF% and virtually asymptomatic SCD.

References

ARNAUD, L., SAISON, C., HELIAS, V., LUCIEN, N., STESCHENKO, D., GIARRATANA, M. C., PREHU, C., FOLIGUET, B., MONTOUT, L., DE BREVERN, A. G., FRANCINA, A., RIPOCHE, P., FENNETEAU, O., DA COSTA, L., PEYRARD, T., COGHLAN, G., ILLUM, N., BIRGENS, H., TAMARY, H., IOLASCON, A., DELAUNAY, J., TCHERNIA, G. & CARTRON, J. P. 2010. A dominant mutation in the gene encoding the erythroid transcription factor KLF1 causes a congenital dyserythropoietic anemia. *Am J Hum Genet,* 87**,** 721-7.

BHATNAGAR, P., PURVIS, S., BARRON-CASELLA, E., DEBAUN, M. R., CASELLA, J. F., ARKING, D. E. & KEEFER, J. R. 2011. Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J Hum Genet,* 56**,** 316-23.

BIEKER, J. J. 2010. Putting a finger on the switch. *Nat Genet,* 42**,** 733-4.

BORG, J., PAPADOPOULOS, P., GEORGITSI, M., GUTIERREZ, L., GRECH, G., FANIS, P., PHYLACTIDES, M., VERKERK, A. J., VAN DER SPEK, P. J., SCERRI, C. A., CASSAR, W., GALDIES, R., VAN IJCKEN, W., OZGUR, Z., GILLEMANS, N., HOU, J., BUGEJA, M., GROSVELD, F. G., VON LINDERN, M., FELICE, A. E., PATRINOS, G. P. & PHILIPSEN, S. 2010. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat Genet,* 42**,** 801-5.

BORG, J., PATRINOS, G. P., FELICE, A. E. & PHILIPSEN, S. 2011. Erythroid phenotypes associated with KLF1 mutations. *Haematologica,* 96**,** 635-8.

ESTEGHAMAT, F., GILLEMANS, N., BILIC, I., VAN DEN AKKER, E., CANTU, I., VAN GENT, T., KLINGMULLER, U., VAN LOM, K., VON LINDERN, M., GROSVELD, F., BRYN VAN DIJK, T., BUSSLINGER, M. & PHILIPSEN, S. 2013. Erythropoiesis and globin switching in compound Klf1::Bcl11a mutant mice. *Blood,* 121**,** 2553-62.

GALLIENNE, A. E., DREAU, H. M., SCHUH, A., OLD, J. M. & HENDERSON, S. 2012. Ten novel mutations in the erythroid transcription factor KLF1 gene associated with increased fetal hemoglobin levels in adults. *Haematologica,* 97**,** 340-3.

JAFFRAY, J. A., MITCHELL, W. B., GNANAPRAGASAM, M. N., SESHAN, S. V., GUO, X., WESTHOFF, C. M., BIEKER, J. J. & MANWANI, D. 2013. Erythroid transcription factor EKLF/KLF1 mutation causing congenital dyserythropoietic anemia type IV in a patient of Taiwanese origin: review of all reported cases and development of a clinical diagnostic paradigm. *Blood Cells Mol Dis,* 51**,** 71-5.

LIU, D., ZHANG, X., YU, L., CAI, R., MA, X., ZHENG, C., ZHOU, Y., LIU, Q., WEI, X., LIN, L., YAN, T., HUANG, J., MOHANDAS, N., AN, X. & XU, X. 2014. KLF1 mutations are relatively more common in a thalassemia endemic region and ameliorate the severity of beta-thalassemia. *Blood,* 124**,** 803-11.

