2018

# Investigating cis- and trans-acting elements involved in regulating fetal hemoglobin gene expression using high throughput genetic data

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

# INVESTIGATING CIS- AND TRANS-ACTING ELEMENTS INVOLVED IN REGULATING FETAL HEMOGLOBIN GENE EXPRESSION USING HIGH THROUGHPUT GENETIC DATA

by

**ELMUTAZ SHAIKHO ELHAJ MOHAMMED**

B.S., University of Khartoum, 2002
M.S., University of the Sciences in Philadelphia, 2013

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2018

Approved by

First Reader: _____

Martin H. Steinberg, Ph.D.
Professor of Medicine

Second Reader: _____

Paola Sebastiani, Ph.D.
Professor of Biostatistics

## DEDICATION

In the Name of Allah, the All-Merciful, the Most Compassionate.

To my parents, who define every bit of good in me; to my siblings, my anchors in a fragile world; and to my friends, the delights of life.

## ACKNOWLEDGMENTS

I would like to thank the members of my committee for their patience, advice, and support through my research. I am hugely appreciative to Dr. Yan Dai for performing the experimental validation of *BCL2L1* and *HBG* mRNA correlation. I am also grateful to the co-authors of the following papers that represent the core of this dissertation.

1- Shaikho EM, Farrell JJ, Alsultan A, Qutub H, Al-Ali AK, Figueiredo MS, Chui DHK, Farrer LA, Murphy GJ, Mostoslavsky G, Sebastiani P, Steinberg MH. **A phased SNP-based classification of sickle cell anemia HBB haplotypes. BMC Genomics.** 2017 Aug 11;18(1):608. doi: 10.1186/s12864-017-4013-y. PubMed PMID: 28800727; PubMed Central PMCID: PMC5553663.

2- Habara AH, Shaikho EM, Steinberg MH. **Fetal hemoglobin in sickle cell anemia: The Arab-Indian haplotype and new therapeutic agents.** Am J Hematol. 2017 Jul 24. doi: 10.1002/ajh.24872. [Epub ahead of print] Review. PubMed PMID: 28736939.

3- Shaikho EM, Farrell JJ, Alsultan A, Sebastiani P, Steinberg MH. **Genetic determinants of HbF in Saudi Arabian and African Benin haplotype sickle cell** anemia. Am J Hematol. 2017 Sep;92(9):E555-E557. doi: 10.1002/ajh.24822. Epub 2017

4- Shaikho EM, Habara AH, Alsultan A, Al-Rubaish AM, Al-Muhanna F, Naserullah Z,

Alsuliman A, Qutub HO, Patra PK, Sebastiani P, Baltrusaitis K, Farrell JJ, Jiang Z, Luo HY, Chui DH, Al-Ali AK, Steinberg MH. **Variants of ZBTB7A (LRF) and its β-globin gene cluster binding motifs in sickle cell anemia.** Blood Cells Mol Dis. 2016 Jul;59:49-51. doi: 10.1016/j.bcmd.2016.04.001. Epub 2016 Apr 13. PubMed PMID: 27282567.

Funded in part by R01 HL 068970, RC2 HL 101212, R01 8768, T32 HL007501, T32 GM074905 (KB) from the NIH Bethesda, MD, and the University of Dammam, SP 11/2011, Office of Collaboration and Knowledge Exchange, University of Dammam.

# INVESTIGATING CIS- AND TRANS-ACTING ELEMENTS INVOLVED IN REGULATING FETAL HEMOGLOBIN GENE EXPRESSION USING HIGH THROUGHPUT GENETIC DATA

**ELMUTAZ SHAIKHO ELHAJ MOHAMMED**

Boston University

Graduate School of Arts and Sciences and College of Engineering, 2018

Major Professor: Martin H Steinberg, MD, Professor of Medicine

ABSTRACT

Sickle cell anemia is caused by a single mutation in the β-hemoglobin gene, *HBB*. The disease originated in Africa and affects millions of people worldwide. Sickle hemoglobin tetramers polymerize upon deoxygenation and lead to hemolysis and vaso-occlusion. Patients with high fetal hemoglobin (HbF) can have milder disease. The only FDA-approved drug is hydroxyurea that increases HbF. HbF modulates the disease by preventing the polymerization of sickle hemoglobin and reduces the pain episodes, anemia, and organ damage associated with the disease. There are five common haplotypes associated with the HbS gene and that are very loosely associated with disease severity and HbF. Understanding the genetic bases of HbF regulation is a key factor to identify potential drug targets to induce HbF for therapeutic purposes.

To fully understand the mechanism behind HbF regulation, developing a fast and accurate computational method for sickle cell haplotype classification is useful for examining the variability of HbF among sickle cell patients. Moreover, investigating the

cis and trans-acting regulators of HbF gene expression to pinpoint the mechanism through which they regulate HbF is essential to develop a successful treatment. The availability of high-throughput genetic data provides an excellent opportunity to study HbF regulation in sickle cell patients and normal people comprehensively.

The work reported in this thesis describes a fast and accurate method for sickle cell *HBB* haplotype classification. I also examine the differential effect of cis and trans-acting HbF hemoglobin regulators on γ-globin gene expression using the GTEx database and identify *BCL2L1* as a new potential trans-regulator of HbF.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# Chapter 1. Introduction

## Sickle cell anemia

Sickle cell anemia is a genetic disorder of hemoglobin that is a result of homozygosity for a GAG-GTG mutation in the 6$^{th}$ position of the β-globin gene (*HBB*) that codes for sickle hemoglobin (HbS) (Ingram 1957). The disease affects millions worldwide and had its origins in Africa, the Middle East, and India (Flint et al. 1998; Kulozik et al. 1986). Upon deoxygenation, sickle hemoglobin tetramers polymerize and lead to many complications including hemolysis and vaso-occlusion as shown in **Figure 1**. Sickle cell anemia is characterized primarily by chronic anemia and periodic episodes of pain. The disease phenotypes are heterogeneous and have severe complications affecting the pulmonary, cardiovascular, renal, and other systems including lung injury, systemic and pulmonary hypertension, stroke, cutaneous leg ulceration, kidney injury, proteinuria, and osteonecrosis (Kato et al. 2017). At the moment, no cure is available other than by hematopoietic stem cell transplantation in rare patients. Otherwise, treatment strategies include, blood transfusion, induction of fetal hemoglobin by hydroxyurea, management of vaso-occlusive crisis and chronic pain syndromes with analgesics, prevention and treatment of infections, and treating various organ damage syndromes associated with the disease like stroke and pulmonary hypertension(Piel et al. 2017). Allogeneic bone marrow transplantation (BMT) can cure sickle cell disease, BMT requires aggressive chemical intervention and about 5% of patients die with complications of transplantation;

the procedure is also very expensive and finding identical matched sibling donors is difficult (Walters et al. 1996).

**Fetal hemoglobin in sickle cell anemia**

Fetal hemoglobin (HbF) is a well-known modulator of the pathophysiology of sickle cell disease (Habara and Steinberg 2016; Rees et al. 2010). HbF is mostly excluded from the deoxy sickle hemoglobin (HbS) polymer. This reduces the tendency of deoxyHbS to polymerize, in turn diminishing polymer-induced damage to the sickle erythrocyte and the many downstream consequences of this pathophysiologic event. Patients with high levels of HbF can have milder disease, but HbF does not ameliorate equally all disease complications. This is a result of the complex pathophysiology of disease that is caused by sickle vaso-occlusion and intravascular hemolysis, each with a different effect on disease complications, and the variation in the distribution of the concentrations of HbF amongst HbF containing erythrocytes, or F-cells. The latter results in in varying red cell protection from the adverse effects of HbS polymerization because of the heterogeneous cellular concentrations of HbF (Kato et al. 2007; Steinberg et al. 2014). For example, high HbF is associated with fewer acute painful episodes but an effect of HbF on stroke is less apparent and patients with the same HbF concentration can have different disease severity because the distribution of HbF amongst erythrocytes is very different (Kato et al. 2017). Key to devising novel means of inducing therapeutically high levels of HbF is a detailed understanding of the mechanisms underlying the normal switch from HbF synthesis in the fetus to HbA synthesis in adults.

**Hemoglobin switching in the β-globin gene cluster**

One feature of the transition from embryonic to fetal to extra-uterine life is the synthesis of globin polypeptides characteristic of each developmental stage. Switching in the *HBB* cluster is controlled by transcription factors that interact directly with these genes, or indirectly with other transcription factors to affect gene expression. **Figure 2** displays the chromosomal order of the β-like globin genes and the associated cis-acting elements that contain binding sites for the trans-acting elements shown in **Table 1** and in this figure (Cui et al. 2014; Cui et al. 2015; Ngo and Steinberg 2015). *BCL11A*, a major repressor of γ-globin gene (*HBG2, HBG1*) transcription, has been intensively studied (Lettre and Bauer 2016; Sankaran et al. 2008; Sankaran and Weiss 2015; Zhou et al. 2010). Polymorphic variants in the erythroid-specific enhancers of this gene are being targeted using gene editing as a potential cell-based approach to increase HbF (Bauer et al. 2013). Recently BCL11A binding sites in the *HBG* promoters were presumed to be the key loci through which this transcription factor exerted its effects (Nan Liu 2017). Leukemia/lymphoma-related factor (LRF), encoded by *ZBTB7A*, another zinc finger transcription factor, is another important repressor of *HBG* transcription (Masuda et al. 2016). *ZBTB7A* and *BCL11A* acted independently as *HBG* silencers and are parts of different nucleosome remodeling deacetylase (NuRD) complexes (**Figure 2**).

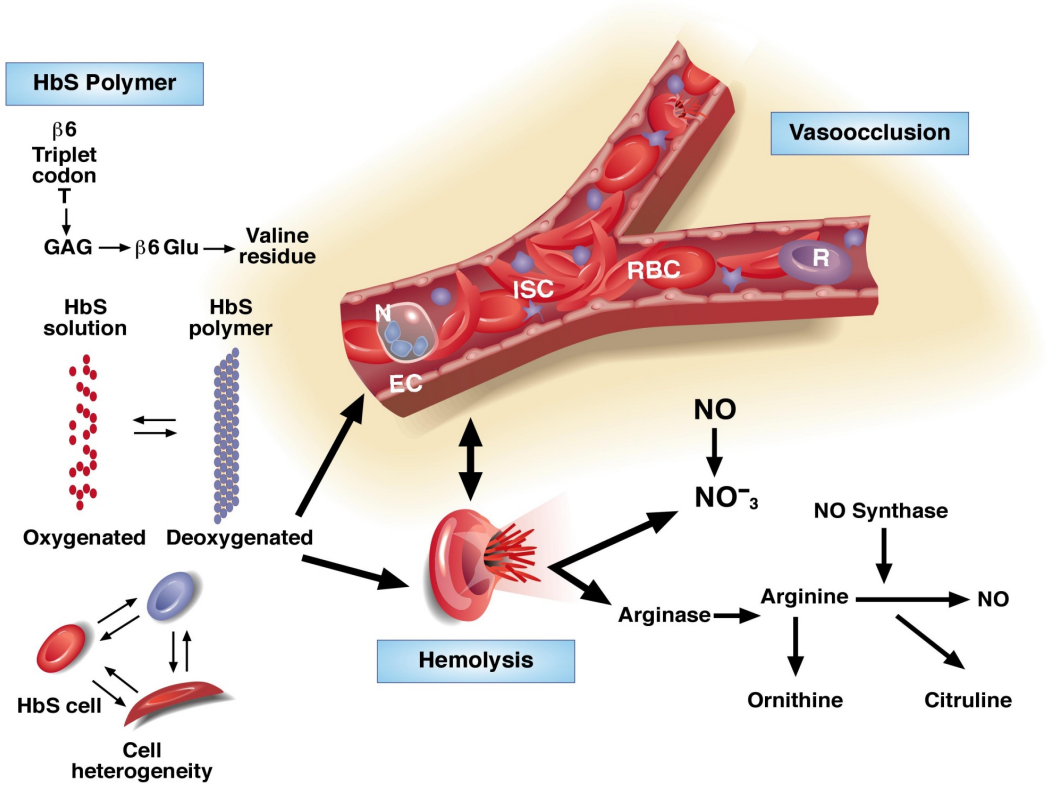



Figure 1. The pathophysiology of sickle cell disease as a result of mutation of glu6val which leads to polymerization of hemoglobin upon deoxygenation. The sickle polymer causes irreversible damage of erythrocytes, vasoocclusion and hemolysis (Steinberg 2008).

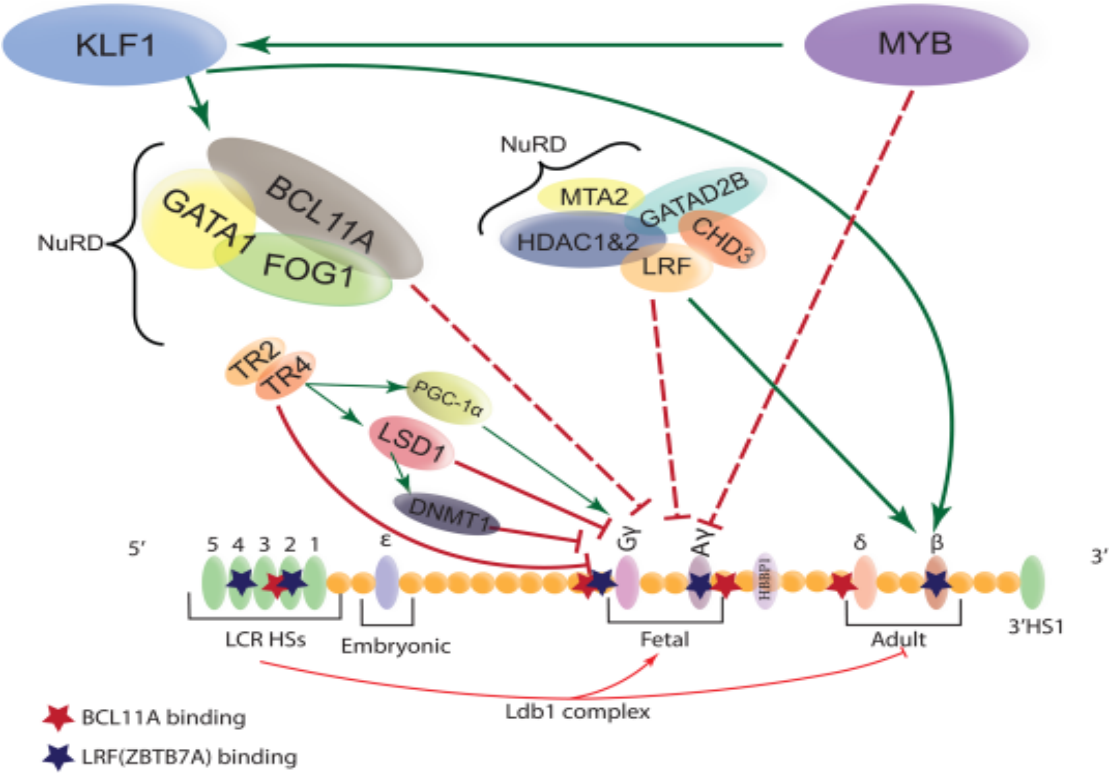



Figure 2: Cis- and trans-acting effectors of gene expression within the β-globin gene cluster. The Ldb1 complex: (Ldb1/LMO2/GATA1/Tal 1) occupies the locus control region (LCR) and gene promoters and facilitates looping to globin gene promoters that could be modulated by cis-acting variants associated with HbS gene haplotypes. Throughout the *HBB* gene complex and its flanking regions are binding domains for BCL11A and ZBTB7A (LRF) shown as red and blue stars, respectively. Separate nucleosome remodeling deacetylase (NuRD) complexes are associated with BCL11A and LRF (ZBTB7A). *MYB* has a direct effect on HbF gene expression and also acts indirectly through *KLF1* and *BCL11A*. Suppression of the HbF genes is depicted by dashed lines. This figure is not drawn to scale; based on Figure 3 from (Sankaran and Weiss 2015).

| Transcript/Gene | Effect | Comment |
|---|---|---|
| BCL11A | Direct inhibition | Master controller of *HBG* to *HBB* switching |
| KLF1 | Indirect inhibition | Affects *BCL11A* |
| MYB | Indirect inhibition | Affects *KLF1* |
| LRF (ZBTB7A) | Direct inhibition | Causes active chromatin formation |
| LSD1 | Direct inhibition | Silences the *HBG* promotor region |
| ANTXR1a | Inhibition | Effects on *HBG* only hypothesized |
| PPARGC1A (PGC-1α)b | Direct induction | Induces *HBG* with TR2 & TR4. Overexpression of PGC-1α has been shown to induce *HBG* in Lin- murine bone morrow cells. |

Table 1: Some genes with established roles in *HBG* expression

### *ZBTB7A* and Sickle Cell Anemia

Fetal hemoglobin (HbF) is the major modulator of the phenotype of sickle cell anemia and increased concentrations can reduce disease severity. HbF levels are controlled by genetic elements linked to the haplotype of the *HBB gene* cluster and by trans-acting quantitative trait loci (QTL). *ZBTB7A* (LRF) is a suppressor of γ-globin gene (*HBG2, HBG1*) expression. HbF levels were 70% after *ZBTB7A* knockout in a human erythroid cell line (HUDEP-2) that normally expressed adult HbA on terminal differentiation (Masuda et al. 2016). Knocking out both *ZBTB7A* and *BCL11A* increased HbF to more than 90% of total hemoglobin (Masuda et al. 2016). *ZBTB7A* and *BCL11A* acted independently as *HBG* silencers. In previous genome-wide association studies (GWAS) variants of *ZBTB7A* or in linkage disequilibrium (LD) with this gene were not associated with HbF (Lettre et al. 2008; Mtatiro et al. 2014; Ngo et al. 2013; Solovieff et al. 2010; Uda et al. 2008). Because of the profound effect of *ZBTB7A* on HbF expression we asked whether polymorphisms of this gene or its promoters and proximal enhancer elements, and in putative binding motifs for ZBTB7A in and adjacent to the *HBB* gene cluster are associated with HbF levels in sickle cell anemia.