MENZEL, S., GARNER, C., GUT, I., MATSUDA, F., YAMAGUCHI, M., HEATH, S., FOGLIO, M., ZELENIKA, D., BOLAND, A., ROOKS, H., BEST, S., SPECTOR, T. D., FARRALL, M., LATHROP, M. & THEIN, S. L. 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet,* 39**,** 1197-9.
MENZEL, S., ROOKS, H., ZELENIKA, D., MTATIRO, S. N., GNANAKULASEKARAN, A., DRASAR, E., COX, S., LIU, L., MASOOD, M., SILVER, N., GARNER, C., VASAVDA, N., HOWARD, J., MAKANI, J., ADEKILE, A., PACE, B., SPECTOR, T., FARRALL, M., LATHROP, M. & THEIN, S. L. 2014. Global genetic architecture of an erythroid quantitative trait locus, HMIP-2. *Ann Hum Genet,* 78**,** 434-51.
MILLER, I. J. & BIEKER, J. J. 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the *Krüppel* family of nuclear proteins. *Molecular and Cellular Biology,* 13**,** 2776-2786.
MTATIRO, S. N., SINGH, T., ROOKS, H., MGAYA, J., MARIKI, H., SOKA, D., MMBANDO, B., MSAKI, E., KOLDER, I., THEIN, S. L., MENZEL, S., COX, S. E., MAKANI, J. & BARRETT, J. C. 2014. Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One,* 9**,** e111464.
SATTA, S., PERSEU, L., MOI, P., ASUNIS, I., CABRIOLU, A., MACCIONI, L., DEMARTIS, F. R., MANUNZA, L., CAO, A. & GALANELLO, R. 2011. Compound heterozygosity for KLF1 mutations associated with remarkable increase of fetal hemoglobin and red cell protoporphyrin. *Haematologica,* 96**,** 767-70.
SHARIATI, L., KHANAHMAD, H., SALEHI, M., HEJAZI, Z., RAHIMMANESH, I., TABATABAIEFAR, M. A. & MODARRESSI, M. H. 2016. Genetic disruption of the KLF1 gene to overexpress the gamma-globin gene using the CRISPR/Cas9 system. *J Gene Med,* 18**,** 294-301.
SIATECKA, M., LOHMANN F., BAO S. & BIEKER J. J. 2010. EKLF Directly Activates the $p21^{WAF1/CIP1}$ Gene by Proximal Promoter and Novel Intronic Regulatory Regions during Erythroid Differentiation. *Mol Cell Biol,* 30, 2811-22.
SIATECKA, M. & BIEKER, J. J. 2011. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood,* 118**,** 2044-54.
SOLOVIEFF, N., MILTON, J. N., HARTLEY, S. W., SHERVA, R., SEBASTIANI, P., DWORKIS, D. A., KLINGS, E. S., FARRER, L. A., GARRETT, M. E., ASHLEY-KOCH, A., TELEN, M. J., FUCHAROEN, S., HA, S. Y., LI, C. K., CHUI, D. H., BALDWIN, C. T. & STEINBERG, M. H. 2010. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood,* 115**,** 1815-22.
VIPRAKASIT, V., EKWATTANAKIT, S., RIOLUEANG, S., CHALAOW, N., FISHER, C., LOWER, K., KANNO, H., TACHAVANICH, K., BEJRACHANDRA, S., SAIPIN, J., JUNTHARANIYOM, M., SANPAKIT, K., TANPHAICHITR, V. S., SONGDEJ, D., BABBS, C., GIBBONS, R. J., PHILIPSEN, S. & HIGGS, D. R. 2014. Mutations in Kruppel-like factor 1 cause transfusion-dependent hemolytic anemia and persistence of embryonic globin gene expression. *Blood,* 123**,** 1586-95.
WAYE, J. S. & ENG, B. 2015. Kruppel-like factor 1: hematologic phenotypes associated with KLF1 gene mutations. *Int J Lab Hematol,* 37 Suppl 1**,** 78-84.
WONKAM, A., NGO BITOUNGUI, V. J., VORSTER, A. A., RAMESAR, R., COOPER, R. S., TAYO, B., LETTRE, G. & NGOGANG, J. 2014. Association of variants at BCL11A and HBS1L-MYB with hemoglobin F and hospitalization rates among sickle cell patients in Cameroon. *PLoS One,* 9**,** e92506.
ZHOU, D., LIU, K., SUN, C. W., PAWLIK, K. M. & TOWNES, T. M. 2010. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. *Nat Genet,* 42**,** 742-4.

Appendix 1
Protocol for RNA preservation from reticulocytes

Equipment

- 15ml Falcon tubes
- Pasteur pipettes
- Centrifuge

Reagents

- Fresh whole blood in EDTA 4.5ml
- Phosphate buffered saline (PBS)
- Tri reagent

Method

- Tip 4.5ml fresh whole blood into 15ml Falcon and clearly label with study number and date
- Pipette PBS into sample tube and tip into 15ml falcon until total volume 14ml
- WASH ONE: suspend blood by inverting tube, then centrifuge at 3000rpm for 10mins at 4°C
- Pipette and discard off PBS supernatant (do not disturb buffy coat)
- WASH TWO: top-up falcon tube with PBS supernatant, re-suspend blood by inverting tube, then centrifuge again at 3000rpm for 10mins at 4°C
- Pipette and discard off PBS supernatant (do not disturb buffy coat)
- WASH TWO: top-up falcon tube with PBS supernatant, re-suspend blood by inverting tube, then centrifuge again at 4000rpm for 30mins at 4°C
- Pipette and discard off PBS supernatant (do not disturb buffy coat)
- Using a pipette, aspirate the buffy coat in one swooping manoeuvre, and discard
- Using another pipette, aspirate the top 0.5ml of underlying red cells into a fresh 15ml Falcon tube
- Immediately add 1.5ml TRI reagent and mix thoroughly by Vortexing. The mixture becomes brown and viscous.
- Immediately store at -80°C

Appendix 2
Protocol for RNA extraction

Equipment

- 1.5ml Eppendorf tubes
- Pipettes/pipette tips
- Desktop microfuge capable of spinning at 14000rpm at 4°C
- Nanodrop

Reagents

- Bromochloro Fresh whole blood in EDTA 4.5ml
- Isopropranol Phosphate buffered saline
- 75% ethanol Tri reagent
- Molecular grade water

Method

- Defrost the TRI-preserved RNA and use immediately
- Add 500µl to a 1.5ml Eppendorf
- PHASE SEPARATION
    - Add 37.5µl (10% TRI volume) of bromochloro, vortex and leave for 5 mins standing
    - Centrifuge for 15 mins at 14000rpm at 4°C
    - Note outcome of top (aqueous) layer which contains the RNA; DNA and proteins in lower organic layers
- RNA PRECIPITATION
    - Pipette off as much aqueous phase into a second 1.5ml Eppendorf without disturbing lower layers, and note volume
    - Add 3x aqueous phase volume transferred of isopropranol, vortex and leave for 10 mins standing
    - Centrifuge for 10 mins at 14000rpm at 4°C
    - Note (pellet size) indicative of volume RNA precipitated
    - Remove supernatant but leave pellet
- RNA WASH
    - Add 500µl of 75% ethanol, and dislodge pellet
    - Centrifuge for 5 mins at 4000rpm at 4°C
    - Remove ethanol, but leave pellet and then repeat RNA wash step once
    - Leave to dry out for about 10 minutes (don't overdry the pellet)
- WATER SOLUBILISATION
    - Add 30µl molecular grade water and pipette pellet up and down to solubilise the RNA
- RNA QUANTIFICATION
    - Use the Nanodrop to measure the concentration and contamination (260/280 and 260/230 ratios) and note values
- Immediately store at -80°C

Chapter 7: Conclusions

7.1. Summary of findings 254
7.1.1. Genotype-phenotype association studies 254
7.1.2. Phenotyping 254
7.1.3. Genotyping 255
7.1.4. Association analysis 255
7.1.5. KLF1 lab projects 257

7.2. Work in context 257

7.3. Future work 258
7.3.1. Replication of signals of association 258
7.3.2. Beyond GWAS: from association to function 258
7.3.3. Specific work on PK-LR 259
7.3.4. Improvements to statistical modelling 259
7.3.5. New genetic projects (non-association analyses) 260

References 261

7.1. Summary of findings

7.1.1. Genotype-phenotype association studies

Information-rich genomic data has transformed the study of genetic variants, especially in white populations. Analyses in African-heritage and mixed populations are one step behind European populations: not only are there fewer studies, but analysis needs to take account of the increased genetic complexity. The (recent) advent of both African-specific genome-wide micro-arrays and African genomes in public genetic reference panels means the tools are now available to analyse these non-white populations more accurately.