Genetic data from GWAS and next generation sequencing from diverse patients who were homozygous for the sickle hemoglobin gene and had varying levels of HbF were available for analysis (**Table 2**). Included were 21 individuals studied by whole genome sequencing (WGS) and 15 Saudi Benin and 8 Saudi AI haplotype homozygotes studied using exome sequencing GWAS were available from 822 African American HbS

homozygotes of diverse haplotypes (Cooperative Study of Sickle Cell Disease or CSSCD cohort), 104 Saudi AI haplotype homozygotes who originated from the Eastern Province and 71 Saudi Benin haplotype homozygotes who originated from the Southwestern Province. GWAS data were also imputed to the 1000 Genomes (Phase 3) reference. All studies were approved by the Institutional Review Boards of the participating institutions.

To detect variants in *ZBTB7A* and its promoters or putative proximal enhancers that might be associated with HbF we searched 100 kb upstream and 100 kb downstream of the *ZBTB7A* coding sequences using data from WGS. Using IMPUT2 the SNPs in this region were imputed to the 1000 Genomes (Phase 3) reference panel in the CSSCD and Saudi cohorts that were studied by GWAS in order to test associations of HbF with these variants using an additive genetic model. The most significant association with HbF in the CSSCD cohort did not pass the correction for multiple testing (rs114623325, p-value 0.002). Two of the 23 Saudi patients studied by exome sequencing with HbF levels of 6.9% and 14.3% were heterozygous for a CGC insertion polymorphism (transcript NM_015898, c.539_540insCGC (p.Ala181_Ser182insAla) that appeared to be a neutral variant. These data suggest that it is unlikely than common variants in *ZBTB7A* or its promoters and proximal enhancers accounted for HbF variation in sickle cell anemia.

ZBTB7A affects HbF gene silencing through its binding in and about the *HBB* gene cluster. Accordingly, using the permissive motif ([GAC][ACG][GTAC][AC]CC[CAG][CTA]) as a target, we searched for variants in

these putative binding motifs in the interval downstream of *OR51V1* (5' olfactory receptor gene cluster) and upstream of *OR51B4* (3' olfactory gene cluster) flanking the *HBB*-like genes and its locus control region (LCR) in chromosome 11p15.5. Within the putative binding motifs, we found 8 motif-modifying SNPs. The binding domains identified by bioinformatics analysis overlapped some of the ZBTB7A binding occupancy data based on ChIP-seq signals that were previously reported (Masuda et al. 2016). **Figure 3** displays the position of these SNPs in putative binding motifs along with the experimentally defined ZBTB7A occupancy regions. **Table 3** provides additional details of these SNPs. Rs16912979 is present in hypersensitive site (HS)-4 of the LCR; and rs7119428 and rs9736333, which are in perfect LD, are located in HS-2. The alternative allele (C) of both HS-2 SNPs, was present only in African American sickle cell anemia samples of diverse African-origin haplotypes and are in adjacent potential ZBTB7A binding sites at coordinates 11:5302080-5302087 and 11:5302057-11:5302064. HS-2 and HS-4 also contain GATA1 and TAL1 binding sites that were shown to be involved in hemoglobin switching. The alternative C allele of rs7119428 that was found in HS-2 had a minor allele frequency of about 0.22 in the HapMap Yoruban sample and 0.17 in another African population reported in dbSNP; the major A allele was monomorphic in Europeans, Asians and Indians as reported in dbSNP. Rs7119428, which was not represented on the Human610-Quad array (~600,000 SNPs), was imputed with very high confidence in cohorts studied with this chip (Table 2). The C allele had a frequency of 0.56 in 822 cases of the African American CSSCD cohort, in which all *HBB* haplotypes except for the AI haplotype were represented. This SNP was included on the

Illumina Human Omni Express BeadChip (~700,000 SNPs). Its allele frequency was 0.88 in Saudis with the Benin haplotype who were studied using this chip. There was a significant association of this variant with HbF in the CSSCD cohort after adjusting for age and sex (beta coefficient = -0.05, p-value 0.02) but the effect was small. In 15 Saudi patients with the Benin haplotype who were heterozygous for the C allele, HbF was 13.7%; 85 C homozygotes had HbF of 10.2% however this difference was not significant (p-value 0.14). These preliminary results suggest that the C alleles of both rs7119428 and rs9736333 in HS-2, and also other SNPs in LD with these alleles might be play some role in the decreased HbF in sickle cell anemia of African descent compared with AI haplotype sickle cell anemia. Although rs7119428 was rare in Saudi AI sickle cell anemia we found 3 heterozygotes for the C allele and their HbF was about 10% compared with nearly 20% in homozygotes for the common A allele.

A putative ZBTB7A binding motif was found 5' to *HBG2* although this region was devoid of uniquely mapped binding sites as reported in (Masuda et al. 2016). This motif contained rs7482144 (G/A) in the positive strand or (C/T) in the negative strand, the well-studied Xmn1 restriction site that is associated with high HbF in the African Senegal and AI haplotypes (Nagel et al. 1985). Variants were also present in *HBG1* and *HBB*. The T allele of rs567305547 and rs537552941 in the small intron of *HBG1*, with an alternative allele frequency of about 0.40 in Africans, was found in 2 of 3 African Benin haplotype samples studied with WGS. All AI haplotype patients were monomorphic for the G allele, a result confirmed by Sanger sequencing in a subset of cases. These 2

intronic SNPs were not on the haplotype reference panel used for imputation and therefore they could not be imputed so any association with HbF could not be tested.

SNPs in putative ZBTB7A binding sites distinguish the high HbF AI haplotype from African origin haplotypes of sickle cell anemia where HbF is usually lower. They are present in sites with characteristics of active enhancers like transcription factor binding and epigenetic marks and are therefore candidates for the functional elements of this haplotype. Perhaps the variants of this haplotype alter looping of the LCR to globin gene promoters, however mechanistic studies are required to validate these genetic associations.

| Cohorts | **n** | **Age (y)** | **HbF (%)** |
|---|---|---|---|
| Saudi Benin Haplotype-Exome-Seq | 15 | 21.5±10.5 | 9.8±3.7 |
| Saudi Benin Haplotype-GWAS (700K) | 100 | 18.6±11.0 | 10.8±4.6 |
| African American Benin-WGS | 3 | 26.0±3.6 | 19.8±0.4 |
| African American Senegal- WGS | 1 | 5.9 | 16.0 |
| Saudi AI Haplotype-Exome-Seq | 8 | 32.3±11.1 | 12.7±5.5 |
| Saudi AI Haplotype cohort 1-GWAS(600K) | 42 | 26.4±11.1 | 17.6±5.018 |
| Saudi AI Haplotype cohort 2-GWAS(700K) | 62 | 23.9±9.6 | 18.8±7.5 |
| Saudi AI Haplotype-WGS | 7 | 25.9±6.8 | 23.5±2.6 |
| Saudi AI Haplotype-WGS | 7 | 34.1±10.3 | 8.2±1.3 |
| Indian AI haplotype-WGS | 3 | 22.7±5.5 | 26.0±4.5 |
| CSSCD-GWAS (600K) | 822 | 13.6±11.3 | 5.2±5.6 |

Table 2. Cohorts analyzed. Fourteen Saudi and 3 Indian AI haplotype HbS homozygotes, 3 African Americans with the Benin haplotype (selected because of their unusually high HbF) and 1 African American with the Senegal haplotype had whole genome sequencing (WGS). Saudi AI patients who had WGS all had the same minor alleles for *BCL11A* and *MYB*. The other cohorts included African Americans of diverse *HBB* haplotypes who were participants in the Cooperative Study of Sickle Cell Disease (CSSCD) and Saudi patients with the AI and Benin haplotype. The approximate number of variants represented on the 2 Illumina arrays used for genome-wide association studies (GWAS) is shown in parentheses.

| Site | Location (v37) | S | Ref-seq | Benin | Senegal | AI | RSID | REF allele | ALT allele |
|---|---|---|---|---|---|---|---|---|---|
| HS-4 | 11:5309693-5309700 | + | GG**T**C CCCA | GGCCC CCA | GGCC CCCA | GGTCC CCA | rs16912979 | T | C |
| HS-2 | 11:5302080-5302087 | - | **A**GGG GCCT | CGGG GCCT | AGGG GCCT | AGGG GCCT | rs7119428 | A | C |
| HS-2 | 11:5302057-5302064 | - | GGGG G**T**GG | GGGG GCGG | GGGG GTGG | GGGG GTGG | rs9736333 | T | C |
| *HBG2* 5' | 11:5276167-5276174 | + | GG**G**A CCGT | GGGA CCGT | GGAA CCGT | GGAA CCGT | rs7482144 | G | A |
| *HBG1* intron 1 | 11:5270900-5270907 | - | AG**GG** TCCT | AGTTT CCT | AGGG TCCT | AGGG TCCT | rs567305547 rs537552941 | G | T |
| *HBG1* intron 2 | 11:5269931-5269938 | + | **G**CCA CCAT | GCCAC CAT | CCCAC CAT | CCCAC CAT | rs2187608 | G | C |
| *HBB* intron 2 | 11:5247791-5247798 | + | **C**GTCC CAT | GGTCC CAT | GGTCC CAT | GGTCC CAT | rs10768683 | C | G |

Table 3. SNPs in putative ZBTB7A binding motifs in the *HBB* gene cluster (RSID) along with their genomic locations (v37), reference sequence (ref-seq and variation according to haplotype (Benin, Senegal, AI). Underlined and bolded are the variant alleles. S denotes positive (+) or negative (-) strand location of the putative binding motif.

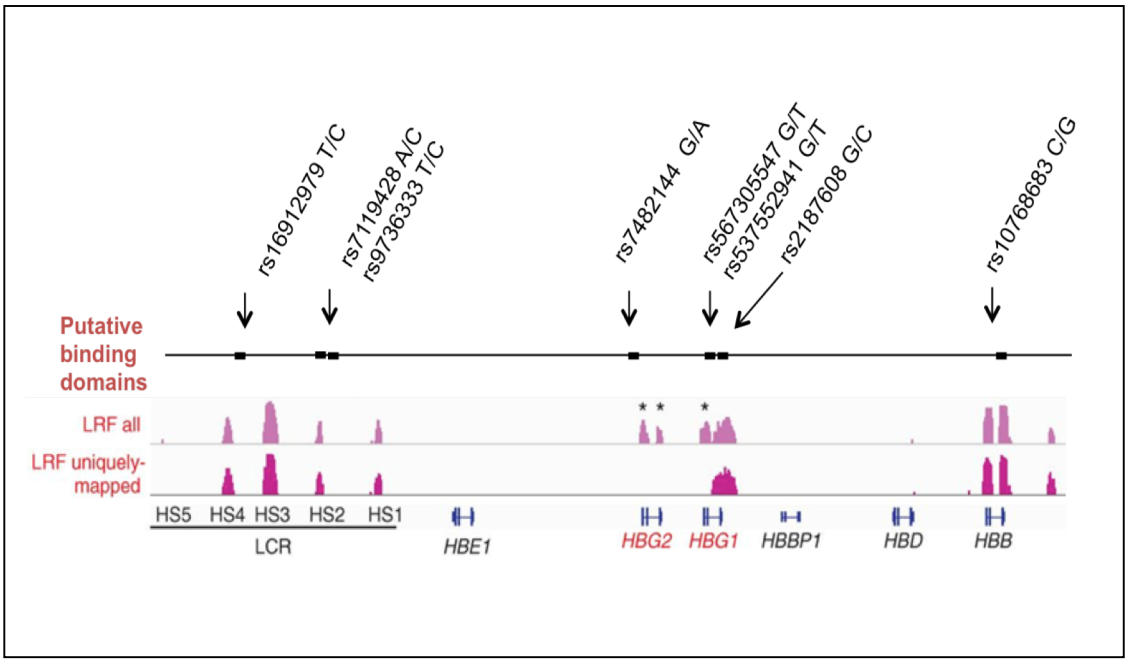



Figure 3. Approximate locations of the putative binding motifs for ZBTB7A (first track) that contained polymorphisms (shown above track) and were located between *OR51V1* in the 5' olfactory receptor gene cluster and *OR51B4* in the 3' olfactory gene cluster flanking the *HBB*-like genes and its locus control region (LCR) on chromosome 11p15.5. SNPs in the binding motifs are shown (arrows) with their major/alternative alleles. ChIP-seq data (tracks 2 and 3) is taken from (Masuda et al. 2016) along with the relative locations of globin genes and the LCR (track 4). Extensive homology between *HBG2* and *HBG1* makes mapping to these regions difficult and ChIP-seq data only showed uniquely-mapped binding in *HBG1* (track 3)

**Haplotypes of the HbS gene and HbF**

The first suggestion that the cis-acting elements of the *HBB* gene complex influenced HbF levels was the discovery that the HbS gene originated on five common *HBB* haplotypes. Although each haplotype was associated with a characteristic average level of HbF (Akinsheye et al. 2011; Bhagat et al. 2013; Ngo et al. 2013), among patients homozygous for any haplotype, HbF levels varied. The AI haplotype was prevalent in patients from the Arabian Peninsula and in India and might have originated in India and migrated to the Arabian Peninsula. It is associated with the highest HbF level of all haplotypes. After HbF levels stabilized at about age 10 years, adults with the AI haplotype had an average HbF level approaching 20%. Four other common *HBB* haplotypes, the Senegal, Benin, Cameroon and Bantu or Central African Republic haplotype originated in Africa. These were associated with mean HbF levels of 10%, 7%, 7%, and 5%, respectively; in these patients HbF levels become stable after five years of age (Chang et al. 1995; Green et al. 1993).

*HBB* haplotypes were first categorized by examination of a limited number of restriction endonuclease cleavage sites, most of which are SNPs. With one exception, little evidence existed that any of these polymorphisms were functionally important. The exception was the Xmn1 restriction site polymorphism (rs7482144), located in the promoter of *HBG2*, 158 bp upstream of the transcription start site. This SNP was exclusive to carriers of the AI and Senegal haplotypes. Resequencing part of the *HBB* gene cluster suggested that rs10128556, a SNP in linkage disequilibrium (LD) with rs7482144, was the functional

element of the Senegal haplotype (Galarneau et al. 2010). However, a mechanism whereby the locus of this SNP influenced HbF was unclear, and unlike rs7482144, this SNP was not present in a putative transcription factor binding site (**Figure 3**).

To fully understand the mechanism behind HbF regulation, developing a fast and accurate method for haplotype classification is crucial as well as studying the cis and transacting elements that regulate expression of HbF. This dissertation will cover the major projects:

1. Developing a fast and accurate method for classifying sickle cell *HBB* haplotypes.
2. Studying the genetic determinants of HbF in Saudi Arabian and African Benin haplotype sickle cell.
3. Investigating cis and trans-acting regulators of HbF expression to identify quantitative trait loci associated *with HBG* expression on a genome-wide scale.

Understanding the genetic bases of HbF expression is a key element in developing therapeutic agents. Achieving the goals of these projects will give the medical community a better understanding of HbF genetic modulators, and might lead to the identifications of potential drug targets or drug candidates.