Successful genotype-phenotype association studies require (1) clear and meaningful phenotyping (of heritable traits), (2) accurate genotyping and (3) appropriate statistical analysis for association.

7.1.2. Phenotyping

Phenotyping is important for all clinical research, as well as for genotype/phenotype association studies. In SCD, there is no single, accepted definition or marker of severity. I have defined and applied several clinical endpoints as phenotypes in genetic association studies.

These phenotypes reflect the global spectrum of disease: *haemolytic index* (which reflects the degree of haemolysis); *hospitalisation rate* (which reflects frequency of pain episodes) and *mortality*. I have also considered the well-established intermediate phenotype in SCD, *fetal haemoglobin* (HbF), and *proteinuria* (via urinary albumin creatinine ratio, uACR), which reflects the severity of the commonest end organ damage: sickle nephropathy. I have used data collected over a 10-year period, and taken averages to ensure values reflect baseline status. For laboratory data, I have calculated arithmetic means wherever multiple data points were available for a patient and for hospitalisation rate, I have accounted for an individual's observation period over the 10-year study period.

I published a survival study for patients with SCD at KCH. This demonstrated that estimated median survival of 67 years in HbSS disease – a significant improvement in survival compared to other recent estimates in the USA and Jamaica. We also confirmed associations between mortality in SCD and markers of cardiorespiratory dysfunction, renal impairment, and haemolysis as well as frequent hospitalisation rate.

Using multiple severity indicators as phenotypes, rather than a single 'severity index', better reflects the clinical complexity of SCD and allows us to evaluate the genetic architecture of

specific phenotypes. Furthermore, by comparing results between phenotypes, one may be able to gain some understanding of global mechanisms contributing towards severity in SCD. Moreover, the use of multiple phenotypes potentially confers increased statistical power as "severity" is defined in different ways.

7.1.3. Genotyping

Genome-wide micro-arrays are the ideal platform for evaluating markers in the human genome, enabling identification of genetic variants amongst millions of markers.

Illumina's MEGA chip includes markers chosen on the basis of genetic polymorphisms in an African-American cohort, and its use in African-heritage populations better reflects genetic variation in this population than European-specific panels.

I have created and curated a large genetic database of over 15 million markers for our regional sickle cohort. This was time-consuming and required bioinformatics experience. Quality control is imperative for genetic association analysis to prevent false-positive and false-negative results. It must take account of multiple aspects of the quality of samples and genetic markers.

Imputation is now standard practice in genome-wide association analyses. Imputation enhances the sensitivity of association analyses (increasing power) and facilitates fine-mapping to get closer to the causative mutation. Imputation also facilitates meta-analysis: we are collaborating with a Tanzanian group who have undertaken micro-array analysis on a sickle cohort.

I have assessed relatedness via generation of a "genetic relatedness matrix". This takes account of both near (cryptic) relatedness as well as issues of population stratification by quantifying the pair-wise relatedness between all individuals. I have used this to identify duplicates in our sickle cohort (this mostly reflects the same individual attending two hospital clinics), and thus one of a pair of close relatives can be removed for some analyses.

7.1.4. Association analysis

I undertook genetic association analysis using two approaches: genome-wide association analysis (GWAS) and candidate gene analysis. For both, the use of the genome-wide data in a linear mixed model allowed me to take account of relatedness. Notably, this has been a failing

of many previous studies, where identified genetic associations have been (or may be) confounded by population stratification (ethnicity).

For GWAS, I evaluated fetal haemoglobin, hospitalisation rates, a haemolytic index and uACR as quantitative markers of severity of SCD. My analyses replicated previous findings for HbF, hospitalisation rates, haemolytic index and uACR.