# Chapter 2. A Phased SNP-based Classification of Sickle Cell Anemia *HBB* Haplotypes

## Background

Haplotypes of sickle cell anemia were first ascertained by analysis of restriction fragment length polymorphisms (RFLPs) in the *HBB* gene cluster (Sutton et al. 1989). Classification of patients' haplotype is useful for prognostic purposes and for studying the genetic differences that contribute to the HbF variability among these haplotypes. RFLP classification was based on detecting whether or not cleavage occurred at five to eight restriction sites when DNA was digested with restriction endonucleases, as shown in **Figure 4** (Antonarakis et al. 1982; Joly et al. 2011; Rezende et al. 2016). This method is time-consuming and can lead to error (Joly et al. 2011). Fluorescence resonance energy transfer coupled with high-resolution melting (HRM) assay is another method to classify sickle cell haplotypes, but it is also labor intensive and requires multiple laboratory assays (Joly et al. 2011). Neither method is capable of differentiating between parental and maternal alleles in an individual so that without informative genetic data from family members, the phasing of restriction patterns is not possible, and in many cases ascertainment of a haplotype is either equivocal or impossible. We used genome-wide association study (GWAS) data imputed to a reference panel to obtain a phased output. The phased GWAS data allowed assigning SNPs to parental chromosomes, which facilitated the classification procedure using fewer SNPs.

## Methods

### *Haplotype Classification*

GWAS data were available for patients with sickle cell anemia from the CSSCD (Solovieff et al. 2010). SNP array data containing 588451 markers were evaluated using PLINK to identify and remove SNPs with minor allele frequency (MAF) < 0.01, violated Hardy-Weinberg Equilibrium (HWE), and had more than 0.05 missing genotype information (Purcell et al. 2007). Genotypes for a total of 560170 SNPs were imputed using the Michigan Imputation Server (Das et al. 2016), the 1000 Genomes Phase 3 v5 reference panel, and the Eagle phasing algorithm to obtain phased output (Loh et al. 2016; The Genomes Project 2015). We developed a Python script based on VCF and PYSAM Python modules to read SNP information and assign the haplotype accordingly (Casbon 2017; Pysam-developers/pysam.). Code and an example are available on GitHub (https://github.com/eshaikho/haplotypeClassifier). We used this script to classify 1394 samples that were previously classified by RFLP in the CSSCD. We selected four SNPs (rs3834466, rs28440105, rs10128556, and rs968857) which define all of the haplotypes spanning the β-globin gene cluster (**Table 4**).

### *Calculation of HbF Average per Haplotype*

To check the consistency of classification and average HbF for each haplotype, we used samples with available HbF level information including 559 of the 813 samples that were successfully classified with RFLPs, 916 of the samples classified with the SNP-based method, and 252 samples that were either partially classified or failed classification with

RFLPs. We calculated the average HbF level for each haplotype using psych R package (Revelle 2017), and generated a boxplot for the most common haplotypes (Benin homozygotes [BEN/BEN], Benin/Central African Republic compound heterozygotes [BEN/CAR], Benin/Senegal compound heterozygotes [BEN/SEN], Benin/Cameroon compound heterozygotes [BEN/CAM], Central African Republic homozygotes [CAR/CAR]) in this cohort to show the consistency of HbF levels across the three groups (five RFLP classification, SNP-based classification, and the group that failed five RFLP classification but were able to be classified with the SNP-based method).

***Classification of haplotypes in Saudi sickle cell anemia patients and in a library of sickle cell anemia induced pluripotent stem cells (iPSCs)***

Since CSSCD patients are mostly African American, we tested our method using data obtained from sickle cell anemia patients from the Eastern and Southwestern Provinces of Saudi Arabia. Eastern Province patients tend to have the autochthonous AI haplotype as the major haplotype, while Southwestern Province patients mostly have the BEN haplotype that was introduced from Africa. The HbF levels in Saudi Benin patients is twice as high as African American patients with this haplotype (Akinsheye et al. 2011; Alsultan et al. 2011). To further test our method on a mixed population of diverse ethnicity we reclassified haplotypes originally ascertained using RFLPs in a library of sickle cell anemia-derived iPSCs (Park et al. 2017).

## Results

### *Haplotype Classification*

Of 371 CSSCD patients classified as BEN/BEN using five RFLPs, we achieved a concordance of 98% (367/371) using four phased SNPs. We achieved >99% concordance for patients classified as BEN/CAR using RFLPs. For BEN/SEN, BEN/CAM, CAR/CAR, CAR/SEN, CAR/CAM, CAR/AI, SEN/SEN, SEN/CAM, SEN/AI, and CAM/CAM haplotypes our concordance with the RFLP method was 100% although the numbers of patients in each category was smaller. Two patients classified originally as BEN/AI failed reclassification (**Table 5**). Discordance between our method and the five RFLP method occurred in only eight of 813 patients providing an overall concordance rate >99%. Two patients classified as BEN/AI with RFLP were reclassified as UNKNOWN/SEN. Four patients classified as BEN/BEN were reclassified as UNKNOWN/BEN, CAR/CAR, CAM/BEN, and SEN/BEN. The last two patients were classified as CAR/SEN and SEN/BEN with our methods instead of BEN/CAR according to RFLP analysis. Importantly, we were able to assign a haplotype to 86% (343/ 395) of samples CSSCD that failed classification using RFLPs.

### *Calculation of HbF Average per Haplotype*

The average haplotype HbF level of patients classified with our method is consistent with average HbF in haplotypes reported in literature based on RFLPs (**Table 6**) (Perrine et al. 1972; Powars 1991). The average haplotype HbF for samples unclassifiable using RFLPs, but classified using phased SNP data matched the average HbF for each known

haplotype (**Table 6**). Boxplots of the most common five haplotypes in the CSSCD cohort show the consistency of HbF levels across the three classification groups (**Figure 5)**.

***Classification of Haplotype in Saudi Sickle Cell Anemia and Sickle iPSCs***

Haplotypes among 55 Southwestern Province patients classified using the RFLP method included 39 BEN/BEN, 11 CAR/CAR, 2 BEN/SEN, 1 BEN/CAR, 1 SEN/SEN and one unknown. The distribution of haplotypes for these subjects derived using the SNP method was 48 BEN/BEN, 3 BEN/UNKNOWN, 2 BEN/CAM, and one AI/AI (**Table 6**). The concordance between RFLP and SNP-based classification was 67%. For the 30 Eastern Province patients, we had 100% concordance since all patients reclassified as AI/AI (**Table 6**).

In a library of sickle cell anemia iPSCs there was high concordance between the two methods of haplotype ascertainment. The only discordance was in two patients classified originally as BEN/BEN that according to SNP-based reclassification were CAM/BEN and CAR/SEN (**Table 7**). Importantly, we were able to assign a haplotype to 15 of 17 iPSC samples that were classified as either atypical or were indeterminate using RFLPs.

**Discussion**

In adults, homozygotes for the BEN, CAR, and CAM haplotypes were associated with HbF of 5-7% of total hemoglobin; SEN and AI haplotypes had HbF levels of about 10% and 20%, respectively. Using GWAS data we were able to classify with high accuracy and time efficiency the haplotype of sickle cell anemia patients using four SNPs. The

primary feature of our classification method is a phasing step after genotype imputation where SNP alleles are assigned to parental chromosomes, and the haplotype of each chromosome is assigned independently. This method was superior to ascertaining haplotype by RFLP using unphased SNPs at five sites and was successfully applied in a few seconds on a personal computer.

Haplotyping errors can occur using the SNP-based method because of the SNP genotyping platform, imputation errors, and ambiguities arising from phasing algorithms. Nevertheless, in African-origin patient samples, we were able to achieve a concordance of 99% percent (805/813) between 4-SNP haplotypes derived from a phasing algorithm using GWAS data with 5-SNP haplotypes determined using restriction analysis. Haplotype assignment in a sickle cell anemia iPSC library also showed high concordance and demonstrated the efficiency of SNP-based method to classify samples that failed RFLP classification. The discordance between SNP-based and RFLP ascertainment most likely resulted from errors in the RFLP classification that is sensitive to the presence of other SNPs in the restriction sites and the vagaries of restriction enzyme analysis and Southern blotting that was used for haplotype analysis in the CSSCD.

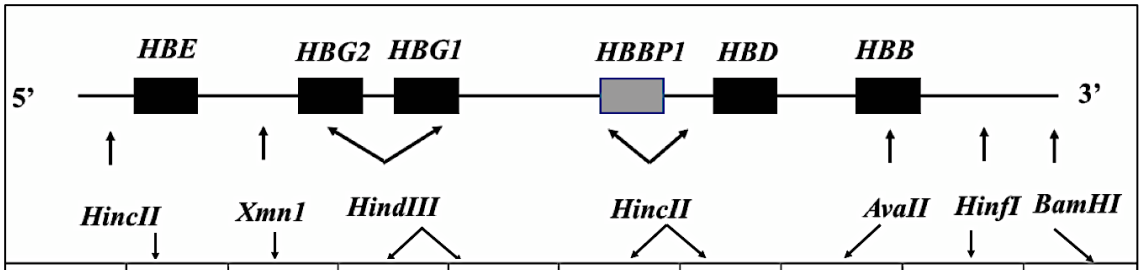
In 30 Saudi East patients where the AI haplotype was ascertained by genotyping rs7482144 (Xmn1 5′ to *HBG2*), rs3834466 (Hinc2 5′ to *HBE1*), and rs549964658 (5′ to *HBD*) we had 100% concordance. The major discordance between RFLP and SNP-based analysis for classification of Saudi Southwestern patients occurred among 11 subjects classified as CAR/CAR. Eight of these 11 were reclassified as BEN/BEN and three as

BEN heterozygotes. The only difference between CAR and BEN haplotypes is the SNP rs968857 at the *HincII* site 5' of *HBD* (**Figure 4, Table 4**). It is most likely that this discrepancy was a result of an error in RFLP analysis. If the discordance was due to imputation quality, the error rate would probably match the imputation error. There is 100% discordance at this *HincII* site while the imputation quality score of rs968857 is $R^2$=0.99. One patient with HbF of 20.4% originally as SEN/SEN by RFLP was reclassified as AI/AI.

To investigate the discordance in Southwestern Province patients, we examined the genotype data of the *HBB* gene cluster downstream of *OR51V1* (5' olfactory receptor gene cluster) and upstream of *OR51B4* (3' olfactory gene cluster) in both patient groups. The SNP genotypes of the 11 CAR/CAR that we reclassified as BEN/BEN or BEN heterozygous had the same SNP genotype of BEN/BEN patients that were classified as such with both methods. The genotype data, average HbF, and the imputation quality score of rs968857 suggest that the high discordance in Southwestern Saudi patients is due to RFLP errors.

A limitation of our method is the dependency on the availability of GWAS data for many SNPs in the β-globin gene region. However, many large patient cohorts have been genotyped using genome-wide SNP arrays and the cost of a gene array is less than the cost of genotyping by RFLP. In these patients, haplotype information might be useful as a covariate in a genetic risk analysis. RFLP analysis might be suitable for a small

number of patients but requires optimization of all of the individual assays. The main advantage of our haplotype determination method is the rapid classification and high accuracy. This method can also be used for whole genome sequence data classification after SNP calling and phasing. Moreover, it is not sensitive to SNPs that alter the restriction enzyme recognition sequence that can lead to error using RFLPs




| RSID | rs3834466 | rs7482144 | rs2070972 | rs28440105 | rs10128556 | rs968857 | rs10768683 | rs10837631 | No RSID |
|---|---|---|---|---|---|---|---|---|---|
| *Gene* | *ε (HBE1)* | *Gγ (HBG2)* | *Gγ (HBG2)* | *Aγ (HBG1)* | *ψβ (HBBP1)* | *5' δ (HBD)* | *β (HBB)* | *3'β (HBB)* | *3' β (HBB)* |
| Restriction Enzyme | *HincII* | *Xmn1* | *HindIII* | *HindIII* | *HincII* | *HincII* | *AvaII* | *HinfI* | *BamHI* |
| SEN | - | + | + | - | + | + | + | + | + |
| CAR | - | - | + | - | - | - | + | + | + |
| CAM | - | - | + | + | - | + | + | - | - |
| BEN | - | - | - | - | - | + | + | + | + |
| AI | + | + | + | - | + | + | + | - | - |

Figure 4. Restriction Enzyme Recognition Sites in the β-Globin Gene Cluster. RSIDs of SNPs present in restriction endonuclease sites in the β-globin-like gene cluster. (+) denotes the presence of the corresponding enzyme site while (-) denotes the absence in the five-major sickle cell haplotype (Benin (BEN), Central African Republic (CAR), Senegal (SEN), Cameroon (CAM) and Arab–Indian (AI).

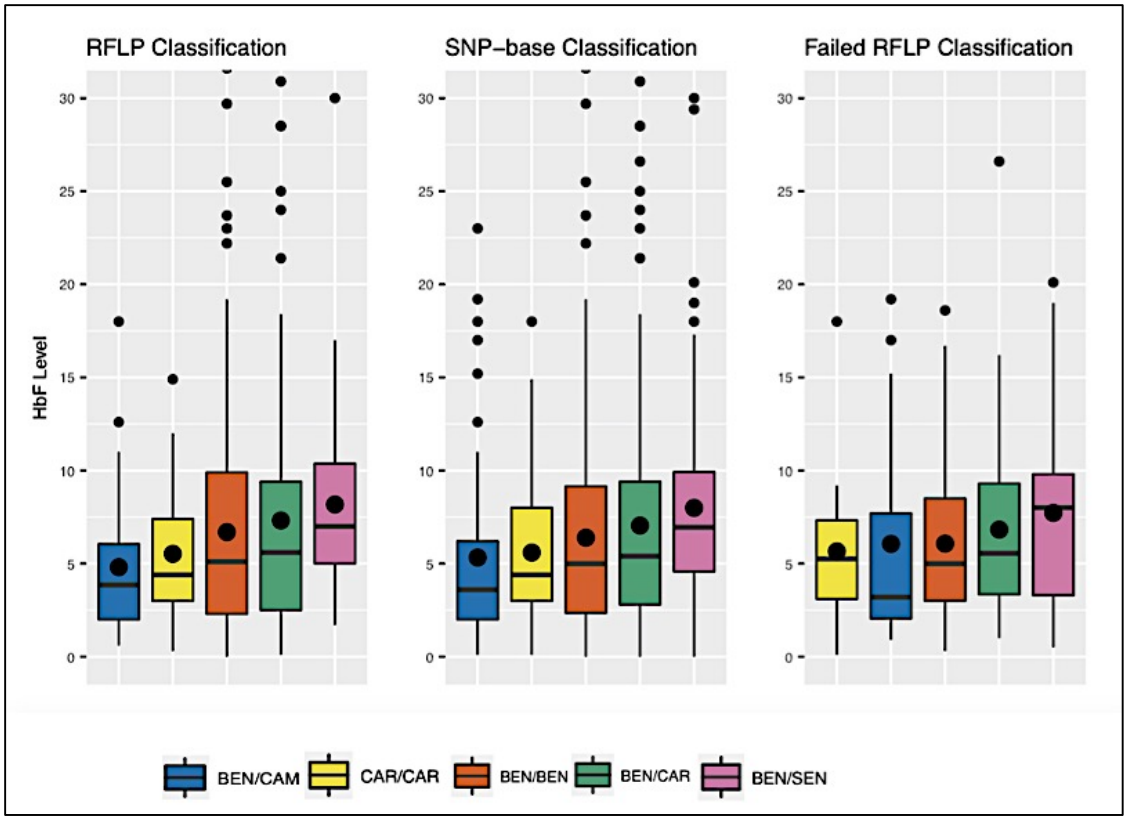



Figure 5. Boxplots of the Common Haplotypes in the CSSCD. Shown are the HbF levels according to haplotype defined by RFLPs and SNP-based methodology. The third panel shows data from patients not classified by RFLP but successfully classified using the SNP-based method. The black dots in the middle of the boxplots represent mean HbF level, while the black horizontal lines represent the median HbF level.

| Haplotype | rs3834466 | rs28440105 | rs10128556 | rs968857 |
|---|---|---|---|---|
| AI | GT | C | T | T |
| SEN | G | C | T | T |
| BEN | G | C | C | T |
| CAR | G | C | C | C |
| CAM | G | A | C | T |

Table 4. Five Major Haplotypes and Alleles of the four SNPs defining the haplotype.

| Haplotype | 5 RFLPs | 4 SNPs | Concordance |
|---|---|---|---|
| BEN/BEN | 371 | 367 | 0.989218329 |
| BEN/CAR | 226 | 224 | 0.991150442 |
| BEN/SEN | 91 | 91 | 1 |
| BEN/CAM | 41 | 41 | 1 |
| CAR/CAR | 31 | 31 | 1 |
| CAR/SEN | 17 | 17 | 1 |
| CAR/CAM | 14 | 14 | 1 |
| SEN/CAM | 9 | 9 | 1 |
| SEN/SEN | 8 | 8 | 1 |
| BEN/AI | 2 | 0 | 0 |
| CAR/AI | 1 | 1 | 1 |
| SEN/AI | 1 | 1 | 1 |
| CAM/CAM | 1 | 1 | 1 |

Table 5. Comparison of Haplotype Classification Methods for 813 Sickle Cell Anemia Patients from the CSSCD. 5 RFLP represents haplotypes previously assigned by RFLPs using information from 5 restriction sites based on Southern blotting. 4 SNPs represents the number of haplotypes assigned using the phased SNP-based classification.