These data were also used to construct a polygenic score model for HbF% which uses four variants in a quantitative trait to estimate specifically the *genetic* component of HbF%. We have termed this summary variable for the genetics of HbF% ***"HbFg"*** and suggest its use as a covariate in future clinical and genetic studies in SCD.

The GWAS analysis also identified some tentative novel loci for hospitalisation rates and haemolytic index. For hospitalisation rates, peak signals for three regions were at *rs75904749* on chromosome 5, *rs10792490* on chromosome 11 and *rs510384* on chromosome 12. The latter is within an intron of gene *SLC6A13*, a sodium-dependent GABA and taurine transporter. For haemolytic index, peak signals were a copy number variant in *HBA2* and rs4695226 on chromosome 4.

The genome-wide analyses remain under-powered in our relatively small SCD cohort, so I progressed to candidate gene studies which have greater power to detect new risk variants.

Using a candidate gene approach, I assessed several regions of interest in association with the severity indices hospitalisation rate and haemolytic index. First, candidate genes with the potential to increase red cell 2,3-DPG levels (and increase intracellular deoxy-HbS) and, second, complement-related genes. I identified risk variants in the red cell pyruvate kinase gene *PK-LR* associated with hospitalisation rate in SCD. I used our HbSS patients as the discovery set, and validated four variants in *PK-LR* in our HbSC cohort, and a meta-analysis of *all* SCD patients showed improved p-values. If this association is validated in further cohorts, these intronic risk variants are less likely to represent the causal element itself than to be in linkage disequilibrium with the actual functional DNA change. However, and notably, rs8177970 is strongly predicted to create a new 5' splice site only when the variant is present. These risk variants are all common in African populations (14-15% in African populations in 1000 Genomes Project).

I also investigated specific genetic variants and their association with sickle complications. The association of the G1 variants in *APOL1* with proteinuria was replicated. However, the *DARC* promoter variant rs2814778 was not associated with any severity index. rs2814778 is an ancestry-informative marker only seen in African-heritage populations. Positive findings in previous studies may have been artefacts due to confounding by admixture and population stratification. In our study, I accounted for relatedness (including population structure) in our statistical modelling, thus removing the artefact.

7.1.5. KLF1 lab projects

I investigated the role of KLF1 in HbF% levels in patients with SCD. I have provided evidence that a very rare variant in KLF1 zinc finger is associated with increased γ-globin expression via reduced *KLF1* and *BCL11A* expression.

A common *KLF1* intronic polymorphism was also evaluated for its association with HbF% levels in SCD, but no link was identified in the larger population I assessed.

7.2. Work in context

Many genotype-phenotype association studies in SCD have been published in the last decade. As in the non-sickle setting, many of the resulting findings have not been replicated. Notable exceptions to the large number of marginal findings in the field of β-haemoglobinopathies is the discovery of BCL11A as a key repressor of γ-globin, and the *HBS1L-MYB* (*HMIP*) region as a quantitative trait locus for HbF%. As in many GWA studies, the *HMIP* locus lies in a gene-free region, and studies have now provided evidence that the inter-genic region contains causative variants in erythroid-specific enhancers that not only regulate HbF but also other pleiotropic haematological parameters via *MYB*, one of the flanking genes.

Genetic association studies today must take account of genome-wide information on relatedness – we are now in the era of linear mixed model analysis.

While the proof of any novel genetic association is intrinsically linked to showing replication in an independent dataset, efforts to optimise all stages of phenotype-genotype association can be made. I have taken care over each aspect of association. First, I have used quality phenotypes: I have collected data over a 10-year period and considered traits that are variable. Second, careful genotyping is required: I have made use of a new African-specific genotyping array (Illumina's MEGA chip) as well as the recent availability of multi-ethnic reference panels for imputation. I have been complete and conservative with quality control measures. Third, I

have carefully built a statistical-analysis strategy, using up-to-date knowledge in the field to construct sophisticated analysis models to take account of relatedness, including population structure. Linear mixed modelling is now becoming the gold standard for genetic association studies. Optimising each of these three steps helps increase the power of the study to disentangle true signals from those due to chance.

7.3. Future work

7.3.1. Replication of signals of association

Signals of association should be validated through replication and/or meta-analysis. To confirm positive association signals from an initial study, it is imperative to replicate the result in *independent* samples (either from the same or different populations). More widely in GWA studies, replication of positive association signals has not proved to be easy for logistical and statistical reasons: it depends on the power of both initial and replication studies. As in our candidate gene studies, another approach is to genotype a subset (in our case, HbSS) and follow-up the strongest signals of association in a different subset (HbSC). Collaboration between international groups studying the same trait allows for *in silico* replication.

In replication and meta-analyses, rather than direct data exchange, summary statistics are provided instead: for each genetic variant, the "risk" allele, p-value, β +/- standard error are exchanged for quantitative traits. Recall that imputation can be used to combine different GWAS to enable direct comparison of different genotype sets.

7.3.2. Beyond GWAS: from association to function

After identification of new GWAS loci, it is tempting to speculate on the molecular pathways by which these loci affect trait risk. However, one must be cautious and apply a methodical approach before making assertions. Multiple steps are required in post-GWAS analyses in order to identify the functional variant(s) responsible for the observed risk-differences, and to dissect the molecular pathways underlying their effects(Edwards et al., 2013). First, fine-scale mapping of the associated locus using imputed genotypes must be undertaken to identify a set of associated variants. As in our studies, imputation may have contributed to fine-mapping the causal variant already. Second, prioritisation of putative functional variants (using epidemiological information and bioinformatics) to select candidates that could contribute to the trait. Some identified loci are near well-characterised genes, which may be good candidates for being the causal gene. Other loci lie in 'gene deserts' with no nearby genes. Third, *in vitro* and *in vivo* functional analyses must then take place to confirm and further elucidate the mechanisms by which the causal variants are acting. Fundamentally, in order to

establish causality formally (of a given variant for a trait), one must show that recreating the risk variants in human cells or animal models generates analogous phenotypes in the model.

7.3.3. Specific work on PK-LR

We plan to continue to investigate these *PK-LR* risk variants and their association with hospitalisation rate in SCD. First, we aim to replicate our findings in other cohorts, but are limited by sickle cohorts where these data are available (North America and Cameroon). We must be careful of interpretation of results for hospitalisation rate as there are complex non-genetic determinants of this trait, including socio-cultural and economic considerations, which are different to the UK. Furthermore, many North American sickle cohorts comprise paediatric not adult patients.

We also aim to do more genotyping of the *PK-LR* region both to confirm findings and look for potential associated functional variants in this very polymorphic gene.

Measuring pyruvate kinase (PK) protein levels in the red cells of patients seems intuitive but could prove difficult; there are no formal reference ranges of red cell PK in SCD. In SCD, red cells with low PK are not removed by the (functionally impaired) spleen. This also means that, *in SCD*, a genetic diagnosis of PK deficiency (however mild) is likely to be more appropriate than a proteomic diagnosis.

If these findings corroborate PK-LR as a genetic disease severity modifier, we must consider functional studies using cell or animal models of SCD with degrees of PK deficiency.

7.3.4. Improvements to statistical modelling

Our statistical modelling could be revisited to consider improvements. Modifications to the statistical modelling could include adding in more covariates, including ***HbFg*** (our summary genetic variable for the three HbF% loci), haemoglobin levels or haemolytic index. In this way, other aspects of SCD pathophysiology can be controlled for while trying to understand the genetic determinants of our trait of interest.

Alternative methods of analysis have also been proposed since the modelling was undertaken. While linear mixed models (LMM) with genetic relatedness matrices (GRM) to control for all forms of relatedness is now common in genetic analyses, more recently the addition of principal components (PC) to a LMM with a GRM has been advocated. There are two key reasons for this.

First, PC capture additional variance, resulting from non-sample relatedness such as genotyping batch effects. Some very recent studies recommend including principal components as fixed effects in addition to the random effect of the GRM in mixed association models(Yang et al., 2014, Zhang and Pan, 2015).