| *CSSCD | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *RFLP classification* | | | *SNP-based classification* | | | *Failed RFLP classification* | | |
| Haplotype | n | mean | sd | n | mean | sd | n | mean | sd |
| BEN/BEN | 253 | 6.69 | 5.58 | 379 | 6.39 | 5.14 | 89 | 6.07 | 4.32 |
| BEN/CAR | 157 | 7.32 | 7.63 | 261 | 7.04 | 6.77 | 74 | 6.83 | 5.37 |
| BEN/SEN | 62 | 8.18 | 4.75 | 116 | 7.99 | 5.32 | 33 | 7.72 | 5.50 |
| BEN/CAM | 32 | 4.81 | 3.89 | 53 | 5.33 | 5.19 | 15 | 6.06 | 6.18 |
| CAR/CAR | 21 | 5.52 | 3.71 | 41 | 5.59 | 3.90 | 16 | 5.66 | 4.47 |
| CAR/SEN | 11 | 9.80 | 3.48 | 24 | 8.91 | 7.59 | 7 | 11.24 | 13.14 |
| CAR/CAM | 9 | 3.80 | 2.51 | 18 | 3.96 | 3.24 | 8 | 4.41 | 4.16 |
| SEN/SEN | 7 | 9.24 | 5.30 | 13 | 8.74 | 4.35 | 5 | 7.58 | 3.35 |
| CAM/SEN | 3 | 7.43 | 6.23 | 9 | 8.61 | 6.02 | 5 | 9.24 | 7.17 |
| BEN/AI | 2 | 4.20 | 5.37 | - | - | - | - | - | - |
| CAM/CAM | 1 | 7.00 | - | 1 | 7.00 | - | - | - | - |
| CAR/AI | 1 | 16.10 | - | 1 | 16.10 | - | - | - | - |
| **SW Saudi | | | | | | | | | |
| | *RFLP classification* | | | *SNP-based classification* | | | *Failed RFLP classification* | | |
| Haplotype | n | mean | sd | n | mean | sd | n | mean | sd |
| BEN/BEN | 39 | 11.22 | 5.32 | 48 | 10.28 | 4.19 | - | - | - |
| CAR/CAR | 11 | 9.35 | 5.07 | - | - | - | - | - | - |
| BEN/SEN | 2 | 8.65 | 0.64 | - | - | - | - | - | - |
| BEN/CAR | 1 | 6.50 | - | 1 | 3.10 | - | - | - | - |
| SEN/SEN | 1 | 20.40 | - | - | - | - | - | - | - |
| UNK | 1 | 12.60 | - | - | - | - | - | - | - |
| AI/AI | - | - | - | 1 | 20.40 | - | - | - | - |
| BEN/CAM | - | - | - | 2 | 10.15 | 9.69 | - | - | - |
| BEN/UNK | - | - | - | 3 | 20.20 | 8.64 | - | - | - |
| ***E Saudi | | | | | | | | | |
| | *RFLP classification* | | | *SNP-based classification* | | | *Failed RFLP classification* | | |
| Haplotype | n | mean | sd | n | mean | sd | n | mean | sd |
| AI/AI | 30 | 18.03 | 5.39 | 30 | 18.03 | 5.39 | - | - | - |

Table 6. Mean HbF Levels Among Haplotypes. Included are 559 patients with HbF phenotypes classified with RFLP method, 916 patients with HbF phenotypes classified with SNP-based method, and 252 samples that failed RFLP classification but were classifiable with SNP-based method. (n) is the number of patients in the corresponding haplotype class, (mean) is the mean of HbF per haplotype, (sd) is the standard deviation. * CSSCD are African Americans; ** SW Saudi is Saudi patients from Southwestern Province; ***E Saudi are Saudi patients from the Eastern Province. UNK-unknown

| Name of Line | Gender | Nationality of Origin | Age | Haplotype | 4SNPs |
|---|---|---|---|---|---|
| SA108 | male | Saudi Arabia | 9 | AI/AI | AI/AI |
| SA50-1 | female | Saudi Arabia | NA | AI/AI | AI/AI |
| SA106-1 | female | Saudi Arabia | NA | AI/AI | AI/AI |
| SA170-1 | male | Saudi Arabia | 3 | AI/AI | AI/AI |
| SS2-1 | female | US | 32 | UNK/UNK | BEN/SEN |
| SS2-1GAG (CRISPR corrected) | female | US | 32 | UNK/UNK | BEN/SEN |
| SS12-1 | female | US | 27 | UNK/UNK | BEN/CAR |
| SS18-1 | female | US | 23 | UNK/UNK | UNK/UNK |
| SS28-1 | female | US | 25 | UNK/UNK | BEN/BEN |
| SS36 | male | US | 38 | UNK/UNK | BEN/CAR |
| SS41-1 | male | US | 21 | UNK/UNK | CAR/CAR |
| SS45-1 | female | US | 37 | UNK/UNK | BEN/BEN |
| SS47-1 | female | US | 42 | UNK/UNK | BEN/CAR |
| SS48-1 | male | US | 30 | UNK/UNK | CAR/CAR |
| SA5-1 | female | Saudi Arabia | 9 | UNK/UNK | BEN/BEN |
| SA53-1 | male | Saudi Arabia | 14 | UNK/UNK | BEN/CAR |
| SA208 | male | Saudi Arabia | 7 | UNK/UNK | UNK/BEN |
| SA138-1 | male | Saudi Arabia | 16 | UNK/UNK | AI/AI |
| BR-SP-21-1 | female | Brazil | 20 | UNK/UNK | BEN/CAR |
| BR-SP-37-1 | female | Brazil | 20 | UNK/UNK | CAR/CAM |
| BR-SP-45-1 | female | Brazil | 20 | UNK/UNK | CAR/CAR |
| SS24-1 | male | US | 24 | CAR/CAR | CAR/CAR |
| SS25-1 | female | US | 22 | CAR/CAR | CAR/CAR |
| BR-SP-3-1 | female | Brazil | 34 | CAR/CAR | CAR/CAR |
| BR-SP-23-1 | female | Brazil | 23 | CAR/CAR | CAR/CAR |
| BR-SP-25-1 | male | Brazil | 34 | CAR/CAR | CAR/CAR |
| BR-SP-41-1 | male | Brazil | 22 | CAR/CAR | CAR/CAR |
| BR-SP-43-1 | male | Brazil | 21 | CAR/CAR | CAR/CAR |
| SS9-1 | female | US | 29 | BEN/CAR | BEN/CAR |
| SS13-1 | female | US | 25 | BEN/CAR | BEN/CAR |

| | | | | | |
|---|---|---|---|---|---|
| SS15-1 | female | US | 28 | BEN/CAR | BEN/CAR |
| SS35 | male | US | 50 | BEN/CAR | BEN/CAR |
| BR-SP-29-1 | male | Brazil | 20 | BEN/CAR | BEN/CAR |
| BR-SP-33-1 | female | Brazil | 53 | BEN/CAR | BEN/CAR |
| BR-SP-39-1 | male | Brazil | 22 | BEN/CAR | BEN/CAR |
| SS5-1 | male | US | 32 | BEN/BEN | BEN/BEN |
| SS14-1 | female | US | 39 | BEN/BEN | BEN/BEN |
| SS16-1 | female | US | 36 | BEN/BEN | BEN/BEN |
| SS19-1 | male | US | 30 | BEN/BEN | BEN/BEN |
| SS29-1 | female | US | 32 | BEN/BEN | BEN/BEN |
| SS32 | female | US | 33 | BEN/BEN | BEN/BEN |
| SS37 | female | US | 37 | BEN/BEN | BEN/BEN |
| SS38 | male | US | 26 | BEN/BEN | BEN/BEN |
| SS44-1 | female | US | 23 | BEN/BEN | BEN/CAM |
| SS49-1 | male | US | 31 | BEN/BEN | CAR/SEN |
| SA36 | female | Saudi Arabia | 26 | BEN/BEN | BEN/BEN |
| SA40-1 | male | Saudi Arabia | 20 | BEN/BEN | BEN/BEN |
| SA64 | male | Saudi Arabia | 14 | BEN/BEN | BEN/BEN |
| SA82-2 | male | Saudi Arabia | 24 | BEN/BEN | BEN/BEN |
| SA209-1 | male | Saudi Arabia | 12 | BEN/BEN | BEN/BEN |
| SA210-1 | male | Saudi Arabia | 9 | BEN/BEN | BEN/BEN |
| BR-SP-31-1 | male | Brazil | 35 | BEN/BEN | BEN/BEN |
| SS4-1 | male | US | 30 | BEN/SEN | BEN/SEN |
| SS8-2 | female | US | 31 | SEN/SEN | SEN/SEN |
| SS43-2 | female | US | 32 | SEN/SEN | SEN/SEN |

Table 7. Sickle Anemia-Specific iPSC Library. Haplotype denotes the haplotype classification and 4 SNPs is SNP-based reclassification. The table is modified from (Park et al. 2017).

# Chapter 3. Genetic Determinants of HbF in Saudi Arabian and African Benin Haplotype Sickle Cell Anemia

## Background

Each of the five major haplotypes of the *HBB* gene cluster (Benin, Car, Cameroon, Senegal, and AI) in sickle cell anemia are associated with different HbF levels. Although HbF is the major modulator of disease severity, the genetic elements that underlie the association of HbF and *HBB* haplotypes are not fully understood (Steinberg et al. 2009).

Saudi sickle cell patients from the Southwestern Province, whose *HBB* gene cluster is of African origin, have HbF levels of about 10%; African origin patients with the Benin haplotype have HbF levels of about 6%. Saudi patients have a genetic population structure similar to other Arabs, which does not resemble African-origin patients (Alsultan et al. 2011). We hypothesized that while Saudi and African American Benin haplotype homozygotes have similar *HBB* clusters, there might be common variants in the Saudi Benin patients that are associated with their increased HbF; conversely, African American Benin patients might have common variants that are associated with reduced HbF relative to Saudi patients.

**Methods**

To study the genetic differences between Saudi and African American Benin haplotype patients that might be associated with HbF, we imputed genome-wide association study (GWAS) data from the CSSCD and patients from the Southwestern Province of Saudi Arabia to 1000 Genomes Phase 3 v5 reference panel. Pre-imputation quality control was performed using PLINK (Purcell et al. 2007), and imputation was carried out through the Michigan Imputation server and phased imputed data was obtained using Eagle (Loh et al. 2016). We classified haplotypes using haplotypeClassifier available from https://github.com/eshaikho/haplotypeClassifier. Only homozygous Benin haplotype cases were selected for downstream analysis. We then removed related samples and outliers (based on their genetic markers) in the first 20 principal components (PCs) using King and EIGENSOFT, respectively (Manichaikul et al. 2010; Patterson et al. 2006). There were 293 African American patients not taking hydroxyurea, 153 males, and 140 females, aged between 2 and 68 years with an average HbF of 6.38%, and 63 Saudi Benin haplotype patients 27 of them are taking hydroxyurea, 36 males and 27 females, aged between 4 and 43 years, with an average HbF of 10.38%. A linear model adjusted for age and sex to predict the effect of PCs HbF level showed insignificance of the first 20 components indicating absence of population substructure within each population; the first five PCs were used in the final GWAS analysis to account for any potential bias due to these components. $Log_{10}$ HbF levels were employed as a quantitative phenotype to find the most significant SNPs in each cohort. Efficient and Parallelizable Association Container Toolbox (EPACTS; https://github.com/statgen/EPACTS) adjusted for sex and

the first five PCs were used for the final analysis. To avoid false associations due to small sample size we only considered SNPs with minor allele frequency (MAF) greater than 0.05. Including age as covariate does not improve the goodness of the model fit, thus it was excluded from the final model. Both cohorts were analyzed separately, and subjected to the same analytical methods except for hydroxyurea adjustment in Saudi Benin patients.

**Results and Discussion**

In Saudi Benin cases, there were no associations with HbF meeting GWAS significance levels; however, the small sample size reduced the statistical power of the study to detect an association (**Figure 6** & **Table 8**). In African American Benin haplotype patients, only rs1427407 in *BCL11A* met GWAS significance levels for association with HbF (**Figure 7**). Six intronic SNPs in *BCL11A* had marginal genome-wide significance; 3 intronic SNPS in *LARGE1*, *NEDD9* and *PAK2* also showed marginal GWAS significance with p-values between 8.97E-07 and 2.61E-07 (**Table 9**). The allele frequencies for the top 10 associated SNPs in African American patients were similar to that in Saudi patients except for rs6706648 and rs7606173 where the MAFs were 0.4 and 0.44 in African American and 0.11 and 0.25 in Saudi cases. To examine the effect of rs1427407, rs6706648, and rs7606173 we examined the distribution of HbF levels by the genotypes of these SNPs. We took advantage of the phased imputed data to examine the three SNPs haplotype effect on HbF. Homozygosity for a TCG haplotype of rs1427407, rs6706648 and rs7606173, respectively, was associated with 10 % HbF in African American

patients. This haplotype was found in 29% of Saudi and 24% of African American Benin haplotype patients. Homozygosity for the T allele rs1427407 was always associated with homozygosity for the C allele of rs6706648 and G allele of rs7606173. Homozygosity for a GTC haplotype of these same three SNPs was associated with 4.5% HbF in African American Benin and had a frequency of 0.40 in African American Benin, and 0.11 in Saudi Benin patients. Even when accounting for the population frequency differences of the TCG and GTC haplotypes, *BCL11A* variants do not explain the difference in HbF level seen between Saudi Benin and African America Benin sickle cell anemia.

A 3-base pair deletion in the *HBS1L-MYB* intergenic region is likely to be the functional element accounting for increased HbF associated with this QTL and is in high LD with rs9399137 (Farrell et al. 2011). The MAF of this SNP was 0.037 and 0.047 in Saudi Benin and African American Benin patients respectively suggesting that the 3-bp deletion is unlikely to explain the HbF difference between Saudi Benin and African American Benin patients.

The difference in HbF between Saudi Benin and African Americans Benin may due to one or more variants in Saudi Benin patients that could not be detected due to the small number of cases available for study. Whole genome sequencing might help identify genetic differences between these populations that account for their disparate HbF levels.

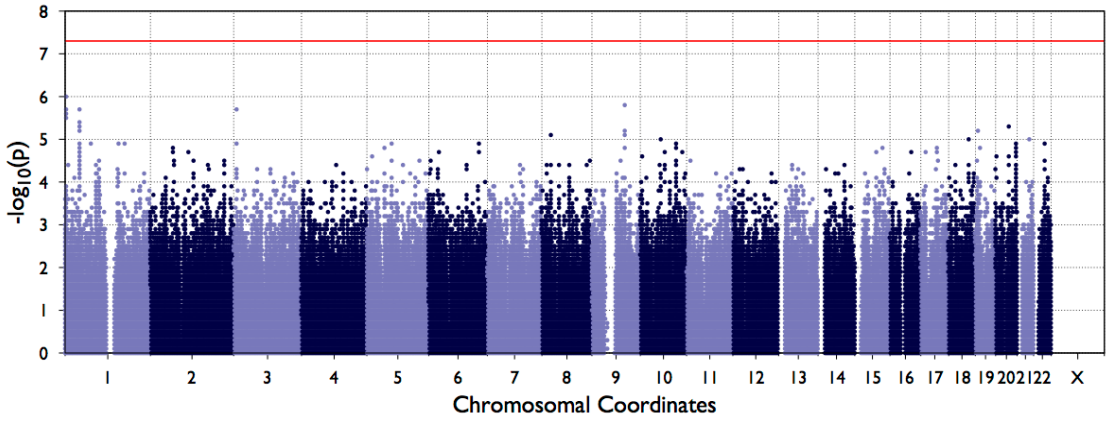



Figure 6. Manhattan plot for Saudi Benin haplotype and HbF. There is no signal at chromosome 2 which corresponds to *BCL11A* SNPs.

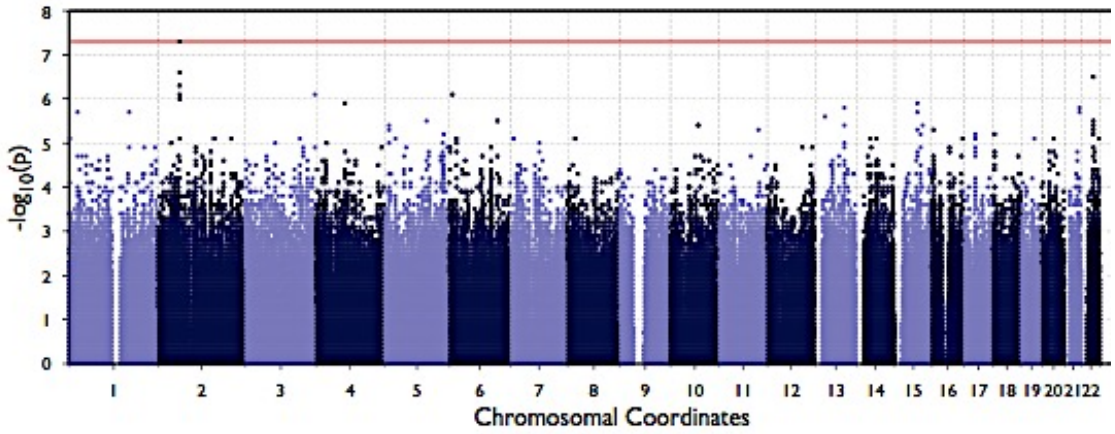



Figure 7. Manhattan plot for African American Benin haplotype and HbF. There is a clear signal at chromosome 2 which corresponds to *BCL11A* SNPs.

| CHROM | BEGIN | END | MARKER_ID | NS | MAF | PVALUE | BETA |
|---|---|---|---|---|---|---|---|
| **1** | 5504553 | 5504553 | 1:5504553_G/A_Intergenic | 63 | 0.2381 | 9.74E-07 | -0.19241 |
| **9** | 97412837 | 97412837 | 9:97412837_C/T_Intergenic | 63 | 0.16667 | 1.47E-06 | 0.22112 |
| **3** | 9067668 | 9067668 | 3:9067668_C/T_Intron:SRGAP3 | 63 | 0.14286 | 1.84E-06 | 0.24474 |
| **1** | 5489136 | 5489136 | 1:5489136_C/T_Intergenic | 63 | 0.34921 | 1.88E-06 | -0.17445 |
| **1** | 44630318 | 44630318 | 1:44630318_C/T_Intergenic | 63 | 0.29365 | 2.18E-06 | -0.17355 |
| **1** | 5488947 | 5488947 | 1:5488947_GA/G_Intergenic | 63 | 0.36508 | 2.46E-06 | -0.15751 |
| **1** | 5516961 | 5516961 | 1:5516961_T/A_Intergenic | 63 | 0.34921 | 3.25E-06 | 0.15378 |
| **1** | 44633681 | 44633681 | 1:44633681_C/T_Intergenic | 63 | 0.28571 | 4.26E-06 | -0.17587 |
| **1** | 44745624 | 44745624 | 1:44745624_C/T_Intron:ERI3 | 63 | 0.31746 | 5.48E-06 | -0.16997 |
| **1** | 44751206 | 44751206 | 1:44751206_A/G_Intron:ERI3 | 63 | 0.31746 | 5.48E-06 | -0.16997 |

Table 8. Top associations with HbF in Saudi patients homozygous for the Benin haplotype with MAF = [0.05,0.5]. NS denotes number of samples and BETA represents the effect of a SNP.

| CHR | BEGIN | END | RSID | MARKER_ID | NS | MAF | PVALUE | BETA |
|---|---|---|---|---|---|---|---|---|
| 2 | 60718043 | 60718043 | rs1427407 | 2:60718043_T/G_Intron:BCL11A | 293 | 0.25 | 5.44E-08 | -0.20 |
| 2 | 60722040 | 60722040 | rs6706648 | 2:60722040_C/T_Intron:BCL11A | 293 | 0.40 | 2.61E-07 | -0.16 |
| 22 | 33862330 | 33862330 | rs557939075 | 22:33862330_G/GT_Insertion:LARGE | 293 | 0.12 | 3.05E-07 | -0.25 |
| 2 | 60725451 | 60725451 | rs7606173 | 2:60725451_G/C_Intron:AC009970.1\|BCL11A | 293 | 0.44 | 5.13E-07 | -0.16 |
| 2 | 60724086 | 60724086 | rs1896295 | 2:60724086_T/C_Intron:AC009970.1\|BCL11A | 293 | 0.27 | 7.52E-07 | -0.18 |
| 2 | 60724087 | 60724087 | rs1896296 | 2:60724087_G/T_Intron:AC009970.1\|BCL11A | 293 | 0.27 | 7.52E-07 | -0.18 |
| 6 | 11287332 | 11287332 | rs4713339 | 6:11287332_C/T_Intron:NEDD9 | 293 | 0.23 | 7.93E-07 | -0.18 |
| 3 | 196544117 | 196544117 | rs13080125 | 3:196544117_T/C_Intron:PAK2 | 293 | 0.35 | 8.29E-07 | 0.16 |
| 2 | 60719970 | 60719970 | rs766432 | 2:60719970_C/A_Intron:BCL11A | 293 | 0.27 | 8.97E-07 | -0.18 |
| 2 | 60720951 | 60720951 | rs4671393 | 2:60720951_A/G_Intron:BCL11A | 293 | 0.27 | 8.97E-07 | -0.18 |

Table 9. Top associations with HbF in African Americans homozygous for the Benin haplotype with MAF = [0.05,0.5]. NS denotes number of samples and BETA represents the effect of a SNP.

# Chapter 4. Investigating Cis- and Trans-Acting Regulators of HbF Expression Using GTEx

## Background

GWAS and genetics studies have detected many SNPs in chromosomes 2, 6, and 11 that were associated with HbF levels in normal individuals and in patients with sickle cell anemia and β thalassemia (Bhanushali et al. 2015; Craig et al. 1996; Garner et al. 1998; Mtatiro et al. 2014; Solovieff et al. 2010; Thein et al. 1987; Uda et al. 2008). These studies used either HbF protein levels (expressed as a percentage of total hemoglobin levels) or the percentage of all erythrocytes that were F-cells (erythrocytes containing sufficient HbF to be detected by immunofluorescence) as a quantitative trait. HbF is a tetramer composed of two α and two γ polypeptide subunits. The γ-globin subunits, which characterize HbF, are encoded by two closely linked genes, from 5' to 3', *HBG2* and *HBG1*. The respective polypeptides of the γ-globin genes differ by only a single amino acid in position 136; glycine the $^{G}\gamma$ chain (HBG2) and alanine in the $^{A}\gamma$ chain (HBG1) (Schroeder et al. 1968). In adults, these genes are expressed at a ratio of 2:3 (Terasawa et al. 1980). The differential expression of these γ-globin genes might provide an understanding of whether the expression quantitative trait loci (eQTL) associated with *HBG* expression affect the expression of one or both genes and therefore, provide additional insight into their mechanisms of action. eQTLs are genomic loci that lead to variability in mRNAs expression levels. Analyzing *HBG2* and *HBG1*expression

separately should allow the detection of eQTLs that affect the expression of a single or both γ-globin genes.

RNA sequence, whole genome sequences, or genotypes from erythroid progenitor cells expressing *HBG* would provide the ideal data set to study cis and trans-acting elements that regulate HbF expression. However, such a data set is not publicly available. The Genotype-Tissue Expression (GTEx) -project provides an alternative resource to study HbF gene expression, the regulatory elements modulating *HBG* (both γ-globin genes) expression, and any effects of genetic variation in these regulatory elements on *HBG* expression (GTEx_Consortium 2013).The GTEx project was launched in 2010 by the National Institute of Health (NIH) to create a publicly available database and tissue bank. The GTEx samples are either from deceased organ/tissue or surgical donors. GTEx contains RNA sequencing data and genome-wide SNP analysis from multiple human tissues. These tissues include peripheral blood that contains reticulocytes that express hemoglobin genes and synthesize about 10% of total hemoglobin. The primary limitation of using GTEx data is that the blood samples obtained were likely to have reticulocyte counts of <1% of total red cells. However, reticulocytes are the only cells in the blood that express hemoglobin genes. eQTLs can be identified by analyzing global RNA expression using the expression levels of genes as quantitative phenotypes.

We performed a genome-wide association study (GWAS) to detect eQTLs associated with expression of *HBG* in whole blood, other genes of the β-globin (*HBB*) gene cluster

and with known HbF cis- and trans-acting regulators. In addition, we examined genes co-expressed with *HBG1* to discover enriched pathways, studied upstream regulators of HbF co-expressed genes and performed correlation analysis between *HBG* and known HbF regulators. A similar correlation analysis included *BCL2L1*, a potential HbF activator we found in our analysis. We hypothesized that trans-acting elements that are transcription factors affect the expression of both *HBG2* and *HBG1*, while the known cis-acting QTL in the promoter of *HBG2* is likely to be associated with the expression of this gene only.

## Methods

### *Genome-wide eQTL association analysis*

Whole blood normalized RNA-seq data, 1000 Genome imputed genotypes, and the covariates file of 338 donors were downloaded from the GTEx portal version 6. The covariates used in the original GTEx data analysis included 3 genome-wide genotype principal components (PCs), 35 Probabilistic Estimation of Expression Residuals (PEER) factors to be used as confounders, genotyping platform (Illumina HiSeq 2000 or HiSeq X), and sex (Consortium 2017). We performed genome-wide eQTL analysis using Efficient and Parallelizable Association Container Toolbox (EPACTS) (EPACTS 2017) and selected SNPs with minor allele frequency (MAF) $\geq 0.01$ and imputation quality score (R^2) $\geq 0.4$ to reduce the rate of false association. We used the first 3 PCs to adjust for population substructure. We also used sex and platform as covariates to remove any bias introduced by these two components. The 35 PEER factors were used to adjust for batch effects and experimental confounders. This standard set of covariates were used in

the models to detect eQTLs for *HBG1*, *HBG2*, *BCL11A*, *KLF1*, *MYB*, *HBB*, and *HBD*. To detect any significant association after adjusting for the most significant SNPs on chromosomes 2, 6, and 11, we tested association of normalized expression of *HBG1* and *HBG2* conditioned on the genotypes of rs7482144, rs1427407, and rs66650371. Only SNPs that reached genome-wide significance level of p-value $\leq 5 \times 10^{-8}$ were considered statistically significant.

***Preliminary studies***

To reproduce the genetic association of known QTL variants with *HBG* expression using GTEx RNA-seq data from reticulocytes, three SNPs were selected based on their prior association with HbF levels in GWAS. We chose rs1427407 in the *BCL11A* erythroid-specific enhancer on chromosome 2p (Bauer et al. 2013), rs10128556, which is in high linkage disequilibrium (LD) with the Xmn1 restriction polymorphism (rs7482144) on chromosome 11p (Galarneau et al. 2010), and rs9399137 that is in perfect LD with the 3-base pair deletion (rs66650371) in the *HBS1L-MYB* intergenic polymorphisms (HMIP) region on chromosome 6q (Farrell et al. 2011). SNPs in LD with rs66650371and rs7482144 were chosen because GTEx pre-calculated eQTL did not contain these SNPs. The GTEx eQTL calculation tool was used to validate these associations in 338 whole blood samples available in the GTEx portal. We sought to validate the effect of a *BCL11A* erythroid-specific-enhancer SNP on *HBG2* and *HBG1* to confirm that genotype can be used to predict phenotype and vice versa.

***Association of the erythroid-specific BCL11A enhancer variant rs1427407 with whole blood gene expression***

We extracted the genotype of rs1427407 form GTEx data set version 6 and used these data to predict the impact of this SNP on global gene expression. The same covariates were used in the eQTL analysis, and Linear Models for Microarray and RNA-seq Data (limma) were used to perform linear regression (Ritchie et al. 2015). We used false discovery rate (FDR) to correct for multiple hypothesis testing.

***HBG1 differential co-expression analysis***

Since *HBG2* expression is significantly associated with the genotype of rs7482144, we used *HBG1* expression as a marker of expression of both γ-globin genes. A gene expression matrix of normalized log expression of RNA-seq along with covariates including 35 PEER factors and sex from 338 whole blood samples were used as input for limma to perform differential analysis. We used FDR to correct for multiple testing. Using Ingenuity Pathways Analysis (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/) we performed pathway analysis using the top100 co-expressed genes to see whether there was any enrichment in pathways involved in HbF regulation. In addition, we examined upstream regulators of the genes in the top100 *HBG1* co-expressed gene list to identify HbF transcriptional regulators
that can explain the observed gene expression changes.

***Correlation analysis between known and potential HbF regulators and HBG expression***

We performed Pearson correlation between *HBG* expression and the expression of *BCL11A*, *KLF1, MYB, ZBTB7A, SIRT1* and *BCL2L1* on 338 GTEx whole blood samples. These genes were selected because of their role or putative role as modulators of *HBG* expression, except for *BCL2L1* which was selected based on our *HBG1* co-expression and upstream regulators analysis (Dai et al. 2017; Jiang et al. 2006; Masuda et al. 2016; Zhou et al. 2010).

To replicate parts of our GTEx-based analysis, the same correlation analysis was done on data from erythroid samples acquired from Gene Expression Omnibus (GEO; accession number GSE59089) (Xu et al. 2015). These data contained RNA-seq derived from primary human fetal liver proerythroblasts studied at different developmental stages after shRNA-mediated knockdown of Polycomb Repressive Complex 2 (PRC2) core subunits.

**Results**

***Preliminary studies***

GTEx whole blood RNA-seq data detected HbF eQTLs in *BCL11A*, HMIP and in the *HBB* gene cluster using HbF gene expression as a surrogate for HbF protein levels, validating the utility of this data set for examining HbF-associated eQTL (**Table 10 and Fig. 8A-F)**.

***Genome-wide eQTL association analysis***

Twenty-seven SNPs were eQTL associated with *HBG1* (p-value <= 5 × 10−8). Of these 27, 21 were intergenic variants in the HMIP region on chromosome 6p, five were *BCL11A* variants on chromosome 2p, and one was on chromosome 11p (**Fig. 9A, Table S1**). The SNP rs66650371, the functional 3-bp deletion in the *MYB* enhancer, is one of the most significant SNPs in the HMIP region. The most significant SNP of the five *BCL11A* variants was rs1427407, which in other studies was the functional SNP in the erythroid-specific enhancer of this gene (Bauer et al. 2013).

Forty-nine SNPs meeting genome-wide significance levels were eQTL for *HBG2*. Seventeen were intergenic variants in the HMIP region on chromosome 6, five were *BCL11A* variants on chromosome 2p, 26 were on chromosome 11p (**Fig 9B, Table S2**) and one was on chromosome 1. The most significant SNP in chromosome 11 was rs7482144 and this was associated only with the expression of *HBG2*; rs16912979, which is located in hypersensitive site (HS) 4 of the locus control region (LCR) of the *HBB* gene complex, was also associated solely with *HBG2* expression. The SNP rs66650371 (*MYB*) and rs1427407 (*BCL11A*) were significantly associated with both *HBG2* and *HBG1*. The conditional analysis of *HBG1* showed that after adjusting for the effects of rs1427407, rs66650371, and rs7482144, there was no association on chromosome 2 or chromosome 6 (**Fig. 9C, Table S3**).

Conditional analysis of *HBG2* showed that no SNP was significantly associated with *HBG2* expression after adjusting for rs7482144, rs1427407, and rs66650371 (**Fig. S1, Table S4**). There was no SNP with genome-wide significance associated with expression of *BCL11A* and *ZBTB7A* (**Figs. S2, S3; Tables S5, S6**). One SNP on chromosome 6 was associated with *MYB* expression, another on chromosome 6 was associated with *HBD* expression, and one on chromosome 12 was associated with *KLF1* expression (**Figs. S4, S5, S6; Tables S7, S8, S9**). Variants on chromosome 1, 2, 6, 7, 14, and 20 were significantly associated with *HBB* expression (**Fig. S7, Table S10)**.

***Association of the erythroid-specific BCL11A enhancer variant rs1427407 with whole blood gene expression***

Linear regression analysis showed that the genotype of rs1427407 was significantly associated with *HBG1* and *HBG2* expression in whole blood (**Table 11**) after correction for multiple testing using FDR, p-values were 2.04E-06 and 3.63E-06 for *HBG1* and *HBG2*, respectively. This SNP was not a eQTL for any other gene expressed in peripheral blood.

***HBG1 differential co-expression analysis***

Twelve genes were differentially co-expressed with *HBG1* after FDR p-value adjustment (**Table S11**). The pathway enrichment analysis for the top100 co-expressed genes did not include any known pathway involved in HbF regulation among the top 10 enriched pathways. Of 452 upstream regulators that regulate the top100 *HBG1* co-expressed

genes, 360 regulated *BCL2L1* (**Table S12**). *GATA1*and *KLF1* were also in top five statistically significant upstream regulators.

***Correlation analysis between known and potential HbF regulators and HBG***

The correlation between *HBG1 and HBG2* normalized gene expression and the expression of *BCL11A*, *KLF1*, *MYB*, *ZBTB7A*, *SIRT1* and *BCL2L1* in the GTEx whole blood data set, showed that *BCL2L1* had correlation coefficients of 0.59 and 0.55 for *HBG1* and *HBG2*, respectively with a p value of 2.2E-16. (**Table 12**).

In primary human fetal liver proerythroblasts, *BCL2L1*, *KLF1*, and *SIRT1* were positively correlated with *HBG* expression with correlation coefficients of 0.93, 0.92, and 0.62, respectively; *BCL11A* and *MYB* were negatively correlated with *HBG1* with correlation coefficients of -0.63, and -0.64, respectively; *ZBTB7A* was weakly correlated with *HBG1*, and the p-value for association was not significant (**Table 13)**.