Second, the setting of extreme differences in allele frequencies in different populations may undermine the validity of GRMs(Price et al., 2010). By adding PCs to a LMM, the PCs can take account of significant differences. While our measures of model accuracy (lambda GC measuring genomic inflation, and QQ plots) were satisfactory (indeed, excellent in some cases), this may improve modelling and therefore power. However, of note, as part of the modelling, I undertook subgroup analysis by sickle genotype to test the sensitivity of the results. HbSS sub-analysis yielded results that were largely concordant with those obtained from ALL sickle genotypes.

As an aside, the issue of subgroup analysis versus whole-cohort analysis highlights issues around power achieved in larger (less homogeneous) populations versus smaller (more homogeneous) populations. This balance is different for GWA studies and candidate gene analyses, which have more power to identify variants.

Finally, there are now new statistical methods available to take account of multiple phenotypes to increase the power in studies of association between variants and phenotypes. This way, correlated traits (such as our severity indices) can be used in a multi-phenotypic statistical model to evaluate association with genetic variants. *MultiPhen* is an R package which performs association testing between genetic variants and multiple phenotypes(O'Reilly et al., 2012). It performs regression analysis where variants are treated as the *outcome* and multiple phenotypes as the *predictors*; this can result in large increases in statistical power to detect genotype-phenotype associations over the univariate approach.

7.3.5. New genetic projects (non-association analyses)

The substantial genetic dataset assembled, in an unusual population, opens up many new avenues for genetic analyses, not necessarily incumbent on my present work. This includes evaluation of the *heritability* of traits using genome-wide data (in contrast to twin studies). These analyses using genome-wide data are relatively new and have been criticised for overestimating heritability, however, they offer an insight into the degree with which a trait is heritable in order to drive the search for variants. As phenotypes are better understood, one

can create *polygenic score models* – estimating the effects of genetic variants jointly – as we have done for HbF%. As the genetic basis of our more experimental phenotypes becomes better characterised, a polygenic model could be created to explain these outcomes, too. Taking this further, the ultimate goal would be a *genetic prediction model* which robustly differentiates "mild" and "severe" phenotypes. Stratification on a genetic basis would have particular clinical significance in SCD: for those identified as severe, disease-modifying therapies can be considered early (hydroxycarbamide, transfusion programmes) or even haematopoietic stem cell transplantation. Crucially, early identification of "severe" cases means that these (potentially toxic) therapies can be offered to at-risk patients prior to the onset of end-organ damage.

An alternate approach to take this project forward would be to consider different, improved phenotypes. We have strength in the quality of the clinical data of our adult cohorts, with results from over a decade long period. These longitudinal data could be harvested to evaluate rate of progression of complications. This potentially gives more meaningful genetic association of clinical phenotypes which could transfer back to the clinic to identify mechanisms for disease progression.

Finally, a further potential field of study will be the testing for gene-environment interactions. The environmental risk factors for a disease may act at least partly via interactions with genetic risk determinants. Therefore, the validity of the genetic findings will be enhanced by accounting for these environmental exposures as either potential confounders or effect modifiers. By not accounting for environmental effects, one is exposed to errors that lead to both false positives and false negative results. In order to manage or investigate environmental exposures, depending on whether they act as confounders or effect modifiers, they can be incorporated into analyses by adjusting for their effect (as a fixed covariate) or by assessing their potential interaction with genetic region of interest.

References

EDWARDS, S. L., BEESLEY, J., FRENCH, J. D. & DUNNING, A. M. 2013. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet,* 93**,** 779-97.

O'REILLY, P. F., HOGGART, C. J., POMYEN, Y., CALBOLI, F. C., ELLIOTT, P., JARVELIN, M. R. & COIN, L. J. 2012. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One,* 7**,** e34861.

PRICE, A. L., ZAITLEN, N. A., REICH, D. & PATTERSON, N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet,* 11**,** 459-63.

YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. & PRICE, A. L. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet,* 46**,** 100-6.

ZHANG, Y. & PAN, W. 2015. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genet Epidemiol,* 39**,** 149-55.