**Discussion**

Among eQTL significantly associated with *HBG* expression 21 were intergenic variants in HMIP on chromosome 6p. The functional 3-bp deletion (rs66650371) was one of the most significant eQTL in this *MYB* enhancer (Farrell et al. 2011; Stadhouders et al. 2014). The SNP rs66650371 was in strong LD with 20 other SNPs in chromosome 6p in Europeans, South and East Asians, and admixed Americans (**Table S13**). A long non-coding or lncRNA containing the site of rs66650371was transcribed from this enhancer.

Its downregulation in an erythroid cell line that expressed adult hemoglobin (HbA) was associated with a 200-fold increase in *HBG* expression and a 20-fold increase in HbF (Morrison et al. 2017). The eQTL at rs66650371 further supports the functional importance of this SNP. Five *BCL11A* SNPs in chromosome 2p were in very strong LD across populations (**Table S14**). The most significant, rs1427407, is a functional erythroid-specific enhancer variant that altered the expression of *BCL11A* (Bauer et al. 2013). Both rs1427407 and rs66650371are trans-acting elements and had the same magnitude of effect on *HBG1* and *HBG2* (**Fig. 8C, 8D; Fig. 10**). These observations are consistent with our hypothesis that trans-acting elements affect expression of both γ-globin genes.

The SNP rs10128556 in the pseudogene *HBBP1* on chromosome 11p was in strong LD with rs7482144 and was reported to have an effect independent of rs7482144 in African American sickle cell disease patients and be the likely functional variant modulating cis-acting HbF gene expression (Galarneau et al. 2010). However, rs7482144 is located within the *HBG2* promoter and highly linked to rs368698783 in *HBG1* promoter. The SNP rs368698783 together with rs7482144 are associated with reduced methylation in six CpG sites flanking the transcription start site of *HBG* (Chen et al. 2017). Moreover, rs7482144 alters a putative binding motif for ZBTB7A, a silencer of *HBG* expression (Masuda et al. 2016; Shaikho et al. 2016).

Strong evidence supporting differential regulation of *HBG* by cis- and trans-acting elements comes from clinical observations of patients with the Arab Indian (AI) and

Senegal haplotypes of the sickle hemoglobin (HbS) gene and from the reports of hereditary persistence of HbF (HPFH) caused by point mutations in *HBG* promoters. Patients with the AI and Senegal haplotypes, which are the only *HBB* haplotypes containing rs7482144, had increased levels of HbF that was predominantly of the $^{G}\gamma$-globin type (Ballas et al. 1991; Nagel et al. 1985; Rahimi et al. 2015). Many mutations have been described in the promoters of *HBG2* and *HBG1* that cause the phenotype of HPFH. For each of these mutations, depending on the affected gene, either HBG2 or HBG1usually comprises more than 90% of total γ globin (Wood 2001). The T-C mutations at -175 and -173 in the promoter of *HBG1* reactivated $^{A}\gamma$-globin gene expression in adult erythrocytes of transgenic mice while altering the binding of GATA-1 and Oct-1 and (Liu et al. 2005). Moreover, a C-T polymorphism that is identical to rs7482144 but at position -158 relative to *HBG1* affected *HBG1* expression (Patrinos et al. 1998).

These results suggest that rs7482144 rather than rs10128556 is either the functional cis-acting element regulating *HBG2* expression or the best surrogate for this element. In addition, rs16912979 in HS-4 of the LCR is also solely associated with *HBG2* further supporting the effect of cis-acting elements on a single γ-globin gene (**Fig. 11, 12**). This SNP is a member of the 3-SNP T/A/T haplotype (rs16912979, rs7482144, rs10128556) that is exclusive to the AI haplotype and is associated with their high HbF (Vathipadiekal et al. 2016). The T allele of rs16912979 tags a predicted binding site for runt-related transcription factor 1 (RUNX1) in the palindromic region of 5′ HS-4. RUNX1 plays an

important role in hematopoiesis and electromobility shift assays (EMSA) suggested an allele-specific binding of RUNX1 to 5′ HS-4 (Dehghani et al. 2016).

No common variants affected the expression of *BCL11A*. The *BCL11A* hypersensitive site variant rs1427407 is erythroid specific and in whole blood samples would impact *BCL11A* expression only in reticulocytes. However, *BCL11A* is expressed in leukocytes that are abundant in the blood. *BCL11A* expression in these nucleated cells with their greater transcriptional activity than reticulocytes would render an erythroid-specific signal impossible to detect. Our inability to find genes known to influence *HBG1* in the co-expression analysis and a lack of correlation of known HbF regulators with *HBG* in GTEx data despite a strong relationship in fetal liver proerythroblasts might have a similar explanation and illustrate the main limitation of using GTEx data where globin synthesizing reticulocytes were likely to be <1% of all red blood cells(Means RT 2009). Although few in number, reticulocytes are the only cells in the blood expressing globin genes, and RNA-seq provides sufficient data for most statistical analysis.

The unexpected correlation between *BCL2L1* and *HBG* expression (r 0.58 & 0.55; p=2.30E-16) was validated in fetal proerythroblasts (r> 0.9). *BCL2L1* is a member of bcl-2 gene family involved in anti-apoptotic activities with an important role in erythropoiesis. Induction of bcl-$x_L$ (*BCL2L1*) expression by erythropoietin and GATA1 was critical for the survival of late proerythroblasts and early normoblasts (Gregory et al. 1999).The upstream regulation analysis shows that *BCL2L1* was regulated by 80 percent

of the regulators that regulate the top 100 co-expressed genes. *GATA1* and *KLF13* are among these upstream regulators. These data and the correlation of *BCL2L1* with *HBG* suggest a role for *BCL2L1* in HbF regulation. Experimental validation to confirm the relationship between *BCL2L1* and *HBG* and to pinpoint the mechanism through which *BCL2L1* regulates *HBG* expression could lead to the discovery of potential new drug targets or molecules that can be used to induce HbF in sickle cell anemia and β thalassemia.

Cis- and trans-acting regulators have a differential effect on *HBG1 and HBG2* expression. The trans-acting eQTLs of *BCL11A* and *MYB* enhancers affect expression of both γ-globin genes. In contrast, only a single γ-globin gene is affected by the cis-acting eQTL. HbF is the major modulator of the severity of both sickle cell anemia and β thalassemia and considerable effort is being made to develop methods to increase *HBG* expression for therapeutic gain (Deng et al. 2014; Lettre and Bauer 2016). Although clinical observations suggest that an increase of HBG2 alone has some benefit (Alsultan et al. 2014; Ballas et al. 1991; Teixeira et al. 2003), modification of trans-acting elements that affect both HbF genes might have a greater effect on HbF levels and help force a more pancellular distribution that could be therapeutically important (Guda et al. 2015; Masuda et al. 2016; Sankaran et al. 2011; Steinberg et al. 2014)**.**

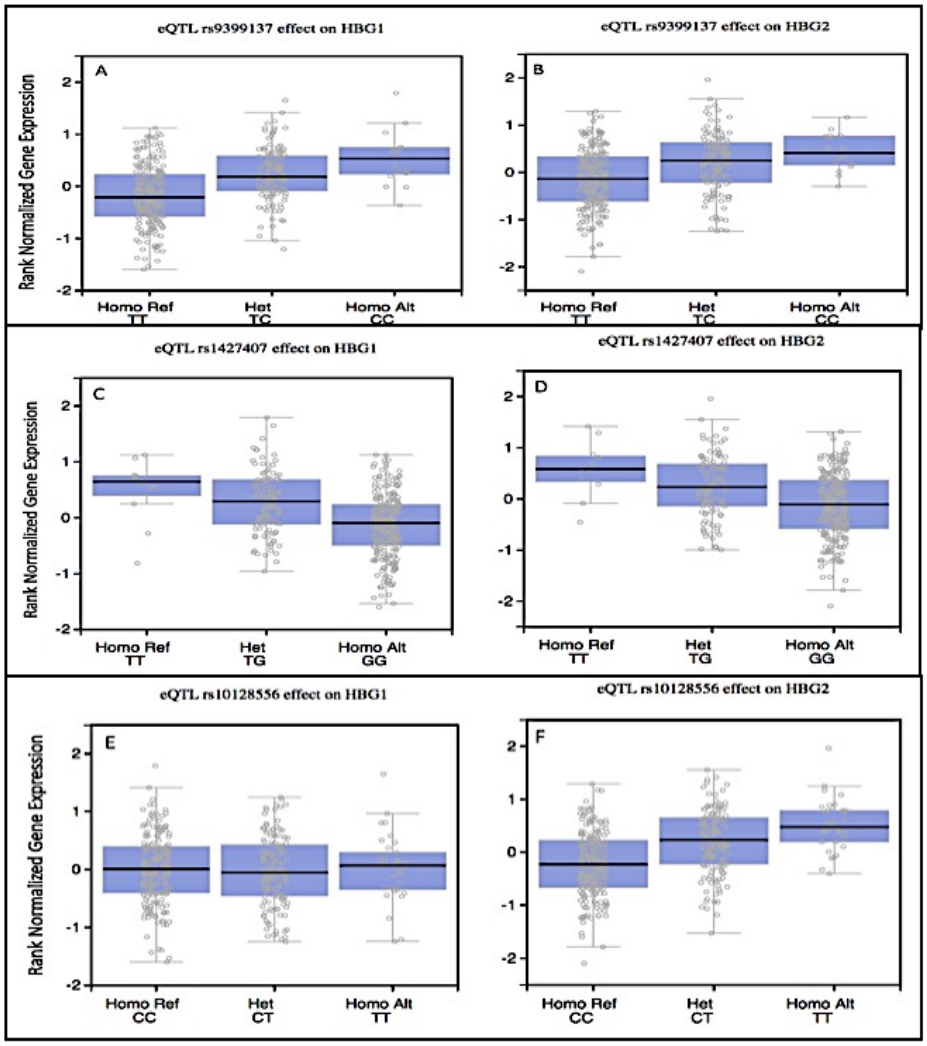



Figure 8. Effect of rs1427407 (2:60718043_T/G), rs10128556 (11:5263683_C/T), and rs9399137 (6:135418632_TTAC/T) on HBG1 and HBG2 expression in GTEx data set version 6. P-values are above genome-wide significance for rs1427407 (BCL11A, chr2) and rs9399137 (MYB, chr6) association with expression of both HBG genes. Rs10128556 (chr 11) is significantly associated with only HBG2 (p-value of 2.60E-16) while it has no effect on HBG1(p-value 0.89) Het denotes heterozygous, and Homo, homozygous.

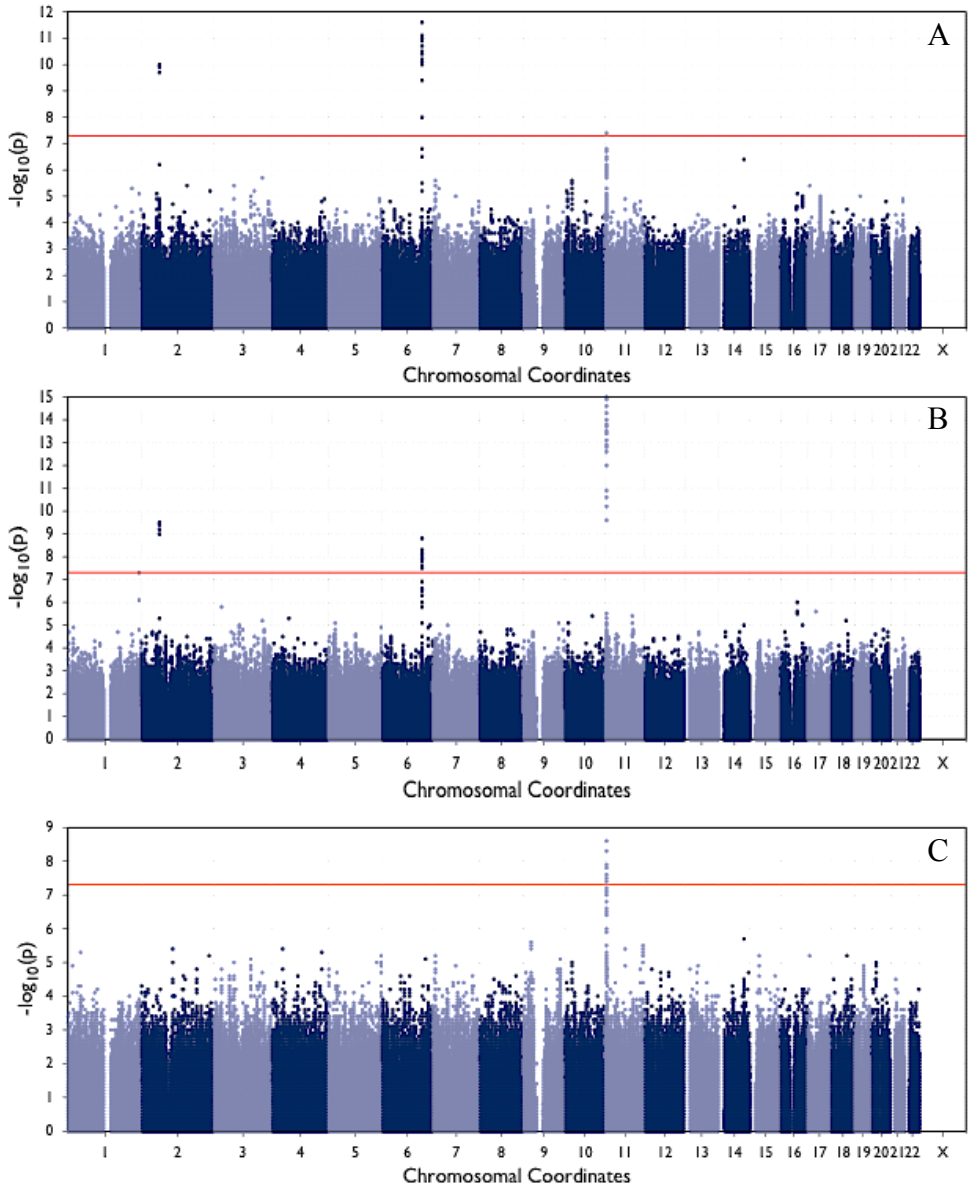



Figure 9. Manhattan plots for *HBG* eQTLs in 338 whole blood samples. Fig. 9A shows *HBG1* eQTL; Fig. 9B shows *HBG2* eQTL; Fig. 9C shows *HBG1* eQTL conditioned on rs66650371, rs1427407 and rs7482144 genotype. The red line indicates genome-wide significance levels

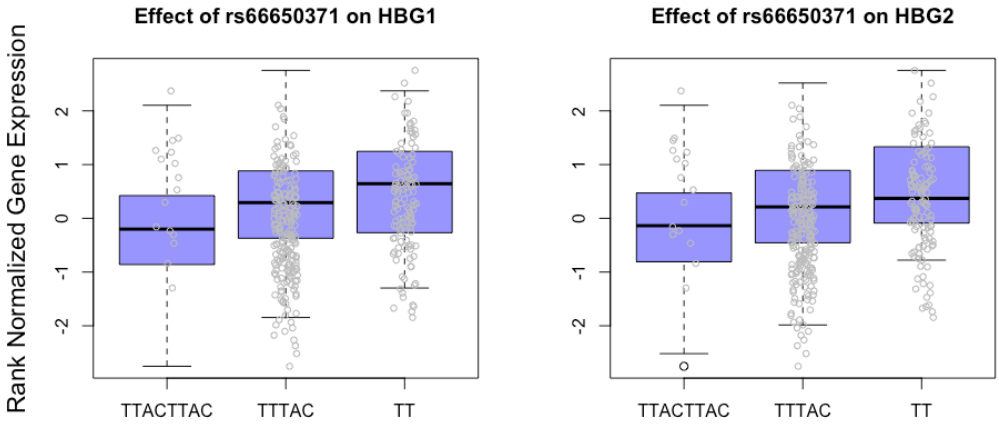



Figure 10. Effect of rs66650371 (6:135418632_TTAC/T) genotypes on *HBG1* and *HBG2* expression in 338 whole blood samples. P-values of genome-wide association are 6.49E-11 and 5.86E-09 for *HBG1* and *HBG2*, respectively.

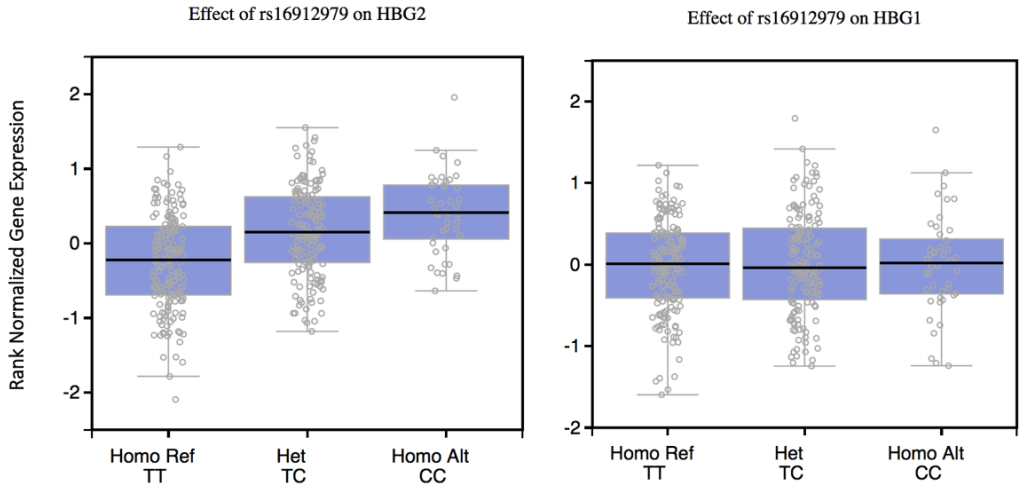



Figure 11. Effect of rs16912979 (11_5309695_T_C) on *HBG2* and *HBG1* expression in GTEx data set. p-values of genome-wide significance are 7.0e-14 and 0.77, respectively. Het denotes heterozygous, and Homo, homozygous.

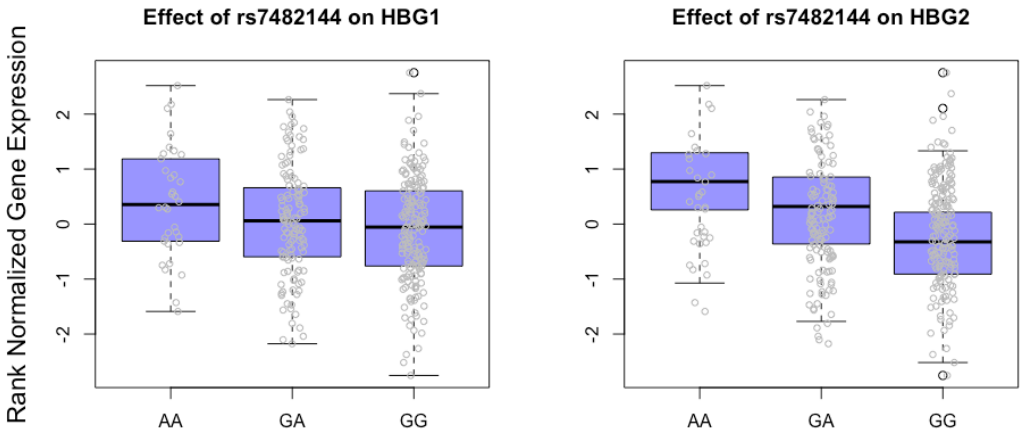



Figure 12. Effect of rs7482144 (11:5276169_G/A) genotypes on *HBG1* and *HBG2* expression in 338 whole blood samples. P-values of genome-wide association are 0.038593 and 9.49E-16 for *HBG1* and *HBG2*, respectively.

| Gene Symbol | Variant Id | SNP | P-Value | Effect Size | T-Statistic | Standard Error |
|---|---|---|---|---|---|---|
| ***HBG1*** | 2_60718043_T_G_b37 | rs1427407 | 1.20E-11 | -0.46 | -7.1 | 0.065 |
| ***HBG1*** | 11_5263683_C_T_b37 | rs10128556 | 0.89 | -0.0085 | -0.14 | 0.059 |
| ***HBG1*** | 6_135419018_T_C_b37 | rs9399137 | 2.20E-15 | 0.51 | 8.4 | 0.061 |
| ***HBG2*** | 2_60718043_T_G_b37 | rs1427407 | 1.10E-09 | -0.46 | -6.3 | 0.073 |
| ***HBG2*** | 11_5263683_C_T_b37 | rs10128556 | 2.60E-16 | 0.51 | 8.7 | 0.058 |
| ***HBG2*** | 6_135419018_T_C_b37 | rs9399137 | 1.70E-09 | 0.44 | 6.2 | 0.07 |

Table 10. SNPs from previous GWAS and their association with *HBG1* and *HBG2* expression in whole blood samples available in the GTEx portal.

| ENS_ID | Gene Symbol | logFC | AveExpr | P.Value | adj.P.Val |
|---|---|---|---|---|---|
| ENSG00000213934.5 | HBG1 | 0.502491942 | -4.32E-06 | 8.51E-11 | 2.04E-06 |
| ENSG00000196565.8 | HBG2 | 0.507932715 | -7.56E-06 | 3.03E-10 | 3.63E-06 |
| ENSG00000187017.10 | ESPN | 0.307905604 | -1.45E-07 | 1.80E-05 | 0.143668299 |
| ENSG00000239405.1 | TMED10P2 | 0.362195583 | 0.002468525 | 7.07E-05 | 0.423972382 |
| ENSG00000010626.10 | LRRC23 | 0.277170914 | 2.20E-07 | 9.30E-05 | 0.427249648 |
| ENSG00000248309.1 | MEF2C-AS1 | 0.320432287 | -9.14E-08 | 0.000106933 | 0.427249648 |
| ENSG00000160401.10 | CFAP157 | -0.342869883 | -8.98E-07 | 0.000164101 | 0.561998024 |
| ENSG00000135414.5 | GDF11 | -0.163176468 | -3.51E-08 | 0.000286461 | 0.777799594 |
| ENSG00000255318.1 | - | -0.358828046 | 0.000665361 | 0.000316391 | 0.777799594 |

Table 11. Effect of rs1427407 on whole blood gene expression. ENS_ID is ensemble gene ID, LogFC is log fold change, aveExpr is average expression, and adj.P.Val is FDR adjusted p-value.

| Regulators | *HBG1* Correlation | P-value | *HBG2* Correlation | P-value |
|---|---|---|---|---|
| *BCL2L1* | 0.59 | 2.20E-16 | 0.55 | 2.20E-16 |
| *KLF1* | 0.47 | 2.20E-16 | 0.44 | 2.20E-16 |
| *MYB* | 0.27 | 2.94E-07 | 0.27 | 2.8E-07 |
| *BCL11A* | 0.15 | 0.003914 | 0.17 | 0.001694 |
| *SIRT1* | -0.06 | 0.238 | -0.06 | 0.29 |
| *ZBTB7A* | -0.0045 | 0.9334 | 0. 067 | 0.212 |

Table 12. Pearson correlation between *HBG* and known or potential HbF regulators using RNA-seq from 338 whole blood samples from GTEx.

| Regulators | *HBG1* Correlation | P-value | *HBG1* Correlation | P-value |
|---|---|---|---|---|
| *BCL2L1* | 0.93 | 3.85E-07 | 0.93 | 3.85E-07 |
| *KLF1* | 0.92 | 9.24E-07 | 0.92 | 9.24E-07 |
| *MYB* | -0.64 | 0.01 | -0.63 | 0.01 |
| *BCL11A* | -0.63 | 0.01 | -0.63 | 0.01 |
| *SIRT1* | 0.62 | 0.01 | 0.62 | 0.01 |
| *ZBTB7A* | -0.38 | 0.17 | -0.38 | 0.16 |

Table 13. Pearson correlation between *HBG* and known or potential HbF regulators using RNA-seq primary human fetal liver proerythroblasts.

# Chapter 5. Conclusion

## *A phased SNP-based classification of sickle cell anemia HBB haplotypes*

The main advantage of the phased SNP-based haplotype determination method is the rapid classification and high accuracy. This method can also be used for whole genome sequence data classification after SNP calling and phasing. Moreover, it is not sensitive to SNPs that alter the restriction enzyme recognition sequence that can lead to error using RFLPs. A limitation of the phased SNP-based haplotype determination method is the dependency on the availability of GWAS data for many SNPs in the β-globin gene region. However, the cost of genome-wide SNP arrays has fallen dramatically and haplotype ascertainment by GWAS is likely to be less expensive than traditional methods. Also, patient cohorts with genome-wide SNP array data or whole genome sequencing data are increasing available. The method is very fast and more accurate than the RFLP method and very useful in generating haplotype information that can be used as a covariate in a genetic risk analysis.

## *Genetic Determinants of HbF in Saudi Arabian and African Benin Haplotype Sickle Cell Anemia*

Based on the GWAS analysis for Saudi and African American Benin haplotype, homozygosity for a *BCL11A* enhancer haplotype of T (rs1427407), C (rs6706648) and G (rs7606173) in the enhancer elements +62, +58 and +55, was associated with 10 % HbF in African American. Homozygosity for the GTC haplotype of these same three SNPs

was associated with 4.5% HbF in African American Benin and had a frequency of 0.40 in African American Benin, and 0.11 in Saudi Benin patients. In another study that used focused genotyping of the *BCL11A* enhancer SNPs, enhancer haplotypes associated with lower HbF were almost three times as common in the African Benin haplotype compared with the Saudi Benin haplotype (Sebastiani et al. 2015). Even when accounting for the population frequency differences of *BCL11A* enhancer and the 3-base pair deletion in the *HBS1L-MYB* intergenic polymorphisms (HMIP) haplotypes, *BCL11A* and HMIP variants did not explain the difference in HbF levels in Saudi Benin and African America Benin sickle cell anemia. The small number of Saudi Benin haplotype cases studied diminished the power of this study to find novel associations, particularly if the minor allele frequency was low and the effects of the variant on HbF were small.

***Cis and Trans-Acting Regulators of HbF Expression***

A sub-haplotype of rs16912979, rs7119428, and rs7482144 of the seven SNPs located within putative ZBTB7A binding motifs differentiates AI, Senegal, and Benin haplotypes. Homozygosity for minor alleles at rs16912979, rs7119428, and rs7482144 (T/A/T) is exclusive to the AI haplotype, and this sub-haplotype might represent a functional cis-acting domain modulating *HBG2* expression. Rs16912979 in the HS-4 region has strong binding signals for GATA1, GATA2, and POLR2A. Senegal haplotype carriers are homozygous for the T allele of rs7482144 but do not have the minor allele of rs16912979 and rs7119428 (C/C/T), while Benin haplotype is (C/C/C). Perhaps this divergence accounts in part for the differences in HbF between Senegal and AI

haplotypes. In summary, SNPs in putative ZBTB7A binding sites distinguish the AI haplotype from African-origin *HBB* haplotypes. As these variants are present in regions with characteristics of active enhancers, like transcription factor binding and epigenetic marks, they are candidates for the functional elements of this haplotype. Perhaps this haplotype alters looping of the LCR to globin gene promoters. It must be emphasized however that mechanistic studies are required to prove the functionality of these variants.

eQTL analysis of γ-globin genes expression suggests a differential effect of cis- and trans-acting HbF regulators. The trans-acting eQTLs (rs1427407 in the *BCL11A* enhancer and the HMIP eQTL rs66650371) influence the expression both *HBG* genes; cis-acting eQTLs (rs7482144 and rs16912979) on chromosome 11) only affect *HBG2*.

The strong correlation between *BCL2L1* and *HBG* expression in the validation data set (primary human fetal liver proerythroblasts) and the upstream regulation analysis suggest that BCL2L1 is a potential new HbF regulatory element.

***Current Work and Future Directions***

To begin further studies on the effects of B*CL2L1* on *HBG* expression we examined expression levels of this gene and percent HbF in sickle cell anemia. HbF expression was obtained by qRT-PCR analysis of mRNA isolated from cultured erythroid progenitor cells from eight patients. The preliminary results show a good correlation between *BCL2L1* mRNA and *HBG* mRNA (R2 0.72, $p < 0.05$).

These promising preliminary studies suggest future work aimed at experimental and functional validation of *BCL2L1* and its effect on *HBG*, and the mechanism through which *BCL2L1* regulates HbF gene expression. Using such methods as overexpression and knockdown of *BCL2L1* in primary erythroid progenitors and cell lines that synthesize adult and fetal hemoglobins will help us better define and understand the role of this gene in modulating *HBG* expression and possibly targeting this gene for therapeutic purposes.

## BIBLIOGRAPHY

Akinsheye I, Alsultan A, Solovieff N, Ngo D, Baldwin CT, Sebastiani P, Chui DH, Steinberg MH (2011) Fetal hemoglobin in sickle cell anemia. Blood 118: 19-27. doi: 10.1182/blood-2011-03-325258

Alsultan A, Alabdulaali MK, Griffin PJ, Alsuliman AM, Ghabbour HA, Sebastiani P, Albuali WH, Al-Ali AK, Chui DH, Steinberg MH (2014) Sickle cell disease in Saudi Arabia: the phenotype in adults with the Arab-Indian haplotype is not benign. British Journal of Haematology 164: 597-604. doi: 10.1111/bjh.12650

Alsultan A, Solovieff N, Aleem A, AlGahtani FH, Al-Shehri A, Elfaki Osman M, Kurban K, Bahakim H, Kareem Al-Momen A, Baldwin CT (2011) Fetal hemoglobin in sickle cell anemia: Saudi patients from the Southwestern province have similar HBB haplotypes but higher HbF levels than African Americans. American Journal of Hematology 86: 612-614.

Antonarakis SE, Boehm CD, Giardina PJ, Kazazian HH, Jr. (1982) Nonrandom association of polymorphic restriction sites in the beta-globin gene cluster. Proceedings of the National Academy of Sciences of the United States of America 79: 137-41.

Ballas SK, Talacki CA, Adachi K, Schwartz E, Surrey S, Rappaport E (1991) The Xmn I site (-158, C----T) 5' to the G gamma gene: correlation with the Senegalese haplotype and G gamma globin expression. Hemoglobin 15: 393-405.

Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC, Pinello L, Sabo PJ, Vierstra J, Voit RA, Yuan GC, Porteus MH, Stamatoyannopoulos JA, Lettre G, Orkin SH (2013) An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science 342: 253-7. doi: 10.1126/science.1242088

Bhagat S, Patra PK, Thakur AS (2013) Fetal Haemoglobin and β-globin Gene Cluster Haplotypes among Sickle Cell Patients in Chhattisgarh. Journal of clinical and diagnostic research: JCDR 7: 269.

Bhanushali AA, Patra PK, Pradhan S, Khanka SS, Singh S, Das BR (2015) Genetics of fetal hemoglobin in tribal Indian patients with sickle cell anemia. Translational Research 165: 696-703. doi: https://doi.org/10.1016/j.trsl.2015.01.002

Casbon J (2017) PyVCF https://github.com/jamescasbon/PyVCF/. Accessed 20 May 2017

Chang Y, Smith KD, Moore R, Serjeant GR, Dover GJ (1995) An analysis of fetal hemoglobin variation in sickle cell disease: the relative contributions of the X-linked factor, beta-globin haplotypes, alpha-globin gene number, gender, and age. Blood 85: 1111-1117.

Chen D, Zuo Y, Zhang X, Ye Y, Bao X, Huang H, Tepakhan W, Wang L, Ju J, Chen G, Zheng M, Liu D, Huang S, Zong L, Li C, Chen Y, Zheng C, Shi L, Zhao Q, Wu Q, Fucharoen S, Zhao C, Xu X (2017) A Genetic Variant Ameliorates beta-Thalassemia Severity by Epigenetic-Mediated Elevation of Human Fetal

Hemoglobin Expression. American Journal of Human Genetics 101: 130-138. doi: 10.1016/j.ajhg.2017.05.012

Consortium GT (2017) Genetic effects on gene expression across human tissues. Nature 550: 204. doi: 10.1038/nature24277

https://www.nature.com/articles/nature24277 - supplementary-information

Craig JE, Rochette J, Fisher CA, Weatherall DJ, Marc S, Lathrop GM, Demenais F, Thein S (1996) Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. Nature Genetics 12: 58-64. doi: 10.1038/ng0196-58

Cui S, Tanabe O, Lim K-C, Xu HE, Zhou XE, Lin JD, Shi L, Schmidt L, Campbell A, Shimizu R (2014) PGC-1 coactivator activity is required for murine erythropoiesis. Molecular and Cellular Biology 34: 1956-1965.

Cui S, Tanabe O, Sierant M, Shi L, Campbell A, Lim K-C, Engel JD (2015) Compound loss of function of nuclear receptors Tr2 and Tr4 leads to induction of murine embryonic β-type globin genes. Blood 125: 1477-1487.

Dai Y, Chen T, Ijaz H, Cho EH, Steinberg MH (2017) SIRT1 activates the expression of fetal hemoglobin genes. American Journal of Hematology 92: 1177-1186. doi: 10.1002/ajh.24879

Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C (2016) Next-generation genotype imputation service and methods. Nature Genetics 48: 1284-1287. doi: 10.1038/ng.3656

http://www.nature.com/ng/journal/v48/n10/abs/ng.3656.html - supplementary-information

Dehghani H, Ghobakhloo S, Neishabury M (2016) Electromobility Shift Assay Reveals Evidence in Favor of Allele-Specific Binding of RUNX1 to the 5′ Hypersensitive Site 4-Locus Control Region. Hemoglobin 40: 236-239. doi: 10.1080/03630269.2016.1189931

Deng W, Rupon JW, Krivega I, Breda L, Motta I, Jahn KS, Reik A, Gregory PD, Rivella S, Dean A, Blobel GA (2014) Reactivation of developmentally silenced globin genes by forced chromatin looping. Cell 158: 849-860. doi: 10.1016/j.cell.2014.05.050

EPACTS (2017) EPACTS: Efficient and Parallelizable Association Container Toolbox. https://github.com/statgen/EPACTS.

Farrell JJ, Sherva RM, Chen ZY, Luo HY, Chu BF, Ha SY, Li CK, Lee AC, Li RC, Yuen HL, So JC, Ma ES, Chan LC, Chan V, Sebastiani P, Farrer LA, Baldwin CT, Steinberg MH, Chui DH (2011) A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. Blood 117: 4935-45. doi: 10.1182/blood-2010-11-317081

Flint J, Harding RM, Boyce AJ, Clegg JB (1998) The population genetics of the haemoglobinopathies. Bailliere's Clinical Haematology 11: 1-51.

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

Garner C, Mitchell J, Hatzis T, Reittie J, Farrall M, Thein SL (1998) Haplotype mapping of a major quantitative-trait locus for fetal hemoglobin production, on chromosome 6q23. American Journal of Human Genetics 62: 1468-74. doi: 10.1086/301859

Green NS, Fabry ME, Kaptue-Noche L, Nagel RL (1993) Senegal haplotype is associated with higher HbF than Benin and Cameroon haplotypes in African children with sickle cell anemia. American Journal of Hematology 44: 145-146.

Gregory T, Yu C, Ma A, Orkin SH, Blobel GA, Weiss MJ (1999) GATA-1 and Erythropoietin Cooperate to Promote Erythroid Cell Survival by Regulating bcl-xL Expression. Blood 94: 87-96.

GTEx_Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nature Genetics 45: 580-5. doi: 10.1038/ng.2653

Guda S, Brendel C, Renella R, Du P, Bauer DE, Canver MC, Grenier JK, Grimson AW, Kamran SC, Thornton J, de Boer H, Root DE, Milsom MD, Orkin SH, Gregory RI, Williams DA (2015) miRNA-embedded shRNAs for Lineage-specific BCL11A Knockdown and Hemoglobin F Induction. Molecular Therapy 23: 1465-74. doi: 10.1038/mt.2015.113

Habara A, Steinberg MH (2016) Minireview: Genetic basis of heterogeneity and severity in sickle cell disease. Experimental Biology and Medicine (Maywood, N.J.) 241: 689-96. doi: 10.1177/1535370216636726

Ingram VM (1957) Gene Mutations in Human Hæmoglobin: the Chemical Difference Between Normal and Sickle Cell Hæmoglobin. Nature 180: 326. doi: 10.1038/180326a0

Jiang J, Best S, Menzel S, Silver N, Lai MI, Surdulescu GL, Spector TD, Thein SL (2006) cMYB is involved in the regulation of fetal hemoglobin production in adults. Blood 108: 1077-83. doi: 10.1182/blood-2006-01-008912

Joly P, Lacan P, Garcia C, Delasaux A, Francina A (2011) Rapid and reliable beta-globin gene cluster haplotyping of sickle cell disease patients by FRET Light Cycler and HRM assays. Clinica Chimica Acta 412: 1257-61. doi: 10.1016/j.cca.2011.03.025

Kato GJ, Gladwin MT, Steinberg MH (2007) Deconstructing sickle cell disease: Reappraisal of the role of hemolysis in the development of clinical subphenotypes. Blood Reviews 21: 37-47. doi: 10.1016/j.blre.2006.07.001

Kato GJ, Steinberg MH, Gladwin MT (2017) Intravascular hemolysis and the pathophysiology of sickle cell disease. The Journal of clinical investigation 127: 750-760.

Kulozik AE, Wainscoat JS, Serjeant GR, Kar BC, Al-Awamy B, Essan GJF, Falusi AG, Haque SK, Hilali AM, Kate S, Ranasinghe W, Weatherall DJ (1986) Geographical survey of βS-globin gene haplotypes: Evidence for an independent Asian origin of the sickle-cell mutation. American Journal of Human Genetics 39: 239-44.

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

Lettre G, Sankaran VG, Bezerra MA, Araujo AS, Uda M, Sanna S, Cao A, Schlessinger D, Costa FF, Hirschhorn JN, Orkin SH (2008) DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proceedings of the National Academy of Sciences of the United States of America 105: 11869-11874.

Liu LR, Du ZW, Zhao HL, Liu XL, Huang XD, Shen J, Ju LM, Fang FD, Zhang JW (2005) T to C substitution at -175 or -173 of the gamma-globin promoter affects GATA-1 and Oct-1 binding in vitro differently but can independently reproduce the hereditary persistence of fetal hemoglobin phenotype in transgenic mice. Journal of Biological Chemistry 280: 7452-9. doi: 10.1074/jbc.M411407200

Loh P-R, Palamara PF, Price AL (2016) Fast and accurate long-range phasing in a UK Biobank cohort. Nature Genetics 48: 811-816. doi: 10.1038/ng.3571
http://www.nature.com/ng/journal/v48/n7/abs/ng.3571.html - supplementary-information

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867-2873. doi: 10.1093/bioinformatics/btq559

Masuda T, Wang X, Maeda M, Canver MC, Sher F, Funnell AP, Fisher C, Suciu M, Martyn GE, Norton LJ, Zhu C, Kurita R, Nakamura Y, Xu J, Higgs DR, Crossley M, Bauer DE, Orkin SH, Kharchenko PV, Maeda T (2016) Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. Science 351: 285-9. doi: 10.1126/science.aad3312

Means RT GB, General Considerations., Greer JP, Foerster J, Rodgers GM, eds (2009) Anemia. Wintrobe's Clinical Hematology, 12th edn, vol Vol. 1. Lippincott Williams and Wilkins, Philadelphia, PA

Morrison TA, Wilcox I, Luo HY, Farrell JJ, Kurita R, Nakamura Y, Murphy GJ, Cui S, Steinberg MH, Chui DHK (2017) A long noncoding RNA from the HBS1L-MYB intergenic region on chr6q23 regulates human fetal hemoglobin expression. Blood Cells, Molecules, and Diseases 69: 1-9. doi: 10.1016/j.bcmd.2017.11.003

Mtatiro SN, Singh T, Rooks H, Mgaya J, Mariki H, Soka D, Mmbando B, Msaki E, Kolder I, Thein SL, Menzel S, Cox SE, Makani J, Barrett JC (2014) Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. PloS One 9: e111464. doi: 10.1371/journal.pone.0111464

Nagel RL, Fabry ME, 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. New England Journal of Medicine 312: 880-4. doi: 10.1056/nejm198504043121403

Nan Liu P, Victoria V Hargreaves, PhD, Jiyoung Hong, PhD, Woojin Kim, PhD, Jesse Kurland, PhD, Qian Zhu, PhD, Falak Sher, PhD, Claudio Macias-Trevino, BS, Jill Rogers, PhD, Guo-Cheng Yuan, PhD, Daniel E. Bauer, MD, PhD, Jian Xu, PhD, Martha L Bulyk, PhD and Stuart Orkin, MD. (2017) Fetal Hemoglobin (HbF) Silencer BCL11A Acts through a Novel DNA-Motif in the Gamma-Globin

Promoters, Simplifying the Model for Hemoglobin Switching. ASH, Atalanta, GA

Ngo D, Bae H, Steinberg MH, Sebastiani P, Solovieff N, Baldwin CT, Melista E, Safaya S, Farrer LA, Al-Suliman AM, Albuali WH, Al Bagshi MH, Naserullah Z, Akinsheye I, Gallagher P, Luo HY, Chui DH, Farrell JJ, Al-Ali AK, Alsultan A (2013) Fetal hemoglobin in sickle cell anemia: genetic studies of the Arab-Indian haplotype. Blood Cells, Molecules, and Diseases 51: 22-6. doi: 10.1016/j.bcmd.2012.12.005

Ngo DA, Steinberg MH (2015) Genomic approaches to identifying targets for treating beta hemoglobinopathies. BMC Medical Genomics 8: 44. doi: 10.1186/s12920-015-0120-2

Park S, Gianotti-Sommer A, Molina-Estevez FJ, Vanuytsel K, Skvir N, Leung A, Rozelle SS, Shaikho EM, Weir I, Jiang Z, Luo HY, Chui DHK, Figueiredo MS, Alsultan A, Al-Ali A, Sebastiani P, Steinberg MH, Mostoslavsky G, Murphy GJ (2017) A Comprehensive, Ethnically Diverse Library of Sickle Cell Disease-Specific Induced Pluripotent Stem Cells. Stem Cell Reports 8: 1076-85. doi: 10.1016/j.stemcr.2016.12.017

Patrinos GP, Kollia P, Loutradi-Anagnostou A, Loukopoulos D, Papadakis MN (1998) The Cretan type of non-deletional hereditary persistence of fetal hemoglobin [A gamma-158C-->T] results from two independent gene conversion events. Human Genetics 102: 629-34.

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genetics 2: e190. doi: 10.1371/journal.pgen.0020190

Perrine RP, Brown MJ, Clegg JB, Weatherall DJ, May A (1972) Benign sickle-cell anaemia. The Lancet 300: 1163-1167. doi: http://dx.doi.org/10.1016/S0140-6736(72)92592-5

Piel FB, Steinberg MH, Rees DC (2017) Sickle Cell Disease. New England Journal of Medicine 376: 1561-1573. doi: 10.1056/NEJMra1510865

Powars DR (1991) Beta s-gene-cluster haplotypes in sickle cell anemia. Clinical and hematologic features. Hematology/Oncology Clinics of North America 5: 475-93.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559-75. doi: 10.1086/519795

Pysam-developers/pysam. https://github.com/pysam-developers/pysam/. Accessed May 13 2017

Rahimi Z, Vaisi-Raygani A, Merat A, Haghshenass M, Rezaei M (2015) Level of Hemoglobin F and Gg Gene Expression in Sickle Cell Disease and Their Association with Haplotype and XmnI Polymorphic Site in South of Iran

Rees DC, Williams TN, Gladwin MT (2010) Sickle-cell disease. The Lancet 376: 2018-2031. doi: 10.1016/s0140-6736(10)61029-x

Revelle W (2017) psych: Procedures for Psychological, Psychometric, and Personality Research. https://cran.r-project.org/package=psych. Accessed 21 May 2017

Rezende PV, Costa KS, Domingues Junior JC, Silveira PB, Belisário AR, Silva CM, Viana MB (2016) Clinical, hematological and genetic data of a cohort of children with hemoglobin SD. Revista Brasileira de Hematologia e Hemoterapia 38: 240-246. doi: 10.1016/j.bjhh.2016.05.002

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43: e47. doi: 10.1093/nar/gkv007

Sankaran VG, Menne TF, Šćepanović D, Vergilio J-A, Ji P, Kim J, Thiru P, Orkin SH, Lander ES, Lodish HF (2011) MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13. Proceedings of the National Academy of Sciences. doi: 10.1073/pnas.1018384108

Sankaran VG, Menne TF, Xu J, Akie TE, Lettre G, Van Handel B, Mikkola HK, Hirschhorn JN, Cantor AB, Orkin SH (2008) Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science 322: 1839-1842.

Sankaran VG, Weiss MJ (2015) Anemia: progress in molecular mechanisms and therapies. Nature Medicine 21: 221-230.

Schroeder WA, Huisman TH, Shelton JR, Shelton JB, Kleihauer EF, Dozy AM, Robberson B (1968) Evidence for multiple structural genes for the gamma chain of human fetal hemoglobin. Proceedings of the National Academy of Sciences of the United States of America 60: 537-44.

Sebastiani P, Farrell J, Alsultan A, Wang S, Edward H, Shappell H, Bae H, Milton J, Baldwin C, Al-Rubaish A (2015) BCL11A enhancer haplotypes and fetal hemoglobin in sickle cell anemia. Blood Cells, Molecules, and Diseases 54: 224-230.

Shaikho EM, Habara AH, Alsultan A, Al-Rubaish A, Al-Muhanna F, Naserullah Z, Alsuliman A, Qutub HO, Patra P, Sebastiani P (2016) Variants of ZBTB7A (LRF) and its β-globin gene cluster binding motifs in sickle cell anemia. Blood cells, molecules & diseases 59: 49.

Solovieff N, Milton JN, Hartley SW, Sherva R, Sebastiani P, Dworkis DA, Klings ES, Farrer LA, Garrett ME, Ashley-Koch A, Telen MJ, Fucharoen S, Ha SY, Li CK, Chui DH, Baldwin CT, Steinberg MH (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. doi: 10.1182/blood-2009-08-239517

Stadhouders R, Aktuna S, Thongjuea S, Aghajanirefah A, Pourfarzad F, van Ijcken W, Lenhard B, Rooks H, Best S, Menzel S, Grosveld F, Thein SL, Soler E (2014) HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. The Journal of clinical investigation 124: 1699-710. doi: 10.1172/JCI71520

Steinberg MH (2008) Sickle Cell Anemia, the First Molecular Disease: Overview of Molecular Etiology, Pathophysiology, and Therapeutic Approaches. The Scientific World Journal 8. doi: 10.1100/tsw.2008.157

Steinberg MH, Chui DH, Dover GJ, Sebastiani P, Alsultan A (2014) Fetal hemoglobin in sickle cell anemia: a glass half full? Blood 123: 481-485.

Steinberg MH, Forget BG, Higgs DR, Weatherall DJ (2009) Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management, 2 edn. Cambridge University Press, Cambridge

Sutton M, Bouhassira EE, Nagel RL (1989) Polymerase chain reaction amplification applied to the determination of beta-like globin gene cluster haplotypes. American Journal of Hematology 32: 66-9.

Teixeira SM, Cortellazzi LC, Grotto HZ (2003) Effect of hydroxyurea on G gamma chain fetal hemoglobin synthesis by sickle-cell disease patients. Brazilian Journal of Medical and Biological Research 36: 1289-92.

Terasawa T, Ogawa M, Porter P, Karam J (1980) G gamma and A gamma globin-chain biosynthesis by adult and umbilical cord blood erythropoietic bursts and reticulocytes. Blood 56: 93-97.

The Genomes Project C (2015) A global reference for human genetic variation. Nature 526: 68-74. doi: 10.1038/nature15393

http://www.nature.com/nature/journal/v526/n7571/abs/nature15393.html - supplementary-information

Thein SL, Wainscoat JS, Sampietro M, Old JM, Cappellini D, Fiorelli G, Modell B, Weatherall DJ (1987) Association of thalassaemia intermedia with a beta-globin gene haplotype. British Journal of Haematology 65: 367-73.

Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, Usala G, Busonero F, Maschio A, Albai G, Piras MG, Sestu N, Lai S, Dei M, Mulas A, Crisponi L, Naitza S, Asunis I, Deiana M, Nagaraja R, Perseu L, Satta S, Cipollina MD, Sollaino C, Moi P, Hirschhorn JN, Orkin SH, Abecasis GR, Schlessinger D, Cao A (2008) Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proceedings of the National Academy of Sciences of the United States of America 105: 1620-1625.

Vathipadiekal V, Alsultan A, Baltrusaitis K, Farrell JJ, Al-Rubaish A, Al-Muhanna F, Naserullah Z, Alsuliman A, Patra P, Milton JN, Farrer LA, Chui DHK, Al-Ali AK, Sebastiani P, Steinberg MH (2016) Homozygosity for a Haplotype in the HBG2-OR51B4 Region is Exclusive to Arab-Indian Haplotype Sickle Cell Anemia. American Journal of Hematology 91: E308-11. doi: 10.1002/ajh.24368

Walters MC, Patience M, Leisenring W, Eckman JR, Scott JP, Mentzer WC, Davies SC, Ohene-Frempong K, Bernaudin F, Matthews DC, Storb R, Sullivan KM (1996) Bone Marrow Transplantation for Sickle Cell Disease. New England Journal of Medicine 335: 369-376. doi: 10.1056/nejm199608083350601

Wood W (2001) Hereditary Persistence of Fetal Hemoglobin and Thalassemia, in, Disorders of Hemoglobin: Genetics, Pathophysiology and Clinical Management, Chapter15, Steinberg MH, Forget BG, Higgs DR, Nagel RL eds. Cambridge University Press, pp 356-388

Xu J, Shao Z, Li D, Xie H, Kim W, Huang J, Taylor JE, Pinello L, Glass K, Jaffe JD, Yuan GC, Orkin SH (2015) Developmental control of polycomb subunit

composition by GATA factors mediates a switch to non-canonical functions. Molecular Cell 57: 304-316. doi: 10.1016/j.molcel.2014.12.009

Zhou D, Liu K, Sun C-W, Pawlik KM, Townes TM (2010) KLF1 regulates BCL11A expression and γ- to β-globin gene switching. Nature Genetics 42: 742. doi: 10.1038/ng.637

https://www.nature.com/articles/ng.637 - supplementary-information

# CURRICULUM VITAE